IAB DiscussionPaper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 5/2005

Analyzing the Changing Gender Wage Gap based on Multiply Imputed Right Censored Wages

Hermann Gartner and Susanne Rässler

Bundesagentur für Arbeit

Analyzing the Changing Gender Wage Gap based on Multiply Imputed Right Censored Wages

Hermann Gartner and Susanne Rässler (IAB)

Analyzing the Changing Gender Wage Gap based on Multiply Imputed Right Censored Wages*

Hermann Gartner and Susanne Rässler†

February 15, 2005

In order to analyze the gender wage gap with the German IAB-employment register we have to solve the problem of censored wages at the upper limit of the social security system. We treat this problem as a missing data problem. We regard the missingness mechanism as not missing at random (NMAR, according to Little and Rubin, 1987, 2002) as well as missing by design. The censored wages are multiply imputed by draws of a random variable from a truncated distribution. The multiple imputation is based on Markov chain Monte Carlo (MCMC) technique. We complete the dataset with this technique in order to apply a Juhn-Murphy-Pierce-decomposition. As the main sources for the narrowing gender wage gap from 1991 to 2001 we identify an improvement of women's position within the wage distribution.

JEL-code: C15, J16

Keywords: Juhn-Murphy-Pierce-decomposition, multiple imputation, missing data

1 Introduction

For studying the sources of the changing gender wage gap we need exact information about the distribution of the wages of males and females. If we knew the position for each male and female worker within the wage distribution, we could use a Juhn-Murphy-Pierce-decomposition to identify the sources of the changing gender wage gap. The gender wage gap can, for example, decrease because the position of women in the wage distribution relative to men is improved. Another possible reason for a decreasing

*We thank the DFG (project Al 393 / 6-3 – Gender-Specific Wages and Organisations) for financial support and Lutz Bellmann for helpful hints.

†Institute for Employment Research - Nuremberg, Germany.

gender gap is, that the entire distribution becomes mores compressed and therefore also the differences between male and female wages get lower.

We want to investigate the gender wage gap with the German IAB employment sample[1]. But the dataset lacks information about the entire wage distribution. Since the data stems from the social security accounts, the wages are only given up to the contribution limit according to the social security system, thus, the wages are right censored. For employees with wages above the limit, only the limit is reported. In some other countries similar problems exist with censored wages of administrative datasets.

To allow the analysis as mentioned above based on such censored data, we treat this problem as a missing data problem. In this special case we regard the missingness mechanism as not missing at random (NMAR, according to Little and Rubin, 1987, 2002) as well as missing by design. The first because the missingness depends on the value itself, i.e., if the limit is exceeded the true value will not be reported but the limit, say a. The latter because the data are missing due to the fact that they were not asked. A common approach to handle missing data is multiple imputation which means that every missing value is randomly imputed for m times, Rubin (1978, 1987, 1996). In our case this basically contains draws of the wages whenever the limit is reported. Thus, random draws of a random variable from a truncated distribution have to be performed.

The aim of this paper is to present a refined multiple imputation technique based on a suggestion of Chib (1992) to impute wages above the limit according to the social security system and to use this completed dataset to decompose the gender wage gap with the Juhn-Murphy-Pierce-technique. For this purpose we use data from the IAB employment sample for the years 1991 and 2001.

2 Analyzing the Gender Wage Gap: Juhn-Murphy-Pierce-Decomposition

We use a decomposition technique proposed by Juhn et al. (1993) to analyze the source of the change in the gender wage gap. The technique is used in several studies by Blau and Kahn (1994, 1996, 1997) to compare gender earning differences across time and across countries. The dependent variable of our estimates is the gross daily wage. We compare the gender wage gap of full-time employees in 1991 with the gap in 2001.

In the first step, we estimate an augmented Mincerian wage equation (Mincer, 1974) for males, which contains as proxies for human capital the potential work experience and education. Further we include dummies for 15 industrial sectors and for 12 firm size categories. We compute the potential experience according to *experience = age-6-years of schooling*. The years of schooling are as follows: For lower or intermediate secondary school only: 10 years; lower or intermediate secondary school with vocational training: 12.125 years; academic secondary school: 13 years; academic secondary school with vocational training: 15.125 years; college: 15 years; university: 18 years.

