Johannes Gutenberg Universität Mainz
Fachbereich Wirtschaftswissenschaften

Skript zur Voresung

Mikroökonometrie

Dr. Andrey Launov

Das vorliegende Manuskript ist für Studierende, die die Vorlesung *Mikroökonometrie* besuchen, als Begleitmaterial gedacht, um den Mitschrieb zu erleichtern. Es sollte nicht als Ersatz für die Teilnahme an den Lehrveranstaltungen gesehen werden. Ich kann nicht für die Vollständigkeit und die Fehlerfreiheit des Manuskripts garantieren.

WS 2009-10

Inhaltsverzeichnis

Notation

Wir werden versuchen an die folgende Notation zu halten:

- Zufallsvariablen

Y ist eine Zufallsvariable,

y ist eine Ausprägung der Zufallsvariable Y,

Die Zufallsvariable Y hat die kumulative Dichtefunktion $F(y)$, d.h. die Wahrscheinlichkeit, daß eine Ausprägung nicht Größer als y ist, ist $P(Y \leq y) = F(y)$,

Die entsprechende Dichtefunktion für Y ist $f(y)$, d.h. $F(y) = \int_{x \leq y} f(x)\, dx$. Zum Beispiel, wenn $y > 0$, dann $F(y) = \int_0^y f(x)\, dx$.

- Skalaren, Vektoren und Matrizen

x_i ist ein Skalar,

$\mathbf{x}_i$ ist ein Vektor (ein Spaltenvektor)

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,k} \end{bmatrix}$$

$\mathbf{X}$ ist eine Matrix. Diese Matrix besteht aus n transponierten Spaltenvektoren $\mathbf{x}_i$, $i = 1, ..., n$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ ... \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11} & ... & x_{1k} \\ x_{21} & ... & x_{2k} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{nk} \end{bmatrix}$$

- Funktionen

$\log(x)$ und $\ln(x)$ bedeuten das gleiche

Teil I

Theoretische Grundlagen

1 Was ist Mikroökonometrie?

- Mikroökonometrie = Theorie und Praxis der Modellierung von Mikrodaten.
- Mikrodaten = Unabhängig gezogene Einzelbeobachtungen auf dem nicht-aggregierten Ebene.

Typische Mikrodaten beinhalten in sich die Information über eine kleine Einheit, wie z.B., Beobachtungen über Individuen, Beobachtungen über Haushalte, Beobachtungen über Firmen, Transaktionen usw.

Das Ziel jedes mikroökonometrischen Verfahrens ist: Die Zusammenhänge zwischen den Erklärenden Variablen und der zuerklärenden Variable herauszufinden und zu quantifizieren.

Das Vorgehen ist, wie immer: Eine Stichprobe ziehen und aus der Dateninformation in dieser Stichprobe die obige zusammenhänge hinausbekommen.

Typische zuerklärende Variablen in den mikrodaten kommen in den verschiedensten Formaten vor

- Qualitative Daten
  - Binär

    Arbeitslos / Beschäftigt, Kreditwürdig / nicht Kreditwürdig
  - Multinomial (polytom)

    Bier / Wein / Whiskey, Vorlesung besuchen / Vorlesung nicht Besuchen, wenn nicht: Schlafen / Lesen / Spazieren gehen
  - Geordnet (polytom)

    Rating eines Unternehmens, Zufriedenheit mit der Regierung, mit dem/der Freund/in

- Quantitative / Unvollständig beobachtbare Daten
  - Zähldaten

    Anzahl Arztbesuche, Anzahl Kinder in der Familie
  - Unbeschränkte nichtnegative Daten

    Preise, Löhne
  - Beschränkt beobachtete Daten

    Löhne unter Beitragsbemessungsgrenze, Noten zu einer Studiengang zugelassenen Studenten

Unvollkommenheiten bei der zuerklärenden Variablen führen üblicherweise zu den nichtlinearen Zusammenhängen zwischen den Modelparameter und dem Erwartungswert der zuerklärenden Variable. Darum benötigen wir die entsprechende Modellen, die diese nichtlineare Zusammenhänge richtig spezifizieren und schätzen können.

Was kennen wir? - Das klassische lineare Regressionsmodell:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i,$$

$E(\varepsilon_i) = 0 \quad \Rightarrow \quad E(y_i) = \mathbf{x}_i'\beta$, d.h. der Erwartungswert der zuerklärenden Variable ist eine lineare Funktion der Parameter. Schätzmethode? *KQ*!

Was haben wir nun? - Das algemeine nichtlineare Regressionsmodell!

