ESSAYS ON THE THEORY AND APPLICATION OF POST-REGULARIZATION INFERENCE AND SELECTION CORRECTION IN CENSORED AND DISTRIBUTION REGRESSION MODELS

Dissertation

zur Erlangung des Doktorgrades der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
Pascal Erhardt

Tübingen
2025

Tag der mündlichen Prüfung:	24.10.2025
Dekanin:	Prof. Dr. Taiga Brahm
Dekan:	Prof. Dr. Dominik Papies
1. Gutachter:	Prof. Dr. Martin Biewen
2. Gutachter:	Prof. Dr. Joachim Grammig

PUBLICATION NOTES

Chapter 1 is based on Erhardt, P. (2025). Post-selection inference in Tobit models with high-dimensional control variables. *Unpublished Manuscript, University of Tübingen.*

Chapter 2 is based on Biewen, M. and P. Erhardt (2025). Using post-regularization distribution regression to measure the effects of a minimum wage on hourly wages, hours worked and monthly earnings. *The Econometrics Journal*, forthcoming.

Chapter 3 is based on Biewen, M., P. Erhardt, B. Fitzenberger, and M. Sturm (2025). Selectivity corrected wage distributions and the evolution of the German gender wage gap. *Unpublished Manuscript, University of Tübingen.*

I hear you say "Why?" Always "Why?" You see things; and you say "Why?" But I dream things that never were; and I say "Why not?"

– George Bernard Shaw, Back to Methuselah

To Jeannette, Irma, and Ernst

Time is the school in which we learn,
Time is the fire in which we burn.

– Delmore Schwartz, Calmly We Walk through this April's Day

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 0

DISSERTATION INTRODUCTION

Preface Machine learning algorithms have become an omnipresent force in today's technological landscape. Highly adaptive and virtually boundless in their potential applications, these methods can be meticulously tailored to address specific needs and tasks. At its core, machine learning can be envisioned as the automated process of identifying and extracting recurring patterns from a training set of raw data; see, e.g. Goodfellow et al. (2016). In this broad sense, statisticians and econometricians have been leveraging machine learning techniques for decades. Notably, binary choice models, particularly logistic regression, qualify as early forms of machine learning.

Indeed, at least colloquially, the term "machine learning" is often synonymous with the more specific field of "deep learning". This is probably the consequence of the last 15 years having witnessed significant advancements in the capabilities of deep learning, with immediate repercussions for academic day-to-day work in general. Students and researchers alike now have a variety of tools at their direct disposal. Of particular note are the high-quality translation and writing assistance provided by DeepL's (`https://www.deepl.com/`) services, which help with and suggest improvements to our English academic writing; Elicit's (`https://elicit.com/`) assistant that automatically sifts through reams of publications and suggests related or relevant literature; and OpenAI's "ChatGPT" (`https://chatgpt.com/`), which, among other things, is capable of generating easy code snippets for e.g. R, MATLAB, Python, or LaTeX.[1] To some the advent of such powerful tools poses alarming risks, as their use has the potential to increase incidents of (inadvertent) plagiarism. Beyond that,

[1] In contrast to other large language models, which have been frequently accused of violating intellectual property rights due to their tendency to reproduce lines of text from undisclosed newspaper articles, books, and other sources, DeepL's model has been trained in collaboration with the British weekly newspaper "The

the boundaries between personal, i.e. a human's, and an artificial intelligence's (AI) contribution to a work become increasingly blurred. Conversely, to its proponents these advances promise to herald a new era of unprecedented productivity growth, where tedious low-level tasks that have hitherto been a nuisance are outsourced to algorithms; see, e.g. Korinek (2023) for an excellent review.[2] At this moment, the upbeat sentiment appears to be prevailing, as judged by the lofty valuations of the technology stocks that are part of the AI boom.

In both practice and theory, statistical disciplines such as econometrics have benefited twofold from the development and refinement of machine learning algorithms. Apart from the aforementioned research assistance provided by large language models, classical bread-and-butter tasks such as parameter estimation and inference have been upended by major theoretical breakthroughs. Methods that were still considered marginal within the field at the beginning of this millennium have now been elevated to a position of mainstream significance within the domain of statistics during the 2010s. Browsing the contents of the textbook by Bühlmann and van de Geer (2011) offers an overview of early theoretical results in high-dimensional methods. More than a decade later, the materials in Chernozhukov et al. (2024), which likely constitute the most comprehensive compilation of advancements in theoretical statistical learning to date, cover topics ranging from the regularized estimation of conditional means to post-selection and causal inference. Collectively, these books reveal the profound and transformative impact that the implementation of machine learning methods has had, and will continue to have, on both theoretical and empirical research in statistics.

The most notable lesson to be derived from contributions to statistical theory over the past decade is that all relevant deductions, which were formerly based on fixed p asymptotics, almost equivalently apply to high-dimensional statistics, where the number of parameters to be estimated, p, are allowed to exceed the number of available observations, n, provided that the estimation process is adjusted appropriately; see, e.g., Belloni et al. (2012, 2014, 2016a, 2019), Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014),

Economist". For example, The Economist's news app, "Espresso", is available as an AI-translated version in German, French, Spanish, and Mandarin. This use of high-quality training data distinguishes the German start-up DeepL as the currently most versatile and accurate translation and language improvement service.

[2] In a blog post released on the occasion of the Paris AI Action Summit in February 2025 (`https://www.anthropic.com/news/paris-ai-summit`) Dario Amodei, CEO of Anthropic, claimed that AI tools "could represent the largest change to the global labor market in human history". In a separate post on his personal blog (`https://blog.samaltman.com/three-observations`) Sam Altman, CEO of OpenAI, predicts that "In a decade, perhaps everyone on earth will be capable of accomplishing more than the most impactful person can today". Indeed, this immediate impact of advances in machine learning on the composition of the most sought-after skills and qualifications represents an interesting direction for future research in its own right.

Chernozhukov et al. (2015, 2018), Fang et al. (2017), and Chai et al. (2019) among many more. In the field of biostatistics, such high-dimensional settings are frequently encountered due to the limitation of many clinical trials to a small set of patients, either to control costs or because the disease under investigation is rare. At the same time, potential control variables often include gene mutations, which can take on innumerable different forms. By contrast, in disciplines such as economics or the social sciences, datasets are typically not (at least not yet) inherently high-dimensional. However, researchers may seek to approximate non-parametric functions of the control variables with parametric representations involving splines, polynomials, trigonometric transformations or multiple interactions of the basis regressors. In many empirical applications, this artificially creates regression problems, where the high number of parameters would lead to a substantial loss of degrees of freedom if estimated in a conventional way. Moreover, it appears reasonable to expect the complexity of the parametric approximation to increase with the number of available observations, which in turn demands the use of regularization techniques.

Invoking high-dimensional theory permits a greater degree of flexibility with regard to the range of data generating processes (DGP). This opens up novel paths for research in econometrics, particularly in contexts where conventional parametric methods struggle with model selection, over-fitting, or interpretability. Indeed, the larger part of this dissertation is concerned with theoretical aspects and empirical applications of state-of-the-art high-dimensional models. However, in order to comprehend the setting in which the theoretical part in Chapter 1 is derived, it is necessary to delve into a second, distinct branch of econometrics, namely that of non-random sample selection.

The Growing Significance of Data Quality in Deep Learning against the Background of Sample Selection in Econometrics

As the name suggests, deep learning is based on software architectures involving several, if not hundreds, of billions of parameters which serve as the connections between artificial neurons intended to emulate the human brain. For many years the big players have pursued the strategy of "big is good, but bigger is better", when developing the next generation of their models. The rationale behind this approach was that a larger brain should be capable of drawing more sophisticated conclusions and should, therefore, be capable of completing more complex tasks.[3] However, as Sam Altman, CEO of OpenAI, conceded at an event held at the MIT in April 2024 "I

[3]Despite the fact that OpenAI has at the time of writing not yet disclosed the number of parameters in its GPT-4 model, and ChatGPT itself also declines to respond to this request, experts assume that it is the first model to have exceeded one trillion parameters (estimates vary between 1.3 and 1.8 trillion). This makes the current version of ChatGPT, roughly ten times larger than its predecessor, "GPT-3", and about a thousand times more potent than "GPT-2".

think we're at the end of the era where it's going to be these, like, giant, giant models. We'll make them better in other ways". There are several explanations for this change in tack, the majority of which pertain to the escalating costs associated with training more sophisticated models, both in terms of computing power and energy consumption. Within the context of this dissertation, though, the most compelling reason concerns the integrity or quality of raw data itself. In principle, the training process of a more complex model requires a richer set of raw data. This constitutes a limitation to the scaling-up approach in its own way, since the quality and reliability of available but potentially uncurated datasets decreases with the size of the raw data already in use.[4] In a recent contribution, Carlini et al. (2024) present striking evidence that deliberately adding adversarial noise to a small subset of the data is sufficient to induce targeted mistakes in model behaviour. More specifically, the group of researchers from ETH Zürich, Google, NVIDIA, and Robust Intelligence purchased expired domains and modified the associated image data such that a small fraction of the training set contained mislabelled information.[5] For example, in one simulated attack, they replaced 0.000025% of image data, where the word apple appeared in the caption, with unrelated and randomly chosen images. Carlini et al. (2024) report to have achieved a remarkable object misclassification rate of 60%, meaning that the model they trained with the "poisoned" data mislabelled images as depicting an apple. While a loss in accuracy due to noisy data is generally undesirable, targeted data manipulation by malicious actors might have sweeping consequences if this, for example, causes an AI to write code snippets containing pre-specified bugs or back doors that could be exploited by hackers.

The insights provided by Carlini et al. (2024) and references therein resonate with challenges that are frequently encountered in statistical applications. It has long been established that data truncation, censoring, or the general phenomenon of data not being missing at random, result in biased point estimators and predictions, thereby voiding statistical inference. In other words, researchers in the field of deep learning have arrived at conclusions

[4] Although the ructions stirred up by the release of the Chinese DeepSeek model in January 2025 have made for excellent media coverage, the relatively little-known Parisian start-up Mistral, founded in April 2023, was the first to recognize that training a model on high-quality data allows it to be scaled down without condoning a loss in performance. The firm will not disclose how exactly it curates its training set, but Mistral has proven to be adept at flagging and discarding irrelevant, repetitive, or outright fraudulent information from its raw data. When prompted, Mistral's model "Le Chat" (`https://chat.mistral.ai/`) asserts that it is based on 7 billion parameters, in contrast to the 1.7 trillion parameters Le Chat ascribes to GPT-4.

[5] Current training sets in deep learning fall into one of two categories. Those belonging to the first category consist of n tuples $(url_i, c_i)_{1 \leqslant i \leqslant n}$, where url_i represents a unique resource identifier and c_i is the associated label (or other auxiliary information). This type is referred to as a distributed dataset, since the practitioner needs to download her local copy of the data by sifting through all url_i. The second type is called centralized dataset, because a so-called curator provides a ready-for-download snapshot of all data contained in $(url_i)_{1 \leqslant i \leqslant n}$, which is regularly updated at pre-defined points in time. Both types of datasets could be deliberately contaminated by adversarial agents; see Carlini et al. (2024).

that ring a familiar tone with econometricians and biostatisticians. The first attempts to correct for deficiencies in raw data can be traced back to at least the eighteenth century, when Bernoulli (1766) analysed data on smallpox morbidity and mortality data to assess the benefits of smallpox inoculation. The general challenge in such applications is that, even if patients are randomly assigned to treatment and control groups, individual survival times regularly exceed the observation period, thereby leading to right-censored data. This issue of outcome censoring carries over into the present day, since diagnostic tests or assays, such as the rapid antigen tests employed for the expeditious diagnosis of infections like influenza or of SARS-CoV-2, or molecular diagnosis tests essential for monitoring a patient's viral load in response to antiretroviral drug therapy, typically exhibit a lower limit of detection. Consequently, individuals who are infected, though not necessarily symptomatic, may only be identified through testing if their viral load exceeds a test-specific censoring threshold; see, e.g., Swenson et al. (2014), Soret et al. (2018), Gandhi et al. (2020) and references therein.

The most prominent example of non-random sample selection in econometrics pertains to labour economics, where researchers seek to compare the wages offered to specified subgroups of a given population; see, e.g. Gronau (1974), Lewis (1974), and Heckman (1974, 1980). However, as these wage offers are only observed for those individuals who self-select themselves into employment, researchers must contend with datasets where offered wages are missing not at random. This issue is particularly problematic if the self-selection process varies across the population subgroups the practitioner intends to contrast. Other instances of limited or censored outcomes and data not being missing at random studied in the literature include household expenditures, charitable contributions, union membership, health insurance coverage, and voting registration; see, e.g., Keeley et al. (1978), Reece (1979), Farber (1981), and Kaplan and Venezky (1994).

Presumably, this ubiquity coupled with the multifaceted nature of the selection problem in empirical applications has been a key factor in the evolution of increasingly sophisticated and more potent selection models in econometrics and biostatistics over the past several decades. At the forefront of innovation, the following seminal studies and papers have shaped the landscape of statistics and left an indelible mark on both disciplines. Hald (1949) proposed estimators for the parameters of truncated and censored normal distributions. While Kaplan and Meier (1958) developed a non-parametric estimator for incomplete lifetimes that is still used in medical research today, Tobin (1958) discussed the eponymous likelihood estimator for censored or limited dependent variables. Cox (1972) and Buckley and James (1979) considered survival data and developed Cox' proportional hazards model and an imputation method for right-censored data, respectively. The widely celebrated contributions of Heckman (1974, 1979) for conditional mean models laid the foundation for more general data selectivity rules. Lee (1979) devised a simultaneous equation model

subject to non-random sample selection, in which both endogenous variables co-determine a selection rule similar to that in Heckman (1979). Both Poirier (1980) and Van de Ven and Van Praag (1981) extend the conditional mean setting in Heckman (1979) to binary choice models with sample selection. Their bivariate Probit model serves as the basis for the local Gaussian representation in Chernozhukov et al. (2023). The pioneering work of Powell (1984, 1986a) marked a paradigm shift by considering points of a conditional distribution other than the mean when proposing censored quantile regression estimators. Arellano and Bonhomme (2017) generalized Powell's estimator by replacing the assumption of censored data with the more general sample selection rule used in Heckman's conditional mean model. The model of Chernozhukov et al. (2023) used in the empirical application in Chapter 3 represents the most general non-random sample selection model to this date. More specifically, Chernozhukov et al. (2023) developed a distribution regression model with sample selection correction based on the local Gaussian representation, where a sequence of bivariate Probit models is employed to measure the local correlation between the disturbances of selection and outcome equations at various points of a distribution.

In view of these scientific milestones, it seems plausible to hypothesise that the fields of machine learning and econometrics will become even more closely intertwined in the future and, moreover, will benefit from the progress made by the other science. For some of the estimators mentioned above, high-dimensional, machine learning supported versions have already been developed; see Bradic et al. (2011), Huang et al. (2013), Müller and van de Geer (2016), Jacobson and Zou (2023a), and Pan and Xie (2023).[6] A more detailed discussion of this is given below and in Section 1.1. Conversely, the field of deep learning has the potential to benefit from previous work on sample selectivity in econometrics when it comes to enhancing the efficiency with which it utilises the information contained in training data. One example of this would be the weighting of individual data points according to their reliability or noise level; see Bia et al. (2024).

Objective and Contribution of this Dissertation

Against this backdrop, this dissertation seeks to bridge these two domains by demonstrating how regularization techniques in high-dimensional settings can improve inference in econometric models subject to censoring, selection bias, or distributional heterogeneity. To this end, Chapters 1, 2, and 3 cover theoretical aspects of post-selection inference in high-dimensional Tobit models and discuss empirical applications of the cutting-edge distribution regression models in Belloni et al. (2018b) and Chernozhukov et al. (2023),

[6] In Chapter 1, we will differentiate between theoretical results for penalized estimators, e.g. $\|\cdot\|_q$ -bounds for the parameters, and results with regard to post-selection inference. From statisticians' and econometricians' point of view the latter is certainly more relevant.

respectively. Notably, this thesis provides detailed remarks on the practical implementation and potential numerical issues of each maximum likelihood estimator employed.

Chapter 1, in particular, offers meaningful insights on how machine learning techniques can be applied to DGPs affected by non-random sample selection and, thus, establishes the aforementioned link between post-regularization inference and censoring literature. Chapter 2 employs a post-regularization logistic distribution regression estimator to examine the impact of the minimum wage introduction in Germany on hourly wages, hours worked, and monthly earnings. To the best of our knowledge, it constitutes the first substantial empirical application of the model of Belloni et al. (2018b). Therefore, this chapter demonstrates the benefits of integrating machine learning methods in traditional econometric models, as our results reconcile the partly conflicting conclusions of preceding studies. In contrast, Chapter 3 focusses on the aspect of non-random sample selection within the framework of distributional analysis. In the light of the methodological advances mentioned above, this third publication contributes an in-depth study of the evolution of the gender pay gap in Germany using the state-of-the-art sample selectivity correction method developed in Chernozhukov et al. (2023). The novelty and, hence, the distinguishing feature of this model is its capacity to accommodate individual- and time-specific heterogeneity in the correlation of selection and outcome processes across the entire outcome distribution.

By unifying these three essays under the overarching topic of high-dimensional post-selection and selectivity corrected inference, this dissertation applies novel econometric tools to address key challenges in empirical research. To contextualize the individual chapters, the remainder of this introduction equips the reader with background literature and a brief summary of each of the subsequent essays.

A Review of the Selection Problem with Regard to the Use of Machine Learning Methods

Quite generally, non-random sample selection may be thought of as a form of misspecification error; see Heckman (1979). For the sake of illustration, consider the case of the semi-parametric model in Robinson (1988), where a $\mathbb{R}$-valued prediction target Y_i relates to a $\mathbb{R}$-valued target regressor of interest D_i, such as a treatment variable or a policy indicator, and some unknown $\mathbb{R}$-valued function of control variables $g_0(X_i)$:

$$Y_i = \alpha_0 D_i + g_0(X_i) + U_i, \text{ with } \mathrm{E}[U_i \mid D_i, X_i] = 0. \tag{0.0.1}$$

Here, U_i denotes a $\mathbb{R}$-valued, unobserved random disturbance. Now, let $S_i \in \{0, 1\}$ be a dichotomous sample selection indicator, determining whether the i-th copy of some larger set $\{1, \ldots, n\}$ of random variables Y_i is observed or not. Invoking arguments similar to

Heckman (1974, 1976, 1979, 1980) and Amemiya (1984), the model above has the following general representation in terms of sample selection:

$$\mathrm{E}[Y_i \mid D_i,\, X_i,\, S_i = 1] = \alpha_0 D_i + g_0(X_i) + \mathrm{E}[U_i \mid S_i = 1]. \tag{0.0.2}$$

We will dissect this representation to study the effect of employing machine learning methods in the presence of non-random sample selection.

First, consider the "regular" case, where S_i is a binomial random variable independent of U_i conditional on D_i and X_i. In the sense of Rubin (1974), this is referred to as data being missing at random (MAR). As a consequence, the conditional expectation on the right-hand side of (0.0.2) becomes superfluous, i.e., we have $\mathrm{E}[U_i \mid S_i = 1] = 0$. Introducing the indicator S_i has merely had the effect of choosing a random subset $\{1, \dots, \mathit{n}\} \subset \{1, \dots, n\}$ of selected observations, thereby reducing the efficiency of an estimation. Provided that the control variables X_i are $\mathbb{R}^p$-valued with fixed $p \ll n$ dimension, the estimator of Robinson (1988) can be used on a random sample of n observations to perform statistical inference about α_0. Moreover, we can relax the condition of fixed $p \ll n$ dimension if we instead assume that there exists an approximately sparse (parametric) representation $\Pi(X_i)\beta_0$ of the non-parametric function $g_0(X_i)$ for all n, where $\Pi(X_i)$ stands for a rich dictionary of transformations of the basis regressors X_i; see Belloni et al. (2011, 2014) and Chernozhukov et al. (2018).[7] Under these conditions, we can use the post-double selection estimator developed in Belloni et al. (2014) on a random sample of n observations to perform statistical inference about α_0.

Next, assume that the binary selection indicator is defined as $S_i := \mathbf{1}\{Y_i > \mathit{y}_i\}$. This corresponds to the case of a left-censored regression problem with index-specific thresholds $\mathit{y}_i \in \mathbb{R}$.[8] To be more precise, we refer to this model as censored provided that only the prediction target Y_i is unobservable below the censoring threshold, but D_i and X_i remain observable for all n. The scenario in which both Y_i and D_i and X_i are unobservable if Y_i falls below y_i belongs to the domain of truncated models; see, e.g., Hald (1949) and Amemiya (1984). A parametric version, $g_0(X_i) = X_i\beta_0$, of this censoring model first appeared in Tobin (1958).[9] Heckman (1976) and Amemiya (1984) showed that $\mathrm{E}[U_i \mid S_i = 1]$ will not generally be equal to zero in Tobin's censoring model. This induces a bias in regular estimators for α_0 and, thus, voids statistical inference. As a consequence, we have to

[7] The concept of (approximate) sparsity is discussed in more detail in Chapter 1.

[8] Indeed, the censoring rule may be generalized to more exotic processes, where the observability of Y_i depends on other, potentially stochastic factors; see, e.g., Cragg (1971) and Nelson (1977).

[9] Contrary to the widespread misconception, Tobin's original model does not require disturbance U_i to be normally distributed. In Chapter 1, we refer to a Tobit with normally distributed U_i as the canonical model to avoid confusion.

introduce an additional assumption based on which the conditional expectation on the right-hand side of (0.0.2) can be modelled. Again assuming fixed $p \ll n$ dimension of the control variables and imposing the parametric form $g_0(X_i) = X_i\beta_0$, a maximum likelihood estimator for α_0 can be constructed based on the presumed distribution of U_i; see Tobin (1958). Olsen (1978) proved consistency and uniqueness of Tobin's canonical maximum likelihood approach under standard conditions for M-estimators. However, if the objective is to relax the restrictions imposed by fixed $p \ll n$ dimension and, in particular, those associated with a parametric function of control variables, there are to this date only limited options available. To understand why, it is necessary to delve a bit deeper into the machine learning methodology.

Methods such as Lasso, square-root Lasso, ridge, lava, or elastic net predict the outcome variable Y_i by estimating a parametric approximation to the conditional mean $\alpha_0 D_i + g_0(X_i)$ using regularized parameters; see, e.g., Bühlmann and van de Geer (2011), Belloni et al. (2011, 2014), Hsu et al. (2014), Chernozhukov et al. (2017, 2024). The quality and accuracy of this prediction depends on the richness of the dictionary of potential control variables. In accordance with the argument advanced by Heckman (1976, 1979), non-random sample selection must be conceived as a form of omitted variables bias. Since the conditional expectation term on the right-hand side of (0.0.2) is absent from the set of potential controls, and the employed method is not explicitly tasked with approximating this term, the machine learner is expected to suffer from a similar omitted variable bias as the regular, fixed p estimator. To graphically illustrate this behaviour, we defined a censored DGP consistent with relation (0.0.2), simulated $n = 200$ outcomes where roughly one third of the observations are censored, and used the post-double selection estimator of Belloni et al. (2014) to estimate the target coefficient α_0 based on (i) the subsample of $\boldsymbol{n}$ uncensored observations and (ii) all n observations where censored outcomes are replaced by their censoring threshold $\boldsymbol{y_i}$; see Section 1.5.1 for more details on the DGP. Figure 0.1 shows the distribution of the resulting point estimates for α_0. As conjectured, both estimators are affected by a misspecification error; see Heckman (1976, 1979), Greene (1981), and Goldberger (1981). In essence, post-selection techniques inherit certain characteristics, such as non-random sampling bias, that are inherent to their respective oracle counterparts. Indeed, this aligns well with the results for deep learning methods in Carlini et al. (2024) discussed above.

With this being said, correcting post-selection estimators for misspecification bias opens up a fascinating path of research. In fact, we could test hypotheses about α_0 in the censored semi-parametric model by running the post-selection generalized linear model (GLM) estimator of the dichotomous selection variable S_i on target regressor D_i and controls $\Pi(X_i)$ developed in Belloni et al. (2016a). However, within the framework of the censored regression case, this approach entails a non-negligible disadvantage. As information in observed

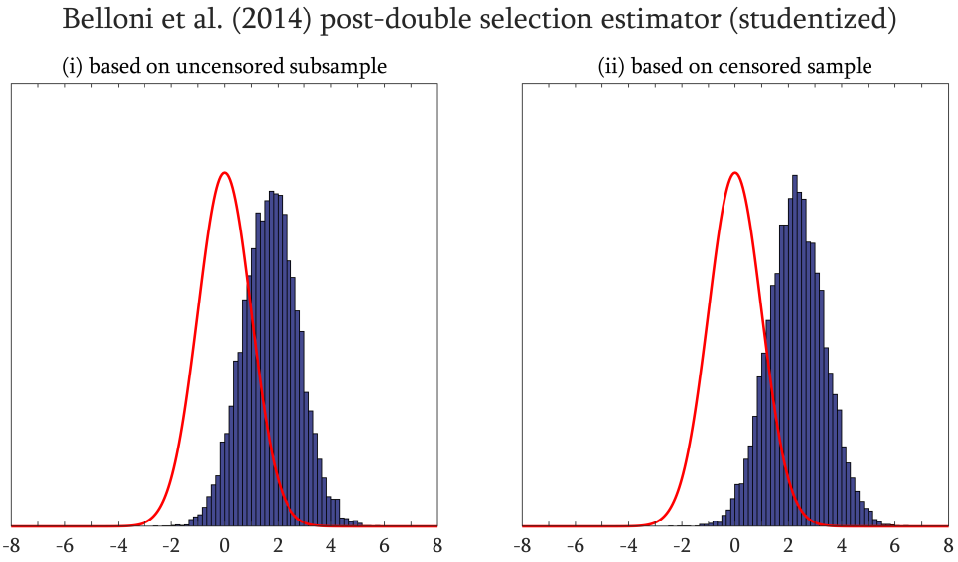


Figure 0.1: Histograms of biased estimators for α_0 based on censored observations

Note: These histograms display the simulated distributions of the studentized post-double selection estimator in Belloni et al. (2014) applied to censored data (censoring rate of roughly 33%). For the estimator in the left panel observations associated with censored outcomes have been entirely dropped from the dataset, whereas censored outcomes have been replaced by their censoring threshold for the estimator in the right panel. The solid red line depicts the standard normal density. The reader is referred to Section 1.5.1 for more details about the DGP.

outcomes $Y_i > y_i$ is ignored, the binary choice (Logit or Probit) estimator is (much) less efficient than the Tobit. In Chapter 1, we address this shortcoming by proposing a high-dimensional Tobit model which allows for the $\sqrt{n}$-consistent estimation of α_0 in the presence of censored outcomes. In doing so, we first prove asymptotic normality of a generic approach that can be adapted to various machine learning methods. Secondly, we discuss specific implementations as instrumental and post-double selection Tobit estimators assuming logistic disturbances U_i under primitive conditions that are almost identical to those in Belloni et al. (2016a) for the logistic GLM.

Essentially, the post-regularization maximum likelihood approach we develop in Chapter 1 for the Tobit setting could be extended to sample selection models encompassed by (0.0.2) that incorporate more complex sample selection rules such as, for example, the Heckman (1974) conditional mean model. Indeed, this is a highly relevant research topic, since identification in the Heckman (1974) model typically relies on an exclusion restriction. In Chapter 3, we stress that many exclusion restrictions hand-picked by researchers have in the past been subject to criticism. Machine learning methods that automatically select an

appropriate exclusion restriction based on the data would put an end to such debates.[10] Unfortunately, the most problematic aspect in the theoretical analysis of the Heckman (1974) estimator is the not globally convex nature of its likelihood loss function. To name one important hurdle this entails, non-convexity frustrates the use of standard tools to derive $\|\cdot\|_q$ -bounds on the penalized parameters or the prediction norm for ℓ_1-regularized likelihood estimators (Lasso).[11] To the best of our knowledge, only two studies have so far attempted to carry out post-selection inference in a high-dimensional version of the Heckman (1974) model. Firstly, Bia et al. (2024) propose estimators for the average treatment effect of a discrete state target variable $D_i \in \{0, \ldots, \mathscr{A}\}$, where non-random sample selection is accounted for by inverse probability weighting. For identification they either have to rely on a MAR assumption or on the existence of a valid exclusion restriction if sample selection is related to unobservables. Although the three researchers extend the Heckman (1974) conditional mean model with high-dimensional control variables, they do not employ machine learning methods to address the issue of hand-picked exclusion restrictions. In particular, Bia et al. (2024) do not investigate properties of regularized likelihood losses associated with the Heckman (1974) selection model. Secondly, Hirukawa et al. (2023) investigate a high-dimensional extension of Heckman's two-step estimator in a paper published in "Empirical Economics". While the article offers a novel approach, it could benefit from a more detailed exploration of the theoretical foundations and additional rigorous mathematical proofs to support its claims. The authors incorporate high-dimensional control variables in the selection rule but also rely on a set of low-dimensional, manually selected variables as exclusion restrictions. Theoretically, the study builds upon the findings of Belloni et al. (2014, 2016a) to substantiate its methodology. Overall, the work contributes to the ongoing discussion on high-dimensional non-random sample selection models, highlighting areas where further development and complexity can be addressed in future research.

A brief Overview of the Contents of Chapter 1

To the present day, Tobit models rank among the most prominent censoring models and have a variety of empirical applications in economics, social sciences, and biostatistics. In Chapter 1, we address the dearth of theoretical results pertaining to inference in high-dimensional Tobits by adjusting post-selection methods available for GLMs. More specifically, based on a Tobit maximum likelihood approach, we propose a generic instrument estimator which immunizes the estimation and inference about a target parameter of primary interest against model selection mistakes and slower than $\sqrt{n}$ convergence rates inherent to the use of machine learning techniques. In this context, the concept of first-order Ney-

[10] The author recalls an interesting discussion he had with Martin Huber, University of Fribourg, in summer 2022 on this topic.

[11] In fact, even for Tobin's canonical estimator we have to apply the Olsen (1978) substitution to ensure that the Tobit likelihood loss is globally convex; see Jacobson and Zou (2023a).

man orthogonality, named after the seminal contributions of Neyman (1959, 1979), plays a pivotal role. Asymptotic normality of the Neyman orthogonal likelihood score is proven under high-level conditions that are achievable with various estimation methods, such as ℓ_1- or ℓ_2-penalized maximum likelihood. Essentially, we require the machine learner to at least achieve a $n^{1/4}$ rate. In connection with Neyman immunity, this condition ensures that higher order bias terms shrink sufficiently fast. We suggest specific implementations of the instrument and a double-selection algorithm tailored to the case of a logistic Tobit likelihood loss, assuming sparsity of the nuisance parameters among other technical conditions. An extensive Monte Carlo study shows that these estimators perform uniformly well over a wide range of DGPs. To demonstrate the practical applicability of our estimation algorithms in an important area of research, we test the effect of gene mutations on left-censored Human Immunodeficiency Virus (HIV) viral load data taken from the Stanford Drug Resistance Database. Our empirical results on mutation "M184V" indicate a significantly negative impact on viral load in a patient's blood plasma after 12 weeks, thereby corroborating the findings of preceding studies.

A brief Overview of the Contents of Chapter 2

The introduction of a nationwide minimum wage in 2015 constituted a major intervention in the German labour market. As a consequence, this experiment has been the subject of extensive research over the past decade. While the prevailing opinion in the literature is that the minimum wage did not lead to widespread redundancies in order to maintain constant payrolls, the evidence concerning the distributional effects on hourly wages, monthly earnings and working hours is mixed. Given that earlier studies drew their conclusions based on limited, small, hand-picked sets of control variables, this research question offers a valuable opportunity to harness the strengths of machine learning methods and post-selection inference. The use of machine learning methods is particularly advantageous in this context, as it appears reasonable to assume that the set of relevant control variables differs across multiple points of the outcome distributions. To test hypotheses concerning potentially heterogeneous treatment effects, we use the post-double selection logistic distribution regression approach proposed by Belloni et al. (2018b). This estimator allows for uniformly valid inference about the target coefficients of our low-dimensional treatment variables across the entire outcome distributions.

The data for this study were sourced from the German Socio-Economic Panel (SOEP). The SOEP is characterised by a moderate sample size, yet it contains a large number of potential control variables. To estimate the base and treatment effects of our continuous target variable, we employ a difference-in-differences set-up. More precisely, the continuous treatment variable in question represents the proportion of workers in specific population

subgroups who earn wages below the minimum wage level. This proportion is measured in the period prior to the introduction of the minimum wage. The idea is that changes induced by the minimum wage should be largest in population subgroups that had the strongest exposure to the new minimum wage level, controlling for other characteristics. Our empirical results indicate that the minimum wage displaced hourly wages below its minimum threshold, benefited monthly wages in the lower-middle but not the lowest part of the distribution, and did not significantly distort the distribution of working hours. These findings help reconcile the previously conflicting conclusions reached in the literature based on alternative data sources.

A brief Overview of the Contents of Chapter 3

In Chapter 3, we employ the cutting-edge selection model developed by Chernozhukov et al. (2023) to investigate the impacts of unobserved, potentially heterogeneous selectivity. This allows us to conduct a thorough analysis of wage disparities between male and female workers in Germany, utilizing high-quality administrative data. Additionally, we are among the first to delve deeply into the unobserved selectivity patterns in both full-time and part-time employment. Our findings reveal significant variations in selectivity patterns across different wage distributions.

For full-time male workers, unobserved selectivity is positive at the lower end of the wage distribution but negative elsewhere, with this trend becoming more pronounced over time. Conversely, for full-time female workers, unobserved selectivity is generally negative, possibly due to assortative matching and household dynamics, where women with superior unobservables may not need to contribute to household income. Our model indicates that a substantial portion of the full-time gender wage gap can be attributed to differences in unobserved selectivity between men and women. Overall, the full-time wage gap between men and women significantly narrowed between 2000-2005 and 2012-2017, largely due to decreasing differences in unobserved selectivity and improved observables for women.

In the context of part-time employment, we find that men who work part-time constitute a distinct subset of those not pursuing regular full-time employment. Historically, the share of men working part-time in Germany was small, making this group highly specific in terms of both observed characteristics and unobservables. However, as part-time employment among men has increased, male selectivity in part-time work has become less pronounced, although it remains less common than for women. For women, part-time work exhibits a complex selectivity pattern, shifting from negative at the lower end to positive at the upper end of the wage distribution. This pattern can be explained by assortative matching, where women in the lower part of the distribution need to contribute to household

income, while those in the upper part only work if they receive very high wage offers. Recent declines in female part-time selectivity may reflect improvements in public childcare and shifting social norms that support post-childbirth employment. Similar to full-time employment, we observe a convergence of male and female part-time wage distributions, primarily explained by declining differences in wage returns and unobserved selectivity.

In conclusion, our study provides compelling evidence of substantial heterogeneity in selectivity patterns across full-time and part-time wage distributions, a phenomenon that has not been adequately acknowledged in previous research.

Chapter 1

POST-SELECTION INFERENCE IN TOBIT MODELS WITH HIGH-DIMENSIONAL CONTROL VARIABLES*

1.1. INTRODUCTION

Regression problems involving outcome variables, which are unobservable if their realizations fall below a certain limit of detection, are encountered frequently and in a variety of contexts. In general, we differentiate between situations, in which information on observations with outcomes beyond such a threshold is lost entirely, and those, where we retain information on our set of explanatory variables. For the latter, Hald (1949) coined the term censoring while he referred to the former as truncation. Censoring, which this study will focus on, arises either as a consequence of naturally limited response variables or from a selective sampling process. Examples of responses with a natural zero lower bound include household expenditures, labour supply, and charitable contributions; see, e.g. Tobin (1958), Keeley et al. (1978), and Reece (1979), respectively. In clinical trials, by contrast, prediction targets such as Human Immunodeficiency Virus (HIV) viral load in blood plasma; see, e.g. Swenson et al. (2014), or individual survival times following the diagnosis and treatment of cancer or other severe diseases; see, e.g. Cox (1972) and Buckley and James (1979), involve selective sampling processes because biological assays cannot detect concentrations below a certain threshold or a patient's survival time exceeds the observation period.

*This chapter is version 3.2.1 of an unpublished working paper as at 26 February 2025.

If the presence of censoring is ignored, we generally cannot expect to consistently estimate parameters or other quantities since the selectivity of observed realizations induces a substantial, non-vanishing bias in standard estimation methods. As a result, many models have been devised to address parameter estimation when observed outcomes are censored. Tobin's maximum likelihood estimator; see, e.g. Tobin (1958) and Amemiya (1973, 1984), enjoys enduring popularity in economics, social sciences and, as of late, also in biostatistics; see, e.g. Jacobson and Zou (2023a). Tobit estimators rely on a distributional assumption. Although the canonical model assumes the regression disturbances to be normally distributed, Tobin's original idea does not hinge on normality. Indeed, several variations of the standard model have been studied, among them Tobit models with Burr Type II and Student-t distributions; see, e.g. Fry (1991) and Arellano-Valle et al. (2012).

Powell (1984) developed censored least absolute deviations (LAD) and Powell (1986a) generalized this concept to arbitrary regression quantiles. Chernozhukov and Hong (2002) provide a revised version of Powell's censored quantile regression in the form of a three-step estimator. Powell (1986b) proposed a least squares estimator that balances left-censored distributions by symmetrically clipping the upper tail. This idea is further refined in Honoré and Powell (1994) who considered pairwise trimming the distribution of the residuals conditional on covariates to restore symmetry. These methods, although more robust to outliers and less dependent on distributional assumptions, are numerically intricate and computationally much more costly than maximum likelihood.

Survival data are commonly dealt with by Cox' proportional hazards model; see Cox (1972), Buckley-James imputation; see Buckley and James (1979), or the accelerated failure time (AFT) model, where the outcome represents the logarithm of the time to an event and censoring is accounted for by Kaplan-Meier weights; see, e.g. Stute (1993).

The growing accessibility of high-dimensional data and the emergence of regularization techniques to handle them have prompted a shift in focus towards extending existing censoring models to such high-dimensional settings. Tibshirani (1997) initiated a fast-growing field of research when he combined the original Cox model with Lasso penalty. As a consequence, regularization methods for the Cox model have been subject to comprehensive studies. Interested readers are referred to Bradic et al. (2011), Huang et al. (2013), and references therein for thorough discussions on both computational and theoretical properties of penalized versions of Cox' model.

Buckley-James methods with Lasso penalty or in a Dantzig selector framework can be found in Wang et al. (2008), Johnson (2009), Li et al. (2014), and Soret et al. (2018). Liu

et al. (2013), Zhou et al. (2013), and Müller and van de Geer (2016) covered Lasso and group Lasso for Powell's censored LAD estimator. With the exception of Müller and van de Geer (2016), aforementioned studies mainly focus on feature selection properties. In contrast, Müller and van de Geer (2016) derive bounds for excess risk (Kullback-Leibler information) and parameters in ℓ_1-norm. However, they do not address the numerical challenges due to non-convexity that the ℓ_1-penalized censored LAD method inherits from Powell's original estimator. Pan and Xie (2023) evade this technical difficulty by instead focussing on a ℓ_1-penalized version of the pairwise difference LAD estimator of Honoré and Powell (1994). They cover both computational aspects by developing less demanding optimization algorithms based on the general procedure of the alternating direction method of multipliers (ADMM) and theoretical guaranties in deriving bounds for the penalized estimator in prediction and ℓ_2-norm.

Notwithstanding the enduring popularity of Tobin's original maximum likelihood estimator, penalized counterparts have thus far received scant coverage in the rapidly evolving literature on high-dimensional censoring models. Aydin et al. (2021) discuss a ℓ_2-penalized Tobit estimator. Indeed, Jacobson and Zou (2023a) were the first to investigate the theoretical properties of Tobit Lasso. They show that the likelihood score with respect to the regression parameters in Tobin's canonical model with normal disturbances is sub-Gaussian. Based on this result and other technical conditions, they provide a ℓ_2-bound for the parameter vector of the normal Tobit Lasso estimator.

To date, post-regularization inference in high-dimensional censoring models has only been investigated in a limited number of studies. For the Cox model, Fang et al. (2017) use a one-step correction based on a linearised, decorrelated score function which serves as an approximately unbiased estimating equation. Their approach is closely related to the Neyman orthogonal likelihood setting outlined in Chernozhukov et al. (2015) and used by Belloni et al. (2016a) for generalized linear models (GLM). In the AFT framework, Chai et al. (2019) propose quasi normal equations for Stute's least squares estimator with Kaplan-Meier weights. The resulting orthogonality between weighted control variables and a suitably constructed instrument for the target regressors of interest is analogous to a weighted version of Neyman conditions for linear regression models, as developed in Belloni et al. (2012, 2014), and Zhang and Zhang (2014). Bradic and Guo (2019) developed a one-step de-biasing method for generalized M-estimators, which can also be applied to censored data. Essentially, they utilize a system of smoothed estimating equations to update an initial, regularized estimator in the semi-parametric censored regression model. By choosing the LAD loss, the authors derive asymptotically valid confidence intervals for the parameter vector, suggesting the use of the ℓ_1-regularized estimator of Müller and van de Geer (2016) as an initial plug-in estimate.

1.1.1. Contribution and Structure of the Article

The present study addresses the dearth of theoretical results pertaining to post-selection inference in high-dimensional Tobit models. To this end, we essentially contribute twofold. Firstly, we derive asymptotic normality of an instrumental Tobit estimator under high-level conditions that permit the use of various machine learning methods in providing plug-in estimates for the high-dimensional nuisance parts of the model. In this regard, the Neyman orthogonality property, which establishes first-order immunity with respect to nuisance functions, plays a pivotal role. By leveraging the same asymptotic Taylor expansion arguments as in Belloni et al. (2016a, 2019), we show that due to this first-order immunity, higher-order bias terms vanish at a rate faster than $\sqrt{n}$ and, consequently, do not impact the limiting distribution of our estimating function for the target parameter of interest. Our proof allows us to cover Tobit likelihood losses based on a range of disturbance distributions, provided that the density is three times continuously differentiable and that certain Lipschitz conditions on the likelihood and its first and second derivatives are satisfied. These requirements, although considerably more stringent, are analogous to the conditions imposed on the link function of a GLM; see van de Geer (2008) and Belloni et al. (2016a).

Secondly, we demonstrate that the high-level conditions set out for asymptotic normality are implied by a set of primitive assumptions and the choice of the logistic distribution in location-scale parametrization. Among more technical conditions, we impose sparsity on the nuisance parameters, thereby allowing the number of overall control variables to be substantially larger than the number of observations $p \gg n$, but requiring the essential number of controls relevant to accurately approximate the data generating process for a given n to be much smaller $s \ll n$. Both $p = p_n$ and $s = s_n$ may grow according to the condition $s^2 \log^2(p) = o(n)$. The theoretical analysis of our estimators with logistic likelihood loss is primarily based on the following pillars. The likelihood loss comprises the softplus function $\log(1 + \exp\{t\})$ and a linear component. Therefore, we can build on the results of Bach (2010) regarding the modified self-concordance property of the softplus function and of Belloni and Chernozhukov (2011) on the non-linear impact coefficient. In this context, the presence of an additional scale parameter associated with observed values of the outcome variable mandates the introduction of augmented design quantities (matrices and eigenvalues). Additionally, we rely on the results for weighted post-Lasso found in Belloni et al. (2016b, 2019). Our theoretical contributions demonstrate that $\sqrt{n}$ consistency and asymptotic normality of a post-ℓ_1 instrumental logistic Tobit can be established under almost identical conditions as those required for the logistic GLM in Belloni et al. (2016a).

We stress that none of the results listed above has previously been established for Tobit models. As highlighted in the literature review, Jacobson and Zou (2023a) derived bounds

on the parameters in a ℓ_1-regularized version of Tobin's canonical model with normal disturbances assuming a fixed design, i.e., assuming non-random regressors. While there are similarities in terms of the general idea behind both models, our work does not rely on any specific results obtained by Jacobson and Zou (2023a). Moreover, we present conditions under which statistical inference can be performed in the presence of censored outcomes, a topic that Jacobson and Zou (2023a) identified as an interesting direction for future research in their concluding remarks.

The remainder of this article is organized as follows. Section 1.2.1 provides an overview of the general set-up and illustrates the main idea behind simultaneously countering the impact of regularization and censoring bias. In Section 1.2.2, specific estimation algorithms based on the ℓ_1-penalized logistic Tobit are outlined. The optimization of Tobit likelihood losses and related issues are addressed in Section 1.3. Section 1.4 presents our main theoretical results, including conditions for asymptotic normality. The finite sample performances of our estimators are investigated in extensive Monte Carlo experiments, the results of which are summarized in Section 1.5.1. To demonstrate the practical applicability of our estimators, we apply them in Section 1.5.2 to the important use case of testing the effect of gene mutations on left-censored HIV viral load data from AIDS Clinical Trails Group 5241 from the Stanford HIV Drug Resistance Database. Eventually, Section 1.6 concludes.

1.1.2. Notation

Symbol $\mathrm{E}[\cdot]$ denotes the expectation operator with respect to the underlying probability measure P that characterizes the distribution of the data. We use $\mathbb{E}_n[\cdot] = n^{-1}\sum_{i=1}^n[\cdot]$ as a shorthand notation for the empirical mean over indices $i \in \{1,\dots,n\}$. Moreover, let $\overline{\mathrm{E}}[\cdot] = \mathbb{E}_n[\mathrm{E}[\cdot]]$. For example, $\overline{\mathrm{E}}[d_i^2] = n^{-1}\sum_{i=1}^n \mathrm{E}[d_i^2]$. For a function $f : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^p \to \mathbb{R}$, we write $\mathbb{G}_n(f) = \sqrt{n}\,\mathbb{E}_n[f(y_i,d_i,x_i) - \mathrm{E}[f(y_i,d_i,x_i)]]$. For a vector $\xi \in \mathbb{R}^p$ we denote the ℓ_1-norm as $\|\xi\|_1$, the ℓ_2-norm as $\|\xi\|$, and the ℓ_∞-norm as $\|\xi\|_\infty$. For a sequence $(m_i)_{1\leqslant i\leqslant n}$, we define $\|m_i\|_{2,n} = \sqrt{\mathbb{E}_n[m_i^2]}$. For example, $\|x_i\xi\|_{2,n} = \sqrt{\mathbb{E}_n[(x_i\xi)^2]}$ represents the prediction norm of ξ. The support of vector ξ defines the set of indices associated with non-zero elements, i.e., $\mathrm{support}(\xi) = \{j \in \{1,\dots,p\} : \xi_j \neq 0\}$ and $\|\xi\|_0$ yields the cardinality of this set. We write $\xi_{\mathcal{S}}$ to create a vector ξ^* containing the components ξ_j in index positions $j \in \mathcal{S} \subseteq \{1,\dots,p\}$, i.e., $\xi_j^* = \xi_j$ for all $j \in \mathcal{S}$ and $\xi_j^* = 0$ else. $a \lesssim b$ means $a \leqslant c\cdot b$ for some constant $c > 0$ that does not depend on the sample size n. Similarly, $a \lesssim_{\mathrm{P}} b$ indicates that a is bounded in probability $a = O_{\mathrm{P}}(b)$. For a differentiable map $g : \mathbb{R}^p \to \mathbb{R} : \xi \mapsto g(\xi)$ expression $\partial_\xi\, g$ abbreviates $\partial g/\partial\xi$. In particular, $\partial_\xi\, g(\hat{\xi})$ means $\partial_\xi\, g(\xi)|_{\xi=\hat{\xi}}$. For the chain $f(g(\xi))$, we write $f'(g(\hat{\xi})) = \partial_{g(\xi)}\, f(g(\hat{\xi}))$ to indicate the outer derivative.

1.2. SET-UP AND OUTLINE OF METHOD

1.2.1. General High-dimensional Tobit

Consider the case of a linear regression model, where the left-censored outcome of interest y_i relates to a scalar target regressor d_i, and p-dimensional control variables x_i:

$$\gamma_0 y_i = \max\{\gamma_0 y_i^* \,,\, \text{ɣ}_i\} = \max\{\alpha_0 d_i + x_i\beta_0 + \gamma_0 u_i \,,\, \text{ɣ}_i\}, \text{ with } \mathrm{E}[u_i \mid d_i, x_i] = 0. \tag{1.2.1}$$

The observed outcome y_i is censored unless the latent variable y_i^* exceeds the (rescaled) limit of detection $\gamma_0^{-1}\text{ɣ}_i$. We assume this censoring threshold to be known and equal to zero $\text{ɣ}_i = \text{ɣ} = 0$ for all $i \in \{1, \ldots, n\}$. This simplification does not affect the generality of the results presented below, since non-zero limits ɣ_i could be subtracted from the constant column included in x_i. Moreover, right-censored outcomes can be treated as left-censored by reversing the sign of y_i; see, e.g. Amemiya (1973).

In this context, $\alpha_0 \in \mathbb{R}$ is our target parameter, $x_i\beta_0$ with $\beta_0 \in \mathbb{R}^p$ is the nuisance regression function, and $\gamma_0 \in \mathbb{R}^+$ represents the reciprocal of the scale parameter of the distribution of the disturbances u_i, where we apply the well-known substitution of Olsen (1978): $\gamma_0 = 1/\sigma_0$, $\alpha_0 = \alpha_0^*/\sigma_0$, and $\beta_0 = \beta_0^*/\sigma_0$. We assume that disturbances u_i are $i.i.d.$ and follow the absolutely continuous probability distribution $\mathrm{P}(u_i/\sigma_0 \leqslant t) = F(t)$ (CDF) with three times continuously differentiable density $f(t)$ (PDF), where $\sigma_0 = 1/\gamma_0 > 0$ and $\sigma_0 \lesssim \mathrm{E}[u_i^2]^{1/2}$.[1] For example, in the logistic case discussed below, we have $\sigma_0 = \pi^{-1}\{3\,\mathrm{E}[u_i^2]\}^{1/2}$. In Condition ITob 4.1–(i), we list several conditions which directly translate into restrictions on the set of potential choices for CDF $F(t)$. Also note that (1.2.1) contains the canonical model of Tobin (1958) in Olsen's parametrization as a special case; see Olsen (1978).

Let $(m_i)_{1\leqslant i\leqslant n} := (y_i, d_i, x_i)_{1\leqslant i\leqslant n}$ be a random sample, independent across i, and obeying the model (1.2.1). Then, the empirical likelihood loss (negative log-likelihood) function associated with distribution function $F(t)$ and density $f(t)$ is given by:

$$\begin{aligned}\Lambda(\alpha, \beta, \gamma) \;=\; \mathbb{E}_n[\Lambda_i(\alpha, \beta, \gamma)] \;\;=\;\; & -\mathbb{E}_n[(1 - s_i)\log\{F(-\alpha d_i - x_i\beta)\}] \\ & -\mathbb{E}_n[s_i \log\{f(\gamma y_i - \alpha d_i - x_i\beta)\}] - \mathbb{E}_n[s_i \log\{\gamma\}],\end{aligned} \tag{1.2.2}$$

[1] The distribution function $F(t)$ serves as the (inverse) link function of a binary choice model, where the indicator of observed outcomes $\mathbf{1}\{y_i^* > 0\}$ relates to the target regressor d_i and the high-dimensional controls x_i by the equality $\mathrm{E}[\mathbf{1}\{y_i^* > 0\} \mid d_i, x_i] = F(\alpha_0 d_i + x_i\beta_0)$. In Remark 1.2.2 below, we specifically address the case of the logistic link function and compare the logistic Tobit to a logistic binary choice model.

where the dichotomous variables $s_i := \mathbf{1}\{y_i^* > 0\}$ and $(1 - s_i) := \mathbf{1}\{y_i^* \leqslant 0\}$ mark observed and censored outcomes, respectively. Based on loss function (1.2.2), we define the following "outer score" function in our target αd_i:

$$\begin{aligned} g(\gamma y_i - \alpha d_i - x_i\beta) &= (1 - s_i) g_1(-\alpha d_i - x_i\beta) + s_i g_2(\gamma y_i - \alpha d_i - x_i\beta) \\ &= (1 - s_i) \frac{f(-\alpha d_i - x_i\beta)}{F(-\alpha d_i - x_i\beta)} + s_i \frac{f'(\gamma y_i - \alpha d_i - x_i\beta)}{f(\gamma y_i - \alpha d_i - x_i\beta)}. \end{aligned} \tag{1.2.3}$$

We aim to perform statistical inference about the target coefficient α_0 that continues to be valid in a high-dimensional setting, i.e., when the overall number of control variables exceeds the sample size $p \gg n$, or when we we have to employ regularization techniques to estimate the nuisance term $x_i\beta_0$. To this end, we rely on the estimating function

$$\psi(m_i, z_i, q_i, \alpha, \beta, \gamma, \mu) = g(\gamma y_i - \alpha d_i - x_i\beta)(z_i - \mu q_i) - \mu s_i/\gamma, \tag{1.2.4}$$

which is designed to be robust regarding moderate model selection mistakes or, quite generally, slower than $\sqrt{n}$ convergence rates that invariably accompany regularized estimators. More specifically, our methods described below explicitly or implicitly construct instruments $z_{0i} = z_0(d_i, x_i)$ and $q_{0i} = q_0(y_i, d_i, x_i)$ such that:

$$\begin{aligned} \mathrm{E}[\psi(m_i, z_{0i}, q_{0i}, \alpha_0, \beta_0, \gamma_0, \mu_0)] &= \underbrace{\mathrm{E}[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta) z_{0i}]}_{=0} \\ &\quad - \mu_0 \underbrace{\mathrm{E}[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta) q_{0i} + s_i/\gamma_0]}_{=0} = 0, \end{aligned} \tag{1.2.5}$$

$$\partial_\alpha \, \mathrm{E}[\psi(m_i, z_{0i}, q_{0i}, \alpha, \beta_0, \gamma_0, \mu_0)]\big|_{\alpha=\alpha_0} = \underbrace{\mathrm{E}[w_i z_{0i} d_i]}_{>0} - \mu_0 \underbrace{\mathrm{E}[w_i q_{0i} d_i]}_{=0} > 0, \tag{1.2.6}$$

$$\partial_\beta \, \mathrm{E}[\psi(m_i, z_{0i}, q_{0i}, \alpha_0, \beta, \gamma_0, \mu_0)]\big|_{\beta=\beta_0} = \underbrace{\mathrm{E}[w_i z_{0i} x_i]}_{=0} - \mu_0 \underbrace{\mathrm{E}[w_i q_{0i} x_i]}_{=0} = 0, \tag{1.2.7}$$

$$\partial_\gamma \, \mathrm{E}[\psi(m_i, z_{0i}, q_{0i}, \alpha_0, \beta_0, \gamma, \mu_0)]\big|_{\gamma=\gamma_0} = \mu_0 \underbrace{\mathrm{E}[w_i q_{0i} y_i + s_i/\gamma_0^2]}_{>0} - \mathrm{E}[w_i z_{0i} y_i] = 0, \tag{1.2.8}$$

where we define:

$$w_i := -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0), \tag{1.2.9}$$

which represent data-dependent weighting factors in the moment conditions. In Condition ITob 4.1–(i) below, we require $0 < w_i \leqslant C$ for some suitable constant C. Indeed, if the distribution of u_i satisfies $\{f(t)\}^2 > f'(t)F(t)$ and $\{f'(t)\}^2 > f''(t)f(t)$ for all $t \in \mathbb{R}$, we guarantee $w_i > 0$ and ensure strict convexity of likelihood loss function (1.2.2) in all arguments.

Relations (1.2.5) and (1.2.6) yield conditions for the estimating equation for the target parameter of interest α_0. Conditions (1.2.7) and (1.2.8) allow us to control the impact of estimating the nuisance function and the inverse scale parameter by imposing first-order

orthogonality with respect to β_0 and γ_0. In other words, we immunize the estimation of α_0 against small perturbations of $x_i\beta_0$ and γ_0.[2] While instruments z_{0i} and q_{0i} are specifically constructed in such a way that orthogonality in β_0 is achieved, we require an additional parameter, μ_0, to establish orthogonality in γ_0. Indeed, (1.2.8) provides the solution $\mu_0 = \mathrm{E}[w_i q_{0i} y_i + s_i/\gamma_0^2]^{-1}\mathrm{E}[w_i z_{0i} y_i]$. In addition, since $\mathrm{E}[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)] = 0$ holds by including an intercept, the estimation of α_0 is also insensitive with respect to perturbations of instruments z_0 and q_0.

Estimating function (1.2.4) serves as a generic vehicle for the construction of a variety of estimators for α_0. This encompasses the estimation of the nuisance part or any of the instruments by different machine learning techniques. Depending on the chosen method, there will often be alternative, asymptotically equivalent approaches that establish the required orthogonality conditions. To gain a more concrete understanding of the estimation algorithms described below, consider the following decompositions of the weighted main regressor and the weighted censored outcome variable:

$$\sqrt{w_i}d_i = \sqrt{w_i}x_i\eta_0 + v_i, \text{ with } \mathrm{E}[\sqrt{w_i}v_i x_i] = 0, \text{ and} \tag{1.2.10}$$

$$\sqrt{w_i}y_i = \sqrt{w_i}(d_i,\, x_i)\theta_0 + r_i, \text{ with } \mathrm{E}[\sqrt{w_i}(d_i,\, x_i)\, r_i] = 0. \tag{1.2.11}$$

Note that the moment conditions in (1.2.10) and (1.2.11) satisfy those in (1.2.7) and the second term on the right-hand side of (1.2.6), provided that the instruments are built according to:

$$z_{0i} := \frac{v_i}{\{-g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}^{1/2}} = \frac{v_i}{\sqrt{w_i}} = d_i - x_i\eta_0, \text{ and} \tag{1.2.12}$$

$$q_{0i} := \frac{r_i}{\{-g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}^{1/2}} = \frac{r_i}{\sqrt{w_i}} = y_i - (d_i,\, x_i)\theta_0. \tag{1.2.13}$$

This choice of z_{0i} and q_{0i} leads to the following representation of (1.2.5) as one specific linear combination of optimality conditions of the population loss function, namely:

$$\mathrm{E}[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta)\{(1+\mu_0\theta_{01})d_i - x_i(\eta_0 - \mu_0\theta_{0\setminus 1}) - \mu_0 y_i\} - \mu_0 s_i/\gamma_0] = 0, \tag{1.2.14}$$

where θ_{01} represents the first element of column vector θ_0, while $\theta_{0\setminus 1}$ denotes all components of θ_0, with the exception of the first. This relates our approach to the concept of Neyman orthogonal likelihood scores in Chernozhukov et al. (2015). In this context, the true orthogonalization parameter associated with the nuisance function $x_i\beta_0$ is itself a linear combination of individual parameters $(\eta_0 - \mu_0\theta_{0\setminus 1})$, since the control variables occur in both

[2] Although γ_0 is a scalar and can be estimated at the negligible cost of losing one degree of freedom, we consider it a nuisance parameter due to the fact that, in the presence of the p-dimensional parameter β_0, it is not generally possible to estimate γ_0 at $\sqrt{n}$ rate without imposing further assumptions; see, e.g. Yu and Bien (2019).

instruments z_{0i} and q_{0i}. By adding up z_{0i} and q_{0i}, (1.2.14) dispenses with the need to differentiate between these two distinct instruments. In particular, relation (1.2.14) motivates the use of estimators which separately create orthogonality in each element or specific blocks of the columns of $(d_i,\, x_i,\, y_i)$; see Algorithms 1.2.1 and 1.2.2, and Remark 1.2.1.

By combining equation (1.2.6) and orthogonality conditions (1.2.7), (1.2.8) with the definitions of our instruments in (1.2.10) and (1.2.11), we have:

$$\begin{aligned} \mathrm{E}[w_i(d_i,\, x_i)\, q_{0i}] &= \mathrm{E}[w_i\{y_i - (d_i,\, x_i)\theta_0\}(d_i, x_i)] = 0, && (1.2.15)\\ \mathrm{E}[w_i z_{0i} x_i] &= \mathrm{E}[w_i(d_i - x_i\eta_0)x_i] = 0, && (1.2.16)\\ \mathrm{E}[w_i q_{0i} y_i + s_i/\gamma_0^2] &= \mathrm{E}[\sqrt{w_i} r_i y_i] + \mathrm{E}[s_i/\gamma_0^2] > 0. && (1.2.17) \end{aligned}$$

The first two relations may be conceived as the optimality conditions of linear projections of the weighted regressor of interest $\sqrt{w_i}d_i$ on the weighted control variables $\sqrt{w_i}x_i$ and, analogously, $\sqrt{w_i}y_i$ on $\sqrt{w_i}(d_i,\, x_i)$. As such, we can apply regularization methods and derive estimators for instruments z_{0i} and q_{0i}.

In Section 1.2.2 below, we exploit sparsity of the high-dimensional nuisance parameters by imposing $\|\beta_0\|_0 \leqslant s$ and $\|\eta_0\|_0 \leqslant s$, where sparsity index s and the overall number of potential control variables p may grow with the sample size n. In particular, we require the triple (s, p, n) to obey the growth condition $s^2 \log^2(p)/n \to 0$. Since ℓ_1-regularization methods typically miss "weak signals" – coefficients that are indistinguishable from zero for a given sample size n – a naïve, i.e. non-orthogonal, approach suffers from omitted variable bias. Formally, the use of regularization methods induces a bias $\|\hat{\beta} - \beta_0\|$, which typically does not vanish at $\sqrt{n}$ rate. By contrast, the orthogonality conditions presented above guarantee that estimating the nuisance term has an asymptotically negligible effect on the estimating equation for α_0, because we immunize against first-order bias and the second-order term $\|\hat{\beta} - \beta_0\|^2$ shrinks sufficiently fast. In combination with a ℓ_1-regularization penalty level of order $\sqrt{\log(p)/n}$, the growth condition defined above warrants the estimation of the nuisance parameters at $n^{1/4}$ rate. The resulting estimator for α_0 asymptotically behaves as if the true values β_0 and γ_0 were taken. Therefore, orthogonality conditions (1.2.7) and (1.2.8) are indispensable ingredients in deriving the asymptotic validity of the proposed statistical inference on α_0. In Section 1.4, we present a set of high-level conditions that extend beyond the sparse setting and permit the estimation of instruments using a variety of machine learning methods.

1.2.2. Implementation as Sparse Logistic Tobit

Next, we discuss specific implementations of the general set-up described above for the case of a logistic Tobit. Under the assumption of logistic disturbances u_i, we have pdf $f(t) = \exp\{t\}/(1+\exp\{t\})^2$ and cdf $F(t) = \exp\{t\}/(1+\exp\{t\})$. From the definitions in (1.2.3) and (1.2.9), we obtain expressions

$$g(t) \;=\; \frac{1 - s_i \exp\{t\}}{1+\exp\{t\}}, \quad -g'(t) \;=\; \frac{(1+s_i)\exp\{t\}}{(1+\exp\{t\})^2}, \tag{1.2.18}$$

which satisfy $\sup_{t\in\mathbb{R}} |g(t)| = 1$, $\sup_{t\in\mathbb{R}} |g'(t)| = 1/2$, $\sup_{t\in\mathbb{R}} |g''(t)| = 3^{-3/2} < 1/5$ and therefore conform to the requirements in Condition ITob 4.1–(i) below. Moreover, as shown in Appendix A.4, the choice of the logistic distribution reduces model (1.2.1) to the simple representation:

$$\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] \;=\; \log\big(1+\exp\{\alpha_0 d_i + x_i\beta_0\}\big), \tag{1.2.19}$$

where the softplus function $\log(1+\exp\{t\})$ relates the conditional expected value of censored outcomes $\gamma_0 y_i$ to the target regressor d_i and high-dimensional controls x_i.

In Algorithms 1.2.1 and 1.2.2, we give detailed instructions on two procedures that estimate α_0. There, the empirical likelihood loss associated with our logistic Tobit is:

$$\begin{aligned}\Lambda(\alpha,\beta,\gamma) \;=\; \mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)] \;=\; \mathbb{E}_n[(1+s_i)\log(1+\exp\{-\gamma y_i + \alpha d_i + x_i\beta\})] \\ - \mathbb{E}_n[s_i(-\gamma y_i + \alpha d_i + x_i\beta + \log\{\gamma\})],\end{aligned} \tag{1.2.20}$$

which can be broken down into two distinct parts. The first term consists of the softplus function, whereas the second term is linear in the regression error $\gamma y_i - \alpha d_i - x_i\beta$. In the theoretical analysis of the ℓ_1-penalized logistic Tobit estimator, we utilize the modified self-concordance property of the softplus function, as defined in Bach (2010). In particular, we can bound (1.2.20) from below by a suitable quadratic function, employing the non-linear impact coefficient introduced in Belloni and Chernozhukov (2011) to control the quality of this minoration.

Our estimators mainly comprise three steps. The first step is to estimate the nuisance terms $x_i\beta_0$ and γ_0 by predicting the censored outcomes y_i using a post-ℓ_1 regularized logistic Tobit. In the second step, we estimate instruments z_{0i} and q_{0i} that are crucial for establishing orthogonality by weighted post-Lasso. Note that the weights involved are data-dependent and, therefore, must be estimated based on the results from Step 1. The third step updates the initial estimate for the parameter of interest α_0 by combining the estimates of nuisance terms and instruments.

Algorithm 1.2.1 *(Instrument Estimator for Logistic Tobit).*

Step 1. Run the post-ℓ_1 penalized logistic Tobit of y_i on d_i and x_i to obtain:

$$\begin{aligned}(\hat{\alpha},\,\hat{\beta},\,\hat{\gamma}) &\in \underset{\alpha,\beta,\gamma}{\arg\min}\,\mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)] + \frac{\lambda_y}{n}\big\|(\alpha,\,\beta^T)^T\big\|_1,\\ (\tilde{\alpha},\,\tilde{\beta},\,\tilde{\gamma}) &\in \underset{\alpha,\beta,\gamma}{\arg\min}\,\mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)]\ :\ \text{support}(\beta)\ \subseteq\ \text{support}(\hat{\beta}).\end{aligned}$$

For $i \in \{1,\ldots,n\}$ save $x_i\tilde{\beta}$, $\tilde{\gamma}y_i$, and compute weights $\widehat{w}_i = -g'(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})$.

Step 2. *(a)* Run the weighted post-Lasso estimator of $\sqrt{\widehat{w}_i}d_i$ on $\sqrt{\widehat{w}_i}x_i$:

$$\begin{aligned}\hat{\eta} &\in \underset{\eta}{\arg\min}\,\frac{1}{2}\,\mathbb{E}_n\big[\widehat{w}_i(d_i - x_i\eta)^2\big] + \frac{\lambda_d}{n}\big\|\widehat{\Psi}\eta\big\|_1.\\ \tilde{\eta} &\in \underset{\eta}{\arg\min}\,\mathbb{E}_n\big[\widehat{w}_i(d_i - x_i\eta)^2\big]\ :\ \text{support}(\eta)\ \subseteq\ \text{support}(\hat{\eta}).\end{aligned}$$

(b) (optional) Run the WLS estimator of $\sqrt{\widehat{w}_i}y_i$ on $\sqrt{\widehat{w}_i}(d_i,\,x_i)$:

$$\tilde{\theta} \in \underset{\theta}{\arg\min}\,\mathbb{E}_n\big[\widehat{w}_i\{y_i - (d_i,\,x_i)\theta\}^2\big]\ :\ \text{support}(\theta)\ \subseteq\ \text{support}(\hat{\beta}).$$

For $i \in \{1,\ldots,n\}$ compute instruments $\hat{z}_i = d_i - x_i\tilde{\eta}$ and optionally $\hat{q}_i = y_i - (d_i,\,x_i)\tilde{\theta}$ and orthogonalization parameter $\hat{\mu} = \mathbb{E}_n[\widehat{w}_i\hat{q}_i y_i + s_i/\tilde{\gamma}^2]^{-1}\mathbb{E}_n[\widehat{w}_i\hat{z}_i y_i]$.

Step 3. Run the instrumental Tobit estimator using $\hat{z}_i$ as instrument for d_i:

$$\check{\alpha} \in \underset{\alpha\in\mathcal{A}}{\arg\inf}\,\frac{\mathbb{E}_n[g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})\hat{z}_i]^2}{\mathbb{E}_n[\{g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})\}^2\hat{z}_i^2]},$$

or, if Step 2 (b) has been executed, run the alternative instrumental estimator:

$$\check{\alpha} \in \underset{\alpha\in\mathcal{A}}{\arg\inf}\,\frac{\mathbb{E}_n[g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}]^2}{\mathbb{E}_n[\{g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}\}^2]},$$

where $\mathcal{A} = \{\alpha \in \mathbb{R} : |\alpha - \tilde{\alpha}| \leqslant C/\log(n)\}$.

Step 4. Choose a confidence level $\zeta \in (0,1)$ and compute the associated interval:

$$\mathcal{CR}_I \;=\; \{\alpha \in \mathbb{R} : |\alpha - \check{\alpha}| \leqslant \widehat{\Sigma}_n\Phi^{-1}(1-\zeta/2)/\sqrt{n}\},$$

where $\widehat{\Sigma}_n^2 = \max\big\{\widehat{\Sigma}_{1n}^2, \widehat{\Sigma}_{2n}^2\big\}$ with either $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\widehat{w}_i d_i\hat{z}_i]^{-2}\mathbb{E}_n[\{g(\tilde{\gamma}y_i - \check{\alpha}d_i - x_i\tilde{\beta})\hat{z}_i\}^2]$, or $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\widehat{w}_i d_i\hat{z}_i]^{-2}\mathbb{E}_n[\{g(\tilde{\gamma}y_i - \check{\alpha}d_i - x_i\tilde{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}\}^2]$, and $\widehat{\Sigma}_{2n}^2 = \mathbb{E}_n[\widehat{w}_i\hat{z}_i^2]^{-1}$.

To provide a more formal explanation of how this instrumental estimator works, we examine the optimality conditions of Step 1 in greater detail. To this end, let $\widehat{\mathcal{J}} = \text{support}(\hat{\beta}) = \{j \in \{1,\ldots,p\} \mid \hat{\beta}_j \neq 0\}$ denote the set of indices selected in Step 1. By the first-order con-

ditions of the post-ℓ_1 logistic Tobit, we have

$$\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})(d_i,\, x_{i\widehat{\mathcal{J}}})^T] = 0,\ -\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})y_i + s_i/\tilde{\gamma}] = 0, \tag{1.2.21}$$

which create an orthogonal relation to any linear combination of $(-y_i,\, d_i,\, x_{i\widehat{\mathcal{J}}})$. In particular, we implicitly build instrument $-\mu\hat{q}_i$ by choosing $(-y_i,\, d_i,\, x_{i\widehat{\mathcal{J}}})(\mu,\, \mu\tilde{\theta}^T)^T$ such that

$$\begin{aligned}\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})(-\mu y_i + \mu\tilde{\theta}_1 d_i + \mu x_{i\widehat{\mathcal{J}}}\tilde{\theta}_{\setminus 1}) - \mu s_i/\tilde{\gamma}] \\ = -\mu\,\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})\hat{q}_i + s_i/\tilde{\gamma}] = 0.\end{aligned} \tag{1.2.22}$$

Step 1 of Algorithm 1.2.1 automatically estimates the nuisance parameters β_0 and γ_0 in such a way that the empirical analogue of the second term on the right-hand side of relation (1.2.5) becomes zero. Moreover, there is no need to concern ourselves with the estimation of μ_0, as (1.2.22) is satisfied for any value of μ, most importantly $\hat{\mu}$ or μ_0. Consequently, we may concentrate our efforts on attaining the sample equivalent of the first condition on the right-hand side of (1.2.5). In Step 3, the point estimate for α_0 is refined from the initial, naïve estimate $\tilde{\alpha}$ to the robust $\check{\alpha}$ by replacing the regressor of interest d_i with the estimated instrument $\hat{z}_i$. Hypotheses about α_0 can be tested based on the asymptotic normality of the estimator $\check{\alpha}$ using the estimated variance in Step 4.