The wage equation is given by

[1] A documentation of the dataset can be found in Bender et al. (2000).

$$\ln wage_{it} = X'_{it}\beta_t + \sigma_t\theta_{it} \tag{1}$$

where $wage_{it}$ is the daily wage of person i and X_{it} is the vector of covariates. $t \in (91;01)$ stands for the years 1991 and 2001. σ_t is the standard deviation of male's residuals. θ_{it} is the standardized residual. It is calculated by $\frac{\ln wage_{it} - X_{it}\beta_t}{\sigma_t}$.

The mean gender wage gap at year t is given by

$$D_t = \overline{\ln wage}_{mt} - \overline{\ln wage}_{ft} = \Delta X'_t\beta_t + \sigma_t\Delta\theta_t. \tag{2}$$

Δ indicates the difference between the means of the variables for males and females: $\Delta x \equiv \overline{x}_m - \overline{x}_f$. As implied by the OLS-principle, the mean of the standardized error terms θ_{it} for males is zero, because the estimation of β_t is only done for males. Therefore $\sigma_t\Delta\theta_t = -\sigma_t\overline{\theta}_{ft}$.

We can decompose the change of the gender wage gap after some manipulation of (2):

$$D_{01} - D_{91} = \underbrace{(\Delta X_{01} - \Delta X_{91})'\beta_{01}}_{\text{endowment effect}} + \underbrace{\Delta X'_{91}(\beta_{01} - \beta_{91})}_{\text{observed price effect}} + \underbrace{(\overline{\theta}_{f91} - \overline{\theta}_{f01})\sigma_{01}}_{\text{gap effect}} + \underbrace{\overline{\theta}_{f91}(\sigma_{91} - \sigma_{01})}_{\text{unobserved price effect}} \tag{3}$$

The term $\overline{\theta}_{f91}\sigma_{01}$ is constructed as follows: Each woman in year 1991 is assigned to the percentile of the male residuals of the same year[2]. These women get the residual of males in 2001 at the same percentile. The mean of these residuals is $\overline{\theta}_{f91}\sigma_{01}$.

The endowment effect is the change in wage inequality attributed to changes in gender specific endowment. For example: If women's relative endowment with human capital rises, the wage gap will decrease. The endowment effect then is negative.

The second term is the observed price effect. It captures changes in the evaluation of the endowment. If males are better endowed with human capital, an increasing reward of one unit human capital will rise the wage gap. This leads to a positive unobserved price effect.

The gap effect represents the change in the relative position of females within the wage distribution of males after adjusting for differences in observed endowment. If for example discrimination decreases, the position of women become better. This lowers the wage gap; the gap effect would be negative.

The unobserved price effect captures the change of the wage gap attributed to the change of the variance of wages controlled for observed endowment. If the variance of wages rises, then the wage gap would rise and the observed price effect would be positive.

3 Multiple Imputation

To start with, let $Y = (Y_{obs}, Y_{mis})$ denote the random variables concerning the data with observed and missing parts. In our specific situation this means that for all units with

[2]We split the male residuals in 100 percentiles.

wages below the limit a each data record is complete, i.e., $Y = (Y_{obs}) = (X, wages)$. For every unit with a value of the limit a for its wage information we treat the data record as partly missing, i.e., $Y = (Y_{obs}, Y_{mis}) = (X, ?)$. Thus, we have to multiply impute the missing data $Y_{mis} = wage$.

3.1 The Basic Principle

The theory and principle of multiple imputation (MI) originates from Rubin (1978). The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually also valid from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$ of the missing data given the observed data. Since it is often difficult to draw from $f_{Y_{mis}|Y_{obs}}$ directly, a two-step procedure for each of the m draws is useful:

(a) First, we make random draws of the parameters Ξ according to their observed-data posterior distribution $f_{\Xi|Y_{obs}}$,

(b) then, we perform random draws of Y_{mis} according to their conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$.

