Ludsteck, Johannes; Drechsler, Jörg

Conference Paper

Dealing with Rightcensored Wages through Imputation

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Ludsteck, Johannes; Drechsler, Jörg (2023) : Dealing with Rightcensored Wages through Imputation, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277650

Dealing with Rightcensored Wages through Imputation – preliminary and incomplete – do not cite without author's permission

Johannes Ludsteck*, Joerg Drechsler†

Abstract

The paramount importance of the German employment register data (BeH) is documented by a large number of papers on labour market topics which are based on this data set. Many analyses of wages are hampered, however, by right-censoring of wages at the social contribution threshold (Beitragsbemessungsgrenze).

In order to free researchers from the burden to tackle this problem, we develop and implement an imputation algorithm which solves the well-known problem that the density of imputed wages shows sizeable kinks and bumps at the censoring threshold. We identify the dependence of the regression model coefficients on the wage quantiles as cause of the problem and solve it by using censored quantile regressions and Tobit models with additional left-censoring. The problem that no variant of the proposed estimators (quantile regression and extended Tobit models) dominates the other for all subsamples of the data is addressed with an automatic selection procedure based on a smoothness criterion.

Our approaches are applicable for other data bases and right-censoring problems.

1 Introduction

Censoring from above is a notorious problem when analyzing wage data. To protect confidentiality, statistical agencies typically apply top-coding strategies when disseminating wage information to the public. Furthermore, in many administrative data sources the wage information is only collected up to an administrative limit such as the maximum taxable amount or the maximum contribution to the social security system. With top-coding, values above a

*Corresponing author. IAB Nuremberg

†IAB Nuremberg, University of Mannheim, Ludwig-Maximilians-University Munich, University of Maryland. We thank Andreas Moczall and Alexandra Schmucker for helpfulll discussions. All remaining errors are our owns.

predefined threshold are not revealed. Values above the threshold are typically replaced with the value of the threshold or sometimes with the mean of all units affected by top-coding. For example, in the Current Population Survey (CPS), the U.S. Census Bureau is planning to top-code the top three percent of reported earnings in the future U.S. Census Bureau 2022a (the currently employed top-coding rules are slightly more complicated U.S. Census Bureau 2022b). While three percent of the data might not seem much, the rate will be much higher when analyzing specific subgroups of the data, such as highly educated respondents. Other prominent surveys that use top-coding include the American Community Survey (ACS), the Panel Study of Income Dynamics (PSID), or the Survey of Income and Program Participation (SIPP). Outside the U.S. top-coding is used for example in the UK Household Logitudinal Study and the Quarterly Labour Force Survey in the UK.

When working with administrative data, similar problems arise but typically for a different reason. In these data income information is often collected for administrative purposes such as setting the amount of social security payments. Since there typically is an upper limit regarding these payments, income information is often only reported up to the contribution limit. In fact, this problem motivated the research presented in this paper. The Employment History Data (BeH) at the Institute for Employment Research is a very rich administrative data source containing detailed employment information such as duration of employment, earnings, occupation, industry of the employer etc. for all German employees covered by the Social Security System. The data are based on notifications that every employee in Germany has to provide regarding his employees on a regular basis. Since the administrative purpose of this database is to set the social security payments, all wage information is only collected up to the contribution limit. This implies that the wage information is censored from above. Censoring shares range (depending on the year) between about 10 and 12 percent for full-time employed men on average but exceed 30 percent for men with a college degree.

Similar problems arise in the Austrian Social Security Database, which also only provides wage information up to the contribution limit. In the U.S., the Earnings Public-Use Microdata File published by the Social Security Administration based on social security tax records are censored at the maximum taxable earnings level of the social security Compson 2011.

The problem of wage censoring is especially severe for analyses of wage inequality since movements at the upper end of the income distribution became more important in the recent past.

Two general strategies are commonly applied to deal with this problem. Either the censoring is directly taken into account when analyzing the data or a two-stage procedure is employed in which all censored values are imputed first before applying standard analysis procedures using the imputed data (potentially accounting for the extra uncertainty from imputation). For the first approach, different strategies are applied depending on the type of analysis to be conducted. If wages are treated as the dependent variable in a regression context, the most common approach is to replace the linear regression by Tobit

models. If the censored variable is used as a predictor, most researchers follow Chow 1979; Anderson, Basilevsky, and Hum 1983 who propose to interact the censored regressor with a dummy which indicates whether the observation is censored and to add this dummy and the interaction term to the model. As discussed for example in Jones 1996 the second approach should generally be avoided as it can introduce bias in the estimated regression coefficient of the censored variable, which may translate to biases in the coefficients of the other variables (depending on their partial correlations with the censored regressor).

But even the Tobit model, which provides unbiased results as long as the model assumptions are fulfilled, has the drawback that it can only be used to obtain estimates for the linear regression model. If interest lies on other aspects such as studying income inequality the Tobit model will not be helpful.