Remark 1.2.1 *(Equivalence of Variations of the Instrument Estimator).*

Equation (1.2.5) states that an estimator for α_0 must create an orthogonal relation in both instruments. By exempting the inverse scale parameter from regularization and the first-order conditions of the likelihood loss minimization in Step 1 of Algorithm 1.2.1, the estimator automatically achieves the empirical analogue to the second condition on the right-hand side of (1.2.5). In fact, set $\widehat{\mathcal{J}}$ contains all indices relevant for the construction of instrument $\hat{q}_i$. To see that both instrumental estimators in Step 3 of Algorithm 1.2.1 are asymptotically equivalent, consider the first-order Taylor expansion in α with second-order Lagrange remainder of the Neyman orthogonal score function about the post-ℓ_1 estimate $\tilde{\alpha}$ from Step 1. We have that

$$\begin{aligned}\mathbb{E}_n[g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}] = \mathbb{E}_n[g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})\hat{z}_i] \\ -\hat{\mu}\,\mathbb{E}_n[\widehat{w}_i\hat{q}_i d_i](\alpha - \tilde{\alpha}) + O_{\mathrm{P}}\big(|\alpha - \tilde{\alpha}|^2\big),\end{aligned}$$

because of the first-order conditions of the post-ℓ_1 minimization in Step 1. Additionally, by the construction of $\hat{q}_i$ according to Step 2 (b), the in-sample orthogonality condition holds $\mathbb{E}_n[\widehat{w}_i\hat{q}_i d_i] = 0$; see the second term on the right-hand side of (1.2.6). As a consequence, both instrumental approaches in Step 3 are equivalent up to the second-order term on the right-hand side, which we control by construction of $\mathcal{A}$ and the post-ℓ_1 rates on $\tilde{\alpha}$ so that $|\alpha - \tilde{\alpha}|^2 \lesssim n^{-1/2}\log^{-1}(n)$; see Step 8 in Appendix A.2.

A second estimator, that builds upon the idea of the double selection method proposed in Belloni et al. (2014), is presented in Algorithm 1.2.2.

Algorithm 1.2.2 *(Post-double Selection Estimator for Logistic Tobit).*

Step 1. Run the post-ℓ_1 penalized logistic Tobit of y_i on d_i and x_i to obtain:

$$\begin{aligned}(\hat{\alpha},\, \hat{\beta},\, \hat{\gamma}) &\in \underset{\alpha,\beta,\gamma}{\arg\min}\, \mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)] + \frac{\lambda_y}{n}\big\|(\alpha,\, \beta^T)^T\big\|_1,\\ (\tilde{\alpha},\, \tilde{\beta},\, \tilde{\gamma}) &\in \underset{\alpha,\beta,\gamma}{\arg\min}\, \mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)] \,:\, \text{support}(\beta) \,\subseteq\, \text{support}(\hat{\beta}).\end{aligned}$$

For $i \in \{1,\ldots,n\}$ save $x_i\tilde{\beta}$, $\tilde{\gamma}y_i$, and compute weights $\widehat{w}_i = -g'(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})$.

Step 2. *(a)* Run the weighted post-Lasso estimator of $\sqrt{\widehat{w}_i}d_i$ on $\sqrt{\widehat{w}_i}x_i$:

$$\begin{aligned}\hat{\eta} &\in \underset{\eta}{\arg\min}\, \frac{1}{2}\,\mathbb{E}_n\big[\widehat{w}_i(d_i - x_i\eta)^2\big] + \frac{\lambda_d}{n}\big\|\widehat{\Psi}\eta\big\|_1.\\ \tilde{\eta} &\in \underset{\eta}{\arg\min}\, \mathbb{E}_n\big[\widehat{w}_i(d_i - x_i\eta)^2\big] \,:\, \text{support}(\eta) \,\subseteq\, \text{support}(\hat{\eta}).\end{aligned}$$

(b) (optional) Run the weighted post-Lasso estimator of censored outcomes $\sqrt{\widehat{w}_i}y_i$ on $\sqrt{\widehat{w}_i}d_i$ and $\sqrt{\widehat{w}_i}x_i$:

$$\begin{aligned}\hat{\theta} &\in \underset{\theta}{\arg\min}\, \frac{1}{2}\,\mathbb{E}_n\big[\widehat{w}_i\{y_i - (d_i, x_i)\theta\}^2\big] + \frac{\lambda_d}{n}\big\|\widehat{\Psi}^*\theta\big\|_1,\\ \tilde{\theta} &\in \underset{\theta}{\arg\min}\, \mathbb{E}_n\big[\widehat{w}_i\{y_i - (d_i,\, x_i)\theta\}^2\big] \,:\, \text{support}(\theta) \,\subseteq\, \text{support}(\hat{\theta}).\end{aligned}$$

For $i \in \{1,\ldots,n\}$ compute instruments $\hat{z}_i = d_i - x_i\tilde{\eta}$ and $\hat{q}_i = y_i - (d_i,\, x_i)\tilde{\theta}$.

Step 3. Run the logistic Tobit of y_i on d_i and the union of controls $\check{\mathcal{J}} = \text{support}(\hat{\beta}) \cup \text{support}(\hat{\eta}) \cup \text{support}(\hat{\theta}_{\setminus 1})$ selected in either Step 1 or 2 to obtain:

$$(\check{\alpha},\, \check{\beta},\, \check{\gamma}) \;\in\; \underset{\alpha,\beta,\gamma}{\arg\min}\, \mathbb{E}_n[\Lambda_i(\alpha,\beta,\gamma)] \,:\, \text{support}(\beta) \,\subseteq\, \check{\mathcal{J}}.$$

Step 4. Choose a confidence level $\zeta \in (0,1)$ and compute the associated interval:

$$\mathcal{CR}_{DS} \;=\; \{\alpha \in \mathbb{R} : |\alpha - \check{\alpha}| \leqslant \widehat{\Sigma}_n\Phi^{-1}(1-\zeta/2)/\sqrt{n}\},$$

where $\widehat{\Sigma}_n^2 = \max\big\{\widehat{\Sigma}_{1n}^2, \widehat{\Sigma}_{2n}^2\big\}$, and for the $(\check{s}+1)$-dimensional quadratic zero matrix $\mathbf{0}_{\check{s}+1}$ with $\check{s} = \text{card}(\check{\mathcal{J}})$ we define

$$\widehat{\Sigma}_{2n}^2 = \left\{\mathbb{E}_n\big[\check{w}_i(d_i,\, x_{i,\check{\mathcal{J}}},\, -y_i)^T(d_i,\, x_{i,\check{\mathcal{J}}},\, -y_i)\big] + \begin{bmatrix}\mathbf{0}_{\check{s}+1} & 0\\ 0 & \mathbb{E}_n[s_i]/\check{\gamma}^2\end{bmatrix}\right\}^{-1}_{1,1},$$

$\check{w}_i = -g'(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})$, $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\check{w}_i d_i\hat{z}_i]^{-2}\mathbb{E}_n[\{g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})(\hat{z}_i - \check{\mu}\hat{q}_i) - \check{\mu}s_i/\check{\gamma}\}^2]$, and $\check{\mu} = \mathbb{E}_n[\check{w}_i\hat{q}_i y_i + s_i/\check{\gamma}^2]^{-1}\mathbb{E}_n[\check{w}_i\hat{z}_i y_i]$.

Our post-double selection implementation aims at directly solving the empirical counterpart to the Neyman orthogonal likelihood score (1.2.14). To this end, let $\check{\mathcal{J}} = \text{support}(\check{\beta}) = \text{support}(\hat{\beta}) \cup \text{support}(\hat{\eta}) \cup \text{support}(\hat{\theta}_{\setminus 1})$ be defined as the union of indices selected in either Step 1 or 2. By the same argument underlying (1.2.22) above, we can construct a likelihood estimator that yields the first-order conditions

$$\mathbb{E}_n[g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})(d_i,\, x_{i\check{\mathcal{J}}})^T] = 0,\; -\mathbb{E}_n[g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})y_i + s_i/\check{\gamma}] = 0. \tag{1.2.23}$$

Remark 1.2.2 *(Comparison to Instrumental Logit of Belloni et al., 2016a).*

As emphasized in footnote 1, we could perform statistical inference on α_0 using the binary choice framework described in Belloni et al. (2016a). To do this, one needs to replace the post-ℓ_1 and instrumental or post-double logistic Tobits in Steps 1 and 3 of Algorithms 1.2.1 and 1.2.2 with post-ℓ_1 and instrumental or post-double Logits of the censoring indicator s_i on d_i and x_i. This alternative approach has the advantage of dispensing with the estimation of γ_0, but at the expense of sacrificing efficiency, as outcomes above the detection limit are exchanged for unit values, and thus the information in the data is made use of at a coarser level, since the variation in observed y_i is ignored. On the other hand, estimating γ_0 does not present a significant challenge to the construction of the estimator for two reasons. Firstly, γ_0 is a scalar and its estimation only causes the loss of a single degree of freedom, which does not affect convergence rates for the parameter vector. Secondly, γ_0 appears as the coefficient in front of y_i, which is the primary prediction target. Chernozhukov et al. (2015) draw a connection between the construction of the instruments in Step 2 and the process of "partialing-out" the effect of the high-dimensional nuisance part. In this spirit, we implicitly residualize y_i by predicting the outcome variable. When estimating α_0, it therefore suffices to control the "direct channel", through which the nuisance function $x_i\beta_0$ affects the estimating function.
Against this background, our logistic Tobit constitutes an extension to the binary choice framework for settings where outcomes exceeding a threshold are observed rather than classified as "successes" (in the sense of a binomial model). In fact, we may conceive the estimating function of our Tobit as that of a GLM with separate link functions for censored and observed outcomes.

If we choose the linear combination $(d_i,\, x_{i\check{\mathcal{J}}},\, -y_i)(\{1 + \mu\tilde{\theta}_1\},\, \{\mu\tilde{\theta}_{\setminus 1}^T - \tilde{\eta}^T\},\, \mu)^T$, we have

$$\begin{aligned}\mathbb{E}_n[g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})\{(1 + \mu\tilde{\theta}_1)d_i - x_{i\check{\mathcal{J}}}(\tilde{\eta} - \mu\tilde{\theta}_{\setminus 1}) - \mu y_i\} - \mu s_i/\check{\gamma}] \\ = \mathbb{E}_n[g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})(\hat{z}_i - \mu\hat{q}_i) - \mu s_i/\check{\gamma}] = 0,\end{aligned} \tag{1.2.24}$$

where now both instruments $\hat{z}_i$ and $\hat{q}_i$ are built implicitly. As in the instrumental estimator case, the optimality condition holds for any value of μ. This is an immediate consequence

of the fact that the inverse scale parameter is automatically estimated, which creates orthogonality with respect to γ_0. Moreover, the post-double selection method re-estimates the nuisance function and replaces the point estimates $\tilde{\beta}$ and $\tilde{\gamma}$ with their respective post-double equivalents $\check{\beta}$ and $\check{\gamma}$. Indeed, as pointed out in Belloni et al. (2016a), a post-double selection estimator can be seen as an iterated version of the instrumental estimator because $\check{\alpha}$ minimizes the criterion function

$$L_n(\alpha) \;:=\; \frac{\mathbb{E}_n[g(\check{\gamma}y_i - \alpha d_i - x_i\check{\beta})(\hat{z}_i - \check{\mu}\hat{q}_i) - \check{\mu}s_i/\check{\gamma}]^2}{\mathbb{E}_n[\{g(\check{\gamma}y_i - \alpha d_i - x_i\check{\beta})(\hat{z}_i - \check{\mu}\hat{q}_i) - \check{\mu}s_i/\check{\gamma}\}^2]}$$

over $\alpha \in \mathbb{R}$.

Remark 1.2.3 *(Data-Driven Choice of Penalty Parameters).*

The appropriate choice of penalty levels λ_y and λ_d is of utmost importance to attain good theoretical properties for our ℓ_1-penalized estimators. The standard principle to derive bounds on the parameters, which we also pursue, is to ensure that the event

$$\lambda_y/n \geqslant c\left\|\nabla_{(\alpha,\,\beta^T,\,\gamma)^T}\,\mathbb{E}_n[\Lambda_i(\alpha_0,\beta_0,\gamma_0)]\right\|_\infty \quad \text{for some } c>1$$

holds with probability $1-o(1)$; see, e.g., Bickel et al. (2009), Belloni et al. (2011, 2016a, 2019). To achieve this, we recommend to normalize the censored outcomes and the features so that $\mathbb{E}_n[y_i^2] = \mathbb{E}_n[d_i^2] = \mathbb{E}_n[x_{ij}^2] = 1$ for all $j \in \{1,\dots,p\}$, to set the penalty levels $\lambda_y = 0.78\sqrt{n}\,\Phi^{-1}(1-0.05/\{p\log(n)\})$, $\lambda_d = 1.1\sqrt{n}\,\Phi^{-1}(1-0.05/\{p\log(n)\})$, and to compute penalty loadings $\widehat{\Psi}$ according to Algorithm A.4.5, which is an adjusted version of the procedure described in Belloni et al. (2016b). The latter choice relies on moderate deviation theory for self-normalizing sums applied to the score of weighted post-Lasso, as proposed in the references above. The choice of λ_y is also based on moderate deviation theory, this time applied to the logistic Tobit likelihood score with respect to the nuisance parameter $\partial_\beta\,\mathbb{E}_n[\Lambda_i(\alpha_0,\beta_0,\gamma_0)]$, and the fact that $0 \leqslant \sqrt{w_i} \leqslant \sqrt{1/2}$; see Lemma 11 in Belloni et al. (2016b). Note that the inverse scale parameter $\tilde{\gamma}$ by construction belongs to the active set, while target coefficient $\tilde{\alpha}$ can be set active without affecting the rates of convergence. Furthermore, the normalization of y_i does not affect the point estimates for α_0 or β_0, but rescales γ_0. Since the variance of $\partial_\gamma\,\mathbb{E}_n[\Lambda_i(\alpha_0,\beta_0,\gamma_0)]$ depends on γ_0, we use this normalization to control the noise induced by this element of the score in finite samples.

The use of alternative methods to choose λ_y and λ_d, such as cross-validation, seems plausible but is not covered by the theory developed here. In the real data application presented in Section 1.5.2, we compare the inferential results obtained with different estimators and penalty choices, including stratified cross-validation. We find that the cross-validated Tobit Lasso tends to under-penalize, which is in line with the theoretical analysis provided in Chetverikov et al. (2021) for standard linear Lasso.

Step 2 (b) of Algorithm 1.2.2 is optional as it only serves to establish orthogonality in the score function of the inverse scale parameter γ_0 with respect to nuisance parameter β_0. We stress that this additional orthogonalization step is not required to prove asymptotic normality of the orthogonal score function of α_0. However, in certain cases of our simulation experiments it enhanced the finite sample performance of the post-double selection estimator by improving the quality of the point estimate for γ_0. Another aspect, which is not covered by the theory developed in this study, concerns joint inference about α_0 and γ_0. Using Algorithm 1.2.2, we necessarily re-estimate γ_0. Developing the theory behind hypothesis tests about γ_0 would be an interesting subject for future research. In this context, Step 2 (b) is likely to play a central role.

1.3. OPTIMIZATION OF PENALIZED TOBIT LIKELIHOOD LOSSES

For computational purposes, we introduce the augmented design matrix

$$A := \begin{bmatrix} d_{\mathcal{C}} & x_{\mathcal{C}} & 0 \\ d_{\bar{\mathcal{C}}} & x_{\bar{\mathcal{C}}} & -y_{\bar{\mathcal{C}}} \\ 0 & 0 & 1 \end{bmatrix}, \tag{1.3.1}$$

where we define the index sets $\mathcal{C} := \{i \in \{1, \dots, n\} \mid s_i = 0\}$, $\bar{\mathcal{C}} = \{1, \dots, n\} \setminus \mathcal{C}$, and $n_0 := \text{card}(\mathcal{C})$. In effect, the random sample $(m_i)_{1 \leqslant i \leqslant n}$ is sorted according to censoring indicator s_i. Then, the gradient of the empirical Tobit likelihood loss function, as defined in equation (1.2.2), can be expressed in terms of the definitions of g_1 and g_2 in (1.2.3):

$$\nabla_\Theta \Lambda(\Theta) = \mathbb{E}_n \begin{bmatrix} g(\gamma y_i - \alpha d_i - x_i\beta) d_i \\ g(\gamma y_i - \alpha d_i - x_i\beta) x_i^T \\ -g(\gamma y_i - \alpha d_i - x_i\beta) y_i - s_i/\gamma \end{bmatrix} = A^T G(\Theta)/n, \tag{1.3.2}$$

where

$$G(\Theta) := \begin{bmatrix} g_1(-\alpha d_{\mathcal{C}} - x_{\mathcal{C}}\beta) \\ g_2(\gamma y_{\bar{\mathcal{C}}} - \alpha d_{\bar{\mathcal{C}}} - x_{\bar{\mathcal{C}}}\beta) \\ (n_0 - n)/\gamma \end{bmatrix} \tag{1.3.3}$$

and $\Theta = (\alpha,\, \beta^T,\, \gamma)^T$. The same argument applies to the empirical Hessian:

$$\triangle_\Theta \Lambda(\Theta) = A^T W(\Theta) A/n, \;\; \text{where} \;\; W(\Theta) := \text{diag} \begin{bmatrix} -g_1'(-\alpha d_{\mathcal{C}} - x_{\mathcal{C}}\beta) \\ -g_2'(\gamma y_{\bar{\mathcal{C}}} - \alpha d_{\bar{\mathcal{C}}} - x_{\bar{\mathcal{C}}}\beta) \\ (n - n_0)/\gamma^2 \end{bmatrix}. \tag{1.3.4}$$

Consequently, imposing $g_1'(t) \leqslant 0$ and $g_2'(t) \leqslant 0$ for $t \in \mathbb{R}$ suffices to ensure that the optimization of (1.2.2) constitutes a convex problem. Based on that, the following result states a versatile minimization procedure.

Proposition 1.3.1 *(Iterated Weighted Least Squares for Tobit Models). Let the Tobit model and its associated likelihood loss be defined as in (1.2.1) and (1.2.2), respectively. For* $s_i = \mathbf{1}\{y_i > 0\}$*, the weighting function*

$$w(t) = (1 - s_i)\frac{\{f(t)\}^2 - f'(t)F(t)}{\{F(t)\}^2} + s_i\frac{\{f'(t)\}^2 - f''(t)f(t)}{\{f(t)\}^2}$$

satisfies $\inf_{t\in\mathbb{R}} w(t) > 0$*. Define* $\Theta = (\alpha, \beta^T, \gamma)^T$ *and let* $\Theta^{[0]}$ *be a feasible initial point. The augmented empirical Gram matrix* A^TA *is invertible. Then, the iterative procedure* $\Theta^{[k]} \leftarrow \Theta^{[k+1]}$ *for* $k \in \mathbb{N}_0$*, which solves the sequence of weighted least squares problems*

$$\Theta^{[k+1]} = \big[A^TW(\Theta^{[k]})A\big]^{-1}\big[A^TW(\Theta^{[k]})\big\{A\Theta^{[k]} - W^{-1}(\Theta^{[k]})G(\Theta^{[k]})\big\}\big],$$

converges to a global minimizer of (1.2.2).

Proof. Because of the restriction imposed on $w(t)$, the empirical Hessian is globally positive semi-definite. If the augmented design matrix A is of full column rank, the empirical Hessian is invertible and repeated Newton-Raphson updates find a unique minimum:

$$\begin{aligned}\Theta^{[k+1]} &= \Theta^{[k]} - \big[A^TW(\Theta^{[k]})A\big]^{-1}A^TG(\Theta^{[k]}) \\ &= \big[A^TW(\Theta^{[k]})A\big]^{-1}\big[A^TW(\Theta^{[k]})A\,\Theta^{[k]} - A^TG(\Theta^{[k]})\big] \\ &= \big[A^TW(\Theta^{[k]})A\big]^{-1}\big[A^TW(\Theta^{[k]})\big\{\underbrace{A\Theta^{[k]} - W^{-1}(\Theta^{[k]})G(\Theta^{[k]})}_{\text{“IWLS response”}}\big\}\big].\end{aligned}$$

■

Proposition 1.3.1 extends the iterated weighted least squares (IWLS) representation of GLMs discussed in, e.g. Friedman et al. (2010) to convex Tobit problems. In the $p \gg n$ setting, the empirical Hessian is rank deficient and therefore not invertible. However, the IWLS procedure can be combined with any convex penalty term, such as Lasso, ridge, or elastic net. In particular, since Proposition 1.3.1 provides the solution to a local quadratic problem, any algorithm designed to solve penalized weighted least squares can be applied to minimize penalized Tobit loss functions. We implemented our estimation Algorithms 1.2.1 and 1.2.2 as Matlab routines using both a Gauss-Seidel (GS) procedure and the alternating direction method of multipliers (ADMM).[3] In our empirical applications, where the sample size is below 500, the GS based minimizer consistently outperformed ADMM by quite a margin. An open and intriguing question is, whether this performance gap narrows or even reverses for large-scale problems.

[3] A comprehensive account of the general idea behind ADMM is provided in Boyd et al. (2010).

Indeed, there are a number of options available to reduce the computational burden when n is large. Most notably, updating the elements of the diagonal matrix $W\left(\Theta^{[k]}\right)$ involved in the construction of the Hessian constitutes a computational bottleneck. Given that the Hessian only determines the size of the Newton-Raphson step, it is possible to use methods requiring fewer arithmetic operations to approximate this matrix. For example, $W\left(\Theta^{[k]}\right)$ can be recomputed at intervals of two or three iteration steps, or alternatively, $W\left(\Theta^{[k]}\right)$ can be replaced by $\sup_{t\in\mathbb{R}} w(t)$ (if finite), or any other reasonable value; see, e.g. Yang and Zou (2013).[4]

Remark 1.3.4 *(Augmentation of the Design Matrix).*

Vertically and horizontally concatenating the original design matrix with an additional column and row allows us to treat the inverse scale parameter γ as a regression coefficient. As such, it is iteratively updated alongside α and β. Moreover, by setting individual penalty loadings, we can exempt γ from regularization at the negligible cost of increasing the cardinality of the set of non-zero coefficients by one. Provided that at least a few outcome observations remain uncensored $\mathbb{E}_n[s_i] \geqslant \underline{c} > 0$ so that $n_0 < n$, $\sigma \geqslant \underline{c} > 0$ and $\gamma \geqslant \underline{c} > 0$, the augmentation does not cause the empirical Hessian to be become degenerate.

In the theoretical analysis of the ℓ_1-penalized Tobit, we impose restrictions on the matrix

$$\begin{bmatrix} \mathbb{E}_n[(d_i,\, x_i)^T(d_i,\, x_i)] & -\mathbb{E}_n[(d_i,\, x_i)^T y_i] \\ -\mathbb{E}_n[y_i(d_i,\, x_i)] & \mathbb{E}_n[y_i^2] \end{bmatrix}. \tag{1.3.5}$$

Specifically, we require its minimal and maximal Cs-sparse eigenvalues for an appropriate constant C to be bounded from below and above, respectively. As opposed to the binary choice logistic model of Belloni et al. (2016a), these conditions need to hold for the augmented Gram matrix, as the additional column associated with y_i is always part of the $(Cs \times Cs)$ sub-matrices. Notably, by the interlacing theorem – an application of the Courant-Fischer theorem – the minimal and maximal sparse eigenvalues of the upper left block associated with $(d_i,\, x_i)$ are bounded by the sparse eigenvalues of the augmented Gram matrix.

[4] In our simulation experiments, the empirical median of $w(t)$ over the n data points performed quite well.

1.4. MAIN THEORETICAL RESULTS

1.4.1. General Tobit Under High-level Conditions

In this section, we establish $\sqrt{n}$ consistency and asymptotic normality for an estimator $\check{\alpha}$ of α_0 associated with the Tobit model (1.2.1) based on high-level assumptions. These conditions impose stringent restrictions on the probability distribution of the disturbances u_i. For a given distribution, these conditions can be verified for a variety of different estimators including the post-ℓ_1 methods described in Algorithms 1.2.1 and 1.2.2. In particular, Theorem 1.4.1 applies to the generic orthogonal score function, which serves as a basis for the construction of instruments z_{0i} and q_{0i} using various machine learning methods, and might therefore be of independent interest.

The estimated quantities $\hat{z}_i = \hat{z}(d_i, x_i)$, $\hat{q}_i = \hat{q}(y_i, d_i, x_i)$ and expectations below are evaluated at the given parameter estimates. Furthermore, let $\delta_n \searrow 0$ and $\Delta_n \searrow 0$ be sequences of positive constants. Additionally, fix some constants $0 < \underline{c} < C < \infty$.

Condition ITob 4.1–(i) restricts the set of potential disturbance distributions by requiring the ratios of CDF, PDF and its derivatives to be bounded. Indeed, this condition is considerably more stringent than the Lipschitz continuity imposed on the link functions for GLMs in Belloni et al. (2016a). Given that $\sup_{t\in\mathbb{R}} |g(t)| \leqslant \bar{L}$ is satisfied if the CDF and PDF of u_i obey the linear differential inequalities $f(t) \leqslant \bar{L} \cdot F(t)$ and $|f'(t)| \leqslant \bar{L} \cdot f(t)$, it must not surprise that the logistic link – a function involving the exponential of a linear argument t – complies with ITob 4.1–(i). With regard to the proof of Theorem 1.4.1, we expect that certain constraints on $g(t)$, $g'(t)$, and $g''(t)$ can be either relaxed, or replaced with requirements that are tailored to the specific distribution under consideration. For example, the normal distribution violates the boundedness of $|g(t)|$, because the inverse Mill's ratio is not bounded from above. However, for the normal distribution, $|g(t)|$ asymptotically grows linearly in t which suggests imposing a growth condition on the largest argument $|t|$ over the n sample points. This is a topic for future research.

ITob 4.1–(ii) requires orthogonality, non-zero variances, and the existence of moments up to the fourth order. ITob 4.1–(iii) requires that the nuisance parameters and instruments are estimated at least at $n^{1/4}$ rate, and that cross-terms are estimated at least at $n^{1/2}$

rate. These constraints can be verified for various machine learning methods beyond the ℓ_1-penalized estimators covered here. As Proposition 1.3.1 provides a weighted least squares representation of our Tobit problem, we can combine the logistic Tobit likelihood loss with other convex penalties such as ridge or elastic net.[5]

Condition ITob 4.1 *(High-Level Assumptions). (i) The data sequence* $m_i = (y_i, d_i, x_i)$ *is independent across* $i \in \{1, \dots, n\}$ *and obeys model (1.2.1), where* $i.i.d.$ *disturbances* u_i *follow the absolutely continuous probability distribution* $\mathrm{P}(\gamma_0 u_i \leqslant t) = F(t)$ *with three times continuously differentiable density* $f(t)$ *and inverse scale parameter* $\gamma_0 = 1/\sigma_0 \geqslant \underline{c} > 0$*, where* $\underline{c} \leqslant \sigma_0^2 \lesssim \mathrm{Var}[u_i] \leqslant C$*. For* $s_i = \mathbf{1}\{y_i > 0\}$*, function*

$$g(t) = (1 - s_i)\frac{f(t)}{F(t)} + s_i \frac{f'(t)}{f(t)}$$

satisfies $\sup_{t\in\mathbb{R}} |g(t)| \leqslant \bar{L}$, $\sup_{t\in\mathbb{R}} |g'(t)| \leqslant \bar{L}'$, $\sup_{t\in\mathbb{R}} |g''(t)| \leqslant \bar{L}''$ *such that* $\bar{L} \vee \bar{L}' \vee \bar{L}'' \leqslant C$*. (ii) Let* $w_i = -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0) > 0$ *be defined as in (1.2.9). The following moment conditions hold* $\mathrm{E}[w_i z_{0i} x_i] = 0$, $\mathrm{E}[w_i(d_i,\, x_i)q_{0i}] = 0$, $|\overline{\mathrm{E}}[w_i z_{0i} d_i]| \geqslant \underline{c}$, $|\overline{\mathrm{E}}[w_i q_{0i} y_i + s_i/\gamma_0^2]| \geqslant \underline{c}$, $\overline{\mathrm{E}}[\{\psi(m_i, \alpha_0, \beta_0, \gamma_0, z_{0i}, q_{0i}, \mu_0)\}^2] \geqslant \underline{c}$, $\overline{\mathrm{E}}[d_i^4] \leqslant C$, $\overline{\mathrm{E}}[z_{0i}^4] \leqslant C$, $\overline{\mathrm{E}}[q_{0i}^4] \leqslant C$, $\overline{\mathrm{E}}[y_i^2] \leqslant C$, $\overline{\mathrm{E}}[y_i^2] \geqslant \underline{c}$, *and* $\overline{\mathrm{E}}[\{x_i\xi\}^4] \leqslant C$ *for all* $\|\xi\| = 1$*. (iii) With probability at least* $1 - \Delta_n$*, the estimates* $\check{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{z}$, $\hat{q}$, $\hat{\mu}$ *satisfy*

$$\begin{aligned}
&\|\hat{\beta} - \beta_0\| \leqslant \delta_n n^{-1/4},\ |\hat{\gamma} - \gamma_0| \leqslant \delta_n n^{-1/4},\ |\hat{\mu} - \mu_0| \leqslant \delta_n,\ \hat{\gamma} \geqslant \underline{c}, \qquad (1.4.1)\\
&\qquad \left\{\overline{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\right\}^{1/2} \leqslant \delta_n,\ \left\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} \leqslant \delta_n,\\
&\|\hat{\beta} - \beta_0\|\left\{\overline{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\right\}^{1/2} \leqslant \delta_n n^{-1/2},\ |\hat{\gamma} - \gamma_0|\left\{\overline{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\right\}^{1/2} \leqslant \delta_n n^{-1/2},\\
&\|\hat{\beta} - \beta_0\|\left\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} \leqslant \delta_n n^{-1/2},\ |\hat{\gamma} - \gamma_0|\left\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} \leqslant \delta_n n^{-1/2},\\
&|\hat{\mu} - \mu_0|\left\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} \leqslant \delta_n n^{-1/2},\ |\hat{\gamma} - \gamma_0|\,|\hat{\mu} - \mu_0| \lesssim \delta_n n^{-1/2},
\end{aligned}$$

$$\sup_{\alpha:|\alpha-\alpha_0|\leqslant\delta_n} \left|(\mathbb{E}_n - \overline{\mathrm{E}})\left[\psi(m_i, \alpha, \beta_0, \gamma_0, z_{0i}, q_{0i}, \mu_0) - \psi(m_i, \alpha, \hat{\beta}, \hat{\gamma}, \hat{z}_i, \hat{q}_i, \hat{\mu})\right]\right| \leqslant \delta_n n^{-1/2}, \qquad (1.4.2)$$

$$|\check{\alpha} - \alpha_0| \leqslant \delta_n, \qquad \left|\mathbb{E}_n[g(\hat{\gamma} y_i - \check{\alpha} d_i - x_i\hat{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu} s_i/\hat{\gamma}]\right| \leqslant \delta_n n^{-1/2}. \qquad (1.4.3)$$

(iv) With probability at least $1-\Delta_n$ *we have* $\|\{\widehat{w}_i - w_i\}\{1 \vee |d_i|\}\|_{2,n} \leqslant \delta_n$, $\|\hat{z}_i - z_{0i}\|_{2,n} \leqslant \delta_n$, $\|d_i(\hat{z}_i - \hat{\mu}\hat{q}_i)\|_{2,n} \leqslant C$, $\|\hat{q}_i - q_{0i}\|_{2,n} \leqslant \delta_n$, $\|x_i\{\hat{\beta} - \beta_0\}(z_{0i} - \mu_0 q_{0i})\|_{2,n} \leqslant \delta_n$, *and* $\|y_i(z_{0i} - \mu_0 q_{0i})\|_{2,n} \leqslant C$.

The following theorem states asymptotic normality of the generic instrumental Tobit estimator based on the high-level assumptions above.

[5] Inspired by the theoretical results of Hsu et al. (2014) for ridge regression, we illustrate the implementation of a ℓ_2-based instrument Tobit in our empirical example in Section 1.5.2.

Theorem 1.4.1 *(Asymptotic Normality Under High-Level Assumptions). Under Condition ITob 4.1–(i),(ii),(iii) we have*

$$\frac{\overline{\mathrm{E}}[w_i d_i z_{0i}]\sqrt{n}(\check{\alpha}-\alpha_0)}{\overline{\mathrm{E}}[\{\psi(m_i,\alpha_0,\beta_0,\gamma_0,z_{0i},q_{0i},\mu_0)\}^2]^{1/2}} = \frac{\sum_{i=1}^{n}\{g(\gamma_0 y_i-\alpha_0 d_i-x_i\beta_0)(z_{0i}-\mu_0 q_{0i})-\mu_0 s_i/\gamma_0\}}{\sqrt{n}\,\overline{\mathrm{E}}[\{\psi(m_i,\alpha_0,\beta_0,\gamma_0,z_{0i},q_{0i},\mu_0)\}^2]^{1/2}}+o_{\mathrm{P}}(1)$$

and

$$\overline{\mathrm{E}}[w_i d_i z_{0i}]\,\overline{\mathrm{E}}[\{\psi(m_i,\alpha_0,\beta_0,\gamma_0,z_{0i},q_{0i},\mu_0)\}^2]^{-1/2}\sqrt{n}\,(\check{\alpha}-\alpha_0)\rightsquigarrow\mathcal{N}(0,1).$$

Additionally, if Condition ITob 4.1–(iv) holds, we can replace the variance by the consistent estimator

$$\frac{\mathbb{E}_n[\{\psi(m_i,\check{\alpha},\hat{\beta},\hat{\gamma},\hat{z}_i,\hat{q}_i,\hat{\mu})\}^2]}{\mathbb{E}_n[\widehat{w}_i d_i\hat{z}_i]^2}=\frac{\overline{\mathrm{E}}[\{\psi(m_i,\alpha_0,\beta_0,\gamma_0,z_{0i},q_{0i},\mu_0)\}^2]}{\overline{\mathrm{E}}[w_i d_i z_{0i}]^2}+o_{\mathrm{P}}(1).$$

1.4.2. Logistic Tobit Under Primitive Conditions

Next, we present and discuss a set of primitive assumptions that imply Condition ITob 4.1 by choosing the logistic distribution. The results in Theorems 1.4.2 and 1.4.3 immediately follow from Theorem 1.4.1 and the definitions of instruments $z_{0i}=d_i-x_i\eta_0$ and $q_{0i}=y_i-(d_i,\,x_i)\theta_0$ in (1.2.12) and (1.2.13).

Condition ITob 4.2 *(Primitive Assumptions). (i) The data sequence $m_i=(y_i,d_i,x_i)$ is independent across $i\in\{1,\ldots,n\}$ and obeys models (1.2.1), (1.2.10), and (1.2.11) with $i.i.d.$ logistic disturbances $\mathrm{P}(\gamma_0 u_i\leqslant t)=F(t)=(1+\exp\{-t\})^{-1}$, where the scale parameter satisfies $\sigma_0\geqslant\underline{c}$ and $1/\sigma_0:=\gamma_0\geqslant\underline{c}$. Let functions $g(t)$ and $g'(t)$ be given by (1.2.18). There exists $s=s_n$, so that $\|\beta_0\|_0+\|\eta_0\|_0+\|\theta_0\|_0\leqslant s$ and $|\alpha_0|+\|\beta_0\|+\|\eta_0\|+\|\theta_0\|\leqslant C$. (ii) The following moment conditions hold $\overline{\mathrm{E}}[\{(d_i,\,x_i,\,-y_i)\xi\}^4]\leqslant C\|\xi\|^4$, $\overline{\mathrm{E}}[w_i\{(d_i,\,x_i,\,-y_i)\xi\}^2]\geqslant\underline{c}\|\xi\|^2$, $\overline{\mathrm{E}}[s_i]\geqslant\underline{c}$, and $\mathrm{E}[\min_{1\leqslant i\leqslant n}w_i]\geqslant\underline{c}$, where w_i is defined as in (1.2.9). Additionally, we have that $\min_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[w_i^2x_{ij}^2z_{0i}^2]\geqslant\underline{c}$, $\overline{\mathrm{E}}[w_i^2d_i^2q_{0i}^2]\wedge\min_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[w_i^2x_{ij}^2q_{0i}^2]\geqslant\underline{c}$, $\big\{\max_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[|w_ix_{ij}z_{0i}|^3]^{1/3}\big\}\sqrt{\log(p\vee n)}\leqslant\delta_n n^{1/6}$, $\big\{\overline{\mathrm{E}}[|w_id_iq_{0i}|^3]^{1/3}\vee\max_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[|w_ix_{ij}q_{0i}|^3]^{1/3}\big\}\sqrt{\log(p\vee n)}\leqslant\delta_n n^{1/6}$. (iii) The following holds $K_1^2s^2\log^2(p\vee n)\leqslant\delta_n n$ and $K_4^4 s\log(p\vee n)\log^3(n)\leqslant\delta_n n$, where $K_a=\mathrm{E}[\max_{1\leqslant i\leqslant n}\|(d_i,\,x_i,\,y_i,\,z_{0i},\,q_{0i})\|_\infty^a]^{1/a}$ and $p=p_n$.*

The stated assumptions are almost identical to those of the logsitic GLM; see Belloni et al. (2016a). Specifically, Condition ITob 4.2–(i) assumes an independent random sample, a logistic disturbance distribution and sparsity of the nuisance terms which allows α_0 to be consistently estimated in a $p\gg n$ setting. ITob 4.2–(ii) imposes moment restrictions, re-

quires that the censoring probabilities stay away from one, and that weights w_i stay away from zero. ITob 4.2–(iii) imposes growth conditions on the triple (s, p, n). Essentially, these assumptions ensure that the condition numbers of augmented sub-matrices of the empirical Gram matrix in (1.3.5) are bounded even if the complete matrix is rank deficient as $p > n$; see Rudelson and Vershynin (2008) for a general discussion. The boundedness of minimal and maximal sparse eigenvalues guarantees that the absolute distance between estimated and true model parameters is bounded by the rate for λ/n.

The following theorem states asymptotic normality of the instrumental logistic Tobit estimator based on the assumptions above.

Theorem 1.4.2 *(Inference on α_0 Based on Asymptotic Normality of Instrumental Logistic Tobit). Assume that an triangular array of data $(y_i, d_i, x_i)_{1 \leqslant i \leqslant n}$ satisfies Condition ITob 4.2 for all $n \geqslant 1$. As $n \to \infty$, the logistic Tobit instrument estimator for α_0, outlined in Algorithm 1.2.1 and Remark 1.2.1, follows*

$$\Sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) = Z_n + o_{\mathrm{P}}(1) \; with \; Z_n \rightsquigarrow \mathcal{N}(0, 1),$$

where

$$Z_n := \frac{\Sigma_n}{\sqrt{n}} \sum_{i=1}^{n} g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0) z_{0i} \; and \; \Sigma_n^2 := \overline{\mathrm{E}}[w_i z_{0i}^2]^{-1}.$$

Moreover, Σ_n^2 can be replaced by either $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\widehat{w}_i d_i \hat{z}_i]^{-2}\mathbb{E}_n[\{g(\tilde{\gamma} y_i - \check{\alpha} d_i - x_i\tilde{\beta})\hat{z}_i\}^2]$ in the case of Algorithm 1.2.1–(a) and $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\widehat{w}_i d_i \hat{z}_i]^{-2}\mathbb{E}_n[\{g(\tilde{\gamma} y_i - \check{\alpha} d_i - x_i\tilde{\beta})(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu} s_i/\tilde{\gamma}\}^2]$ for the alternative instrument in Algorithm 1.2.1–(b), or by $\widehat{\Sigma}_{2n}^2 = \mathbb{E}_n[\widehat{w}_i \hat{z}_i^2]^{-1}$ without affecting the result.

Theorem 1.4.2 formalizes that the studentized point estimator $\sqrt{n}\,\widehat{\Sigma}_n^{-1}(\check{\alpha} - \alpha_0)$ converges in distribution to a standard normal one. Importantly, this result holds although the ℓ_1-penalized estimator misses "weak signals" and small coefficients are erroneously treated as zeros. As a consequence, the model framework outlined in Section 1.2.1 can be generalized to that of a partially linear model with censored outcomes:

$$\begin{aligned} \gamma_0 y_i &= \max\{\alpha_0 d_i + M_1(x_i) + \gamma_0 u_i \, , \, \mathcal{Y}_i\} \\ &= \max\{\alpha_0 d_i + \Pi_1(x_i)\beta_0 + \epsilon_{yi} + \gamma_0 u_i \, , \, \mathcal{Y}_i\}, \end{aligned} \tag{1.4.4}$$

where $\mathrm{E}[u_i \mid d_i, x_i] = 0$ and $\Pi_1(x_i)$ represents a rich dictionary of non-linear transformations, e.g. polynomials, splines, trigonometric functions, interactions or interval indicators of the basic controls x_i necessary to adequately approximate $M_1(x_i)$ by a parametric function that is linear in nuisance parameter β_0 with $\|\beta_0\|_0 \leqslant s$. Here, ϵ_{yi} quantifies the approximation error; see, e.g. Belloni et al. (2011, 2014) for a detailed discussion on partially linear

models and approximate sparsity. Furthermore, we impose the same approximately sparse structure on the weighted decomposition of the target regressor:

$$\sqrt{w_i}d_i \;=\; \sqrt{w_i}M_2(x_i) + v_i = \sqrt{w_i}\Pi_2(x_i)\eta_0 + \epsilon_{di} + v_i, \tag{1.4.5}$$

where $\mathrm{E}[\sqrt{w_i}v_i\Pi_2(x_i)] = 0$ and $\|\eta_0\|_0 \leqslant s$. Analogously to the logistic regression case, the result of Theorem 1.4.2 continues to hold under approximate sparsity provided that the approximation errors satisfy

$$\overline{\mathrm{E}}[\epsilon_{yi}^2]^{1/2} \lesssim \sqrt{s/n}, \quad \overline{\mathrm{E}}[\epsilon_{di}^2]^{1/2} \lesssim \sqrt{s/n}, \text{ and } \left|\overline{\mathrm{E}}[\sqrt{w_i}v_i\epsilon_{yi}]\right| \leqslant \delta_n n^{-1/2};$$

see Belloni et al. (2014, 2016a, 2019).

Lastly, the following theorem states asymptotic normality of the post-double selection logistic Tobit estimator based on Condition ITob 4.2.

Theorem 1.4.3 *(Inference on α_0 Based on Asymptotic Normality of Post-Double Selection Logistic Tobit). Assume that an triangular array of data $(y_i, d_i, x_i)_{1\leqslant i\leqslant n}$ satisfies Condition ITob 4.2 for all $n \geqslant 1$. As $n \to \infty$, the post-double selection Tobit estimator for α_0, outlined in Algorithm 1.2.2, follows*

$$\Sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) = Z_n + o_{\mathrm{P}}(1) \; with \; Z_n \rightsquigarrow \mathcal{N}(0,1),$$

where

$$Z_n := \frac{\Sigma_n}{\sqrt{n}}\sum_{i=1}^{n}\left\{g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(z_{0i} - \mu_0 q_{0i}) - \mu_0 s_i/\gamma_0\right\} \; and \; \Sigma_n^2 := \overline{\mathrm{E}}[w_i z_{0i}^2]^{-1}.$$

Moreover, Σ_n^2 can be replaced by either $\widehat{\Sigma}_{1n}^2 = \mathbb{E}_n[\check{w}_i d_i \hat{z}_i]^{-2}\mathbb{E}_n[\{g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})(\hat{z}_i - \check{\mu}\hat{q}_i) - \check{\mu}s_i/\check{\gamma}\}^2]$ or by

$$\widehat{\Sigma}_{2n}^2 = \left\{\mathbb{E}_n\big[\check{w}_i(d_i,\, x_{i,\check{\mathcal{J}}},\, -y_i)^T(d_i,\, x_{i,\check{\mathcal{J}}},\, -y_i)\big] + \begin{bmatrix}\mathbf{0}_{\check{s}+1} & 0\\ 0 & \mathbb{E}_n[s_i]/\check{\gamma}^2\end{bmatrix}\right\}^{-1}_{1,1},$$

where $\check{w}_i = -g'(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})$, and $\check{\mu} = \mathbb{E}_n[\check{w}_i\hat{q}_i y_i + s_i/\check{\gamma}^2]^{-1}\mathbb{E}_n[\check{w}_i\hat{z}_i y_i]$ without affecting the result.

1.5. EMPIRICAL PERFORMANCE

1.5.1. Monte Carlo Experiments: Uniformity

In this section, we summarize the results of extensive Monte Carlo experiments. We contrast the finite sample behaviour of the instrumental and post-double selection estimators discussed in Section 1.2.2 with the post-naïve selection approach, which is defined as running a logistic Tobit based on the set of control variables selected by the ℓ_1-penalized Tobit in Step 1 of Algorithms 1.2.1 and 1.2.2.

Our simulations are based on the set-up used by Belloni et al. (2011, 2016a, 2019) to assess the finite sample performance of high-dimensional linear, logistic (binary choice) and approximately sparse quantile regression, respectively. The data generating processes (DGP) are given by:

$$y = \max\{\alpha_0 d + x(c_y\nu_y) + \tilde{u}\ ,\ \tilde{y}\}, \tag{1.5.1}$$
$$d = x(c_d\nu_d) + \tilde{v}, \tag{1.5.2}$$

where the coefficient vectors ν_y and ν_d are set to

$$\nu_y = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, 0, \ldots, 0)^T,$$
$$\nu_d = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, 0, \ldots, 0)^T$$

for all cases with exact sparsity. In a second set of simulations, we also consider an alternative, approximately sparse design, where the coefficients

$$(\nu_y)_j = (\nu_d)_j = j^{-2} \text{ for } j \in \{1, \ldots, p\}$$

of all features are allowed to be different from zero but decay polynomially.[6] The dimension of the control variables is $p = 250$ and the sample size equals $n = 200$. We include an intercept $x = (1,\, x^*)$ and the controls are jointly normal $x^* \sim \mathcal{N}(0, \Sigma)$, where the covariance matrix exhibits a Toeplitz structure, i.e., $\Sigma_{i,j} = \varphi^{|i-j|}$ for all $i, j \in \{1\ \ldots, p\}$ and $\varphi \in (-1, 1)$. In accordance with Belloni et al. (2011, 2016a, 2019) we choose $\varphi = 0.5$ for all designs. The censoring threshold $\tilde{y}$ is adjusted to induce a specific censoring rate $c \in (0, 1)$. For example, with the above DGPs, $\tilde{y} = 0$ provokes a censoring rate of 50%, whereas $\tilde{y} < 0$

[6] The reader is referred to Belloni et al. (2014) for a thorough discussion of "smoothness" requirements and general information on the rate of decay of sorted coefficient vectors in approximately sparse designs.

($\tilde{y} > 0$) leads to $c < 0.5$ ($c > 0.5$).

The error $\tilde{v}$ is $i.i.d.$ standard normal noise $\tilde{v} \sim \mathcal{N}(0, 1)$, while $\tilde{u}$ satisfies the model assumptions in being $i.i.d.$ logistic noise. We standardize $\tilde{u}$ to have a unit variance by setting the scale parameter equal to $\sigma_0 = \sqrt{3}/\pi$. Standardizing the error distributions of both $\tilde{u}$ and $\tilde{v}$ does not affect our simulation results as we introduce two scaling factors $c_y > 0$ and $c_d > 0$. Scalar c_d is used to control the coefficient of determination R^2, denoted by R_d^2, in the auxiliary equation which relates the main regressor d to the high-dimensional control variables, and c_y is used to control the R^2, denoted by R_y^2, in the reduced regression equation: $y^* - \alpha_0 d = x(c_y \nu_y) + \tilde{u}$.

In the exactly sparse scenario, the predictive power of controls x^* rapidly decays and the smallest non-zero coefficients are difficult to differentiate from zero for the given sample size. As a consequence, ℓ_1-based methods invariably miss "weak signals" and we expect these selection mistakes to translate into a substantial bias of the post-naïve approach. Moreover, to demonstrate that the derived statistical inference for our proposed estimators is asymptotically uniformly valid, we systematically vary c_d and c_y, and experiment with a wide range of DGPs, where model selection becomes more or less difficult for ℓ_1-penalized methods. As shown below, the performance of the post-naïve estimator varies considerably across these designs. As opposed to that, our immunized (orthogonal) approaches are invariant and, thus, perform similarly well over this large collection of DGPs.

In total, we consider the grid

$$\left(R_y^2,\, R_d^2\right) \in \{0, 0.1, 0.2, \ldots, 0.9\}^2 \tag{1.5.3}$$

of 100 DGPs and vary the size of the target coefficient $\alpha \in \{-0.25, 0, 0.25, 0.5\}$ for three degrees of censoring severity $c \in \{0.25, 0.5, 0.75\}$. For conciseness, we limit ourselves to presenting the most interesting insights but emphasize that there were no meaningful differences between the results for these designs. Simulation results are reported for:

(a) the naïve post-selection estimator – the estimator of α_0 based on a logistic Tobit after the naïve selection step using the ℓ_1-penalized Tobit,

(b) the instrument estimator – the estimator of α_0 based on the instrumental logistic Tobit described in Algorithm 1.2.1 (a), and

(c) the post-double selection estimator – the estimator of α_0 based on the logistic Tobit after double selection using the ℓ_1-penalized Tobit in Step 1 and both weighted post-Lassos in Steps 2 (a), (b) of Algorithm 1.2.2.

Table 1.1: Summary of the simulation illustrations in Figures 1.1–1.4

estimator	bias	RMSE	variance	median	$\text{rp}_{0.05}$
	exact sparsity: $\alpha_0 = 0.2$, $\tilde{y} = 0$, $c_d = 1$, $c_y = 0.75$				
post-naïve selection	0.2662	0.3337	0.0405	0.4617	0.5093
instrument Algorithm 1.2.1	0.0274	0.1638	0.0261	0.2274	0.0679
post-double Algorithm 1.2.2	0.0376	0.1715	0.028	0.2388	0.0684
	approx. sparsity: α_0 = -0.25, $\tilde{y} = 0$, $c_d = 1.43$, $c_y = 0.48$				
post-naïve selection	0.248	0.3047	0.0313	0.0463	0.5643
instrument Algorithm 1.2.1	0.0232	0.1442	0.0202	-0.2237	0.0517
post-double Algorithm 1.2.2	0.0196	0.1413	0.0196	-0.2297	0.0564

Note: This table summarizes the bias $\mathbb{E}_m[\check{\alpha}_r - \alpha_0]$, root mean squared error (RMSE) $\|\check{\alpha}_r - \alpha_0\|_{2,m}$, variance $\|\check{\alpha}_r - \mathbb{E}_m[\check{\alpha}_r]\|_{2,m}^2$, median point estimate, and the rejection frequency at 5% level ($\text{rp}_{0.05}$) of post-naïve, post-double, and instrumental Tobit estimators for target parameter α_0. The size of the simulated samples is $n = 200$, while the number of potential controls equals $p = 250$. In the exactly sparse szenario, the number of true non-zero coefficients is set to $s = 10$. In the approximatedly sparse setting all 250 coefficients are different from zero but decay polynomially. The respective distributions associated with the first DGP are illustrated in Figures 1.1 and 1.2, while those for the second DGP are displayed in Figures 1.3 and 1.4.

For each replication of the simulations we draw new realizations for the control variables x^* and errors $\tilde{u}$ and $\tilde{v}$. To assess the performance of the estimators, we conduct the standard hypothesis test for the true value α_0 and "studentize" the point estimates by computing the t-statistic:

$$t_r = \frac{\check{\alpha}_r - \alpha_0}{(\widehat{\Sigma}_n)_r} \text{ for } r \in \{1, \dots, m\}, \tag{1.5.4}$$

where $(\widehat{\Sigma}_n)_r$ represents the estimated standard deviation of the respective estimator for replication r out of m simulation repetitions. Since t_r is asymptotically standard normal distributed by Theorems 1.4.2 and 1.4.3 presented in the previous section, we compare the empirical distributions of the test statistics to a standard normal one. In particular, we count the share of simulations in which the true value α_0 is rejected based on the typical 1.96 critical value of the standard normal distribution and call this the rejection frequency at a 5% nominal level ($\text{rp}_{0.05}$).

Figures 1.1 and 1.2 provide a first visual impression of our simulation results, where we replicate the example in Belloni et al. (2016a) with $\alpha_0 = 0.2$, $R_d^2 = R_y^2 = 0.75$, induced by setting $c_d = 1$, $c_y = 0.75$, and $\tilde{y} = 0$ ($\mathscr{c} = 0.5$). Additionally, Figures 1.3 and 1.4 illustrate the simulation results for an approximately sparse design with $\alpha_0 = -0.25$, $R_d^2 = 0.75$,

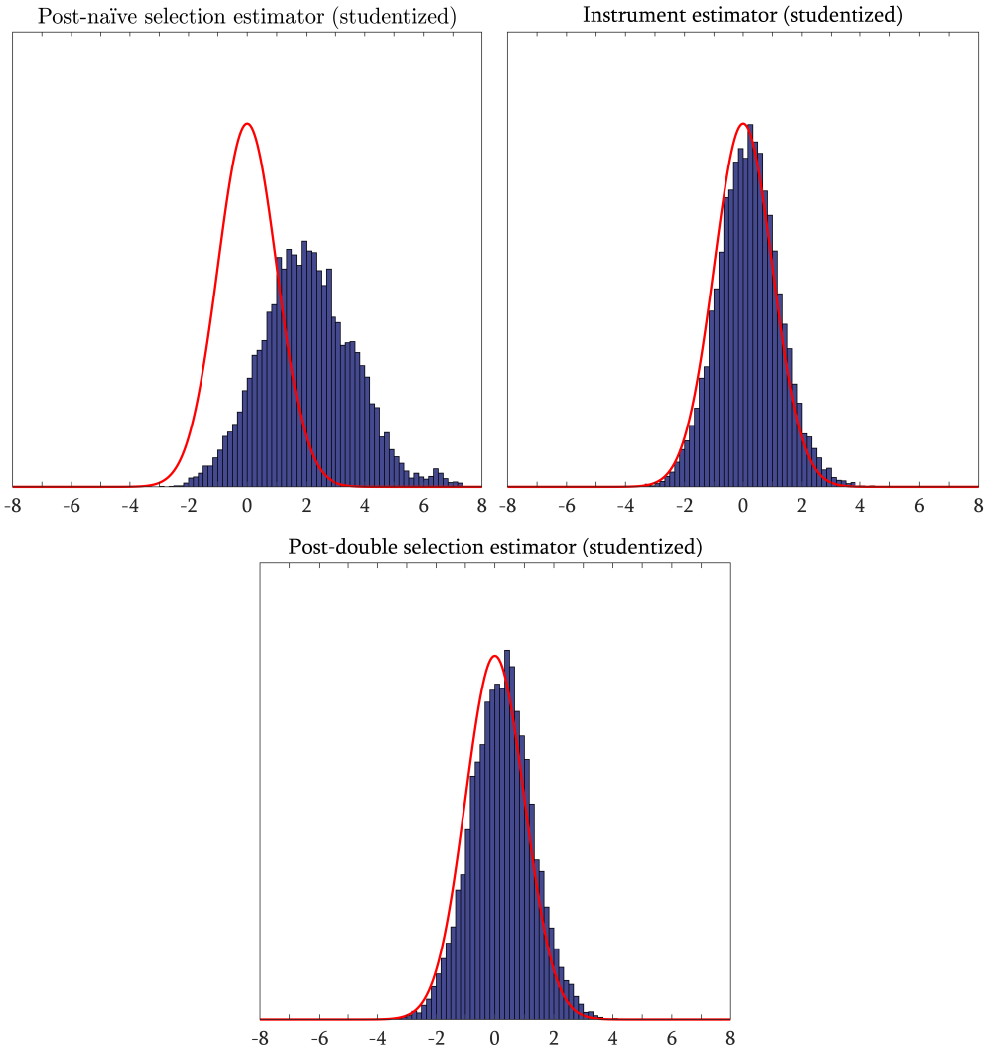


Figure 1.1: Histograms of studentized estimators for α_0 with exactly sparse DGP

Note: These histograms display the simulated distributions of the post-naïve selection estimator (top-left panel), the instrumental Tobit in Algorithm 1.2.1 (top-right panel), and the post-double selection estimator in Algorithm 1.2.2 (bottom panel) for the DGP example in Belloni et al. (2016a) with $\alpha_0 = 0.2$, $\tilde{y} = 0$ (censoring rate of 50%), $c_d = 1$, and $c_y = 0.75$. For comparison, the solid red line represents a standard normal PDF. Additional information on bias, RMSE, variance, median point estimate, and rejection frequencies at 5% nominal level is provided in Table 1.1.

$R_y^2 = 0.25$, induced by setting $c_d = 1.43$, $c_y = 0.48$, and $\tilde{y} = 0$ ($c = 0.5$). These additional simulations indicate that our results presented above are robust concerning moderate violations of the assumption of exact sparsity. All histograms in Figures 1.1, 1.3 and the normal-quantile plots in Figures 1.2, 1.4 are based on $m = 10^4$ Monte Carlo replications.

The distributions associated with the post-naïve estimator displayed in the top-left tile of Figures 1.1 and 1.3 differ considerably from a standard normal distribution (solid red line). In fact, these distributions are not only shifted to the right but also appear flat and bimodal.

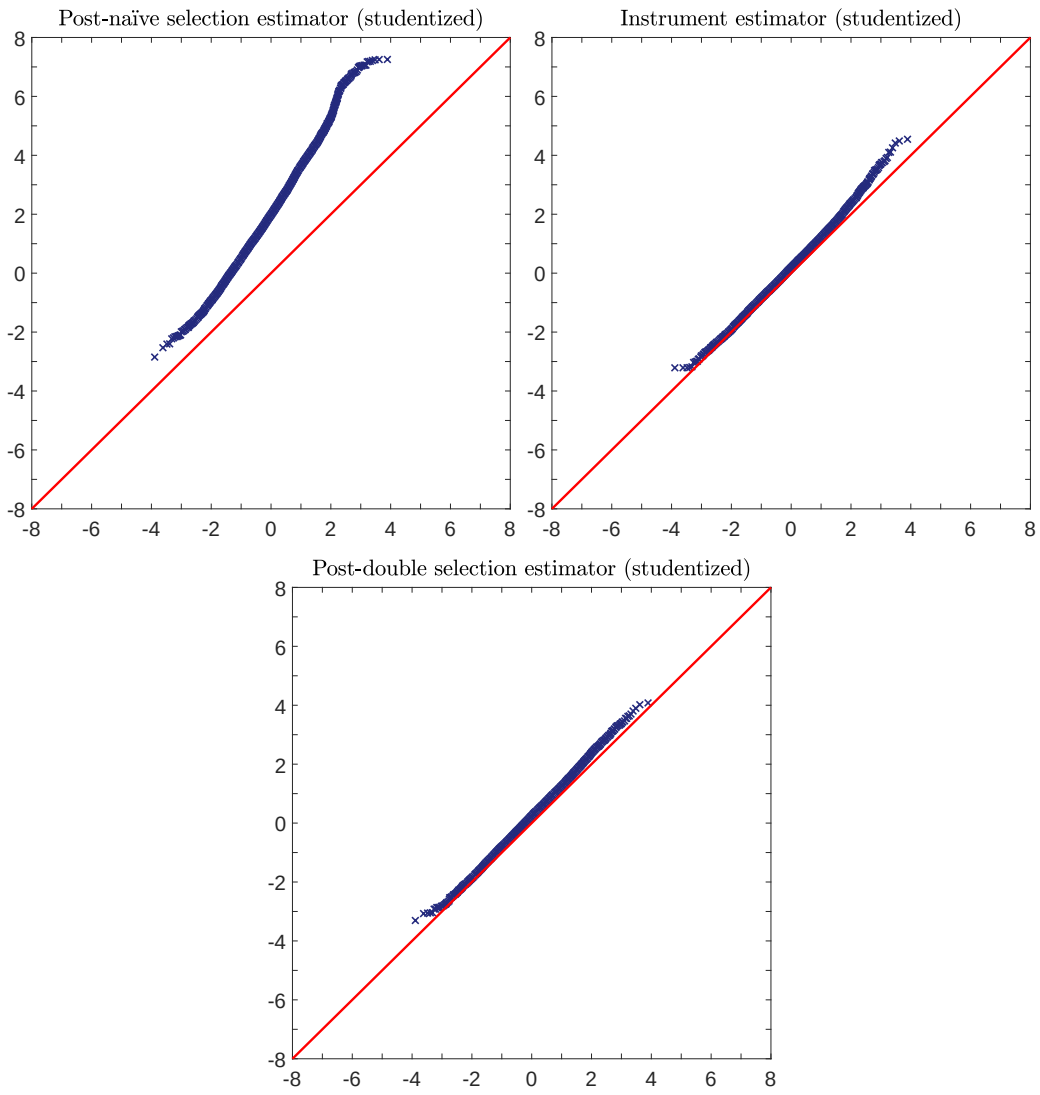


Figure 1.2: Normal-quantile plots of estimators for α_0 with exactly sparse DGP

Note: These normal-quantile plots display the quantiles of the simulated distributions of the post-naïve selection estimator (top-left panel), the instrumental Tobit in Algorithm 1.2.1 (top-right panel), and the post-double selection estimator in Algorithm 1.2.2 (bottom panel) for the DGP example in Belloni et al. (2016a) with $\alpha = 0.2$, $\tilde{y} = 0$ (censoring rate of 50%), $c_d = 1$, and $c_y = 0.75$. For comparison, the quantiles of a standard normal distribution follow the solid red diagonal line.

This occurs because the DGPs do not allow for perfect model selection. As a consequence, the naïve point estimates are systematically biased and have a large root mean squared error (see columns bias and RMSE in Table 1.1). In particular, studentizing the point estimates

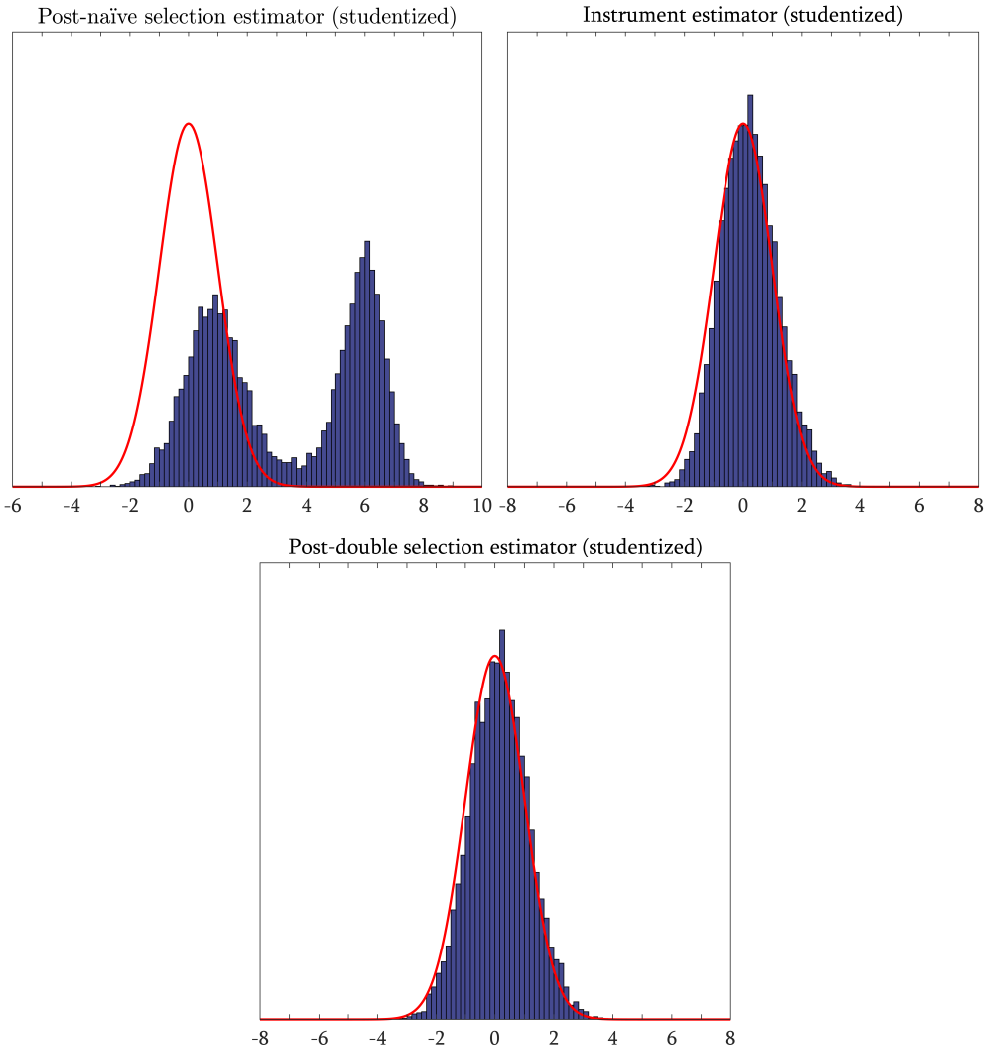


Figure 1.3: Histograms of studentized estimators for α_0 with approximately sparse DGP

Note: These histograms display the simulated distributions of the post-naïve selection estimator (top-left panel), the instrumental Tobit in Algorithm 1.2.1 (top-right panel), and the post-double selection estimator in Algorithm 1.2.2 (bottom panel) for the second DGP example with $\alpha_0 =$ -0.25, $\tilde{y} = 0$ (censoring rate of 50%), $c_d = 1.43$, and $c_y = 0.48$. For comparison, the solid red line represents a standard normal PDF. Additional information on bias, RMSE, variance, median point estimate, and rejection frequencies at 5% nominal level is provided in Table 1.1.

with the naïve variance estimator does not result in a sequence with unit variance.[7] Thus, the distributions of the naïve test statistic appear platykurtic. Such shapes of the finite sam-

[7]The naïve variance estimator is given by

$$\widehat{\Sigma}_n^2 = \left\{ \mathbb{E}_n\big[\widehat{w}_i(d_i,\, x_{i,\widehat{J}},\, -y_i)^T(d_i,\, x_{i,\widehat{J}},\, -y_i)\big] + \begin{bmatrix} \mathbf{0}_{\widehat{s}+1} & 0 \\ 0 & \mathbb{E}_n[s_i]/\tilde{\gamma}^2 \end{bmatrix} \right\}^{-1}_{1,1},$$

where $\widehat{s} = \text{card}(\widehat{J})$ and $\mathbf{0}_{\widehat{s}+1}$ is a $(\widehat{s}+1)$-dimensional quadratic zero matrix. The naïve confidence regions below are constructed as $\{\alpha \in \mathbb{R} : |\alpha - \tilde{\alpha}| \leqslant \widehat{\Sigma}_n \Phi^{-1}(1-\zeta/2)/\sqrt{n}\}$.

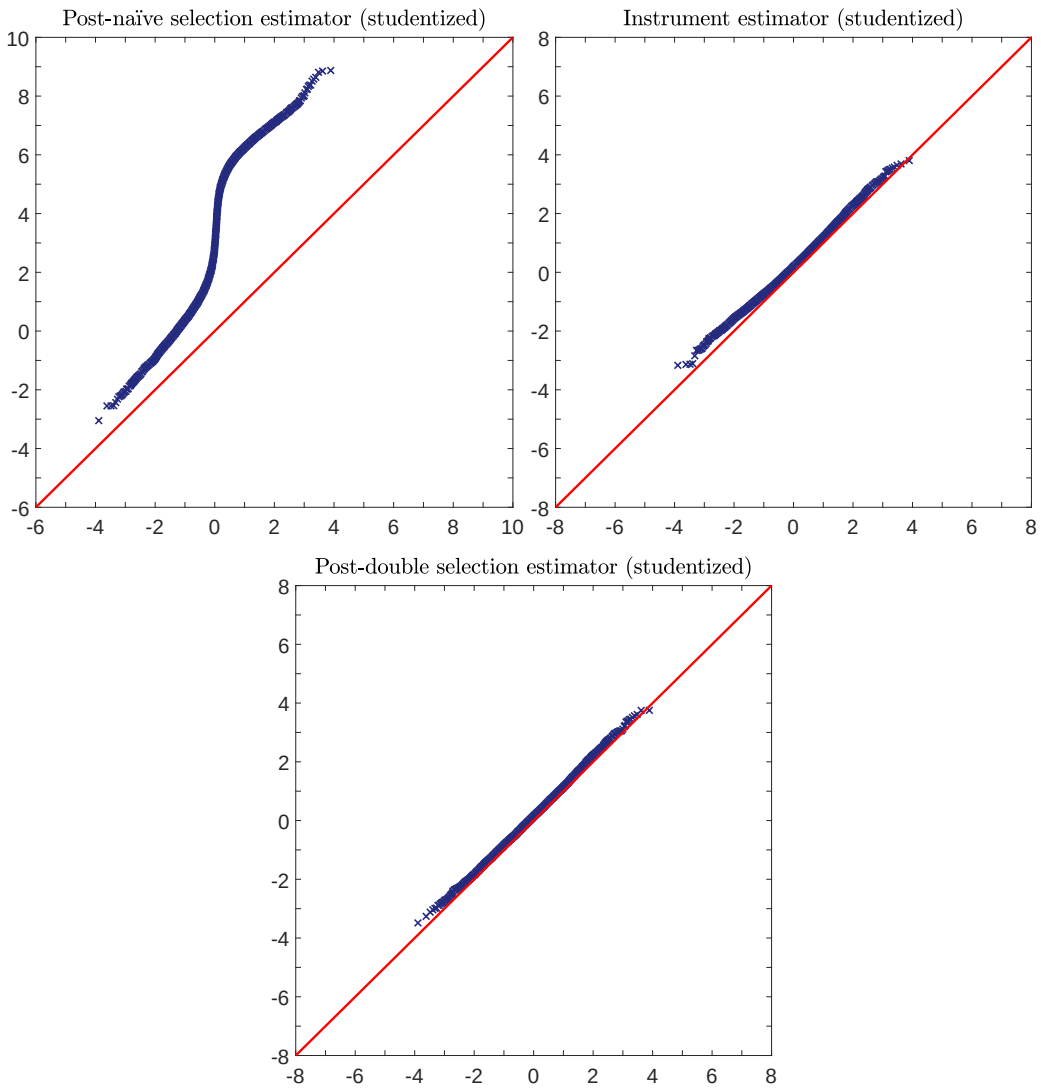


Figure 1.4: Normal-quantile plots of estimators for α_0 with approximately sparse DGP

Note: These normal-quantile plots display the quantiles of the simulated distributions of the post-naïve selection estimator (top-left panel), the instrumental Tobit in Algorithm 1.2.1 (top-right panel), and the post-double selection estimator in Algorithm 1.2.2 (bottom panel) for the second DGP example with α = -0.25, $\tilde{y} = 0$ (censoring rate of 50%), $c_d = 1.43$, and $c_y = 0.48$. For comparison, the quantiles of a standard normal distribution follow the solid red diagonal line.

ple distributions of naïve methods are in line with theoretical arguments given in Leeb and Pötscher (2005) for linear models and have also been observed for other models; see, e.g. Belloni et al. (2016a, 2019). If we tested hypotheses at a 5% level based on the quantiles of a normal distribution, we would reject the true value in about 50.9% and 56.4% of the cases (see column rp(0.05) in Table 1.1).