Der Erwartungswert der zuerklärenden Variable ist eine nichtlineare Funktion der Parameter, z.B. $E(y_i) = h(\mathbf{x}_i'\beta; \gamma)$, die Möglicherweise nicht nur von den Parameter β, sondern auch von den weiteren Parameter γ abhängig ist. Definiert man einen gesamten Parametervektor $\theta = \{\beta, \gamma\}$, dann, schreibt man $E(y_i) = h(\mathbf{x}_i; \theta)$ und somit

$$y_i = h(\mathbf{x}_i; \theta) + \varepsilon_i, \quad E(\varepsilon_i) = 0.$$

Schätzmethode?

1. *Nicht-lineare KQ*: wie früher, minimieren wir die quadrierte Abweichungen zwischen den Ausprägungen y_i und deren Mittelwerten $h(\mathbf{x}_i; \theta)$ um θ zu schätzen.

2. Wir bemerken, daß nicht nur der Störterm, sondern auch die zuerklärende Variable Y, deren Ausprägungen y_i wir beobachten, auch ihre eigene Verteilung hat. Der Erwartungswert dieser Verteilung ist genau $E(y_i) = h(\mathbf{x}_i; \theta)$; die Varianz und

weitere Momenten sind ebenfalls die funktionen von θ (und, möglicherweise $\mathbf{x}_i$). Dies erlaubt uns zu sagen, daß die Wahrscheinlichkeit eine gegebene Ausprägung y_i zu beobachten, ist

$$P\left(Y = y_i\right) = g\left(y_i; \mathbf{x}_i, \theta\right).$$

Dabei steht $g\left(y_i\right)$ für die Wahrscheinlichkeitsdichte der zuerklärenden Variable. Wir verwenden die Information über den Zusammenhang zwischen y_i, $\mathbf{x}_i$ und θ, die diese Dichte zusammenfasst, um θ zu schätzen. Dies impliziert eine neue Schätzmethode – die *Maximum Likelihood Schätzung.*

Beispiel 1 (Nichtlineare Zusammenhänge) *Betrachten wir eine binäre Variable y_i, z.B.*

$$y_i = \begin{cases} 1 & \textit{für Vorlesung besuchen} \\ 0 & \textit{für Vorlesung nicht besuchen} \end{cases}$$

y_i als Zufallsvariable folgt der Bernoulli Verteilung:

$$y_i = \begin{cases} 1 & \textit{mit Wahrscheinlichkeit } p \\ 0 & \textit{mit Wahrscheinlichkeit } 1-p \end{cases}$$

Kompakt lässt sich die Wahrscheinlichkeitsfunktion $g(y_i) = p^{y_i}(1-p)^{1-y_i}$ schreiben. Wie groß ist der Erwartungswert?

$$\begin{aligned} E(y_i) &= 0 \cdot P(y_i = 0) + 1 \cdot P(y_i = 1) \\ &= 0 \cdot (p^0(1-p)^{1-0}) + 1 \cdot p^1(1-p)^{1-1} \\ &= p = P(y_i = 1) \end{aligned}$$

Der bedingte Erwartungswert ist dementsprechend $E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i)$. Da $P(y_i = 1|\mathbf{x}_i)$ eine Wahrscheinlichkeit ist liegt sie zwischen 0 und 1, was klar auf die Nichtlinearität zeigt. Es liegt nahe, $F(\boldsymbol{x}_i\boldsymbol{\beta})$ als eine Verteilungsfunktion zu spezifizieren, da $F(z) = P(Z < z) \in (0, 1)$. Somit

$$E(y_i|\mathbf{x}_i) = F(\mathbf{x}_i'\beta)$$

was den Erwartungswert der zuerklärenden Variable als die nichtlineare Funktion der Modellparameter darstellt.

2 Maximum Likelihood Schätzung

2.1 Definition und Erläuterung des Schätzers

Die Parameter eines stochastischen Modells müssen geschätzt werden. Hierzu braucht man Schätzprinzipien:

- klassisches lineares Regressionsmodell: KQ-Methode
- allgemeines nichtlineares Regressionsmodell: ML-Methode

Der Ausgangspunkt ist immer der gleiche:

a) Der Parametervektor ist unbekannt

b) Es liegt eine Stichprobe vor, die Information über den unbekannten Parametervektor in sich beinhaltet

Angenommen wir haben eine Zufallsvariable Y mit den Ausprägungen y und der Verteilungsfunktion $G(y;\theta)$

$$G(y;\theta) = \int_{x \leq y} g(x;\theta)\, dx,$$

wobei $g(x;\theta)$ ist die entsprechende Dichtefunktion und θ ist der Parametervektor, $\theta \in \mathbf{\Theta}$. Die Stichprobe mit Beobachtungsumfang n wird interpretiert als $Y_1, Y_2, \ldots, Y_n$, wobei für jede Zufallsvariable Y_i eine Realisierung beobachtet wird. Die gemeinsame Wahrscheinlichkeitsfunktion für die gesamte beobachtete Stichprobe, somit, ist $g(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n)$.