Because

$$f_{Y_{mis}|Y_{obs}}(y_{mis}|y_{obs}) = \int f_{Y_{mis}|Y_{obs},\Xi}(y_{mis}|y_{obs},\xi) f_{\Xi|Y_{obs}}(\xi|y_{obs}) d\xi \tag{4}$$

holds, with (a) and (b) we achieve imputations of Y_{mis} from their posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$. Due to the data generating model used, for many models the conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$ is rather straightforward. Often it can be formulated for each unit with missing data easily.

In contrast, the corresponding observed-data posteriors $f_{\Xi|Y_{obs}}$ usually are difficult to derive for those units with missing data, especially when the data have a multivariate structure and different missing data patterns. The observed-data posteriors are often no standard distributions from which random numbers can easily be generated. However, simpler methods have been developed to enable multiple imputation based on Markov chain Monte Carlo (MCMC) techniques.[3] In MCMC the desired distributions $f_{Y_{mis}|Y_{obs}}$ and $f_{\Xi|Y_{obs}}$ are achieved as stationary distributions of Markov chains which are based on the complete-data distributions, that is easier to compute.

To proceed further, let θ denote a scalar quantity of interest that is to be estimated, such as a mean, variance, or correlation coefficient. Notice that now θ can be completely different from the data model used before to create the imputations. Although θ (analysis) could be an explicit function of ξ (imputation), one of the strengths of the multiple imputation approach is that this need not be the case. In fact, θ (analysis) could even be the parameter of the imputation model, then the imputation and the analysis model are the same and are said to be congenial (Meng 1995). However, multiple imputation is designed for situations where the analyst and the imputer are different, thus, the analyst's

[3]These are extensively discussed by Schafer (1997).

model could be quite different from the imputer's model. As long as the two models are not overly incompatible or the fraction of missing information is not high, inferences based on the multiply imputed data should still be approximately valid. Moreover, if the analyst's model is a sub-model of the imputer's model, i.e., the imputer uses a larger set of covariates than the analyst and the covariates are good predictors of the missing values, then MI inference is superior to the best inference possible using only the variables in the analyst's model. This property is called superefficiency by Rubin (1996). On the other hand, if the imputer ignores some important correlates of variables with missing data, but these variables are used in the analyst's model, then the results will be biased.

Now let $\widehat{\theta} = \widehat{\theta}(Y)$ denote the statistic that would be used to estimate θ if the data were complete. Furthermore, let $\widehat{var}(\widehat{\theta}) = \widehat{var}(\widehat{\theta}(Y))$ be the variance estimate of $\widehat{\theta}(Y)$ based on the complete dataset. We also assume that with complete data, tests and interval estimates which are based on the normal approximation

$$(\widehat{\theta} - \theta)/\sqrt{\widehat{var}(\widehat{\theta})} \sim N(0,1) \tag{5}$$

should work well. Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions.

Suppose now that the data are missing and we make $m > 1$ independent simulated imputations $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, $\ldots$, $(Y_{obs}, Y_{mis}^{(m)})$ enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(Y_{obs}, Y_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(Y_{obs}, Y_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. From these m imputed datasets the multiple imputation estimates are computed.

The MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{6}$$

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$, we first calculate the "between-imputation" variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1}\sum_{t=1}^{m} (\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{7}$$

and then the "within-imputation" variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m}\sum_{t=1}^{m} \widehat{var}(\widehat{\theta}^{(t)})\,. \tag{8}$$

Finally, the estimated total variance is defined by

$$\begin{aligned} \widehat{var}(\widehat{\theta}_{MI}) &= T = \widehat{var}(\widehat{\theta})_{within} + (1 + \frac{1}{m})\widehat{var}(\widehat{\theta})_{between} \\ &= W + \frac{m+1}{m}B. \end{aligned} \tag{9}$$

The term $((m+1)/m)B$ enlarges the total variance estimate T compared to the usual analysis of variance with $T = B + W$; $(m+1)/m$ is an adjustment for finite m. An estimate of the fraction of missing information γ about θ due to nonresponse is given by

$$\widehat{\gamma} = \frac{(1+1/m)B}{T}. \tag{10}$$

For large sample sizes, tests and two-sided $(1-\alpha)100\%$ interval estimates can be based on the Student's t-distribution

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{11}$$

with the degrees of freedom

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2 \tag{12}$$

From (11) we can see that the multiple imputation interval estimate is expected to produce a larger interval than an estimate based only on one single imputation (SI). The multiple imputation interval estimates are widened to account for the missing data uncertainty and simulation error. Using only one singly imputed dataset, in general, will lead to an underestimation of uncertainty and thus produce variance estimates that are too low and p-values that are too significant.