As opposed to that, the instrumental and post-double selection Tobit estimators have a small bias and low RMSE. Notably, the biases of the instrumental and post-double To-

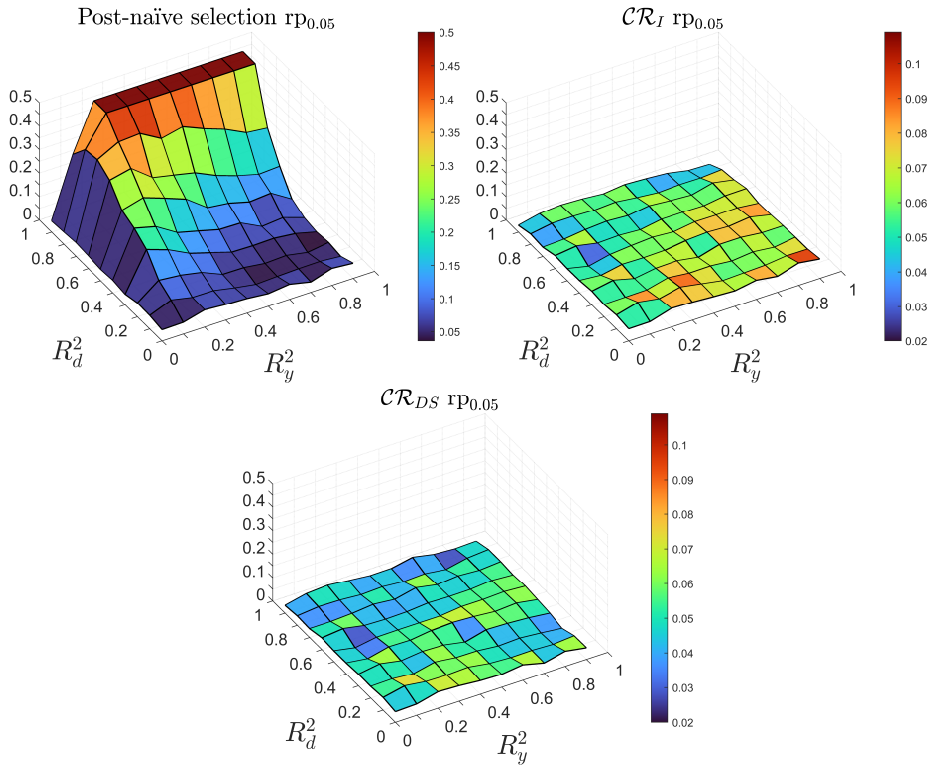


Figure 1.5: Surface plots of rejection frequencies of $H_0 : \alpha = \alpha_0$ under exactly sparse, 25% censored DGP

Note: These surface plots display the rejections frequencies at 0.05 level ($rp_{0.05}$) of the confidence regions based on the post-naïve selection (top-left panel), the instrumental Tobit in Algorithm 1.2.1 $\mathcal{CR}_I$ (top-right panel), and the post-double selection $\mathcal{CR}_{DS}$ (bottom panel). The ideal plot would show a flat horizontal plane at height 0.05. The grid comprises 100 different DGPs, where the assumption of exact sparsity holds, with $\alpha_0 = 0.5$ and censoring rate $c = 0.25$. The rejection frequencies are based on $m = 1000$ replications for each design.

bit estimators for the first design match those reported in Belloni et al. (2016a) for logistic regression. The variances (and RMSEs) of the Tobit estimators, however, are considerably lower than for the Logit estimators. This suggests an efficiency gain compared to logistic regression – at least if the assumptions of model 1.2.1 are satisfied. The median point estimates roughly coincide with the true value α_0 (see column median in Table 1.1). Most importantly, we obtain reliable inferential results since the true value is rejected in between 5.2% and 6.8% of the cases, which comes close to the envisaged 5% level.

To study the performance of these four estimators more systematically, we now turn to the simulation results from our grid of 100 different (R_y^2, R_d^2) points. To visualize our results, we compute the rejection frequencies of the t-statistic in (1.5.4) at a 5% level based on $m = 1000$ replications for all 100 grid points and display these in a surface plot. For example, $\mathcal{CR}_{DS}$ $rp_{0.05}$ denotes the empirical rejection frequencies based on the confidence

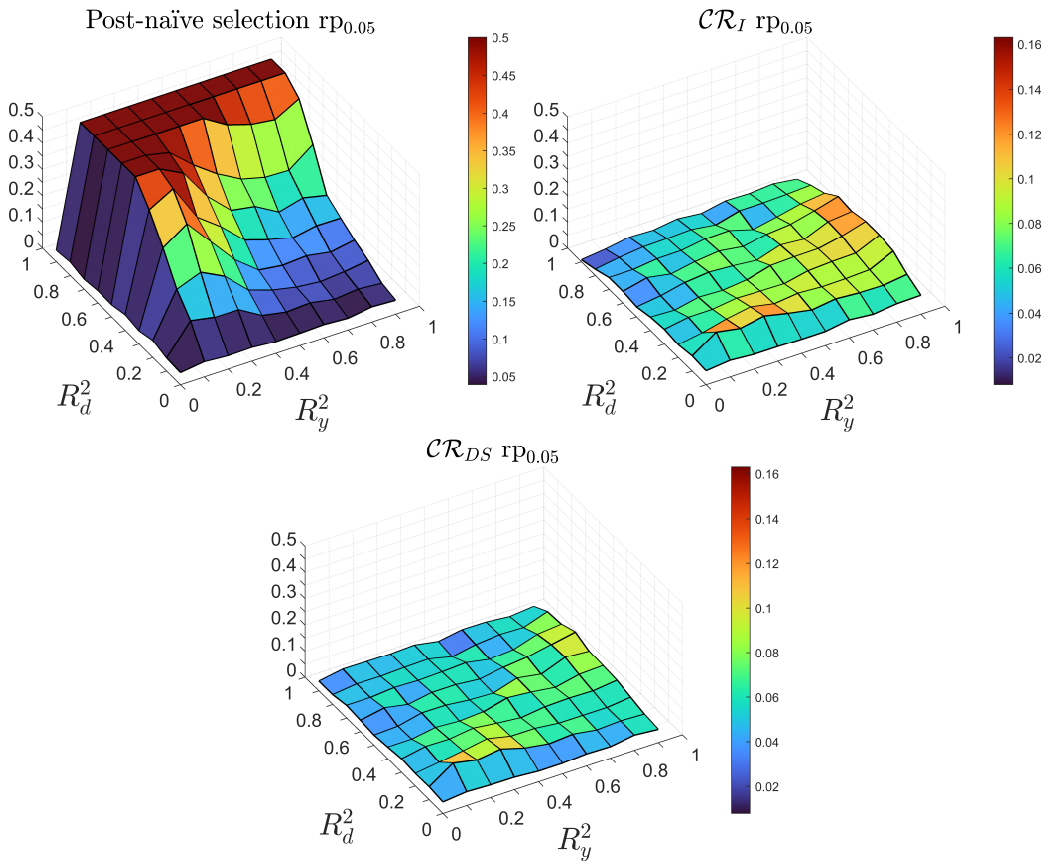


Figure 1.6: Surface plots of rejection frequencies of $\mathrm{H}_0 : \alpha = \alpha_0$ under exactly sparse, 50% censored DGP

Note: These surface plots display the rejections frequencies at 0.05 level ($\mathrm{rp}_{0.05}$) of the confidence regions based on the post-naïve selection (top-left panel), the instrumental Tobit in Algorithm 1.2.1 $\mathcal{CR}_I$ (top-right panel), and the post-double selection $\mathcal{CR}_{DS}$ (bottom panel). The ideal plot would show a flat horizontal plane at height 0.05. The grid comprises 100 different DGPs, where the assumption of exact sparsity holds, with $\alpha_0 = 0$ and censoring rate $c = 0.5$. The rejection frequencies are based on $m = 1000$ replications for each design.

interval of the post-double selection estimator in Step 4 of Algorithm 1.2.2 at a 5% confidence level. The optimal plot associated with an oracle estimator would show a horizontal plane at height 0.05.

As expected, the rejection frequencies of the post-naïve selection methods in the top-left tile of Figures 1.5–1.8 exceed the nominal 0.05 level by a substantial margin over large parts of the grid. The stronger the correlation between the target regressor of interest d and the high-dimensional controls x, the more the simulated rejection frequencies deviate away from the ideal plane. By contrast, the confidence regions of the instrumental Tobit in the top-right tile and the post-double selection Tobit in the bottom tile resemble flat surfaces close to the desired height of 0.05. The rejection frequencies of the majority of grid points

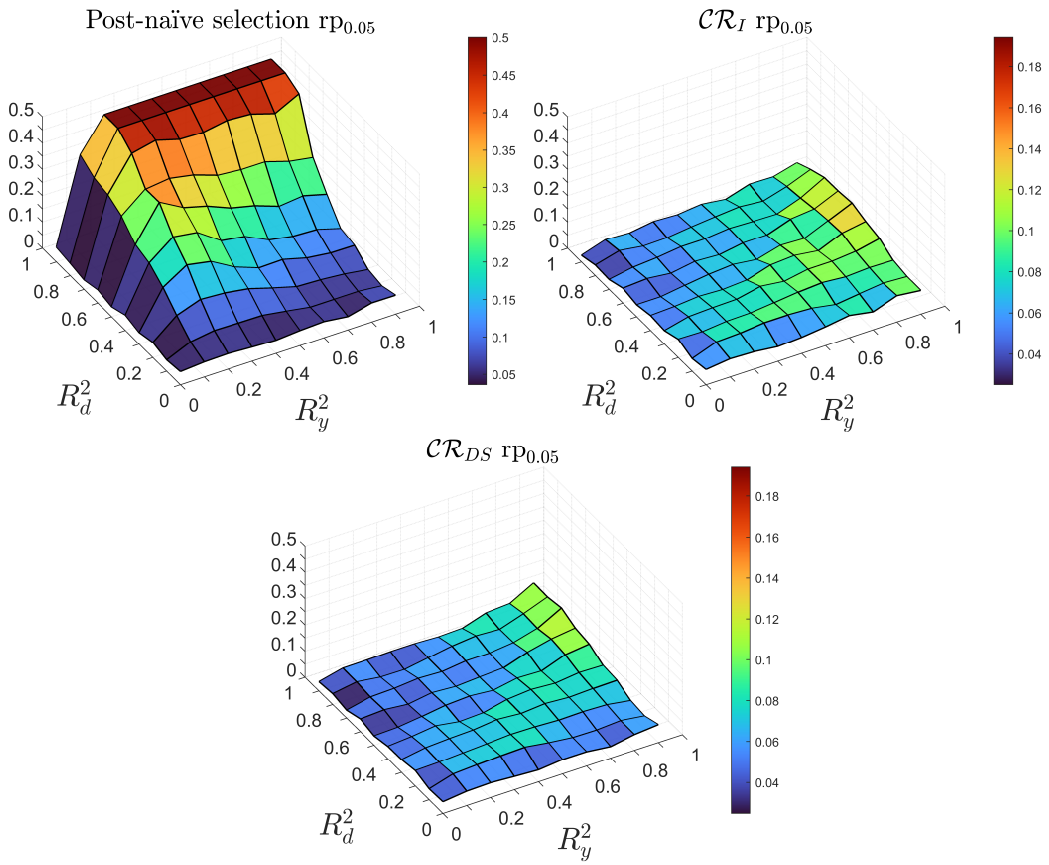


Figure 1.7: Surface plots of rejection frequencies of $\mathrm{H}_0 : \alpha = \alpha_0$ under exactly sparse, 75% censored DGP

Note: These surface plots display the rejections frequencies at 0.05 level ($\mathrm{rp}_{0.05}$) of the confidence regions based on the post-naïve selection (top-left panel), the instrumental Tobit in Algorithm 1.2.1 $\mathcal{CR}_I$ (top-right panel), and the post-double selection $\mathcal{CR}_{DS}$ (bottom panel). The ideal plot would show a flat horizontal plane at height 0.05. The grid comprises 100 different DGPs, where the assumption of exact sparsity holds, with $\alpha_0 = 0.25$ and censoring rate $c = 0.75$. The rejection frequencies are based on $m = 1000$ replications for each design.

fall within the range of 0.03 to 0.08, with the exception of a few isolated instances where the rejection frequencies exceed 0.1. Notably, these outliers occur for extreme DGPs with particularly high values of R_y^2 or nearly orthogonal designs with R_d^2 values close to zero. (Belloni et al. 2016a, 2019 report similar ranges for instrumental Logit and post-double quantile regressions.)

A comparison of the simulated rejection frequencies of DGPs with different censoring rates (25% to 50% and 75%) reveals no discernible difference in the quality of the inferential results of our Tobit estimators.[8]

[8]We remark that the current framework does not allow to investigate the behaviour of our estimators for censoring rates approaching one, i.e., $c \to 1$.

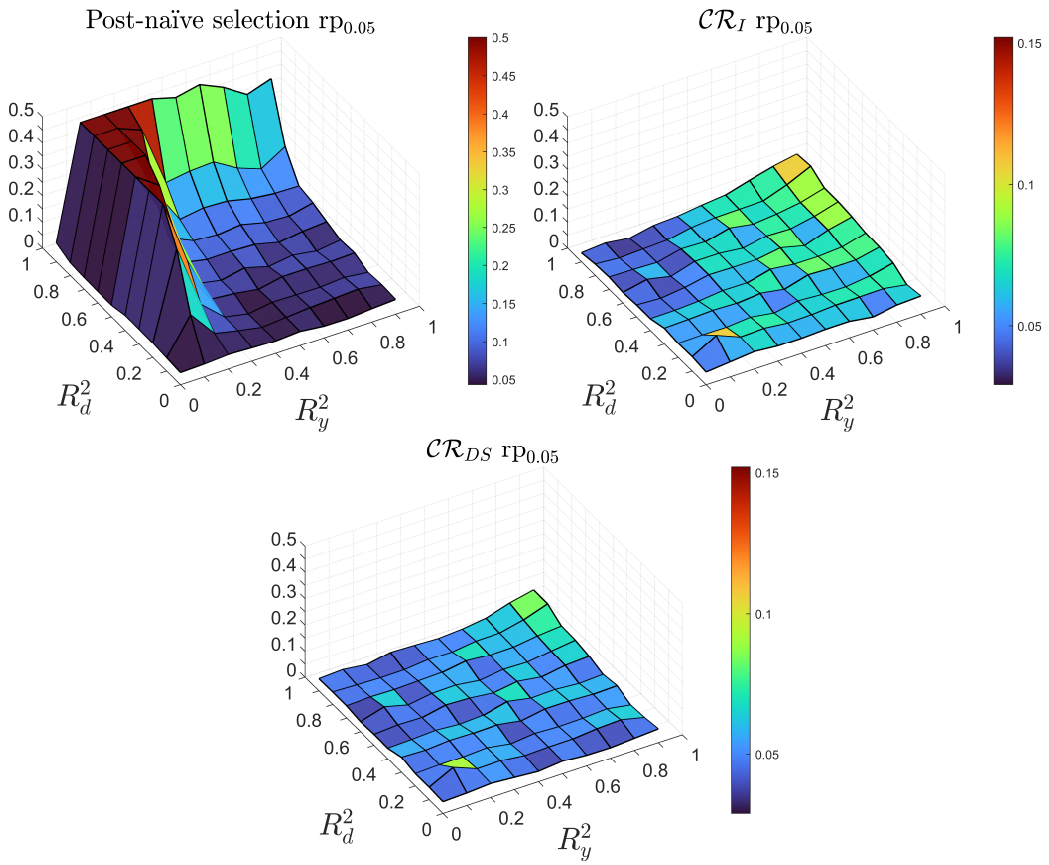


Figure 1.8: Surface plots of rejection frequencies of $\text{H}_0 : \alpha = \alpha_0$ under approximately sparse, 50% censored DGP

Note: These surface plots display the rejections frequencies at 0.05 level ($\text{rp}_{0.05}$) of the confidence regions based on the post-naïve selection (top-left panel), the instrumental Tobit in Algorithm 1.2.1 $\mathcal{CR}_I$ (top-right panel), and the post-double selection $\mathcal{CR}_{DS}$ (bottom panel). The ideal plot would show a flat horizontal plane at height 0.05. The grid comprises 100 different DGPs with $\alpha_0 = -0.25$ and censoring rate $c = 0.5$ in an approximately sparse setting. The rejection frequencies are based on $m = 1000$ replications for each design.

Overall, the post-double selection procedure emerges as the clear winner from our simulations, demonstrating superior finite sample performance compared to its instrumental peer. We attribute its high robustness to the refined estimation of the nuisance function $x\check{\beta}$ and the inverse scale parameter $\check{\gamma}$ in Step 3 of Algorithm 1.2.2. In sparse settings, we therefore advocate using the post-double selection Tobit estimator. On the other hand, the instrumental Tobit permits the estimation of instruments z_0 and q_0 by alternative methods, such as elastic net or ridge regressions, which is more appropriate in "dense" settings.

1.5.2. Testing the Effect of Gene Mutation ``M184V'' on HIV Load

HIV infections in humans with drug-susceptible strains have a good chance of being contained. A patient's response to antiretroviral drug regimens, however, depends on the presence or development of resistant gene mutations. Therefore, testing for drug-resistant mutations is crucial for attuning medication since inappropriate treatment may result in incomplete viral suppression, which in turn facilitates genetic variation and the selection of resistant variants during therapy. To measure the impact of specific mutations, researchers are interested in mapping observed genetic mutations on HIV concentrations in patients' blood plasma; see, e.g., Swenson et al. (2014) and Soret et al. (2018).

We downloaded data from AIDS Clinical Trials Group 5241 from the Stanford HIV Drug Resistance Database.[9] These files record blood plasma HIV viral load in decadic logarithm of copies/mL, HIV subtype, antiretroviral treatment history, and reverse transcriptase (RT) and protease (PT) mutations of 412 patients at multiple time points. The dataset is inherently high-dimensional because one-hot encoding observed RT and PT mutations generates more than three times as many potential control variables as available observations. We choose measurements of HIV viral load at week 12 of the trial as our outcome variable, because Gandhi et al. (2020) define virological failure in terms of the decline between baseline and week 12 viral load. We intend to test hypotheses about the effect of a particular gene mutation on viral load at week 12, after using ℓ_1-regularization to find a sparse set of necessary control variables. In this context, the selection of appropriate control variables is crucial, since some (primary) mutations induce resistance by themselves, while other (secondary) mutations are known to enhance the effect of primary mutations; see, e.g. Shafer (2002). The biological assay used to measure viral load in this clinical trial could not detect concentrations below a threshold of 50 copies/mL ($\log_{10}(50) \approx 1.7$), leading to a selective sampling process with censored outcomes. Thus, this dataset perfectly fits our model framework with a high-dimensional feature set and left-censored outcomes.

We drop participants who did not return for their week 12 examination and also omit binary features that identify less than 1% of the data (rare gene mutations). This leaves us with $n = 407$ observations, of which approximately 37% are censored below the detection limit of 1.7 $\log_{10}$ copies/mL. To account for individual differences in patients' infection histories, we also include baseline, i.e., week 0 viral load measurements as a potential feature.[10] Since baseline viral load is the sole non-binary variable in our feature set, we generate a sixth-order polynomial to allow for a potentially non-linear impact. In addition, we inter-

[9] Follow https://hivdb.stanford.edu/ACTG5241.html to access the data. We last checked the validity of this link in July 2023. For more information on this dataset the reader is referred to Gandhi et al. (2020).

[10] Note that all patients' baseline HIV concentrations exceed the limit of detection.

act the baseline viral load measurement with all antiretroviral drugs, RT, and PT mutations. In total, we end up with $p = 1296$ potential features. As our target regressor of interest, we choose RT mutation "M184V" because previous studies have documented an ongoing immunological treatment benefit despite the rapid development of drug resistance conferred by "M184V". From the perspective of fine-tuning medication, this makes "M184V" an interesting target to study as we would expect to find a significant negative effect on viral load independent of drug resistance; see, e.g. Miller et al. (2002) and Gallant (2006).

Table 1.2: Impact of gene mutation "M184V" on HIV viral after 12 weeks

	p	$\check{\alpha}$	$\widehat{\Sigma}_{1n}$	$\widehat{\Sigma}_{2n}$	t-statistic	p-value	card($\check{\mathcal{J}}$)
ℓ_1-based instr. Tobit	1296	-0.6403	0.2153	0.202	-2.9735	0.0029	21
ℓ_2-based instr. Tobit	1296	-0.6594	0.2518	0.2159	-2.6188	0.0088	1296
double selection Tobit	1296	-0.6869	0.2146	0.2124	-3.2014	0.0014	9
double selection Logit	1228	-0.8817	0.27	0.2719	-3.2422	0.0012	5

Note: We report point estimates $\check{\alpha}$ for the target parameter in front of RT mutation "M184V", its respective standard errors $\widehat{\Sigma}_{1n}$ based on the inverse Hessian and $\widehat{\Sigma}_{2n}$ based on the Neyman orthogonal score, t-statistics, p-values, and the cardinality of the set of selected controls card($\check{\mathcal{J}}$). The nuisance term $x_i\beta$ and the instrument z_i used as plug-in estimates in the instrumental Tobits were estimated by ℓ_1- and ℓ_2-regularized methods (Tobit and weighted least squares), where the penalty term is chosen by stratified 10-fold cross-validation; see Figure 1.9 below. To run the post-double selection Logit of Belloni et al. (2016a), we had to exclude 68 rare gene mutations from our set of potential control variables due to perfect prediction/classification of the binary outcome variable.

In Table 1.2, we report point estimates $\check{\alpha}$ for the coefficient in front of our target regressor of interest "M184V" based on two separate implementations of our instrumental Tobit in Algorithm 1.2.1, the post-double selection Tobit in Algorithm 1.2.2, and as a comparison, the binary choice Logit of Belloni et al. (2016a). For the latter, we generate the binary outcome $\mathbf{1}\{\log_{10}(\text{viral load}) > 1.7\}$, which classifies week 12 viral loads exceeding the censoring threshold as "successes". Due to this artificial reduction in the variation of outcomes we have to exclude 68 features from the set of potential controls for the Logit, because these gene mutations perfectly predict/classify the binary outcome; see column p in Table 1.2. For the instrumental Tobits we employ both ℓ_1- and ℓ_2-regularization to obtain plug-in estimates for the nuisance term $x\beta_0$, the inverse scale parameter γ_0 and instrument z_0. The penalty levels for Tobit Lasso, weighted linear Lasso, ridge Tobit, and weighted ridge regression are determined by 10-fold cross-validation. To preserve the censoring rate in all folds, we perform a stratified sample split. As measures of the generalization errors, we take the mean squared prediction error (MSE) averaged over all left-out folds for the weighted least squares

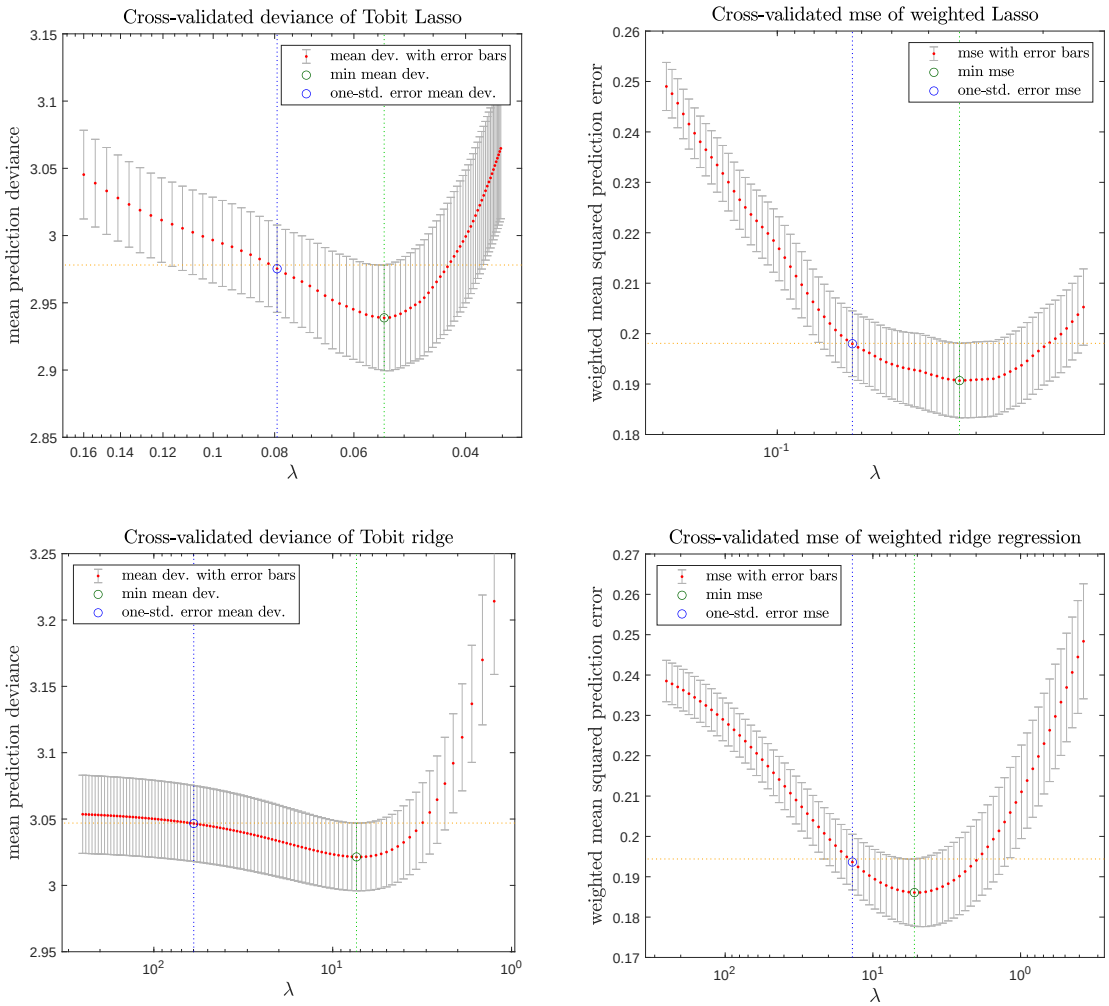


Figure 1.9: Generalization error of Lasso and ridge estimators for HIV data example

Note: These graphs display the mean prediction deviance of ℓ_1-regularized Tobit (top-left panel) and ℓ_2-regularized Tobit (bottom-left panel), as well as the mean squared prediction error (MSE) of the weighted linear Lasso (top-right panel) and the weighted ridge regression (bottom-right panel) computed for a sequence of on a logarithmic scale equidistant penalty terms λ by stratified 10-fold cross-validation. Both the nuisance term $x_i\beta$ and the instrument z_i are re-estimated on the full sample using the one-standard error rule penalty (dotted blue line). Resulting point estimates of the instrumental Tobits are given in Table 1.2.

estimator, and the mean prediction deviance, defined as minus two times the log-likelihood, averaged over all left-out folds for the Tobit estimators. The models are selected according to the "one-standard error rule", where we choose the largest penalty level within the error bar of the minimum of the respective accuracy measure. Figure 1.9 illustrates the results of the model selection process by cross-validation. We also experimented with elastic net penalties and ran a grid search over different combinations of penalty level and the weight assigned to the Lasso part of the penalty. However, for this dataset the minimum mean prediction deviance is attained with a Tobit with pure Lasso penalty. This suggests that a sparse representation is better suited to approximate the DGP than a dense one. We nevertheless

report the results obtained with ℓ_2-penalized estimators for the sake of illustration.

The estimated coefficients associated with the three Tobit specifications are of a similar magnitude. Compared to the post-double selection algorithm, the instrumental Tobit based on the Lasso penalty selected by cross-validation uses information of roughly twice the number of control variables. This reflects the fact that a cross-validated penalty level is typically smaller than one based on the asymptotic plug-in rule, and that the cross-validated nuisance term is predicted with penalized coefficients; see Remark 1.2.3 and Chetverikov et al. (2021). Indeed, the majority of active controls are selected in the second step where the instrument for the target regressor is built.

Remark 1.5.5 *(Interpretation of α_0 in terms of partial effects).*

The general interpretability and meaning of the magnitude of parameters estimated by a Tobit maximum likelihood approach depend on the specific research question investigated by the practitioner. By construction of model 1.2.1, we have $\mathrm{E}[\gamma_0 y_i^* \mid d_i,\, x_i] = \alpha_0 d_i + x_i\beta_0$, such that the parameter α_0 represents the partial effect of the target regressor d_i on the conditional mean of the re-scaled, uncensored outcome $\gamma_0 y_i^*$ given d_i and x_i. To obtain the effect on y_i^*, one needs to reverse Olsen's substitution: $\alpha_0^* = \alpha_0\sigma_0$. It should be noted that this interpretation is only valid if y_i^* is not merely a hypothetical construct. In the HIV data example, y_i^* is defined as a measurement of HIV concentrations in a patient's blood on a logarithmic scale to the base 10. Thus, values of y_i^* below $\mathcal{y}_i = 1.7$ exist and occur, but are undetectable. In other applications of Tobit models, such as to charitable contributions or household expenditures mentioned in Section 1.1, values of y_i^* below the threshold $\mathcal{y}_i$ are purely hypothetical. For example, a non-smoking household's expenditures on cigarettes cannot be negative, and yet two households with zero expenditures may differ in their propensities to smoke. In such cases, $\mathrm{E}[\gamma_0 y_i^* \mid d_i,\, x_i]$ lacks an immediate interpretation. However, the effect of d_i on $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0]$ or $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i]$ might be of interest; see McDonald and Moffitt (1980). As shown in Appendix A.4, for the logistic Tobit we have $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] = \big(1 + \exp\{-\alpha_0 d_i - x_i\beta_0\}\big)\log\big(1 + \exp\{\alpha_0 d_i x_i\beta_0\}\big)$ and, in particular, $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] = \log\big(1 + \exp\{\alpha_0 d_i x_i\beta_0\}\big)$. Both relations can be used to compute average partial effects.

We compute standard errors using the Jacobian of the estimating function for the instrumental Tobits and the Hessian of the post-double selection estimators $\widehat{\Sigma}_{1n}$. Additionally, we provide a second point estimate $\widehat{\Sigma}_{2n}$ based on the Neyman orthogonal score for all four estimators. There is no meaningful difference between both alternatives, which suggests that the information matrix equality holds. The fact that the estimated standard errors of the post-double Logit are larger than those of the post-double Tobit is in line with the argu-

ment put forth in Remark 1.2.2, namely, that the Tobit uses the information in the data, i.e. the non-censored outcomes, more efficiently. Although the coefficient estimated with the Logit is slightly larger by absolute value than that of the post-double Tobit, we reach the same conclusion: the presence of RT mutation "M184V" in a patient's blood sample significantly increases the chance of suppressing the HIV viral load to below the limit of detection after 12 weeks. This finding is consistent with previous research on "M184V". Explanations for this mechanism include hyper-susceptibility to some drugs and the delayed immunologic progression due to reduced HIV replication capacity; see, e.g. Miller et al. (2002), Gallant (2006) and references therein.

1.6. CONCLUDING REMARKS

Notwithstanding the enduring popularity of Tobin's original maximum likelihood estimator, post-regularization methods for Tobit models have received scant coverage and are, thus, under-represented in the rapidly growing field of research on high-dimensional methods. This study addressed the dearth of theoretical results pertaining to post-selection inference in Tobit models in several ways. Firstly, based on the general framework of Neyman orthogonal likelihood scores in Chernozhukov et al. (2015) and existing adaptations to GLMs in Belloni et al. (2016a) and quantile regression in Belloni et al. (2019), we proposed an adjusted Tobit score function which allows us to consistently estimate a scalar target parameter of interest at $\sqrt{n}$ rate despite the presence of high-dimensional nuisance terms. The estimating function for the target parameter is based on a class of distributions in location-scale parametrization, where the well-known Olsen (1978) substitution $\gamma_0 = 1/\sigma_0$ was applied. The proof of asymptotic normality rests on the same Taylor series arguments as for GLMs. However, the necessary restrictions imposed on the CDF of the disturbance distribution and its derivatives are considerably more stringent.

Secondly, we showed that the logistic distribution satisfies these assumptions, in that logistic CDF and PDF obey first-order differential inequalities. In this context, it was found that the logistic Tobit likelihood loss can be decomposed into parts that resemble those of the logistic binary choice likelihood loss. As a result, asymptotic normality for the logistic Tobit estimator was established under primitive assumptions that are almost identical to those required for the logistic GLM. The additional scale parameter was considered a separate nuisance, but did not pose a significant challenge. Our likelihood approach allows for

the joint estimation of all nuisance parameters at $n^{1/4}$ rate and the scale parameter was exempted from regularization at the negligible cost of losing one degree of freedom without affecting the convergence rate. To create an orthogonal relation in the scale parameter, a second instrument was explicitly or implicitly constructed as a weighted decomposition of the censored outcome observations.

Additionally, we provided a detailed discussion on different implementations as sparse instrument and post-double selection Tobit estimators. To illustrate the practical applicability of these algorithms in an important area of research, we estimated the effect of mutation "M184V" on blood plasma HIV concentrations using data from AIDS Clinical Trials Group 5241 from the Stanford HIV Drug Resistance Database. Our analysis revealed a significant negative effect that was consistent across different estimators and regularization methods. Both the hypothesis test result and the sign of the effect align with findings from previous studies on mutation "M184V".

Chapter 2

USING POST-REGULARIZATION DISTRIBUTION REGRESSION TO MEASURE THE EFFECTS OF A MINIMUM WAGE ON HOURLY WAGES, HOURS WORKED AND MONTHLY EARNINGS†

2.1. INTRODUCTION

The introduction of Germany's statutory minimum wage on January 1, 2015 was a significant policy experiment. While industry-specific minimum wages had existed before 2015, Germany was among the few countries worldwide without a universal minimum wage. The imposition of a nationwide minimum wage of 8.50 euros/hour in 2015 represented a major intervention in the German labour market, affecting around 4 million workers (more than 11% of the workforce) who earned less than 8.50 euros/hour before its introduction; see Mindestlohnkommission (2020).[1]

Based on different data sets, previous contributions have examined various aspects of the German minimum wage introduction. As to potential employment effects, the literature has

†This chapter is a modified version of an article that will be published in "The Econometrics Journal".

[1]See Caliendo et al. (2019) for a more detailed overview of the institutional details of the minimum wage introduction.

reached the consensus that these were non-existent or very small; see, e.g. Caliendo et al. (2019), Dustmann et al. (2022a), Bossler and Schank (2023), and Link (2024). By contrast, the literature appears to have reached conflicting results about the distributional effects of the minimum wage, i.e., its effects on the distributions of hourly wages, monthly earnings and working hours (the latter including potential shifts between full-time, part-time and marginal part-time work). Using register data, Bossler and Schank (2023) find that the minimum wage significantly reduced inequality in monthly wages. On the basis of survey data, however, Burauel et al. (2019a,b) and Caliendo et al. (2022) conclude that the minimum wage introduction also reduced working hours, neutralizing its effect on monthly wages. Given that German register data do not include information on working hours, Biewen et al. (2022) analyse large-scale data from the statistical offices to conclude that working hours were not causally affected by the minimum wage so that increased hourly wages should fully translate into changes in monthly earnings.

Given the inconclusive evidence, this paper reexamines the effects of the minimum wage on the distributions of hourly wages, monthly earnings, and working hours. We use the same survey data analysed in Burauel et al. (2019a,b) and Caliendo et al. (2022). Based on modern machine learning methods that allow us to examine the effects of the minimum wage across all points of the distribution, we reach the conclusion that the minimum wage replaced hourly wages below the minimum threshold, increased monthly earnings in the lower-middle segment but not at the very bottom of the distribution (consistent with Bossler and Schank, 2023), and did not significantly affect the distribution of working hours. Our results help reconcile the conflicting results in the literature using different data sources mentioned above.

Our econometric analysis is based on the distribution regression approach introduced by Foresi and Peracchi (1995) and developed by Chernozhukov et al. (2013). Distribution regression involves performing numerous binary regressions, each modelling the probability that the outcome falls below a specific threshold, across a finely spaced grid covering the entire distribution. Compared to alternative methods such as conditional or unconditional quantile regression, distribution regression directly targets nominal points in the outcome distribution. It is therefore ideally suited to study changes in distributions, such as hourly wages, working hours or monthly wages, whose quantiles typically change over time, complicating the interpretation of quantile regression results if more than one time period is involved. Moreover, distribution regression easily deals with discrete mass points, see Chernozhukov et al. (2013), which is particularly relevant when dealing with discrete distributions (working hours) or distributions with severe heaping (hourly wages, especially after the introduction of a minimum wage). This is in contrast to conditional or unconditional quantile regression that are based on the assumption of continuous distributions.

A key challenge in distribution regressions is the need to specify and estimate multiple binary models. The different binary regression models should take into account potentially varying sets of covariates as different covariates may matter at different points of the distribution. This challenge is particularly pronounced when using a sample with a moderate sample size but a large number of potential covariates as it will become inevitable to select relevant covariates at each distributional threshold to save degrees of freedom. In light of the fact that this is generally required to be carried out across a substantial number of thresholds, a hand-picked approach might result in a considerable amount of arbitrary specification search with unknown consequences for the potential bias of estimated coefficients and their estimated standard errors. Another challenge is the likely high correlation between regression results at different thresholds, which makes inference across multiple points in the distribution more complex. Apart from inferential aspects, the sheer practical task of specification searches for a large number of parallel regressions suggests the use of machine learning techniques such as Lasso to separately predict nuisance terms at the large number of distributional thresholds.

Both the practical and the inferential aspect have been addressed by recent advances in econometric machine learning. In a recent contribution, Belloni et al. (2018b) showed how to employ a sequence of ℓ_1-regularized logistic regressions such that inferences about the coefficients of target regressors are valid both point-wise, i.e., individually in each regression model, as well as uniformly across a large number of models. Their proposed algorithm is closely related to the concept of "partialling-out" in the econometrics literature, where one removes nuisance terms that are either related to the outcome or the treatment variables. Against this background, the estimators used in the present study belong to the class of post-regularization methods commonly referred to as post-double selection algorithms; see Belloni et al. (2014).

Picking covariates with the Lasso also entails functional form specification as the features offered to the Lasso may include arbitrary transformations of variables (logs, polynomials, indicators for particular values, interaction terms). The possibility to obtain valid inference after large-scale automatic specification search is a remarkable achievement of the recent econometric machine learning literature. It represents a major improvement over the often arbitrary and undocumented specification searches carried out by individual researchers, which are typically influenced in unknown ways by the propagation of pre-tested control variables used in previous research.

A key assumption of the ℓ_1-regularized methods used by us is approximate sparsity, i.e., the sequence of coefficients of potential confounder terms sorted by absolute value decays

quickly enough, but does not have to be exactly equal to zero. This is a natural assumption in substantive applications where one would like to control for relevant confounding information, but does not rule out the existence of factors whose influence may be negligible for the inferential purpose at hand. In the following, we employ the method described in Belloni et al. (2018b), albeit with certain modifications to address the fact that our data and research design feature observation clusters and sampling weights.

2.2. ECONOMETRIC METHODS

We aim to measure the effects of the minimum wage introduction across the distribution of an outcome variable $Y \in \{$"hourly wage", "hours worked", "monthly earnings"$\}$. To this end, let $F(t) = \exp\{t\}/(1+\exp\{t\})$ for all $t \in \mathbb{R}$ denote the logistic (inverse) link function. We use the logistic distribution regression model:

$$\mathrm{P}(Y < u \mid D, X) = \mathrm{E}[\mathbf{1}\{Y < u\} \mid D, X] = F(D\theta_u + X\beta_u) \quad \forall\, u \in \mathcal{U}, \tag{2.2.1}$$

which measures the effects of target variables $D = (D_1, \ldots, D_{\tilde{p}})$, representing the difference--in-differences specification described below, across a set of grid points $u \in \mathcal{U}$ in the support of the outcome distribution. In order to isolate the effect of the target variables on the likelihood of falling below a particular threshold u, it is necessary to control for confounders X, which may vary across different points of the outcome distribution. This motivates the use of a separate Lasso procedure to select a set of relevant control variables at each point $u \in \mathcal{U}$.

For the rest of this article, let the vector of potential control variables be given by $X = (X_1, \ldots, X_p)$, the vector of target coefficients of interest by $\theta_u = (\theta_{u1}, \ldots, \theta_{u\tilde{p}})^T$, the vector of nuisance parameters by $\beta_u = (\beta_{u1}, \ldots, \beta_{up})^T$, and the sequence of threshold indicators by $Y^u = \mathbf{1}\{Y < u\}$. Our application comprises clusters $g \in \{1, \ldots, G\}$ of observations $W_{ig} = (Y_{ig}, D_{ig}, X_{ig}) = (Y_{ig}, D_{1ig}, \ldots, D_{\tilde{p}ig}, X_{1ig}, \ldots, X_{pig})$, assuming that observations are independent across clusters but are potentially correlated within clusters. The number of observations within cluster g is denoted by n_g such that $\sum_{g=1}^{G} n_g = n$. Furthermore, we make use of deterministic sampling weights v_{ig}, which are normalized to sum up to the total number of observations $\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig} = n$.

2.2.1. Parameter Estimation

The post-double Lasso method for the logistic regression model and other generalized linear models was developed by Belloni et al. (2016a). Belloni et al. (2018b) extended the method to cover uniform inference for functional parameters, e.g., for coefficients of many parallel Logit models as needed in our application. Belloni et al. (2016), Chiang (2020), and Ahrens et al. (2020) considered extensions of the Lasso method to clustered data.

Algorithm 2.2.3 *(Post-double Selection Logistic Distribution Regression – essentially Algorithm 2 in Belloni et al., 2018b).*

Step 1. Run the post-ℓ_1 penalized logistic regression of Y_{ig}^u on $(D_{ig},\ X_{ig})$:

$$\begin{aligned}
(\hat{\theta}_u, \hat{\beta}_u) &\in \underset{\theta,\beta}{\arg\min}\ \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\Lambda_u(W_{ig},\theta,\beta) + \frac{\lambda_1}{G}\big\|\widehat{\Psi}_u(\theta^T,\ \beta^T)^T\big\|_1,\\
(\tilde{\theta}_u, \tilde{\beta}_u) &\in \underset{\theta,\beta}{\arg\min}\ \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\Lambda_u(W_{ig},\theta,\beta)\ :\ \text{support}(\theta,\beta)\subseteq\widehat{\mathcal{S}}_u,\\
\Lambda_u(W_{ig},\theta,\beta) &= \log(1+\exp\{D_{ig}\theta + X_{ig}\beta\}) - Y_{ig}^u\cdot(D_{ig}\theta + X_{ig}\beta),
\end{aligned}$$

where $\widehat{\mathcal{S}}_u := \text{support}(\hat{\theta}_u,\hat{\beta}_u) = \{l\in\{1,\dots,\tilde{p},\dots,\tilde{p}+p\}\mid(\hat{\theta}_u^T,\hat{\beta}_u^T)_l^T\neq 0\}$. Penalty parameter λ_1 and the entries of diagonal penalty loadings matrix $\widehat{\Psi}_u$ are chosen according to Algorithm B.1.6 explained in Appendix B.1.
For $i=1,\dots,n$ compute $\widehat{f}_{u,ig}^2 = F'(D_{ig}\tilde{\theta}_u + X_{ig}\tilde{\beta}_u)$.

Step 2. For all $j\in\mathcal{J} := \{1,\dots,\tilde{p}\}$, define $\tilde{X}_{ig}^j = (D_{ig\mathcal{J}\setminus j}, X_{ig})$ and run the weighted post-Lasso estimator of target $\widehat{f}_{u,ig}D_{igj}$ on $\widehat{f}_{u,ig}D_{ig\mathcal{J}\setminus j}$ and $\widehat{f}_{u,ig}X_{ig}$:

$$\begin{aligned}
\hat{\gamma}_u^j &\in \underset{\gamma}{\arg\min}\ \frac{1}{2G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\widehat{f}_{u,ig}^2\big(D_{igj}-\tilde{X}_{ig}^j\gamma\big)^2 + \frac{\lambda_2}{G}\big\|\widehat{\Psi}_u^j\gamma\big\|_1,\\
\tilde{\gamma}_u^j &\in \underset{\gamma}{\arg\min}\ \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\widehat{f}_{u,ig}^2\big(D_{igj}-\tilde{X}_{ig}^j\gamma\big)^2\ :\ \text{support}(\gamma)\subseteq\text{support}(\hat{\gamma}_u^j),
\end{aligned}$$

where $\text{support}(\hat{\gamma}_u^j) = \{l\in\{1,\dots,\tilde{p}+p-1\}\mid(\hat{\gamma}_u^j)_l\neq 0\}$. Additionally, define $\hat{\mathcal{S}}_u^j := \{l\in\{\tilde{p}+1,\dots,\tilde{p}+p\}\mid(\hat{\gamma}_u^j)_{l-1}\neq 0\}$. Penalty parameter λ_2 and the entries of diagonal penalty loadings matrix $\widehat{\Psi}_u^j$ are chosen according to Algorithm B.1.7 explained in Appendix B.1.

Step 3. Run the logistic regression of Y_{ig}^u on the union of target variables D_{ig} and the set of all control variables selected in either Step 1 or 2:

$$(\check{\theta}_u,\check{\beta}_u)\ \in\ \underset{\theta,\beta}{\arg\min}\ \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\Lambda_u(W_{ig},\theta,\beta)\ :\ \text{support}(\theta,\beta)\subseteq\Big(\mathcal{J}\cup\hat{\mathcal{S}}_u\bigcup_{j=1}^{\tilde{p}}\hat{\mathcal{S}}_u^j\Big).$$

Essentially, applying Lasso to clustered data involves the treatment of blocks of observations belonging to the same cluster as "super-observations", thereby "collapsing" these blocks and computing penalty levels based on partial sums of observations. Following Belloni et al. (2018b) and applying modifications for clustering similar to Chiang (2020), as well as for sampling weights, we use the post-double selection procedure described in Algorithm 2.2.3 to estimate the process of target coefficients $(\theta_{uj})_{u\in\mathcal{U},j\in\mathcal{J}}$ indexed by the distributional thresholds of outcome variable Y and the set of target variables $\mathcal{J} = \{1,\dots,\tilde{p}\}$. The resulting post-double selection point estimate $\check{\theta}_u = (\check{\theta}_{u1},\ \dots,\ \check{\theta}_{u\tilde{p}})^T$ measures the impact of target regressors $D_1,\dots,D_{\tilde{p}}$ on Y^u at threshold u.

It should be noted that our approach differs slightly from that of Belloni et al. (2018b) in that we jointly estimate all target parameters in Step 3. As pointed out in, e.g. Appendix B of Belloni et al. (2016a), we can add additional variables to support(θ,β) by including covariates selected in a fixed number of individual orthogonalization steps. In the present application, this is a computationally attractive modification, as we avoid running separate estimations for all our target variables over all thresholds. Moreover, this greatly facilitates the computation of the influence functions in (2.2.11), since we can pre-compute certain quantities based on the joint estimation in Step 3 and, thus, process the Jacobian and the outer product of the clustered orthogonal score in a vectorized manner. Generally, the post-double selection method simultaneously establishes in-sample orthogonality with respect to the instruments for all target variables by controlling for the effect of all selected confounders. Given that the cardinality of the union of all selected variables is still far smaller than the sample size, we expect the inferential results obtained by a joint estimation to be more robust when compared to target-by-target computations; see, e.g. Appendix L in Belloni et al. (2018a).

2.2.2. Uniform Inference

In order to compute pointwise and simultaneous confidence intervals for subsets of the coefficient processes $(\theta_{\mathcal{U}'\mathcal{J}'})$ indexed by $\mathcal{U}' \subseteq \mathcal{U}$ and $\mathcal{J}' \subseteq \mathcal{J}$, we estimate the following quantities. The Neyman-orthogonal moment function for target parameter θ_{uj} is given by

$$\psi_j(W,\theta_u,\eta_u) = \big\{Y^u - F(D\theta_u + X\beta_u)\big\}\cdot\big(D_j - \tilde{X}^j\gamma_u^j\big), \tag{2.2.2}$$

where the nuisance parameters are collectively defined as $\eta_u = (\beta_u^T,\ \gamma_u^{1T},\ \dots,\ \gamma_u^{\tilde{p}T})^T$. Let

$$\psi(W,\theta_u,\eta_u) = \big(\psi_1(W,\theta_u,\eta_u),\ \dots,\ \psi_{\tilde{p}}(W,\theta_u,\eta_u)\big)^T \tag{2.2.3}$$

be the vector of moment functions of all target parameters and its Jacobian matrix

$$J(W,\theta_u,\eta_u) = \frac{\partial\,\psi(W,\theta_u,\eta_u)}{\partial\,\theta_u^T}. \tag{2.2.4}$$

As shown in Belloni et al. (2018a), the first-order conditions of the post-double selection estimator in Step 3 of Algorithm 2.2.3 implicitly create an orthogonal relation such that

$$\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\psi(W_{ig},\check{\theta}_u,\check{\eta}_u) = 0, \tag{2.2.5}$$

where the moment functions are evaluated at $\check{\eta}_u = (\check{\beta}_u^T,\, \tilde{\gamma}_u^{1T},\, \ldots,\, \tilde{\gamma}_u^{\tilde{p}T})^T$. A first-order expansion in the target parameters $\check{\theta}_{u1},\ldots,\check{\theta}_{u\tilde{p}}$ yields a consistent estimate of their asymptotic covariance matrix

$$\big(\widehat{\Sigma}^u\big)^2 = \big(\widehat{J}_u^{-1}\big)\,\widehat{B}_u\big(\widehat{J}_u^{-1}\big)^T, \tag{2.2.6}$$

where

$$\widehat{J}_u^{-1} \;=\; \Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,J(W_{ig},\check{\theta}_u,\check{\eta}_u)\Big]^{-1} \tag{2.2.7}$$

$$\widehat{B}_u \;=\; \Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g} v_{ig}\,\psi(W_{ig},\check{\theta}_u,\check{\eta}_u)\Big)\Big(\sum_{i=1}^{n_g} v_{ig}\,\psi(W_{ig},\check{\theta}_u,\check{\eta}_u)\Big)^T\Big]. \tag{2.2.8}$$

The inner part of matrix $\widehat{B}_u$ accounts for the clustering by "collapsing" the blocks of observations within a cluster g. The estimated asymptotic variance associated with target parameter θ_{uj} is given by the j-th diagonal element $\hat{\sigma}_{uj}^2 = \big(\widehat{\Sigma}_{j,j}^u\big)^2$.

Note that the Neyman-orthogonal moment function for the target parameters $\check{\theta}_u = (\check{\theta}_{u1},\, \ldots,\, \check{\theta}_{u\tilde{p}})^T$ is constructed such that

$$\frac{\partial}{\partial\,\eta^T}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\psi(W_{ig},\check{\theta}_u,\eta)\Big]\Big|_{\eta=\check{\eta}_u} = 0. \tag{2.2.9}$$

This implies that the estimating equations are first-order immune with respect to the nuisance terms. In other words, by constructing an instrument for D_j in (2.2.2), one has "partialled-out" the effect of covariates $\tilde{X}^j$. In particular, this relation also applies to all elements $\mathcal{J}\setminus j$ of the target vector that are associated with other target variables

$$\frac{\partial}{\partial\,\theta_{\mathcal{J}\setminus j}^T}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} v_{ig}\,\psi_j(W_{ig},\theta,\check{\eta}_u)\Big]\Big|_{\theta=\check{\theta}_u} = 0. \tag{2.2.10}$$

This stresses the validity of our modification to Step 3 of Algorithm 2.2.3.

For the multiplier bootstrap procedure, define the $\tilde{p}$-dimensional estimated influence function as

$$IF(W_{ig},\check{\theta}_u,\check{\eta}_u) = -\widehat{J}_u^{-1}\psi(W_{ig},\check{\theta}_u,\check{\eta}_u). \tag{2.2.11}$$

The multiplier bootstrap critical value c_α is computed as the $(1-\alpha)$-quantile of the distribution of the supremum statistic

$$\mathcal{S} := \sup_{u \in \mathcal{U}', j \in \mathcal{J}'} \frac{1}{\sqrt{G}\hat{\sigma}_{uj}} \sum_{g=1}^{G} \xi_g \cdot \Big[\sum_{i=1}^{n_g} v_{ig} \cdot IF_j(W_{ig}, \check{\theta}_u, \check{\eta}_u) \Big] \tag{2.2.12}$$

where $IF_j(W_{ig}, \check{\theta}_u, \check{\eta}_u)$ is the j-th element of $IF(W_{ig}, \check{\theta}_u, \check{\eta}_u)$. The distribution of $\mathcal{S}$ can be simulated by repeatedly drawing $i.i.d.$ weights $\xi_g \sim \mathcal{N}(0,1)$ from the standard normal distribution. The bootstrap critical value c_α is then used as a scaling factor for the point-wise confidence regions which results in a simultaneous confidence band covering multiple target parameters $(\theta_{uj}), j \in \mathcal{J}' \subseteq \mathcal{J}$ at multiple distribution thresholds $u \in \mathcal{U}' \subseteq \mathcal{U}$. We use these confidence intervals primarily to test hypotheses. To exemplify, we seek to test hypotheses of the form that multiple target regressors $j \in \mathcal{J}'$ have no effect $\mathrm{H}_{0,\mathcal{U}'\mathcal{J}'} : \theta^0_{uj} = 0$ for all thresholds $u \in \mathcal{U}'$, based on the asymptotic relation:

$$\mathrm{P}\Big(\check{\theta}_{uj} - c_\alpha \frac{\hat{\sigma}_{uj}}{\sqrt{G}} \;\leq\; \theta^0_{uj} \;\leq\; \check{\theta}_{uj} + c_\alpha \frac{\hat{\sigma}_{uj}}{\sqrt{G}} \; \forall\, u \in \mathcal{U}', j \in \mathcal{J}' \Big) \approx 1 - \alpha. \tag{2.2.13}$$

2.3. DATA AND IMPLEMENTATION

2.3.1. Data Sources and Specification

Our empirical analysis is based on the German Socio-Economic Panel Study (SOEP, v35) which is a long-running survey providing representative information about the German population; see, e.g. Schröder et al. (2020). We utilize information from 2011 to 2018, which encompasses the years leading up to and following the implementation of the German minimum wage on January 1st, 2015. The strength of a survey like the SOEP is the wealth of information that can be used as covariates. A weakness is the moderate sample size of around 11,000 wage earners per year, which motivates the use of specification selection methods. After applying selection criteria, our final sample includes approximately 90,000 observations. We rely only on cross-sectional information in the SOEP, as frequent refreshment samples and permanent dropout make the panel highly unbalanced. However, we fully account for longitudinal correlation in our data by adjusting our inference and penalty selection procedures as described in section 2.3.3. In addition, we use the sampling weights provided with the SOEP which ensure that cross-sectional information is representative for the German population in the given year. We exclude from our sample individuals

who are not subject to the minimum wage (the self-employed, students, apprentices, interns and similar groups).

Following the seminal work by Card (1992) on minimum wages, we assess the effects of the minimum wage introduction using a continuous treatment indicator, called the minimum wage bite MWB_{it}, which represents the share of workers in specific population subgroups who earned less than 8.50 euros/hour before the policy was implemented. The minimum wage is expected to have its most significant impact on subgroups with the highest pre-reform exposure, when other relevant factors are controlled for.[2] As the sample size of the SOEP would be too low to construct reliable bite measures for small population subgroups, we take our bite measure from a larger data set, the German Structure of Earnings Survey (GSES). The bite measure used here is defined at the 2-digit industry level differentiated by East/West Germany. Defining the bite measure at the industry level is well aligned with the structure of industrial relations in Germany, where a substantial part of wage bargaining takes place at the industry level. Our bite measure varies between .003 and .701, providing large variation to measure the effects induced by the exposure to the newly introduced minimum wage.[3] Note that MWB_{it} is indexed in both i and t because the minimum wage bite, measured at the industry level in 2014, is assigned annually based on each individual's industry affiliation. Consequently, our estimations also capture wage effects for individuals who switch industries.

We measure the effects of the minimum wage introduction on the distribution of our outcome variables $Y \in \{$"hourly wage", "hours worked", "monthly earnings"$\}$ by the following difference-in-differences specification:

$$\begin{aligned} \mathrm{P}(Y_{it} < u \mid D_{it}, X_{it}) \;=\; & F\big(\theta_{u,bite} \cdot MWB_{it} + \theta_{u,2011/12} \cdot \mathbf{1}\{t = 2011/12\} \\ & + \theta_{u,2015/16} \cdot \mathbf{1}\{t = 2015/16\} + \theta_{u,2017/18} \cdot \mathbf{1}\{t = 2017/18\} \\ & + \Theta_{u,2011/12} \cdot MWB_{it} \cdot \mathbf{1}\{t = 2011/12\} \\ & + \Theta_{u,2015/16} \cdot MWB_{it} \cdot \mathbf{1}\{t = 2015/16\} \\ & + \Theta_{u,2017/18} \cdot MWB_{it} \cdot \mathbf{1}\{t = 2017/18\} + X_{it}\beta_u\big). \end{aligned} \tag{2.3.1}$$

In order to keep the number of target coefficients low, we combine two adjacent years. The

[2] We thank a reviewer for suggesting a robustness check adjusting the minimum wage bite for counterfactual wage growth due to inflation or productivity gains. However, this adjustment is unlikely to affect our results, as inflation was near zero at the time of the minimum wage introduction and productivity growth at the lower end of the wage distribution is negligible.

[3] For more details on the bite measure used here, see Biewen et al. (2022). An alternative would be to define the bite at the level of labour market regions but this faces the difficulty that the coverage of labour market regions in a survey like the SOEP is patchy and that regional information in the SOEP can only be processed on-site with limited computational facilities.

relevant periods are: $t = 2013/14$ representing the reference period, i.e. the period immediately before the minimum wage introduction, $t = 2015/16$ representing short-term effects after the introduction, $t = 2017/18$ representing medium-term effects after the introduction, and $t = 2011/12$ representing the pre-test period. Upper-case coefficients $\Theta_{u,2015/16}$ and $\Theta_{u,2017/18}$ correspond to the short-term and medium-term treatment effects of the minimum wage introduction on the likelihood of falling below a particular threshold u in the outcome distribution. They measure to what extent, e.g. hourly wages below a particular level u became more or less frequent after the minimum wage introduction per unit of exposure to the newly introduced minimum wage MWB_{it}, controlling for time effects $\mathbf{1}\{t = 2011/12\}$, $\mathbf{1}\{t = 2015/16\}$, $\mathbf{1}\{t = 2017/18\}$, base effects MWB_{it}, and for all other characteristics X_{it} such as work experience, education and occupational characteristics, that are relevant for explaining that a particular wage observation falls below threshold u. Upper-case coefficient $\Theta_{u,2011/12}$ provides a pre-treatment test as it measures to what extent differences already emerged between high and low exposure groups in the pre-treatment period, i.e. between 2011/12 and the reference period 2013/14. The main (uninteracted) effects in the difference-in-differences specification control for general time differences $\theta_{u,2011/12}$, $\theta_{u,2015/16}$, $\theta_{u,2017/18}$, and for differences $\theta_{u,bite}$ between high and low bite groups of falling below threshold u that are time-invariant. All other terms X_{it} of the regression are selected by the Lasso procedure.

2.3.2. Variables and Feature Engineering

Our dependent variables are derived from the survey information on monthly earnings and actual hours worked per week, including overtime. Monthly earnings and actual hours worked per week are taken as they appear in the survey. Hourly wages are computed as monthly earnings divided by monthly hours worked (defined as weekly hours multiplied by the factor 4.345).

Table B.2.1 in Appendix B.2 describes the information that is used to construct the set of potential control variables from which the ℓ_1-methods can choose relevant elements for predicting the nuisance terms at each threshold. The total number of features constructed in this way is in the order of several thousands, as we not only include transformations of continuous variables (e.g. polynomial terms, square root, log) and indicators for potentially important individual values of continuous variables (e.g. an indicator for having an unemployment experience of zero years), but also interactions and full sets of indicators for all our categorial variables. To illustrate this, consider an educational classification with five categories. In this case, we include a full set of five indicators describing the membership in each category (no omitted category). The Lasso can then flexibly pick the indicators that help to remove the omitted variable bias for explaining the effect of the treatment variables

at a particular threshold.[4] It is important not to omit a reference category when constructing sets of such indicators as exactly the omitted category could be the one preferred by the Lasso. The information represented by the omitted category could be re-constructed as a linear combination of other categories, but this runs counter to the idea of finding a sparse approximation for the nuisance term. In a similar way, we offer to the Lasso nested or overlapping information from classifications of higher or lower aggregation levels from which it can choose the information that is most suitable to remove omitted variable bias. For example, we include occupation codes at different aggregation levels (1-digit, 2-digit etc.) and nested or partly overlapping education classifications that offer finer or more coarse information.

In order to arrive at the final set of potential covariates offered to the Lasso, we eliminate from the full set of features described in Table B.2.1 i) constant features, ii) duplicates/multiples of other features, iii) features that uniquely characterize less than 1 percent of our sample. We can relax restriction iii) to a certain extent without affecting the results. However, our experience suggests that doing so can result in an increased likelihood of perfect prediction problems and convergence issues in the Logit models. This is in contrast to the primary motivation of this paper, which is to identify a fully automatic method for selecting controls at the typically large number of thresholds without the need to manually fix problems or eliminate features at individual thresholds. Applying the above criteria, the final number of features included in our estimations was around $p = 2{,}500$. The exact number of features depends on the outcome variable, since features related to working hours cannot be included for monthly earnings and hours worked due to perfect prediction issues. This is clearly too large for individual specification searches at one given threshold, let alone at the typically around 40-50 thresholds used per dependent variable in our application.

2.3.3. Details on Lasso Implementation

As described in Section 2.2, our methods allow for clustering of observations in two ways: i) for statistical inference and ii) for the choice of Lasso penalties. Concerning the former, it is well-known that in difference-in-differences-like designs, it is necessary to cluster at the level of the treatment variable; see, e.g. Abadie et al. (2022). Our treatment variable MWB_{it} is based on the combination of 2-digit industries and East/West information. This provides 152 population subgroups at the level of which we cluster when computing the variance matrices and when drawing multiplier bootstraps. We initially also tried to cluster the Lasso penalties at this level but found that this led to quite erratic and volatile results

[4] For categorial variables, we also define a category "missing value" that may also be picked by the Lasso if it helps to predict the treatment or the outcome variable. This also helps to conserve the number of observations as observations with missing values in these variables do not have to be discarded.

across different thresholds and coefficients which did not seem plausible. This behaviour of the Lasso is not surprising given the relatively low number of clusters in our application and their sometimes chunky nature. For computing Lasso penalty loadings, we therefore clustered at the level of the panel units, which is standard for panel data; see, e.g. Belloni et al. (2016), Ahrens et al. (2020).

It is well known that the Lasso solution need not be unique if the feature set contains mostly discrete (binary) variables as in our case; see Tibshirani (2013). To evade numerical difficulties, we implemented our post-ℓ_1-estimators using the Moore-Penrose pseudo-inverse.

In our application, the cardinality of the active set of control variables was between 100 and 120 depending on the threshold. The double-selected features consistently included information on educational qualifications, work experience as well as additional controls that differed across thresholds in plausible ways, e.g. indicators for low occupational positions/job types at lower thresholds, information on firm characteristics or particular educational/occupational qualifications at medium or upper thresholds, interactions of such characteristics with gender or East/West Germany at particular thresholds.

2.4. EMPIRICAL RESULTS AND ECONOMETRIC ANALYSIS

We begin by examining how the minimum wage affected the likelihood that hourly wages fall below certain thresholds. Figure 2.1 displays the treatment effect coefficients in the pre-treatment period $\Theta_{u,2011/12}$ and in the two evaluation periods $\Theta_{u,2015/16}$ and $\Theta_{u,2017/18}$. The results in top-right and bottom panels of Figure 2.1 show that, as intended by policymakers, the likelihood of having hourly wages below 8.50 euros per hour declined in groups with high minimum wage exposure after the introduction. In the pre-test period, the top-left panel of Figure 2.1 reveals no significant differences between treated and untreated individuals. In 2017/18, there is a slight indication of spill-over effects above the minimum level of 8.50 euros/hour. However, these effects are not statistically significant.[5]

[5]Note that the minimum wage level was initially set to 8.50 euros/hour in 2015, but increased to 8.84 euros/hour in 2017.

Indeed, the pattern in the lower part of the distribution in Figure 2.1 might also reflect disemployment effects, since our analysis includes only employed workers. In separate analyses, however, Caliendo et al. (2019), Dustmann et al. (2022a), and Bossler and Schank (2023) have found little evidence for such disemployment effects. Link (2024) concludes that firms affected by the minimum wage mostly increased prices but did not cut employment. Below, we also show that there were no minimum wage effects on low levels of working hours, which is what would be expected if the minimum wage systematically displaced low-wage workers.

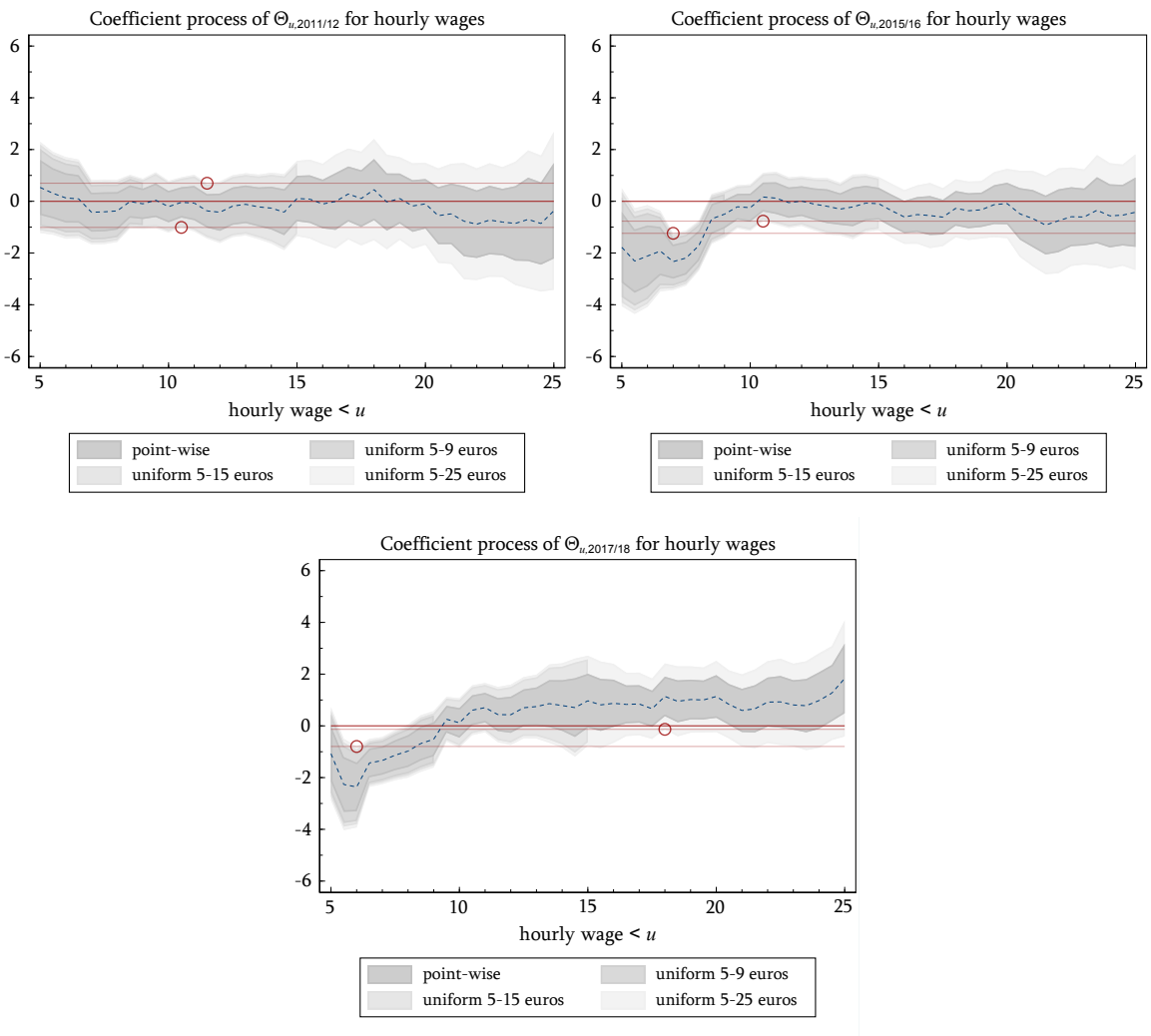


Figure 2.1: Coefficient processes of estimated minimum wage effects $\check{\Theta}_{u,t}$ on the hourly wage distribution

Note: These graphs show the processes of estimated coefficients indexed by thresholds $u \in \mathcal{U}$. Grey shaded areas represent the 90% uniform confidence bands based on 100,000 multiplier bootstrap replications. The circles mark the infimum of the upper confidence bounds and the supremum of the lower confidence bounds.

To assess the economic magnitude of the effects in Figures 2.1 note that Logit coefficients indicate changes in the log-odds ratio. For a worker i in year t with an initial probability p_{it} of falling below a wage threshold u, an increase of Δ_{MWB} in treatment intensity changes the log-odds ratio as follows:

$$\log\left(\frac{p'_{it}}{1-p'_{it}}\right) - \log\left(\frac{p_{it}}{1-p_{it}}\right) = \check{\Theta}_{u,t} \cdot \Delta_{MWB}, \tag{2.4.1}$$

where $\check{\Theta}_{u,t}$ represents the estimated Logit coefficient for the treatment effect, and p'_{it} is the updated probability of falling below a threshold u. This updated probability p'_{it} is computed as:

$$p'_{it} = \frac{\exp\{m_{it}\}}{1+\exp\{m_{it}\}} \quad \text{with} \quad m_{it} = \log\left(\frac{p_{it}}{1-p_{it}}\right) + \check{\Theta}_{u,t} \cdot \Delta_{MWB}. \tag{2.4.2}$$

As an example, consider the Logit coefficients of around minus two for wage thresholds below 8.50 euros in the top-right panel of Figure 2.1. For a worker i in year t, whose initial probability of falling below a threshold is $p_{it} = 0.7$, an increase in the exposure to the minimum wage of $\Delta_{MWB} \in \{0.1, 0.2, 0.3\}$ reduces the probability of falling below a threshold to $p'_{it} \in \{0.656, 0.610, 0.561\}$. These are substantial reductions.

Figure 2.1 presents simultaneous confidence intervals, computed using the multiplier bootstrap over an increasing range of thresholds. By design, these bands widen as more coefficients are included, most notably when shifting from pointwise to simultaneous intervals. The graphs also show a solid red line representing a zero effect. If there is at least one threshold, at which the simultaneous confidence band does not include this zero line, we will reject the hypothesis that the minimum wage had no effect on the distribution of hourly wages. This hypothesis is not rejected in the pre-test period 2011/12 displayed in the top-left panel of Figure 2.1, but rejected in the post-introduction periods 2015/16 and 2017/18 as displayed in the top-right and bottom panels of Figure 2.1. Analogously, we can also test whether effects are homogeneous across the distribution. To do so, the infimum of the upper confidence bounds has to be compared to the supremum of the lower confidence bounds. In the graphs, these points are symbolized by small circles. If the supremum circle lies above the infimum circle, the hypothesis of a constant effect over the entire distribution is rejected. Such effect homogeneity can be rejected for both post-introduction periods 2015/16 and 2017/18, but not for the pre-test period 2011/12.