Annahmen ($i.i.d.$-Annahme)

- Unabhängigkeit: $P(A \cap B) = P(A) \cdot P(B)$

$$g(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) = g_1(y_1;\theta_1) \cdot g_2(y_2;\theta_2) \cdot \ldots \cdot g_n(y_n;\theta_n) = \prod_{i=1}^{n} g_i(y_i;\theta_i)$$

- Identische Verteilung: $g_1(\cdot;\theta_1) = g_2(\cdot;\theta_2) = \ldots = g_n(\cdot;\theta_n) = g(\cdot;\theta)$

Die gemeinsame Wahrscheinlichkeitsfunktion für die gesamte beobachtete Stichprobe unter i.i.d. Annahme ist dann

$$g(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} g(y_i; \theta).$$

Die Likelihoodfunktion $\mathcal{L}$ ist die gemeinsame Wahrscheinlichkeitsfunktion, die als die Funktion der Modellparameter gegeben die beobachtete Stichprobe definiert wird.

$$\mathcal{L}(\theta; y_1, \ldots, y_n) = g(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} g(y_i; \theta).$$

Definition 1 *Der ML Schätzer ist der Parametervektor $\hat{\theta}$, der die Likelihoodfunktion $\mathcal{L}(\theta; \mathbf{y})$ maximiert*

$$\hat{\theta} = \arg\max_{\theta \in \mathbf{\Theta}} \mathcal{L}(\theta; \mathbf{y}) \quad \Leftrightarrow \quad \left. \frac{\partial \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} \right|_{\hat{\theta}} = \mathbf{0}.$$

Da die Likelihoodfunktion bzgl. des unbekannten Parameters maximiert wird und die Bedingung 1. Ordnung Ableitungen erfordert, muss die Produktregel angewendet werden, denn $\mathcal{L}$ stellt ein Produkt dar. Um dies zu vereinfachen, wird die log-Likelihood Funktion verwendet. Die logarithmische Transformation ist zulässig, da $g(y_i) > 0$. Sie ändert nichts an der Lage des Optimums, und wir haben

$$\arg\max_{\theta \in \mathbf{\Theta}} \mathcal{L}(\theta; \mathbf{y}) = \arg\max_{\theta \in \mathbf{\Theta}} \ln \mathcal{L}(\theta; \mathbf{y}) = \hat{\theta}.$$

Die log-Likelihoodfunktion, somit, ist

$$\ln \mathcal{L}(\theta; \mathbf{y}) = \ln \left(\prod_{i=1}^{n} g(y_i; \theta) \right) = \sum_{i=1}^{n} \ln g(y_i; \theta).$$

Beispiel 2 (Bernoulli Verteilung) *Stichprobe: $Y_1, \ldots, Y_n$ i.i.d. mit $Y_i \sim B(p)$, d.h. es gibt nur zwei Ausprägungen*

$$Y_i = \begin{cases} 1, \text{ mit Wahrscheinlichkeit } p \\ 0, \text{ mit Wahrscheinlichkeit } 1-p \end{cases} \qquad i = 1, ..., n.$$

Die Dichte und die Likelihoodfunktion:

$$g(y_i) = p^{y_i} (1-p)^{1-y_i} \qquad \text{und} \qquad \mathcal{L}(p; \mathbf{y}) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}.$$

Die log-Likelihoodfunktion:

$$\ln \mathcal{L}(p; \mathbf{y}) = \sum_{i=1}^{n} \ln \left(p^{y_i} (1-p)^{1-y_i} \right) \quad \Leftrightarrow$$
$$\ln \mathcal{L}(p; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \ln p + (1-y_i) \ln (1-p) \right]$$

Bedingung erster Ordnung für log-Likelihood:

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial p} &= \frac{\partial}{\partial p} \left(\sum_{i=1}^{n} \left[y_i \ln p + (1-y_i) \ln (1-p) \right] \right) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial p} \left[y_i \ln p + (1-y_i) \ln (1-p) \right] \\ &= \sum_{i=1}^{n} \frac{y_i}{p} + \frac{1-y_i}{1-p}(-1) \stackrel{!}{=} 0 \end{aligned}$$

