FDZ Methodenreport

Methodische Aspekte zu Arbeitsmarktdaten

Nr. 2/2005

The imputation of wages above the contribution limit with the German IAB employment sample

Hermann Gartner

Bundesagentur für Arbeit

Contents

Data access

The data used in this Methodenreport relate to all public available IABS and LIAB versions. For more information on the data, please have a look at http://fdz.iab.de at the category “Personendaten” for the IABS and at the category “Integrierte Betriebs- und Personendaten“ for the LIAB (currently only in German language).

1 Introduction

Researchers analyzing wages with the German IAB employment sample are confronted with a censoring problem: The data comes from the social security accounts. Therefore the wages are only given up to the contribution limit of the social security (A documentation of the data set is found in Bender, Haas, and Klose (2000)). To avoid biased estimation the censored wages can be imputed with estimated wages. The estimation can be done with a tobit model. The independent variables of the estimation are for example qualification, region and branch.

But for the imputation of the wage you should not use the expected value of the log wage

$$E(lnw_i) = x_i'\widehat{\beta}. \quad (1)$$

If you use the expected value, the correlation of the imputed wages with the covariates would be higher than the correlation of the true unobserved wages with the covariates. The standard errors of estimations with the imputed data set would then be to low (A discussion of the problems with imputation of missing data is found in Little and Rubin (1987); Schafer (1997)).

A better imputation procedure is: Draw a random value lnw^{imp} from a normal distribution $N(x'\widehat{\beta}, \sigma^2)$ [1]. In other words: Add to the expected wage an error term η with the standard deviation σ.

For σ we can use the estimated σ from the tobit estimation.

$$lnw_i^{imp} = x_i'\widehat{\beta} + \eta_i \quad (2)$$

We know: The true value is above the contribution limit of the social security. Therefore we have to draw the imputed wage lnw^{imp} so, that it is also above the limit. We need a drawing of a random variable from a truncated distribution.

[1] An assumption is that the wages follow a log-normal distribution. Empirically this is a good approximation.

2 Drawings From a Truncated Distribution

Suppose 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:

$$g(\epsilon) = \phi(\epsilon) \tag{3}$$

$\phi(x)$ 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

$$g(\epsilon | \epsilon > \alpha) = \frac{f(\epsilon)}{1 - \Phi(\alpha)}, \qquad \epsilon > \alpha. \tag{4}$$

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

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

$$G(\epsilon) = \int_\alpha^\epsilon \frac{\phi(z)}{1 - \Phi(\alpha)} dz. \tag{5}$$

Splitting the integral

$$G(\epsilon) = \frac{1}{1 - \Phi(\alpha)} \Big(\int_{-\infty}^\epsilon \phi(t)dt - \int_{-\infty}^\alpha \phi(t)dt \Big) \tag{6}$$

leads to

$$G(\epsilon) = \frac{1}{1 - \Phi(\alpha)} (\Phi(\epsilon) - \Phi(\alpha)). \tag{7}$$

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:

$$Y(1 - \Phi(\alpha)) + \Phi(\alpha) = \Phi(\epsilon) \tag{8}$$

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

$$\Phi^{-1}(Y(1 - \Phi(\alpha)) + \Phi(\alpha)) = \epsilon. \tag{9}$$

Thus ϵ can be generated in STATA with:

$$\epsilon\texttt{=invnorm(uniform()*(1-norm(}\alpha\texttt{))+norm(}\alpha\texttt{))} \qquad (10)$$

$Y \in [0,1]$ is substituted by `uniform()`, which generate an unique distribution on the interval [0, 1].

3 The Imputation of Wages

In our example ϵ is the imputed log wage lnw^{imp}, μ is the expected log wage $x'\widehat{\beta}$. a is the log of upper contribution limit $\ln(limit)$. For σ we use the estimated value $\widehat{\sigma}$. If we substitute this for the definition of ϵ we get:

$$\epsilon_i = \frac{lnw_i^{imp} - x_i'\widehat{\beta}}{\widehat{\sigma}} \qquad (11)$$

or

$$lnw_i^{imp} = \epsilon_i\widehat{\sigma} + x_i'\widehat{\beta} \qquad (12)$$

The censored wages can then be imputed with the command:

$$lnw_i^{imp}\texttt{=invnorm(uniform()*(1-norm(}\alpha_i\texttt{))+norm(}\alpha_i\texttt{))}\widehat{\sigma} + x_i'\widehat{\beta}, \qquad (13)$$