In order to illustrate the benefits of our method, we also show the results for the main (time and base) effects in the difference-in-differences specification. The time effects displayed in the top-left, top-right and bottom-left panels of Figure 2.2 indicate uniform wage growth over time. Wages were uniformly lower in the pre-test period 2011/12 compared to the reference period 2013/14 as shown in the top-left panel, higher in the first post-reform period 2015/16 as shown in the top-right panel and even more so in the second

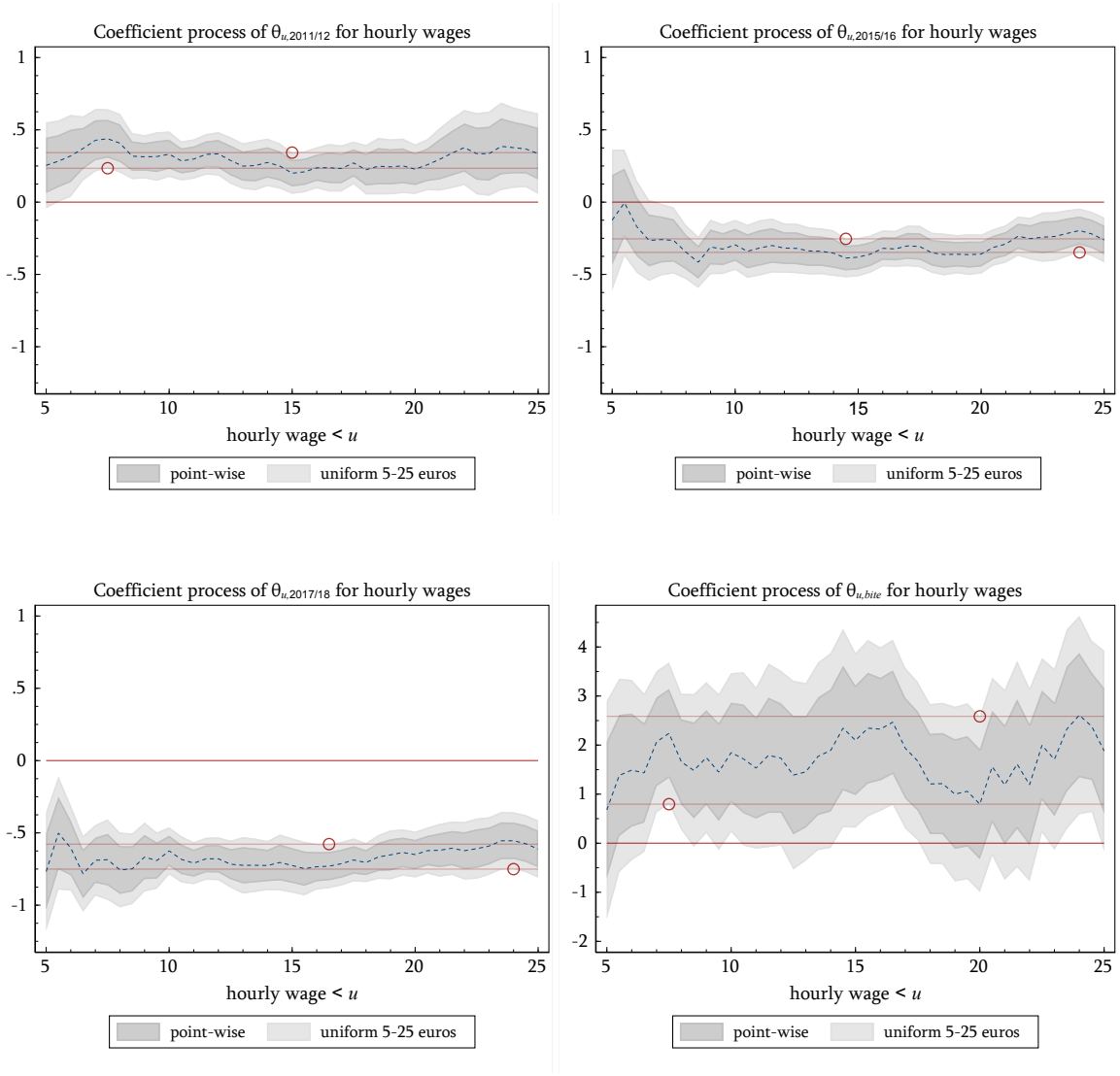


Figure 2.2: Coefficient processes of estimated time and base effects $\check{\theta}_{u,t}$ and $\check{\theta}_{u,bite}$ on the hourly wage distribution

Note: These graphs show the processes of estimated coefficients indexed by thresholds $u \in \mathcal{U}$. Grey shaded areas represent the 90% uniform confidence bands based on 100,000 multiplier bootstrap replications. The circles mark the infimum of the upper confidence bounds and the supremum of the lower confidence bounds.

post-reform period 2017/18 as shown in the bottom-left panel.[6] There is an indication of less wage growth at the lower end compared to the rest of the distribution in 2015/16, but homogeneity cannot be rejected. The base effect of the difference-in-differences specification displayed in the bottom-right panel of Figure 2.2 shows that wages were predominantly lower in high-bite groups, both preceding and succeeding the implementation of the reform. This is what one would expect because wages in high-bite industries are likely to be generally lower.

[6]Note that positive/negative coefficients indicate a higher/lower likelihood of falling below a particular wage threshold, i.e., lower/higher wages.

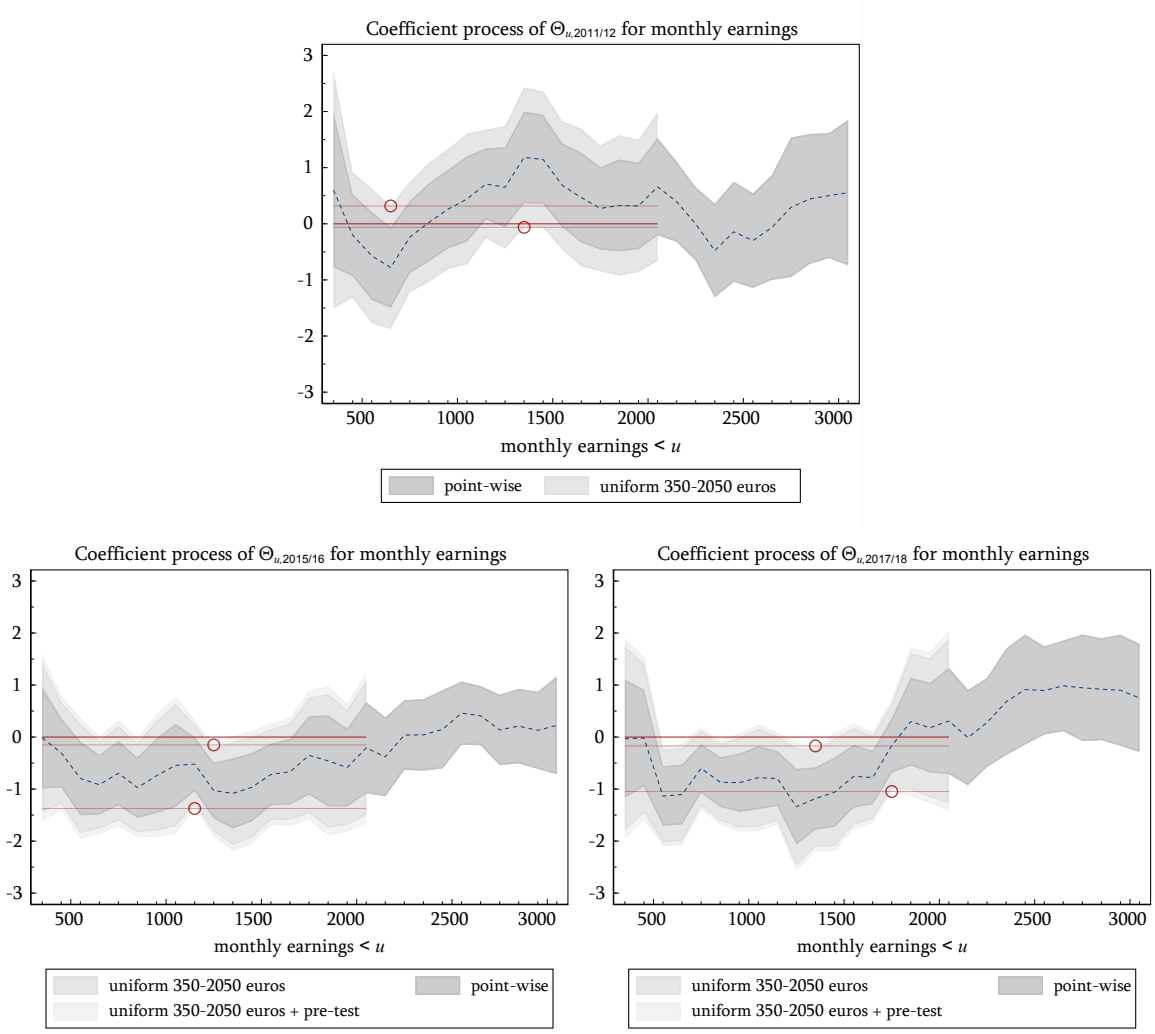


Figure 2.3: Coefficient processes of estimated minimum wage effects $\check{\Theta}_{u,t}$ on the monthly earnings distribution

Note: These graphs show the processes of estimated coefficients indexed by thresholds $u \in \mathcal{U}$. Grey shaded areas represent the 90% uniform confidence bands based on 100,000 multiplier bootstrap replications. The circles mark the infimum of the upper confidence bounds and the supremum of the lower confidence bounds.

Figure 2.3 examines how the minimum wage affected the distribution of monthly earnings, shedding light on which segments of the personal income distribution benefited most. The simultaneous confidence intervals in these figures extend up to 2,050 euros/month, since it is implausible that minimum wage recipients earn more than this amount. In fact, earning over 2,050 euros/month at the minimum wage rate would require working more than 50 hours a week. Our results suggest that the largest gains occurred not among individuals with very low monthly wages, i.e., marginal part-time workers, but among those earning between 700 and 1,700 euros/month, as indicated by the negative coefficients in the post-treatment periods. It is important to note, however, that while the coefficients turn negative

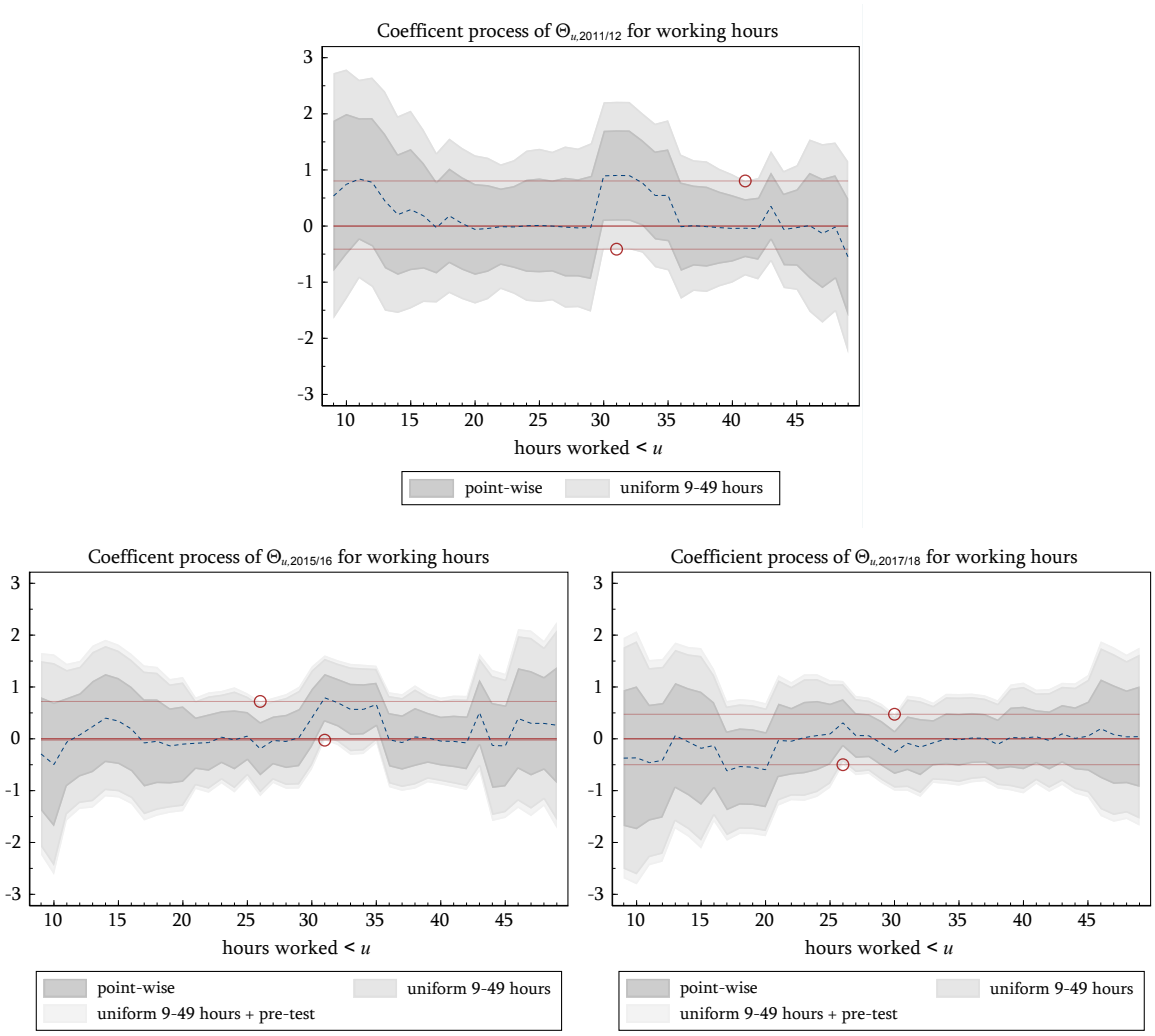


Figure 2.4: Coefficient processes of estimated minimum wage effects $\check{\Theta}_{u,t}$ on the working hours distribution

Note: These graphs show the processes of estimated coefficients indexed by thresholds $u \in \mathcal{U}$. Grey shaded areas represent the 90% uniform confidence bands based on 100,000 multiplier bootstrap replications. The circles mark the infimum of the upper confidence bounds and the supremum of the lower confidence bounds.

in the post-treatment periods, and the hypothesis of no effect can thus be rejected, the uniform confidence intervals are wide, rendering these estimates only marginally significant. This is the likely reason why previous contributions based on the same data set struggled to establish significant effects of the minimum wage introduction on monthly earnings; see Burauel et al. (2019a,b) and Caliendo et al. (2022).

To further illustrate the usefulness of the framework proposed by Belloni et al. (2018b), we demonstrate in the bottom panels of Figure 2.3 the point made by Roth (2022) that conditioning on the result of a pre-test leads to an under-coverage of confidence intervals for difference-in-differences effects. In accordance with equation (2.2.13), confidence intervals

pertaining to the process of a single target coefficient over arbitrary sets of distributional thresholds, as well as the processes of multiple target coefficients, can be constructed. In the bottom-left and bottom-right panels of Figure 2.3, we incorporate the coefficients for potential pre-trends from the top panel of Figure 2.3 into the simultaneous interval. This widens the confidence intervals and demonstrates the "cost" in interval coverage for the treatment effect coefficients associated with testing for the existence of pre-trends. Although this cost is small in our application, it is non-negligible and further reduces the significance of the treatment effect estimates.

An open question arising from Figures 2.1 and 2.3 is the extent to which the observed difference in the effects of the minimum wage on monthly and hourly earnings can be explained by potential changes in working hours, as suggested by Burauel et al. (2019a,b) and Caliendo et al. (2022). The results depicted in Figure 2.4 provide no evidence of significant changes in the distribution of working hours as a result of the minimum wage. This is particularly evident in the second post-reform period 2017/18. Here, we observe a uniform zero effect, as the zero line is fully encompassed in both the point-wise and uniform confidence intervals. In summary, our results indicate that the introduction of the minimum wage neither led to significant reductions in weekly working hours, which would help maintain monthly wage bills at a constant level, nor triggered substantial transitions between employment categories (part-time, marginal part-time, and full-time).

2.5. CONCLUDING REMARKS

This paper uses a distribution regression model to evaluate the effects of the introduction of the German minimum wage in 2015 on the distribution of hourly wages, hours worked and monthly earnings. Our data source is the German Socio-Economic Panel (SOEP) which is characterized by a moderate sample size but a large number of potential control variables. We measure the effects of the minimum wage at each point of our outcome distributions employing flexible machine learning methods recently developed by Belloni et al. (2018b). These methods allow us to automatically specify a large number of parallel Logit models over a fine grid of distributional thresholds, while providing valid statistical inference across ranges of thresholds, after a thorough, machine-led specification search. Our distribution regression analysis provides a more comprehensive picture about the points of the distribution at which the minimum wage had an effect compared to previous contributions. It also

allows us to assess the information content of our data base in an objective manner, unaffected by subjective decisions (p-hacking) or pre-tested specification choices. Our findings suggest that the minimum wage displaced hourly wages below its minimum level, benefited monthly wages in the lower-middle segment of the distribution but not at the lowest end, and did not significantly alter the distribution of working hours. These findings help reconcile conflicting results in the literature regarding the effects of the German minimum wage, which had previously varied depending on the data source used.

Chapter 3

SELECTIVITY CORRECTED WAGE DISTRIBUTIONS AND THE EVOLUTION OF THE GERMAN GENDER WAGE GAP[‡]

3.1. INTRODUCTION

Since the seminal works of Heckman (1974, 1979), the issue of unobserved selectivity has been a central concern in labour economics. It is well-established that selection on unobservables can bias the measurement of wage disparities between groups of workers if one group is more positively or negatively selected than the other. For instance, the wage gap between men and women will be underestimated if women with less favourable unobserved characteristics are more likely to stay out of the labour market. A growing body of literature has examined the effects of such unobserved selectivity on the gender gap in full-time wages; see, e.g. Mulligan and Rubinstein (2008), Olivetti and Petrongolo (2008), Albrecht et al. (2009), and Arellano and Bonhomme (2017). For a long time, selection correction was targeted at women, whose self-selection behaviour and the resulting labour market participation – which is traditionally lower than that of men – was typically assumed to be the main driver of the bias. This view has shifted in light of the fluctuations in male employment shares in many industrialised countries caused by events such as the global financial crisis; see, e.g. Maasoumi and Wang (2019), Dolado et al. (2020), and Elass (2024). A distinguishing feature of recent contributions is that the methodological advances of Arellano and

[‡]This chapter is a modified version of an unpublished working paper as at 14 February 2025.

Bonhomme (2017) and Chernozhukov et al. (2023) allow researchers to investigate the impact of unobserved selectivity in a manner that extends beyond mean or median outcomes, encompassing the entire outcome distributions.

This paper contributes to the expanding literature by providing new evidence on the role of unobserved selectivity in shaping male and female wage distributions in Germany. Specifically, we contribute threefold. Firstly, in contrast to the majority of preceding studies on this topic, the present paper appears to be the first empirical application of a distribution regression model with sample selection correction, as proposed by Chernozhukov et al. (2023). By choosing this method, we are able to flexibly investigate the evolution of unobserved selectivity patterns across the entire outcome distribution. Secondly, our analysis is based on high-quality German administrative data that have been used in a number of influential papers on the wage distribution, including those by Dustmann et al. (2009), Card et al. (2013), Dustmann et al. (2022b), and Bossler and Schank (2023). One challenge in using administrative data to study unobserved selectivity is the lack of household variables that can be used as instruments. We propose using group variables based on labour market dynamics as instruments for unobserved selectivity. In addition, our study spans distinct phases of the business cycle (recession 2000-2005 and labour market boom 2012-2017), enabling us to examine the influence of the cycle on unobserved selectivity. Thirdly, in contrast to the majority of previous studies, we consider unobserved selectivity in both full-time and part-time wages, whereas the existing literature has almost exclusively focused on the full-time case. Given the substantial role of female part-time employment in many industrialised countries and the growing importance of male part-time employment, the part-time case appears increasingly relevant.

Our results suggest an important role for unobserved selectivity in male and female wages in Germany. For full-time men, selection on unobservables was largely neutral over a wide range of the distribution in the past, but turned negative as more men were drawn into a booming labour market after 2012. At the lower end of the wage distribution male selectivity tends to be positive, potentially due to the protective effect of Germany's generous social safety net. For full-time women, we find generally negative unobserved selectivity, which may be attributed to assortative matching in combination with the German tax and transfer system that discourages secondary earnings. Again, selectivity is less negative towards the bottom of the distribution. In the case of part-time work, our results show negative selectivity for men, while for women, selectivity patterns are more complex, shifting from negative in the lower half to positive in the upper part of the distribution. Our findings reveal pronounced heterogeneity in selectivity patterns across the wage distribution, which has not been recognised by the previous literature. Gender gaps in full-time as well as in part-time wages have narrowed between 2000-2005 and 2012-2017. For the full-time gen-

der wage gap, declining differences in unobserved selectivity between men and women and improved observables for women explain most of the change. For part-time work, declining differences in unobserved selectivity also play a significant role, but most of the changes are explained by a convergence in wage returns to male and female part-time jobs.

3.2. RELATED LITERATURE

A growing body of literature has studied wage differences between men and women accounting for unobserved selectivity. There is a broad consensus in the literature that conventional measures of the gender pay gap may over- or underestimate inequality between men and women if such selectivity is not considered. Researchers have proposed different approaches to recover wage distributions that are purged of selection. Important early contributions focussing on differences in mean or median wages include Mulligan and Rubinstein (2008) and Olivetti and Petrongolo (2008). Employing selectivity corrections based on Heckman (1979), Mulligan and Rubinstein (2008) find that female selectivity in full-time work in the US shifted from negative to positive in the 1990s, contributing to the narrowing of the observed gender gap.[1] Olivetti and Petrongolo (2008) impute the unobserved wages of the non-employed to study the impact on selection on wage gaps in OECD countries. Their findings indicate limited effects in most countries, with the exception of southern Europe, where positive selection of women reduced observed gender gaps. A recent study following a similar methodology is Blau et al. (2024). They conclude that correcting for unobserved selection leads to larger declines of the US gender wage gap over time than without correction, implying that changes in selectivity played a role in narrowing the gap.

Extending the analysis beyond the median gender gap, Albrecht et al. (2009) and Chzhen and Mumford (2011) study the impact of unobserved selectivity across the entire wage distribution. In an application for the Netherlands, Albrecht et al. (2009) find positive selection of women into full-time work, reducing observed wage differences between men and women. Chzhen and Mumford (2011) obtain similar results for the UK. Both studies apply a selectivity correction for quantile regression models developed by Buchinsky (1998, 2001). This correction method was later shown to be highly restrictive by Huber and Melly (2015). Biewen et al. (2020) propose a modification to the method of Albrecht et al. (2009)

[1] The reader is referred to Beblo et al. (2003) for an equivalent study for European countries.

in order to address the problem pointed out by Huber and Melly (2015). Their application to the full-time gender wage gap in Germany still produces similar findings as in Albrecht et al. (2009) and Chzhen and Mumford (2011). Employing a related method, Fitzenberger and de Lazzer (2022) examine the effect of unobserved selectivity on male full-time wages in Germany. Their results suggest positive selection of men into full-time work, which became less positive as more men entered the expanding labour market.

In a significant methodological advance, Arellano and Bonhomme (2017) developed a direct method for correcting entire distributions of outcomes for unobserved selectivity. Their approach entails the selection specific rotation of the quantile indices of individual observations. This is achieved by modelling the joint distribution of unobservables in participation and wage equation by a bivariate copula; see Biewen and Erhardt (2021). Arellano and Bonhomme (2017) apply their quantile regression method to male and female wages in the UK. In contrast to previous contributions, they also explicitly model unobserved selectivity for males, finding positive selection on unobservables for both men and women. Maasoumi and Wang (2019) apply Arellano and Bonhomme's estimator to US data. Their results point to negative male and female selection into full-time work that turned positive in the 1990s. Chen et al. (2024) propose a modification of Arellano and Bonhomme's original method. They obtain significant negative selection among males and positive selection for females based on the same data as Arellano and Bonhomme (2017). Arellano and Bonhomme's estimator has been applied to data from various countries, leading to diverse findings. Dolado et al. (2020) use the method to examine employment and wage patterns in EU countries before and after the Great Recession. The authors obtain varied results for different countries, suggesting that male selection became positive during the recession, while positive female selection decreased in some countries due to an added-worker effect (low-ability women joined the labour market in the recession). Also using the method developed by Arellano and Bonhomme (2017), Elass (2024) presents results for France, Finland, and the UK: her analysis implies positive male and female selection on unobservables for France, whereas it is negative for Finland and the UK. Pereda-Fernandez (2024) applies the Arellano and Bonhomme estimator to the US, finding positive selection for both men and women, but a different evolution of male and female selectivity patterns reduces the observed gender wage gap over time.

The quantile regression approach of Arellano and Bonhomme (2017) imposes the restriction of a constant degree of selectivity over all quantile indices, as represented by the scalar copula parameter. Recently, Chernozhukov et al. (2023) have proposed an alternative model based on a bivariate Probit that allows for heterogeneity in the selection structure to be captured. In Section 3.3, we give a more detailed description of this method. In their application to the UK, Chernozhukov et al. (2023) find significant differences in selectivity

across the distribution: in the past, male unobserved selectivity was negative at the bottom and positive at the top, but became more uniformly positive in more recent years. In contrast, female selectivity is generally negative, although it has recently become less severe and more uniform. These results indicate that methods assuming homogeneous selectivity may miss important aspects of the selection behaviours of men and women. The present paper contributes to this recent strand of the literature by employing the method developed by Chernozhukov et al. (2023) to study selectivity patterns in male and female wage distributions in Germany. Apart from full-time wages, we also examine selectivity in part-time wages. To the best of our knowledge, only a limited number of previous contributions have considered the issue of unobserved selectivity with regard to part-time wages. With the exception of Gallego-Granados (2019), who employs a distributional imputation technique due to Melly and Santangelo (2014), the majority of these studies use elementary techniques for selection correction; see, e.g. Manning and Petrongolo (2008), Bardasi and Gornick (2008), Matteazzi et al. (2014).[2]

We conclude by mentioning a related body of literature that focuses on selectivity at the intensive margin (working hours). For further details, the reader is referred to Fernandez-Val et al. (2023), Fernandez-Val et al. (2024a), and Fernandez-Val et al. (2024b). We emphasize that our dataset only allows for a distinction between full-time and part-time, which prevents us from applying these techniques.

3.3. ECONOMETRIC METHOD

This section presents a short description of the method proposed by Chernozhukov et al. (2023) to estimate selectivity corrected wage distributions. This method allows for more flexible patterns of selectivity compared to the approaches of Heckman (1974), Buchinsky (1998), and Albrecht et al. (2009) as the sign and the magnitude of selection may differ across the outcome distribution.

[2]Gallego-Granados and Wrohlich (2020) apply the method of Melly and Santangelo (2014) to full-time wages.

Consider the following selection process:

$$\begin{aligned} D &= \mathbf{1}\{D^* \leq 0\}, \\ Y &= Y^*, \text{ if } D = 1, \end{aligned}$$

where Y^* is the wage a person would receive if she decided to work. The observability of the wage offer is contingent upon the individual being employed, which in turn depends on the latent propensity to work D^*. This is related to the difference between the wage offer and the reservation wage of the person. It is crucial to account for selectivity as the distribution of observed outcomes Y typically differs from the distribution of latent outcomes Y^* due to the endogeneity of employment decisions D, since employment decisions depend on wage offers Y^*. For example, if, in a group of potential workers, individuals with low wage offers opt out of employment, the resulting wage distribution appears more favourable than it is in reality, as lower wage levels are not represented among those who are actually employed.

Let F_{Y^*} and F_{D^*} denote the marginal cumulative distribution functions (CDF) of Y^* and D^*, respectively, and let F_{Y^*,D^*} be their joint CDF. Chernozhukov et al. (2023) show that the latter can generally be represented by a standard bivariate normal distribution with local correlation parameter ρ evaluated at point $(y, d) \in \mathcal{Y} \times \mathcal{D} \subset \mathbb{R}^2$. Notably, this representation does not assume the joint distribution of Y^* and D^* to be normal. Moreover, it establishes that any joint distribution function of two random variables can be represented by a sequence of bivariate normals. Chernozhukov et al. (2023) call this unique representation the Local Gaussian Representation (LGR):

$$F_{Y^*,D^*}(y, d) = \Phi_2\big(\mu(y), \nu(d); \rho(y, d)\big), \tag{3.3.1}$$

where $\Phi_2(\cdot)$ denotes the bivariate standard normal CDF, $\Phi(\cdot)$ represents the univariate standard normal CDF, and $\mu(y) = \Phi^{-1}(F_{Y^*}(y))$, $\nu(d) = \Phi^{-1}(F_{D^*}(d))$. The parameter $\rho(y, d) \in [-1, 1]$ measures the local dependence between the two dichotomous variables $\mathbf{1}\{Y^* \leq y\}$ and $\mathbf{1}\{D^* \leq d\}$ at the threshold tuple $(y, d) \in \mathcal{Y} \times \mathcal{D}$. In the context of our present labour market application, parameter $\rho(y, 0)$ quantifies the degree to which the occurrence of an individual being employed, $\mathbf{1}\{D^* \leq 0\}$, and concurrently receiving a wage offer below a specified threshold y, $\mathbf{1}\{Y^* \leq y\}$, are correlated. Non-zero correlations cause a selection bias in the sense that the distribution of observed wages F_Y differs from the distribution of wage offers F_{Y^*}, that would be observed in the hypothetical scenario in which every individual is employed.

Chernozhukov et al. (2023) demonstrate that point identification of the parameters in equation (3.3.1) is achieved using at least one binary instrumental variable Z_1. To this end, define a vector of covariates X determining wage offers Y^* and let $Z = (Z_1, X)$. We impose

the exclusion restriction that instruments Z_1 influence the propensity to work, but not the wage offers, nor their correlation with the propensity to work. It follows that

$$\begin{aligned} F_{Y^*,D^*|Z}(y,d \mid z) &= \Phi_2\big(\mu(y \mid z), \nu(d \mid z); \rho(y,d \mid z)\big) \\ &= \Phi_2\big(\mu(y \mid x), \nu(d \mid z); \rho(y,d \mid x)\big). \end{aligned} \tag{3.3.2}$$

Chernozhukov et al. (2023) propose to parametrise the unknown parameters in (3.3.2) by $\mu(y \mid x) = -x\beta(y)$, $\nu(d \mid z) = -z\pi$ and $\rho(y,d \mid x) = \rho(x\delta(y)) = \tanh(x\delta(y))$ by the Fisher link function. For the remainder, these will be referred to as outcome, selection, and sorting equations, respectively.[3] The joint CDF of (Y^*, D^*) conditional on Z can be represented in terms of its LGR:

$$F_{Y^*,D^*}(y,0 \mid Z=z) = \Phi_2\big(-x\beta(y), -z\pi; \rho(x\delta(y))\big) \tag{3.3.3}$$

$$= \mathrm{P}(Y \leq y, D=1 \mid Z=z), \tag{3.3.4}$$

where the second line of this equation implies that the joint distribution $F_{Y^*,D^*|Z}$ can be recovered by estimating a series of selection corrected bivariate Probit models for wages being below a fine grid of thresholds in the support of the outcome $y \in \mathcal{Y}$. This representation of the selection problem motivates the use of a distribution regression approach of the form:

$$F_{Y|X}(y \mid z) = \mathrm{E}[\mathbf{1}\{Y \leq y\} \mid Z=z, D=1], \tag{3.3.5}$$

where Z denotes covariates and instruments, and $F_{Y|Z}$ stands for the CDF of observed wages Y conditional on Z. This representation of the selection problem is highly flexible, as it allows for heterogeneous returns $\beta(y)$ as well as for heterogeneous selection sorting $\rho(x\delta(y))$ across the wage distribution. Moreover, selection patterns may vary depending on the covariates, adding another layer of heterogeneity.

This model can be used to recover a variety of distributions of interest. Firstly, one may wish to recover the wage offer distribution, i.e., the distribution of wages that is free of selection bias:

$$F_{Y^*}(y) = \int \Phi(y \mid Z=z)\,\mathrm{d}F_Z(z) = \int \Phi\big(-x\beta(y)\big)\,\mathrm{d}F_X(x), \tag{3.3.6}$$

which can be estimated by its empirical counterpart

$$\widehat{F}_{Y^*}(y) = N^{-1}\sum_{i=1}^{N} \Phi\big(-X_i\hat{\beta}(y)\big). \tag{3.3.7}$$

[3] Chernozhukov et al. (2023) add the minus signs to facilitate the interpretation of the respective parameters in terms of comparability with the classical Heckman selection model, where selection is defined by $D^* > 0$ rather than by $D^* \leq 0$.

Below, we will use this distribution to describe differences in the distribution of male and female wages free of selection bias, and to describe changes in wage selectivity over time.[4] The distribution of the observed outcome Y implied by the model is given by

$$\begin{aligned} F_Y(y) &= \int \frac{\Phi_2\big(-x\beta(y), -z\pi; \rho(x\delta(y))\big)}{\Phi(-z\pi)} \, \mathrm{d}F_Z(z \mid D=1) \\ &= \frac{\int \Phi_2\big(-x\beta(y), -z\pi; \rho(x\delta(y))\big) \, \mathrm{d}F_Z(z)}{\int \Phi(-z\pi) \, \mathrm{d}F_Z(z)}, \end{aligned} \tag{3.3.8}$$

the sample analogue of which is

$$\widehat{F}_Y(y) = \frac{\sum_{i=1}^{N} \Phi_2\big(-X_i\hat{\beta}(y), -Z_i\hat{\pi}; \rho(X_i\hat{\delta}(y))\big)}{\sum_{i=1}^{N} \Phi(-Z_i\hat{\pi})}. \tag{3.3.9}$$

Equation (3.3.9) can be used for the construction of counterfactual distributions of the observed outcome, which may serve as the basis of decomposition exercises exploring the sources of wage differences between groups of workers (men and women), or across periods.

Algorithm 3.3.4 *(Two-Step Distribution Regression with Sample Selection – essentially Algorithm 3.1 in Chernozhukov et al., 2023).*

Step 1. Run the Probit estimator of D_i on $Z_i = (Z_{1i},\, X_i)$ to obtain:

$$\hat{\pi} = \arg\max_{\pi} \sum_{i=1}^{N} \big\{ D_i \log \Phi(Z_i\pi) + (1 - D_i) \log \Phi(-Z_i\pi) \big\}.$$

For $i \in \{1, \ldots, N\}$ compute and save $Z_i\hat{\pi}$.

Step 2. For a fine grid of points $y \in \mathcal{Y}$, run the bivariate Probit of $I_{yi} = \mathbf{1}\{Y_i \le y\}$ if $D_i = 1$ on covariates X_i using the plug-in estimate $Z_i\hat{\pi}$ from Step 1 to obtain:

$$\begin{aligned} \big(\hat{\beta}(y), \hat{\delta}(y)\big) = \arg\max_{(\beta,\delta)} \sum_{i=1}^{N} D_i \big\{ & I_{yi} \log \Phi_2\big(-X_i\beta, Z_i\hat{\pi}; -\rho(X_i\delta)\big) \\ & + (1 - I_{yi}) \log \Phi_2\big(X_i\beta, Z_i\hat{\pi}; \rho(X_i\delta)\big) \big\}. \end{aligned}$$

Depending on the question at hand, the desired counterfactuals can be constructed by combining the estimated $\hat{\beta}(y)$, $\hat{\pi}$ and $\rho(x\hat{\delta}(y))$, as well as the marginal distributions of covariates, $\widehat{F}_Z$, of the two groups. For example, in the context of the gender wage gap, we may ask what the distribution of female wages would look like if women sorted into employment like men, thereby counterfactually changing $\rho(x\hat{\delta}(y))$ to that of men. Equation (3.3.9) also serves as a specification check for the model, as this distribution should coincide with the empirically

[4] In order to ensure the monotonicity of constructed distributions (3.3.7), we use the method of monotone rearrangement as in Chernozhukov et al. (2023).

observed outcome distribution if the model is correctly specified. In the empirical analysis described below, this is the case.

Remark 3.3.6 *(Optimization of the Log-Likelihood).*

As opposed to the univariate Probit likelihood in Step 1, the log-likelihood of the bivariate Probit in Step 2 of Algorithm 3.3.4 is not globally concave. In fact, indefiniteness of the Hessian regularly originates in column indices associated with the selection sorting parameters $\delta(y)$. To handle such cases if they occur, or rather, to prevent them altogether, we start by estimating a constant $\rho(y)$ model, where $\delta(y)$ only consists of an intercept. For this simplified model, Step 2 can be replaced by a standard Probit of $I_{yi} = \mathbf{1}\{Y_i \leq y\}$ if $D_i = 1$ on covariates X_i and additionally the inverse Mill's ratio $\phi(Z_i\hat{\pi})/\Phi(Z_i\hat{\pi})$. Because of the concavity of the Probit log-likelihood, this initial estimation always converges to a maximum and yields useful starting values for $\beta(y)$ and $\delta(y)$ (provided that Z_1 is a valid instrument). Based on these, we can evaluate the bivariate log-likelihood in Step 2 and either refine the starting values by cycling through a fixed number of updates of the constant $\rho(y)$ model, or immediately optimize the actual, more complex likelihood, where all parameter indices of $\delta(y)$ except for the constant are initialized at zero. The exact strategy should always depend on the topography of the log-likelihood about the current parameter values. To identify non-concave regions, we use the standard test for negative definiteness by attempting to compute the Cholesky decomposition of the negative Hessian; see, e.g. Higham (1988) and Higham et al. (2016). If this test goes through, we can compute and take the standard Newton-Raphson update. However, if the negative Hessian cannot be decomposed, we have to modify the matrix by backing it up with a positive definite "donor-matrix". The most common approach involves shrinkage to artificially create a positive definite matrix, as discussed in the seminal contributions of Levenberg (1944) and Marquardt (1963). Note that we only consider the optimization to be successfully completed, if the Hessian does not have to be backed up in the optimum. In this case, we reached that local maximum which is closest to the initial starting values. To avoid excess step sizes, we additionally employ a line search as described in, e.g. Grippo et al. (1986).

To estimate the model parameters π, $\big(\beta(y)\big)_{y\in\mathcal{Y}}$, and $\big(\delta(y)\big)_{y\in\mathcal{Y}}$, we use the computationally attractive two-step method summarized in the Algorithm 3.3.4. Step 1 is analogous to the first step in the classical Heckman or Arellano and Bonhomme's selection model and consists of a univariate Probit model of the selection into employment to estimate the parameters π in the selection equation. The second step comprises a sequence of selection corrected, bivariate Probit regressions over a fine grid of thresholds $y \in \mathcal{Y}$ to estimate $\beta(y)$ and $\rho(x\delta(y))$. Notably, the first-stage parameter estimates $\hat{\pi}$ are held fixed in this second step

to reduce the computational burden, since the Probit from Step 1 already yields consistent estimates for the selection equation.

A frequent source of convergence issues in the second step is the flexibility of the selection sorting parametrization. The evaluation of the bivariate CDF fails, if the linear combination $X_i\delta$ exceeds a certain threshold by absolute value. To ensure numerical stability, it is necessary to either limit $|X_i\delta| \leqslant \bar{c}$, where e.g. $\bar{c} = 5$, thereby artificially creating flat regions in the log-likelihood, or preferably, to reduce the dimension of parameter δ in the selection sorting equation.[5] The latter approach amounts to imposing further "exclusion restrictions" on the selection sorting equation. In effect, the local correlation parameter is restricted to vary only on a specific subset of the covariates, e.g., marital status or year indicators.

Pointwise standard errors of the model parameters in all equations can be computed based on the usual asymptotic expansions of the likelihood score. The estimated standard errors of $\beta(y)$ and $\rho(x\delta(y))$ need to be adjusted for the use of the first-stage plug-in estimates. Uniform confidence bands for functionals of the parameters are obtained by the multiplier bootstrap described in Chernozhukov et al. (2013, 2023). In Appendix C.1.8, we provide a detailed overview of the algorithms we apply.

3.4. DATA

Our analysis is based on administrative data from the Sample of Integrated Labour Market Biographies (SIAB), a 2% random sample of the Integrated Employment Biographies (IEB). The IEB contain the administrative records of all employees liable to social security contributions in Germany and report precise to-the-day information on individual employment status and the daily wage received.[6] Starting in the year 1975 for West Germany

[5]The principal issue is that the Fisher link $\rho(X_i\delta) = \tanh(X_i\delta)$ rapidly flattens out ($\tanh(5) \approx 0.9999$) thus limiting its effective range. In the context of the simplified bivariate Probit with scalar correlation ρ, Terza and Tsai (2006) propose the re-parametrization $\beta^* := \beta/\sqrt{1-\rho^2}$, $\rho^* := \rho/\sqrt{1-\rho^2}$, where $\rho^* \in \mathbb{R}$ is unrestricted. However, as we intend to estimate index-specific correlations $\rho(X_i\delta)$, it is not possible to apply this substitution in the LGR setting.

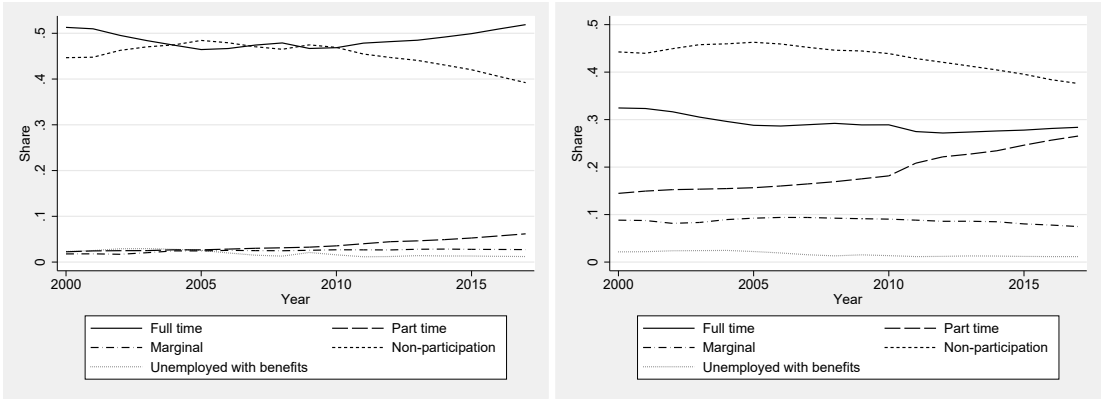
[6]Wages are only reported up to the social security contributions ceiling, resulting in right-censoring that affects roughly 5–12% of all wage observations in each year. We impute these wages following Gartner (2005), as is common practice in the literature.

and in 1992 for East Germany, the SIAB records the entire employment histories of approximately 1.8 million individuals. Since 1999, the SIAB additionally reports episodes in so-called "minijobs" (also referred to as "marginal employment"), which are not subject to social security contributions and pay a low wage not exceeding a certain threshold ("minijob" threshold). For our analysis, we determine each individual's main employment status on June 30th of a given year. This status is then assigned to one of three employment categories: "full-time", "part-time" and "non-participation". The categories of "full-time" and "part-time" refer to employment subject to social security contributions, while the category of "non-participation" refers to non-participation in any form of employment subject to social security contributions. Throughout this paper, we refer to employment subject to social security contributions as "regular" employment.

The SIAB is widely considered the most informative data source on employment and wages in Germany. Its administrative nature and large sample size make it an obvious choice for the given application, particularly because it is less prone to limitations typically associated with survey data. Since the reported wages serve as the basis for the calculation of pensions within the German social security system, the wage and employment data are necessarily free from measurement error. Similarly, attrition is not a concern since an individual's entire employment history is tracked as soon as the person's first employment episode is recorded. It thus follows that in the event of a gap occurring before, after, or between two adjacent employment episodes, the individual in question cannot have been employed subject to social security contributions or in a "minijob" during this time frame. In our selection model, such gaps are generally labelled as "non-participation". It is important to mention that the data generally do not allow us to distinguish between different forms of non-participation in regular employment. Individuals in the pool of "non-participants" may be unemployed or not active in the labour market, including those who are currently in education or on parental leave. In fact, they may work for pay, but are not liable to social security contributions. This applies to civil servants, the self-employed and those in undeclared work or work abroad. Figure C.1 in the appendix shows that 4–5% of the German workforce are civil servants. Moreover, self-employment is much more prevalent among men; about 12.5–14% over the observation period vs. 7.5% for women.

For our main selection models we use a pooled sample of all years from 2000 to 2017, but we estimate the LGR separately for men and women. We construct our sample in the following way. An individual is included in the sample as soon as we observe a single employment spell in either "full-time", "part-time", or "marginal part-time" in any of the years 1975 to 2017. Furthermore, individuals are only included for periods in which they are aged between 20 and 60 years. For periods during which the individual is not observed in regular full-time or part-time employment, we assign them the "non-participation" status.

Marginal employment is also assigned to the "non-participation" state as we only distinguish between regular full-time, regular part-time, and non-participation in regular employment. Note that our sample excludes individuals who never participated in one of the employment forms registered in our data. This includes individuals who never actively participated in the labour market or who always worked as civil servants or self-employed. The broad definition of sample membership in our analysis, which considers an individual to be a sample member as soon as there is one observed employment spell at any time, is motivated by the fact that it is only in this way that we can capture many forms of female non-participation, which often takes the form of employment interruptions due to family-related reasons. However, it should be noted that our data does not include direct information on events such as child-birth, transitions into self-employment or civil service, or general household information.



(a) Men (b) Women

Figure 3.1: The evolution of employment shares in Germany

Note: These graphs illustrate the evolution from 2000 to 2017 of the shares in full-time, part-time and marginal employment, as well as the shares of non-participation and the unemployed for men in the left and for women in the right panel.

Figure 3.1 provides an overview of employment trends in our final sample. The period we study spans from 2000 to 2017, including a time characterised by rising unemployment peaking in 2005, followed by a recovery phase that remained uninterrupted by the global financial crisis. This led to an unprecedented employment boom after 2010. The diversity of business cycle phases within our time frame makes it especially suitable for analysing the impact of shifting labour market conditions on employment sorting. The graph displays the shares of persons in full-time, part-time, marginal and non-participation as well as of those currently unemployed with regular unemployment benefits "Arbeitslosengeld I".[7] The left panel shows that male full-time employment shares were at their lowest point in 2005, where both male unemployment and non-employment were the highest, whereas

[7]The figures for unemployment and marginal employment are only presented for descriptive purposes. In our analysis, we subsume these labour states in the "non-participation" category.

they steadily increased after 2008. Likewise, the growth in male part-time employment picked up in 2010, while non-employment declined significantly.

The evolution of female employment shares followed a somewhat different pattern. The right panel shows that, the employment boom was particularly driven by an increase in part-time employment, with its share rising sharply after 2010, while full-time employment for women saw only a modest recovery over the same period. By 2017, the shares of regular part-time and full-time employment had nearly converged. Marginal part-time employment was much more common among women, though its share remained stable at just under 10% throughout the observation period. Similar to men, non-employment among women declined significantly between 2005 and 2017. These shifts in employment shares for both men and women likely affected the composition of the employed and non-employed populations, potentially influencing inequality measures like the gender pay gap. Our analysis seeks to identify and adjust for such compositional changes in both genders.

Recall that the LGR selection model comprises three equations: the outcome equation, the selection equation, and the sorting equation. The dependent variable in the outcome equation is a binary for receiving a log real daily wage below a particular threshold $y \in \mathcal{Y}$ in a given employment state. An indicator of this employment state, in turn, serves as the dependent variable in the selection equation. Since we carry out separate estimations for full-time and part-time employment, the pools of selected and non-selected individuals differ depending on the employment type considered. The two groups need to be chosen carefully since their exact definitions bear important implications for the estimated latent distributions. In our model of full-time employment, we include those in part-time, marginal part-time and non-participation into the non-selected group, such that the resulting wage offer distribution is one that would result if all individuals in the population received full-time wages. In contrast, we exclude full-time workers from our analysis of part-time wages, considering only individuals in marginal part-time employment and those not participating in the regular labour force as the non-selected group. Consequently, the latent distribution of part-time wages reflects what would occur if all workers in the residual group of non-full-time employees were to receive part-time wages. This specification was chosen to avoid creating an inconsistent comparison group for part-time employees that would arise if both full-time workers and non-employed individuals were classified as non-selected in part-time employment.

As covariates determining the wage offer, we include six categories of age (25-29, 30-34, …, 55-60), second-order polynomials of work experience in full-time, part-time and marginal employment measured in years as well as indicators for zero experience in each of the three categories, regional (state-level) dummies, an indicator for German nation-

ality, and the following categories of educational attainment: (1) lower/middle secondary education only (reference category), (2) lower/middle secondary education and completed vocational training, (3) upper secondary education ("Abitur") only, (4) upper secondary education and completed vocational training, (5) degree from a university of applied sciences ("Fachhochschule") and (6) university degree. Tables C.3.1 and C.3.2 in Appendix C.3 provide summary statistics for our selected and non-selected groups by gender. To allow for time effects in a flexible way, the outcome equation also includes indicators for the years 2001 to 2017. Similarly, a full set of year dummies is included in the sorting equation, which thereby determine the sign and magnitude of unobserved selectivity across years and distributional thresholds. A complete overview of the covariates used in each of the three equations is given in Table C.3.3 in Appendix C.3.

As discussed in Section 3.3, and as in the vast majority of contributions in the literature, we leverage instrumental variables in order to estimate selectivity.[8] Suitable instruments for selection can be hard to find depending on the nature of the data used. Instrumental variables that have been used in the literature include the number of small children in the household, often in combination with marital status, the husband's income, or potential out-of-work income; see Mulligan and Rubinstein (2008), Albrecht et al. (2009), Maasoumi and Wang (2019), Buchinsky (2001), Arellano and Bonhomme (2017) and Elass (2024). The first two options are not feasible in our analysis because our data lack household information. Additionally, we cannot utilise potential out-of-work income, as it is closely tied to earnings potential, i.e. prior earnings, and household characteristics within the German unemployment insurance system. Moreover, family-related instruments suffer from the essential drawback that they typically apply better to women than to men, while one of our objectives is to describe selectivity patterns for both men and women.

Instead, we opt for a group-instrument strategy that leverages variation in employment dynamics across groups of workers and over time. To the best of our knowledge Fitzenberger and de Lazzer (2022) was the first study using this type of instrumentation strategy, which addresses the fact that administrative data typically lack other information that can be used for instrumentation. In our application, we exploit employment dynamics within worker cells defined by sex, year, age, educational attainment, and local labour markets "Raumordnungsregionen" (ROR), in order to explain varying selection into full-time and part-time

[8] A notable and very recent exception is the approach suggested by Chen et al. (2024), who adapt the Arellano and Bonhomme (2017) method by replacing the binary selection equation with a censored equation of hours worked, rendering the use of instruments redundant. Since our administrative data only includes full- and part-time status, but not hours worked, this approach is not feasible in our application. D'Haultfœuille et al. (2018) also do not invoke exclusion restrictions, but rely on extreme wage observations for identification. Such observations are are not available in our data due to censoring. Both approaches also do not allow one to study the strength of selection at different points in the distribution.

employment at the level of the individual. We use 15 different indicators which we compute by worker cell: full-time share, part-time share, minijob share, annual first differences of these shares, as well as the full matrix of annual transition rates between full-time employment, part-time employment and non-participation. The reader is referred to table C.3.3 in Appendix C.3 for more details. As additional instruments, we include indicators that mark persons of three different age groups who are likely to be students in the given year: persons who did not hold a college degree in that year, but are reported to have one at a later stage in their employment history.

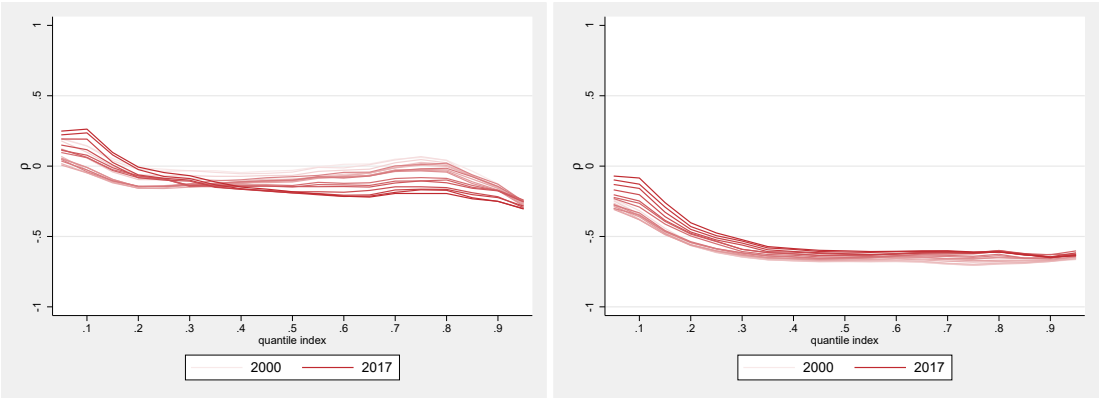
Our instrumental variables are valid provided that the employment dynamics represented by the above indicators influence wages only with a sufficient time lag. We assume this to be the case in the German labour market, where wage rigidities prevent short-term adjustments of wages; see, e.g. Bauer et al. (2007). Wage contracts typically span multiple years, and the remuneration of new employees is usually aligned with the rates paid to existing staff. Additionally, collective bargaining occurs at the level of the industry or region, with a notable lag, and is unlikely to respond directly to developments within the disaggregated worker cells that we have considered. Apart from variation between worker groups and regions, our instruments represent rich variation over time reflecting the different phases of the business cycle (rising unemployment 2000-2005, stagnation 2006-2009, boom 2010-2017; see figures C.2 and C.3 in the appendix). We consider the exclusion restriction underlying our instruments not to be stronger than the assumptions made in other studies in the literature. For example, the presence of children and a husband's income have been criticised by Huber and Mellace (2014), while Blundell et al. (2007) question the validity of out-of-work income.

3.5. EMPIRICAL RESULTS

We present three sets of results, both for full-time and for part-time wages. Firstly, we describe male and female patterns of unobserved selectivity and how these evolved over time. Secondly, we decompose the gender gap in latent wages into their main sources of observable worker characteristics and returns to these characteristics. Finally, we decompose differences in observed wage distributions between men and women and across time periods into the different components implied by our rich distributional selection model.

3.5.1. Selection into Full-time Employment

Figure 3.2 presents the estimated sorting parameters for full-time employment against all other employment states, where the sorting equations $\rho(x\hat{\delta}(y))$ include a constant and dummies for the years 2001 through 2017. As in Chernozhukov et al. (2023), we show the sorting parameters at wage thresholds defined by a fine grid of percentiles of the distribution of log real wages with indices $\{0.05, 0.10, \ldots, 0.90, 0.95\}$ computed in the pooled sample of men and women. The change in color saturation, moving from pale to more vibrant red hues, reflects the progression of time from the year 2000 to 2017.



(a) Men (b) Women

Figure 3.2: Correlation coefficient process $\rho(x\hat{\delta}(y))$ of selection sorting into full-time work at increasing quantiles of the log daily wage distribution

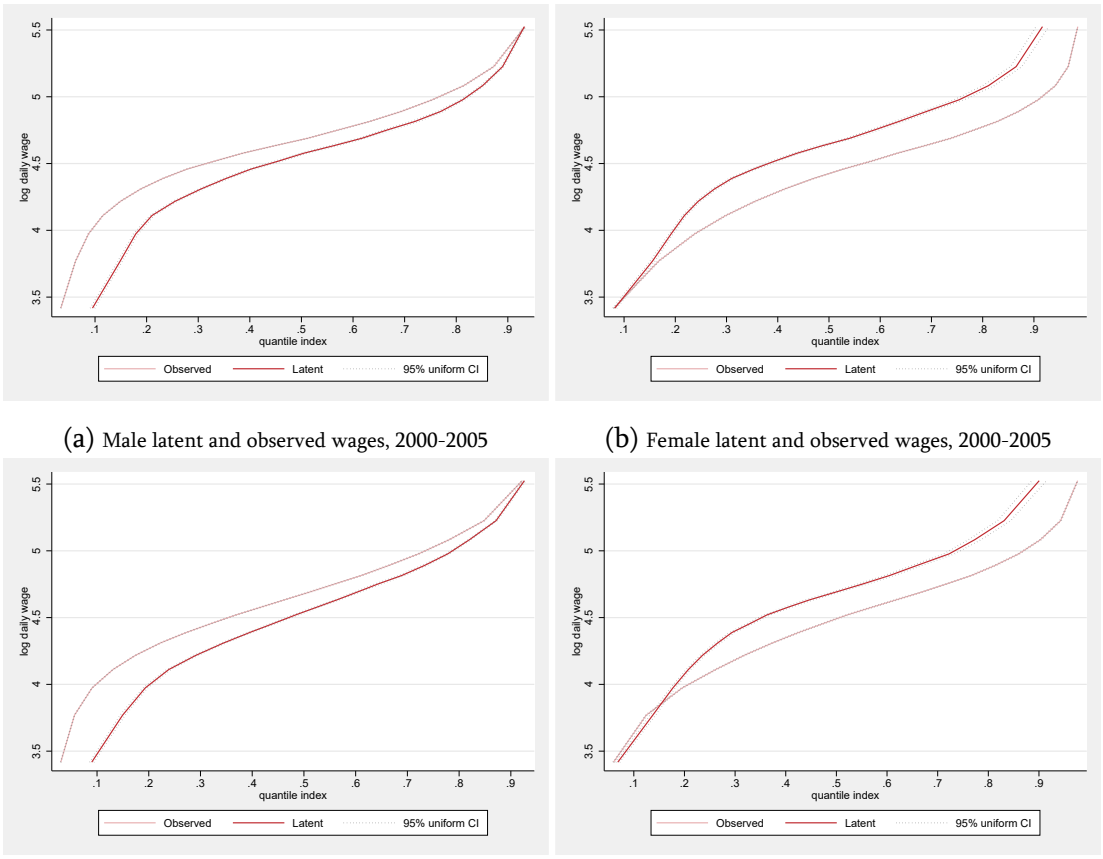
Note: These graphs show the estimated process of the selection sorting coefficients into full-time employment of the year indicators ranging from 2000 to 2017 over increasing quantile indices of the log daily wage distribution for men in the left and for women in the right panel.

Both graphs reveal sizeable heterogeneities in sorting across the wage distribution for both genders, where the estimated selection bias increases in the quantile index by absolute value towards the top for women and changes its sign for men. This challenges the results obtained from other methods that only estimate one single sorting parameter for the whole distribution. Secondly, Figure 3.2 shows non-negligible female and male selection, which has thus far received limited attention in the literature. Thirdly, the observed selection patterns become somewhat more heterogeneous over time, most notably for men. While the selection bias tended to be relatively small across the distribution directly after the millennium, sorting became larger in magnitude as well as increasingly heterogeneous towards more recent years. Similarly, there is a tendency for increasingly heterogeneous sorting from 2000 to 2017 for women, as implied by the steeper $\rho(x\hat{\delta}(y))$ curve in the right panel, but the general pattern does not change as much as it does for men.

Overall, Figure 3.2a suggests that in the early 2000's, a period characterised by high unemployment, men were positively selected at the bottom as well as between the 60th and 80th percentiles and almost non-selected in other parts of the wage distribution. In contrast, male selection is only positive at the bottom, but negative in other parts towards the more recent years that have witnessed record employment levels. This inter-temporal pattern supports the hypothesis that the composition of the workforce in terms of unobservables becomes less favourable during periods of high employment as individuals with less favourable unobservables enter the labour market; see Dolado et al. (2020). Alternatively, individuals with less favourable unobservables are less likely to exit the labour market; see Riphahn and Schrader (2020). Empirically, a similar result was obtained by Fitzenberger and de Lazzer (2022) who conclude that unemployed individuals are negatively selected in times of high employment. A potential explanation for male negative selection in our sample is the fact that our non-selected group partly contains the self-employed and civil servants, who may be positively selected against the employed subject to social security. In Figure 3.2a, we observe positive unobserved selectivity for men at the very bottom of the distribution. A potential explanation for positive selectivity in this part of the distribution is the generous German social safety net, which may disincentivise employment for individuals with the lowest wage offers.

In contrast to men, women experience negative selection across the entire distribution in all years. Selection is close to zero at the bottom, especially in the later years. Negative female selection has been found both by Elass (2024) for Finland and the United Kingdom over a period covering the Great Recession, and by Maasoumi and Wang (2019) for the United States until the 1990s. Likewise, our results align with those of Chernozhukov et al. (2023) who document increasingly negative sorting of British women towards the top of the distribution. From a theoretical point of view, Ermisch and Wright (1994) argue that negative sorting can be plausible when there is a high positive correlation between wage offers and reservation wages. One possible explanation for this is assortative matching on the marriage market, with couples becoming increasingly homogeneous in terms of education, income, and likely also unobserved ability; see, e.g. Calvo et al. (2024). After the birth of a child, in particular, assortative matching may result in high-potential women taking on the bulk of care work because their high-potential partner's income is sufficient to maintain household income, while women with less successful partners face the necessity to contribute to family income. Negative sorting of married women towards the top of the distribution is also consistent with Germany's tax and transfer system, which over-proportionally favours single-income marriages and marriages with a high inter-couple income gap since the secondary earner often faces high marginal tax rates; see Bach et al. (2013). According to Figure 3.2b, female selectivity became somewhat less negative in more recent years. This may be due to improvements in public child care and changing social norms which facilitated the

return to full-time work after child birth; see Geyer et al. (2015).



(a) Male latent and observed wages, 2000-2005

(b) Female latent and observed wages, 2000-2005

(c) Male latent and observed wages, 2012-2017

(d) Female latent and observed wages, 2012-2017

Figure 3.3: Latent and observed full-time log daily wages with 95% uniform confidence band

Note: These graphs show male and female wage distributions in the left and right column, respectively. The period of 2000-2005 is illustrated in the top row and the period of 2012-2017 in the bottom row.

As explained in Section 3.3, our estimates for the selection patterns can be used to recover the latent wage offer distributions for men and women, which can be compared to the observed distribution. Figure 3.3 presents these comparisons separately for the period 2000-2005 and 2012-2017 in order to assess potential changes over time. For men, the left panels show that the observed distribution of full-time wages first-order stochastically dominates the latent distribution, i.e., observed wages are higher than latent wages. This may be surprising given the negative selection pattern for men above the 30th percentile shown in Figure 3.2a. Recall however, that the latent distribution of full-time wages also incorporates observables for men who do not select themselves into full-time work. These are much less favourable than for men who work full-time. In particular, men participating in regular full-time employment are more favourably selected with respect to work experience, education and nationality.

For women, the right panels of Figure 3.3 paint a different picture. With the exception of the lowest income levels, where the latent and observed distributions coincide, the latent full-time wage is higher than the observed wage. This is in line with their distinctively negative selection on unobservables as shown in Figure 3.2b, which dominates the positive selection on observables. For both men and women, latent full-time wage distributions are relatively stable, despite some temporal change in selection on unobservables.

Next, we decompose the gender gap in latent full-time log daily wages between men (group 0) and women (group 1) into a contribution due to differences in wage structures and a contribution due to differences in the composition of worker characteristics:

$$F_{Y^*\langle 1,1\rangle} - F_{Y^*\langle 0,0\rangle} = \underbrace{F_{Y^*\langle 1,1\rangle} - F_{Y^*\langle 0,1\rangle}}_{\text{wage structure}} + \underbrace{F_{Y^*\langle 0,1\rangle} - F_{Y^*\langle 0,0\rangle}}_{\text{worker characteristics}}, \tag{3.5.1}$$

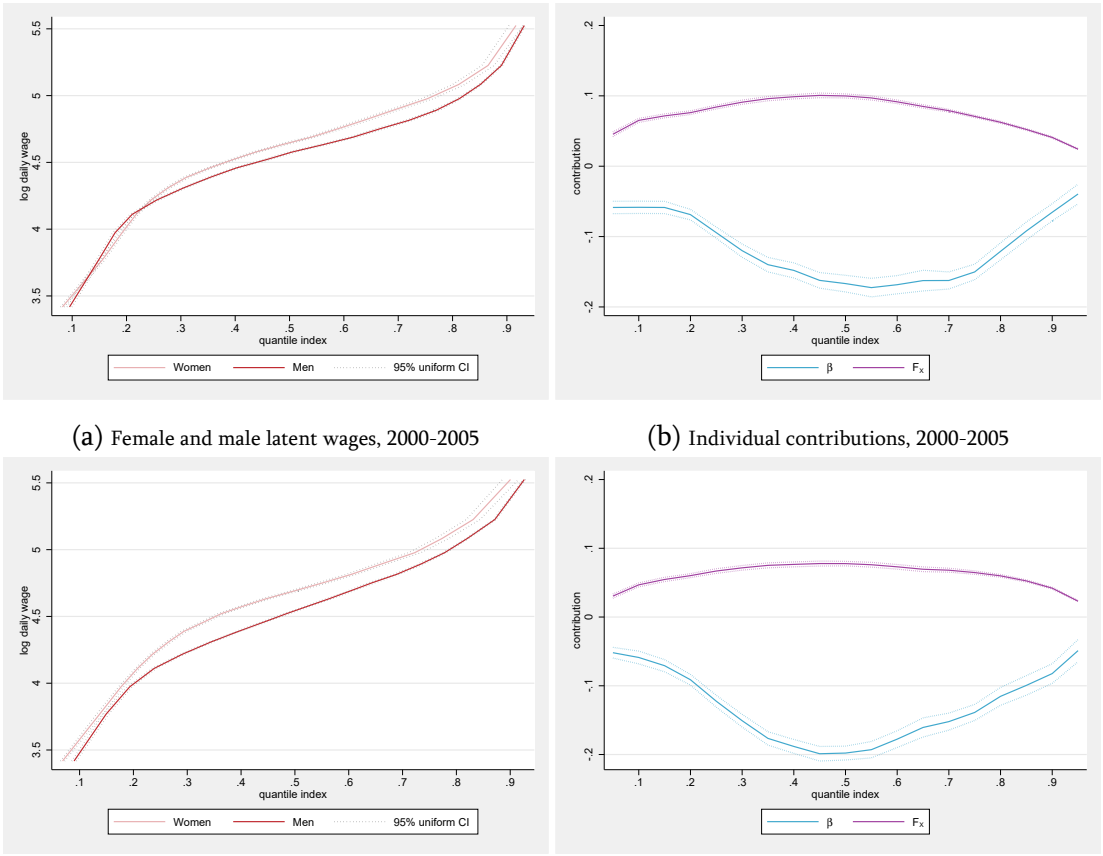
where

$$F_{Y^*\langle j,k\rangle}(y) = \int \Phi\big(-x\beta_j(y)\big)\,\mathrm{d}F_{X_k}(x), \tag{3.5.2}$$

where $\beta_j(y)$ denotes the coefficients in the wage equation for group $j \in \{0,1\}$, and F_{X_k} stands for the distribution of characteristics of group $k \in \{0,1\}$. Note that our conclusions are unaffected by the ordering of the decomposition.

Perhaps surprisingly, the latent full-time wage distribution of women is slightly above that of men as illustrated in the left panel of Figure 3.4). The contributions to the difference between the latent CDF of men and women, as defined in (3.5.1), are shown in the right panels of Figure 3.4. In order to facilitate interpretation, we switch the signs of the contributions in the graphs, so that a positive contribution means that the wages of group 1 (women) are lifted upwards, i.e. their CDF is pulled downwards, if a given decomposition factor is replaced by its group 0 (male) counterpart. The right-hand panels of Figure 3.4 show that there are two countervailing effects. Applying male wage returns to women pulls their wage distribution downwards, while applying the more favourable distribution of male observables shifts them upwards. The observation that returns to characteristics are less positive for men than for women may be surprising. Recall however, that these returns are corrected for selectivity. As described above, selectivity on unobservables is distinctively negative for women, as shown in Figure 3.2b. This downward biases estimated wage returns when sample selection is not corrected for. An interpretation of this finding is that, if institutional regulations limit observable pay differences between two worker groups, as represented by returns uncorrected for selectivity, this may mask higher actual returns for the more negatively selected worker group.

We now turn to differences in observed full-time wages between men and women. Note that this analysis differs from the previous one in that it only refers to men and women



(a) Female and male latent wages, 2000-2005

(b) Individual contributions, 2000-2005

(c) Female and male latent wages, 2012-2017

(d) Individual contributions, 2012-2017

Figure 3.4: Decomposition of differences in the latent full-time log daily wage distributions of men and women with 95% uniform confidence bands

Note: In the left panels male and female latent full-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panels, the differences between male and female latent distributions are decomposed into differences due to wage structure $\beta(y)$ and worker characteristics F_X.

observed in full-time employment, whereas the previous analysis referred to the full population of all men and women. Similar to Chernozhukov et al. (2023), we decompose these differences into five components implied by the distributional selection model: (i) sorting on unobservables, (ii) differences in employment structure, (iii) differences in the wage structure, (iv) differences in observables, and (v) differences in labour market dynamics.[9] The

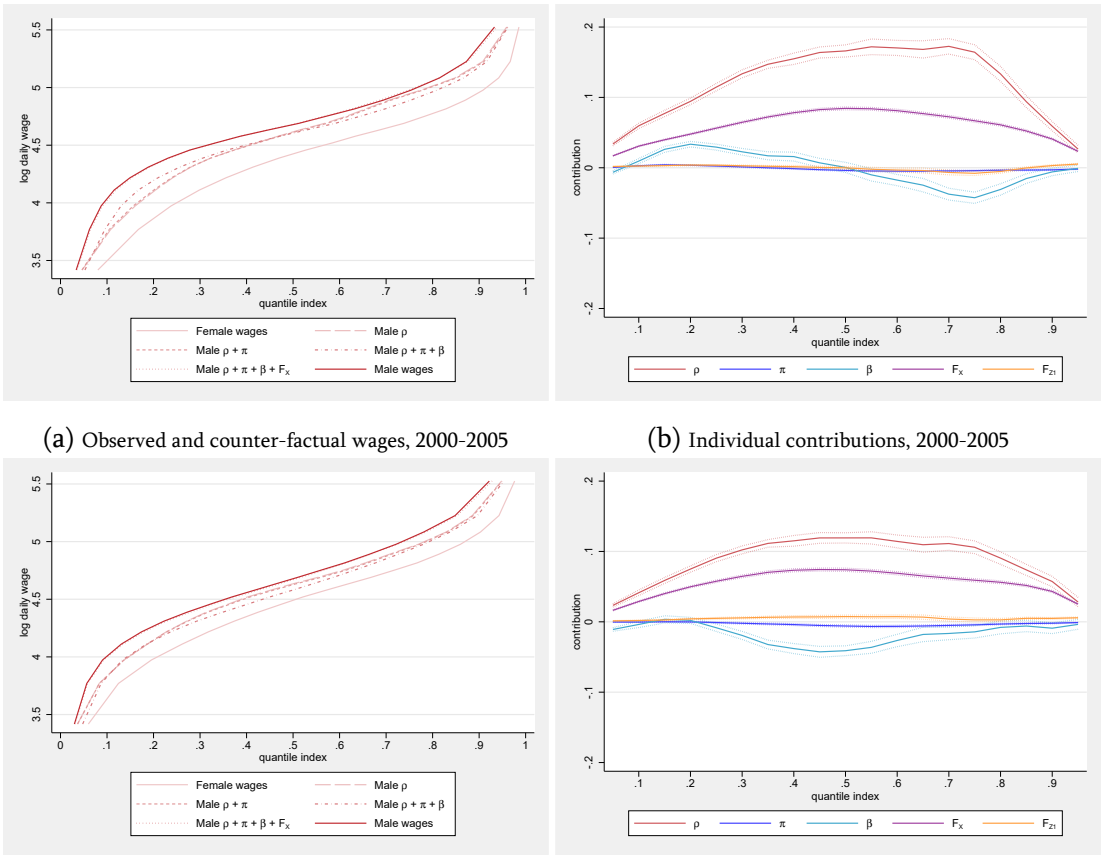
[9]Compared to Chernozhukov et al. (2023), we add component (v), as we can readily swap the values for the labour market instruments between the genders.

decomposition is given by

$$\begin{aligned} F_{Y\langle 1,1,1,1,1\rangle} - F_{Y\langle 0,0,0,0,0\rangle} \;=\; & \underbrace{F_{Y\langle 1,1,1,1,1\rangle} - F_{Y\langle 0,1,1,1,1\rangle}}_{\text{selection sorting } \rho(X\delta(y))} + \underbrace{F_{Y\langle 0,1,1,1,1\rangle} - F_{Y\langle 0,0,1,1,1\rangle}}_{\text{selection structure } \pi} \\ & + \underbrace{F_{Y\langle 0,0,1,1,1\rangle} - F_{Y\langle 0,0,0,1,1\rangle}}_{\text{wage structure } \beta(y)} + \underbrace{F_{Y\langle 0,0,0,1,1\rangle} - F_{Y\langle 0,0,0,0,1\rangle}}_{\text{worker characteristics } F_X} \\ & + \underbrace{F_{Y\langle 0,0,0,0,1\rangle} - F_{Y\langle 0,0,0,0,0\rangle}}_{\text{labour market dynamics } F_{Z_1}}, \end{aligned} \tag{3.5.3}$$

where

$$F_{Y\langle t,s,r,k,l\rangle}(y) = \frac{\int \Phi_2\big(-x\beta_s(y), z\pi_s; -\rho(x\delta_t(y))\big)\,\mathrm{d}F_{Z_{1l},X_k}(z_1,x)}{\int \Phi(z\pi_s)\,\mathrm{d}F_{Z_{1l},X_k}(z_1,x)}. \tag{3.5.4}$$



(a) Observed and counter-factual wages, 2000-2005

(b) Individual contributions, 2000-2005

(c) Observed and counter-factual wages, 2012-2017

(d) Individual contributions, 2012-2017

Figure 3.5: Detailed decomposition of differences in the observed full-time log daily wage distributions of men and women with 95% uniform confidence bands

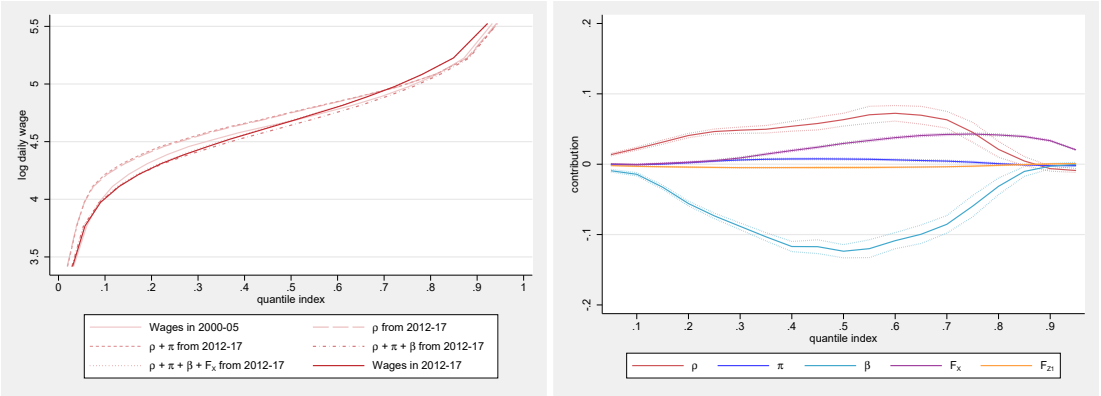
Note: In the left panels male and female observed and counter-factual full-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panels, the differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.