3.2 Imputation Model

We assume that for person i the wage in logs is given by

$$y_i^* = x_i'\beta + \epsilon_i \tag{13}$$

where $\epsilon \overset{iid}{\sim} N(0, \tau^{-2})$

We observe the wage $y_{obs} = y_i^*$ only if the wage is under the threshold a. If the wage is above a, we observe a instead of y_i^*:

$$y_i = \begin{cases} y_{obs} & \text{if } y_i^* \leq a \\ a & \text{if } y_i^* > a \end{cases} \tag{14}$$

We impute for the a estimations z of the true wages. Thus, we define $y = (y_{obs}, a)$ and $y_z = (y_{obs}, z)$. Then, z is a truncated variable in the range (a, ∞) and its conditional predictive distribution is given by

$$f(z|y, \beta, \tau^2) = \frac{f_N(z|x'\beta, \tau^{-2})}{1 - \Phi(\tau a - \tau x'\beta)} \tag{15}$$

where $a < z < \infty$. According to Chib (1992) we get a data augmentation algorithm and Gibbs sampler based on the full conditional distributions according to

$$f(\beta|y, z, \tau^2) = f_N(\beta|\hat{\beta}_z, \tau^{-2}(X'X)^{-1}) \tag{16}$$

$$f(\tau^2|y,z,\beta) = f_G(\tau^2|n/2, \sum_{i=1}^{n}(y_z - x'\beta_z)^2/2) \tag{17}$$

where $\widehat{\beta}_z^{(t)} = (X'X)^{-1}X'y_z^{(t)}$ is the usual OLS estimate based on the complete dataset.

To receive valid imputations and random draws of the parameters from their observed data distribution according to the rule presented in (4), we finally propose a MCMC technique as mentioned earlier. To start the chain we adopt the starting values $\beta^{(0)}, \tau^{2(0)}$ from a ML tobit estimation.

Imputation-Step:

First, we randomly draw values for the missing variables from the truncated distribution according to

$$z_i^{(t)} \sim N(x_i'\beta, \tau^{-2(t)}) \tag{18}$$

Note that alternatively a accept-rejection algorithm could be applied instead of drawing directly from the truncated distribution. But the computational time gets too large with such an amount of missing data and these large datasets. The proposed new algorithm is computationally by far friendlier and described in the appendix.

Then the OLS regression is computed based on the imputed datasets according to

$$\widehat{\beta}_z^{(t)} = (X'X)^{-1}X'y_z^{(t)} \tag{19}$$

Then we produce new random draws for the parameters according to their complete data posterior distribution. Since drawings from a gamma distribution are complicated to compute with STATA we use a slight modification of (17).

Posterior-Step:

$$g \sim \chi^2(n-k) \tag{20}$$

$$\tau^{2(t+1)} = \frac{g}{RSS} \tag{21}$$

where RSS is the residual sum of squares: $RSS = \sum_{i=1}^{n}\left(y_{z_i}^{(t)} - x_i'\widehat{\beta}_z^{(t)}\right)^2$

$$\beta^{(t+1)} \sim N\left(\widehat{\beta}_z^{(t)}, \tau^{-2(t+1)}(X'X)^{-1}\right) \tag{22}$$

k is the number of columns of X. The covariates contained in X are: potential experience (linear, quadratic and cubic), 6 educational levels, 11 occupational group according to Blossfeld (1985), 12 categories of firm size, 15 industrial categories. We repeat the imputation-step and the probability-step 11,000 times and use ($z_i^{(2,000)}$, $z_i^{(3,000)}$, ... , $z_i^{(11,000)}$) to obtain 10 completed datasets. The imputation is done separately for males and females and for 1991 and 2001. For each dataset the imputation routine requires

Table 1: Wage of males and females, 1991 and 2001

Year	$\ln wage_m$	$\ln wage_f$	$\Delta \ln wage$	σ_t
1991	4.481	4.126	0.355	0.335
2001	4.490	4.204	0.286	0.346

Notes: log of daily wage of fulltime employees in Euros, western Germany. Source: German IAB employment sample (IABS)

about six hours. Different analyses of the convergence of the chains do not show any problems.