Thus, the imputation approach is often preferred in practice as it offers full flexibility regarding the type of analysis conducted on the imputed data. This strategy has been used in various contexts, most importantly when studying wage inequality, e.g in Dustmann, Ludsteck, and Schönberg 2009; Card, Heining, and Kline 2013. An additional important advantage of the imputation approach is the possibility of reusing the imputed variable for other research projects. This can reduce the burden for future projects that otherwise always have to come up with their own strategy how to deal with the censoring problem. Furthermore, given the larger potential benefits it justifies some extra efforts to carefully design and evaluate the model used for imputation. However, there is an important caveat to this approach. The imputed values will only reflect those relationships that were built into the imputation model. Meng Meng 1994 coined the term *uncongeniality* to describe the situation if the modeling assumptions differ between the imputer and the analyst. If the imputer and the analyst are different individuals, uncongeniality is almost inevitable. Uncongeniality is especially problematic if variables or interaction terms that are included in the analysis model are not included in the imputation model. The regression coefficients of these variables will be attenuated after imputation unless the implicit assumption of the imputation model is satisfied that these variables are no longer correlated with the dependent variable given the variables included in the model.

A general recommendation in the imputation literature is therefore to always use inclusive models based on a rich set of predictors. The more of the variability of the dependent variable can be explained by the predictors the smaller the possible attenuation bias for any variables not included in the model. In the context of wage regression this implies that person level and establishment level fixed effects should always be included in the imputation model. The inclusion of these effects is important for two reasons: First, the effects control for all time-invariant individual and establishment level effects avoiding omitted variable bias and substantially improve the model fit. For example Abowd, Kramarz, and Margolis 1999 find that the R^2 in a wage regression improves from about 0.4 to 0.9 if these fixed effects are included. Second, including the establishment level effects will implicitly control for all regional and industry level effects as well. Since establishments rarely move their geographical location or change

their main activity to the extend that they would be classified as belonging to a different industry, the regional and industry effects are simple aggregates of the establishment effects and are thus already taken into account by the inclusion of the establishment level effects.

However, directly including the fixed effects as dummies is often not feasible in practice because of the large number of parameters that would need to be estimated. Furthermore, the commonly applied within-transformation strategy, which helps to reduce the number of parameters, can also not be employed, as the necessary average wages on the individual and establishment level are not available due to censoring. We address this problem by adopting and improving on an imputation strategy first discussed in Card, Heining, and Kline 2013. They suggest to approximate the fixed effects by leave-on-out-means (LOOMs). However, they do not account for the fact that these LOOMs will also be affected by censoring. To overcome this problem, we propose a two-stage imputation routine, in which censored values are replaced with imputed values based on a model without fixed effects on the first-stage. These imputed values are used as input for the second stage, which includes the LOOMs computed based on the imputed values from the previous round. We also identify another problem: simply using all information below the contribution limit, will introduce bias in the imputed values generating a heap in the imputed data immediately above the contribution limit. Since this heap occurs in two different datasets that we use in our evaluations, we believe that this is a general problem, which is not limited to our data. We demonstrate that the problem arises since the assumption of constant regression coefficients is violated and propose two alternative methods to cope with it: (1) A doubly censored Tobit model which reduces the bias by reducing the contribution of observations in the lower tail of the distribution of the dependent, and (2) estimating the regression coefficients near the censoring limit using censored quantile regressions.

2 Imputation Model

2.1 Basic Setup

Similar to early works (e.g. Dustmann, Ludsteck, and Schönberg 2009; Card, Heining, and Kline 2013) our imputation model is based on the Tobit model. To maximize the explanatory power and to fully account for the hierarchical structure of the data (employment spells nested within individuals and individuals nested within establishments and within occupations), the ideal model would be given as

$$w_{siet} = \max\left\{c_t, \min\left\{C_t, x_{siet}\, b_t + \mu_i + \eta_{et} + \omega_{ot} + u_{siet}\right\}\right\} \tag{1}$$

where

- s, i, e, o, t denote identifiers for spells, persons, establishments, occupations and time, respectively.

- w_{siet} denotes the natural logarithm of daily (pre-tax) wages, censored at the social contribution assessment ceiling C_t.
- C_t denotes the social contribution assessment ceiling.[1]
- c_t denotes artificial censoring introduced to ensure stable estimates of the regression coefficients as explained below.
- x_{siet} denotes a row vector of predictors that vary over spells, persons, establishments and time.
- μ_i denote fixed person effects.
- $\tilde{\eta}_{et}$ denote time-varying fixed establishment effects.
- $\tilde{\omega}_{ot}$ denote time-varying fixed occupation effects.
- u_{siet} denote residuals.

However, as discussed in the introduction, estimating the true fixed effects μ_i, η_{et} and ω_{ot} by adding dummies for the respective groups computationally infeasible[2] and yields biased estimates[3].