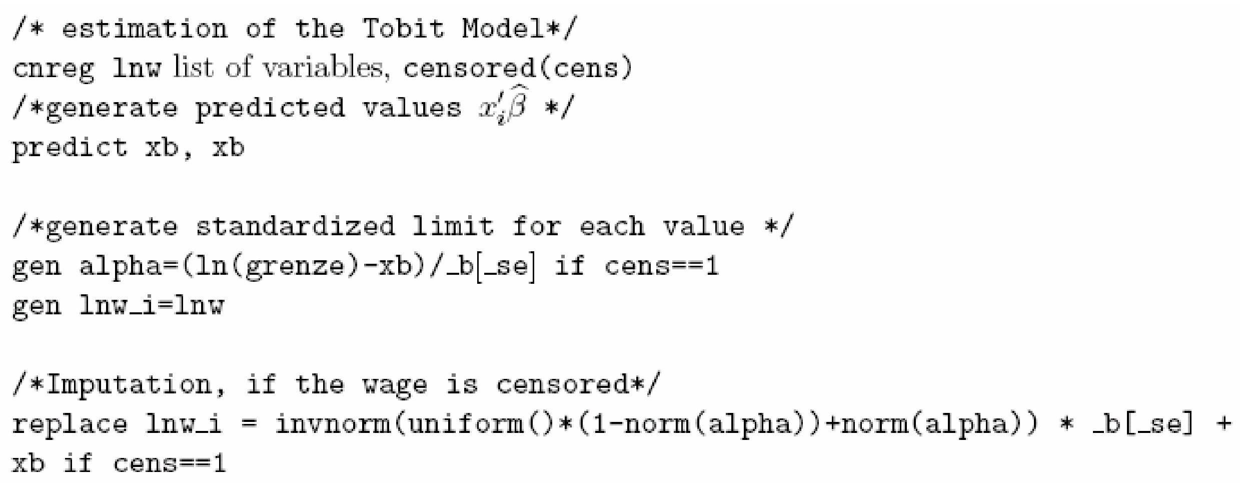
where

$$\alpha_i = \frac{\ln(limit) - x_i'\widehat{\beta}}{\widehat{\sigma}}. \qquad (14)$$

4 A Program to Impute Wages

The imputation procedure can be done with the following STATA-commands:

`cens = 1` if the wage is censored, else 0
`lnw` = observed log wage
`limit` = upper limit for social contribution



```
/* estimation of the Tobit Model*/
cnreg lnw list of variables, censored(cens)
/*generate predicted values $x_i'\widehat{\beta}$ */
predict xb, xb

/*generate standardized limit for each value */
gen alpha=(ln(grenze)-xb)/_b[_se] if cens==1
gen lnw_i=lnw

/*Imputation, if the wage is censored*/
replace lnw_i = invnorm(uniform()*(1-norm(alpha))+norm(alpha)) * _b[_se] +
xb if cens==1
```

This procedure is implemented in the ado-file `imputw`.

5 The Syntax

`imputw` *depvar varlist* [if *exp*] cens(*varname*) grenze(*varname*) [outvar(*varname*)]

`by...:` is possible.

depvar is the depend variable.

varlist is the list of indepent variables used in the tobit regression.

`cens(`*varname*`)` *varname* is a Dummy variable indication whether the wage is censored (1 if censored, 0 else).

`grenze(`*varname*`)` *varname* is the upper limit of the censored values (not in logs!).

`outvar(`*varname*`)` *varname* is the name of the new variable including the uncensored and the imputed values. The defauld name is *lnw_i*.

6 A Final Remark

Little and Rubin (1987) suggest that imputation should be done multiple and in a bayesian way. A bayesian estimation requires less restrictive distributional assumptions. Multiple imputation accounts for the additional uncertainty caused by the estimation with imputed values instead of the true values. In Gartner and Rässler (2005) we develop such a procedure for impute censored wages multiple. The procedure is based on a refined data augmentation (Markov-Chain-Monte-Carlo) algorithm. But this algorithm is somewhat more computational intensive than the above algorithm.

References

S. Bender, 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.

H. Gartner, and S. Rässler, 2005, “*Analyzing the changing gender wage gap based on multiply imputed right censored wages*,” IAB Discussion paper 5/05.

R. J. A. Little, and D. R. Rubin, 1987, “*Statistical Analysis with Missing Data”*, John Wiley, New York, 1 edn.

J. L. Schafer, 1997, “*Analysis of Incomplete Multivariate Data”*, Chapman & Hall, London.

Impressum

FDZ Methodenreport
Nr. 2/2005

Herausgeber
Forschungsdatenzentrum (FDZ)
der Bundesagentur für Arbeit
Regensburger Str. 104
D-90478 Nürnberg

Redaktion
Dagmar Herrlinger, Annette Kohlmann

Technische Herstellung
Dagmar Herrlinger

Rechte
Nachdruck – auch auszugsweise – nur mit
Genehmigung des FDZ gestattet

Bezugsmöglichkeit
Volltext-Download dieses Datenreports unter
http://doku.iab.de/fdz/reporte/2005/MR_2.pdf

FDZ im Internet
http://fdz.iab.de/

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