A common finding is that observed male wages are higher than those for females. This is also confirmed in our case, as shown by the bold and light red lines in Figure 3.5. Decom-

position (3.5.3) starts with the female full-time wage distribution $F_{Y\langle 1,1,1,1,1\rangle}$ and assigns the much less negative sorting behaviour of men to women. This shifts the wage distribution upwards, as shown in Figure 3.5a. Again, we represent this as a positive contribution in the right panel figure 3.5b. In the next decomposition step, we additionally apply the male selection structure effects π. This does not change the distribution in a meaningful way. Next, we assign the male wage structure, which raises female wages in the lower half but dampens them in the upper half of the distribution (blue line in Figure 3.5b). In Figure 3.5d for the period 2012 to 2017, this contribution is mostly negative, i.e. favourable for females. Also adding the male observables distribution leads to a significant upward shift of wages, which is not surprising given that male full-time employees have better observables than their female counterparts. Finally, switching female labour market dynamics, as represented by the instruments, to those of men does not lead to any noticeable changes in Figure 3.5a for the period 2000 to 2005. However, there is a slightly positive contribution for 2012 to 2017 owing to the fact that the development of full-time employment was more favourable for men than for women over this period.

Taken together, our decomposition suggests that a significant portion of the observed gap in full-time wages between men and women can be attributed to the less negative unobserved selectivity of men compared to women. Additionally, a substantial part of the gap arises from men in full-time employment having better observable characteristics than women. When comparing the periods from 2000-2005 and 2012-2017, the wage gap between men and women has substantially narrowed. As shown in the right panels of Figure 3.5, this reduction is largely driven by declining differences in unobserved selectivity. Women's selectivity became less negative, while men's selectivity turned more negative, as shown in Figure 3.2. Another, smaller contribution comes from returns to observable characteristics, which became more favourable for women than for men in 2012-2017 compared to 2000-2005 (blue lines in Figures 3.5b and 3.5d).

Finally, we use decomposition formula (3.5.3) to examine changes in full-time wages over time. For this, we estimate our selection model separately for the periods 2000-2005 and 2012-2017 and by gender. Figure 3.6 displays the results for men, where the left panel successively moves from the factual distribution in 2000-2005 (group 1) to that of 2012-2017 (group 0). It shows that the lower half of the 2012-2017 distribution of real full-time wages lies below that of 2000-2005, but that the upper half lies slightly above it. This is in line with earlier contributions showing that the development after 2005 was characterised by stagnation or decline of real wages in the lower half of the distribution; see Dustmann et al. (2014), Baumgarten et al. (2020), and Biewen and Sturm (2022). Figure 3.6b suggests that this was the result of a complex mix of different effects: declining real wage returns in the middle of the distribution, improved selection in the lower and middle part of the distribution, and



(a) Male observed and counter-factual wages (b) Individual contributions

Figure 3.6: Detailed decomposition of the time difference between the observed full-time log daily wage distributions of men for periods 2000-2005 and 2012-2017 with 95% uniform confidence bands

Note: In the left panel male observed full-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panel, the time differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.

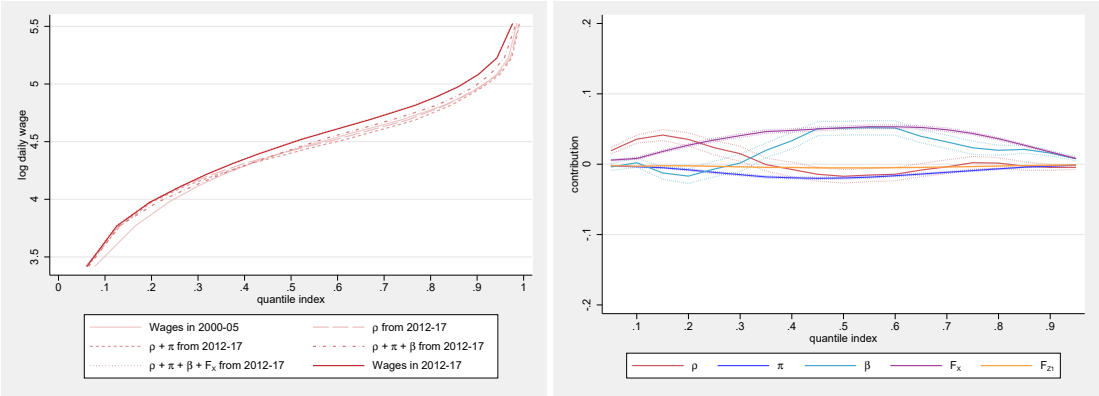
better observables at the top of the distribution. The strong pattern of declining real returns is consistent with evidence in Dustmann et al. (2014), who concluded that wage setting in Germany became significantly more flexible after 2005.

Unlike men, women experienced moderate real wage gains across all parts of the distribution from 2000-2005 to 2012-2017, as illustrated in Figure 3.7a. Figure 3.7b shows that this is the result of more favourable observables, increased real wage returns in the middle of the distribution, and, to a smaller extent, less negative selection at the lower end of the distribution. The strongest effect comes from improved observables, in particular from a higher proportion of women with a university degree and higher levels of full-time work experience.

3.5.2. Selection into Part-time Employment

We repeat the analysis for part-time wages. Recall from Section 3.4 that we model selection into part-time employment for the residual population of men and women who do not work in regular full-time. For these individuals, we consider participation in part-time employment compared to all other labour market states.

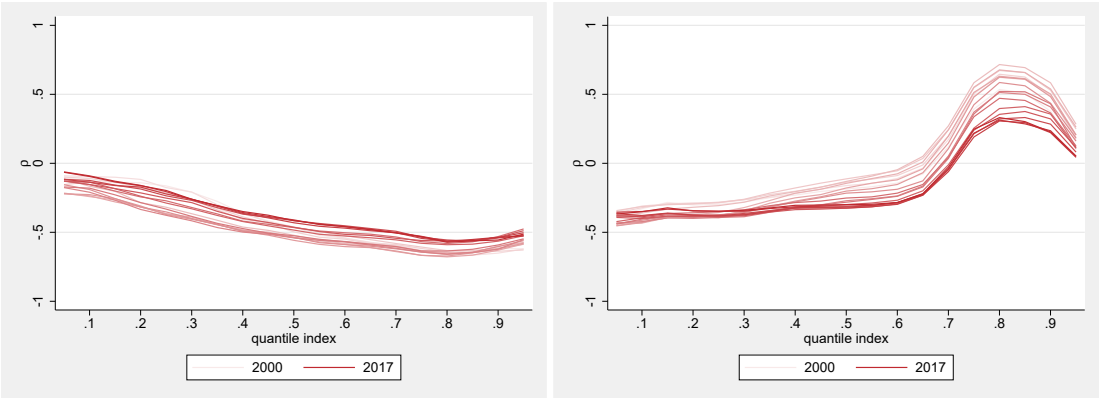
For men, Figure 3.8a suggests negative unobserved selectivity in part-time work, which



(a) Female observed and counter-factual wages (b) Individual contributions

Figure 3.7: Detailed decomposition of the time differences between the observed full-time log daily wage distributions of women for periods 2000-2005 and 2012-2017 with 95% uniform confidence bands

Note: In the left panel female observed full-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panel, the time differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.



(a) Men (b) Women

Figure 3.8: Correlation coefficient process $\rho(x\hat{\delta}(y))$ of selection sorting into part-time work at increasing quantiles of the log daily wage distribution

Note: These graphs show the estimated process of the selection sorting coefficients into part-time employment of the year indicators ranging from 2000 to 2017 over increasing quantile indices of the log daily wage distribution for men in the left and for women in the right panel.

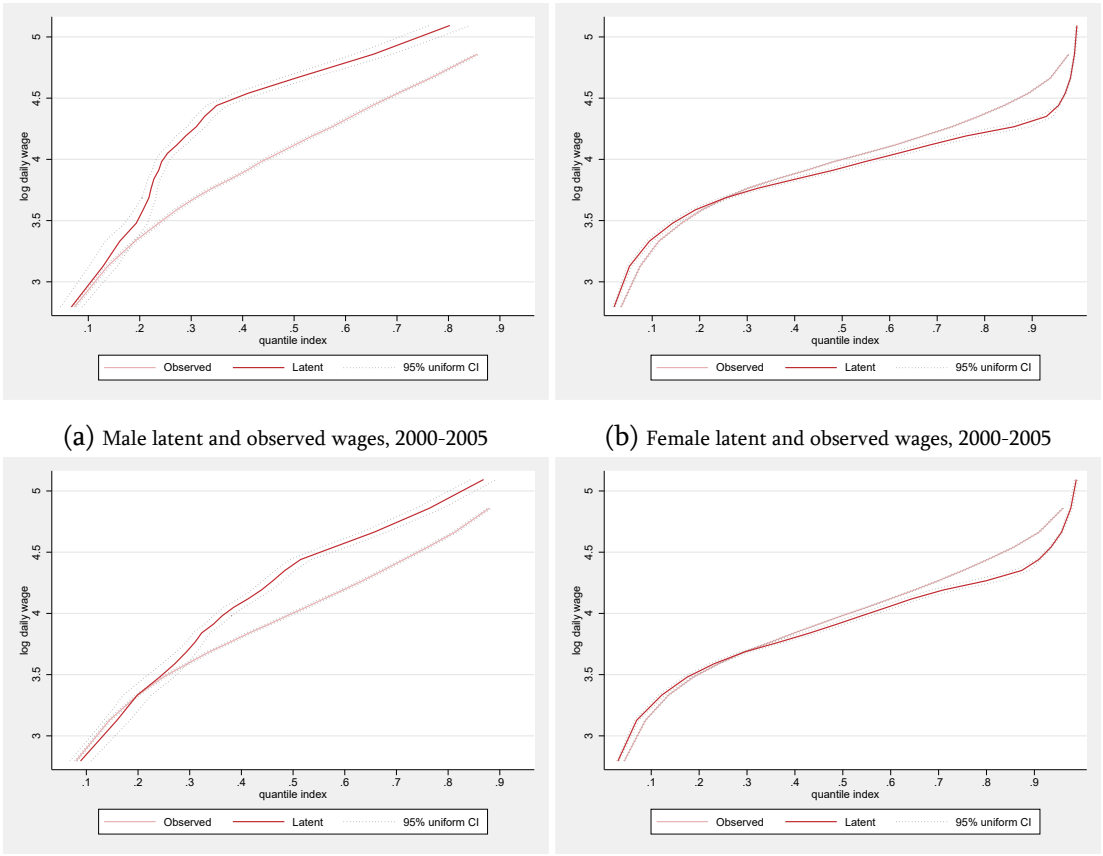
became less negative towards more recent years. The phenomenon of negative selection into part-time employment can be explained by the fact that the residual population of men who do not work in regular full-time employment comprises a significant proportion of individuals who at some point moved into self-employment, civil service or work abroad. Unfortunately, we cannot differentiate these cases with our data set. Compared to this group, men

in part-time employment may be a negative selection with respect to unobservables, especially for higher levels of pay. Out of the residual group of men not working full-time, only a small fraction actually chose to take up regular part-time employment (see Table C.3.2). However, this fraction significantly increased over time. Both in 2000-2005 and 2012-2017, around 1.3 million men in our sample did not work full-time. Of these, around 150 thousand worked part-time in 2012-2017, compared to around 75 thousand in 2000-2005. The fact that part-time work became more common among men may have also contributed to the decline in negative selectivity as shown in Figure 3.8a.

The corresponding Figure 3.8b for women reveals a complex sorting pattern into part-time work with a change of sign between the .6 and .7 quantile indices. A potential explanation for the observed pattern may again be related to assortative matching. In the lower to middle part of the distribution, even women with less favourable unobservables would often be forced to contribute to household income because their partners have low income. On the other hand, women with more favourable unobservables may generally be disincentivised by the structure of the German tax system which favours large earnings differentials between partners. These disincentives are particularly strong in the upper part of the distribution, so that only women with the highest wage offers decide to take up additional part-time employment; see Bick and Fuchs-Schündeln (2017). This pattern of positive selection at the top weakened in later years, possibly because female part-time work became increasingly common, as shown in Figure 3.1, or because changes in social norms and public child care made employment after child care more acceptable.

Figure 3.9 compares latent and observed part-time wage distributions for men and women. For men, the pronounced negative selection on unobservables, as shown in Figure 3.8a, clearly dominates the somewhat positive selection on observables so that latent wages lie substantially above observed part-time wages (see left panel of Figure 3.9). For women, negative selection on unobservables balances with positive selection on observables up to around the 70th percentile, above which positive selectivity on unobservables pushes observed wages above latent wages (see right panel of Figure 3.9). For both men and women, differences between latent and observed part-time wage distributions narrowed in recent years, driven by differences in unobserved selectivity over time (see Figure 3.8).

Figure 3.10 presents the gender gap in latent part-time wage distributions. Again, these distributions are constructed by assigning all individuals out of the residual group of men and women not working in regular full-time the part-time wages of the respective group. This leads to very large differences between men and women. As the right-hand side of Figure 3.10 shows, the large differences are mainly due to the much better wage returns for men in part-time work compared to women, while differences in the distribution of observables



(a) Male latent and observed wages, 2000-2005

(b) Female latent and observed wages, 2000-2005

(c) Male latent and observed wages, 2012-2017

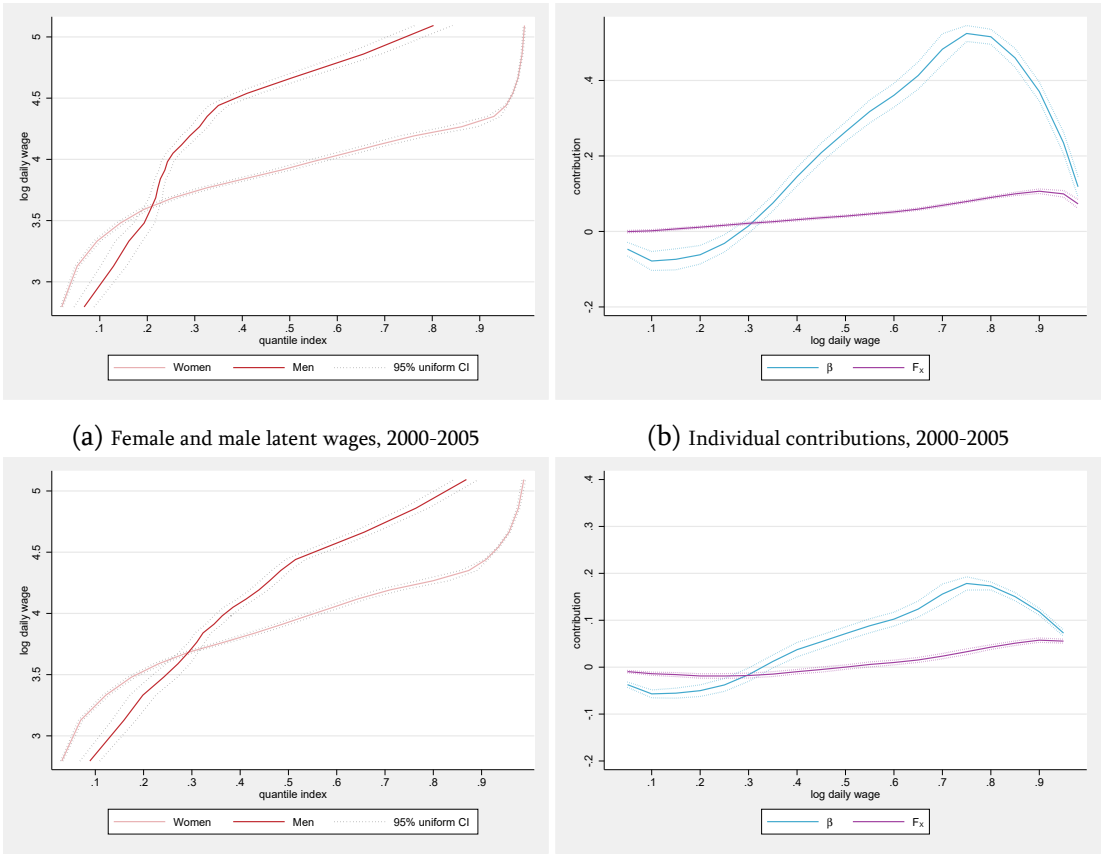
(d) Female latent and observed wages, 2012-2017

Figure 3.9: Latent and observed part-time log daily wages with 95% uniform confidence band

Note: These graphs show male and female part-time wage distributions in the left and right column, respectively. The period of 2000-2005 is illustrated in the top row and the period of 2012-2017 in the bottom row.

play a minor role. A potential explanation for this very pronounced wage structure effect is that our set of observables contains important characteristics such as age, education and experience but lacks more detailed characteristics of male and female part-time job profiles. It may be the case that the relatively small number of part-time men are employed in roles that require higher qualifications, as opposed to part-time women whose range of part-time employments tends to be much wider. The bottom panel of Figure 3.10 again shows signs of convergence between male and female distributions, driven by a convergence in wage returns (see blue lines in the right column of Figures 3.10). This convergence is a likely consequence of part-time work becoming more common among men, thereby expanding the range of job types typically available to men in part-time roles.

Figure 3.11 examines differences in observed part-time wage distributions between men and women, as given by decomposition (3.5.3). Male part-time wages dominate female part-time wages in the upper part of the distribution, while the reverse is true in the lower part of



(a) Female and male latent wages, 2000-2005

(b) Individual contributions, 2000-2005

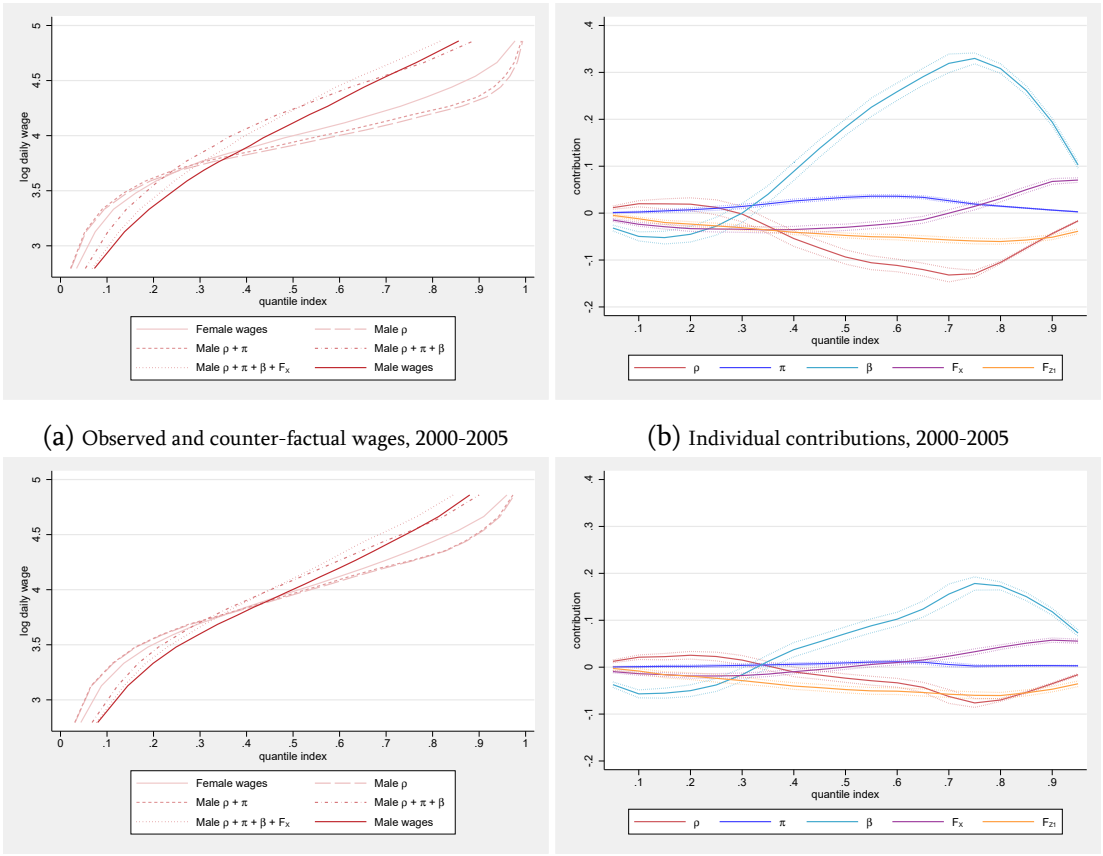
(c) Female and male latent wages, 2012-2017

(d) Individual contributions, 2012-2017

Figure 3.10: Decomposition of differences in the latent part-time log daily wage distributions of men and women with 95% uniform confidence bands

Note: In the left panels male and female latent part-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panels, the differences between male and female latent distributions are decomposed into differences due to wage structure $\beta(y)$ and worker characteristics F_X.

the distribution. Again, it turns out that part-time men in the upper half of the distribution enjoy much higher wage returns than part-time women (see blue lines in figures 3.11b and 3.11d), which is the likely result of better job characteristics not controlled for in our set of observables. The effect of higher male wage returns is partly counteracted by more positive female selection on unobservables (see red lines in Figures 3.11b and 3.11d). Observed male part-time wages only dominate female part-time wages above the 40th percentile, as shown in the left column of Figure 3.11. In the lower part of the distribution, women have better observables and face slightly better wage returns than men (see blue and purple lines in Figures 3.11b and 3.11d). Differences in labour market dynamics between men and women also play a small role, boosting observed part-time wages of women (see orange line in Figures 3.11b and 3.11d). The comparison between the upper and the bottom row of Figure 3.11 again indicates convergence between male and female part-time wage distributions over time, mostly due to diminishing differences in wage returns (see blue lines in right column)



(a) Observed and counter-factual wages, 2000-2005

(b) Individual contributions, 2000-2005

(c) Observed and counter-factual wages, 2012-2017

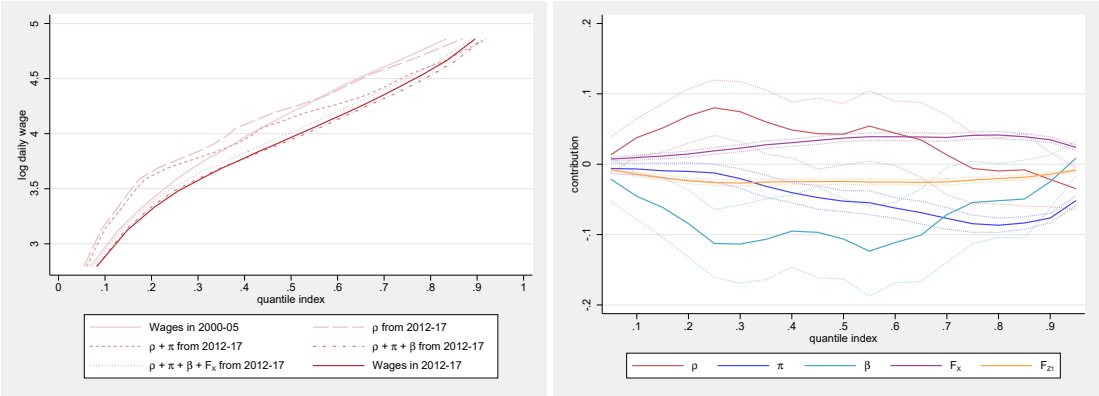
(d) Individual contributions, 2012-2017

Figure 3.11: Detailed decomposition of differences in the observed part-time log daily wage distributions of men and women with 95% uniform confidence bands

Note: In the left panels male and female observed and counter-factual part-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panels, the differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.

and declining differences in selectivity on unobservables (see red lines in right column).

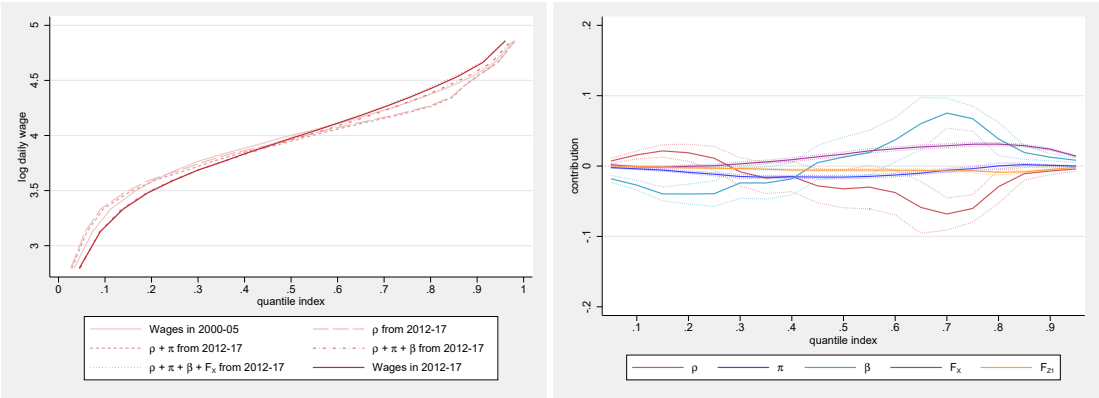
Our final analysis concerns the decomposition of the evolution in observed male and female part-time wage distributions over time, as shown in Figures 3.12 and 3.13). For men Figure 3.12 indicates a deterioration of part-time wages between 2000-2005 and 2012-2017 since the distribution of 2000-2005 lies above that of 2012-2017, as indicated by the fine and bold red lines in Figure 3.12. The right panel of Figure 3.12 suggests that, although there were improvements in selection on unobservables and selection on observables (see red and purple lines in Figure 3.12b), these were more than offset by a decline in part-time wage returns (see blue line) and a selection structure effect (see dark blue line in Figure 3.12b). This deterioration in part-time wage returns is in line with the above conjecture that the proportionally massive expansion of male part-time employment between 2000-2005 and



(a) Male observed and counter-factual wages (b) Individual contributions

Figure 3.12: Detailed decomposition of the time difference between the observed part-time log daily wage distributions of men for periods 2000-2005 and 2012-2017 with 95% uniform confidence bands

Note: In the left panel male observed part-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panel, the time differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.



(a) Female observed and counter-factual wages (b) Individual contributions

Figure 3.13: Detailed decomposition of the time differences between the observed part-time log daily wage distributions of women for periods 2000-2005 and 2012-2017 with 95% uniform confidence bands

Note: In the left panel female observed part-time wage distributions are compared for the periods 2000-2005 and 2012-2017. In the right panel, the time differences between both observed distributions are decomposed into differences due to selection sorting $\rho(X\delta(y))$, selection structure π, wage structure $\beta(y)$, worker characteristics F_X, and labour market dynamics F_{Z_1}.

2012-2017 involved a change in job profiles not captured by our observables. Furthermore, the selection structure effect, as depicted in the dark blue lines, indicates that male selection into part-time employment underwent a transformation whereby observable characteristics,

specifically age, education, and experience, were increasingly matched with less favourable part-time roles, particularly at the upper end of the part-time distribution. The much lower statistical precision of the point estimates in Figure 3.12b reflects the relatively small sample size of this group of workers (see Table C.3.2).

The analysis of female part-time wages over time is given in Figure 3.13. Apart from some gains above the 60th percentile and some losses below, the female part-time distribution remained relatively stable between 2000-2005 and 2012-2017. Figure 3.13b suggests that the gains were the result of improved wage returns to observed characteristics (see blue line), which overcompensated the decline in positive selection on unobservables (see red line). Better observable characteristics, in particular, more work experience and a much higher proportion of workers with university degrees (see Table C.3.2), also contributed to higher female part-time wages in 2012-2017 compared to 2000-2005 (see purple line in Figure 3.13). The corresponding pattern is strikingly similar to that observed for males in Figure 3.12, which suggests a general upgrading of part-time jobs in Germany. At the lower end of the female part-time wage distribution, there was an erosion of wage returns (see blue line), which was not compensated by slight improvements in unobserved selectivity (see red line in Figure 3.13b).

3.6. CONCLUSIONS

Based on high-quality administrative data, this paper examines male and female wage distributions in Germany while flexibly accounting for selection on unobservables. Our findings suggest that male selectivity in full-time work is positive at the lower end of the distribution but becomes negative throughout the rest of it, a trend that intensified over time. This aligns with the hypothesis that the post-2012 full-time employment boom drew individuals with less favourable unobservables into the workforce. Positive selectivity at the bottom of the distribution may be explained by the generosity of Germany's social safety net, which discourages employment among those with very low wage offers. For women, full-time employment generally exhibits negative selectivity, potentially driven by assortative matching and household dynamics, where women with better unobservables may not need to contribute to household income. Recent improvements in public childcare and evolving social norms appear to have mitigated this negative selectivity by encouraging women's return to full-time work after childbirth.

A significant portion of the full-time gender wage gap is explained by differences in unobserved selectivity between men and women. Better observables of full-time men, especially work experience and education, explain another large share of the gap, while selectivity-corrected wage returns work in favour of women. Between 2000-2005 and 2012-2017, male full-time wages rose in the upper half of the distribution but declined in the lower half, reflecting wage erosion across broad segments of the distribution. We attribute this to wage restraints and increased wage flexibility; see Dustmann et al. (2014). At the top of the distribution, improved selection on unobservables and better observables contributed to higher male wages in the later period. By contrast, female full-time wages grew between 2000-2005 and 2012-2017 across the whole distribution as a result of better observables, increased wage returns in the middle and less negative selection at the lower end of the distribution. Overall, the full-time wage gap between men and women significantly narrowed between 2000-2005 and 2012-2017, largely due to declining differences in unobserved selectivity and improved observables for women.

Our paper is among the first to provide an in-depth examination of unobserved selectivity in part-time employment. We find that men working part-time represent a negatively selected subset of the group of men not pursuing regular full-time employment. This may be explained by the fact that the group of men in our data not observed in full-time employment also include the self-employed and other individuals working for pay, who are not subject to social security contributions. In the past, only very few men worked part-time in Germany, making this group of workers a highly specific one. Male selectivity in part-time work became less negative as part-time employment among men has grown, although it remains much less common than for women. At the lower end of the wage distribution, negative selectivity is less pronounced, possibly due to the same social safety net effects observed in full-time employment. For women, part-time work exhibits a complex selectivity pattern turning from negative at the bottom to positive at the top. Our explanation for this pattern is that assortative matching forces women in the lower part of the distribution to contribute to household income, while women in the upper part only work if they receive very high wage offers. Recent declines in female part-time selectivity may reflect improved public childcare and shifting social norms supporting post-childbirth employment.

We find that male part-time wages are higher than those of females in the upper half but lower in the bottom half. This is due to higher wage returns in male part-time jobs, likely driven by the unique job profiles of male part-time workers. At the upper end, men also possess better observables, but this is offset by positive unobserved selectivity among women. The male part-time wage distribution shifted downwards between 2000-2005 and 2012-2017, likely due to worsening wage returns and less specific job profiles. Female part-

time wages were more stable over this period, with modest improvements in wage returns but less positive selection in the upper part of the distribution. We observe a convergence of male and female part-time wage distributions, mainly explained by declining differences in wage returns and unobserved selectivity.

All told, our study offers compelling evidence for substantial heterogeneity in selectivity patterns across full-time and part-time wage distributions. This phenomenon has not been adequately acknowledged in previous research.

Chapter 4

DISSERTATION CONCLUSION

This dissertation has sought to bridge the methodological divide between modern machine learning techniques and traditional econometric approaches, with a particular focus on settings characterized by non-random sample selection and distributional heterogeneity. Encompassing both theoretical aspects and empirical applications, our contributions have identified several research areas, where classical econometrics and machine learning theory have scope for further integration.

In this context, Chapter 1 served as a valuable illustration of how machine learning techniques can be employed to uncover points of latent conditional distributions in the presence of high-dimensional control variables. Specifically, the study established asymptotic normality of an estimator measuring the effect of a target regressor of interest – such as a treatment variable or policy indicator – on the conditional mean of an outcome variable when realizations below a certain threshold remain unobserved. Starting point for this endeavour was the maximum likelihood estimator originally proposed by Tobin (1958). Tobin's seminal study is still widely regarded as a foundational piece of research, having attracted a significant level of academic attention from scholars and practitioners in various disciplines. His pioneering contribution is justly credited with having triggered further research into non-random sample selection, which has had a profound and lasting effect on the field of economics; see, e.g. Heckman (1976) and Amemiya (1984) among others. In adapting Tobin's estimator to post-regularization inference, we have built on the approaches employed by Fang et al. (2017) for the Cox model and Belloni et al. (2016a) for GLMs. More precisely, we have adjusted the general concept of "decorrelating" the score, as termed by Fang et al. (2017), also known as "Neyman orthogonalization" in the parlance of Chernozhukov et al. (2015), to immunize the likelihood score of the target parameter with respect to the high-dimensional

nuisance part of the model. In econometric terms, the methods described in Chapter 1 create an instrument for the target regressor, where the effects of high-dimensional control variables have been "partialled-out". In combination with more technical conditions, this allows us to test hypotheses about the target parameter of interest in a conventional way. Taking existing double machine learning estimators as a benchmark, Algorithms 1.2.1 and 1.2.2 proved to be as straightforward to implement and computationally robust as those for the logistic GLM in Belloni et al. (2016a). Notwithstanding the deficiencies inherent to Tobin's maximum likelihood approach – most notably its incapacity to accommodate heteroscedasticity – our work provides a foundation for future research into more sophisticated sample selection models that leverage machine learning to overcome the limitations of traditional econometric methods.

Expanding upon these theoretical findings, Chapter 2 showcased the various means by which machine learning approaches complement and enhance traditional statistical tools. Focusing on a nationwide minimum wage reform in Germany, we employed a post-double selection logistic distribution regression model to examine the heterogeneity in treatment effects across the distributions of hourly wages, working hours, and monthly earnings. In comparison to the first chapter, which has sought to explore an entirely new research trajectory, this study utilised modern machine learning techniques to revisit a pre-existing yet ongoing research question that has garnered substantial attention in recent years. In principle, the divergence of previous results in the literature can be attributed to the inherent quality of the datasets, or to different methodological choices. While the former is often given and fixed, the latter is subject to discretionary decisions made by the researchers. The implementation of machine learning, on the other hand, provides a flexible approach that curbs direct interventions by the practitioner, as the majority of steps taken by estimation algorithms are automated. This ensures that the researcher has limited opportunity to influence the outcome. Leveraging this integrity, the analysis presented in Chapter 2 has managed to successfully reconcile conflicting findings from previous articles and underscored the efficacy of machine learning techniques in strengthening empirical insights within labour economics. Particularly, the capacity to identify and extract relevant signals from datasets with a rich set of potential control variables exemplifies the relevancy of these state-of-the-art methods in addressing complex real-world scenarios. As a consequence, Chapter 2 has further substantiated a central thesis of this dissertation: integrating machine learning approaches into econometric analyses can reveal nuanced structural relationships that remain undetected by conventional methodologies.

The third study in this dissertation has examined the gender pay gap – probably the prototype of research questions in empirical economics affected by non-random sample selection bias. Indeed, as an article recently published in the British weekly newspaper "The

Economist" defiantly quipped, the observation that "women earn less than men in rich countries is so well-known it is often met with a shrug".[1] When interpreted literally, this statement suggests that personal beliefs and convictions concerning pay inequalities have become so deeply entrenched that the findings of even the most rigorous empirical research can, under public scrutiny, be reduced to oversimplified narratives. To address such remarks at a scientific level, Chapter 3 has provided novel insights by adding an additional layer to the decomposition of wage disparities in Germany. By employing the cutting-edge distribution regression model with sample selection correction developed in Chernozhukov et al. (2023), our study has revealed nuanced patterns in the heterogeneity of selectivity across both full-time and part-time wage distributions. These insights contribute to a deeper understanding of the sources and mechanism of pay inequality in Germany.

Together, these essays not only charted a new course in (high-dimensional) econometric theory but have also implemented and illustrated practical tools that capture the subtle dynamics of real-world data, thereby challenging the constraints of conventional methods. Chapters 1 and 2 have demonstrated that integrating machine learning techniques into classical econometric tasks, such as estimation and hypothesis testing, not only improves accuracy and efficiency compared to conventional methods, but often also helps to alleviate issues arising from minor violations of the underlying model assumptions. Leveraging rich datasets and highly adaptive algorithms, machine learning sidesteps the painstaking task of manually selecting models. As a consequence, it both simplifies the analysis and minimizes the risk of costly misspecification errors in model designs. Beyond its immediate applications, the integration of machine learning algorithms in econometrics broadens the scope for future research. Further exploration of high-dimensional extensions to classical models should equip practitioners from various fields with more versatile statistical tools to tackle issues ranging from non-random sample selection to heterogeneous treatment effects in increasingly complex frameworks.

[1] The interested reader is referred to the article entitled "No longer narrowing. The pay gap between men and women won't go away", published online in the The Economist's Business Section on 6 March 2025.

Chapter A

MATHEMATICAL PROOFS TO CHAPTER 1: ``POST-SELECTION INFERENCE IN TOBIT MODELS WITH HIGH-DIMENSIONAL CONTROL VARIABLES´´

The proofs in this section are based on the proofs in Belloni et al. (2016b, 2019) for GLMs and quantile regression. We extend their steps to allow for the additional instrument and the inverse scale parameter associated with the censored outcome variable.

A.1. PROOF OF THEOREM 1.4.1

Let $y \in \mathcal{Y}$ and $(d, x) \in \mathcal{D} \times \mathcal{X}$. For notational convenience we define $m := (y, d, x)$ and $\tilde{h} = (\tilde{\beta}, \tilde{\gamma}, \tilde{z}, \tilde{q}, \tilde{\mu})$, where $\tilde{z} : \mathcal{D} \times \mathcal{X} \to \mathbb{R} : (d, x) \mapsto \tilde{z}(d, x)$ and $\tilde{q} : \mathcal{Y} \times \mathcal{D} \times \mathcal{X} \to \mathbb{R} : (y, d, x) \mapsto \tilde{q}(y, d, x)$ are functions, for which we introduce the shorthand notations $\tilde{z}_i = \tilde{z}(d_i, x_i)$ and $\tilde{q}_i = \tilde{q}(y_i, d_i, x_i)$, respectively.

Our estimating function for observation i evaluated at $\tilde{\alpha}$, $\tilde{h}$ is

$$\begin{aligned} \psi_{\tilde{\alpha},\tilde{h}}(m_i) &= \psi_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{z},\tilde{q},\tilde{\mu}}(m_i) = g(\tilde{\gamma} y_i - \tilde{\alpha} d_i - x_i\tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i) - \tilde{\mu} s_i/\tilde{\gamma} \\ &= \{(1 - s_i) g_1(-\tilde{\alpha} d_i - x_i\tilde{\beta}) + s_i g_2(\tilde{\gamma} y_i - \tilde{\alpha} d_i - x_i\tilde{\beta})\}(\tilde{z}_i - \tilde{\mu}\tilde{q}_i) - \tilde{\mu} s_i/\tilde{\gamma}. \end{aligned} \tag{A.1.1}$$

Note that for $h_0 = (\beta_0, \gamma_0, z_0, q_0, \mu_0)$, we have $\mathrm{E}[\psi_{\alpha_0,h_0}(m_i)] = 0$ because of (1.2.5). For fixed $\tilde{\alpha} \in \mathbb{R}$, $\tilde{\beta} \in \mathbb{R}^p$, $\tilde{\gamma} \in \mathbb{R}^+$, $\tilde{z}_i$, $\tilde{q}_i$, and $\tilde{\mu} \in \mathbb{R}$ we define

$$\Gamma(\tilde{\alpha}, \tilde{h}) := \overline{\mathrm{E}}[\psi_{\tilde{\alpha},\tilde{h}}(m_i)], \tag{A.1.2}$$

$$\Gamma_1(\tilde{\alpha}, \tilde{h}) := \partial_\alpha \, \overline{\mathrm{E}}[\psi_{\alpha,\tilde{h}}(m_i)]\Big|_{\alpha=\tilde{\alpha}}, \tag{A.1.3}$$

$$\Gamma_{1,2}(\tilde{\alpha}, \tilde{h}) := \partial_\alpha \, \partial_\alpha \, \overline{\mathrm{E}}[\psi_{\alpha,\tilde{h}}(m_i)]\Big|_{\alpha=\tilde{\alpha}}. \tag{A.1.4}$$

The directional derivative with respect to $[\hat{h} - h_0]$ evaluated at $(\tilde{\alpha}, \tilde{h})$ is given by

$$\Gamma_2(\tilde{\alpha}, \tilde{h})[\hat{h} - h_0] = \lim_{t \to 0} \frac{\Gamma(\tilde{\alpha}, \tilde{h} + t[\hat{h} - h_0]) - \Gamma(\tilde{\alpha}, \tilde{h})}{t}.$$

Moreover, we write $\Gamma_{2,2}(\tilde{\alpha}, \tilde{h})[\hat{h} - h_0, \hat{h} - h_0]$ to denote the second directional derivative evaluated at $(\tilde{\alpha}, \tilde{h})$.

Following the definition in (1.2.9), the true weights w_i are

$$\begin{aligned} w_i &= -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0) \\ &= (s_i - 1)g_1'(-\alpha_0 d_i - x_i\beta_0) - s_i g_2'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0). \end{aligned} \tag{A.1.5}$$

In the following steps, we also need the orthogonalization parameter

$$\mu_0 = \frac{\overline{\mathrm{E}}[w_i z_{0i} y_i]}{\overline{\mathrm{E}}[w_i q_{0i} y_i + s_i/\gamma_0^2]}, \tag{A.1.6}$$

which is bounded by $|\mu_0| \leqslant \underline{c}^{-1} |\overline{\mathrm{E}}[w_i z_{0i} y_i]| \lesssim \overline{\mathrm{E}}[z_{0i}^2]^{1/2} \overline{\mathrm{E}}[y_i^2]^{1/2}$, because $|\overline{\mathrm{E}}[w_i q_{0i} y_i + s_i/\gamma_0^2]| \geqslant \underline{c} > 0$ by Condition ITob 4.1–(ii).

In Steps 1 and 2 below we assume that the estimated parameters $\check{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{\mu}$ and the estimated functions $\hat{z}$, $\hat{q}$ satisfy Condition ITob 4.1–(i-iii). In Step 3 we additionally need Condition ITob 4.1–(iv).

Step 1: (Main Step for Normality of $\mathbb{E}_n[\psi_{\alpha_0,h_0}(m_i)]$)

Observe that by definition of $\mathbb{G}_n(f(m_i)) := \sqrt{n}\,\mathbb{E}_n\big[f(m_i) - \mathrm{E}[f(m_i)]\big]$, we have

$$\begin{aligned} 0 &= \mathbb{G}_n(\psi_{\check{\alpha},\hat{h}}(m_i) - \psi_{\alpha_0,h_0}(m_i)) + \mathbb{G}_n(\psi_{\check{\alpha},h_0}(m_i) - \psi_{\check{\alpha},\hat{h}}(m_i)) \\ &\quad + \mathbb{G}_n(\psi_{\alpha_0,h_0}(m_i) - \psi_{\check{\alpha},h_0}(m_i)) \qquad (A.1.7) \\ \underbrace{\mathbb{E}_n[\psi_{\alpha_0,h_0}(m_i)]}_{:=(\emptyset)} &= \underbrace{\mathbb{E}_n[\psi_{\check{\alpha},\hat{h}}(m_i)]}_{:=(I)} - \underbrace{\Gamma(\check{\alpha},\hat{h})}_{:=(II)} + \underbrace{\overline{\mathrm{E}}[\psi_{\alpha_0,h_0}(m_i)]}_{=0} \\ &\quad + \underbrace{\mathbb{G}_n(\psi_{\check{\alpha},h_0}(m_i) - \psi_{\check{\alpha},\hat{h}}(m_i))/\sqrt{n}}_{:=(III)} \\ &\quad + \underbrace{\mathbb{G}_n(\psi_{\alpha_0,h_0}(m_i) - \psi_{\check{\alpha},h_0}(m_i))/\sqrt{n}}_{:=(IV)} \end{aligned}$$

(a) By Condition ITob 4.1–(iii), (1.4.3), with probability at least $1 - \Delta_n$ we have $|(I)| \lesssim \delta_n/\sqrt{n}$.

(b) By equation (A.1.14) in Step 2 below, we have

$$(II) = \overline{\mathrm{E}}[w_i z_{0i} d_i](\check{\alpha} - \alpha_0) + O_{\mathrm{P}}\big(\delta_n|\check{\alpha} - \alpha_0| + \delta_n/\sqrt{n}\big).$$

(c) By Condition ITob 4.1–(iii), (1.4.2), with probability at least $1 - \Delta_n$ we have $|(III)| \lesssim \delta_n/\sqrt{n}$.

(d) To control $|(IV)|$, note that since

$$\begin{aligned} \big|\psi_{\alpha,h_0}(m_i) - \psi_{\alpha_0,h_0}(m_i)\big| &= \big|\{g(\gamma_0 y_i - \alpha d_i - x_i\beta_0) \\ &\qquad - g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}(z_{0i} - \mu_0 q_{0i})\big| \\ &\leqslant \bar{L}' \cdot |\alpha - \alpha_0| \cdot |d_i(z_{0i} - \mu_0 q_{0i})|, \end{aligned}$$

we have

$$\overline{\mathrm{E}}[\{\psi_{\alpha,h_0}(m_i) - \psi_{\alpha_0,h_0}(m_i)\}^2] \leqslant (\bar{L}' \cdot |\alpha - \alpha_0|)^2\, \overline{\mathrm{E}}[d_i^2(z_{0i} - \mu_0 q_{0i})^2].$$

Then, because of $|\check{\alpha} - \alpha_0| \leqslant \delta_n$ in Condition ITob 4.1–(iii), (1.4.3), we have

$$\sup_{|\alpha - \alpha_0| \leqslant \delta_n} \mathbb{E}_n[\{\psi_{\alpha,h_0}(m_i) - \psi_{\alpha_0,h_0}(m_i)\}^2] \leqslant (\bar{L}'\delta_n)^2\, \mathbb{E}_n[d_i^2(z_{0i} - \mu_0 q_{0i})^2].$$

Applying Markov's inequality for some $\varepsilon > 0$, we get

$$\mathrm{P}\big(\mathbb{E}_n[d_i^2(z_{0i} - \mu_0 q_{0i})^2] \geqslant \overline{\mathrm{E}}[d_i^2(z_{0i} - \mu_0 q_{0i})^2]/\varepsilon\big) \leqslant \varepsilon,$$

and therefore

$$(\bar{L}'\delta_n)^2\, \mathbb{E}_n[d_i^2(z_{0i} - \mu_0 q_{0i})^2] \lesssim_{\mathrm{P}} (\bar{L}'\delta_n)^2\, \overline{\mathrm{E}}[d_i^2(z_{0i} - \mu_0 q_{0i})^2].$$

Using Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b) with $W_i^2 = d_i^2(z_{0i} - \mu_0 q_{0i})^2$ and $\mathcal{T} = \{(\alpha - \alpha_0) \in \mathbb{R} : |\alpha - \alpha_0| \leqslant \delta_n\}$, we have

$$\begin{aligned} |(IV)| &\lesssim_{\mathrm{P}} \sup_{|\alpha-\alpha_0|\leqslant\delta_n} \left|\mathbb{G}_n(\psi_{\alpha_0,h_0}(m_i) - \psi_{\alpha,h_0}(m_i))/\sqrt{n}\right| \\ &\lesssim_{\mathrm{P}} \sup_{|\alpha-\alpha_0|\leqslant\delta_n} |\alpha - \alpha_0|\overline{\mathrm{E}}[d_i^2(z_{0i} - \mu_0 q_{0i})^2]^{1/2}/\sqrt{n} \lesssim \delta_n/\sqrt{n}. \end{aligned} \tag{A.1.8}$$

(e) Combining our results for (I), (II), (III), and (IV), we obtain

$$(\emptyset) = \overline{\mathrm{E}}[w_i z_{0i} d_i](\check{\alpha} - \alpha_0) + O_{\mathrm{P}}(\delta_n/\sqrt{n}) + O_{\mathrm{P}}(\delta_n)|\check{\alpha} - \alpha_0|.$$

Since $|\overline{\mathrm{E}}[w_i z_{0i} d_i]| \geqslant c > 0$ is bounded away from zero by Condition ITob 4.1–(ii), we have

$$\overline{\mathrm{E}}[w_i z_{0i} d_i](\check{\alpha} - \alpha_0) = \mathbb{E}_n[\psi_{\alpha_0,h_0}(m_i)] + O_{\mathrm{P}}(\delta_n/\sqrt{n}) + O_{\mathrm{P}}(\delta_n)|\check{\alpha} - \alpha_0|,$$

which verifies the first claim of Theorem 1.4.1. Moreover, note that $\overline{\mathrm{E}}[\psi_{\alpha_0,h_0}(m_i)] = 0$ and

$$\begin{aligned} \overline{\mathrm{E}}[|\psi_{\alpha_0,h_0}(m_i)|^3] &= \overline{\mathrm{E}}[|g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(z_{0i} - \mu_0 q_{0i}) - \mu_0 s_i/\gamma_0|^3] \\ &\leqslant \overline{\mathrm{E}}[|g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(z_{0i} - \mu_0 q_{0i})|^3] \\ &\quad + 3\,\overline{\mathrm{E}}[|g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(z_{0i} - \mu_0 q_{0i})|^2|\mu_0 s_i/\gamma_0|] \\ &\quad + 3\,\overline{\mathrm{E}}[|g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(z_{0i} - \mu_0 q_{0i})||\mu_0 s_i/\gamma_0|^2] \\ &\quad + \overline{\mathrm{E}}[|\mu_0 s_i/\gamma_0|^3] \\ &\leqslant \bar{L}^3\big\{\overline{\mathrm{E}}[|z_{0i}|^3] + 3\,|\mu_0|\,\overline{\mathrm{E}}[z_{0i}^4]^{1/2}\overline{\mathrm{E}}[q_{0i}^2]^{1/2} \\ &\qquad + 3\,\mu_0^2\,\overline{\mathrm{E}}[z_{0i}^2]^{1/2}\overline{\mathrm{E}}[q_{0i}^4]^{1/2} + |\mu_0|^3\,\overline{\mathrm{E}}[|q_{0i}|^3]\big\} \\ &\quad + 3\,|\mu_0|/\gamma_0\;\bar{L}^2\big\{\overline{\mathrm{E}}[z_{0i}^2] + 2\,|\mu_0|\,\overline{\mathrm{E}}[z_{0i}^2]^{1/2}\overline{\mathrm{E}}[q_{0i}^2]^{1/2} + \mu_0^2\,\overline{\mathrm{E}}[q_{0i}^2]\big\} \\ &\quad + |\mu_0|^3/\gamma_0^3 \\ &\lesssim C, \end{aligned} \tag{A.1.9}$$

because of the moment conditions in Condition ITob 4.1–(ii) and (A.1.6) above. Then, by Lyapunov's central limit theorem and ITob 4.1–(ii), we have

$$\sqrt{n}\,(\emptyset) = \sqrt{n}\,\mathbb{E}_n[\psi_{\alpha_0,h_0}(m_i)] \rightsquigarrow \mathcal{N}\big(0, \overline{\mathrm{E}}[\{\psi_{\alpha_0,h_0}(m_i)\}^2]\big),$$

which verifies the second claim of Theorem 1.4.1.

Step 2: (Bounding $\Gamma(\alpha, \hat{h})$ for $|\alpha - \alpha_0| \leqslant \delta_n$)

By expanding $\Gamma(\alpha, \hat{h})$ we get

$$\Gamma(\alpha, \hat{h}) = \Gamma(\alpha, h_0) + \big\{\Gamma(\alpha, \hat{h}) - \Gamma(\alpha, h_0) - \Gamma_2(\alpha, h_0)[\hat{h} - h_0]\big\} + \Gamma_2(\alpha, h_0)[\hat{h} - h_0]. \tag{A.1.10}$$

(a) (Expansion of $\Gamma(\alpha, h_0)$) We use our definitions for $\Gamma_1(\alpha, \tilde{h})$ and $\Gamma_{1,2}(\alpha, \tilde{h})$ in (A.1.3) and (A.1.4). By a first-order Taylor expansion with second-order Lagrange remainder about α_0, there exists some point $\tilde{\alpha} \in [\alpha_0, \alpha]$ such that the first term in (A.1.10) becomes

$$\begin{aligned}\Gamma(\alpha, h_0) &= \underbrace{\Gamma(\alpha_0, h_0)}_{=0} + \Gamma_1(\alpha_0, h_0)(\alpha - \alpha_0) + \Gamma_{1,2}(\tilde{\alpha}, h_0)(\alpha - \alpha_0)^2/2 \quad \text{(A.1.11)} \\ &= \big\{\overline{\mathrm{E}}[w_i z_{0i} d_i] - \mu_0 \underbrace{\overline{\mathrm{E}}[w_i q_{0i} d_i]}_{=0}\big\}(\alpha - \alpha_0) + \Gamma_{1,2}(\tilde{\alpha}, h_0)(\alpha - \alpha_0)^2/2 \\ &= \overline{\mathrm{E}}[w_i z_{0i} d_i](\alpha - \alpha_0) + O(\delta_n |\alpha - \alpha_0|)\end{aligned}$$

because $\overline{\mathrm{E}}[w_i(d_i,\, x_i) q_{0i}] = 0$ by Condition ITob 4.1–(ii), $|\alpha - \alpha_0| \leqslant \delta_n$, and

$$\begin{aligned}\big|\Gamma_{1,2}(\tilde{\alpha}, h_0)\big| &= \big|\overline{\mathrm{E}}[g''(\gamma_0 y_i - \tilde{\alpha} d_i - x_i \beta_0) d_i^2 (z_{0i} - \mu_0 q_{0i})]\big| \\ &\leqslant \bar{L}''\big\{\overline{\mathrm{E}}[|d_i^2 z_{0i}|] + |\mu_0|\, \overline{\mathrm{E}}[|d_i^2 q_{0i}|]\big\} \\ &\leqslant \bar{L}''\big\{\overline{\mathrm{E}}[d_i^4]^{1/2} \overline{\mathrm{E}}[z_{0i}^2]^{1/2} + |\mu_0|\, \overline{\mathrm{E}}[d_i^4]^{1/2} \overline{\mathrm{E}}[q_{0i}^2]^{1/2}\big\}\end{aligned}$$

is bounded by the moment restrictions in Condition ITob 4.1–(ii).

(b) (Bounding $\Gamma_2(\alpha, h_0)$) The directional derivative Γ_2 with respect to the direction $[\hat{h} - h_0]$ evaluated at some point $\tilde{h} = (\tilde{\beta}, \tilde{\gamma}, \tilde{z}, \tilde{q}, \tilde{\mu})$ is given by

$$\begin{aligned}\Gamma_2(\alpha, \tilde{h})[\hat{h} - h_0] &= -\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i \tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i) x_i \{\hat{\beta} - \beta_0\}] \\ &\quad + \overline{\mathrm{E}}[\{g'(\tilde{\gamma} y_i - \alpha d_i - x_i \tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i) y_i + \tilde{\mu} s_i/\tilde{\gamma}^2\}\{\hat{\gamma} - \gamma_0\}] \\ &\quad + \overline{\mathrm{E}}[g(\tilde{\gamma} y_i - \alpha d_i - x_i \tilde{\beta})\{\hat{z}_i - z_{0i}\}] \\ &\quad - \overline{\mathrm{E}}[g(\tilde{\gamma} y_i - \alpha d_i - x_i \tilde{\beta})\tilde{\mu}\{\hat{q}_i - q_{0i}\}] \\ &\quad - \overline{\mathrm{E}}[\{g(\tilde{\gamma} y_i - \alpha d_i - x_i \tilde{\beta})\tilde{q}_i + s_i/\tilde{\gamma}\}\{\hat{\mu} - \mu_0\}]\end{aligned}$$

Note that if Γ_2 is evaluated at (α_0, h_0) all terms on the right-hand side vanish, i.e., $\Gamma_2(\alpha_0, h_0)[\hat{h} - h_0] = 0$ because of orthogonality $\overline{\mathrm{E}}[w_i z_{0i} x_i] = 0$, $\overline{\mathrm{E}}[w_i q_{0i} x_i] = 0$, $\overline{\mathrm{E}}[g(\gamma_0 y_i - \alpha_0 d_i - x_i \beta_0) q_{0i} + s_i/\gamma_0] = 0$ in Condition ITob 4.1–(ii), by construction of our instruments z_{0i}, q_{0i}, and definitions of w_i and μ_0 in (A.1.5) and (A.1.6):

$$\begin{aligned}\overline{\mathrm{E}}[w_i z_{0i} x_i]\{\hat{\beta} - \beta_0\} &= 0, \\ \mu_0 \overline{\mathrm{E}}[w_i q_{0i} x_i]\{\hat{\beta} - \beta_0\} &= 0, \\ \{\overline{\mathrm{E}}[w_i z_{0i} y_i] - \mu_0 \overline{\mathrm{E}}[w_i q_{0i} y_i + s_i/\gamma_0^2]\}\{\hat{\gamma} - \gamma_0\} &= 0.\end{aligned}$$

Therefore, we obtain the following bound for Γ_2

$$
\begin{aligned}
\left|\Gamma_2(\alpha, h_0)[\hat{h} - h_0]\right| &= \left|\Gamma_2(\alpha, h_0)[\hat{h} - h_0] - \Gamma_2(\alpha_0, h_0)[\hat{h} - h_0]\right| \qquad \text{(A.1.12)} \\
&= \Big| - \overline{\mathrm{E}}[\{g'(\gamma_0 y_i - \alpha d_i - x_i\beta_0) + w_i\}(z_{0i} - \mu_0 q_{0i})x_i\{\hat{\beta} - \beta_0\}] \\
&\quad + \overline{\mathrm{E}}[\{g'(\gamma_0 y_i - \alpha d_i - x_i\beta_0) + w_i\}(z_{0i} - \mu_0 q_{0i})y_i\{\hat{\gamma} - \gamma_0\}] \\
&\quad + \overline{\mathrm{E}}[\{g(\gamma_0 y_i - \alpha d_i - x_i\beta_0) \\
&\qquad\qquad - g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}\{\hat{z}_i - z_{0i}\}] \\
&\quad - \overline{\mathrm{E}}[\{g(\gamma_0 y_i - \alpha d_i - x_i\beta_0) \\
&\qquad\qquad - g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}\mu_0\{\hat{q}_i - q_{0i}\}] \\
&\quad - \overline{\mathrm{E}}[\{g(\gamma_0 y_i - \alpha d_i - x_i\beta_0) \\
&\qquad\qquad - g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)\}q_{0i}\{\hat{\mu} - \mu_0\}]\Big| \\
&\leqslant \bar{L}''|\alpha - \alpha_0|\, \overline{\mathrm{E}}[|d_i\{z_{0i} - \mu_0 q_{0i}\}|\, |x_i\{\hat{\beta} - \beta_0\}|] \\
&\quad + \bar{L}''|\alpha - \alpha_0|\, \overline{\mathrm{E}}[|d_i\{z_{0i} - \mu_0 q_{0i}\}|\, |y_i\{\hat{\gamma} - \gamma_0\}|] \\
&\quad + \bar{L}'|\alpha - \alpha_0|\, \big\{\overline{\mathrm{E}}[|d_i|\, |\hat{z}_i - z_{0i}|] \\
&\qquad\qquad + \mu_0\bar{L}'|\alpha - \alpha_0|\, \overline{\mathrm{E}}[|d_i|\, |\hat{q}_i - q_{0i}|]\big\} \\
&\quad + \bar{L}'|\alpha - \alpha_0|\, \overline{\mathrm{E}}[|d_i q_{0i}|\, |\hat{\mu} - \mu_0|] \\
&\leqslant \bar{L}''|\alpha - \alpha_0|\, \overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^2]^{1/2} \\
&\qquad\qquad \cdot\big\{\overline{\mathrm{E}}[d_i^2 z_{0i}^2]^{1/2} + \mu_0\overline{\mathrm{E}}[d_i^2 q_{0i}^2]^{1/2}\big\} \\
&\quad + \bar{L}''|\alpha - \alpha_0|\, \overline{\mathrm{E}}[y_i^2]^{1/2}|\hat{\gamma} - \gamma_0| \\
&\qquad\qquad \cdot\big\{\overline{\mathrm{E}}[d_i^2 z_{0i}^2]^{1/2} + \mu_0\overline{\mathrm{E}}[d_i^2 q_{0i}^2]^{1/2}\big\} \\
&\quad + \bar{L}'|\alpha - \alpha_0|\, \overline{\mathrm{E}}[d_i^2]^{1/2}\big\{\overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} \\
&\qquad\qquad + \mu_0\overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2}\big\} \\
&\quad + \bar{L}'|\alpha - \alpha_0|\, \overline{\mathrm{E}}[d_i^2]^{1/2}\overline{\mathrm{E}}[q_{0i}^2]^{1/2}\, |\hat{\mu} - \mu_0| \\
&\lesssim |\alpha - \alpha_0|\delta_n
\end{aligned}
$$

because $w_i = -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)$ in (1.2.9), $\overline{\mathrm{E}}[d_i^4] \leqslant C$, $\mathrm{E}[y_i^2] \leqslant C$, $\mathrm{E}[z_{0i}^4] \leqslant C$, $\overline{\mathrm{E}}[q_{0i}^4] \leqslant C$, $\overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^2]^{1/2} \lesssim \|\hat{\beta} - \beta_0\| \leqslant \delta_n n^{-1/4}$, $|\hat{\gamma} - \gamma_0| \leqslant \delta_n n^{-1/4}$, $|\hat{\mu} - \mu_0| \leqslant \delta_n$, $\overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} \leqslant \delta_n$, and $\overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2} \leqslant \delta_n$ in Condition ITob 4.1–(ii) and ITob 4.1–(iii), (1.4.1).

(c) (Bounding $\Gamma_{2,2}(\alpha, \tilde{h})$) For some $\tilde{h} \in [\hat{h}, h_0]$, we have

$$\Gamma(\alpha, \hat{h}) = \Gamma(\alpha, h_0) + \Gamma_2(\alpha, h_0)[\hat{h} - h_0] + \Gamma_{2,2}(\alpha, \tilde{h})[\hat{h} - h_0, \hat{h} - h_0]/2.$$

Thus, the second term on the right-hand side of (A.1.10) is bounded by the second directional derivative $\Gamma_{2,2}$, namely

$$\left|\Gamma(\alpha, \hat{h}) - \Gamma(\alpha, h_0) - \Gamma_2(\alpha, h_0)[\hat{h} - h_0]\right| \leqslant \left|\Gamma_{2,2}(\alpha, \tilde{h})[\hat{h} - h_0, \hat{h} - h_0]\right|.$$

The second directional derivative $\Gamma_{2,2}$ with respect to the direction $[\hat{h} - h_0]$ evaluated at $\tilde{h} = (\tilde{\beta}, \tilde{\gamma}, \tilde{z}, \tilde{q}, \tilde{\mu})$, in turn, can be bounded by

$$
\begin{aligned}
\left|\Gamma_{2,2}(\alpha, \tilde{h})\right| \;=\; & \Big| \overline{\mathrm{E}}[g''(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i)(x_i\{\hat{\beta} - \beta_0\})^2] \\
& - 2\,\overline{\mathrm{E}}[g''(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i)(x_i\{\hat{\beta} - \beta_0\})(y_i\{\hat{\gamma} - \gamma_0\})] \\
& - 2\,\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\{\hat{z}_i - z_{0i}\}x_i\{\hat{\beta} - \beta_0\}] \\
& + 2\,\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\{\hat{q}_i - q_{0i}\}\tilde{\mu} x_i\{\hat{\beta} - \beta_0\}] \\
& + 2\,\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\tilde{q}_i x_i\{\hat{\beta} - \beta_0\}\{\hat{\mu} - \mu_0\}] \\
& + \overline{\mathrm{E}}[\{g''(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})(\tilde{z}_i - \tilde{\mu}\tilde{q}_i) - \tilde{\mu} s_i/\tilde{\gamma}^3\}(y_i\{\hat{\gamma} - \gamma_0\})^2] \\
& + 2\,\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\{\hat{z}_i - z_{0i}\}y_i\{\hat{\gamma} - \gamma_0\}] \\
& - 2\,\overline{\mathrm{E}}[g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\{\hat{q}_i - q_{0i}\}\tilde{\mu} y_i\{\hat{\gamma} - \gamma_0\}] \\
& - \overline{\mathrm{E}}[\{g'(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\tilde{q}_i y_i - s_i/\tilde{\gamma}^2\}\{\hat{\gamma} - \gamma_0\}\{\hat{\mu} - \mu_0\}] \\
& - 2\,\overline{\mathrm{E}}[g(\tilde{\gamma} y_i - \alpha d_i - x_i\tilde{\beta})\{\hat{q}_i - q_{0i}\}\{\hat{\mu} - \mu_0\}]\Big| \\
\leqslant\; & \bar{L}''\,\overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^4]^{1/2}\big\{\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\} \\
& + 2\,\bar{L}''\,\overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^4]^{1/4}\overline{\mathrm{E}}[(y_i\{\hat{\gamma} - \gamma_0\})^4]^{1/4} \\
& \qquad\qquad \cdot\big\{\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\} \\
& + 2\,\bar{L}'\,\overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^2]^{1/2}\big\{\overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2}\big\} \\
& + 2\,\bar{L}'\,|\hat{\mu} - \mu_0|\,\overline{\mathrm{E}}[(x_i\{\hat{\beta} - \beta_0\})^2]^{1/2}\overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2} \\
& + \bar{L}''\,\overline{\mathrm{E}}[(y_i\{\hat{\gamma} - \gamma_0\})^4]^{1/2}\big\{\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\} \\
& + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[(y_i\{\hat{\gamma} - \gamma_0\})^2]/\tilde{\gamma}^3 + |\hat{\gamma} - \gamma_0|\,|\hat{\mu} - \mu_0|/\tilde{\gamma}^2 \\
& + 2\,\bar{L}'\,\overline{\mathrm{E}}[(y_i\{\hat{\gamma} - \gamma_0\})^2]^{1/2}\big\{\overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2}\big\} \\
& + \bar{L}'\,|\hat{\gamma} - \gamma_0|\,|\hat{\mu} - \mu_0|\,\overline{\mathrm{E}}[y_i^2]^{1/2}\overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2} \\
& + 2\,\bar{L}\,|\hat{\mu} - \mu_0|\,\overline{\mathrm{E}}[|\hat{q}_i - q_{0i}|] \\
\lesssim\; & \big\{\|\hat{\beta} - \beta_0\|^2 + \|\hat{\beta} - \beta_0\|\,|\hat{\gamma} - \gamma_0|\big\}\big\{\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\} \\
& + \big\{\|\hat{\beta} - \beta_0\| + |\hat{\gamma} - \gamma_0|\big\}\big\{\overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2}\big\} \\
& + |\hat{\mu} - \mu_0|\,\overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\{\|\hat{\beta} - \beta_0\| + |\hat{\gamma} - \gamma_0|\big\} \\
& + |\hat{\gamma} - \gamma_0|^2\big\{\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} + |\tilde{\mu}| \cdot \overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2}\big\} + |\hat{\mu} - \mu_0|\,\overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2} \\
& + |\tilde{\mu}| \cdot |\hat{\gamma} - \gamma_0|^2/\tilde{\gamma}^3 + |\hat{\gamma} - \gamma_0|\,|\hat{\mu} - \mu_0|/\tilde{\gamma}^2
\end{aligned}
$$

because of $\overline{\mathrm{E}}[\{x_i\xi\}^4] \leqslant C\|\xi\|^4$ in Condition ITob 4.1–(ii), and the moment restrictions above. Note that since $\tilde{h} \in [\hat{h}, h_0]$, we have $|\tilde{z}_i| \leqslant |z_{0i}| + |\hat{z}_i - z_{0i}|$, $|\tilde{q}_i| \leqslant |q_{0i}| + |\hat{q}_i - q_{0i}|$, and $|\tilde{\mu}| \leqslant |\mu_0| + |\hat{\mu} - \mu_0|$ such that $\overline{\mathrm{E}}[\tilde{z}_i^2]^{1/2} \leqslant \overline{\mathrm{E}}[z_{0i}^2]^{1/2} + \overline{\mathrm{E}}[(\hat{z}_i - z_{0i})^2]^{1/2} \leqslant C + \delta_n$, $\overline{\mathrm{E}}[\tilde{q}_i^2]^{1/2} \leqslant \overline{\mathrm{E}}[q_{0i}^2]^{1/2} + \overline{\mathrm{E}}[(\hat{q}_i - q_{0i})^2]^{1/2} \leqslant C + \delta_n$, and $|\tilde{\mu}| \leqslant C + \delta_n$ by the Minkowski inequality and Condition ITob 4.1–(iii), (1.4.1). Additionally, because of $1/\tilde{\gamma} \in [1/\gamma_0, 1/\hat{\gamma}]$ we have $|1/\tilde{\gamma}| \leqslant |1/\gamma_0| + |1/\hat{\gamma} - 1/\gamma_0|$. Moreover, $|1/\hat{\gamma} - 1/\gamma_0| = |\hat{\gamma} - \gamma_0|/\gamma_0\hat{\gamma} \lesssim \delta_n n^{-1/4}$ by Condition ITob 4.1–(i) and ITob 4.1–(iii), (1.4.1).

As a result, with probability $1 - \Delta_n$

$$\begin{aligned}\left|\Gamma(\alpha, \hat{h}) - \Gamma(\alpha, h_0) - \Gamma_2(\alpha, h_0)[\hat{h} - h_0]\right| \\ \leqslant \sup_{\tilde{h} \in [h_0, \hat{h}]} \left|\Gamma_{2,2}(\alpha, \tilde{h})[\hat{h} - h_0, \hat{h} - h_0]\right| \lesssim \delta_n n^{-1/2}. \quad \text{(A.1.13)}\end{aligned}$$

Combining our results from (a), (b), and (c) we get

$$\begin{aligned}\Gamma(\alpha, \hat{h}) &= \Gamma(\alpha, h_0) + \Gamma(\alpha, \hat{h}) - \Gamma(\alpha, h_0) - \Gamma_2(\alpha, h_0)[\hat{h} - h_0] + \Gamma_2(\alpha, h_0)[\hat{h} - h_0] \\ &= \overline{\mathrm{E}}[w_i z_{0i} d_i](\alpha - \alpha_0) + O(\delta_n |\alpha - \alpha_0| + \delta_n n^{-1/2}), \quad \text{(A.1.14)}\end{aligned}$$

which verifies the assertion in Step 1 (b) above.

Step 3: (Estimation of Variance) We define

$$\widehat{w}_i := -g'(\hat{\gamma} y_i - \hat{\alpha} d_i - x_i \hat{\beta}). \quad \text{(A.1.15)}$$

Note that $|\widehat{w}_i| \leqslant \bar{L}'$ by Condition ITob 4.1–(i).