Daraus folgt, dass

$$\begin{aligned} \frac{1}{p} \sum_{i=1}^{n} y_i = \frac{1}{1-p} \sum_{i=1}^{n} (1-y_i) \quad &\Leftrightarrow \quad (1-p) \cdot n_1 = p(n-n_1) \quad | \, [n_1 = \sum_{i=1}^{n} y_i] \\ &\Leftrightarrow \quad n_1 - n_1 p = np - n_1 p \quad | + (n_1 p) \\ &\Leftrightarrow \quad n_1 = np \quad | : n \\ &\Leftrightarrow \quad \hat{p} = \frac{n_1}{n} \end{aligned}$$

Der ML-Schätzer für p ist die relative Häufigkeit der $y_i = 1$ Beobachtungen in der Stichprobe vom Umfang n.(!)

Der Beispiel oben erläutert das sogenannte *Maximum Likelihood Prinzip*. Das Maximum Likelihood Prinzip lautet:

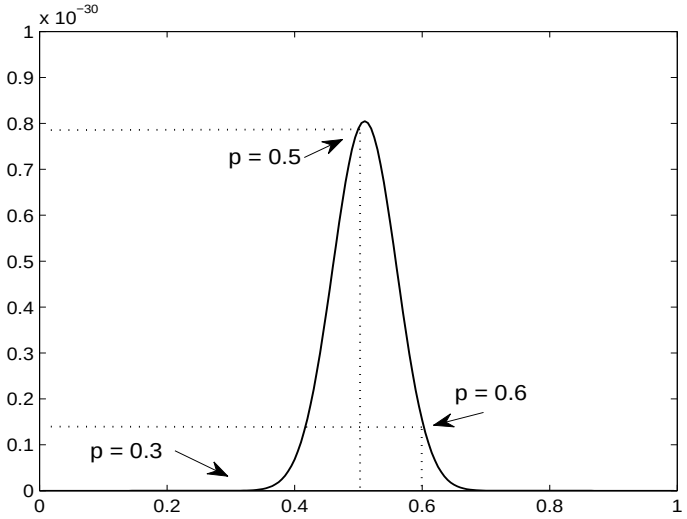
> Der Parametervektor, der die Wahrscheinlichkeit eine gegebene Stichprobe zu beobachten maximiert, ist der Schätzer des wahren unbekannten Parametervektors

Nun, stellen wir uns vor, wir wissen den wahren Parametervektor. Zum Beispiel, wirft man eine Münze, die Wahrscheinlichkeit einen "Kopf" zu beobachten ist 0.5, d.h. der wahre Parametervektor p_0 ist zu 0.5 gleich. Wenn $p_0 = 0.5$, wissen wir, daß die Vorschläge *$\acute{p}$=0.3* und *$\acute{p}$=0.6* klar daneben liegen werden. Wir wissen auch, daß die Anzahl an "Kopf" Ausprägungen der Anzahl an "Zahl" Ausprägungen ungefähr

gleich sein muss. Generieren wir 100 Zufälligen Ziehungen aus der Bernoulli Verteilung mit $p_0 = 0.5$. Wir bekommen 51 "Kopf" und 49 "Zahl" Beobachtungen. Die Likelihoodfunktion ist

$$\begin{aligned}\mathcal{L}(p;\mathbf{y}) &= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &= p^{5020}(1-p)^{4980}\end{aligned}$$

Die Abbildung unten zeigt die $\mathcal{L}(p;\mathbf{y})$ für verschieden Werte des Parameters p.



Die Likelihoodfunktion in *ṕ=0.3* und *p̋=0.6* liegt weit vom Maximum entfernt. Je näher kommt man an den Maximum der Likelihoodfunktion heran, desto näher kommt man an den wahren Parameter 0.5. Laut die Berechnungen oben, $\hat{p} = n_1/n = 0.51$. Der Unterschied 0.01 zwischen p und $\hat{p}$ wird durch die Kleinstichprobenverzerrung Verursacht. Je größer ist der Anzahl der Beobachtungen, desto genauer ist die Schätzung.