We avoid this problem by following Card, Heining, and Kline 2013 who approximate the fixed effects by adding the (spell-duration weighted) leave-one-out means (LOOMs) of daily wages as regressors. Formally, $\tilde{\mu}_{i,-s}$ is the duration-weighted mean over all spells of person i except for spell s. Correspondingly $\tilde{\eta}_{et,-i}$ relates to all co-workers of person i in establishment e and year t. Finally, $\tilde{\omega}_{ot,-i}$ is the average wage of all workers in occupation o (except i). A formal definition of the LOOMS is shifted to the appendix.

After substituting the fixed effects by their approximations the model can be written as

$$w_{siet} = \max\Big\{c_t, \min\left\{C_t, x_{siet}\, b_t + \tilde{\mu}_{i,-s}\, h_\mu + \tilde{\eta}_{et,-i}\, h_\eta + \tilde{\omega}_{ot,-i}\, h_\omega + u_{siet}\right\}\Big\}. \quad (2)$$

In order to render the model flexible, it is estimated separately for subsamples of the register data which are obtained by partitioning the data set by year, gender, four age groups, four education groups and Eastern/Western Germany.

The imputed wages w^I_{siet} are computed as

$$w^I_{siet} := x_{siet}\, \hat{b}_t + \tilde{\mu}_i\, \hat{h}_\mu + \tilde{\eta}_{et}\, \hat{h}_\eta + \tilde{\omega}_{ot}\, \hat{h}_\omega + \tilde{u}_{siet}. \quad (3)$$

[1] The censoring threshold is somewhat smaller (roughly 20 to 30 Euro) in Eastern Germany. We do not use an additional subscript ($C_{t,r}$ for sake of notational simplicity.

[2] Due the large number of groups (several thousand individuals contained in each estimation cell imply that thousands of coefficients would need to be estimated).

[3] Consistent estimation of the person-level dummies is infeasible since the coefficients are based on a small number of observations (on average less than 30 spells per person). Due to the nonlinearity of the Tobit estimator this inconsistency translates to all other coefficients, see, for example, Hsiao 2003, p. 194 and 243. A consistent method-of-moments estimator for Tobit models with large numbers of fixed effects was proposed by Honoré 1993. It is not suitable for imputation purposes since it irretrievably removes the fixed effects.

for censored observations. The error term $\tilde{u}_{siet}$ is sampled from a truncated normal distribution with variance $\hat{\sigma}^2_{siet}$. The estimated variance $\hat{\sigma}^2_{siet}$ includes both the variance of residuals and of the coefficient estimates

$$\hat{\sigma}^2_{siet} = \hat{V}(x_{siet}\,\hat{b}_t + u_{siet}) = x_{siet}\,\hat{V}(\hat{b}_t)\,x^{\top}_{siet} + \hat{V}(u_{siet}) \tag{4}$$

Sampling the error term from a left-truncated distribution ensures that $w^I_{siet} > C_t$ after adding $\tilde{u}_{siet}$. Substitution of definition (2) into the condition $w^I_{siet} > C_t$ and solving for $\tilde{u}_{siet}$ yields

$$\tilde{u}_{siet} > C_t - x_{siet}\,\hat{b}_t - \tilde{\mu}_i\,\hat{h}_\mu - \tilde{\eta}_{et}\,\hat{h}_\eta \tag{5}$$

as the left-truncation threshold.

Note that computing the leave-one-out means directly from the data would imply that the means would be based on censored wages. This may generate considerable bias, especially for persons with large shares of censored wages. To see this, consider the extreme case where all observations of a person are censored. The leave-one-out mean then is the mean of the censoring thresholds C_t for those years in which the person was employed. Since the combination of the leave-one-out means explain a large share of the variance of the dependent variable, the predicted values (before adding $\tilde{u}_{siet}$) are expected to be close to the average censoring limit. We mitigate this effect by obtaining an initial estimate for the censored wage by dropping $\tilde{\mu}_{i,-s}$, $\tilde{\eta}_{et,-i}$ and $\tilde{\omega}_{ot,-i}$ from equation (2). The predicted values from this step are used to compute $\tilde{\mu}_{i,-s}$, $\tilde{\eta}_{et,-i}$ and $\tilde{\omega}_{ot,-i}$ for the final imputation.

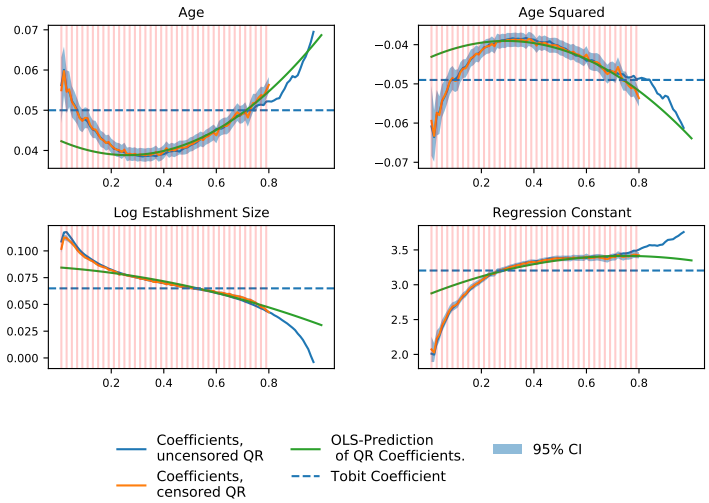
2.2 Remedies – artificial left-censoring and imputation based on quantile regressions