We have assumed a lognormal distribution of the wages. The normal distribution is notoriously sensitive according to outliers (see Gelman et al., 2003, S. 443). Especially by using transformations of a normal distribution this problem may be considerable (as discussed by Rubin, 1983). To touch upon the applicability of our distribution assumption, we compare the distribution of our imputed wages with the distribution calculated with the German socioeconomic Panel (GSOEP). Our imputed wages lay about in the same range as the wages in the GSOEP.[4] Thus, our imputed datasets are used for the analyst's model as described above.

4 The Dataset and Results

The German IAB employment sample (IABS) is a 2 percent random sample of all employees covered by the social security. Accordingly self employed, family workers and civil servants (Beamte) are not included. The dataset represents 80 percent of the employees in Germany. The data are the base for calculating the benefits from the social security system. Therefore they are highly reliable. The IABS includes among others information about age, sex, education, wage, and the occupational group. We exclude for the analysis part time workers, apprentices and all cases where earnings are below twice the limit of minor employment because this wages are implausible for fulltime workers. Further we restrict our data to west German residents. Our dataset contains for the year 1991 223,069 males and 112,694 females, for the year 2001 188,850 males and 94,369 females. In 1991 there are 35,688 (16.0%) censored wages of males and 3,307 (2.9%) of females, in 2001 30,546 (16.2%) of males and 4,293 (4.5%) of females.

Descriptive statistics of the data show: The log wage of men rises over the period (1991-2001) by 0.009 from 4.481 to 4.490, whereas the log wage of women rises more rapidly by 0.078 from 4.126 to 4.204. Therefore the wage gap decreases from 0.355 to 0.286 by 0.0689 (table 1).

The estimation and the decomposition is separately done for each of the ten imputed datasets. The mean and the variance of the decomposition effects are presented in table 2. Changes in the relative endowment of women account for 50.8% (=0.035/0.0689) of the lower gender pay gap. The largest fraction of the endowment effect (0.0309 log

[4] A deeper going comparison of the imputed dataset with other data sources is a task for future research.

Table 2: Decomposition of Change in the Gender Wage Gap, 1991-2001

Effect	mean	Var·10^7
endowment effect	-0.0350	0.0105
human capital	-0.0309	0.0075
firm size	-0.0076	0.0029
industry	-0.0008	0.0066
obs. price effect	0.0144	0.5792
human capital	0.0065	0.0874
firm size	0.0033	0.0161
industry	0.0043	0.2666
gap effect	-0.0728	0.3448
unobs. price effect	0.0245	0.1506
Sum, gender-specific	-0.1122	0.0380
Sum, wage structure	0.0386	0.0732
total sum	-0.0689	0.4301

Notes: Juhn-Murphy-Pierce-decomposition; depend variable of regression: log of daily wage of fulltime employees; covariates: potential work experience, education, dummies for 12 firm size categories and for 15 industrial sectors; western Germany. Source: German IAB employment sample (IABS)

points) is attributable to the improvement of women's endowment with human capital in the 90's.

The endowment effect of the firm size and industrial dummies can be interpreted as changes in gender-specific sorting across firms and industries. Large firms pay higher wages than small firms (a survey on this topic is Oi and Idson, 1999). As males work more frequently in large firms than females, males receive more frequently this additional wage premium. But the share of males in large firms declines from 1991 to 2001. The change in the gender specific sorting between large and small firms contributes to 0.0076 log points of the declining gender gap. Changes in sorting between industries are very small; they account for nearly zero log points of the change of the gender gap.

The observed price effect is positive and amounts 0.0144 log points. This indicates, that changes in the returns alone would have increased the gender wage gap. Changes in the returns on human capital accounts for 0.0065 log points. As the literature about skill-biased technological change found, skill premiums increased in most industrial countries in the last decades because of technological change (for a discussion see Acemoglu, 2002). Because males are better endowed with human capital than females, male's wage rises faster than female's wage and this stretch the gender pay gap.