(a) For the first term of the variance we get

$$\begin{aligned}\left|\mathbb{E}_n[\widehat{w}_i d_i \hat{z}_i] - \overline{\mathrm{E}}[w_i d_i z_{0i}]\right| &= \left|\mathbb{E}_n[\widehat{w}_i d_i \hat{z}_i] - \mathbb{E}_n[w_i d_i \hat{z}_i] - \mathbb{E}_n[\{\widehat{w}_i - w_i\} d_i z_{0i}]\right. \\ &\quad - \mathbb{E}_n[w_i d_i z_{0i}] + \mathbb{E}_n[w_i d_i \hat{z}_i] + \mathbb{E}_n[\{\widehat{w}_i - w_i\} d_i z_{0i}] \\ &\quad \left. + \mathbb{E}_n[w_i d_i z_{0i}] - \overline{\mathrm{E}}[w_i d_i z_{0i}]\right| \quad \text{(A.1.16)} \\ &\leqslant \left|\mathbb{E}_n[\{\widehat{w}_i - w_i\} d_i \{\hat{z}_i - z_{0i}\}]\right| + \left|\mathbb{E}_n[\{\widehat{w}_i - w_i\} d_i z_{0i}]\right| \\ &\quad + \mathbb{E}_n[w_i^2 d_i^2]^{1/2} \mathbb{E}_n[(\hat{z}_i - z_{0i})^2]^{1/2} \\ &\quad + \left|\mathbb{E}_n[w_i d_i z_{0i}] - \overline{\mathrm{E}}[w_i d_i z_{0i}]\right| \\ &\leqslant \mathbb{E}_n[\{\widehat{w}_i - w_i\}^2 d_i^2]^{1/2} \mathbb{E}_n[(\widehat{z}_i - z_{0i})^2]^{1/2} \\ &\quad + \mathbb{E}_n[(\widehat{w}_i - w_i)^2]^{1/2} \mathbb{E}_n[d_i^4]^{1/4} \mathbb{E}_n[z_{0i}^4]^{1/4} \\ &\quad + \mathbb{E}_n[w_i^2 d_i^2]^{1/2} \mathbb{E}_n[(\hat{z}_i - z_{0i})^2]^{1/2} \\ &\quad + \left|\mathbb{E}_n[w_i d_i z_{0i}] - \overline{\mathrm{E}}[w_i d_i z_{0i}]\right| \\ &\lesssim_{\mathrm{P}} \delta_n\end{aligned}$$

because of the moment restrictions in Condition ITob 4.1–(ii) and $\|\{\hat{w}_i - w_i\} d_i\|_{2,n} \leqslant \delta_n$, $\|\hat{w}_i - w_i\|_{2,n} \leqslant \delta_n$, and $\|\hat{z}_i - z_{0i}\|_{2,n} \leqslant \delta_n$ with probability at least $1 - \Delta_n$ in Condition ITob 4.1–(iv).

(b) To control the second term of the variance, note that for $\hat{h} = (\hat{\beta}, \hat{\gamma}, \hat{z}, \hat{q}, \hat{\mu})$

$$\begin{aligned}
|\psi_{\check{\alpha},\hat{h}}(m_i) - \psi_{\alpha_0,h_0}(m_i)| \;\leqslant\; & |\psi_{\check{\alpha},\hat{h}}(m_i) - \psi_{\alpha_0,\hat{h}}(m_i)| + |\psi_{\alpha_0,\hat{h}}(m_i) - \psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,\hat{q},\hat{\mu}}(m_i)| \\
& + |\psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,\hat{q},\hat{\mu}}(m_i) - \psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,q_0,\hat{\mu}}(m_i)| \\
& + |\psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,q_0,\hat{\mu}}(m_i) - \psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,q_0,\mu_0}(m_i)| \\
& + |\psi_{\alpha_0,\hat{\beta},\hat{\gamma},z_0,q_0,\mu_0}(m_i) - \psi_{\alpha_0,\beta_0,\hat{\gamma},z_0,q_0,\mu_0}(m_i)| \\
& + |\psi_{\alpha_0,\beta_0,\hat{\gamma},z_0,q_0,\mu_0}(m_i) - \psi_{\alpha_0,h_0}(m_i)| \\
\leqslant\; & \bar{L}' \, |d_i(\check{\alpha} - \alpha_0)(\hat{z}_i - \hat{\mu}\hat{q}_i)| + \bar{L} \, |\hat{z}_i - z_{0i}| + \bar{L} \, |\hat{\mu}| \, |\hat{q}_i - q_{0i}| \\
& + |\hat{\mu} - \mu_0| \cdot \{\bar{L} \, |q_{0i}| + |s_i/\hat{\gamma}|\} \\
& + \bar{L}' \, |x_i\{\hat{\beta} - \beta_0\}(z_{0i} - \mu_0 q_{0i})| \\
& + \bar{L}' \, |y_i\{\hat{\gamma} - \gamma_0\}(z_{0i} - \mu_0 q_{0i})| + |\mu_0| s_i |1/\hat{\gamma} - 1/\gamma_0|
\end{aligned}$$

and therefore

$$\begin{aligned}
\big| \|\psi_{\check{\alpha},\hat{h}}(m_i)\|_{2,n} - \|\psi_{\alpha_0,h_0}(m_i)\|_{2,n} \big| \;\leqslant\; & \bar{L}' \, \|d_i(\check{\alpha} - \alpha)(\hat{z}_i - \hat{\mu}\hat{q}_i)\|_{2,n} + \bar{L} \, \|\hat{z}_i - z_{0i}\|_{2,n} \\
& + \bar{L} \, |\hat{\mu}| \, \|\hat{q}_i - q_{0i}\|_{2,n} \\
& + |\hat{\mu} - \mu_0| \, \|\bar{L} \, |q_{0i}| + s_i/\hat{\gamma}\|_{2,n} \\
& + \bar{L}' \, \|x_i\{\hat{\beta} - \beta_0\}(z_{0i} - \mu_0 q_{0i})\|_{2,n} \\
& + \bar{L}' \, \|y_i\{\hat{\gamma} - \gamma_0\}(z_{0i} - \mu_0 q_{0i})\|_{2,n} \\
& + |\mu_0| \, |1/\hat{\gamma} - 1/\gamma_0| \\
\lesssim_{\mathrm{P}} \;\; & \delta_n
\end{aligned}$$

because $\|d_i(\hat{z}_i - \hat{\mu}\hat{q}_i)\|_{2,n} \leqslant C$, $\|\hat{z}_i - z_{0i}\|_{2,n} \leqslant \delta_n$, $\|\hat{q}_i - q_{0i}\|_{2,n} \leqslant \delta_n$, $\|\bar{L} \, |q_{0i}| + s_i/\hat{\gamma}\|_{2,n} \leqslant C$, $\|x_i\{\hat{\beta} - \beta_0\}(z_{0i} - \mu_0 q_{0i})\|_{2,n} \leqslant \delta_n$, and $\|y_i(z_{0i} - \mu_0 q_{0i})\|_{2,n} \leqslant C$ by Condition ITob 4.1–(iv). Moreover, by equation (A.1.9) above $\overline{\mathrm{E}}[|\psi_{\alpha_0,h_0}(m_i)|^3] \lesssim C$, we have $|\mathbb{E}_n[\{\psi_{\alpha_0,h_0}(m_i)\}^2] - \overline{\mathrm{E}}[\{\psi_{\alpha_0,h_0}(m_i)\}^2]| \lesssim_{\mathrm{P}} \delta_n$. ■

A.2.

PROOF OF THEOREM 1.4.2

In this section, we show that Condition ITob 4.1 is implied by the choice of the logistic Tobit likelihood loss, the model described by (1.2.1), (1.2.10), (1.2.11) and the set of primitive assumptions in ITob 4.2. The result in Theorem 1.4.2 then follows from Theorem 1.4.1.

Step 1: (Verification of Conditions ITob 4.1–(i),(ii))
Recall from Section 1.2.2 that for the logistic likelihood loss the Lipschitz constants in Condition ITob 4.1–(i) are $\bar{L} = 1$, $\bar{L}' = 1/2$, and $\bar{L}'' = 3^{-3/2}$ such that $\bar{L} \vee \bar{L}' \vee \bar{L}'' \leqslant 1$. The moment conditions $\mathrm{E}[w_i z_{0i} x_i]$ and $\mathrm{E}[w_i(d_i,\, x_i) q_{0i}]$ hold by (1.2.12) and (1.2.13), respectively. Since $z_{0i} = d_i - x_i\eta_0$ and $\mathrm{E}[w_i z_{0i} x_i] = 0$, we have that $\overline{\mathrm{E}}[w_i z_{0i} d_i] = \overline{\mathrm{E}}[w_i z_{0i}^2] = \overline{\mathrm{E}}[w_i(d_i - x_i\eta_0)^2] \geqslant \underline{c}\|(1,\, \eta_0)\|^2$ by assuming $\overline{\mathrm{E}}[w_i\{(d_i,\, x_i,\, -y_i)\xi\}^2] \geqslant \underline{c}\|\xi\|^2$. Analogously, since $q_{0i} = y_i - (d_i,\, x_i)\theta_0$ and $\mathrm{E}[w_i(d_i,\, x_i) q_{0i}] = 0$, we have $\mathrm{E}[w_i q_{0i} y_i + s_i/\gamma_0^2] = \mathrm{E}[w_i q_{0i}^2] + \mathrm{E}[s_i\sigma_0^2] \geqslant \underline{c}\|(1,\, \theta_0)\|^2 + \mathrm{E}[s_i\sigma_0^2]$ if $\overline{\mathrm{E}}[w_i\{(d_i,\, x_i,\, -y_i)\xi\}^2] \geqslant \underline{c}\|\xi\|^2$. Consequently, Conditions ITob 4.1–(i) and ITob 4.1–(ii) hold for the logistic Tobit.

To verify Conditions ITob 4.1–(iii),(iv), we follow the proofs in Appendix A of Belloni et al. (2016b) for the logistic binary choice model. We make adjustments to account for the differences between the Neyman orthogonal score functions associated with logistic Tobit and logistic GLM, specifically to accommodate the inclusion of the inverse scale parameter γ and the censored outcomes y_i in the design quantities.

Step 2 (a): (Discussion of conditions for design quantities; see Bickel et al., 2009 and Belloni et al., 2016b)
For $\tilde{x}_i := (d_i,\, x_i)$, $i \in \{1, \ldots, n\}$, and $\delta := (\delta_d,\, \delta_x^T,\, \delta_y)^T = (\delta_{\tilde{x}}^T,\, \delta_y)^T \in \mathbb{R}^{p+2}$ we define the minimal and maximal m-sparse empirical eigenvalues as

$$\phi^u_{\min,n}(m) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2}{\|\delta\|^2} \quad \text{and} \quad \phi^u_{\max,n}(m) := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2}{\|\delta\|^2}.$$

By Lemma A.6.3, under the growth condition $4\bar{C}^2 K_2^2 s \log(p \vee n) \log(n) \log^2(1+s) \lesssim K_2^2 s \log(p \vee n) \log^3(n) \leqslant \delta_n n$ with $K_2 := \mathrm{E}[\max_{1 \leqslant i \leqslant n} \|(\tilde{x}_i,\, y_i)^2\|_\infty]^{1/2}$, we have that with probability $1 - o(1)$ the sparse eigenvalues of order $k = s\ell_n$ are bounded away from zero and from above by a constant for some sequence $\ell_n \to \infty$, i.e. $0 < \underline{c} \leqslant \phi^u_{\min,n}(s\ell_n) \leqslant \phi^u_{\max,n}(s\ell_n) \leqslant C < \infty$. Moreover, Remark A.6.7 states the following relation for the sparse empirical eigenvalues of augmented and non-augmented design design quantities as defined in Definition A.6.3: $\phi^u_{\min,n}(s\ell_n) \leqslant \phi^{u*}_{\min,n}(s\ell_n) \leqslant \phi^{u*}_{\max,n}(s\ell_n) \leqslant \phi^u_{\max,n}(s\ell_n)$. Consequently, boundedness of the sparse eigenvalues of the augmented empirical Gram matrix implies the boundedness of the sparse eigenvalues of the original Gram matrix (excluding the additional column associated with y_i). From the discussion in Bickel et al. (2009) follows that, if the sparse eigenvalues of order Cs are bounded away from zero and from above for a suitable constant C, the minimal restricted empirical eigenvalue $\kappa^u_{\min,n}(\mathbf{c})$ in Definition A.6.1 is bounded away from zero. Provided that the quotient of weighted and non-weighted empirical design quantities $\nu_{(r),n}(\mathbf{c})$ in Definition A.6.4 is bounded away from zero for $\mathbf{c} = (1+c)/(c-1)$ with $c > 1$, we have that the minimal restricted empirical eigenvalue of the weighted design quantity $\kappa_{\min,n}(\mathbf{c})$ is bounded away from zero. Therefore, under the additional condition

$\min_{1\leqslant i\leqslant n} w_i \geqslant \underline{c}$, we have that $\kappa_{\min,n}(\mathbf{c})$ is bounded away from zero with probability $1-o(1)$ for sufficiently large n; see Appendix A of Belloni et al. (2016b).

Step 2 (b): (Verification of the conditions for post-ℓ_1 rates of logistic Tobit estimator) Step 1 of Algorithms 1.2.1 and 1.2.2 are based on the post-ℓ_1 logistic Tobit estimator. To obtain convergence rates and sparsity bounds, we have to verify the side-conditions in Lemmas A.6.11 and A.6.13. At the negligible cost of increasing the cardinality of the active set by one, we assume that the index of the target regressor d belongs to set $\mathcal{S} = \{j \in \{1,\dots,p+2\} : \Theta_{0j} \neq 0\}$ and $|\mathcal{S}| = s$. Indeed, we have $|\hat{\gamma}-\gamma_0| \leqslant \|\hat{\delta}_{\mathcal{S}}\| = \|\widehat{\Theta}_{\mathcal{S}} - \Theta_{0\mathcal{S}}\|$ by construction. Furthermore, we define $\widehat{\Theta} = (\hat{\alpha},\, \hat{\beta}^T,\, \hat{\gamma})^T$, $\tilde{\Theta} = (\tilde{\alpha},\, \tilde{\beta}^T,\, \tilde{\gamma})^T$ and $\Theta_0 = (\alpha_0,\, \beta_0^T,\, \gamma_0)^T$.

The non-linear impact coefficient $\overline{q}_{\Delta_{\mathbf{c},n}}$ in Definition A.6.2 for set $\Delta_{\mathbf{c}} = \{\delta \in \mathbb{R}^{p+2} : \|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1\}$ can be bounded from below by

$$\begin{aligned}
\inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^3]} &\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2\|\delta_{\mathcal{S}}\|\kappa_{\min,n}(\mathbf{c})}{\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty\|\delta\|_1\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2} \\
&\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\delta_{\mathcal{S}}\|\kappa_{\min,n}(\mathbf{c})}{\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty\sqrt{s}(1+\mathbf{c})\|\delta_{\mathcal{S}}\|} \\
&\gtrsim_{\mathrm{P}} \frac{1}{\mathrm{E}[\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty]\sqrt{s}},
\end{aligned}$$

because for $K_1 = \mathrm{E}[\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty]$ and some $\varepsilon > 0$ we have $\mathrm{P}(\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty \geqslant K_1/\varepsilon) \leqslant \varepsilon$ by Markov's inequality. Moreover, for $\lambda/n \lesssim \sqrt{\log(p\vee n)/n}$ and the growth condition $K_1^2 s^2 \log^2(p\vee n) \leqslant \delta_n n$ in Condition ITob 4.2–(iii), we have

$$\frac{\lambda\sqrt{s}}{n\kappa_{\min,n}(\mathbf{c})} \lesssim \sqrt{\frac{s\log(p\vee n)}{n\{\kappa_{\min,n}(\mathbf{c})\}^2}} \leqslant \frac{\sqrt{\delta_n}}{K_1\sqrt{s}\kappa_{\min,n}(\mathbf{c})} \ll \frac{1}{K_1\sqrt{s}} \lesssim_{\mathrm{P}} \overline{q}_{\Delta_{\mathbf{c},n}} \tag{A.2.1}$$

such that the side-condition $\overline{q}_{\Delta_{\mathbf{c},n}} > 3(1+1/c)\lambda\sqrt{s}/\{n\kappa_{\min,n}(\mathbf{c})\}$ holds with probability $1-o(1)$ as $\delta_n \searrow 0$.

By the results of Lemmas A.6.8, A.6.9, A.6.10, and Remark A.6.8 we have $\lambda_y/n \geqslant c\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty$ with probability at least $1-\Delta$ if $\lambda_y = c\sqrt{2n\log(2(p+1)/\Delta)}$. Therefore, Lemma A.6.11 yields $\|\widehat{\Theta}-\Theta_0\|_1 \lesssim s\sqrt{\log(p\vee n)/n}$, $\|\widehat{\Theta}-\Theta_0\| \lesssim \sqrt{s\log(p\vee n)/n}$ and $\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) \lesssim s\log(p\vee n)/n$. In particular, we have $|\hat{\alpha}-\alpha_0| \leqslant \|\widehat{\Theta}-\Theta_0\|$, $\|\hat{\beta}-\beta_0\|_1 \leqslant \|\widehat{\Theta}-\Theta_0\|_1$ and $|\hat{\gamma}-\gamma_0| \leqslant \|\widehat{\Theta}-\Theta_0\|$.

Additionally, since $\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty \cdot \|\widehat{\Theta}-\Theta_0\|_1 \lesssim_{\mathrm{P}} K_1 s\sqrt{\log(p\vee n)/n} \lesssim \delta_n^{1/2} \leqslant 1$ by Markov's inequality and the growth condition stated above, Lemma A.6.13 yields $\hat{s} = \|\widehat{\Theta}\|_0 \lesssim s$ for sufficiently large n.

To show that $\hat{\gamma} \geqslant \underline{c}$, as required by Condition ITob 4.1–(iii) (1.4.1), note that by the first-order condition $\partial_\gamma \Lambda(\hat{\alpha},\hat{\beta},\gamma)|_{\gamma=\hat{\gamma}} = 0$, we have

$$\begin{aligned}
\mathbb{E}_n[s_i/\hat{\gamma}] &= -\mathbb{E}_n[g(\hat{\gamma}y_i - \hat{\alpha}d_i - x_i\hat{\beta})y_i] \\
|\mathbb{E}_n[s_i]| &= |\mathbb{E}_n[g(\hat{\gamma}y_i - \hat{\alpha}d_i - x_i\hat{\beta})y_i\hat{\gamma}]| \leqslant \hat{\gamma}\,\mathbb{E}_n[|y_i|].
\end{aligned}$$

Since $s_i \geqslant 0$ and $y_i \geqslant 0$, we get $\hat{\gamma} \geqslant \mathbb{E}_n[s_i]/\mathbb{E}_n[y_i] \gtrsim_{\mathrm{P}} \mathbb{E}_n[s_i]/\overline{\mathrm{E}}[y_i] \gtrsim \underline{c}$ by Markov's inequality if we impose $\overline{\mathrm{E}}[y_i] \leqslant C$ and assume that at least some outcome observations remain uncensored, i.e. $\mathbb{E}_n[s_i] > 0$. Indeed, we have $\mathbb{E}_n[s_i] \geqslant \overline{\mathrm{E}}[s_i] - |\mathbb{E}_n[s_i] - \overline{\mathrm{E}}[s_i]| = \overline{\mathrm{E}}[s_i] - O_{\mathrm{P}}(n^{-1/2})$ by Chebyshev's inequality. Note that the same argument applies to the post-ℓ_1 point estimate $\tilde{\gamma} \geqslant \underline{c}$.

To obtain rates for the post-ℓ_1 logistic Tobit estimator, we have to verify the side-condition in Lemma A.6.14

$$\frac{\overline{q}_{A,n}}{6} > \left\{ \sqrt{\frac{s+\hat{s}}{\phi_{\min,n}(s+\hat{s})}} \|\nabla_\Theta \Lambda(\Theta_0)\|_\infty \right\} \vee \left\{ \max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}^{1/2} \right\}, \tag{A.2.2}$$

where $\phi_{\min,n}(m)$ denotes the minimal m-sparse empirical eigenvalue of the weighted design quantity in Definition A.6.3. Because of the sparsity of $\widehat{\Theta}$ (in fact the sparsity of $\hat{\beta}$), it suffices to consider $\hat{s} \leq sC$ for some constant C and, thus, $\overline{q}_{A,n}$ for $A = \{\delta \in \mathbb{R}^{p+2} : \|\delta\|_0 \leqslant s(1+C)\}$. Then, we have

$$\begin{aligned}
\inf_{\delta \in A} \frac{\|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|(\tilde{x}_i, -y_i)\delta|^3]} &\geqslant \inf_{\delta \in A} \frac{\|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}^2 \|\delta\| \sqrt{\phi^u_{\min,n}(s(1+C))}\nu_{(s),n}(s(1+C))}{\max_{1\leqslant i\leqslant n} \|(\tilde{x}_i, y_i)\|_\infty \|\delta\|_1 \|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}^2} \\
&\geqslant \inf_{\delta \in A} \frac{\|\delta\| \sqrt{\phi^u_{\min,n}(s(1+C))}\min_{1\leqslant i\leqslant n}\sqrt{w_i}}{\max_{1\leqslant i\leqslant n} \|(\tilde{x}_i, y_i)\|_\infty \sqrt{s(1+C)}\|\delta\|} \\
&\gtrsim_{\mathrm{P}} \frac{1}{\mathrm{E}[\max_{1\leqslant i\leqslant n} \|(\tilde{x}_i, y_i)\|_\infty]\sqrt{s}},
\end{aligned}$$

because $\phi^u_{\min,n}(s(1+C))$ and $\min_{1\leqslant i\leqslant n}\sqrt{w_i}$ are bounded away from zero. Since $\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty \leqslant \lambda_y/c \lesssim \sqrt{\log(p\vee n)/n}$ and by the definition of the post-ℓ_1 estimator, we have $\Lambda(\tilde{\Theta}) \leqslant \Lambda(\widehat{\Theta})$ such that

$$\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0) \leqslant \Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) \lesssim s\log(p\vee n)/n$$

by the result from Lemma A.6.11 above. Therefore, the right-hand side of (A.2.2) is $\lesssim \sqrt{s\log(p\vee n)/n}$. Analogously to the arguments in (A.2.1) above, the growth condition $K_1^2 s^2 \log(p\vee n) \leqslant \delta_n n$ suffices for (A.2.2) to hold with probability $1 - o(1)$ for sufficiently large n. Lemma A.6.14 yields the bounds $\|\tilde{\Theta} - \Theta_0\|_1 \lesssim s\sqrt{\log(p\vee n)/n}$, $\|\tilde{\Theta} - \Theta_0\| \lesssim \sqrt{s\log(p\vee n)/n}$ and sparsity $\|\tilde{\Theta}\|_0 \leqslant \|\widehat{\Theta}\|_0 \lesssim s$. Below, we need rates on $\|x_i(\tilde{\beta} - \beta_0)\|_{2,n}$ instead of the parameters alone. In fact, we have

$$\begin{aligned}
\|x_i(\tilde{\beta} - \beta_0)\|_{2,n} &\leqslant \|(\tilde{x}_i, -y_i)\delta\|_{2,n} + \|d_i\|_{2,n}|\tilde{\alpha} - \alpha_0| + \|y_i\|_{2,n}|\tilde{\gamma} - \gamma_0| \\
&\leqslant \|(\tilde{x}_i, -y_i)\delta\|_{2,n} + \big(\|d_i\|_{2,n} + \|y_i\|_{2,n}\big)\frac{\|(\tilde{x}_i, -y_i)\delta\|_{2,n}}{\sqrt{\phi^u_{\min,n}(\hat{s}+s)}} \\
&\lesssim \sqrt{s\log(p\vee n)/n}
\end{aligned}$$

since $|\tilde{\alpha} - \alpha_0| \leqslant \|\tilde{\Theta} - \Theta_0\|$, $|\tilde{\gamma} - \gamma_0| \leqslant \|\tilde{\Theta} - \Theta_0\|$, definition (and boundedness) of the minimal sparse eigenvalue and the moment restriction in Condition ITob 4.2-(ii).

Step 3: (Verification of post-Lasso rates using the results of Belloni et al., 2016b)
Step 2 of Algorithms of Algorithms 1.2.1 and 1.2.2 is based on post-Lasso with estimated weights. As the second step of the instrumental logistic Tobit differs from the second step of the instrumental Logit in Belloni et al. (2016a) only with regard to the weights, it suffices to show that our weighting function satisfies Condition WL in Belloni et al. (2016b, 2019), allowing us to apply their results for weighted post-Lasso to our algorithms.

Note that $\widehat{w}_i = -g'(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})$, $w_i = -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)$ and it holds that $0 \leqslant \widehat{w}_i \leqslant 1/2$ and $0 < \underline{c} \leqslant w_i \leqslant 1/2$. In particular, function $-g'(t)$ is Lipschitz with constant $\bar{L}'' = 3^{-3/2} < 1$. Recall that $z_{0i} = d_i - x_i\eta_0$, $\hat{z}_i = d_i - x_i\tilde{\eta}$ and $q_{0i} = y_i - (d_i,\, x_i)\theta_0$, $\hat{q}_i = y_i - (d_i,\, x_i)\tilde{\theta}$. Using the results from Appendix A and Theorem 4 in Belloni et al. (2016b) for Lasso and post-Lasso with estimated weights, we have $\|\tilde{\eta} - \eta_0\| \lesssim \sqrt{s\log(p\vee n)/n}$, $\|\tilde{\theta} - \theta_0\| \lesssim \sqrt{s\log(p\vee n)/n}$ and $\|\tilde{\eta}\|_0 \lesssim Cs$, $\|\tilde{\theta}\|_0 \lesssim Cs$ with probability $1 - o(1)$, provided that the growth conditions $K_4^4 s\log^3(n)\log(p\vee n) \leqslant \delta_n n$ and $K_1^2 s^2\log(p\vee n) \leqslant \delta_n n$ hold for $K_a := \mathrm{E}[\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i,\, z_{0i},\, q_{0i})\|_\infty^a]^{1/a}$, the fourth moment restrictions $\overline{\mathrm{E}}[d_i^4] \leqslant C$, $\overline{\mathrm{E}}[\{x_i\xi\}^4] \leqslant C\|\xi\|^4$ hold, $\min_{1\leqslant i\leqslant n}\sqrt{\widehat{w}_i/w_i}$ is bounded away from zero, and by assuming $\|\eta_0\| \leqslant C$ and $\|\theta_0\| \leqslant C$. Furthermore, the established rates for post-Lasso rely on the iterative procedure outlined in Algorithm B.1 in Belloni et al. (2016b), a adjusted version of which is restated below as Algorithm A.4.5, to compute the diagonal matrix of penalty loadings $\widehat{\Psi}$. In order to satisfy Condition WL (ii) in Belloni et al. (2016b), we additionally assume $\min_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[w_i^2 x_{ij}^2 z_{0i}^2] \geqslant \underline{c} > 0$, $\min_{1\leqslant j\leqslant p+1}\overline{\mathrm{E}}[w_i^2\tilde{x}_{ij}^2 q_{0i}^2] \geqslant \underline{c} > 0$ and $\max_{1\leqslant j\leqslant p}\{\overline{\mathrm{E}}[|w_i x_{ij} z_{0i}|^3]\}^{1/3}\sqrt{\log(p\vee n)} \leqslant \delta_n n^{1/6}$, $\max_{1\leqslant j\leqslant p+1}\{\overline{\mathrm{E}}[|w_i\tilde{x}_{ij} q_{0i}|^3]\}^{1/3}\sqrt{\log(p\vee n)} \leqslant \delta_n n^{1/6}$.

Step 4: (Verification of Condition ITob 4.1–(iii) (1.4.1))
Equipped with the rates for $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{\eta}$ and $\tilde{\theta}$ and under the assumed growth condition $s^2\log^2(p\vee n) \leqslant \delta_n n$ in Condition ITob 4.2–(iii), we have

$$\begin{aligned}
\|\tilde{\beta} - \beta_0\| \leqslant \|\tilde{\Theta} - \Theta_0\| &\lesssim n^{-1/4}\sqrt{\frac{s\log(p\vee n)}{\sqrt{n}}} \lesssim n^{-1/4}\delta_n^{1/4} \\
|\tilde{\gamma} - \gamma_0| \leqslant \|\tilde{\Theta} - \Theta_0\| &\lesssim n^{-1/4}\sqrt{\frac{s\log(p\vee n)}{\sqrt{n}}} \lesssim n^{-1/4}\delta_n^{1/4} \\
\big\{\overline{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\big\}^{1/2} = \overline{\mathrm{E}}[\{x_i(\tilde{\eta} - \eta_0)\}^2]^{1/2} &\lesssim \|\tilde{\eta} - \eta_0\| \lesssim \sqrt{\frac{s\log(p\vee n)}{n}} \lesssim \delta_n^{1/2} \\
\big\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\big\}^{1/2} = \overline{\mathrm{E}}[\{(d_i,\, x_i)(\tilde{\theta} - \theta_0)\}^2]^{1/2} &\lesssim \|\tilde{\theta} - \theta_0\| \lesssim \sqrt{\frac{s\log(p\vee n)}{n}} \lesssim \delta_n^{1/2}
\end{aligned}$$

$$\begin{aligned}
\|\tilde{\beta}-\beta_0\|\left\{\overline{\mathrm{E}}[(\tilde{z}_i-z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\right\}^{1/2} &\lesssim \|\tilde{\beta}-\beta_0\|\|\tilde{\eta}-\eta_0\| \lesssim n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2}\\
|\tilde{\gamma}-\gamma_0|\left\{\overline{\mathrm{E}}[(\tilde{z}_i-z_{0i})^2]\big|_{\tilde{z}=\hat{z}}\right\}^{1/2} &\lesssim |\tilde{\gamma}-\gamma_0|\,\|\tilde{\eta}-\eta_0\| \lesssim n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2}\\
\|\tilde{\beta}-\beta_0\|\left\{\overline{\mathrm{E}}[(\tilde{q}_i-q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} &\lesssim \|\tilde{\beta}-\beta_0\|\|\tilde{\theta}-\theta_0\| \lesssim n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2}\\
|\tilde{\gamma}-\gamma_0|\left\{\overline{\mathrm{E}}[(\tilde{q}_i-q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} &\lesssim |\tilde{\gamma}-\gamma_0|\,\|\tilde{\theta}-\theta_0\| \lesssim n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2}.
\end{aligned}$$

Moreover, note that $\check{\alpha}\in\mathcal{A}\subset\{\alpha:|\alpha-\tilde{\alpha}|\leqslant C\log^{-1}(n)\}\subset\{\alpha:|\alpha-\alpha_0|\leqslant C\log^{-1}(n)\}$ such that $|\check{\alpha}-\alpha_0|\leqslant C\log^{-1}(n)$. In order to show the rate on $|\hat{\mu}-\mu_0|$, note that

$$\begin{aligned}
|\hat{\mu}-\mu_0| &= \left|\frac{\mathbb{E}_n[\widehat{w}_i\hat{z}_iy_i]}{\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]}-\frac{\overline{\mathrm{E}}[w_iz_{0i}y_i]}{\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]}\right|\\
&= \left|\frac{\mathbb{E}_n[\widehat{w}_i\hat{z}_iy_i]\mathrm{E}[w_iq_{0i}y_i+s_i/\gamma_0^2]-\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]\overline{\mathrm{E}}[w_iz_{0i}y_i]}{\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]}\right|\\
&\lesssim |\mathbb{E}_n[\widehat{w}_i\hat{z}_iy_i]-\overline{\mathrm{E}}[w_iz_{0i}y_i]|+|\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]-\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]|, \quad \text{(A.2.3)}
\end{aligned}$$

because $\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]\geqslant\mathbb{E}_n[s_i]/\underline{c}^2>0$ and $\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]\gtrsim\underline{c}\|(1,\,\theta_0)\|^2$ by the same argument, specifically orthogonality, as in Step 1 above. We proceed to separately bound the terms on the right-hand side of (A.2.3). To this end, we use the same approach as in (A.1.16) to control the first term:

$$\begin{aligned}
|\mathbb{E}_n[\widehat{w}_i\hat{z}_iy_i]-\overline{\mathrm{E}}[w_iz_{0i}y_i]| &\leqslant |\mathbb{E}_n[(\widehat{w}_i-w_i)(\hat{z}_i-z_{0i})y_i]|+|\mathbb{E}_n[(\widehat{w}_i-w_i)z_{0i}y_i]|\\
&\quad+|\mathbb{E}_n[w_i(\hat{z}_i-z_{0i})y_i]|+|\mathbb{E}_n[w_iz_{0i}y_i]-\overline{\mathrm{E}}[w_iz_{0i}y_i]|\\
&\leqslant \|(\widehat{w}_i-w_i)y_i\|_{2,n}\|\hat{z}_i-z_{0i}\|_{2,n}+\|\widehat{w}_i-w_i\|_{2,n}\|z_{0i}y_i\|_{2,n}\\
&\quad+\|w_iy_i\|_{2,n}\|\hat{z}_i-z_{0i}\|_{2,n}+|\mathbb{E}_n[w_iz_{0i}y_i]-\overline{\mathrm{E}}[w_iz_{0i}y_i]|,
\end{aligned}$$

which are bounded by the post-ℓ_1 rates derived in preceding paragraphs. Specifically, since $|\widehat{w}_i|\leqslant 1/2$ and $-g'(t)$ is $3^{-3/2}$-Lipschitz, we have $\|\widehat{w}_i-w_i\|_{2,n}\leqslant\|y_i\|_{2,n}|\tilde{\gamma}-\gamma_0|+\|d_i\|_{2,n}|\tilde{\alpha}-\alpha_0|+\|x(\tilde{\beta}-\beta_0)\|_{2,n}\lesssim\sqrt{s\log(p\vee n)/n}\lesssim\delta_n^{1/2}$. Furthermore, $\|z_{0i}y_i\|_{2,n}\lesssim_{\mathrm{P}}\{\overline{\mathrm{E}}[z_{0i}^4]\overline{\mathrm{E}}[y_i^4]\}^{1/4}$, $\|w_iy_i\|_{2,n}\lesssim_{\mathrm{P}}\overline{\mathrm{E}}[y_i^2]^{1/2}$ and

$$\|(\widehat{w}_i-w_i)y_i\|_{2,n}\leqslant\max_{1\leqslant i\leqslant n}|y_i|\|\widehat{w}_i-w_i\|_{2,n}\lesssim_{\mathrm{P}}K_1\sqrt{\frac{s\log(p\vee n)}{n}}\lesssim\delta_n^{1/2}$$

by Markov inequality and Condition ITob 4.2–(iii). Moreover, $\|\hat{z}_i-z_{0i}\|_{2,n}=\|x_i(\tilde{\eta}-\eta_0)\|_{2,n}\lesssim\sqrt{s\log(p\vee n)/n}\lesssim\delta_n^{1/2}$.

Similarly, for the second term on the right-hand side of (A.2.3) we have

$$\begin{aligned}
|\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i+s_i/\tilde{\gamma}^2]-\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]| &\leqslant |\mathbb{E}_n[\widehat{w}_i\hat{q}_iy_i]-\overline{\mathrm{E}}[w_iq_{0i}y_i]|\\
&\quad+|\mathbb{E}_n[s_i/\tilde{\gamma}^2]-\overline{\mathrm{E}}[s_i/\gamma_0^2]|,
\end{aligned}$$

where the first term on the right-hand side can be bounded in the same manner as above:

$$\begin{aligned}|\mathbb{E}_n[\widehat{w}_i\hat{q}_i y_i] - \overline{\mathrm{E}}[w_i q_{0i} y_i]| \;\leqslant\; & \|(\widehat{w}_i - w_i)y_i\|_{2,n}\|\hat{q}_i - q_{0i}\|_{2,n} + \|\widehat{w}_i - w_i\|_{2,n}\|q_{0i}y_i\|_{2,n} \\ & + \|w_i y_i\|_{2,n}\|\hat{q}_i - q_{0i}\|_{2,n} + |\mathbb{E}_n[w_i q_{0i} y_i] - \overline{\mathrm{E}}[w_i q_{0i} y_i]|.\end{aligned}$$

and we have $\|\hat{q}_i - q_{0i}\|_{2,n} = \|(d_i,\, x_i)(\tilde{\theta} - \theta_0)\|_{2,n} \lesssim \sqrt{s\log(p\vee n)/n} \lesssim \delta_n^{1/2}$ and $\|q_{0i}y_i\|_{2,n} \lesssim_{\mathrm{P}} \{\overline{\mathrm{E}}[q_{0i}^4]\overline{\mathrm{E}}[y_i^4]\}^{1/4}$ by the fourth moment restriction. To bound the second term, note that

$$|\mathbb{E}_n[s_i/\tilde{\gamma}^2] - \overline{\mathrm{E}}[s_i/\gamma_0^2]| \leqslant \mathbb{E}_n[s_i]|\tilde{\gamma}^{-2} - \gamma_0^{-2}| + |\mathbb{E}_n[s_i] - \overline{\mathrm{E}}[s_i]|/\gamma_0^2 \lesssim_{\mathrm{P}} \delta_n^{1/2},$$

where $|\tilde{\gamma}^{-2} - \gamma_0^{-2}| = |\gamma_0/\tilde{\gamma} - \tilde{\gamma}/\gamma_0|/(\tilde{\gamma}\gamma_0) \leqslant \underline{\mathrm{c}}^{-2}|(\gamma_0 - \tilde{\gamma})/\tilde{\gamma} - (\tilde{\gamma} - \gamma_0)/\gamma_0| \lesssim |\gamma_0 - \tilde{\gamma}| \lesssim n^{-1/4}\delta_n^{1/4}$ because of the post-ℓ_1 rate above, Chebyshev's inequality and since $\{\tilde{\gamma} \wedge \gamma_0\} \geqslant \underline{\mathrm{c}}$ by the comment in Step 2 (b) above and Condition ITob 4.2–(i).

Thus, with probability $1 - o(1)$ we have

$$\begin{aligned}|\tilde{\gamma} - \gamma_0|\,|\hat{\mu} - \mu_0| \lesssim \sqrt{\frac{s\log(p\vee n)}{n}}\sqrt{\frac{s\log(p\vee n)}{n}} \;&=\; n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2} \\ |\hat{\mu} - \mu_0|\left\{\overline{\mathrm{E}}[(\tilde{q}_i - q_{0i})^2]\big|_{\tilde{q}=\hat{q}}\right\}^{1/2} \;&\lesssim\; n^{-1/2}\frac{s\log(p\vee n)}{\sqrt{n}} \lesssim n^{-1/2}\delta_n^{1/2}.\end{aligned}$$

Step 5: (Verification of Condition ITob 4.1–(iv))
Next, we verify the remaining conditions in ITob 4.1–(iv). By the same argument as above, we have $\|\{\widehat{w}_i - w_i\}d_i\|_{2,n} \leqslant \max_{1\leqslant i\leqslant n}|d_i|\|\widehat{w}_i - w_i\|_{2,n} \lesssim_{\mathrm{P}} K_1\sqrt{s\log(p\vee n)/n} \lesssim \delta_n^{1/2}$. Additionally, $\|x_i\{\tilde{\beta} - \beta_0\}(z_{0i} - \mu_0 q_{0i})\|_{2,n} \lesssim \max_{1\leqslant i\leqslant n}\|(z_{0i},\, q_{0i})\|_\infty\, \|x_i\{\tilde{\beta} - \beta_0\}\|_{2,n} \lesssim_{\mathrm{P}} K_1\sqrt{s\log(p\vee n)/n} \lesssim \delta_n^{1/2}$ since $|\mu_0| \leqslant C$ in (A.1.6). $\|d_i(\hat{z}_i - \hat{\mu}\hat{q}_i)\|_{2,n} \leqslant \|d_i^2\|_{2,n} + \|d_i^2\|_{2,n}^{1/2} \cdot \|\{x_i\tilde{\eta}\}^2\|_{2,n}^{1/2} + \hat{\mu}^2\|d_i^2\|_{2,n}^{1/2}\, \|y_i^2\|_{2,n}^{1/2} + \hat{\mu}^2\|d_i^2\|_{2,n}^{1/2}\, \|\{\tilde{x}_i\tilde{\theta}\}^2\|_{2,n}^{1/2} \lesssim_{\mathrm{P}} C + \delta_n$ and similarly $\|y_i(z_{0i} - \mu_0 q_{0i})\|_{2,n} \leqslant \|y_i^2\|_{2,n}^{1/2}\, \|d_i^2\|_{2,n}^{1/2} + \|y_i^2\|_{2,n}^{1/2}\|\{x_i\eta_0\}^2\|_{2,n}^{1/2} + \mu_0^2\|y_i^2\|_{2,n}^{1/2} + \mu_0^2\|y_i^2\|_{2,n}^{1/2}\, \|\{\tilde{x}_i\theta_0\}^2\|_{2,n}^{1/2} \lesssim_{\mathrm{P}} C$ by the fourth moment restriction with probability $1 - o(1)$.

Step 6: (Verification of Condition ITob 4.1–(iii) (1.4.2))
For this paragraph define $g_i(\alpha) := g(\gamma_0 y_i - \alpha d_i - x_i\beta_0)$ and $\hat{g}_i(\alpha) := g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})$.

Note that

$$
\begin{aligned}
&\sup_{\alpha\in\mathcal{A}}\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[\psi(m_i,\alpha,\beta_0,\gamma_0,z_{0i},q_{0i},\mu_0)-\psi(m_i,\alpha,\hat{\beta},\hat{\gamma},\hat{z}_i,\hat{q}_i,\hat{\mu})\right]\right| \\
&\quad=\sup_{\alpha\in\mathcal{A}}\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[g_i(\alpha)\{z_{0i}-\mu_0 q_{0i}\}-\hat{g}_i(\alpha)\{\hat{z}_i-\hat{\mu}\hat{q}_i\}-\mu_0 s_i/\gamma_0+\hat{\mu}s_i/\tilde{\gamma}\right]\right| && \text{(A.2.4)}\\
&\quad\leqslant\sup_{\alpha\in\mathcal{A}}\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[\{\hat{g}_i(\alpha)-g_i(\alpha)\}(\hat{z}_i-z_{0i})\right]\right| && \text{(A.2.5)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[g_i(\alpha)\{\hat{z}_i-z_{0i}\}\right]\right| && \text{(A.2.6)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[\{\hat{g}_i(\alpha)-g_i(\alpha)\}z_{0i}\right]\right| && \text{(A.2.7)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}|\hat{\mu}|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[\{g_i(\alpha)-\hat{g}_i(\alpha)\}(\hat{q}_i-q_{0i})\right]\right| && \text{(A.2.8)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}|\hat{\mu}|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[g_i(\alpha)\{\hat{q}_i-q_{0i}\}\right]\right| && \text{(A.2.9)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}|\hat{\mu}|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[\{\hat{g}_i(\alpha)-g_i(\alpha)\}q_{0i}\right]\right| && \text{(A.2.10)}\\
&\qquad+\sup_{\alpha\in\mathcal{A}}|\hat{\mu}-\mu_0|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})\left[g_i(\alpha)q_{0i}\right]\right| && \text{(A.2.11)}\\
&\qquad+|\hat{\mu}-\mu_0|\cdot|\gamma_0^{-1}-\tilde{\gamma}^{-1}|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})[s_i]\right| && \text{(A.2.12)}\\
&\qquad+|\gamma_0^{-1}-\tilde{\gamma}^{-1}|\cdot|\mu_0|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})[s_i]\right| && \text{(A.2.13)}\\
&\qquad+|\mu_0-\hat{\mu}|\cdot\left|(\mathbb{E}_n-\overline{\mathrm{E}})[s_i/\gamma_0]\right|. && \text{(A.2.14)}
\end{aligned}
$$

As $0\leqslant|g_i(\alpha)|\leqslant 1$, $0\leqslant|\hat{g}_i(\alpha)|\leqslant 1$ and $g_i(\alpha)$, $\hat{g}_i(\alpha)$ are $1/2$-Lipschitz, we use the approach in Belloni et al. (2016b) to bound (A.2.5) $\lesssim n^{-1/2}\delta_n^{-1/2}$, (A.2.6) $\lesssim n^{-1/2}\delta_n^{-1/2}$, and (A.2.7) $\lesssim n^{-1/2}\delta_n^{-1/2}$. In fact, we again make use of their approach to bound (A.2.8), (A.2.9), and (A.2.10). Below, we restate their proofs by applying their steps to (A.2.8), (A.2.9), and (A.2.10) and emphasize that they analogously apply to (A.2.5), (A.2.6), and (A.2.7).

Firstly, note that $\hat{\mu}$ is bounded by $|\hat{\mu}|\leqslant|\mu_0|+|\hat{\mu}-\mu_0|\leqslant C+\delta_n$ and the boundedness of μ_0 with probability $1-o(1)$. Furthermore, because $|\hat{g}_i(\alpha)-g_i(\alpha)|\leqslant|y_i(\tilde{\gamma}-\gamma_0)|+|x_i(\tilde{\beta}-\beta_0)|$, $|\hat{q}_i-q_{0i}|=|(d_i,\,x_i)(\tilde{\theta}-\theta_0)|$, $\|\tilde{\beta}\|_0+\|\tilde{\theta}\|_0\lesssim s$, and $\phi^u_{\max,n}(2Cs)$ is uniformly bounded, we have by the Cauchy-Schwarz inequality that

$$(A.2.8)\lesssim\|\tilde{\beta}-\beta_0\|\,\|\tilde{\theta}-\theta_0\|\lesssim s\log(p\vee n)/n\lesssim n^{-1/2}\delta_n^{-1/2}$$

with probability $1-o(1)$ under the assumed growth condition $s^2\log^2(p\vee n)\leqslant\delta_n n$.

To bound (A.2.9), we use that with probability $1-o(1)$, $\|\tilde{\theta}-\theta_0\|_1\leqslant Cs\sqrt{\log(p\vee n)/n}$

such that with the same probability

$$
\begin{aligned}
(A.2.9) \;\leqslant\; & \sup_{\alpha\in\mathcal{A}} |\hat{\mu}| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[\{g_i(\alpha) - g_i(\alpha_0)\}(d_i,\, x_i)(\tilde{\theta} - \theta_0)\big]\big| \\
& \qquad + |\hat{\mu}| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[g_i(\alpha_0)(d_i,\, x_i)(\tilde{\theta} - \theta_0)\big]\big| \\
\lesssim_{\mathrm{P}} & \sup_{\alpha\in\mathcal{A}, \|\xi\|_1 = Cs\sqrt{\log(p\vee n)/n}} \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[\{g_i(\alpha) - g_i(\alpha_0)\}(d_i,\, x_i)\xi\big]\big| \\
& \qquad + \sup_{\|\xi\|_1 = Cs\sqrt{\log(p\vee n)/n}} \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[g_i(\alpha_0)(d_i,\, x_i)\xi\big]\big|.
\end{aligned}
$$

Both terms can be separately bounded using Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b). To handle the first term let $W_{ij} = d_i\tilde{x}_{ij}$. Define $r_1 = Cs\sqrt{\log(p\vee n)/n}$, $\mathcal{T}_+ = \{(\alpha - \alpha_0)\xi \in \mathbb{R}^{p+1} : \alpha \geqslant \alpha_0, \alpha \in \mathcal{A}, \|\xi\|_1 = r_1\}$ and $\mathcal{T}_- = \{(\alpha - \alpha_0)\xi \in \mathbb{R}^{p+1} : \alpha < \alpha_0, \alpha \in \mathcal{A}, \|\xi\|_1 = r_1\}$. For $t \in \mathcal{T}_+$ we define $h_i^+(t) = \{g_i(\alpha_0 + \|t\|_1/r_1) - g_i(\alpha_0)\}\tilde{x}_i t r_1/\|t\|_1$. By construction we have $|h_i^+(t)| \leqslant |W_i t| = |(\alpha - \alpha_0)d_i\tilde{x}_i\xi|$. Similarly, for $t \in \mathcal{T}_-$ we define $h_i^-(t) = \{g_i(\alpha_0 - \|t\|_1/r_1) - g_i(\alpha_0)\}\tilde{x}_i t r_1/\|t\|_1$. Therefore, we have

$$
\begin{aligned}
& \sup_{\alpha\in\mathcal{A}, \|\xi\|_1 = Cs\sqrt{\log(p\vee n)/n}} \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[\{g_i(\alpha) - g_i(\alpha_0)\}(d_i,\, x_i)\xi\big]\big| \\
& \qquad \leqslant \sup_{t\in\mathcal{T}_+} \big|(\mathbb{E}_n - \overline{\mathrm{E}})[h_i^+(t)]\big| + \sup_{t\in\mathcal{T}_-} \big|(\mathbb{E}_n - \overline{\mathrm{E}})[h_i^-(t)]\big| \\
& \qquad \lesssim s\sqrt{\frac{\log(p\vee n)}{n}}\sqrt{\frac{\log(p\vee n)}{n}} \lesssim n^{-1/2}\delta_n^{-1/2},
\end{aligned}
$$

where we set $K^2 = C\log(p\vee n)$, $M \lesssim C$ and $\|\mathcal{T}\|_1 \lesssim s\sqrt{\log(p\vee n)/n}$ in Lemma 5 in Belloni et al. (2016b), because $\max_{1\leqslant j\leqslant p+1} \mathbb{E}_n[d_i^2\tilde{x}_{ij}^2] \leqslant \mathbb{E}_n[d_i^4] \vee \max_{1\leqslant j\leqslant p}\{\mathbb{E}_n[x_{ij}^4]\mathbb{E}_n[d_i^4]\}^{1/2}$ is bounded with probability $1 - o(1)$ under $K_4^4\log(p) \leqslant \delta_n n$ by Lemma A.6.2, i.e., Lemma 3 in Belloni et al. (2016b) and $\max_{1\leqslant j\leqslant p+1} \overline{\mathrm{E}}[d_i^2\tilde{x}_{ij}^2] \leqslant \overline{\mathrm{E}}[d_i^4] \vee \max_{1\leqslant j\leqslant p}\{\overline{\mathrm{E}}[x_{ij}^4]\overline{\mathrm{E}}[d_i^4]\}^{1/2} \leqslant C$ by the fourth moment condition.

To bound the second term, we again use Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b) with $t = \xi$ and $h_i(t) = g_i(\alpha_0)(d_i,\, x_i)\xi$, which satisfies $|h_i(t) \leqslant |W_i t| = |\tilde{x}_i t|$ and again $\|\mathcal{T}\|_1 = s\sqrt{\log(p\vee n)/n}$. Thus, we have with probability $1 - o(1)$

$$
(A.2.9) \;\lesssim\; s\sqrt{\frac{\log(p\vee n)}{n}}\sqrt{\frac{\log(p\vee n)}{n}} + s\sqrt{\frac{\log(p\vee n)}{n}}\sqrt{\frac{\log(p\vee n)}{n}} \lesssim n^{-1/2}\delta_n^{-1/2}
$$

under $s^2\log^2(p\vee n) \leqslant \delta_n n$.

To bound (A.2.10), we consider the class of functions pertaining to $\{\hat{g}_i(\alpha) - g_i(\alpha)\}q_{0i}$, specifically

$$
\mathcal{F} = \left\{ \begin{array}{ll} g(\gamma y_i - \alpha d_i - x_i\beta)q_{0i} & : \\ \quad -g(\gamma_0 y_i - \alpha d_i - x_i\beta_0)q_{0i} & : \end{array} \begin{array}{l} \|\beta\|_0 \leqslant Cs, \\ \|(\beta^T,\, \gamma)^T - (\beta_0^T,\, \gamma_0)^T\| \leqslant C\sqrt{s\log(p\vee n)/n} \end{array} \right\}
$$

for some suitably large constant C. Let $h_i(t,\alpha) = g((\gamma + t_y)y_i - \alpha d_i - x_i(\beta + t_x)) - g(\gamma_0 y_i - \alpha d_i - x_i\beta_0)$, where $t = (t_x^T,\, t_y)^T$, so that $|h_i(t,\alpha)| \leqslant |q_{0i}(-x_i,\, y_i)t|$. Therefore $\|\mathcal{T}\|_1 \lesssim s\sqrt{\log(p \vee n)/n}$ and note that $\mathbb{E}_n[y_i^2 q_{0i}^2] \vee \max_{1\leqslant i\leqslant n} \mathbb{E}_n[x_{ij}^2 q_{0i}^2] \lesssim C$ with probability $1 - o(1)$ under $K_4^4 \log(p) \leqslant \delta_n n$ and $\overline{\mathrm{E}}[y_i^2 q_{0i}^2] \vee \max_{1\leqslant i\leqslant n} \overline{\mathrm{E}}[x_{ij}^2 q_{0i}^2] \leqslant \{\overline{\mathrm{E}}[y_i^4]\overline{\mathrm{E}}[q_{0i}^4]\}^{1/2} \vee \max_{1\leqslant i\leqslant n}\{\overline{\mathrm{E}}[x_{ij}^4[\overline{\mathrm{E}}[q_{0i}^4]\}^{1/2} \leqslant C$ is bounded by the fourth moment condition. By Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b), we have

$$(A.2.10) \;\lesssim\; \sqrt{\frac{\log(p \vee n)}{n}}\|\mathcal{T}\|_1 \lesssim \frac{s\log(p \vee n)}{n} \lesssim n^{-1/2}\delta_n^{1/2}$$

with probability $1 - o(1)$ under the assumed growth condition $s^2 \log^2(p \vee n) \leqslant \delta_n n$.

To show the required rate on (A.2.11), note that with probability $1 - o(1)$ we have

$$\begin{aligned}\sup_{\alpha\in\mathcal{A}} |\hat{\mu} - \mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[g_i(\alpha)q_{0i}\big]\big| \;&\leqslant\; \sup_{\alpha\in\mathcal{A}} |\hat{\mu} - \mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[\{g_i(\alpha) - g_i(\alpha_0)\}q_{0i}\big]\big| \\ &\quad + |\hat{\mu} - \mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})\big[g_i(\alpha_0)q_{0i}\big]\big| \\ &\lesssim\; n^{-1/2}\delta_n + n^{-1/2}\delta_n \log^{-1}(n)\end{aligned}$$

where the first term can be bounded using Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b), since $|g_i(\alpha) - g_i(\alpha_0)| \leqslant |\alpha - \alpha_0| d_i$ because $g_i(\alpha)$ is $1/2$-Lipschitz and $\overline{\mathrm{E}}[d_i^2 q_{0i}^2] \leqslant \{\overline{\mathrm{E}}[d_i^4]\overline{\mathrm{E}}[q_{0i}^4]\}^{1/2} \leqslant C$ by the fourth moment condition. Additionally, the second term converges at $\sqrt{n}$ rate by Chebyshev's inequality and we have $|\hat{\mu} - \mu_0| \leqslant \delta_n$ by the result derived above.

To bound (A.2.12), (A.2.13), and (A.2.14) note that for some $\varepsilon > 0$ we have $\mathrm{P}(|\mathbb{E}_n[s_i] - \overline{\mathrm{E}}[s_i]| \geqslant \varepsilon) \leqslant 1/(n\varepsilon^2)$ by Chebyshev's inequality. Furthermore, since $|\mu_0| \leqslant C$, $|\hat{\mu} - \mu_0| \leqslant \delta_n^{-1/2}$ and $|\tilde{\gamma}^{-1} - \gamma_0^{-1}| = |\tilde{\gamma} - \gamma_0|/(\tilde{\gamma}\gamma_0) \lesssim \delta_n^{-1/2}$ as $\gamma_0 \wedge \tilde{\gamma} \geqslant \underline{c} > 0$, we have that (A.2.12) $\lesssim n^{-1/2}\delta_n^{-1/2}$, (A.2.13) $\lesssim n^{-1/2}\delta_n^{-1/2}$, and (A.2.14) $\lesssim n^{-1/2}\delta_n^{-1/2}$.

Step 7: (Verification of Condition ITob 4.1–(iii) (1.4.3))
To show that our set of primitive assumptions also imply requirement (1.4.3), we again proceed analogously to Belloni et al. (2016b) and follow their steps by replacing the estimating function for α_0. Firstly, note that $\mathcal{A} = \{\alpha : |\alpha - \tilde{\alpha}| \leqslant C\log^{-1}(n)\} \supseteq \{\alpha : |\alpha - \alpha_0| \leqslant (C/2)\log^{-1}(n)\}$ for n large enough since $|\tilde{\alpha} - \alpha_0| \lesssim \sqrt{s\log(p \vee n)/n}$ with probability $1 - o(1)$. In particular, $\check{\alpha} \in \mathcal{A}$ implies that $|\check{\alpha} - \alpha_0| \lesssim \log^{-1}(n)$ with probability $1 - o(1)$ for sufficiently large n. Below, we demonstrate that $\mathbb{E}_n[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma}]$ changes its sign over $\alpha \in \mathcal{A}$ with probability $1 - o(1)$, which by the continuity of function $\hat{g}_i(\alpha)$ implies the existence of the zero $\mathbb{E}_n[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma}] = 0$ with probability $1 - o(1)$. To this

end, note that

$$\begin{aligned}
&\mathbb{E}_n[\hat{g}_i(\alpha)(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}] \\
&\quad = (\mathbb{E}_n - \overline{\mathrm{E}})[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma} - g_i(\alpha)\{z_{0i} - \mu_0 q_{0i}\} + \mu_0 s_i/\gamma_0] && \text{(A.2.15)} \\
&\qquad + \overline{\mathrm{E}}[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma}] - \overline{\mathrm{E}}[g_i(\alpha)\{z_{0i} - \mu_0 q_{0i}\} - \mu_0 s_i/\gamma_0] && \text{(A.2.16)} \\
&\qquad + (\mathbb{E}_n - \overline{\mathrm{E}})[g_i(\alpha)\{z_{0i} - \mu_0 q_{0i}\} - \mu_0 s_i/\gamma_0] && \text{(A.2.17)} \\
&\qquad + \overline{\mathrm{E}}[g_i(\alpha)\{z_{0i} - \mu_0 q_{0i}\} - \mu_0 s_i/\gamma_0]. && \text{(A.2.18)}
\end{aligned}$$

By Condition ITob 4.1–(iii) (1.4.2), we have that $|(A.2.15)| \leqslant (A.2.4) \lesssim n^{-1/2}\delta_n^{1/2}$ with probability $1 - o(1)$. Furthermore, by expansion results (A.1.11), $\overline{\mathrm{E}}[g_i(\alpha)\{z_{0i} - \mu_0 q_{0i}\} - \mu_0 s_i/\gamma_0] = \Gamma(\alpha, h_0) = \overline{\mathrm{E}}[w_i z_{0i} d_i](\alpha - \alpha_0) + O(\delta_n|\alpha - \alpha_0| + (\alpha - \alpha_0)^2)$, and (A.1.14), $\overline{\mathrm{E}}[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma}] = \Gamma(\alpha, \hat{h}) = \overline{\mathrm{E}}[w_i z_{0i} d_i](\alpha - \alpha_0) + O(\delta_n|\alpha - \alpha_0| + n^{-1/2}\delta_n)$, we have with probability $1 - o(1)$ that $|(A.2.16)| \lesssim \delta_n|\alpha - \alpha_0| + n^{-1/2}\delta_n$. Moreover,

$$\begin{aligned}
|(A.2.17)| \;&\leqslant\; \sup_{\alpha\in\mathcal{A}} \big|(\mathbb{E}_n - \overline{\mathrm{E}})[\{g_i(\alpha) - g_i(\alpha_0)\}z_{0i}]\big| + \big|(\mathbb{E}_n - \overline{\mathrm{E}})[g_i(\alpha_0)z_{0i}]\big| \\
&\qquad + \sup_{\alpha\in\mathcal{A}} |\mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})[\{g_i(\alpha) - g_i(\alpha_0)\}q_{0i}]\big| + |\mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})[g_i(\alpha_0)q_{0i}]\big| \\
&\qquad + |\mu_0| \cdot \big|(\mathbb{E}_n - \overline{\mathrm{E}})[s_i]/\gamma_0\big| \\
&\lesssim\; n^{-1/2}\delta_n + n^{-1/2}\log^{-1}(n),
\end{aligned}$$

where the first and third term on the right-hand side are bounded by Lemma A.6.4, i.e., Lemma 5 in Belloni et al. (2016b) using $|g_i(\alpha) - g_i(\alpha_0)| \leqslant |\alpha - \alpha_0|d_i$, $|\alpha - \alpha_0| \lesssim \log^{-1}(n)$ for $\alpha \in \mathcal{A}$, and $\overline{\mathrm{E}}[d_i^2 z_{0i}^2] \leqslant \{\overline{\mathrm{E}}[d_i^4]\overline{\mathrm{E}}[z_{0i}^4]\}^{1/2} \leqslant C$, $\overline{\mathrm{E}}[d_i^2 q_{0i}^2] \leqslant \{\overline{\mathrm{E}}[d_i^4]\overline{\mathrm{E}}[q_{0i}^4]\}^{1/2} \leqslant C$ are bounded by the fourth moment restriction. The second, fourth and fifth term converge at $\sqrt{n}$ rate by Chebyshev's inequality and because $|\mu_0| \leqslant C$ is bounded by (A.2.3).

Thus, as $\overline{\mathrm{E}}[g_i(\alpha)(z_{0i} - \mu_0 q_{0i}) - \mu_0 s_i/\gamma_0] = \overline{\mathrm{E}}[w_i z_{0i}^2](\alpha - \alpha_0) + O(|\alpha - \alpha_0|^2)$, we have with probability $1 - o(1)$

$$\begin{aligned}
\mathbb{E}_n[\hat{g}_i(\alpha)\{\hat{z}_i - \hat{\mu}\hat{q}_i\} - \hat{\mu}s_i/\tilde{\gamma}] \;&=\; O\big(n^{-1/2}\log^{-1}(n) \\
&\qquad + \delta_n|\alpha - \alpha_0|\big) + \overline{\mathrm{E}}[g_i(\alpha)\{z_{0i} - \hat{\mu}q_{0i}\} - \mu_0 s_i/\gamma_0] \qquad \text{(A.2.19)} \\
&=\; O\big(n^{-1/2}\log^{-1}(n)\big) \\
&\qquad + (\alpha - \alpha_0)\{\overline{\mathrm{E}}[w_i z_{0i}^2] + O(\delta_n)\} + O(|\alpha - \alpha_0|^2).
\end{aligned}$$

As $\overline{\mathrm{E}}[w_i z_{0i}^2] \geqslant \underline{c} > 0$ and $\delta_n \searrow 0$, we obtain different signs when evaluating (A.2.19) at the bounds $\underline{\alpha} = \inf\mathcal{A}$ and $\overline{\alpha} = \sup\mathcal{A}$, because $\underline{\alpha} - \alpha_0 < 0$ and $\overline{\alpha} - \alpha_0 > 0$ and $|\alpha^\diamond - \alpha_0| \geqslant (C/2)\log^{-1}(n)$ for $\alpha^\diamond \in \{\underline{\alpha}, \overline{\alpha}\}$ if n is sufficiently large.

Step 8: (Application to Instrument Estimator 1.2.1)
Let $\hat{\mathcal{J}} = \text{support}(\hat{\beta})$. By the first-order conditions of the logistic Tobit loss minimization in

Step 1 of Algorithm 1.2.1, we have

$$\mathbb{E}_n\big[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})(d_i,\, x_{i\hat{\mathcal{J}}}),\, g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})y_i + s_i/\tilde{\gamma}\big] = 0. \tag{A.2.20}$$

To construct a suitable instrument $\hat{q}_i$, define

$$\tilde{\theta} \;\in\; \arg\min_{\theta} \mathbb{E}_n\big[\widehat{w}_i\{y_i - (d_i,\, x_i)\theta\}^2\big] \;:\; \text{support}(\theta) \;\subseteq\; \text{support}(\hat{\beta}),$$

and compute $\hat{q}_i = y_i - (d_i,\, x_i)\tilde{\theta}$ and $\hat{\mu} = \mathbb{E}_n[\widehat{w}_i\hat{q}_i y_i + s_i/\tilde{\gamma}^2]^{-1}\mathbb{E}_n[\widehat{w}_i\hat{z}_i y_i]$. Choosing the linear combination $(-\tilde{\theta}^T,\, 1)^T$ of (A.2.20), we ensure that $\hat{q}_i$ satisfies the in-sample orthogonality condition $\mathbb{E}_n[\widehat{w}_i\hat{q}_i(d_i,\, x_i)] = 0$. Moreover, note that by a first-order Taylor expansion with second-order Lagrange remainder about the first Step estimate $\tilde{\alpha}$, there exists some point $\alpha^\diamond \in [\alpha, \tilde{\alpha}]$ such that

$$\begin{aligned}\mathbb{E}_n[\hat{g}_i(\alpha)(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}] \;&=\; \mathbb{E}_n[\hat{g}_i(\alpha)\hat{z}_i] - \hat{\mu}\,\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})\hat{q}_i + s_i/\tilde{\gamma}] \\ &\quad - \hat{\mu}\,\mathbb{E}_n[\widehat{w}_i\hat{q}_i d_i](\alpha - \tilde{\alpha}) \\ &\quad - \hat{\mu}\,\mathbb{E}_n[g''(\tilde{\gamma}y_i - \alpha^\diamond d_i - x_i\tilde{\beta})\hat{q}_i d_i^2](\alpha - \tilde{\alpha})^2.\end{aligned}$$

By the first-order conditions of Step 1, we have $\mathbb{E}_n[g(\tilde{\gamma}y_i - \tilde{\alpha}d_i - x_i\tilde{\beta})\hat{q}_i + s_i/\tilde{\gamma}] = 0$. Due to in-sample orthogonality, we have $\mathbb{E}_n[\widehat{w}_i\hat{q}_i d_i] = 0$. Furthermore, by the boundedness of $|g''(t)| \leqslant \bar{L}''$, the results of Steps 2–5, $|\hat{\mu}| \leqslant |\mu_0| + |\hat{\mu} - \mu_0| \lesssim C + \delta_n$, $|\hat{q}_i| \leqslant |q_{0i}| + |\hat{q}_i - q_{0i}|$, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\big|\mathbb{E}_n[\hat{g}_i(\alpha)(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}]\big| \;&\lesssim\; \big|\mathbb{E}_n[\hat{g}_i(\alpha)\hat{z}_i]\big| \\ &\quad + \{\mathbb{E}_n[q_{0i}^2]^{1/2} + \|\hat{q}_i - q_{0i}\|_{2,n}\}\mathbb{E}_n[d_i^4]^{1/2}(\alpha - \tilde{\alpha})^2.\end{aligned}$$

Therefore, as $\alpha \in \mathcal{A}$, both instrumental Tobit estimators in Step 3 of Algorithm 1.2.1 are equivalent in the sense that

$$\mathbb{E}_n[\hat{g}_i(\alpha)(\hat{z}_i - \hat{\mu}\hat{q}_i) - \hat{\mu}s_i/\tilde{\gamma}] = \mathbb{E}_n[\hat{g}_i(\alpha)\hat{z}_i] + O\big(n^{-1/2}\log^{-1}(n)\big).$$

In particular, by the same logic underlying Step 7 above, $\mathbb{E}_n[g(\tilde{\gamma}y_i - \alpha d_i - x_i\tilde{\beta})\hat{z}_i]$ takes a zero at $\check{\alpha}$ for sufficiently large n. ■

A.3. PROOF OF THEOREM 1.4.3

Let $\check{\mathcal{J}} = \text{support}(\hat{\beta}) \cup \text{support}(\hat{\eta}) \cup \text{support}(\hat{\theta})$. By the first-order conditions of the likelihood loss minimization in Step 3 of Algorithm 1.2.2, we have

$$\mathbb{E}_n\big[g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})(d_i,\, x_{i\check{\mathcal{J}}}),\, g(\check{\gamma}y_i - \check{\alpha}d_i - x_i\check{\beta})y_i + s_i/\check{\gamma}\big] = 0. \tag{A.3.1}$$

To construct suitable instruments $\hat{z}_i$ and $\hat{q}_i$, define

$$\begin{aligned}\hat{\eta}^* &\in \arg\min_{\eta} \big\|x_i(\eta-\eta_0)\big\|_{2,n} \,:\, \text{support}(\eta)\subseteq\check{\mathcal{J}},\\ \hat{\theta}^* &\in \arg\min_{\theta} \big\|(d_i,\,x_i)(\theta-\theta_0)\big\|_{2,n} \,:\, \text{support}(\theta)\subseteq\check{\mathcal{J}}.\end{aligned}$$

We apply the results in the proof of Theorem 1.4.2, using $z_{0i}=d_i-x_i\eta_0$, $q_{0i}=y_i-(d_i,\,x_i)\theta_0$, $\mu_0=\overline{\mathrm{E}}[w_iq_{0i}y_i+s_i/\gamma_0^2]^{-1}\overline{\mathrm{E}}[w_iz_{0i}y_i]$ and the estimated instruments $\hat{z}_i=d_i-x_i\hat{\eta}^*$, $\hat{q}_i=y_i-(d_i,\,x_i)\hat{\theta}^*$, $\check{\mu}=\mathbb{E}_n[\check{w}_i\hat{q}_iy_i+s_i/\check{\gamma}^2]^{-1}\mathbb{E}_n[\check{w}_i\hat{z}_iy_i]$. Note that by (A.3.1), taking the specific linear combination $(\{1+\check{\mu}\hat{\theta}_1^*\},\,\{\check{\mu}\hat{\theta}_{\setminus 1}^{*T}-\hat{\eta}^{*T}\},\,-\check{\mu})^T$ of the optimality conditions, we have

$$\mathbb{E}_n[g(\check{\gamma}y_i-\check{\alpha}d_i-x_i\check{\beta})(\hat{z}_i-\check{\mu}\hat{q}_i)-\check{\mu}s_i/\check{\gamma}]=0.$$

Therefore, the post-double selection $\check{\alpha}$ minimizes the criterion

$$\frac{\mathbb{E}_n[g(\check{\gamma}y_i-\alpha d_i-x_i\check{\beta})(\hat{z}_i-\check{\mu}\hat{q}_i)-\check{\mu}s_i/\check{\gamma}]^2}{\mathbb{E}_n[\{g(\check{\gamma}y_i-\alpha d_i-x_i\check{\beta})(\hat{z}_i-\check{\mu}\hat{q}_i)-\check{\mu}s_i/\check{\gamma}\}^2]}$$

over $\alpha\in\mathbb{R}$.

Convergence rates for the ℓ_1-penalized logistic Tobit, the post-selection logistic Tobit, post-Lasso with estimated weights and associated sparsity bounds are established in Steps 2–5 of the proof of Theorem 1.4.2. Therefore, we have with probability $1-o(1)$ that $\|\hat{\eta}\|_0\lesssim s$, $\|\hat{\theta}\|_0\lesssim s$, $\|\hat{\beta}\|_0\lesssim s$, $\Lambda(\widehat{\Theta})-\Lambda(\Theta_0)\lesssim s\log(p)/n$, $\|\tilde{\eta}-\eta_0\|\leqslant\sqrt{s\log(p\vee n)/n}$ and $\|\tilde{\theta}-\theta_0\|\leqslant\sqrt{s\log(p\vee n)/n}$. The results for Steps 1 and 2 of the instrument estimator therefore carry over to the post-double selection estimator.

Concerning Step 3, the sparsity bounds above imply that $|\check{\mathcal{J}}|\lesssim s$ with probability $1-o(1)$. In particular, since support$(\hat{\beta})\subset\check{\mathcal{J}}$, it holds that

$$\Lambda(\check{\Theta})-\Lambda(\Theta_0)\leqslant\Lambda(\hat{\Theta})-\Lambda(\Theta_0)\lesssim s\log(p)/n.$$

As a result, we can invoke Lemma A.6.14. As the sparse eigenvalues of order $k=s\ell_n$ for some sequence of constants $\ell_n\to\infty$ are bounded from below, Lemma A.6.14 provides the convergence rate of the post-selection logistic Tobit estimator $\|\check{\Theta}-\Theta_0\|\lesssim\sqrt{s\log(p)/n}$, $\|\check{\Theta}-\Theta_0\|_1\lesssim_{\mathrm{P}}s\sqrt{\log(p)/n}$. Furthermore, since support$(\hat{\eta})\subset\check{\mathcal{J}}$ and support$(\hat{\theta})\subset\check{\mathcal{J}}$, we have $\|\hat{z}_i-z_{0i}\|_{2,n}=\|x_i(\hat{\eta}^*-\eta_0)\|_{2,n}\leqslant\|x_i(\hat{\eta}-\eta_0)\|_{2,n}\lesssim\sqrt{s\log(p\vee n)/n}$, $\|\hat{\eta}^*-\eta_0\|_1\lesssim_{\mathrm{P}}\sqrt{s}\|\hat{\eta}^*-\eta_0\|\lesssim\sqrt{s}\|x_i(\hat{\eta}^*-\eta_0)\|_{2,n}/\{\phi^u_{\min,n}(C's)\}^{1/2}\lesssim s\sqrt{\log(p\vee n)/n}$ and $\|\hat{q}_i-q_{0i}\|_{2,n}=\|(d_i,\,x_i)(\hat{\theta}^*-\theta_0)\|_{2,n}\leqslant\|(d_i,\,x_i)(\hat{\theta}-\theta_0)\|_{2,n}\lesssim\sqrt{s\log(p\vee n)/n}$, $\|\hat{\theta}^*-\theta_0\|_1\lesssim_{\mathrm{P}}\sqrt{s}\|\hat{\theta}^*-\theta_0\|\lesssim\sqrt{s}\|(d_i,\,x_i)(\hat{\eta}^*-\eta_0)\|_{2,n}/\{\phi^u_{\min,n}(C's)\}^{1/2}\lesssim s\sqrt{\log(p\vee n)/n}$ with probability $1-o(1)$. Lastly, by the result of Step 4 in the proof of Theorem 1.4.2, we have $|\check{\mu}-\mu_0|\lesssim|\hat{\mu}-\mu_0|\lesssim\sqrt{s\log(p\vee n)/n}$.