Implementing the imputation approach as described in the previous section introduces artifacts (kinks and bumps) in the distribution of the imputed wages above the censoring threshold, see e.g. Figure 2. A possible explanation for these artifacts is variability of the regression coefficients with respect to the quantiles of the dependent variable. We inspect the by estimating (censored) quantile regressions of a parsimonious model (containing only a small subsect of the predictors. This exercise is performed with the Verdienststrukturerhebung (VSE, structure of earnings survey) of the German Federal Office of Statistics Bundesamt 2018. The data will be described in more detail in the Appendix. At this point it suffices to note that the income and some other variables of the VSE are highly similar to the BeH since they follow almost identical definitions. The VSE offers an ideal test bed for evaluating different imputation approaches since income information in the VSE is censored for less than 1% of the data.

Figure 4 shows coefficient profiles from uncensored and censored quantile regressions[4] (blue and orange graphs) together with an extrapolation (green) of the censored QR coefficient profiles into the censored range and the respective

[4]We implement the three step estimator proposed by Chernozhukov and Hong 2002

Figure 1: Profiles of the Coefficients from Censored Quantile Regressions



Source: Own computations

Data Source VSE, subsample: Medium-qualified men (completed apprenticeship), working in western Germany, age group 30 to 64 years.

Tobit coefficients (green horizontal lines). A glance at the figure clearly rejects the assumption of constant regression coefficients. Furthermore the Tobit coefficients deviate significantly from the quantile coefficients in the censored range. This suggests to take them nearer to the relevant values by weighting down the impact of the observations in the lower range of the wage distribution by introducing artificial left-censoring. In the explorative analysis we censor at the 20th percentile of the subset of uncensored observations.

A further self-suggesting alternative strategy to reduce the bias of the imputation is to extrapolate the quantile regression profiles to the censored range and to use them for the imputation. We do this by regressing the coefficient profiles (in the uncensored red-shaded region) on a constant and a quadratic polynomial of the quantile and use the prediction from this regression in order to extrapolate the profiles into the censored range.

We use regularized weighted least squares regressions to fit the quantile profiles of the coefficients. The regularization (with L2-penalty terms and penalty weight 0.002) is used in order to avoid overfitting of the bends in the bottom quantile ranges (first to tenth quantile). Furthermore we weight the observa-

tions with their quantile distance from the censoring limit, i.e. $w_i = q_c - q_i$ where q_c is the quantile of the censoring limit and q_i the quantile of the coefficient estimate. The weighting puts greater weights on the coefficient values near the censoring limit.

The imputation is then performed in two steps:

1. The extrapolation yields quantile coefficient estimates $\tilde{b}(q_i)$ for a narrow grid of quantiles $q_i \in \{0.01, 0.02, \ldots, 0.99, 1.0\}$. We find – for every observation i the smallest quantile q_i^{min} such that

   $$x_i \, \tilde{b}(q_i) \geq C_t$$

   with (year-specific) censoring limit C_t,

2. For every observation i draw a pseudo random number u_i from the truncated uniform distribution in the range $\{q_i^{min}, q_i^{min} + 0.01, \ldots, 1]$ and generate the imputed values

   $$w_i^I = x_i \, \tilde{b}(u_i).$$

The extrapolation may lead us astray if the fit of the quantile coefficient process by a least squares regression is poor, if the censored range is large or if the quantile coefficient path continues not smoothly or contains turning points in the censored range. We abandoned this strategy since we found in exploratory analyses that these problems arise frequently and may induce considerable bias.

An approach that is less prone to this issue is to assume constancy of the quantile coefficients in the censored range and to use quantile regression coefficient values directly below the censoring limit to compute the prediction

$$w_i^I = x_i \, b(q_C) + \epsilon.$$

Here q_C is the greatest uncensored quantile of wages and ϵ is a draw from a left-truncated normal distribution where the truncation limit is chosen such that