Rising wage differentials across industries amplify the gender gap by 0.0043 log points. A trend in rising industrial wage differentials in Germany is already shown by Bellmann and Gartner (2003). This trend widens the gender wage gap, because men works more often in high wage industries then women. We found also an increase in firm size differentials that widens the gender pay gap by 0.003 log points.

The gap effect of -0.0728 indicates, that women have improved the position of their wage residuals in the distribution of male's wage residuals. The reduction of the gender wage gap because of the gap effect is by 5.7% greater then the observed reduction of the wage gap. The gap effect could be caused by both an improvement of women's unobserved productivity or a lowering of discrimination against women. The unobserved price effect works in the opposite direction: The standard deviation of the residuals σ_t rises from 0.335 to 0.346 log points (see table 1). This pumps up the wage gap by 0.0245 log points.

To summarize: in the 1990's there is a general trend of wage structure, caused by rising observed and unobserved prices, that widens the gender wage gap by 0.0384 log points. But improvements in observed and unobserved endowments, a reduction in gender-specific sorting and in discrimination reduce the gender wage gap by 0.1122 log points. Similar as Blau and Kahn (1994, 1997) argue for the US in the 1980s, the trend of the gender wage gap can be described as a swimming upstream against the rising wage inequality.

5 Appendix: Random Draws From a Truncated Distribution

Assuming we have a normal distributed variable $e \sim N(\mu, \sigma^2)$. The lower limit is a. For easier notation we define: $\alpha = \frac{(a-\mu)}{\sigma}$ and $\epsilon = \frac{(e-\mu)}{\sigma}$. ϵ is then standard normal distributed:

$$g(\epsilon) = \phi(\epsilon) \tag{23}$$

$\phi(x)$ is the density function of the standard normal distribution. We have to draw a random value ϵ_i from this distribution under the condition that $\epsilon_i > \alpha$. Therefore we have to draw from a truncated distribution.

The density function of a truncated standard normal distribution is

$$g(\epsilon|\epsilon > \alpha) = \frac{f(\epsilon)}{1 - \Phi(\alpha)}, \qquad \epsilon > \alpha. \tag{24}$$

$\Phi(x)$ is the standard normal distribution function.

The truncated distribution function $G(\epsilon)\epsilon \to Y$ with $Y \in [0, 1]$ is

$$G(\epsilon) = \int_{\alpha}^{\epsilon} \frac{\phi(z)}{1 - \Phi(\alpha)} dz. \tag{25}$$

Splitting the integral

$$G(\epsilon) = \frac{1}{1 - \Phi(\alpha)} \Big(\int_{-\infty}^{\epsilon} \phi(t)dt - \int_{-\infty}^{\alpha} \phi(t)dt \Big) \tag{26}$$

leads to

$$G(\epsilon) = \frac{1}{1 - \Phi(\alpha)} (\Phi(\epsilon) - \Phi(\alpha)). \tag{27}$$

For generating the random variable with STATA we need the invers function $G^{-1}(Y) = \epsilon$. The solution of $Y = \frac{1}{1-\Phi(\alpha)}(\Phi(\epsilon) - \Phi(\alpha))$ for $\Phi(\epsilon)$ is:

$$Y(1 - \Phi(\alpha)) + \Phi(\alpha) = \Phi(\epsilon) \tag{28}$$

If we take on both sides the inverse Φ^{-1} we get

$$\Phi^{-1}(Y(1 - \Phi(\alpha)) + \Phi(\alpha)) = \epsilon. \tag{29}$$

Thus ϵ can be generated in STATA with:

$$\epsilon\texttt{=invnorm(uniform()*(1-norm(}\alpha\texttt{))+norm(}\alpha\texttt{))} \tag{30}$$

$Y \in [0, 1]$ is substituted by `uniform()`, which generate an unique distribution on the interval $[0, 1]$. This command is adopted for the imputation program to draw the censored wages.

References

Acemoglu, D. (2002). Technical change, inequality, and the labour market. *Journal of Economic Literature 40*, 7–72.