The remaining requirements of Condition ITob 4.2 can be verified as in the proof of Theorem 1.4.2 above. ■

A.4. AUXILIARY RESULTS

A.4.1. Partial Effects in the Logistic Tobit

Consider a data generating process which is independent over indices $i \in \{1, \dots, n\}$ and where

$$\mathrm{E}[\gamma_0 y_i^* \mid d_i,\, x_i] = \alpha_0 d_i + x_i \beta_0.$$

Furthermore, let $s_i = \mathbf{1}\{y_i^* > 0\}$ such that $y_i = s_i y_i^*$. As pointed out by McDonald and Moffitt (1980) and Wooldridge (2010, pp. 521–524), in settings with limited dependent variables, such as household expenditures, labour supply or charitable contributions, a practitioner's interest centers on conditional expectations involving y_i, because the unlimited variable y_i^* is rather hypothetical. By the result of Lemma A.4.1 below, the conditional expectation of observed outcomes in a logistic Tobit model can be expressed as

$$\begin{aligned} \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] &= \alpha_0 d_i + x_i \beta_0 + \mathrm{E}[\gamma_0 u_i \mid \gamma_0 u_i > -\alpha_0 d_i - x_i \beta_0] \\ &= (1 + \exp\{-\alpha_0 d_i - x_i \beta_0\}) \log(1 + \exp\{\alpha_0 d_i + x_i \beta_0\}) \\ &> 0. \end{aligned} \tag{A.4.1}$$

Moreover, be the relation

$$\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] = \mathrm{P}(y_i > 0 \mid d_i,\, x_i) \cdot \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] \tag{A.4.2}$$

we have

$$\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] = \log(1 + \exp\{\alpha_0 d_i + x_i \beta_0\}) > 0, \tag{A.4.3}$$

which lends to the softplus function the surprising and novel interpretation as the conditional expectation of censored outcomes given d_i and x_i. Importantly, equations (A.4.1) and (A.4.3) can be used to derive partial effects of a continuous treatment variable d_i:

$$\partial_{d_i} \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] = \alpha_0 \cdot \underbrace{\left[1 - \frac{\log\left(1 + \exp\{\alpha_0 d_i + x_i \beta_0\}\right)}{\exp\{\alpha_0 d_i + x_i \beta_0\}}\right]}_{:= \mathscr{K}}, \tag{A.4.4}$$

$$\partial_{d_i} \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] = \alpha_0 \cdot \frac{\exp\{\alpha_0 d_i + x_i \beta_0\}}{1 + \exp\{\alpha_0 d_i + x_i \beta_0\}} = \alpha_0 \cdot F(\alpha_0 d_i + x_i \beta_0). \tag{A.4.5}$$

Both additional terms on the right-hand side serve as scaling factors $\mathscr{K} \in [0, 1]$, $F(t) \in [0, 1]$ to the target parameter α_0; see Figure A.1. In particular, the direction of both partial

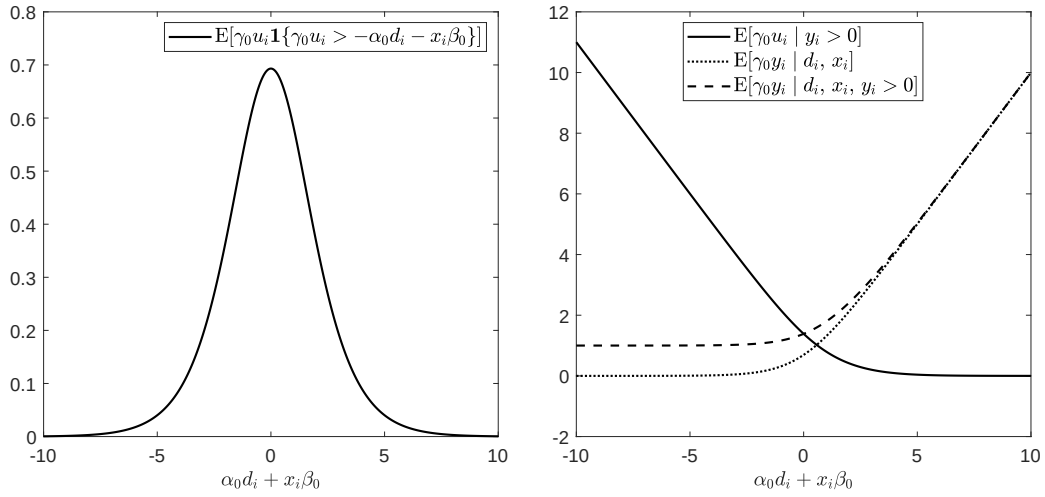


Figure A.1: Conditional expectations of $\gamma_0 u_i$ and $\gamma_0 y_i$ given $y_i > 0$, d_i, and x_i

Note: These graphs show the expectation function $\mathrm{E}[\gamma_0 u_i \mathbf{1}\{\gamma_0 u_i > -\alpha_0 d_i - x_i\beta_0\}]$ in the left-hand panel and the conditional expectation functions $\mathrm{E}[\gamma_0 u_i \mid \gamma_0 u_i > -\alpha_0 d_i - x_i\beta_0]$, $\mathrm{E}[\gamma_0 y_i \mid d_i, x_i, y_i > 0]$, and $\mathrm{E}[\gamma_0 y_i \mid d_i, x_i]$ in the right-hand panel. For example, the solid line in the right-hand panel represents the logistic analogue to the well-known inverse Mill's ratio in Tobin's canonical model. In fact, both functions share certain characteristics such as quasi-linearity for extreme values.

effects of d_i is determined by the sign of α_0. Similarly, the partial effects of a binary treatment variable d_i are obtained by $\mathrm{E}[\gamma_0 y_i \mid d_i = 1,\, x_i,\, y_i > 0] - \mathrm{E}[\gamma_0 y_i \mid d_i = 0,\, x_i,\, y_i > 0]$ and $\mathrm{E}[\gamma_0 y_i \mid d_i = 1,\, x_i] - \mathrm{E}[\gamma_0 y_i \mid d_i = 0,\, x_i]$. Because both (A.4.1) and (A.4.3) are monotonically increasing functions, as illustrated in Figure A.1, the sign of the effect of binary treatment variable d_i equals that of α_0.

For Tobin's canonical model, McDonald and Moffitt (1980) propose the following decomposition based on relation (A.4.2):

$$\begin{aligned} \partial_{d_i} \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i] \;=\; & F(\alpha_0 d_i + x_i\beta_0) \cdot \big\{\partial_{d_i} \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0]\big\} && \text{(A.4.6)} \\ & + \big\{\partial_{d_i} F(\alpha_0 d_i + x_i\beta_0)\big\} \cdot \mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0], && \text{(A.4.7)} \end{aligned}$$

which holds true irrespective of the distributional assumption concerning disturbance u_i. The total change in $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i]$ induced by a marginal change in d_i can be disaggregated into two parts: (A.4.6), the change in the conditional expected value of those above the threshold weighted by the probability of exceeding the threshold, and (A.4.7), the change in the probability of exceeding the threshold weighted by the conditional expectation if being above the limit. The relative magnitudes of these terms have important economic interpretations, depending on the context of the research question; see McDonald and Moffitt (1980).

Lemma A.4.1 *(Conditional Mean Function for Censored Outcomes). The censored outcome variable* $\gamma_0 y_i = \max\{\gamma_0 y_i^*, 0\}$ *obeys model (1.2.1) with logistic cdf* $\mathrm{P}(\gamma_0 u_i \leqslant t) = 1/(1+\exp\{-t\})$. *Then,*

we have that

$$\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] = (1 + \exp\{-\alpha_0 d_i - x_i\beta_0\}) \log(1 + \exp\{\alpha_0 d_i + x_i\beta_0\}).$$

Proof. Define $\tilde{m}_{0i} := \alpha_0 d_i + x_i\beta_0$. Using standard calculus we get

$$\begin{aligned}
\mathrm{E}\big[\gamma_0 u_i \mathbf{1}\{\gamma_0 u_i > -\tilde{m}_{0i}\}\big] &= \int_{-\tilde{m}_{0i}}^{\infty} u \frac{\exp\{u\}}{\big(1+\exp\{u\}\big)^2}\,\mathrm{d}u = \int_{\exp\{-\tilde{m}_{0i}\}}^{\infty} \frac{\log(t)}{(1+t)^2}\,\mathrm{d}t \\
&= \frac{-\tilde{m}_{0i}}{1+\exp\{-\tilde{m}_{0i}\}} + \int_{\exp\{-\tilde{m}_{0i}\}}^{\infty} \frac{1}{(1+t)t}\,\mathrm{d}t \\
&= \frac{-\tilde{m}_{0i}}{1+\exp\{-\tilde{m}_{0i}\}} + \tilde{m}_{0i} + \log(1+\exp\{-\tilde{m}_{0i}\}) \\
&= \tilde{m}_{0i}\frac{\exp\{-\tilde{m}_{0i}\}}{1+\exp\{-\tilde{m}_{0i}\}} + \log(1+\exp\{-\tilde{m}_{0i}\}).
\end{aligned}$$

Since $\mathrm{E}[\gamma_0 u_i \mid \gamma_0 u_i > -\tilde{m}_{0i}] = \mathrm{E}[\gamma_0 u_i \mathbf{1}\{\gamma_0 u_i > -\tilde{m}_{0i}\}]/\mathrm{P}(\gamma_0 u_i > -\tilde{m}_{0i})$, we have

$$\begin{aligned}
\mathrm{E}[\gamma_0 u_i \mid \gamma_0 u_i > -\tilde{m}_{0i}] &= \tilde{m}_{0i}\exp\{-\tilde{m}_{0i}\} + (1+\exp\{-\tilde{m}_{0i}\})\log(1+\exp\{-\tilde{m}_{0i}\}) \\
&= (1+\exp\{-\tilde{m}_{0i}\})\log(1+\exp\{\tilde{m}_{0i}\}) - \tilde{m}_{0i},
\end{aligned}$$

where the last line follows from $\log(1+\exp\{-t\}) = \log(1+\exp\{t\}) - t$. The result follows from $\mathrm{E}[\gamma_0 y_i \mid d_i,\, x_i,\, y_i > 0] = \tilde{m}_{0i} + \mathrm{E}[\gamma_0 u_i \mid \gamma_0 u_i > -\tilde{m}_{0i}]$. ■

A.4.2. Penalties in ℓ_1-regularized Tobits

Without loss of generality, we assume $\|\Theta_0\|_0 = \|(\alpha_0,\, \beta_0^T,\, \gamma_0)^T\|_0 = s \geqslant 1$ and normalized features $\mathbb{E}_n[d_i^2] = 1$, $\mathbb{E}_n[x_{ij}^2] = 1$ for all $j \in \{1, \ldots, p\}$. The parameters $\Theta_0 = (\alpha_0,\, \beta_0^T,\, \gamma_0)^T$ are estimated by a ℓ_1-penalized logistic Tobit:

$$\widehat{\Theta} \in \arg\min_{\Theta} \Lambda(\Theta) + \frac{\lambda}{n}\big\|(\alpha,\, \beta^T)^T\big\|_1. \tag{A.4.8}$$

By a standard principle in the literature on the analysis of ℓ_1-penalized estimators, as found in, e.g., Bickel et al. (2009), Bach (2010), Belloni and Chernozhukov (2011), Belloni et al. (2016a, 2019), Jacobson and Zou (2023a), and Pan and Xie (2023), we can ensure that the logistic Tobit Lasso in (A.4.8) possesses good theoretical properties provided that the penalty level dominates the effective estimation noise:

$$\lambda/n \geqslant c\big\|\nabla_\Theta \Lambda(\alpha_0, \beta_0, \gamma_0)\big\|_\infty \tag{A.4.9}$$

for $c > 1$. As Θ_0 is generally unknown, we content ourselves with setting λ such that the event in (A.4.9) holds with probability $1 - o(1)$, or at least $1 - \Delta$ for some small, fixed

confidence level $\Delta \in (0,1)$. In Remark A.6.8, we provide a discussion on the practical choice for λ.

Algorithm A.4.5 *(Computation of $\widehat{\Psi}$).*

Step 1. Define an iteration limit $\bar{l} \geqslant 1$ and a small tolerance $\varepsilon > 0$. Initialize $l = 1$ and let $\bar{d} := \mathbb{E}_n[\sqrt{\widehat{w}_i} d_i]$. Compute $\|\mathbb{E}_n[\sqrt{\widehat{w}_i} x_i (d_i - \bar{d})]\|_{\infty,k}$, where $\|\xi\|_{\infty,k}$ for some vector $\xi \in \mathbb{R}^p$ and integer $1 \leqslant k \leqslant \bar{k}$ returns the k largest elements by absolute value of this vector. The associated indices $\mathcal{J}^*$ with $|\mathcal{J}^*| = k$ represent the ex-ante, k most promising candidates to join the active set.
Run a weighted least squares estimation of $\sqrt{\widehat{w}_i} d_i$ on $\sqrt{\widehat{w}_i} x_{i\mathcal{J}^*}$:

$$\tilde{\eta}^{[0]} \in \arg\min_{\eta} \mathbb{E}_n\big[\widehat{w}_i (d_i - x_i \eta)^2\big] \ : \ \text{support}(\eta) \ \subseteq \ \mathcal{J}^*.$$

For $j \in \{1, \ldots, p\}$ compute the initial penalty loadings according to

$$\widehat{\Psi}^{[0]}_{j\,j} = \sqrt{\mathbb{E}_n[\widehat{w}_i x_{ij}^2 (d_i - x_i \tilde{\eta}^{[0]})^2]}.$$

Step 2. Run the weighted post-Lasso procedure based on penalty level $\lambda_d = 1.1\sqrt{n}\,\Phi^{-1}(1 - 0.05/\{p \log(n)\})$ and the diagonal matrix $\widehat{\Psi}^{[l-1]} := \text{diag}\big(\widehat{\Psi}^{[l-1]}_{j\,j} \ \forall\ j \in \{1, \ldots, p\}\big)$:

$$\begin{aligned} \hat{\eta}^{[l]} \ &\in \ \arg\min_{\eta} \frac{1}{2} \mathbb{E}_n\big[\widehat{w}_i (d_i - x_i \eta)^2\big] + \frac{\lambda_d}{n} \big\|\widehat{\Psi}^{[l-1]} \eta\big\|_1, \\ \tilde{\eta}^{[l]} \ &\in \ \arg\min_{\eta} \mathbb{E}_n\big[\widehat{w}_i (d_i - x_i \eta)^2\big] \ : \ \text{support}(\eta) \ \subseteq \ \text{support}(\hat{\eta}^{[l]}). \end{aligned}$$

For $j \in \{1, \ldots, p\}$ compute the updated penalty loadings according to

$$\widehat{\Psi}^{[l]}_{j\,j} = \sqrt{\mathbb{E}_n[\widehat{w}_i x_{ij}^2 (d_i - x_i \tilde{\eta}^{[l]})^2]}.$$

If $l < \bar{l}$ and $\max_j \big|\widehat{\Psi}^{[l]}_{j\,j} - \widehat{\Psi}^{[l-1]}_{j\,j}\big| > \varepsilon$, update $l \leftarrow l + 1$ and repeat Step 2. Otherwise stop.

A.5. AUXILIARY INEQUALITIES

Lemma A.6.2 *(Lemma 3 in Belloni et al. (2016b)). Let $X_i \in \mathbb{R}^p$ be independent random variables and let $K = \mathrm{E}[\max_{1\leqslant i\leqslant n} \|X_i\|_\infty^k]$ for some $k \geqslant 1$. Then, we have*

$$\mathrm{E}\big[\max_{1\leqslant j\leqslant p} \big|\mathbb{E}_n[|X_{ij}|^k] - \overline{\mathrm{E}}[|X_{ij}|^k]\big|\big] \lesssim \frac{K\log(p)}{n} + \sqrt{\frac{K\log(p)}{n}\max_{1\leqslant j\leqslant p}\overline{\mathrm{E}}[|X_{ij}|^k]}.$$

Lemma A.6.3 *(Uniform Operator Law of Large Numbers – Theorem 3.6 of Rudelson and Vershynin (2008) in the version of Lemma 4 in Belloni et al. 2016b). Let $(X_i)_{1\leqslant i\leqslant n}$ be independent random vectors in $\mathbb{R}^p$ such that $\sqrt{\mathrm{E}[\max_{1\leqslant i\leqslant n}\|X_i\|_\infty^2]} \leqslant K$. Define*

$$\delta_n = 2\big(\bar{C}K\sqrt{k}\log(1+k)\sqrt{\log(p\vee n)}\sqrt{\log(n)}\big)/\sqrt{n},$$

where $\bar{C}$ is a universal constant. Then,

$$\mathrm{E}\bigg[\sup_{|\theta\|_0\leqslant k,\|\theta\|=1}\big|\mathbb{E}_n\big[(X_i\theta)^2 - \mathrm{E}[(X_i\theta)^2]\big]\big|\bigg] \leqslant \delta_n^2 + \delta_n \sup_{|\theta\|_0\leqslant k,\|\theta\|=1}\overline{\mathrm{E}}[(X_i\theta)^2]^{1/2}.$$

Lemma A.6.4 *(Lemma 5 in Belloni et al. (2016b)). For the random process h_i indexed by $\mathcal{T} \subset \mathbb{R}^p$ and random vector $W_i \in \mathbb{R}^p$, independent across $i \in \{1,\dots,n\}$, let $|h_i(t)| \leqslant |W_i t|$, $\bar{\sigma}^2 := \sup_{t\in\mathcal{T}}\overline{\mathrm{E}}[\{h_i(t)\}^2]$, and $\|\mathcal{T}\|_1 = \sup_{t\in\mathcal{T}}\|t\|_1$. Provided that $K^2\|\mathcal{T}\|_1 M/4 \geqslant \bar{\sigma}^2$, we have*

$$\mathrm{E}\bigg[\sup_{t\in\mathcal{T}}\big|\mathbb{E}_n[h_i(t) - \mathrm{E}[h_i(t)]]\big|\bigg] \leqslant 4\|\mathcal{T}\|_1\mathrm{E}\big[\|\mathbb{E}_n[\varepsilon_i W_i]\|_\infty\big] \qquad \text{and}$$

$$\mathrm{P}\bigg(\sup_{t\in\mathcal{T}}\big|\mathbb{E}_n[h_i(t) - \mathrm{E}[h_i(t)]]\big| > \frac{K\|\mathcal{T}\|_1\sqrt{M}}{\sqrt{n}}\bigg) \leqslant 32p\exp\bigg\{\frac{-K^2}{1024}\bigg\} + \mathrm{P}\bigg(\max_{1\leqslant j\leqslant p}\mathbb{E}_n[W_{ij}^2] > M\bigg),$$

where ε_i are independent Rademacher random variables.

A.6. RESULTS WITH PROOFS FOR LOGISTIC TOBIT

In this section, we establish sparsity and convergence rates for the post-ℓ_1 logistic Tobit estimator. In doing so, we build on related results for the logistic binary choice model in

Belloni et al. (2016b). In particular, the modified self-concordance property of the softplus function $h(t) = \log(1 + \exp(t))$ discussed in Bach (2010), and the non-linear impact coefficient proposed in Belloni and Chernozhukov (2011) are used to bound the logistic Tobit likelihood loss $\Lambda(\alpha, \beta, \gamma)$ defined in (1.2.20).

Define $\Theta := (\alpha,\, \beta^T,\, \gamma)^T \in \mathbb{R}^{p+1} \times \mathbb{R}^+$, $\tilde{x}_i := (d_i,\, x_i)$, and $\delta = (\delta_{\tilde{x}}^T,\, \delta_y)^T = (\delta_d,\, \delta_x^T,\, \delta_y)^T \in \mathbb{R}^{p+1} \times \mathbb{R}^+$. Furthermore, let the support of Θ_0 be given by $\mathcal{S} = \text{support}(\alpha_0,\, \beta_0^T) \cup \{p + 2\} := \mathcal{S}^* \cup \{p + 2\}$. We denote the cardinality of the former set by $|\mathcal{S}| = s$. Note that the index of γ_0 is always contained in $\mathcal{S}$.

The logistic Tobit Lasso estimator is defined as any $\widehat{\Theta}$ such that

$$\widehat{\Theta} \in \arg\min_{\Theta} \Lambda(\Theta) + \frac{\lambda}{n}\big\|(\alpha,\, \beta^T)^T\big\|_1. \tag{A.6.1}$$

Moreover, the post-selection Tobit estimator associated with active set $\text{support}((\hat{\alpha},\, \hat{\beta}^T)^T) =: \widehat{\mathcal{S}^*} \subset \{1, \ldots, p + 1\}$ is defined as

$$\tilde{\Theta} \in \arg\min_{\Theta} \Lambda(\Theta) \;:\; \text{support}((\alpha,\, \beta^T)^T) \subseteq \widehat{\mathcal{S}^*}. \tag{A.6.2}$$

A.6.1. Design Conditions and Relations

Here, we formally introduce relevant quantities which we rely upon at several occasions throughout the following analysis of the ℓ_1-regularized logistic Tobit estimator. The quantities not pertaining to the outcome variable y_i, both weighted and non-weighted, are well documented in the existing literature. However, we also require augmented design quantities that include an additional column associated with y_i. The relevant quantities are either related to the original, empirical Gram matrix $\mathbb{E}_n[\tilde{x}_i^T \tilde{x}_i]$, its weighted $\mathbb{E}_n[w_i \tilde{x}_i^T \tilde{x}_i]$, augmented $\mathbb{E}_n[(\tilde{x}_i,\, -y_i)^T (\tilde{x}_i,\, -y_i)]$, or weighted and augmented $\mathbb{E}_n[w_i (\tilde{x}_i,\, -y_i)^T (\tilde{x}_i,\, -y_i)]$ counterparts, where the weights are defined as in (1.2.9) and satisfy $0 < \underline{c} \leqslant w_i < 1/2$.

As is typical in related literature, the results are obtained under a fixed design, i.e., sequence $(\tilde{x}_i)_{1 \leqslant i \leqslant n}$ is considered non-random. As a consequence, if necessary, the expectation is taken with respect to y_i. However, as the comments below emphasize, when considering a random design where sequence $(\tilde{x}_i)_{1 \leqslant i \leqslant n}$ is generated as independent realizations from a random vector, the presented restrictions are simply imposed on the population counterparts of the respective quantity. To control the empirical quantities, we then require the deviations from the expected version to be bounded (in probability) by a constant (for n large

enough); see, e.g., Rudelson and Vershynin (2008) and van de Geer and Bühlmann (2009). Therefore, the presence of outcomes y_i in the design quantities does not pose a significant challenge when compared to the analysis of ℓ_1-penalized logistic regression in Belloni et al. (2016b).

In order to differentiate between various representations of an arbitrary quantity $t_\circ$, the empirical version is marked by index n, whereas the expected version is indexed by P. Furthermore, the presence of superscript u or an asterisk indicate that the quantity is non-weighted or non-augmented. For example, $t^{u*}_{\circ,n}$ represents the empirical, non-weighted version of $t_\circ$ where the additional column associated with y_i is excluded.

Definition A.6.1 *(Restricted Eigenvalue).* The restricted eigenvalues for the logistic Tobit are defined as

$$\begin{aligned}
\kappa_{\min,n}(\mathbf{c}) &:= \min_{\|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}{\|\delta_{\mathcal{S}}\|}, \\
\kappa^u_{\min,n}(\mathbf{c}) &:= \min_{\|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1} \frac{\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}{\|\delta_{\mathcal{S}}\|}, \\
\kappa_{\min,\mathrm{P}}(\mathbf{c}) &:= \min_{\|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1} \frac{\mathrm{E}[w_i\{(\tilde{x}_i,\, -y_i)\delta\}^2]^{1/2}}{\|\delta_{\mathcal{S}}\|}.
\end{aligned}$$

As pointed out in Jacobson and Zou (2023a), by inclusion of the additional column associated with censored outcomes y_i, the augmented empirical Gram matrix is random, even in a fixed design. Therefore, any quantities related to this matrix, such as the restricted eigenvalues, are random as well. However, because of the following relation

$$\begin{aligned}
\frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}{\|\delta_{\mathcal{S}}\|} &\geqslant \frac{\mathrm{E}[w_i\{(\tilde{x}_i,\, -y_i)\delta\}^2]^{1/2}}{\|\delta_{\mathcal{S}}\|} \\
&\quad - \frac{\big|\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n} - \mathrm{E}[w_i\{(\tilde{x}_i,\, -y_i)\delta\}^2]^{1/2}\big|}{\|\delta_{\mathcal{S}}\|} \\
&\geqslant \kappa_{\min,\mathrm{P}}(\mathbf{c}) - \frac{\big|\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|^2_{2,n} - \mathrm{E}[w_i\{(\tilde{x}_i,\, -y_i)\delta\}^2]\big|^{1/2}}{\|\delta_{\mathcal{S}}\|}
\end{aligned}$$

we have to control the deviation of empirical and expected design quantities to ensure that the empirical eigenvalues are bounded from below. In a fixed design, this condition will hold with high probability, since most random elements of the empirical, weighted, augmented Hessian are bounded and therefore sub-Gaussian; see Jacobson and Zou (2023b) for a discussion on design quantities in the canonical Tobit with fixed design. Since we intend to cover more general designs with random sequences $(\tilde{x}_i)_{1\leqslant i\leqslant n}$ as in Belloni et al. (2016b, 2019), though, we pursue a different approach. Firstly, we need to control the impact the weights w_i have on the design quantities. Then, we invoke a result from Bickel et al. (2009)

that restricted eigenvalues are bounded from below provided that the minimal and maximal eigenvalues of Cs-sparse sub-matrices of the augmented, empirical Gram matrix are well-behaved. Lastly, using Lemma A.6.3 of Rudelson and Vershynin (2008) and Condition ITob 4.2–(ii), we guarantee that the deviation of empirical and expected design quantities is negligible with probability $1 - o(1)$ for sufficiently large n.

Definition A.6.2 *(Non-linear Impact Coefficient).* For a subset $A \subset \mathbb{R}^{p+1} \times \mathbb{R}^+$ let the non-linear impact coefficient be defined as

$$\overline{q}_{A,n} = \inf_{\delta \in A} \frac{\|\sqrt{w_i}(d_i,\, x_i,\, -y_i)\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|(d_i,\, x_i,\, -y_i)\delta|^3]}.$$

The non-linear impact coefficient was originally proposed by Belloni and Chernozhukov (2011) and applied in Belloni et al. (2016b, 2019) and Pan and Xie (2023) to facilitate the analysis of ℓ_1-penalized logistic, quantile regression, and pairwise difference censored median regression, respectively. In general, this quantity controls the quality of the minoration of the objective function (1.2.2) by a suitable quadratic function over a restricted set. Analogously to the situation above, we have to adjust $\overline{q}_{A,n}$ by including the column of censored outcomes y_i. In the Lemmas below, $\overline{q}_{A,n}$ will be applied on the sets $A = \Delta_{\mathbf{c}} = \{\delta \in \mathbb{R}^{p+2} : \|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1\}$ and $A = \{\delta \in \mathbb{R}^{p+2} : \|\delta\|_0 \leqslant Cs\}$.

In extension to the minimal and maximal m-sparse empirical eigenvalues defined in Step 2 (a) of the proof of Theorem 1.4.2, we introduce the weighted and non-weighted, non-augmented m-sparse eigenvalues.

Definition A.6.3 *(Minimal and Maximal m-Sparse Eigenvalues).*

$$\phi_{\text{min},n}(m) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2}{\|\delta\|^2} \quad \text{and} \quad \phi_{\text{max},n}(m) := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}^2}{\|\delta\|^2},$$
$$\phi^{u*}_{\text{min},n}(m) := \min_{1 \leqslant \|\delta_{\tilde{x}}\|_0 \leqslant m} \frac{\|\tilde{x}_i\delta_{\tilde{x}}\|_{2,n}^2}{\|\delta_{\tilde{x}}\|^2} \quad \text{and} \quad \phi^{u*}_{\text{max},n}(m) := \max_{1 \leqslant \|\delta_{\tilde{x}}\|_0 \leqslant m} \frac{\|\tilde{x}_i\delta_{\tilde{x}}\|_{2,n}^2}{\|\delta_{\tilde{x}}\|^2}$$

Remark A.6.7 *(Ordering of Empirical Sparse Eigenvalues of Original and Augmented Gram Matrices).*

To relate the eigenvalues of the original and augmented design quantities, note that the original empirical Gram matrix is a principal submatrix of the augmented one:

$$\mathbb{E}_n[(\tilde{x}_i,\, -y_i)^T(\tilde{x}_i,\, -y_i)] = \mathbb{E}_n \begin{bmatrix} \tilde{x}_i^T\tilde{x}_i & -y_i\tilde{x}_i^T \\ -y_i\tilde{x}_i & y_i^2 \end{bmatrix}.$$

Thus, the interlacing theorem (an application of the Courant-Fischer theorem) implies the following ordering of the m-sparse eigenvalues

$$\phi^u_{\min,n}(m) \leqslant \phi^{u*}_{\min,n}(m) \leqslant \phi^{u*}_{\max,n}(m) \leqslant \phi^u_{\max,n}(m).$$

For a similar implication involving the restricted eigenvalues $\kappa^u_{\min,n}(\mathbf{c}) \leqslant \kappa^{u*}_{\min,n}(\mathbf{c})$, the reader is referred to Bickel et al. (2009) for a thorough discussion. Therefore, we have to establish conditions under which the m-sparse eigenvalues of the augmented empirical Gram matrix are well behaved. ■

Since $\kappa_{\min,n}(\mathbf{c})$ in Definition A.6.1 differs from $\kappa^u_{\min,n}(\mathbf{c})$ and from the restricted eigenvalue for quadratic problems, as introduced in Bickel et al. (2009), by the weights w_i, it is necessary to understand how these weights impact the behaviour of the Gram matrices. To this end, we consider the following quotients.

Definition A.6.4 *(Quotients of Design Quantities).* Let the quotients of weighted and non-weighted expected design quantities be defined as

$$\begin{aligned} \nu_{(r),n}(\mathbf{c}) &:= \min_{\|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}{\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}, \\ \nu_{(s),n}(m) &:= \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}{\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n}}. \end{aligned}$$

The following Lemma establishes lower bounds for the relations between weighted and non-weighted quantities.

Lemma A.6.5 *(Relating Weighted and Non-weighted Design Quantities). Let* $w_i = -g'(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)$ *be defined as in (1.2.18). Then, the following inequalities hold* $\nu_{(r),n}(\mathbf{c}) \geqslant$

$\min_{1\leqslant i\leqslant n}\sqrt{w_i}, \nu_{(s),n}(m) \geqslant \min_{1\leqslant i\leqslant n}\sqrt{w_i}$,

$$\begin{aligned}
\nu_{(r),n}(\mathbf{c}) &\geqslant \frac{\kappa^u_{\min,n}(\mathbf{c})\,\mathbb{E}_n[1/w_i]^{-1/2}}{(1+\mathbf{c})\sqrt{s}\,\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty},\\
\nu_{(s),n}(m) &\geqslant \frac{\sqrt{\phi^u_{\min,n}(m)}\,\mathbb{E}_n[1/w_i]^{-1/2}}{\sqrt{m}\,\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty},\\
\{\nu_{(r),n}(\mathbf{c})\}^2 &\geqslant \frac{\kappa_{\min,n}(\mathbf{c})\,\mathbb{E}_n[1/w_i]^{-1/2}}{(1+\mathbf{c})\sqrt{s}\,\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty},
\end{aligned}$$

Proof. The first four inequalities are essentially due to Lemma 6 in Belloni et al. (2016b). The last relation follows from

$$\begin{aligned}
\mathbb{E}_n[\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/2} &\leqslant \mathbb{E}_n[\sqrt{w_i}/\sqrt{w_i}\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/2}\\
&\leqslant \mathbb{E}_n[w_i\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/4}\,\mathbb{E}_n[\{(\tilde{x}_i,\,-y_i)\delta\}^2/w_i]^{1/4}\\
&\leqslant \mathbb{E}_n[w_i\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/4}\,\mathbb{E}_n[1/w_i]^{1/4}\,\|\delta\|_1^{1/2}\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty^{1/2}\\
\frac{\mathbb{E}_n[\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/2}}{\mathbb{E}_n[w_i\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/4}} &\leqslant \mathbb{E}_n[1/w_i]^{1/4}s^{1/4}(1+\mathbf{c})^{1/2}\|\delta_{\mathcal{S}}\|^{1/2}\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty^{1/2}\\
&\leqslant \mathbb{E}_n[1/w_i]^{1/4}s^{1/4}\sqrt{\frac{1+\mathbf{c}}{\kappa_{\min,n}(\mathbf{c})}}\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty^{1/2}\\
&\qquad\cdot\mathbb{E}_n[w_i\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/4}\\
\frac{\mathbb{E}_n[\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/2}}{\mathbb{E}_n[w_i\{(\tilde{x}_i,\,-y_i)\delta\}^2]^{1/2}} &\leqslant \mathbb{E}_n[1/w_i]^{1/4}s^{1/4}\sqrt{\frac{1+\mathbf{c}}{\kappa_{\min,n}(\mathbf{c})}}\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\,y_i)\|_\infty^{1/2},
\end{aligned}$$

where we used the Cauchy-Schwarz inequality, relation (A.6.10) in Lemma A.6.11 below, namely $\|\delta\|_1 \leqslant \sqrt{s}(1+\mathbf{c})\|\delta_{\mathcal{S}}\|$ for $\delta \in \Delta_{\mathbf{c}}$, and Definition A.6.1 above. The final result follows from inverting the inequality and the fact that it holds for any $\delta \in \Delta_{\mathbf{c}}$. ■

A.6.2. Identification Lemmas

The following two lemmas provide a lower bound on $\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) - \{\nabla_\Theta\Lambda(\Theta_0)\}^T(\widehat{\Theta}-\Theta_0)$. This result proves useful when deriving bounds for penalized and post-model selection parameters. In accordance with Belloni et al. (2016b), we exploit the separability of the likelihood loss across observations and the self-concordance property of the softplus function. This allows us to define a radius over which the criterion function can be bounded by a quadratic function.

Lemma A.6.6 *(Bounds for Univariate Modified Self-Concordant Functions – Lemma 1 in Bach, 2010). Let h be a convex three times differentiable function $h : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|h'''(t)| \leqslant Mh''(t)$, for some $M \geqslant 0$. Then, for all $t \geqslant 0$:*

$$\frac{h''(0)}{M^2}\big(\exp\{-Mt\} + Mt - 1\big) \leqslant h(t) - h(0) - h'(0)t \leqslant \frac{h''(0)}{M^2}\big(\exp\{Mt\} + Mt - 1\big).$$

Indeed, as shown in Lemma 8 in Belloni et al. (2016b) we can further relax the lower bound. Here, we provide an easy-to-grasp alternative to the proof given in Belloni et al. (2016b). By the Taylor series representation of the exponential function, we have for some $\tilde{t} \in [-Mt, 0]$:

$$\exp\{-Mt\} = 1 - Mt + \frac{(Mt)^2}{2} - \frac{(Mt)^3}{6} + \underbrace{\exp\{\tilde{t}\}\frac{(Mt)^4}{24}}_{>0}.$$

Therefore, we get the simplified lower bound

$$\frac{h''(0)}{M^2}\left(\frac{(Mt)^2}{2} - \frac{(Mt)^3}{6}\right) \leqslant \frac{h''(0)}{M^2}\big(\exp\{-Mt\} + Mt - 1\big). \qquad \blacksquare$$

Lemma A.6.7 *(Minoration Lemma). Let the likelihood loss* $\Lambda(\Theta)$ *for* $\Theta = (\alpha, \beta^T, \gamma)^T$ *be defined as in equation (1.2.20). We have that*

$$\begin{aligned}\Lambda(\Theta_0 + \delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \\ \geqslant \Big\{\frac{1}{3}\big\|\sqrt{w_i}(d_i,\, x_i,\, -y_i)\delta\big\|_{2,n}^2\Big\} \wedge \Big\{\frac{\overline{q}_{A,n}}{3}\big\|\sqrt{w_i}(d_i,\, x_i,\, -y_i)\delta\big\|_{2,n}\Big\}\end{aligned}$$

Proof. Essentially, we make use of the result from Lemma 9 in Belloni et al. (2016b) and the fact that the logarithm satisfies $\log(1 + z) \leqslant z$ for $z > -1$.

Step 1: (Minoration). Define $\tilde{x}_i = (d_i,\, x_i)$, $\delta = (\delta_d,\, \delta_x^T,\, \delta_y)^T$ with $\delta_y = \hat{\gamma} - \gamma_0$ and $\hat{\gamma}, \gamma_0 > 0$. Define the maximal radius over which the following criterion function can be bounded from below by a suitable quadratic function

$$r_A = \sup_r \left\{ r : \begin{array}{c} \Lambda(\Theta_0 + \delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \geqslant \frac{1}{3}\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}^2, \\ \forall\, \delta \in A,\, \delta_y > -\gamma_0,\, \big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n} \leqslant r \end{array} \right\}.$$

Step 2 below shows that $r_A \geqslant \overline{q}_{A,n}$. By construction of r_A and convexity of the criterion function $\Lambda(\Theta_0 + \delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta$, we see that

$$\begin{aligned}&\Lambda(\Theta_0 + \delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \\ &\geqslant \frac{\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}^2}{3} \wedge \left\{ \frac{\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}}{r_A} \cdot \inf_{\delta \in A, \|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n} \geqslant r_A} \Lambda(\Theta_0 + \delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \right\} \\ &\geqslant \frac{\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}^2}{3} \wedge \left\{ \frac{\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}}{r_A} \cdot \frac{r_A^2}{3} \right\} \\ &\geqslant \frac{\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}^2}{3} \wedge \frac{\overline{q}_{A,n}\big\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\big\|_{2,n}}{3}.\end{aligned}$$

Step 2: $(r_A \geqslant \overline{q}_{A,n})$. By the definition of the empirical likelihood loss in (1.2.20) and the softplus function $h_i(t) = \log\big(1 + \exp\{(\tilde{x}_i, -y_i)\Theta_0 + t(\tilde{x}_i, -y_i)\delta\}\big)$, we have

$$\begin{aligned}
&\Lambda(\Theta_0+\delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \\
&\quad = \mathbb{E}_n[(1+s_i)h_i(1) - s_i\{(\tilde{x}_i, -y_i)(\Theta_0+\delta) + \log(\gamma_0+\delta_y)\}] \\
&\qquad - \mathbb{E}_n[(1+s_i)h_i(0) - s_i\{(\tilde{x}_i, -y_i)\Theta_0 + \log(\gamma_0)\}] \\
&\qquad - \mathbb{E}_n[(1+s_i)H_i(0)\{(\tilde{x}_i, -y_i)\delta\} - s_i\{(\tilde{x}_i, -y_i)\delta + \delta_y/\gamma_0\}] \\
&\quad = \mathbb{E}_n[(1+s_i)h_i(1) - (1+s_i)h_i(0) - (1+s_i)H_i(0)\{(\tilde{x}_i, -y_i)\delta\}] \\
&\qquad - \mathbb{E}_n[s_i\{\log(1+\delta_y/\gamma_0) - \delta_y/\gamma_0\}] \\
&\quad \geqslant \mathbb{E}_n[(1+s_i)h_i(1) - (1+s_i)h_i(0) - (1+s_i)H_i(0)\{(\tilde{x}_i, -y_i)\delta\}] \\
&\quad = \mathbb{E}_n[(1+s_i)h_i(1) - (1+s_i)h_i(0) - (1+s_i)h_i'(0)\cdot 1],
\end{aligned}$$

where the inequality follows from $\delta_y > -\gamma_0$ and the well known upper bound on the natural logarithm $\log(1+z) \leqslant z$ for $z > -1$. Note that the softplus function $h_i(t)$ is a convex, three times differentiable function that satisfies $|h_i'''(t)| \leqslant |(\tilde{x}_i, -y_i)\delta|\, h_i''(t)$, because for

$$H_i(t) = \frac{\exp\{(\tilde{x}_i, -y_i)\Theta_0 + t(\tilde{x}_i, -y_i)\delta\}}{1 + \exp\{(\tilde{x}_i, -y_i)\Theta_0 + t(\tilde{x}_i, -y_i)\delta\}}$$

we have

$$\begin{aligned}
h_i'(t) &= \{(\tilde{x}_i, -y_i)\delta\} H_i(t) \\
h_i''(t) &= \{(\tilde{x}_i, -y_i)\delta\}^2 H_i(t)[1 - H_i(t)] \\
h_i'''(t) &= \{(\tilde{x}_i, -y_i)\delta\}^3 H_i(t)[1 - H_i(t)][1 - 2\,H_i(t)]
\end{aligned}$$

and $0 \leqslant H_i(t) \leqslant 1$ implies $-1 \leqslant [1 - 2\,H_i(t)] \leqslant 1$. Therefore, by the lower bound in Lemma A.6.6 and the definition of weights w_i in (1.2.9) and (1.2.18) we obtain

$$\begin{aligned}
&(1+s_i)\{h_i(1) - h_i(0) - 1\cdot h_i'(0)\} \\
&\qquad \geqslant \frac{(1+s_i)h_i''(0)}{\{(\tilde{x}_i, -y_i)\delta\}^2}\big(\exp\{-|(\tilde{x}_i, -y_i)\delta|\} + |(\tilde{x}_i, -y_i)\delta| - 1\big) \\
&\qquad \geqslant w_i\left\{\frac{|(\tilde{x}_i, -y_i)\delta|^2}{2} - \frac{|(\tilde{x}_i, -y_i)\delta|^3}{6}\right\}.
\end{aligned}$$

Since the criterion function is separable across observations, we see that

$$\Lambda(\Theta_0+\delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T \delta \geqslant \frac{\mathbb{E}_n[w_i|(\tilde{x}_i, -y_i)\delta|^2]}{2} - \frac{\mathbb{E}_n[w_i|(\tilde{x}_i, -y_i)\delta|^3]}{6}.$$

From the definition of the non-linear impact coefficient in A.6.2, we have that for any $\delta \in A$ such that $\overline{q}_{A,n} \geqslant \|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}$

$$1 \leqslant \frac{\overline{q}_{A,n}}{\|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}} \leqslant \frac{\|\sqrt{w_i}(\tilde{x}_i, -y_i)\delta\|_{2,n}^2}{\mathbb{E}_n[w_i|(\tilde{x}_i, -y_i)\delta|^3]}$$

and thus

$$\mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^3] \leqslant \mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^2].$$

Then, we obtain the following bound

$$\begin{aligned}\Lambda(\Theta_0+\delta) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T\delta &\geqslant \frac{\mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^2]}{2} - \frac{\mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^3]}{6} \\ &\geqslant \frac{\mathbb{E}_n[w_i|(\tilde{x}_i,\, -y_i)\delta|^2]}{3}.\end{aligned}$$

On the other hand, if $\overline{q}_{A,n} \leqslant \|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n}$ we necessarily have $\overline{q}_{A,n} \leqslant r_A$, which verifies the assertion in Step 1 above. The final result follows by differentiating between both cases. ■

A.6.3. Penalty Choice and Rate for ℓ_1-Penalized Logistic Tobit

This section presents the derivation of convergence rates for the ℓ_1-penalized logistic Tobit. On several occasions we require $\lambda_y/n > \|\nabla_\Theta\, \Lambda(\Theta_0)\|_\infty$, which is a standard assumption in the analysis of ℓ_1-regularized estimators. To this end, we provide a collection of probability bounding lemmas for the logistic Tobit likelihood score, the first of which is known from other high-dimensional models. The second and third lemma are new and have been specifically tailored to the additional element in the likelihood score, γ_0.

Lemma A.6.8 *(Concentration of $\nabla_{(\alpha,\,\beta^T)^T}\, \Lambda(\Theta_0)$ Using Hoeffding's Inequality). Assume normalized features $\mathbb{E}_n[d_i^2] = 1$ and $\mathbb{E}_n[x_{i,j}^2] = 1$. For any $t \geqslant 0$, we have*

$$\mathrm{P}(\|\nabla_{(\alpha,\,\beta^T)^T}\, \Lambda(\Theta_0)\|_\infty \geqslant t) \leqslant 2(p+1)\exp\{-t^2 n/2\}.$$

Proof. Recall from definition (1.2.18) that

$$|g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)| = \left|\frac{1 - s_i \exp\{\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0\}}{1 + \exp\{\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0\}}\right| \leqslant 1.$$

Applying the union bound, Hoeffding's two-sided inequality for bounded random variables (see, e.g. Theorem 2.2.6 in Vershynin, 2018, p.16), and exploiting normalization $\mathbb{E}_n[d_i^2] = 1$ and $\mathbb{E}_n[x_i^2] = 1$ of our features, we obtain

$$\begin{aligned}\mathrm{P}(\|\nabla_{(\alpha,\,\beta^T)^T}\, \Lambda(\Theta_0)\|_\infty \geqslant t) &= \mathrm{P}(\|\mathbb{E}_n[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)(\tilde{x}_i)^T]\|_\infty \geqslant t) \\ &\leqslant \mathrm{P}(|\mathbb{E}_n[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)d_i]| \geqslant t) \\ &\quad + p \cdot \max_{1\leqslant j\leqslant p} \mathrm{P}(|\mathbb{E}_n[g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)x_{ij}^T]| \geqslant t) \\ &\leqslant 2(p+1)\exp\{-t^2 n/2\}.\end{aligned}$$

This result yields the standard penalty level $\lambda = \sqrt{2n \log(2(p+1)/\Delta)}$ for some $\Delta \in (0,1)$ known from various other models. Further refinements are possible by, for example, using moderate deviation theory for self-normalizing sums; see Jing et al. (2003) and de la Peña et al. (2009). The steps taken in Lemma 11 of Belloni et al. (2016b) suggest to invoke information matrix equality such that $\lambda = c\sqrt{n/2}\,\Phi^{-1}\big(1 - \Delta/[2(p+1)]\big)$ with $c = 1.1$ and $\Delta = 0.1\,\log^{-1}(n)$ because $w_i \leqslant \sup_{t\in\mathbb{R}} |g'(t)| = 1/2$. ■

Lemma A.6.9 *(Moment Generating Function of $\nabla_\gamma \Lambda_i(\Theta)$). Let $\Lambda_i(\Theta) = \Lambda_i(\alpha, \beta^T, \gamma)^T$ be defined as in (1.2.20) and $\tilde{m}_i = \alpha d_i + x_i\beta$. The censored outcome variable y_i obeys model (1.2.1). Then, the moment generating function $\mathrm{E}\big[\exp\{t\,\nabla_\gamma \Lambda_i(\Theta)\}\big]$ exists for all $|t| < \gamma$ and we have*

$$\begin{aligned}&\mathrm{E}\big[\exp\{t\,\nabla_\gamma \Lambda_i(\Theta)\}\big]\\ &\qquad\leqslant \frac{1}{(1+\mathrm{e}^{\tilde{m}_i})} + \mathrm{e}^{|t|(1+\tilde{m}_i)/\gamma}\left[\mathrm{B}\Big(1+\frac{|t|}{\gamma}, 1-\frac{|t|}{\gamma}\Big) - \mathrm{B}\Big(\frac{1}{(1+\mathrm{e}^{\tilde{m}_i})}; 1+\frac{|t|}{\gamma}, 1-\frac{|t|}{\gamma}\Big)\right],\end{aligned}$$

where for $z \in [0,1]$ and $a, b > 0$ we denote the incomplete beta function by

$$\mathrm{B}(z; a, b) = \int_0^z v^{a-1}(1-v)^{b-1}\,\mathrm{d}v \quad \textit{and} \quad \mathrm{B}(a,b) = \mathrm{B}(1; a, b).$$

Proof. Note that

$$|\nabla_\gamma \Lambda_i(\Theta)| = \left|\frac{\exp\{\gamma y_i - \tilde{m}_i\} - 1}{1 + \exp\{\gamma y_i - \tilde{m}_i\}} y_i - \frac{s_i}{\gamma}\right| \leqslant y_i + s_i/\gamma$$

and therefore

$$\begin{aligned}\mathrm{E}\big[\exp\{t\,\nabla_\gamma \Lambda_i(\Theta)\}\big] &\leqslant \mathrm{E}\big[\exp\{|t|\,|\nabla_\gamma \Lambda_i(\Theta)|\}\big] \leqslant \mathrm{E}\big[\exp\{|t|\,(y_i + s_i/\gamma)\}\big]\\ &= \frac{1}{1+\mathrm{e}^{\tilde{m}_i}} + \mathrm{e}^{|t|/\gamma}\int_0^\infty \mathrm{e}^{|t|y_i}\frac{\gamma\,\mathrm{e}^{\gamma y_i - \tilde{m}_i}}{(1+\mathrm{e}^{\gamma y_i - \tilde{m}_i})^2}\,\mathrm{d}y_i\\ &= \frac{1}{1+\mathrm{e}^{\tilde{m}_i}} + \mathrm{e}^{|t|(1+\tilde{m}_i)/\gamma}\int_{-\tilde{m}_i}^\infty \mathrm{e}^{|t|r/\gamma}\frac{\mathrm{e}^r}{(1+\mathrm{e}^r)^2}\,\mathrm{d}r\\ &= \frac{1}{1+\mathrm{e}^{\tilde{m}_i}} + \mathrm{e}^{|t|(1+\tilde{m}_i)/\gamma}\int_{1/(1+\mathrm{e}^{\tilde{m}_i})}^1 \left(\frac{v}{1-v}\right)^{|t|/\gamma}\mathrm{d}v\\ &= \frac{1}{(1+\mathrm{e}^{\tilde{m}_i})} + \mathrm{e}^{|t|(1+\tilde{m}_i)/\gamma}\left[\mathrm{B}\Big(1+\frac{|t|}{\gamma}, 1-\frac{|t|}{\gamma}\Big) - \mathrm{B}\Big(\frac{1}{(1+\mathrm{e}^{\tilde{m}_i})}; 1+\frac{|t|}{\gamma}, 1-\frac{|t|}{\gamma}\Big)\right].\end{aligned}$$

In particular, by Markov's inequality (a Chernoff bound) we have for some $\lambda > 0$

$$\mathrm{P}(|\mathbb{E}_n[\nabla_\gamma \Lambda_i(\Theta_0)]| \geqslant \lambda/n) \leqslant \inf_{t\in[0,\gamma)} \mathrm{e}^{-\lambda t}\prod_{i=1}^n \mathrm{E}\big[\exp\{t\,|\nabla_\gamma \Lambda_i(\Theta_0)|\}\big].$$ ■

Lemma A.6.10 *(Variance of $\nabla_\gamma \Lambda(\Theta_0)$). Let $g(\gamma_0 y_i - \tilde{m}_{0i})$, $w_i = -g'(\gamma_0 y_i - \tilde{m}_{0i})$, and $\Lambda_i(\Theta) = \Lambda_i(\alpha, \beta, \gamma)$ be defined as in (1.2.18) and (1.2.20), where $\tilde{m}_{0i} = \alpha_0 d_i - x_i\beta_0$. The censored outcome variable y_i obeys model (1.2.1). Then, we have that*

$$\mathrm{P}\big(|\mathbb{E}_n[\nabla_\gamma \Lambda_i(\Theta_0)]| \geqslant t\big) \leqslant \frac{1}{3n\gamma_0^2 t^2}\mathbb{E}_n[\mathrm{e}^{\tilde{m}_{0i}}/(1+\mathrm{e}^{\tilde{m}_{0i}}) - 2\,\mathrm{Li}_2(-\mathrm{e}^{\tilde{m}_{0i}})],$$

where $\mathrm{Li}_2(z)$ represents the dilogarithm

$$\mathrm{Li}_2(z) = \int_z^0 \frac{\log(1-v)}{v}\,\mathrm{d}v$$

and additionally

$$\mathrm{Var}[\nabla_\gamma \Lambda_i(\Theta_0) \,|\, d_i, x_i] = \mathrm{E}[\{g(\gamma_0 y_i - \tilde{m}_{0i})y_i + s_i/\gamma_0\}^2 \,|\, d_i, x_i] = \mathrm{E}[w_i y_i^2 + s_i/\gamma_0^2 \,|\, d_i, x_i].$$

Proof. For $\nabla_\gamma \Lambda_i(\alpha_0, \beta_0, \gamma_0) = -g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)y_i - s_i/\gamma_0$ it must hold that $\mathrm{E}[\nabla_\gamma \Lambda_i(\Theta_0)] = 0$. Therefore,

$$\gamma_0^2\,\mathrm{Var}[\nabla_\gamma \Lambda_i(\Theta_0) \,|\, d_i, x_i] = \int_0^\infty \left(\frac{\exp\{\tilde{y}_i - \tilde{m}_{0i}\} - 1}{1+\exp\{\tilde{y}_i - \tilde{m}_{0i}\}}\tilde{y}_i - 1\right)^2 \frac{\exp\{\tilde{y}_i - \tilde{m}_{0i}\}}{(1+\exp\{\tilde{y}_i - \tilde{m}_{0i}\})^2}\,\mathrm{d}\tilde{y}_i,$$

where we substituted $\tilde{y}_i = \gamma_0 y_i$. Employing standard calculus and integration rules, we obtain the expression

$$\begin{aligned}
&= \lim_{\tilde{y}_i\to\infty}\left[\frac{-2(\mathrm{e}^{\tilde{m}_{0i}}+\mathrm{e}^{\tilde{y}_i})^3\left[\mathrm{Li}_2(-\mathrm{e}^{\tilde{y}_i-\tilde{m}_{0i}}) + \tilde{y}_i\log(1+\mathrm{e}^{\tilde{y}_i-\tilde{m}_{0i}})\right] + \mathrm{e}^{3\tilde{y}_i}\tilde{y}_i^2}{3(\mathrm{e}^{\tilde{m}_{0i}}+\mathrm{e}^{\tilde{y}_i})^3}\right.\\
&\qquad\left. - \frac{\mathrm{e}^{2\tilde{m}_{0i}+\tilde{y}_i}(2-2\tilde{y}_i-3\tilde{y}_i^2) + \mathrm{e}^{\tilde{m}_{0i}+2\tilde{y}_i}(1-2\tilde{y}_i) + \mathrm{e}^{3\tilde{m}_{0i}}}{3(\mathrm{e}^{\tilde{m}_{0i}}+\mathrm{e}^{\tilde{y}_i})^3}\right]\\
&\qquad + \frac{2}{3}\,\mathrm{Li}_2(-\mathrm{e}^{-\tilde{m}_{0i}}) + \frac{\mathrm{e}^{\tilde{m}_{0i}}}{3(1+\mathrm{e}^{\tilde{m}_{0i}})}\\
&= \lim_{\tilde{y}_i\to\infty}\frac{1}{3}\left[(\tilde{y}_i - m)^2 + \frac{\pi^2}{3} - 2\tilde{y}_i^2 + 2m\tilde{y}_i + \tilde{y}_i^2\right] + \frac{2}{3}\,\mathrm{Li}_2(-\mathrm{e}^{-\tilde{m}_{0i}}) + \frac{\mathrm{e}^{\tilde{m}_{0i}}}{3(1+\mathrm{e}^{\tilde{m}_{0i}})}\\
&= \frac{1}{3}\left[\frac{\pi^2}{3} + \tilde{m}_{0i}^2 + 2\,\mathrm{Li}_2(-\mathrm{e}^{-\tilde{m}_{0i}}) + \frac{\mathrm{e}^{\tilde{m}_{0i}}}{1+\mathrm{e}^{\tilde{m}_{0i}}}\right]\\
&= \frac{1}{3}\left[\frac{\mathrm{e}^{\tilde{m}_{0i}}}{1+\mathrm{e}^{\tilde{m}_{0i}}} - 2\,\mathrm{Li}_2(-\mathrm{e}^{\tilde{m}_{0i}})\right],
\end{aligned}$$

where the second and third line follow from $\lim_{a\to\infty} -2[\mathrm{Li}_2(-\mathrm{e}^a) + a\log(1+\mathrm{e}^a)] + a^2 = \pi^2/3$. To show this, we use the asymptotic expansion $\mathrm{Li}_2(-\mathrm{e}^a) = -a^2/2 - \pi^2/6 + O(a^{-1})$ at $a = \infty$. The last equality is due to $-\mathrm{Li}_2(1/z) = \mathrm{Li}_2(z) + \pi^2/6 + \{\log(-z)\}^2/2$. As a result, the variance of the score with respect to element γ is

$$\Sigma^2_{\nabla_\gamma} = \mathrm{Var}[\nabla_\gamma \Lambda(\Theta_0) \,|\, d_i, x_i] = \frac{1}{3n\gamma_0^2}\mathbb{E}_n[\mathrm{e}^{\tilde{m}_{0i}}/(1+\mathrm{e}^{\tilde{m}_{0i}}) - 2\,\mathrm{Li}_2(-\mathrm{e}^{\tilde{m}_{0i}})].$$

For the special case $\tilde{m}_{0i} = 0$ we get a censoring rate of 50%, i.e., $\mathrm{P}(y_i = 0) = \mathrm{P}(y_i^* \leqslant 0) = 1/2$, and the variance is $\Sigma^2_{\nabla_\gamma} = (3+\pi^2)/(18n\gamma_0^2)$. [Compare this to the variance of the mean of n $i.i.d.$ logistic random variables with scale parameter $1/\gamma_0$, which equals $\pi^2/(3n\gamma_0^2)$].

Applying Chebychev's inequality completes the proof of the first assertion

$$\mathrm{P}(|\nabla_\gamma \Lambda(\Theta_0)| \geqslant t) \leqslant t^{-2}\Sigma^2_{\nabla_\gamma}.$$

The second result follows from information matrix equality. Namely,

$$\begin{aligned}
\mathrm{E}[w_i y_i^2 + s_i/\gamma_0^2 \,|\, d_i, x_i] &= \int_0^\infty \left(\frac{2y_i^2 \, \exp\{\gamma_0 y_i - \tilde{m}_{0i}\}}{[1+\exp\{\gamma_0 y_i - \tilde{m}_{0i}\}]^2} + \gamma_0^{-2} \right) \frac{\gamma_0 \exp\{\gamma_0 y_i - \tilde{m}_{0i}\}}{[1+\exp\{\gamma_0 y_i - \tilde{m}_{0i}\}]^2} \, \mathrm{d}y_i \\
&= \frac{2}{\gamma_0^2} \int_0^\infty \tilde{y}_i^2 \left(\frac{\exp\{\tilde{y}_i - \tilde{m}_{0i}\}}{[1+\exp\{\tilde{y}_i - \tilde{m}_{0i}\}]^2} \right)^2 \mathrm{d}\tilde{y}_i + \frac{\gamma_0^{-2} \mathrm{e}^{\tilde{m}_{0i}}}{1+\mathrm{e}^{\tilde{m}_{0i}}} \\
&= \frac{1}{3\gamma_0^2} \left[\frac{\mathrm{e}^{\tilde{m}_{0i}}}{1+\mathrm{e}^{\tilde{m}_{0i}}} - 2\,\mathrm{Li}_2(-\mathrm{e}^{\tilde{m}_{0i}}) \right].
\end{aligned}$$

■

Remark A.6.8 *(Practical choice for $\lambda/n > \|\nabla_\Theta \Lambda(\Theta_0)\|_\infty$).*

By combining the results from Lemmas A.6.8 and A.6.10, we have

$$\begin{aligned}
\mathrm{P}(\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty \geqslant \lambda/n) &\leqslant \mathrm{P}(\|\nabla_{(\alpha,\,\beta^T)^T} \Lambda(\Theta_0)\|_\infty \geqslant \lambda/n) + \mathrm{P}(|\nabla_\gamma \Lambda(\Theta_0)| \geqslant \lambda/n) \\
&\leqslant 2(p+1)\exp\{-\lambda^2/(2n)\} + n^2 \, \Sigma^2_{\nabla_\gamma}/\lambda^2 \leqslant \Delta,
\end{aligned}$$

where $\Sigma^2_{\nabla_\gamma} = \mathrm{Var}[\nabla_\gamma \Lambda(\Theta_0) \,|\, d_i, x_i]$ is defined as in Lemma A.6.10 and $\Delta \in (0,1)$. While the second term on the right-hand side allows $\lambda = O(\sqrt{n})$, the first probability suggests $\lambda = O(\sqrt{\log(p)\, n})$. In particular, exempting target coefficient α and inverse scale parameter γ from regularization does not affect the convergence rate for $\|\widehat{\Theta} - \Theta_0\|$ derived in Lemma A.6.11. To ensure asymptotically good guarantees of our estimators, it therefore suffices to assume $\lambda/n \lesssim \sqrt{\log(p \vee n)/n}$. For practical implementations, we additionally advise to rescale the observed outcome variable y_i. Note that this does not affect the point estimates for α_0 and β_0, since $\gamma y_i = a\gamma(y_i/a) = \gamma^\diamond y_i^\diamond$ where we suggest to normalize y_i by taking $a = \mathbb{E}_n[y_i^2]$. This transformation reduces the perceived noise in $\nabla_\gamma \Lambda(\Theta_0)$. Another approach that was not further pursued here would be to estimate the variance $\Sigma^2_{\nabla_\gamma}$ using the Tobit Lasso coefficients as plug-in estimates for $\tilde{m}_{0i}$. The resulting point estimate, denoted by $\widehat{\Sigma}^2_{\nabla_\gamma}$, could be used to standardize $\nabla_\gamma \Lambda(\widehat{\Theta})$ by rescaling y_i.

Lemma A.6.11 *(Bound for Prediction Norm $\|\cdot\|_{2,n}$). Assume that $c\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty \leqslant \lambda/n$ for $c > 1$. Provided that the non-linear impact coefficient for $A = \Delta_{\mathbf{c}}$ satisfies*

$$\overline{q}_{\Delta_{\mathbf{c}},n} > 3\frac{(1+c)}{c}\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}$$

holds for $\delta \in \Delta_{\mathbf{c}}$, we have

$$\begin{aligned}
\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)(\widehat{\Theta} - \Theta_0)\|_{2,n} &\leqslant 3\frac{(1+c)}{c}\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}, && \text{(A.6.3)}\\
\|\widehat{\Theta} - \Theta_0\|_1 &\leqslant 6\mathbf{c}\frac{\lambda s}{n\,\{\kappa_{\min,n}(\mathbf{c})\}^2}, && \text{(A.6.4)}\\
\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) &\leqslant 3\frac{(1+c)}{c}\left\{\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}\right\}^2. && \text{(A.6.5)}
\end{aligned}$$

Proof. Essentially, we adjust the proof of Lemma 12 in Belloni et al. (2016b) by including y_i and inverse scale parameter γ. Recall the definitions from above: $\mathcal{S} = \text{support}(\alpha_0,\, \beta_0^T) \cup \{p+2\} := \mathcal{S}^* \cup \{p+2\}$ and $\delta = \widehat{\Theta} - \Theta_0$. Because $\widehat{\Theta}$ minimizes the empirical likelihood loss, we have

$$\Lambda(\widehat{\Theta}) + \frac{\lambda}{n}\|(\hat{\alpha},\, \hat{\beta}^T)^T\|_1 \leqslant \Lambda(\Theta_0) + \frac{\lambda}{n}\|(\alpha_0,\, \beta_0^T)^T\|_1$$

and therefore

$$\begin{aligned}
\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) &\leqslant \frac{\lambda}{n}\big(\|(\alpha_0,\, \beta_0^T)^T\|_1 - \|(\hat{\alpha},\, \hat{\beta}^T)^T\|_1 + |\hat{\gamma} - \gamma_0|\big) && \text{(A.6.6)}\\
&= \frac{\lambda}{n}\Big(\|(\alpha_0,\, \beta_0^T)^T_{\mathcal{S}^*} - (\hat{\alpha},\, \hat{\beta}^T)^T_{\mathcal{S}^*} + (\hat{\alpha},\, \hat{\beta}^T)^T_{\mathcal{S}^*}\|_1\\
&\qquad - \|(\hat{\alpha},\, \hat{\beta}^T)^T_{\mathcal{S}^*}\|_1 - \|(\hat{\alpha},\, \hat{\beta}^T)^T_{\mathcal{S}^{*C}}\|_1 + \|\hat{\gamma} - \gamma_0\|_1\Big)\\
&\leqslant \frac{\lambda}{n}\big(\|\widehat{\Theta}_{\mathcal{S}} - \Theta_{0\mathcal{S}}\|_1 - \|\widehat{\Theta}_{\mathcal{S}^C} - \Theta_{0\mathcal{S}^C}\|_1\big)\\
&= \frac{\lambda}{n}\big(\|\delta_{\mathcal{S}}\|_1 - \|\delta_{\mathcal{S}^C}\|_1\big)
\end{aligned}$$

Note that by convexity of the likelihood loss, we have

$$\begin{aligned}
\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) &\geqslant \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T\delta && \text{(A.6.7)}\\
&\geqslant -\big|\big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T\delta\big|\\
&\geqslant -\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty\|\delta\|_1\\
&\geqslant -\frac{\lambda}{nc}\big(\|\delta_{\mathcal{S}}\|_1 + \|\delta_{\mathcal{S}^C}\|_1\big).
\end{aligned}$$

Combining (A.6.6) and (A.6.7) yields

$$\begin{aligned}
-\big(\|\delta_{\mathcal{S}}\|_1 + \|\delta_{\mathcal{S}^C}\|_1\big)/c &\leqslant \|\delta_{\mathcal{S}}\|_1 - \|\delta_{\mathcal{S}^C}\|_1\\
\|\delta_{\mathcal{S}^C}\|_1 &\leqslant \frac{c+1}{c-1}\|\delta_{\mathcal{S}}\|_1 = \mathbf{c}\|\delta_{\mathcal{S}}\|_1. && \text{(A.6.8)}
\end{aligned}$$

Invoking the result from Minoration Lemma A.6.7 with $A = \Delta_{\mathbf{c}}$, we obtain

$$\begin{aligned}
\Big\{\frac{1}{3}\big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n}^2\Big\} \wedge \Big\{\frac{\overline{q}_{A,n}}{3}\big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n}\Big\} & \\
\leqslant\ & \Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) - \big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T\delta \qquad \text{(A.6.9)}\\
\leqslant\ & \frac{\lambda}{n}\big(\|\delta_{\mathcal{S}}\|_1 - \|\delta_{\mathcal{S}^C}\|_1\big) + \big|\big\{\nabla_\Theta \Lambda(\Theta_0)\big\}^T\delta\big| \\
\leqslant\ & \frac{\lambda}{n}\big(\|\delta_{\mathcal{S}}\|_1 - \|\delta_{\mathcal{S}^C}\|_1\big) + \|\nabla_\Theta \Lambda(\Theta_0)\|_\infty\|\delta\|_1 \\
\leqslant\ & \frac{\lambda}{n}\big(\|\delta_{\mathcal{S}}\|_1 - \|\delta_{\mathcal{S}^C}\|_1\big) + \frac{\lambda}{nc}\big(\|\delta_{\mathcal{S}}\|_1 + \|\delta_{\mathcal{S}^C}\|_1\big) \\
\leqslant\ & \frac{(1+c)}{c}\frac{\lambda}{n}\|\delta_{\mathcal{S}}\|_1 \\
\leqslant\ & \frac{(1+c)}{c}\frac{\lambda}{n}\sqrt{s}\|\delta_{\mathcal{S}}\|, \\
\leqslant\ & \frac{(1+c)}{c}\frac{\lambda}{n}\sqrt{s}\frac{\big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n}}{\kappa_{\min,n}(\mathbf{c})},
\end{aligned}$$

where we used relations (A.6.6), (A.6.7), Hölder's inequality, $\frac{(c-1)}{c}\frac{\lambda}{n}\|\delta_{\mathcal{S}^C}\|_1 \geqslant 0$, and $\|\delta_{\mathcal{S}}\|_1 \leqslant \sqrt{|\mathcal{S}|}\,\|\delta_{\mathcal{S}}\|$.

Provided that $\overline{q}_{\Delta_{\mathbf{c}},n} > 3\frac{(1+c)}{c}\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}$, the quadratic term on the left-hand side of (A.6.9) must be the minimum, which verifies the first claim (A.6.3) that

$$\big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n} \leqslant 3\frac{(1+c)}{c}\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}.$$

In order to show (A.6.4) note that for $\delta \in \Delta_{\mathbf{c}}$ we have $\|\delta_{\mathcal{S}^C}\|_1 \leqslant \mathbf{c}\|\delta_{\mathcal{S}}\|_1$ and therefore

$$\|\delta\|_1 = \|\delta_{\mathcal{S}}\|_1 + \|\delta_{\mathcal{S}^C}\|_1 \leqslant (1+\mathbf{c})\|\delta_{\mathcal{S}}\|_1 \leqslant (1+\mathbf{c})\sqrt{s}\|\delta_{\mathcal{S}}\| \qquad \text{(A.6.10)}$$

such that

$$\|\widehat{\Theta} - \Theta_0\|_1 \leqslant \big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n}\frac{(1+\mathbf{c})\sqrt{s}}{\kappa_{\min,n}(\mathbf{c})}.$$

Lastly, by relation (A.6.6)

$$\begin{aligned}
\Lambda(\widehat{\Theta}) - \Lambda(\Theta_0) \ &\leqslant\ \frac{\lambda}{n}\|\delta_{\mathcal{S}}\|_1 - \frac{\lambda}{n}\|\delta_{\mathcal{S}^C}\|_1 \ \leqslant\ \frac{\lambda}{n}\sqrt{s}\|\delta_{\mathcal{S}}\| \\
&\leqslant\ \frac{\lambda}{n}\sqrt{s}\big\|\sqrt{w_i}(\tilde{x}_i,\,-y_i)\delta\big\|_{2,n}/\kappa_{\min,n}(\mathbf{c}) \\
&\leqslant\ 3\frac{(1+c)}{c}\left\{\frac{\lambda\sqrt{s}}{n\,\kappa_{\min,n}(\mathbf{c})}\right\}^2,
\end{aligned}$$

which establishes (A.6.5). ■

A.6.4. Sparsity of Logistic Tobit Lasso

Lemma A.6.12 *(Bounds for $|g(\hat{t}) - g(t_0)|$ – Essentially Lemma 13 in Belloni et al. (2016b)). Let $g(t)$ be defined as in (1.2.18). The score function of the logistic Tobit satisfies*

$$|g(t + t_0) - g(t_0)| \leqslant -g'(t_0)[\exp\{|t|\} - 1].$$

Moreover, if $|t| \leqslant 1$, we have $\exp\{|t|\} - 1 \leqslant 2|t|$.