$$x_i \, b(q_C) + \epsilon > C \iff \epsilon > C - x_i \, b(q_C)$$

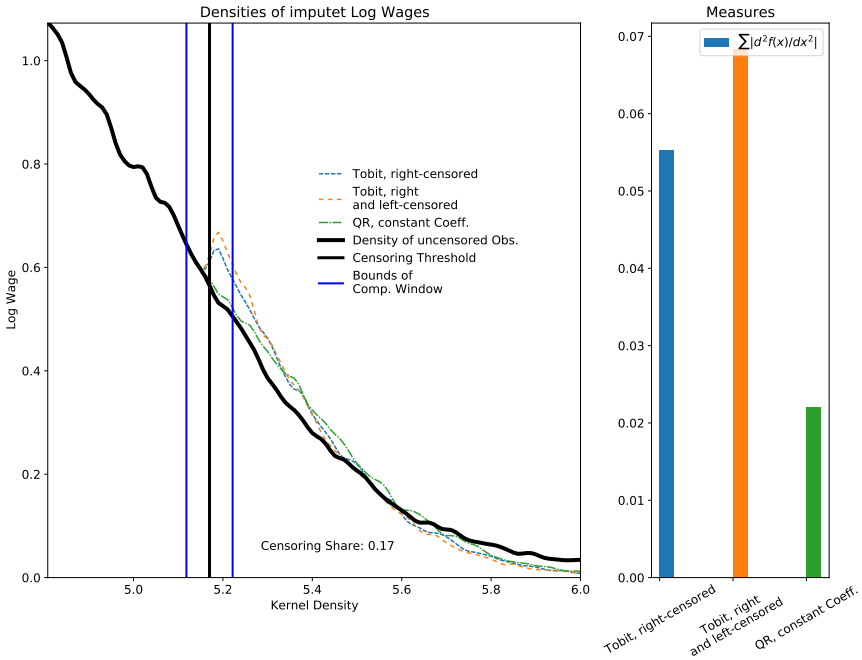
The results from the four imputation approaches are shown in Figure 2. It comprises densities of the respective imputed wages near the censoring threshold on the left hand side.

Unfortunately the ranking of the imputation approaches is less clear than here for many other cells. Furthermore the number of imputation cells is too large[5] to perform the ranking based on a visual inspection of the density plots. Therefore we construct a simple criterion that allows an automated ranking of the approaches. The criterion measures the smoothness of the kernel density near the censoring limit, computed as the sum of absolute values of the second finite derivatives of the kernel density (for each of the approaches), formally:

$$\sum_{i \in G} \left| \frac{\Delta^2 \hat{f}(x_i)}{\Delta x_i^2} \right|,$$

[5] We have 36 cells (gender × east × education × age group) for each year.

Figure 2: Densites for all Imputation Approaches



Source: Own computations

Data Source VSE, subsample: Medium-qualified men (completed apprenticeship), working in western Germany, age group 30 to 64 years.

where the derivatives are evaluated for a grid $G = \{g_{min}, g_{min}+0.001, \ldots, g_{max}-0.001, g_{max}\}$ with $g_min = 0.99 \times C$, $g_{max} = 1.01 \times C$ and $\Delta = 0.001$.[6]

At first glance the ranking of approaches appears somewhat arbitrary as both criteria depend on several parameters, the width of the (window) where they are evaluated, the coarseness of the grid, the bandwidth for the kernel density

[6]We considered another sensible (but in its original definition infeasible) criterion as the weighted (discrete) integral of the (absolute) deviation between the kernel density estimate $f^j(x_i)$ of approach j and the density $f^u(x_i)$ of the uncensored wages (which are available for the VSE only). Formally

$$\sum_{i \in G} \left| f^u(x_i) - f^j(x_i) \right| \times (x_i - x_{i-1}) \times f^u(x_i).$$

This criterion becomes feasible by replacing the true density $f^u(x_i)$ by the extrapolated $\hat{f}^u(x_i)$. We obtained $\hat{f}^u(x_i)$ by approximating the true density in the the interval $\{g_{min}, g_{min} + 0.001, \ldots, C - 0.001\}$ with $g_{min} = 0.9 \times C$ by a linear least squares regression and using $\hat{f}^u(x_i)$ in the comparison interval $G = \{g_{min}, g_{min} + 0.001, \ldots, g_{max}\}$ Since both criteria yielded the same results in almost all situations, the second criterion was abandoned for sake of simplicity.

estimates and the specification of the extrapolation model for the density. After experimenting with small changes of the parameters we found, however, that the ranking of the model specifications is quite robust to such changes.

Getting a smooth density at the censoring threshold appears intuitively coherent. It is, however, unclear whether the imputation model producing the smoothest density yields the least biased results when used in regression models or to computed unconditional statistics like means, variances and quantiles. We assess this by computing measures for the quality of regressions results along several dimensions

1. The Mean Squared Error of prediction (MSE), i.e. the mean of the squared difference between the predicted values from a least squares regression model based on the true uncensored (log) wages and a model based on imputed wages.

2. The Mean Absolute Error of prediction (MAE tStat), i.e. the mean of the absolute difference between the predicted values from a least squares regression model based on the true uncensored (log) wages and a model based on imputed wages.

3. The Mean of the sum of squared differences (MSD tStat) between the t-values from a regression model using the imputed wages and the t-values from a model using the uncensored wages.

4. The Mean of the sum of absolute differences (MAD) between the t-values from a regression model using the imputed wages and the t-values from a model using the uncensored wages.

