Die Imputation von Löhnen oberhalb der Beitragsbemessungsgrenze in der IAB-Beschäftigten-Stichprobe

Hermann Gartner

Methodenreport

Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg

5. November 2004

1 Einführung

Ein häufiges Problem, mit dem man bei Untersuchungen mit der Beschäftigten-Stichprobe konfrontiert ist, besteht darin, dass die Einkommen der Beschäftigten nur bis zur Beitragsbemessungsgrenze gemeldet werden. Die Daten sind daher ab dieser Grenze zensiert (eine Dokumentation des Datensatzes findet sich in Bender et al. (1996)). Um dadurch bedingte Verzerrungen der zu schätzenden Koeffizienten zu vermeiden, können die zensierten Werte durch Schätzungen imputiert werden. Hierzu wird der Lohn mit einem Tobit-Modell geschätzt, wobei die in der Beschäftigten-Stichprobe zur Verfügung stehen Variablen (Qualifikation, Region, Branche usw.) als erklärende Faktoren verwendet werden können. Als Schätzung für den Lohn sollte jedoch nicht der Erwartungswert

$$E(lnw) = x'\widehat{\beta} \tag{1}$$

verwendet werden (lnw ist der logarithmierte Lohn). Erstens würde durch so imputierte Werte eine zu große Korrelation mit den erklärenden Variablen vorgetäuscht (zur Diskussion von Problemen verschiedener Verfahren zur Ersetzung von fehlenden Werten sei auf Schafer (1997) verwiesen). Zweitens kann es Fälle geben, in denen der Erwartungswert unter der Beitragsbemessungsgrenze liegt, obwohl bekannt ist, dass er darüber liegen muss.

Um die korrekte Korrelation des Lohnes mit den Kovariablen zu erhalten, wird einen zufälligen Wert lnw^{imp} aus einer Normalverteilung $N(x'\widehat{\beta}, \sigma^2)$ gezogen[1]. Dem Erwartungswert ist also ein Störterm η mit der Standardabweichung σ hinzuzufügen. Für σ wird der geschätzte Wert aus der Tobit-Regression zugrundegelegt.

$$lnw^{imp} = x'\widehat{\beta} + \eta \tag{2}$$

Da der wahre Wert, der geschätzt wird, über der Beitragsbemessungsgrenze liegt, wird der imputierte Lohn aus einer gestutzten Verteilung gezogen, so dass der Lohn über dieser Schwelle liegt. Es ist also ein Verfahren zur Ziehung aus einer gestutzten Verteilung erforderlich.

2 Ziehungen einer gestutzten Zufallsvariabe

Gegeben sei eine normalverteilte Zufallsvariable $e \sim N(\mu, \sigma^2)$. Die Untergrenze sei a. Zur Vereinfachung der Schreibweise wird normiert: Es gelte $\alpha = \frac{(a-\mu)}{\sigma}$ und $\epsilon = \frac{(e-\mu)}{\sigma}$. ϵ ist somit standardnormalverteilt:

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

$\phi(x)$ ist die Dichtefunktion der Standardnormalverteilung. Aus dieser Verteilung sollen Werte gezogen werden, die über der Grenze α liegen. Es ist also eine gestutzte Verteilung

[1] Vorausgesetzt wird, dass die Verteilung der Löhne einer Lognormal-Verteilung folgt. Empirisch ist dies eine gute Approximation.

zugrundezulegen. Deren Dichtefunktion lautet

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

$\Phi(x)$ ist die Verteilungsfunktion einer standardnormalverteilten Zufallsvariable.

Die zugehörige Verteilungsfunktion ist eine Abbildung der Zufallsvariable ϵ auf einen Intervall zwischen null und eins. $G : \epsilon \to Y$ mit $Y \in [0, 1]$ ist

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

Umformen führt zu

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

oder

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

Um die Zufallswerte in STATA generieren zu können, ist die Inverse dieser Funktion $G^{-1}(Y) = \epsilon$ erforderlich. Auflösen der Gleichung $Y = \frac{1}{1-\Phi(\alpha)}(\Phi(\epsilon) - \Phi(\alpha))$ nach $\Phi(\epsilon)$ führt zu:

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

Daraus folgt

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

In STATA lässt sich die Variable ϵ somit generieren mit:

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

Wobei $Y \in [0, 1]$ durch die Funktion `uniform()` ersetzt wurde, welche eine gleichverteilte Zufallsvariable im Intervall $[0, 1]$ erzeugt.

3 Anwendung auf die Imputation des Lohnes

e ist bei uns der Logarithmus des zu imputierenden Lohnes lnw^{imp}, μ ist dessen Erwartungswert $x'\widehat{\beta}$. a ist die logarithmierte Betragsbemessungsgrenze $\ln(Beitr)$. Statt σ verwenden wir den Schätzer $\widehat{\sigma}$. Es gilt damit:

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

bzw.

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

Die zensierten Werte lassen sich dann imputieren durch den Befehl:

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

wobei gilt

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

4 Das Programm zur Imputation

Die einzelnen Schritte des Imputationsverfahrens lassen sich mit einem STATA-Programm folgender Gestalt durchführen:

`cens = 1` wenn Lohn zensiert ist, sonst 0
`lnw =` beobachteter logarithmierter Lohn
`grenze =` Betragsbemessungsgrenze

```
/* Schätzung des Tobit-Modell*/
cnreg lnw variablenliste, censored(cens)
/*Generieren der Prognose x_i'β̂ */
predict xb, xb

/*standardisierte Grenze für jede Beobachtung wird generiert*/
gen alpha=(ln(grenze)-xb)/_b[_se] if cens==1
gen lnw_i=lnw

/*Imputation, wenn Wert zensiert ist*/
replace lnw_i = invnorm(uniform()*(1-norm(alpha))+norm(alpha)) * _b[_se] +
xb if cens==1
```

Literatur

Bender, S., **Hilzendegen, J.**, **Rohwer, G.** und **Rudolph, H.** (1996). Die IAB-Beschäftigtenstichprobe. *Beiträge zur Arbeitsmarkt- und Berufsforschung*, (197).

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