Bellmann, L. and H. Gartner (2003). Fakten zur Entwicklung der qualifikatorischen und sektoralen Lohnstruktur. *Mitteilungen zur Arbeitsmarkt- und Berufsforschung 4* (493-508).

Bender, S., A. Haas, and C. Klose (2000). IAB employment subsample 1975-1995. opportunities for analysis provided by the anonymised subsample. IZA Discussion Paper No 117. IZA, Bonn.

Blau, F. D. and L. M. Kahn (1994). Rising wage inequality and the U.S. gender gap. *American Economic Review 84* (2), 23–28.

Blau, F. D. and L. M. Kahn (1996). The gender earnings gap: Some international evidence. *Economica 63*, 29–62.

Blau, F. D. and L. M. Kahn (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labour Economics 15* (1), 1–42.

Blossfeld, H.-P. (1985). *Bildungsexpansion und Berufschancen.* Frankfurt am Main: Campus.

Box, G. E. P. and G. C. Tiao (1992). *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Chib, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics 51*, 79–99.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis* (2 ed.). Boca Raton: Chapman & Hall/CRC.

Juhn, C., K. M. Murphy, and B. Pierce (1993). Wage inequality and the rise in return to skill. *The Journal of Political Economy 101*, 410–442.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2 ed.). New York: John Wiley.

Little, R. J. A. and D. R. Rubin (1987). *Statistical Analysis with Missing Data* (1 ed.). New York: John Wiley.

Meng, X. L. (1995). Multiple-imputation inferences with uncongenial source of input (with discussion). *Statistical Science 10*, 538–573.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York: Columbia University Press.

Oi, W. Y. and T. L. Idson (1999). *Firm Size and Wages*, Volume IIIc of *Handbook of Labor Economics*, Chapter 33, pp. 2155–2214. Amsterdam: Elsevier Science.

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Sections of the American Statistica*, 20–40.

Rubin, D. B. (1983). A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. In *Scientific Inference, Data Analysis and Robustness*, pp. 213–244. New York: Academic Press, Inc.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*, 473–489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

In dieser Reihe sind zuletzt erschienen

Recently published

No.	Author(s)	Title	Date
1/2004	Bauer, Th. K., Bender, St., Bonin, H.	Dismissal Protection and Worker Flows in Small Establishments	7/2004
2/2004	Achatz, J., Gartner, H., Glück, T.	Bonus oder Bias? Mechanismen geschlechtsspezifischer Entlohnung	7/2004
3/2004	Andrews, M., Schank, Th., Upward, R.	Practical estimation methods for linked employer-employee data	8/2004
4/2004	Brixy, U., Kohaut, S., Schnabel; C.	Do newly founded firms pay lower wages? First evidence from Germany	9/2004
5/2004	Kölling, A, Rässler, S.	Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models	10/2004
6/2004	Stephan, G, Gerlach, K.	Collective Contracts, Wages and Wage Dispersion in a Multi-Level Model	10/2004
7/2004	Gartner, H. Stephan, G.	How Collective Contracts and Works Councils Reduce the Gender Wage Gap	12/2004
1/2005	Blien, U., Suedekum, J.	Local Economic Structure and Industry Development in Germany, 1993-2001	1/2005
2/2005	Brixy, U., Kohaut, S., Schnabel, C.	How fast do newly founded firms mature? Empirical analyses on job quality in start-ups	1/2005
3/2005	Lechner, M., Miquel, R., Wunsch, C.	Long-Run Effects of Public Sector Sponsored Training in West Germany	1/2005
4/2005	Hinz, Th., Gartner, H.	Lohnunterschiede zwischen Frauen und Männern in Branchen, Berufen und Betrieben	2/2005

Impressum

IABDiscussionPaper
No. 5 / 2005

Herausgeber
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Redaktion
Regina Stoll, Jutta Palm-Nowak

Technische Herstellung
Jutta Sebald

Bezugsmöglichkeit
Volltext-Download dieses DiscussionPaper
unter:
http://doku.iab.de/discussionpapers/2005/dp0505.pdf

IAB im Internet
http://www.iab.de

Rückfragen zum Inhalt an
Hermann Gartner, Tel. 0911/179-3386,
oder e-Mail: hermann.gartner@iab.de