5. The difference between the true uncensored and imputed wages for the quantiles 0.75, 0.90 and 0.99.

Table 1: Selection Criterion and Regression Quality Measures – Subsample: Men, 30-64 Years, Western Germany. Data Source: VSE 2010

Qualification	Completed Apprenticeship			College		
Estimator	Tobit R	Tobit LR	QuantReg	Tobit R	Tobit LR	QuantReg
MSE	0.162	0.162	0.162	0.176	0.176	0.191
MAE	0.299	0.299	0.299	0.305	0.302	0.312
MSD tStat	192.352	196.429	176.416	793.022	830.240	848.300
MAD tStat	10.298	10.439	9.805	22.178	23.114	22.922
Dev. Q(0.75)	-0.000	-0.000	-0.000	0.050	0.002	-0.175
Dev. Q(0.90)	-0.019	-0.024	-0.006	0.019	-0.058	-0.323
Dev. Q(0.99)	-0.266	-0.286	-0.246	-0.374	-0.496	-0.890
SAD2	0.036	0.065	0.009	0.091	0.041	0.740

Legend: MSD (MAD) tStat: Mean Squared (Absolute) Deviations of t Statistics, Dev. $Q(q)$: Deviation between quantile q of uncensored and imputed wages, SAD2: Sum of absolute second finite Derivatives.

Table 1 shows the results for medium and high-qualified men working in Western Germany. The tables for women and Eastern Germany which yield qualitatively similar results are shifted to the appendix.

The last row contains the selection criterion SAD2. It favours the quantile regression for the medium qualifiified (completed Apprenticeship) and the left-righ-censored Tobit for the high qualified (college graduates). Note that the ordering of all quality measures relating to regression models (MSE, MAE, MSD tStat and MAD tStat) is identical to the ordering according to the SAD2, implying that the SAD2 selects the imputation model which yields best results if the imputed values are used for regression models. A glance at the quantile deviations tells us that the results are less clear with respect to unconditional statistics. The favoured quantile regression yields least deviations between the quantiles of the true and the imputed wages for the medium qualified. But this does not apply to the high-qualified where the favoured left-right-censored Tobit model shows greater quantile deviations than the right-censored Tobit model.

3 Conclusion

We obtain multiply applicable/reuseable imputed wages for the German employment register data (BeH) by including (proxies for) fixed effects at several levels and estimating the models separately for narrow cells of years, age, gender and education groups. Inspection of the densities of the imputed wages reveals sizeable kinks and bumps at he censoring threshold. We identify the dependence of regression coefficients as likely cause of these deficiencies and tackle the problem by either additional left-censoring of the wages or using quantile regressions in order to obtain the relevant 'local' regression coefficient values. The best-suited model approach is selected using a simple smoothness criterion based on the second finite derivatives of the kernel density estimates.

The proposing modelling approach appears useful in two respects. First, it is applicable for a wide class of rightcensored variables. Second, it offers a specification test for Tobit models which should be conducted even if using imputations appears not necessary at a glance. Since Tobit models are based on the assumption of constant regression coefficients they yield biased results if it is violated. Computing imputed values an inspection of their density at the censoring threshold may therefore uncover dependency of the regression coefficients on the quantiles of the dependent variable.

A Formal definitions of the leave-one-out-means

Here we provide precise formal definitions of the person-specific ($\tilde{\mu}_{i,-s}$) and establishment-specific ($\tilde{\eta}_{-i,et}$) LOOMs.[7]

$$\tilde{\mu}_{i,-s} = \ln\left(\frac{1}{n_{i,-s}} \sum_{s' \neq s} W_{s'i}\, d_{s'i}\right) \tag{6}$$

$$\tilde{\eta}_{-i,et} = \ln\left(\frac{1}{n_{-i,et}} \sum_{t' \in \theta(t)} \sum_{s \neq s'} \sum_{j \neq i} W_{s'jet'}\, d_{s'jet'}\right). \tag{7}$$

W_{siet} denotes the daily wage here.

$$\theta(t) = \begin{cases} \{t, t+1\}, & \text{is establishment } e \text{ is founed in } t \\ \{t-1, t\} & \text{wenn establishment } e \text{ in closed in } t \\ \{t-1, t, t+1\} & \text{otherwise} \end{cases}$$

$\theta(t)$ denotes the set of years which are averaged d_{siet} denotes the duration of spells, and

$$n_{i,-s} = \sum_{s' \neq s} d_{s'i},$$

$$n_{-i,et} = \sum_{t' \in \theta(t)} \sum_{s' \neq s} \sum_{j \neq i} d_{s'jet'}$$

denote the respective leave-one-out sums of spell durations.

B Comparison of the BeH with the VSE

The VSE is used as a test bed in order to evaluate the imputation models proposed in this paper. Clearly it is suitable for this purpose only if the samples and the available variables are sufficiently similar. Here we describe the main characteristics of the VSE and demonstrate the similarity of both data sets by comparing the frequencies of their age structures and the densities of wages for the subsamples which are used in the figures and tables in the main section 2 (Men, aged 30-64 years, working in Western Germany).