Proof. By definition of function $g(t)$ in (1.2.18), it follows that

$$\frac{|g''(t)|}{-g'(t)} = \left|\frac{\exp\{t\} - 1}{1 + \exp\{t\}}\right| \leqslant 1$$

and, therefore, $-1 \leqslant \mathrm{d}/\mathrm{d}s \log(-g'(s)) = g''(s)/g'(s) \leqslant 1$. Suppose $s \geqslant 0$. By the fundamental theorem of calculus, we get $\int_{t_0}^{s+t_0} \mathrm{d}/\mathrm{d}r \log(-g'(r))\,\mathrm{d}r = \log(-g'(s + t_0)) - \log(-g'(t_0))$, which implies

$$\begin{array}{rcl} -s \leqslant & \log(-g(s + t_0)) - \log(-g'(t_0)) & \leqslant s \\ -g'(t_0)\exp\{-s\} \leqslant & -g(s + t_0) & \leqslant -g'(t_0)\exp\{s\}. \end{array}$$

Integrating over s from 0 to t, we have

$$-g'(t_0)[1 - \exp\{-t\}] \leqslant -g(t + t_0) + g(t_0) \leqslant -g'(t_0)[\exp\{t\} - 1].$$

The result follows, since $1 - \exp\{-|t|\} \leqslant \exp\{|t|\} - 1$ and, thus, $| - g(t + t_0) + g(t_0)| \leqslant -g'(t_0)[\exp\{|t|\} - 1]$. Furthermore, as the exponential function is convex, for values of t such that $|t| \leqslant 1$, the exponential is bounded from above by the secant line connecting the points $(0, 1)$ and $(1, \exp\{1\})$:

$$\exp\{|t|\} \leqslant \exp\{0\} + [\exp\{1\} - \exp\{0\}] \cdot |t| \leqslant 1 + 1.72|t|. \qquad \blacksquare$$

Lemma A.6.13 *(Sparsity). Let $\widehat{\Theta}$ be defined as in (A.6.1), $\widehat{\mathcal{S}}^* = \mathrm{support}((\hat{\alpha},\, \hat{\beta}^T)^T)$, $\hat{s} = |\widehat{\mathcal{S}}^*|$, $\tilde{x}_i := (d_i,\, x_i)$, and $\delta = (\delta_{\tilde{x}}^T,\, \delta_y)^T = (\delta_d,\, \delta_x^T,\, \delta_y)^T$. Assume $\lambda/n \geqslant c\|\nabla_\Theta\, \Lambda(\Theta_0)\|_\infty$. Then*

$$\hat{s} \leqslant \left\{cn/((c-1)\lambda)\right\}^2 \phi_{\max,n}^{u*}(\hat{s})\|(\tilde{x}_i,\, -y_i)(\widehat{\Theta} - \Theta_0)\|_{2,n}^2.$$

Additionally, if $\overline{q}_{\Delta_{\mathbf{c}},n} > 3\frac{(1+c)}{c}\frac{\lambda\sqrt{s+1}}{n\,\kappa_{\min,n}(\mathbf{c})}$

$$\hat{s} \leqslant s\frac{9\mathbf{c}^2}{\{\kappa_{\min,n}(\mathbf{c})\}^2}\frac{\phi_{\max,n}^{u*}(\hat{s})}{\{\nu_{(r),n}(\mathbf{c})\}^2}.$$

Moreover, if $6\mathbf{c}\frac{\lambda s}{n\{\kappa_{\min,n}(\mathbf{c})\}^2}\max_{1\leqslant i\leqslant n}\|(\tilde{x}_i,\, y_i)\|_\infty \leqslant 1$*, we have*

$$\hat{s} \leqslant s\frac{36\mathbf{c}^2\phi^{u*}_{\max,n}(\hat{s})}{\{\kappa_{\min,n}(\mathbf{c})\}^2}.$$

Proof. Essentially, we adjust the proof of Lemma 14 in Belloni et al. (2016b). Define $\hat{g}_i := g(\hat{\gamma}y_i - \hat{\alpha}d_i - x_i\hat{\beta})$ and $g_{0i} := g(\gamma_0 y_i - \alpha_0 d_i - x_i\beta_0)$. By the Karush-Kuhn-Tucker conditions associated with (A.6.1), for any $j \in \widehat{\mathcal{S}}^*$ it holds that $\lambda/n = |\nabla_{\Theta_j}\Lambda(\widehat{\Theta})| = |\mathbb{E}_n[\hat{g}_i\tilde{x}_{ij}]|$. Thus, $\hat{s}\,\lambda/n = \|\mathbb{E}_n[\hat{g}_i\tilde{x}_{i\widehat{\mathcal{S}}^*}]\|_1$. Therefore,

$$\begin{aligned}
\sqrt{\hat{s}}\,\frac{\lambda}{n} &= \big\|\mathbb{E}_n[\hat{g}_i\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\|_1/\sqrt{\hat{s}} \leqslant \big\|\mathbb{E}_n[\hat{g}_i\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| \\
&\leqslant \big\|\mathbb{E}_n[g_{0i}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| + \big\|\mathbb{E}_n[\{\hat{g}_i - g_{0i}\}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| \\
&\leqslant \sqrt{\hat{s}}\,\big\|\mathbb{E}_n[g_{0i}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\|_\infty + \big\|\mathbb{E}_n[\{(y_i,\, -\tilde{x}_i)\delta\}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| \\
&\leqslant \sqrt{\hat{s}}\,\frac{\lambda}{cn} + \sqrt{\phi^{u*}_{\max,n}(\hat{s})}\|(\tilde{x}_i,\, -y_i)\delta\|_{2,n},
\end{aligned}$$

where we used $\|\cdot\|_1 \leqslant \sqrt{\hat{s}}\,\|\cdot\| \leqslant \hat{s}\,\|\cdot\|_\infty$, $\big\|\mathbb{E}_n[g_{0i}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\|_\infty \leqslant \|\nabla_\Theta\Lambda(\Theta_0)\|_\infty \leqslant \lambda/(cn)$, $|\hat{g}_i - g_{0i}| \leqslant |(y_i,\, -\tilde{x}_i)(\widehat{\Theta} - \Theta_0)|$, and definition A.6.3.

The first assertion follows by re-arranging the terms:

$$\hat{s} \leqslant \{cn/((c-1)\lambda)\}^2\phi^{u*}_{\max,n}(\hat{s})\|(\tilde{x}_i,\, -y_i)\delta\|^2_{2,n}.$$

The second result follows from applying Definition A.6.4 and bound (A.6.3) in Lemma A.6.11:

$$\begin{aligned}
\hat{s} &\leqslant \frac{(cn)^2\phi^{u*}_{\max,n}(\hat{s})}{((c-1)\lambda)^2}\|(\tilde{x}_i,\, -y_i)\delta\|^2_{2,n} \\
&\leqslant \frac{(cn)^2\phi^{u*}_{\max,n}(\hat{s})}{((c-1)\lambda)^2}\frac{\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|^2_{2,n}}{\{\nu_{(r),n}(\mathbf{c})\}^2} \\
&\leqslant s\frac{9\mathbf{c}^2\phi^{u*}_{\max,n}(\hat{s})}{\{\kappa_{\min,n}(\mathbf{c})\}^2\{\nu_{(r),n}(\mathbf{c})\}^2}.
\end{aligned}$$

Applying the same logic underlying the proof of the first assertion above, we get

$$\begin{aligned}
\sqrt{\hat{s}}\,\frac{\lambda}{n} &= \big\|\mathbb{E}_n[\hat{g}_i\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\|_1/\sqrt{\hat{s}} \leqslant \big\|\mathbb{E}_n[\hat{g}_i\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| \\
&\leqslant \big\|\mathbb{E}_n[g_{0i}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| + \big\|\mathbb{E}_n[\{\hat{g}_i - g_{0i}\}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\| \\
&\leqslant \sqrt{\hat{s}}\,\big\|\mathbb{E}_n[g_{0i}\tilde{x}_{i\widehat{\mathcal{S}}^*}]\big\|_\infty + \sup_{\|\xi\|_0\leqslant\hat{s},\|\xi\|=1}\mathbb{E}_n[|g_{0i} - \hat{g}_i|\,|\tilde{x}_i\xi|] \\
&\leqslant \sqrt{\hat{s}}\,\frac{\lambda}{cn} + 2\sqrt{\phi^{u*}_{\max,n}(\hat{s})}\|\sqrt{w_i}(\tilde{x}_i,\, -y_i)\delta\|_{2,n},
\end{aligned}$$

where the last line follows from Lemma A.6.12 applied to

$$\big|g_{0i} - \hat{g}_i\big| = \big|g((y_i,\, -\tilde{x}_i)\{\delta + \Theta_0\}) - g((y_i,\, -\tilde{x}_i)\Theta_0)\big| \leqslant -2\,g'((y_i,\, -\tilde{x}_i)\Theta_0)\big|(\tilde{x}_i,\, -y_i)\delta\big|,$$

the definition of weights $w_i = -g'((\tilde{x}_i, -y_i)\Theta_0) = -g'((y_i, -\tilde{x}_i)\Theta_0)$ in (1.2.9), the fact that $0 < w_i < 1/2$ which implies $\sqrt{w_i} \geqslant w_i$, and bound (A.6.4) in Lemma A.6.11, by which we have $\|\delta\|_1 \leqslant 6\mathbf{c}\frac{\lambda s}{n\{\kappa_{\min,n}(\mathbf{c})\}^2}$ such that $\left|(\tilde{x}_i, -y_i)\delta\right| \leqslant \max_{1\leqslant i\leqslant n}\|(\tilde{x}_i, y_i)\|_\infty\|\delta\|_1 \leqslant 1$ by the assumed side-condition above. Using bound (A.6.3) in Lemma A.6.11, we obtain

$$\hat{s} \leqslant s\frac{36\mathbf{c}^2\phi^{u*}_{\max,n}(\hat{s})}{\{\kappa_{\min,n}(\mathbf{c})\}^2}. \qquad \blacksquare$$

A.6.5. Post-model Selection Rate for Logistic Tobit

Lemma A.6.14 *(Bound for Post-selection Prediction Norm $\|\cdot\|_{2,n}$). Let $\tilde{\Theta}$ be defined as in (A.6.2) and $\tilde{s} := |\tilde{\mathcal{S}}|$. Assume $\overline{q}_{A,n} > 6\sqrt{s+\tilde{s}}\,\|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty\{\phi_{\min,n}(s+\tilde{s})\}^{-1/2}$ and $\overline{q}_{A,n} > 6\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}^{1/2}$ hold for $\tilde{\delta} \in A = \{\delta \in \mathbb{R}^{p+2} : \|\delta\|_0 \leqslant \tilde{s}+s\}$. Then, we have*

$$\|\sqrt{w_i}(\tilde{x}_i, -y_i)(\tilde{\Theta} - \Theta_0)\|_{2,n} \leqslant \sqrt{3\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}} + \sqrt{s+\tilde{s}}\,\frac{3\|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty}{\sqrt{\phi_{\min,n}(s+\tilde{s})}}.$$

Proof. Essentially, we adjust the proof of Lemma 15 in Belloni et al. (2016b). Let $\tilde{\delta} = \tilde{\Theta} - \Theta_0$ and $\tilde{t}_{2,n} = \|\sqrt{w_i}(\tilde{x}_i, -y_i)\tilde{\delta}\|_{2,n}$. Invoking the result from Minoration Lemma A.6.7 with $A = \{\delta \in \mathbb{R}^{p+2} : \|\delta\|_0 \leqslant \tilde{s}+s\}$, we have

$$\begin{aligned}\left\{\frac{\tilde{t}^2_{2,n}}{3}\right\} \vee \left\{\frac{\overline{q}_{A,n}\tilde{t}_{2,n}}{3}\right\} &\leqslant \Lambda(\tilde{\Theta}) - \Lambda(\Theta_0) - \nabla_\Theta\,\Lambda(\Theta_0)^T\delta \qquad \text{(A.6.11)}\\ &\leqslant \max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\} + \|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty\sqrt{s+\tilde{s}}\,\|\tilde{\delta}\| \\ &\leqslant \max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\} + \sqrt{s+\tilde{s}}\,\frac{\|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty\tilde{t}_{2,n}}{\sqrt{\phi_{\min,n}(s+\tilde{s})}},\end{aligned}$$

because of $\|\tilde{\Theta}\|_0 = \tilde{s}$, $\|\Theta_0\|_0 = |\mathcal{S}| = s$, and Definition A.6.3. Now, consider two separate cases.

Case 1: if $\overline{q}_{A,n} > \tilde{t}_{2,n}$, the minimum of the left-hand side of (A.6.11) must be the quadratic term. Therefore, we have

$$\tilde{t}^2_{2,n} \leqslant 3\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\} + 3\,\tilde{t}_{2,n}\sqrt{s+\tilde{s}}\,\frac{\|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty}{\sqrt{\phi_{\min,n}(s+\tilde{s})}}$$

and the result follows since for $a, b, c \geqslant 0$ the relation $a^2 \leqslant b + ac$ implies $a \leqslant \sqrt{b+ac} \leqslant \sqrt{b} + \sqrt{ac} \leqslant \sqrt{b} + c$ as, in particular, $a \leqslant c$.

Case 2: if $\overline{q}_{A,n} \leqslant \tilde{t}_{2,n}$, the minimum of the left-hand side of (A.6.11) must be the linear term. Therefore, provided that $\overline{q}_{A,n} > 6\sqrt{s+\tilde{s}}\,\|\nabla_\Theta\,\Lambda(\Theta_0)\|_\infty\{\phi_{\min,n}(s+\tilde{s})\}^{-1/2}$ and $\overline{q}_{A,n} >$

$6\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}^{1/2}$, we have

$$\begin{aligned}
\frac{\overline{q}_{A,n}\tilde{t}_{2,n}}{3} &\leqslant \max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\} + \tilde{t}_{2,n}\sqrt{s+\tilde{s}}\,\frac{\|\nabla_\Theta \Lambda(\Theta_0)\|_\infty}{\sqrt{\phi_{\min,n}(s+\tilde{s})}} \\
&\leqslant \frac{\overline{q}_{A,n}}{6}\sqrt{\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}} + \frac{\overline{q}_{A,n}}{6}\tilde{t}_{2,n} \\
\tilde{t}_{2,n} &\leqslant \sqrt{\max\{\Lambda(\tilde{\Theta}) - \Lambda(\Theta_0), 0\}}
\end{aligned}$$

and the result immediately follows. ■

Chapter B

SUPPLEMENTAL MATERIAL TO CHAPTER 2: ``USING POST-REGULARIZATION DISTRIBUTION REGRESSION TO MEASURE THE EFFECTS OF A MINIMUM WAGE ON HOURLY WAGES, HOURS WORKED AND MONTHLY EARNINGS´´

B.1. CHOICE OF PENALTY LEVELS AND LOADINGS

In order to allow for a different choice of observation clusters when computing penalties as opposed to computing standard errors and confidence intervals, we index clusters by $g \in \{1, \ldots, G^*\}$, $G^* < n$ with observations $i \in \{1, \ldots, n_g^*\}$ so that $\sum_{g=1}^{G^*} n_g^* = n$. In addition, define the $(\tilde{p}+p)$-dimensional vector $\tilde{X} = (D,\, X)$ and the $((\tilde{p}-1)+p)$-dimensional vector $\tilde{X}^j = (D_{\mathcal{J}\setminus j},\, X)$ for $j \in \mathcal{J} = \{1, \ldots, \tilde{p}\}$. Algorithms B.1.6 and B.1.7 are essentially due to Belloni et al. (2018b) with modifications for the clustering of observations as in Chiang (2020), and deterministic sampling weights v_{ig}.

Algorithm B.1.6 *(Penalty Level and Loadings for Post-ℓ_1 Logit – essentially Algorithm 3 in Belloni et al., 2018b).*

Step 1. Define an iteration limit $\bar{m} \geqslant 1$ and a small tolerance $\varepsilon > 0$. Initialize $m = 1$ and let $\overline{Y}^u := n^{-1} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig} Y_{ig}^u$. Compute $\| \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig} (Y_{ig}^u - \overline{Y}^u) \tilde{X}_{ig}^T \|_{\infty,k}$, where $\|\xi\|_{\infty,k}$ for some vector $\xi \in \mathbb{R}^{p+\tilde{p}}$ and integer $1 \leqslant k \leqslant \bar{k}$ returns the k largest elements by absolute value of this vector. The associated indices $\mathcal{S}_u^*$ with $|\mathcal{S}_u^*| = k$ represent the ex-ante, k most promising candidates to join the active set.
Run a logistic regression of Y_{ig}^u on $(D_{ig},\, X_{ig})_{\mathcal{S}_u^*}$:

$$\begin{aligned} (\tilde{\theta}_u, \tilde{\beta}_u)^{[0]} &\in \underset{\theta,\,\beta}{\arg\min} \frac{1}{G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \Lambda_u(W_{ig}, \theta, \beta) \;:\; \text{support}(\theta, \beta) \subseteq \mathcal{S}_u^*, \\ \Lambda_u(W_{ig}, \theta, \beta) &= \log(1 + \exp\{D_{ig}\theta + X_{ig}\beta\}) - Y_{ig}^u \cdot (D_{ig}\theta + X_{ig}\beta). \end{aligned}$$

For $l \in \{1, \ldots, \tilde{p} + p\}$ compute the initial penalty loadings according to

$$\big(\widehat{\Psi}_u^{[0]}\big)_{ll} = \left\{ \frac{1}{G^*} \sum_{g=1}^{G^*} \Big[\sum_{i=1}^{n_g^*} v_{ig}\, (Y_{ig}^u - \Lambda(D_{ig}\tilde{\theta}_u^{[0]} + X_{ig}\tilde{\beta}_u^{[0]}))\tilde{X}_{igl} \Big]^2 \right\}^{1/2}.$$

Step 2. Run the post-ℓ_1 Logit based on penalty level $\lambda_1 = 1.1\sqrt{G^*}\,\Phi^{-1}\big(1 - 0.05/\{\log(G^*)(\tilde{p} + p)\}\big)$ and the diagonal matrix $\widehat{\Psi}_u^{[m-1]} := \text{diag}\big(\big(\widehat{\Psi}_u^{[m-1]}\big)_{ll}\ \forall\ l \in \{1, \ldots, \tilde{p} + p\}\big)$:

$$\begin{aligned} (\hat{\theta}_u, \hat{\beta}_u)^{[m]} &\in \underset{\theta,\beta}{\arg\min} \frac{1}{G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \Lambda_u(W_{ig}, \theta, \beta) + \frac{\lambda_1}{G^*} \big\| \widehat{\Psi}_u^{[m-1]} (\theta^T,\, \beta^T)^T \big\|_1, \\ (\tilde{\theta}_u, \tilde{\beta}_u)^{[m]} &\in \underset{\theta,\,\beta}{\arg\min} \frac{1}{G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \Lambda_u(W_{ig}, \theta, \beta) \;:\; \text{support}(\theta, \beta) \subseteq \widehat{\mathcal{S}}_u^{[m]}, \\ \Lambda_u(W_{ig}, \theta, \beta) &= \log(1 + \exp\{D_{ig}\theta + X_{ig}\beta\}) - Y_{ig}^u \cdot (D_{ig}\theta + X_{ig}\beta), \end{aligned}$$

where $\widehat{\mathcal{S}}_u^{[m]} = \text{support}((\hat{\theta}_u, \hat{\beta}_u)^{[m]})$.
For $l \in \{1, \ldots, \tilde{p} + p\}$ compute the updated penalty loadings according to

$$\big(\widehat{\Psi}_u^{[m]}\big)_{ll} = \left\{ \frac{1}{G^*} \sum_{g=1}^{G^*} \Big[\sum_{i=1}^{n_g^*} v_{ig}\, (Y_{ig}^u - \Lambda(D_{ig}\tilde{\theta}_u^{[m]} + X_{ig}\tilde{\beta}_u^{[m]}))\tilde{X}_{igl} \Big]^2 \right\}^{1/2}.$$

If $m < \bar{m}$ and $\max_l \big|\big(\widehat{\Psi}_u^{[m]}\big)_{ll} - \big(\widehat{\Psi}_u^{[m-1]}\big)_{ll}\big| > \varepsilon$, update $m \leftarrow m + 1$ and repeat Step 2. Otherwise stop.

Algorithm B.1.7 *(Penalty Level and Loadings for WLS Lasso – essentially Algorithm 4 in Belloni et al., 2018b).*

Step 1. Define an iteration limit $\bar{m} \geqslant 1$ and a small tolerance $\varepsilon > 0$. Initialize $m = 1$ and let $\overline{D}_j := n^{-1} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig} \widehat{f}_{u,ig} D_{igj}$ for all $j \in \mathcal{J}$. Compute $\| \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig} \widehat{f}_{u,ig}^2 (D_{igj} - \overline{D}_j) \tilde{X}_{ig}^j \|_{\infty,k}$, where $\|\xi\|_{\infty,k}$ for some vector $\xi \in \mathbb{R}^{\tilde{p}+p-1}$ and integer $1 \leqslant k \leqslant \bar{k}$ returns the k largest elements by absolute value of this vector. The associated indices $\mathcal{S}_u^\diamond$ with $|\mathcal{S}_u^\diamond| = k$ represent the ex-ante, k most promising candidates to join the active set.
Run a weighted least squares estimation of $\widehat{f}_{ig} D_j$ on $\widehat{f}_{u,ig} \tilde{X}_{ig\mathcal{S}_u^\diamond}^j$:

$$\big(\tilde{\gamma}_u^j\big)^{[0]} \in \arg\min_\gamma \frac{1}{G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \widehat{f}_{u,ig}^2 \big(D_{igj} - \tilde{X}_{ig}^j \gamma\big)^2 \ : \ \text{support}(\gamma) \subseteq \mathcal{S}_u^\diamond.$$

For $l \in \{1, \ldots, \tilde{p} + p - 1\}$ compute the initial penalty loadings according to

$$\big(\widehat{\Psi}_{uj}^{[0]}\big)_{ll} = \bigg\{ \frac{1}{G^*} \sum_{g=1}^{G^*} \Big[\sum_{i=1}^{n_g^*} v_{ig}\, \widehat{f}_{u,ig}^2 \, \big(D_{igj} - \tilde{X}_{ig}^j \big(\tilde{\gamma}_u^j\big)^{[0]}\big) \tilde{X}_{igl}^j \Big]^2 \bigg\}^{1/2}.$$

Step 2. Run the weighted post-Lasso procedure based on penalty level $\lambda_2 = 1.1\sqrt{G^*}\, \Phi^{-1}\big(1 - 0.05/\{\log(G^*)(\tilde{p}+p)(\tilde{p}+p-1)\}\big)$ and the diagonal matrix $\widehat{\Psi}_{uj}^{[m-1]} := \text{diag}\big(\big(\widehat{\Psi}_{uj}^{[m-1]}\big)_{ll} \, \forall\, l \in \{1, \ldots, \tilde{p} + p - 1\}\big)$:

$$\begin{aligned}
\big(\hat{\gamma}_u^j\big)^{[m]} \ &\in \ \arg\min_\gamma \frac{1}{2G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \widehat{f}_{u,ig}^2 \big(D_{igj} - \tilde{X}_{ig}^j \gamma\big)^2 + \frac{\lambda_2}{G^*} \big\| \widehat{\Psi}_{uj}^{[m-1]} \gamma \big\|_1, \\
\big(\tilde{\gamma}_u^j\big)^{[m]} \ &\in \ \arg\min_\gamma \frac{1}{G^*} \sum_{g=1}^{G^*} \sum_{i=1}^{n_g^*} v_{ig}\, \widehat{f}_{u,ig}^2 \big(D_{igj} - \tilde{X}_{ig}^j \gamma\big)^2 \ : \ \text{support}(\gamma) \subseteq \text{support}\big(\big(\hat{\gamma}_u^j\big)^{[m]}\big).
\end{aligned}$$

For $l \in \{1, \ldots, \tilde{p} + p - 1\}$ compute the updated penalty loadings according to

$$\big(\widehat{\Psi}_{uj}^{[m]}\big)_{ll} = \bigg\{ \frac{1}{G^*} \sum_{g=1}^{G^*} \Big[\sum_{i=1}^{n_g^*} v_{ig}\, \widehat{f}_{u,ig}^2 \, \big(D_{igj} - \tilde{X}_{ig}^j \big(\tilde{\gamma}_u^j\big)^{[m]}\big) \tilde{X}_{igl}^j \Big]^2 \bigg\}^{1/2}.$$

If $m < \bar{m}$ and $\max_l \ \big|\big(\widehat{\Psi}_{uj}^{[m]}\big)_{ll} - \big(\widehat{\Psi}_{uj}^{[m-1]}\big)_{ll}\big| > \varepsilon$, update $m \leftarrow m + 1$ and repeat Step 2. Otherwise stop.

B.2. CONSTRUCTION OF FEATURE SET

Table B.2.1: Variables and transformations included in Algorithm 2.2.3.

Variable	Type	Transformations included
Worker characteristics 1		
Gender	categorial(2)	indicators for each category
East/West Germany	categorial(2)	indicators for each category
Worker characteristics 2		
Age	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Years of education	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Full-time experience (years)	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Part-time experience (years)	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Full-time + 0.5 Part-time experience	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Tenure (years)	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$
Overtime (hours/week)	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$, indicator for no overtime
Unemployment experience (years)	continuous	4-th order polynomial, $\sqrt{\cdot}$, $\log(\cdot)$ indicator for no unemployment experience
Worker characteristics 3		
Type of school degree	categorial(9)	indicators for each category
Type of vocational training degree	categorial(7)	indicators for each category
Type of tertiary degree	categorial(11)	indicators for each category
Fine type of tertiary degree	categorial(23)	indicators for each category
Variants of no educational degree	categorial(4)	indicators for each category
ISCED classification of educational degree	categorial(10)	indicators for each category
5-group categorization German education system	categorial(5)	indicators for each category
3-group categorization German education system	categorial(3)	indicators for each category
ISCO08 occupation code (2-digit)	categorial(40)	indicators for each category

Continued on following page

Table B.2.1, continued

Variable	Type	Transformations included
ISCO08 occupation code (3-digit)	categorial(121)	indicators for each category
KldB2010 occupation code (1-digit)	categorial(10)	indicators for each category
KldB2010 occupation code (2-digit)	categorial(37)	indicators for each category
Occupational position	categorial(12)	indicators for each category
NACE industry code (1-digit)	categorial(18)	indicators for each category
NACE industry code (2-digit)	categorial(86)	indicators for each category
Full-time/part-time/marginal part-time	categorial(3)	indicators for each category indicator for part-time/marginal part-time combined
Minjob contract	categorial(2)	indicators for each category
Firm size categorization I (coarse)	categorial(5)	indicators for each category
Firm size categorization II (finer)	categorial(8)	indicators for each category
Public sector	categorial(3)	indicators for each category
Federal state	categorial(16)	indicators for each category
Urban area	categorial(2)	indicators for each category
Nationality (continents)	categorial(5)	indicators for each category
Nationality (subcontinents)	categorial(12)	indicators for each category
Nationality (countries)	categorial(116)	indicators for each category indicator for German nationality
Household size	count(16)	1st power, indicator for each category
Partner lives in household	categorial(3)	indicators for each category
Marital status (single/divorced/widowed etc.)	categorial(9)	indicators for each category
Number of children in household aged 0-2 years	count	1st power, indicator for zero
Number of children in household aged 3-5 years	count	1st power, indicator for zero
Number of children in household aged 6-11 years	count	1st power, indicator for zero
Number of children in household aged 12-17 years	count	1st power, indicator for zero
Person in need of care lives in household	categorial(3)	indicators for each category
Homeowner/renter (with subcategories)	categorial(5)	indicators for each category

Continued on following page

Table B.2.1, continued

Variable	Type	Transformations included
Health indicator	ordinal(6)	1st power, indicator for highest two values indicator for lowest two values
Interactions		
Age × Household size	full expansion of age transformations with household size	
Age × (Worker characteristics 2)	full expansion of features with age transformations	
Household size × (Worker characteristics 2)	full expansion of features with household size	
Gender × (Worker characteristics 2)	full expansion of features with gender indicators	
Gender × (Worker characteristics 3)	full expansion of features with gender indicators	
East/West × (Worker characteristics 2)	full expansion of features with East/West indicators	
East/West × (Worker characteristics 3)	full expansion of features with East/West indicators	

Note: This table summarizes and categorizes all variables, their transformations and interactions that were used to construct the set of potential controls. For categorial variables, the number of categories is reported in brackets.

Chapter C

SUPPLEMENTAL MATERIAL TO CHAPTER 3: ``SELECTIVITY CORRECTED WAGE DISTRIBUTIONS AND THE EVOLUTION OF THE GERMAN GENDER WAGE GAP´´

C.1. MULTIPLIER BOOTSTRAP

To obtain standard errors and uniform confidence bands of factual and counterfactual distributions as well as of the contributions of individual decomposition factors, we apply the multiplier bootstrap procedure described in Algorithm C.1.8; see Chernozhukov et al. (2013, 2020, 2023). These steps apply to observations independent across indices $i \in \{1, \dots, N\}$, but could be adjusted in the following manner to account for clusters of observations. Let $\{1, \dots, G\}$ be a set of cluster identifiers with $G < N$, and N_g the associated number of observations in cluster $g \in \{1, \dots, G\}$ so that $\sum_{g=1}^{G} N_g = N$. Whenever sample averages over N data points are computed, the single sum operator is replaced by a double summation. For example, in Step 2 of Algorithm C.1.8, we have $\hat{\theta}_y^b = \hat{\theta}_y + N^{-1} \sum_{g=1}^{G} \omega_g^b \cdot \sum_{i=1}^{N_g} \widehat{\psi}_{ig}(\hat{\theta}_y)$, such that all observations belonging to the same cluster g are weighted by the same bootstrap multiplier ω_g^b. Standard errors have to be adjusted accordingly, by collapsing the blocks of observations for all clusters.

Algorithm C.1.8 *(Multiplier Bootstrap and Uniform Inference with the LGR).*

Step 1. Define a bootstrap replication limit $B \geqslant 2$ and let quantities estimated on a bootstrap sample be superscripted by $b \in \{1, \ldots, B\}$. For each bootstrap replication b draw a random sequence of multiplier weights $(\tilde{\omega}_i^b)_{1 \leqslant i \leqslant N}$ from the standard exponential distribution. Additionally, generate the centred sequence:

$$\omega_i^b = \tilde{\omega}_i^b - \frac{1}{N} \sum_{i=1}^{N} \tilde{\omega}_i^b.$$

For a fine grid of thresholds $y \in \mathcal{Y}' \subseteq \mathcal{Y}$ obtain the bootstrap estimates:

$$\hat{\theta}_y^b = \hat{\theta}_y + \frac{1}{N} \sum_{i=1}^{N} \omega_i^b \cdot \widehat{\psi}_i(\hat{\theta}_y),$$

where $\widehat{\psi}_i(\hat{\theta}_y)$ represents the vector of influence functions for observation i associated with the LGR model parameters $\hat{\theta}_y = \big(\hat{\beta}(y)^T,\, \hat{\pi}^T,\, \hat{\rho}(y)^T\big)^T$ at threshold y.

Step 2. Given the estimated LGR parameters and the sample distribution of observables $\widehat{F}_Z(z) = N^{-1} \sum_{i=1}^{N} \mathbf{1}\{Z_i \leq z\}$, let some functional of interest be denoted as $\varphi(\hat{\theta}_y, \widehat{F}_Z)$. Compute the bootstrap realisation of the supremum statistic by:

$$t_{\mathcal{Y}}^b = \max_{y \in \mathcal{Y}'} \frac{\big|\varphi(\hat{\theta}_y^b, \widehat{F}_Z^b) - \varphi(\hat{\theta}_y, \widehat{F}_Z)\big|}{SE\big(\varphi(\hat{\theta}_y, \widehat{F}_Z)\big)},$$

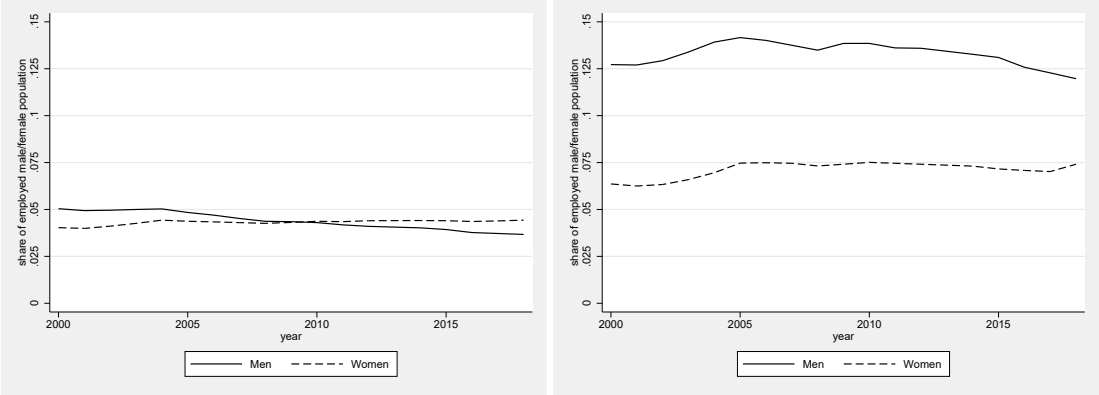
where $\varphi(\hat{\theta}_y^b, \widehat{F}_Z^b)$ is the bootstrapped analogue of $\varphi(\hat{\theta}_y, \widehat{F}_Z)$ based on bootstrap estimate $\hat{\theta}_y^b$ and the bootstrapped distribution of observables $\widehat{F}_Z^b(z) = N^{-1} \sum_{i=1}^{N} \tilde{\omega}_i^b \mathbf{1}\{Z_i \leq z\}$. Examples for such functionals are (3.3.7), (3.3.9) and (3.5.3).

Step 3. Choose a confidence level $\alpha \in (0, 1)$ and compute the associated interval such that

$$\mathrm{P}\Big(\varphi(\theta_y, F_Z) \in \big\{\varphi(\hat{\theta}_y, \widehat{F}_Z) \pm c_\alpha \cdot SE\big(\varphi(\hat{\theta}_y, \widehat{F}_Z)\big)\big\}\ \forall\, y \in \mathcal{Y}'\Big) \to (1 - \alpha)\ as\ N \to \infty,$$

where the critical value c_α is obtained as the $(1 - \alpha)$-quantile of the bootstrapped distribution of the supremum statistic $t_{\mathcal{Y}}$ above.

C.2. SELF-EMPLOYMENT AND CIVIL SERVANTS



(a) Civil servants (b) Self employment

Figure C.1: Shares of civil servants and self-employed by gender

Note: These graphs illustrate the evolution of the shares of civil servants and self-employed individuals from 2000 to 2017 differentiated by gender.

C.3. OBSERVED CHARACTERISTICS

Table C.3.1: Summary statistics for selection into full-time employment

	Men		Women	
	2000-2005	2012-2017	2000-2005	2012-2017
Selected individuals	$N = 1,458,124$	$N = 1,492,091$	$N = 816,786$	$N = 739,634$
Age	39.503 [10.160]	41.277 [11.092]	38.117 [10.874]	39.438 [11.876]
Full-time work experience (years)	14.510 [8.352]	17.309 [10.597]	11.207 [7.458]	13.325 [9.615]
Full-time work experience = 0	0.016 [0.125]	0.020 [0.140]	0.030 [0.169]	0.036 [0.187]

Continued on following page

Table C.3.1, continued

	Men		Women	
	2000-2005	2012-2017	2000-2005	2012-2017
Part-time work experience (years)	0.176 [0.747]	0.375 [1.241]	0.982 [2.559]	1.645 [3.500]
Part-time work experience = 0	0.871 [0.335]	0.774 [0.418]	0.683 [0.465]	0.565 [0.496]
Minijob work experience (years)	0.060 [0.296]	0.813 [1.845]	0.133 [0.478]	1.317 [2.355]
Minijob work experience = 0	0.945 [0.296]	0.612 [0.487]	0.878 [0.327]	0.496 [0.500]
Lower/middle secondary school + voc. training	0.680 [0.466]	0.599 [0.490]	0.639 [0.480]	0.510 [0.500]
Upper secondary school only	0.012 [0.110]	0.022 [0.148]	0.021 [0.144]	0.039 [0.193]
Upper secondary school + voc. training	0.071 [0.257]	0.112 [0.316]	0.116 [0.320]	0.179 [0.383]
"Fachhochschule" degree	0.054 [0.226]	0.022 [0.146]	0.040 [0.196]	0.024 [0.153]
University degree	0.091 [0.288]	0.173 [0.378]	0.071 [0.257]	0.175 [0.380]
German nationality	0.921 [0.270]	0.912 [0.283]	0.944 [0.230]	0.930 [0.254]
Non-selected individuals	$N = 1,329,506$	$N = 1,282,554$	$N = 1,707,996$	$N = 1,788,235$
Age	40.359 [11.823]	42.704 [11.876]	40.502 [10.784]	42.960 [11.179]
Full-time work experience (years)	6.180 [7.172]	6.924 [8.279]	5.184 [5.557]	6.498 [6.673]
Full-time work experience = 0	0.177 [0.382]	0.157 [0.364]	0.194 [0.395]	0.152 [0.360]
Part-time work experience (years)	0.310 [1.221]	0.742 [2.085]	2.471 [4.625]	3.743 [5.691]
Part-time work experience = 0	0.824 [0.380]	0.693 [0.461]	0.523 [0.499]	0.377 [0.485]
Minijob work experience (years)	0.133 [0.516]	0.744 [1.716]	0.463 [1.068]	2.079 [3.340]
Minijob work experience = 0	0.886 [0.318]	0.630 [0.483]	0.750 [0.433]	0.431 [0.495]
Lower/middle secondary school + voc. training	0.546 [0.498]	0.484 [0.500]	0.575 [0.494]	0.515 [0.500]

Continued on following page

Table C.3.1, continued

	Men		Women	
	2000-2005	2012-2017	2000-2005	2012-2017
Upper secondary school only	0.063 [0.244]	0.086 [0.280]	0.045 [0.206]	0.063 [0.243]
Upper secondary school + voc. training	0.060 [0.237]	0.085 [0.279]	0.077 [0.266]	0.124 [0.330]
"Fachhochschule" degree	0.030 [0.171]	0.025 [0.157]	0.025 [0.155]	0.023 [0.149]
University degree	0.085 [0.279]	0.125 [0.331]	0.066 [0.247]	0.128 [0.334]
German nationality	0.750 [0.433]	0.767 [0.423]	0.857 [0.350]	0.867 [0.340]

Note: This table shows summary statistics for our set of covariates differentiated by individuals' selection status into full-time employment. Standard deviations are reported in brackets.

Table C.3.2: Summary statistics for selection into part-time employment

	Men		Women	
	2000-2005	2012-2017	2000-2005	2012-2017
Selected individuals	$N = 75,197$	$N = 148,743$	$N = 402,912$	$N = 642,760$
Age	41.387 [13.009]	40.444 [12.083]	42.828 [9.426]	44.324 [10.038]
Full-time work experience (years)	9.856 [10.031]	9.227 [9.998]	6.541 [5.910]	8.644 [7.048]
Full-time work experience = 0	0.150 [0.357]	0.141 [0.348]	0.127 [0.331]	0.084 [0.277]
Part-time work experience (years)	2.779 [3.302]	3.770 [4.124]	7.154 [6.105]	7.924 [6.723]
Part-time work experience = 0	0.290 [0.454]	0.200 [0.400]	0.010 [0.300]	0.076 [0.265]
Minijob work experience (years)	0.180 [0.530]	1.357 [2.178]	0.222 [0.647]	2.216 [3.263]
Minijob work experience = 0	0.827 [0.378]	0.425 [0.494]	0.830 [0.376]	0.397 [0.489]
Lower/middle secondary school + voc. training	0.502 [0.500]	0.464 [0.499]	0.691 [0.462]	0.604 [0.489]
Upper secondary school only	0.088 [0.283]	0.078 [0.268]	0.015 [0.120]	0.020 [0.139]
Upper secondary school + voc. training	0.110 [0.313]	0.133 [0.339]	0.088 [0.283]	0.156 [0.363]

Continued on following page

Table C.3.2, continued

	Men		Women	
	2000-2005	2012-2017	2000-2005	2012-2017
"Fachhochschule" degree	0.055 [0.229]	0.019 [0.137]	0.034 [0.181]	0.023 [0.149]
University degree	0.139 [0.360]	0.200 [0.400]	0.061 [0.240]	0.133 [0.340]
German nationality	0.885 [0.319]	0.858 [0.349]	0.949 [0.221]	0.931 [0.254]
Non-selected individuals	$N = 1,254,309$	$N = 1,133,811$	$N = 1,305,084$	$N = 1,145,475$
Age	40.297 [11.745]	43.000 [11.816]	39.787 [11.072]	42.195 [11.702]
Full-time work experience (years)	5.959 [6.901]	6.622 [7.977]	4.765 [5.374]	5.294 [6.132]
Full-time work experience = 0	0.179 [0.383]	0.159 [0.366]	0.215 [0.411]	0.191 [0.393]
Part-time work experience (years)	0.162 [0.734]	0.345 [1.152]	1.026 [2.763]	1.396 [3.134]
Part-time work experience = 0	0.857 [0.351]	0.758 [0.429]	0.654 [0.476]	0.546 [0498]
Minijob work experience (years)	0.130 [0.515]	0.664 [1.629]	0.537 [1.157]	2.003 [3.380]
Minijob work experience = 0	0.890 [0.313]	0.657 [0.475]	0.726 [0.446]	0.451 [0.498]
Lower/middle secondary school + voc. training	0.549 [0.498]	0.486 [0.500]	0.539 [0.499]	0.465 [0.499]
Upper secondary school only	0.062 [0.241]	0.087 [0.281]	0.054 [0.226]	0.087 [0.282]
Upper secondary school + voc. training	0.057 [0.231]	0.079 [0.270]	0.073 [0.261]	0.106 [0.308]
"Fachhochschule" degree	0.028 [0.166]	0.026 [0.160]	0.022 [0.146]	0.023 [0.150]
University degree	0.082 [0.274]	0.116 [0.320]	0.067 [0.249]	0.126 [0.331]
German nationality	0.742 [0.437]	0.755 [0.430]	0.828 [0.377]	0.830 [0.375]

Note: This table shows summary statistics for our set of covariates differentiated by individuals' selection status into part-time employment. Standard deviations are reported in brackets.

Table C.3.3: Covariates used in outcome, selection and sorting equations

	Outcome	Selection	Sorting
Constant	✓	✓	✓
Year dummies 2001-2017	✓		✓
Regional (federal state) dummies	✓	✓	
Age 25-29	✓	✓	
Age 30-34	✓	✓	
Age 35-39	✓	✓	
Age 40-44	✓	✓	
Age 45-49	✓	✓	
Age 50-54	✓	✓	
Age 55-60	✓	✓	
Full-time work experience	✓	✓	
Full-time work experience squared	✓	✓	
Full-time work experience $\times$ **1**{East Germany $= 1$}	✓	✓	
Full-time work experience squared $\times$ **1**{East Germany $= 1$}	✓	✓	
1{Full-time work experience $= 0$}	✓	✓	
Part-time work experience	✓	✓	
Part-time work experience squared	✓	✓	
Part-time work experience $\times$ **1**{East Germany $= 1$}	✓	✓	
Part-time work experience squared $\times$ **1**{East Germany $= 1$}	✓	✓	
1{Part-time work experience $= 0$}	✓	✓	
Minijob work experience	✓	✓	
Minijob work experience squared	✓	✓	
Minijob work experience $\times$ **1**{East Germany $= 1$}	✓	✓	
Minijob work experience squared $\times$ **1**{East Germany $= 1$}	✓	✓	
1{Minijob work experience $= 0$}	✓	✓	
Education: lower/middle secondary school + voc. training	✓	✓	
Education: upper secondary school only	✓	✓	
Education: upper secondary school + voc. training	✓	✓	
Education: “Fachhochschule” degree	✓	✓	
Education: university degree	✓	✓	
German nationality	✓	✓	
1{likely student aged 20-25 $= 1$}		✓	
1{likely student aged 26-30 $= 1$}		✓	
1{likely student aged 31-35 $= 1$}		✓	
Full-time share		✓	
Part-time share		✓	
Minijob share		✓	
First difference full-time share		✓	

Continued on following page

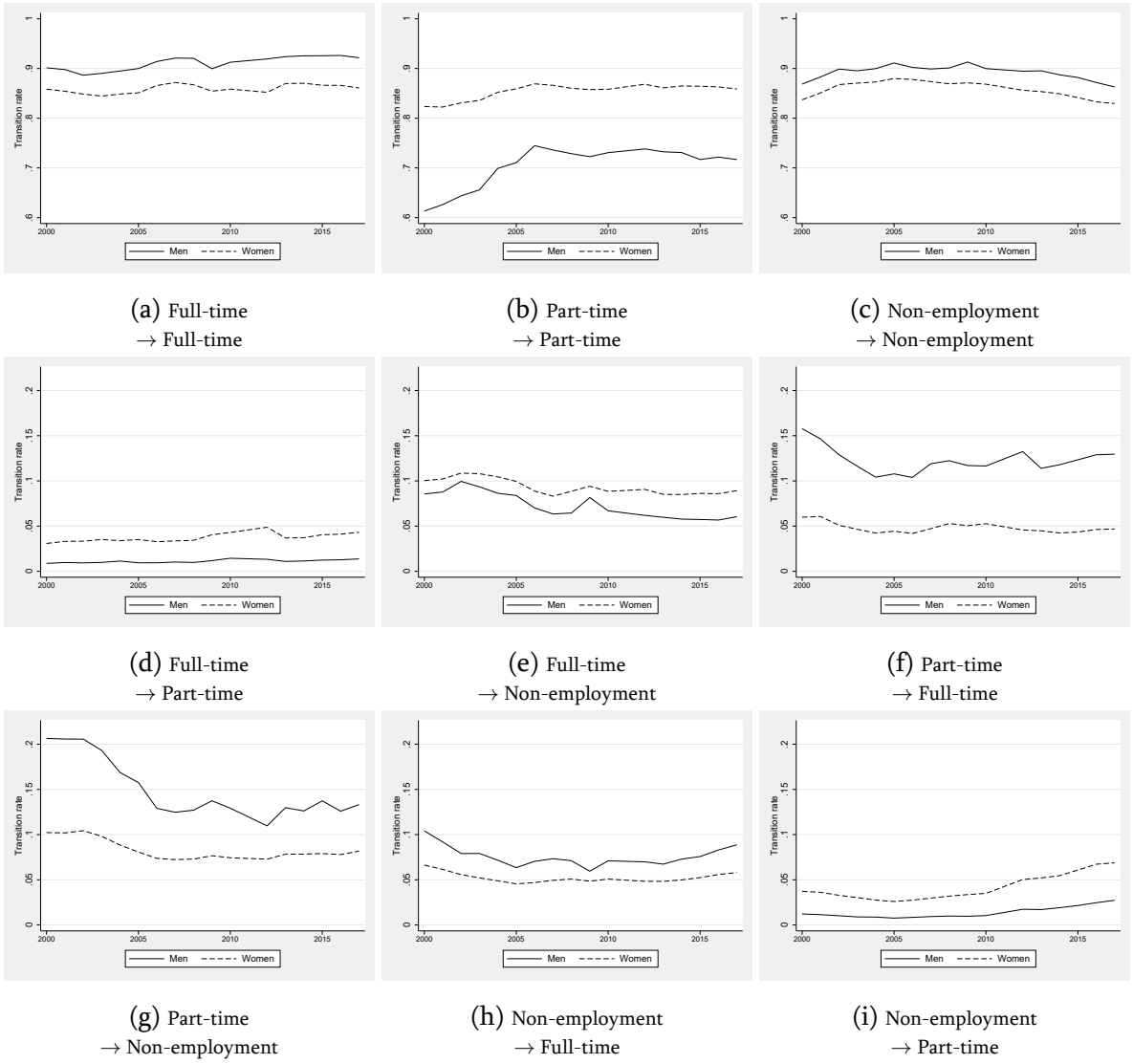
Table C.3.3, continued

	Outcome	Selection	Sorting
First difference part-time share		✓	
First difference minijob share		✓	
Transition rate: Full-time → Full-time		✓	
Transition rate: Full-time → Part-time		✓	
Transition rate: Full-time → Non-employment		✓	
Transition rate: Part-time → Full-time		✓	
Transition rate: Part-time → Part-time		✓	
Transition rate: Part-time → Non-employment		✓	
Transition rate: Non-employment → Full-time		✓	
Transition rate: Non-employment → Part-time		✓	
Transition rate: Non-employment → Non-employment		✓	

Note: This table lists the full set of covariates used in outcome, selection and sorting equations. Labour market variables used as instruments in the selection equation are estimated for cells defined by sex, year, age groups, region "Raumordnungsregion", and educational attainment (group 1: secondary school degree only or less; group 2: vocational training; group 3: university/"Fachhochschule" degree).

C.4.

TIME TRENDS IN INSTRUMENTAL VARIABLES



(a) Full-time → Full-time

(b) Part-time → Part-time

(c) Non-employment → Non-employment

(d) Full-time → Part-time

(e) Full-time → Non-employment

(f) Part-time → Full-time

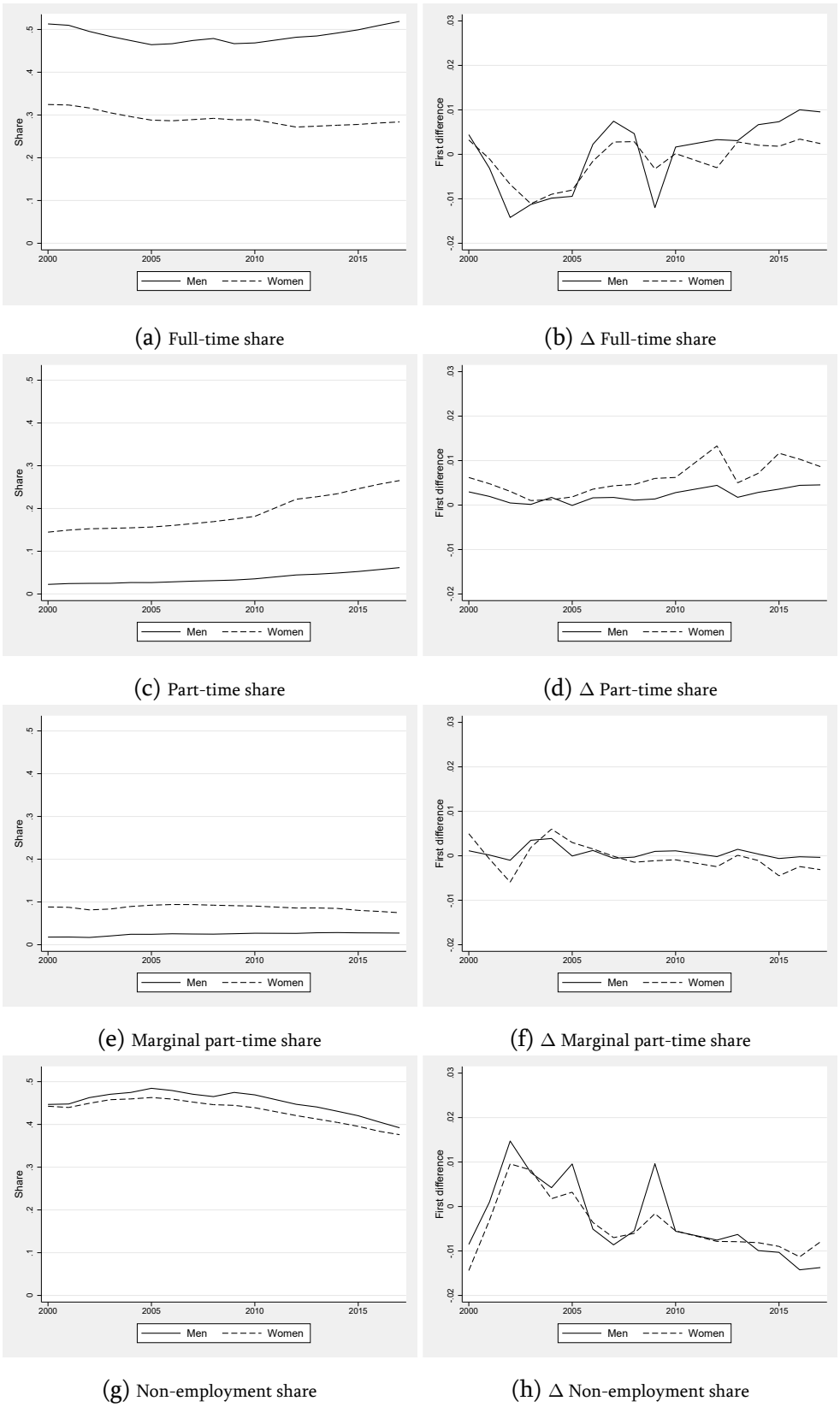
(g) Part-time → Non-employment

(h) Non-employment → Full-time

(i) Non-employment → Part-time

Figure C.2: Evolution of aggregated transition rates by gender

Note: These graphs illustrate the evolution of the transition rates listed in Table C.3.3 from 2000 to 2017 differentiated by gender.



(a) Full-time share (b) Δ Full-time share

(c) Part-time share (d) Δ Part-time share

(e) Marginal part-time share (f) Δ Marginal part-time share

(g) Non-employment share (h) Δ Non-employment share

Figure C.3: Evolution of aggregated employment shares and their first differences by gender

Note: These graphs illustrate the evolution of the employment shares and their first differences listed in Table C.3.3 from 2000 to 2017 differentiated by gender.

BIBLIOGRAPHY

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge (2022). When should you adjust standard errors for clustering? *Quarterly Journal of Economics 138*, 1–35.

Ahrens, A., C. Hansen, and A. Schaffer (2020). lassopack – model selection and prediction with regularized regression in Stata. *The Stata Journal 20*, 176–235.

Albrecht, J., A. Van Vuuren, and S. Vroman (2009). Counterfactual distributions with sample selection adjustments: Econometric theory and an application to the Netherlands. *Labour Economics 16*, 383–396.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica 41*, 997–1016.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics 24*, 3–61.

Arellano, M. and S. Bonhomme (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica 85*, 1–28.

Arellano-Valle, R. B., L. M. Castro, G. González-Farías, and K. A. Muñoz Gajardo (2012). Student-t censored regression model: Properties and inference. *Statistical Methods & Applications 21*, 453–473.

Aydin, D., O. Güneri, and E. Yilmaz (2021). Optimum shrinkage parameter selection for ridge type estimator of Tobit model. *Journal of Statistical Computation and Simulation 91*, 952–975.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics 4*, 384–414.

Bach, S., P. Haan, R. Ochmann, et al. (2013). Taxation of married couples in germany and the UK: One-earner couples make the difference. *International Journal of Microsimulation 6*, 2–20.

Bardasi, E. and J. Gornick (2008). Working for less? women's part-time wage penalties across countries. *Feminist Economics 14*, 37–72.

Bauer, T., H. Bonin, L. Goette, and U. Sunde (2007). Real and nominal wage rigidities and the rate of inflation: evidence from west German micro data. *Economic Journal 117*, 508–529.

Baumgarten, D., G. Felbermayr, and S. Lehwald (2020). Dissecting between-plant and within-plant wage dispersion: evidence from Germany. *Industrial Relations 59*, 85–122.

Beblo, M., D. Beninger, A. Heinze, and F. Laisney (2003). Measuring selectivity-corrected gender wage gaps in the EU. *ZEW Discussion Paper*, 1–33.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*, 2369–2429.

Belloni, A. and V. Chernozhukov (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*, 82–130.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018a). Supplement to "uniformly valid post-regularization confidence regions for many functional parameters in a z-estimation framework". *The Annals of Statistics 46*, 3643–3675.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018b). Uniformly valid post-regularization confidence regions for many functional parameters in a z-estimation framework. *The Annals of Statistics 46*, 3643–3675.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics, 10th World Congress of Econometric Society*, 1–41.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business and Economic Statistics 34*, 590–605.

Belloni, A., V. Chernozhukov, and K. Kato (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association 114*, 749–758.

Belloni, A., V. Chernozhukov, and L. Wang (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics 42*, 757–788.

Belloni, A., V. Chernozhukov, and Y. Wei (2016a). Post-selection inference for generalized linear models with many controls. *Journal of Business and Economic Statistics 34*, 606–619.

Belloni, A., V. Chernozhukov, and Y. Wei (2016b). Supplementary appendix for "Post-selection inference for generalized linear models with many controls". *Journal of Business and Economic Statistics 34*, 1–25.

Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole. *Mémoires de Mathématique et de Physique, tirés des registres de l'Académie Royale des Sciences de l'année 1760*.

Bia, M., M. Huber, and L. Lafférs (2024). Double machine learning for sample selection models. *Journal of Business & Economics Statistics 42*, 958–969.

Bick, A. and N. Fuchs-Schündeln (2017). Quantifying the disincentive effects of joint taxation on married women's labor supply. *American Economic Review 107*, 100–104.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*, 1705–1732.

Biewen, M. and P. Erhardt (2021). arhomme: An implementation of the Arellano and Bonhomme (2017) estimator for quantile regression with selection correction. *The Stata Journal 21*, 602–625.

Biewen, M., B. Fitzenberger, and M. Rümmele (2022). Using distribution regression difference-in-differences to evaluate the effects of a minimum wage introduction on the distribution of hourly wages and hours worked. *IZA Discussion Paper No. 15534*, 1–51.

Biewen, M., B. Fitzenberger, and M. Seckler (2020). Counterfactual quantile decompositions with selection correction taking into account Huber/Melly (2015): An application to the German gender wage gap. *Labour Economics 67*, 101927.

Biewen, M. and M. Sturm (2022). Why a labour market boom does not necessarily bring down inequality: putting together Germany's inequality puzzle. *Fiscal Studies 43*, 121–149.

Blau, F. D., L. M. Kahn, N. Boboshko, and M. Comey (2024). The impact of selection into the labor force on the gender wage gap. *Journal of Labor Economics*, forthcoming.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica 75*, 323–363.

Bossler, M. and T. Schank (2023). Wage inequality in Germany after the minimum wage introduction. *Journal of Labor Economics 41*, 813–855.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*, 1–122.

Bradic, J., J. Fan, and J. Jiang (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics 39*, 3092–3120.

Bradic, J. and J. Guo (2019). Generalized M-estimators for high-dimensional Tobit I models. *Electronic Journal of Statistics 13*, 582–645.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics 13*, 1–30.

Buchinsky, M. (2001). Quantile regression with sample selection: Estimating women's return to education in the US. *Economic Applications of Quantile Regression 26*, 87–113.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*, 429–436.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-dimensional Data*. Berlin, Heidelberg: Springer.

Burauel, P., M. Caliendo, M. Grabka, C. Obs, M. Preuss, and C. Schröder (2019a). The impact of the German minimum wage on individual wages and monthly earnings. *Journal of Economics and Statistics 240*, 201–231.

Burauel, P., M. Caliendo, M. Grabka, C. Obs, M. Preuss, and C. Schröder (2019b). The impact of the minimum wage on working hours. *Journal of Economics and Statistics 240*, 233–267.

Caliendo, M., A. Fedorets, M. Preuss, C. Schröder, and L. Wittbrodt (2022). The short- and medium-term distributional effects of the German minimum wage reform. *Empirical Economics 64*, 1149–1175.

Caliendo, M., C. Schröder, and L. Wittbrodt (2019). The causal effects of the minimum wage introduction in Germany – an overview. *German Economic Review 20*, 257–292.

Calvo, P. A., I. Lindenlaub, and A. Reynoso (2024). Marriage market and labor market sorting. Technical report.

Card, D. (1992). Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial Labor Relations Review 46*, 22–37.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of west German wage inequality. *Quarterly Journal of Economics 128*, 967–1015.

Carlini, N., M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr (2024). Poisoning web-scale training datasets is practical. *IEEE Symposium on Security and Privacy (SP)*.

Chai, H., Q. Zhang, J. Huang, and S. Ma (2019). Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Statistica Sinica 29*, 877–894.

Chen, S., N. Liu, H. Zhang, and Y. Zhou (2024). Estimation of wage inequality in the UK by quantile regression with censored selection. *Journal of Econometrics*, forthcoming.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*, 2786–2819.

Chernozhukov, V., I. Fernandez-Val, and S. Luo (2023). Distribution regression with sample selection and UK wage decomposition. Cemmap working papers, Institute for Fiscal Studies.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*, 2205–2268.

Chernozhukov, V., I. Fernandez-Val, B. Melly, and K. Wüthrich (2020). Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Society 115*, 123–127.

Chernozhukov, V., C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis (2024). *Applied Causal Inference Powered by ML and AI*. https://causalml-book.org/.

Chernozhukov, V., C. Hansen, and Y. Liao (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics 45*, 39–76.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *The Annual Review of Economics 7*, 649–688.

Chernozhukov, V. and H. Hong (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association 97*, 872–882.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics 49*, 1300–1317.

Chiang, H. (2020). *Three essays in cluster robust machine learning and high-dimensional econometrics*. Ph. D. thesis, Graduate School Vanderbilt University.

Chzhen, Y. and K. Mumford (2011). Gender gaps across the earnings distribution for full-time employees in Britain: Allowing for sample selection. *Labour Economics 18*, 837–844.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society 34*, 187–202.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica 39*, 829–844.

de la Peña, V. H., T. L. Lai, and Q.-M. Shao (2009). *Self-Normalized Processes*. Berlin Heidelberg: Springer.

Dolado, J. J., C. García-Peñalosa, and L. Tarasonis (2020). The changing nature of gender selection into employment over the great recession. *Economic Policy 35*, 635–677.

Dustmann, C., B. Fitzenberger, U. Schönberg, and A. Spitz-Oener (2014). From sick man of Europe to economic superstar: Germany's resurgent economy. *Journal of Economic Perspectives 28*, 167–188.

Dustmann, C., A. Lindner, U. Schönberg, M. Umkehrer, and P. vom Berge (2022a). Reallocation effects of the minimum wage. *Quarterly Journal of Economics 137*, 267–328.

Dustmann, C., A. Lindner, U. Schönberg, M. Umkehrer, and P. vom Berge (2022b). Reallocation effects of the minimum wage. *Quarterly Journal of Economics 137*, 267–328.

Dustmann, C., J. Ludsteck, and U. Schönberg (2009). Revisiting the German wage structure. *Quarterly Journal of Economics 124*, 843–881.

D'Haultfœuille, X., A. Maurel, and Y. Zhang (2018). Extremal quantile regressions for selection models and the black-white wage gap. *Journal of Econometrics 203*, 129–142.

Elass, K. (2024). Male and female selection effects on gender wage gaps in three countries. *Labour Economics 87*, 102506.

Ermisch, J. F. and R. E. Wright (1994). Interpretation of negative sample selection effects in wage offer equations. *Applied Economics Letters 1*, 187–189.

Fang, E. X., Y. Ning, and H. Liu (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society 79*, 1415–1437.

Farber, H. S. (1981). Worker preferences for union representation. Working paper 290, Massachusetts Institute of Technology (MIT), Department of Economics.

Fernandez-Val, A. I., van Vuuren, and F. Vella (2024a). Hours worked and the US distribution of real annual earnings 1976-2019. *Journal of Applied Econometrics 39*, 659–678.

Fernandez-Val, A. I., van Vuuren, and F. Vella (2024b). Nonseparable sample selection models with censored selection rules. *Journal of Applied Econometrics 240*, 1–28.

Fernandez-Val, A. I., van Vuuren, F. Vella, and F. Peracchi (2023). Selection and the distribution of female real hourly wages in the United States. *Quantitative Economics 14*, 571–607.

Fitzenberger, B. and J. de Lazzer (2022). Changing selection into full-time work and its effect on wage inequality in Germany. *Empirical Economics 62*, 247–277.

Foresi, S. and F. Peracchi (1995). The conditional distribution of excess returns: an empirical analysis. *Journal of the American Statistical Association 90*, 451–466.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Fry, T. R. L. (1991). A generalized logistic Tobit model. *Monash Working Paper No. 1/91*, 1–15.

Gallant, J. E. (2006). The M184V mutation: what it does, how to prevent it, and what to do with it when it's there. *The AIDS Reader 16*, 556–559.

Gallego-Granados, P. (2019). The part-time wage gap across the wage distribution. *DIW Discussion Paper*, 1–45.

Gallego-Granados, P. and K. Wrohlich (2020). The part-time wage gap across the wage distribution. *SEOPpapers on Multidisciplinary Panel Data Research*.

Gandhi, R. T., K. T. Tashima, L. M. Smeaton, V. Vu, J. Ritz, A. Andrade, J. J. Eron, E. Hogg, and C. J. Fichtenbaum (2020). Long-term outcomes in a large randomized trial of HIV-1 salvage therapy: 96-week results of AIDS clinical trails group A5241 (OPTIONS). *The Journal of Infectious Diseases 221*, 1407–1415.

Gartner, H. (2005). The imputation of wages above the contribution limit with the German IAB employment sample. FDZ Methodenreport.

Geyer, J., P. Haan, and K. Wrohlich (2015). The effects of family policy on maternal labor supply: Combining evidence from a structural model and a quasi-experimental approach. *Labour Economics 36*, 84–98.

Goldberger, A. S. (1981). Linear regression after selection. *Journal of Econometrics 15*, 357–366.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Cambridge, Massachusetts: The MIT Press.

Greene, W. H. (1981). On the asymptotic bias of the ordinary least squares estimator of the Tobit model. *Econometrica 49*, 505–513.

Grippo, L., F. Lampariello, and S. Lucidi (1986). A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis 23*, 707–716.

Gronau, R. (1974). Wage comparisons – a selectivity bias. *Journal of Political Economy 82*, 1119–1143.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal 32*, 119–134.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In S. V. Berg (Ed.), *Annals of Economic and Social Measurement*, Volume 5, Chapter 6, pp. 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Heckman, J. J. (1980). Sample selection bias as a specification error with an application to the estimation of labor supply functions. In J. P. Smith (Ed.), *Female Labor Supply: Theory and Estimation*, Chapter 5, pp. 206–248. Princeton: Princeton University Press.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications 103*, 103–118.

Higham, N. J., N. Strabić, and V. Sego (2016). Restoring definiteness via shrinking, with an application to correlation matrices with a fixed block. *SIAM Review 58*, 245–263.

Hirukawa, M., D. Liu, I. Murtazashvili, and A. Prokhorov (2023). DS-HECK: double-lasso estimation of Heckman selection model. *Empirical Economics 64*, 3167–3195.

Honoré, B. E. and J. L. Powell (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics 64*, 241–278.

Hsu, D., S. M. Kakade, and T. Zhang (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics 14*, 569–600.

Huang, J., T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang (2013). Oracle inequalities for the Lasso in the Cox model. *The Annals of Statistics 41*, 1142–1165.

Huber, M. and G. Mellace (2014). Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics 47*, 75–92.

Huber, M. and B. Melly (2015). A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics 30*, 1144–1168.

Jacobson, T. and H. Zou (2023a). High-dimensional censored regression via the penalized Tobit likelihood. *Journal of Business and Economic Statistics 42*, 286–297.

Jacobson, T. and H. Zou (2023b). Supplementary material for "High-dimensional censored regression via the penalized Tobit likelihood". *Journal of Business and Economic Statistics 42*, 1–35.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*, 2869–2909.

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003). Self-normalized Cramér-type large deviations for independent random variables. *The Annals of Probability 31*, 2167–2215.

Johnson, B. A. (2009). On Lasso for censored data. *Electronic Journal of Statistics 3*, 485–506.

Kaplan, D. and R. L. Venezky (1994). Literacy and voting behaviour: A bivariate probit model with sample selection. *Social Science Research 23*, 350–367.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 457–481.

Keeley, M. C., P. K. Robins, R. G. Spiegelman, and R. W. West (1978). The estimation of labor supply models using experimental data. *The American Economic Review 68*, 873–887.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature 61*, 1281–1317.

Lee, L.-F. (1979). Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica 47*, 977–996.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Manuscript from the Annual Meeting of the American Mathematical Society in Chicago 2*, 164–168.

Lewis, H. G. (1974). Comments on selectivity biases in wage comparisons. *Journal of Political Economy 82*, 1145–1155.

Li, Y., L. Dicker, and S. D. Zhao (2014). The Dantzig selector for censored linear regression models. *Statistica Sinica 24*, 251–268.

Link, S. (2024). The price and employment response of firms to the introduction of the minimum wages. *Journal of Public Economics 239*, 1–18.

Liu, X., Z. Wang, and Y. Wu (2013). Group variable selction and estimation in the Tobit censored response model. *Computational Statistics and Data Analysis 60*, 80–89.

Maasoumi, E. and L. Wang (2019). The gender gap between earnings distributions. *Journal of Political Economy 127*, 2438–2504.

Manning, A. and B. Petrongolo (2008). The part-time penalty for women in Britain. *Economic Journal 118*, 28–51.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics 11*, 431–441.

Matteazzi, E., A. Pailhe, and A. Solaz (2014). Part-time wage penalties for women in prime age. *Industrial and Labor Relations Review 67*, 955–985.

McDonald, J. F. and R. A. Moffitt (1980). The use of tobit analysis. *The Review of Economics and Statistics 62*, 318–321.

Melly, B. and G. Santangelo (2014). The evolution of the gender wage gap: 1968-2008. *Working Paper*.

Miller, V., T. Stark, A. E. Loeliger, and J. M. A. Lange (2002). The impact of the M184V substitution in HIV-1 reverse transcriptase on treatment response. *HIV Medicine 3*, 135–145.

Mindestlohnkommission (2020). Dritter Bericht zu den Auswirkungen des gesetzlichen Mindestlohns. *Bundesamt für Arbeitsschutz und Arbeitsmedizin (BAuA)*.

Müller, P. and S. van de Geer (2016). Censored linear model in high dimensions. *Test 25*, 75–92.

Mulligan, C. B. and Y. Rubinstein (2008). Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics 123*, 1061–1110.

Nelson, F. D. (1977). Censored regression models with unobserved stochastic censoring thresholds. *Journal of Econometrics 6*, 309–327.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (Ed.), *Probability and Statistics, the Harald Cramér Volume*, New York, pp. 213–234. John Wiley and Sons, Inc.

Neyman, J. (1979). C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A 41*(1), 1–21.

Olivetti, C. and B. Petrongolo (2008). Unequal pay or unequal employment? a cross-country analysis of gender gaps. *Journal of Labor Economics 26*, 621–654.

Olsen, R. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica 46*, 1211–1215.

Pan, Z. and J. Xie (2023). ℓ_1-penalized pairwise difference estimation for a high-dimensional censored regression model. *Journal of Business and Economic Statistics 41*, 283–297.

Pereda-Fernandez, S. (2024). Decomposition of differences in distribution under sample selection and the gender wage gap. *Journal of Business and Economics Statistics*, forthcoming.

Poirier, D. J. (1980). Partial observability in bivariate Probit models. *Journal of Econometrics 12*, 209–217.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics 25*, 303–325.

Powell, J. L. (1986a). Censored regression quantiles. *Journal of Econometrics 32*, 143–155.

Powell, J. L. (1986b). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica 54*, 1435–1460.

Reece, W. S. (1979). Charitable contributions: New evidence on household behaviour. *The American Economic Review 69*, 142–151.

Riphahn, R. and R. Schrader (2020). Institutional reforms of 2006 and the dramatic rise in old-age employment in Germany. *Industrial and Labor Relations Review 73*, 1185–1225.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*, 931–954.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights 4*, 305–322.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rudelson, M. and R. Vershynin (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics 61*, 1025–1045.

Schröder, C., J. König, A. Fedorets, J. Goebel, M. Grabka, L. Lüthen, M. Metzing, F. Schikora, and S. Liebig (2020). The economic research potential of the German socio-economic panel study. *German Economic Review 31*, 335–371.

Shafer, R. W. (2002). Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clinical Microbiology Reviews 15*, 247–277.

Soret, P., M. Avalos, L. Wittkop, D. Commenges, and R. Thiébaut (2018). Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. *BMC Medical Research Methodology 18*, 1–13.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis 45*, 89–103.

Swenson, L. C., B. Cobb, A. M. Geretti, P. Harrigan, M. Poljak, C. Seguin-Devaux, C. Verhofstede, M. Wirden, A. A., J. Boni, T. Bourlet, J. B. Huder, J. Karasi, S. Zidovec Lepej, M. M. Lunar, O. Mukabayire, R. Schuurman, J. Tomaz□ič, K. Van Laethem, L. Vandekerckhove, and A. M. J. Wensing (2014). Comparative performances of HIV-1 RNA load assays at low viral load levels: Results of an international collaboration. *Journal of Clinical Microbiology 52*, 517–523.

Terza, J. V. and W.-D. Tsai (2006). Censored probit estimation with correlation near the boundary: A useful reparametrization. *Review of Applied Economics 2*, 1–12.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine 28*, 385–395.

Tibshirani, R. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*, 24–36.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics 36*, 614–645.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*, 1166–1202.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Van de Ven, W. P. M. M. and B. M. S. Van Praag (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics 17*, 229–252.

Vershynin, R. (2018). Concentration of sums of independent random variables. In Z. Ghahramani, R. Gill, F. P. Kelly, B. D. Ripley, S. Ross, and M. Stein (Eds.), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge, pp. 11–37. Cambridge University Press.

Wang, S., B. Nan, J. Zhu, and D. G. Beer (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics 64*, 132–140.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Yang, Y. and H. Zou (2013). An efficient algorithm for computing the HHSVM and its generalizations. *Journal of Computational and Graphical Statistics 22*, 396–415.

Yu, G. and J. Bien (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika 106*, 533–546.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society B 76*, 217–242.

Zhou, Z., R. Jiang, and W. Qian (2013). LAD variable selection for linear models with randomly censored data. *Metrika 76*, 287–300.