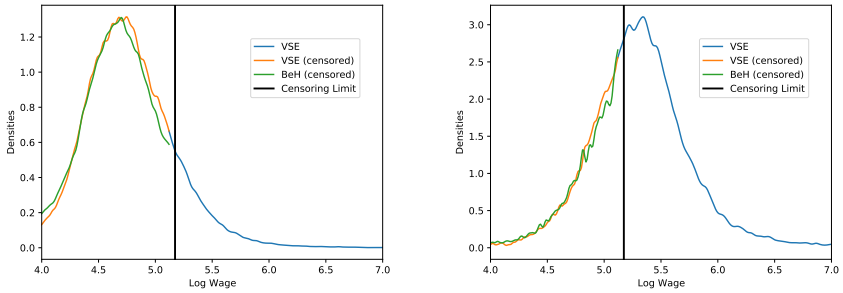
B.1 Main Characteristics of the VSE 2010

The VSE 2010 is a large establishment survey (comprising about 1.9 mio. observations) aimed to assess the wage structure of dependent employees in Germany. Since participation is mandatory, information is highly reliable and nonresponse

[7] The definition of the occupation-specific LOOMs is omitted since it can be obtained by replacing the establishment index by the occupation index in the establishment-specific LOOMs.

bias can be ignored. Comparison with the BeH is hampered as the VSE 2010 excludes establishments with less than ten employees and some industries (mainly those from the services sector). We mimic this by applying the same restrictions to the BeH subsample used for the comparisons.

Figure 3: Comparison of the Densities of Log Wages



Legend: Left hand side: Men aged 30-64 years, Western Germany, medium qualification (completed apprenticeship).
Right hand side: Men aged 30-64 years, Western Germany, medium qualification (college or technical college).

B.2 Tables and Figures

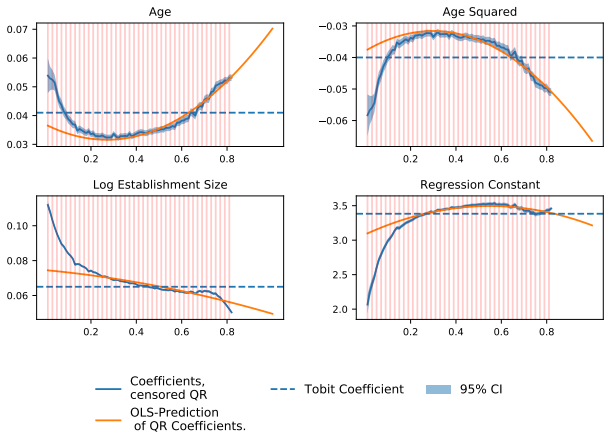
The following Table 2 compares the age structure of both data sets (after applying the VSE sampling weights).

Table 2: Frequencies of Age Groups in the VSE and BeH

	Age Group	Percent VSE	Percent BeH	Difference
3	≤ 19	0.357	0.416	0.059
4	20-24	5.964	5.855	-0.109
5	25-29	10.260	10.190	-0.070
6	30-34	10.984	10.951	-0.033
7	35-39	11.496	11.650	0.155
8	40-44	16.160	16.204	0.044
9	45-49	17.290	17.345	0.055
10	50-54	14.065	14.135	0.070
11	55-59	9.495	9.412	-0.082
12	60-64	3.519	3.429	-0.090

The following Figure 4 reproduces Figure 1 based on the BeH. The coefficient profiles are highly similar to the respective VSE profiles. 1

Figure 4: Profiles of the Coefficients from Censored Quantile Regressions, Obtained from the BeH



Source: Own computations
Data Source BeH, subsample: Medium-qualified men (completed apprenticeship), working in western Germany, age group 30 to 64 years.

C Further Tables for Assessing the Imputation Method Selection Criterion

Table 3: Selection Criterion and Regression Quality Measures – Subsample: Women, 30-64 Years, Western Germany. Data Source: VSE 2010

Qualification	Completed Apprenticeship			College		
Estimator	Tobit R	Tobit LR	QuantReg	Tobit R	Tobit LR	QuantReg
MSE	0.145	0.145	0.145	0.164	0.164	0.164
MAE	0.288	0.288	0.288	0.299	0.299	0.299
MSD tStat	8.652	9.028	6.407	32.401	30.777	44.350
MAD tStat	1.782	1.874	1.870	3.592	3.579	4.026
Dev. Q(0.75)	-0.000	-0.000	-0.000	0.012	0.001	-0.006
Dev. Q(0.90)	0.000	0.000	0.000	-0.005	-0.038	-0.060
Dev. Q(0.99)	-0.096	-0.120	-0.002	-0.236	-0.301	-0.342
SAD2	0.017	0.037	0.048	0.057	0.099	0.048

Legend: MSD (MAD) tStat: Mean Squared (Absolute) Deviations of t Statistics, Dev. Q q Deviation between quantile q of uncensored and imputed wages, SAD2: Sum of Absolute second finite Derivatives.

Table 4: Selection Criterion and Regression Quality Measures – Subsample: Men, 30-64 Years, Eastern Germany. Data Source: VSE 2010

Qualification	Completed Apprenticeship			College		
Estimator	Tobit R	Tobit LR	QuantReg	Tobit R	Tobit LR	QuantReg
MSE	0.162	0.162	0.162	0.207	0.207	0.213
MAE	0.305	0.305	0.305	0.330	0.330	0.333
MSD tStat	11.700	10.984	9.312	27.407	33.315	30.170
MAD tStat	2.734	2.794	2.677	4.293	4.849	4.686
Dev. Q(0.75)	-0.000	-0.000	-0.000	0.026	0.016	-0.078
Dev. Q(0.90)	0.000	0.000	0.000	-0.026	-0.053	-0.230
Dev. Q(0.99)	-0.208	-0.193	-0.123	-0.401	-0.447	-0.753
SAD2	0.037	0.020	0.023	0.061	0.052	0.369

Legend: MSD (MAD) tStat: Mean Squared (Absolute) Deviations of t Statistics, Dev. Q q Deviation between quantile q of uncensored and imputed wages, SAD2: Sum of Absolute second finite Derivatives.

Table 5: Selection Criterion and Regression Quality Measures – Subsample: Women, 30-64 Years, Eastern Germany. Data Source: VSE 2010

Qualification	Completed Apprenticeship			College		
Estimator	Tobit R	Tobit LR	QuantReg	Tobit R	Tobit LR	QuantReg
MSE	0.172	0.172	0.172	0.166	0.166	0.166
MAE	0.332	0.332	0.331	0.312	0.312	0.312
MSD tStat	0.730	0.920	0.328	1.723	2.025	1.446
MAD tStat	0.529	0.604	0.374	1.118	1.105	1.039
Dev. Q(0.75)	0.000	0.000	0.000	0.005	0.001	0.001
Dev. Q(0.90)	-0.000	-0.000	-0.000	-0.008	-0.007	-0.033
Dev. Q(0.99)	-0.039	-0.041	0.063	-0.171	-0.147	-0.276
SAD2	0.010	0.019	0.013	0.116	0.076	0.068

Legend: MSD (MAD) tStat: Mean Squared (Absolute) Deviations of t Statistics, Dev. Q q Deviation between quantile q of uncensored and imputed wages, SAD2: Sum of Absolute second finite Derivatives.

References

Chow, W.K. (1979). "A Look at Various Estimators in Logistic Models in the Presence of Missing Values". In: *Proceedings of Business and Economics Section, American Statistical Association*, pp. 417–420.

Anderson, A. B., A. Basilevsky, and D. P. J. Hum (1983). "Handbook of Survey Research". In: ed. by P. H. Rossi, J.D. Wright, and A. Anderson. Academic Press. Chap. Missing Data: A Review of the Literature, pp. 415–492.

Honoré, Bo (1993). "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects". In: *Econometrica* 60(3), pp. 553–565.

Meng, Xiao-Li (1994). "Multiple-Imputation Inference with Uncongenial Sources of Input (with discussion)". In: *Statistical Science* 9, pp. 538–573.

Jones, Michael P. (1996). "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression Models". In: *Journal of the American Statistical Association* 91.433, pp. 222–230.

Abowd, John, Francis Kramarz, and David Margolis (1999). "High Wage Workers and High Wage Firms". In: *Econometrica* 67.3, pp. 251–333.

Chernozhukov, Victor and Han Hong (2002). "Three-Step Censored Quantile Regressions and Extramarital Affairs". In: *Journal of the American Statistical Association* 97(595), pp. 872–882.

Hsiao, Cheng (2003). *Analysis of Panel Data*. 2nd. Cambridge: Cambridge University Press.

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg (2009). "Revisiting the German Wage Structure". In: *Quarterly Journal of Economics* 124.2, pp. 843–881.

Compson, Michael (2011). "The 2006 Earnings Public-Use Microdata File: An Introduction". In: *Social Security Bulletin* 71.4, pp. 33–59.

Card, David, Jög Heining, and Patrick Kline (2013). "Workplace Heterogeneity and the rise of West German Wage Inequality". In: *Quarterly Journal of Economics* 128(3), pp. 967–1015.

Bundesamt, Statistisches (2018). *Qualitätsbericht Verdiensstrukturerhebung*. Tech. rep. Statistisches Bundesamt.

U.S. Census Bureau (2022a). *Changes to 2022 CPS Public Use Microdata Files*. `https://www.census.gov/content/dam/Census/programs-surveys/cps/updated-2022-cps-puf-changes.pdf`.

— (2022b). *Topcoding of Usual Hourly Earnings*. `https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/topcoding-of-usual-hourly-earnings.html`.