Beispiel 3 (Normalverteilung) *Stichprobe:* $Y_1, \ldots, Y_n$ *i.i.d. mit* $Y_i \sim N(\mu, \sigma^2)$. *Dichte:* $f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right\}$.

$$\begin{aligned}\mathcal{L} &= \mathcal{L}(\mu, \sigma; y_1, \ldots, y_n) \\ &= f(y_1, \ldots, y_n; \mu, \sigma) && \textit{gemeinsame Verteilung der Stichprobe, bzw.} \\ &= f(y_1) \cdot \ldots \cdot f(y_n) && \textit{Annahme einer Verteilung} \\ &= \textstyle\prod_{i=1}^{n} f(y_i) \\ &= \textstyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right\}\end{aligned}$$

Übergang zur Log-Likelihoodfunktion:

$$\begin{aligned}
\ln \mathcal{L} &= \ln \prod_{i=1}^{n} (\cdot) \\
&= \sum_{i=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\} \right) \\
&\quad \ln(a \cdot b) = \ln(a) + \ln(b) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi + (-1) \ln \sigma + \ln \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\} \right) \\
&\quad \ln(a^b) = b \cdot \ln(a) \\
&= -\frac{1}{2} n \ln 2\pi - n \ln \sigma + \sum_{i=1}^{n} \left(-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right)
\end{aligned}$$

Bedingungen 1.Ordnung:

$$\begin{aligned}
I. \quad \frac{\partial \ln \mathcal{L}}{\partial \mu} \stackrel{!}{=} 0 \quad &\Leftrightarrow \quad -\tfrac{1}{2} \textstyle\sum_{i=1}^{n} 2(\frac{y_i - \mu}{\sigma})(\frac{-1}{\sigma}) \stackrel{!}{=} 0 \\
&\Leftrightarrow \quad \tfrac{1}{\sigma^2} \textstyle\sum_{i=1}^{n} (y_i - \mu) = 0 \\
&\Leftrightarrow \quad \textstyle\sum_{i=1}^{n} y_i - n\mu = 0 \\
&\Leftrightarrow \quad \widehat{\mu} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} y_i \\
II. \quad \frac{\partial \ln \mathcal{L}}{\partial \sigma} \stackrel{!}{=} 0 \quad &\Leftrightarrow \quad -n\tfrac{1}{\sigma} + (-\tfrac{1}{2}) \textstyle\sum_{i=1}^{n} 2 \left(\frac{y_i - \mu}{\sigma} \right) \left(\frac{-(y_i - \mu)}{\sigma^2} \right) \stackrel{!}{=} 0 \\
&\Leftrightarrow \quad \tfrac{1}{\sigma^3} \cdot \textstyle\sum_{i=1}^{n} (y_i - \widehat{\mu})^2 = n\tfrac{1}{\sigma} \quad | \cdot \sigma^3 \\
&\Leftrightarrow \quad \textstyle\sum_{i=1}^{n} (y_i - \widehat{\mu})^2 = n\widehat{\sigma^2} \quad | : n \\
&\Leftrightarrow \quad \widehat{\sigma^2} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - \underbrace{\overline{y}}_{\widehat{\mu}})^2
\end{aligned}$$

Die ML-Schätzer für μ und σ^2 bei normalverteilten Zufallsvariablen sind also $\frac{1}{n} \sum_{i=1}^{n} y_i$ und $\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$.

Aus Statistik Basiskurs wissen wir, daß empirische Mittelwert $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ und empirische Varianz $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$ sind, und $E(\overline{y}) = \mu$ bzw., $E(s^2) = \sigma^2$, d.h. die empirische Mittelwert und Varianz sind die unverzerrte Schätzer des wahren Mittelwertes und der wahren Varianz. Offensichtlich ist der Erwartungswert von $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{n-1}{n} \cdot s^2$ nicht gleich σ^2:

$$E\left(\hat{\sigma}^2\right) = E(\underbrace{\frac{n-1}{n}}_{konstant} \cdot s^2) = \frac{n-1}{n} \cdot E(s^2) = \frac{n-1}{n} \cdot \sigma^2$$

Für große Stichproben allerdings, $n \to \infty$ und somit geht der Term $\frac{n-1}{n}$ gegen 1. Dis zeigt wiederum auf die günstige asymptotische Eigenschaften des Schätzers.

2.2 Theoretische Darstellung des Schätzers

Warum die Likelihood maximieren?

Definieren wir θ_0 als den wahren Parametervektor, $\theta_0 \in \mathbf{\Theta}$. Für alle andere Parametervektoren $\theta \in \mathbf{\Theta}$, $\theta \neq \theta_0$ betrachten wir den Erwartungswert $E\left(\mathcal{L}\left(\theta;\mathbf{y}\right)/\mathcal{L}\left(\theta_0;\mathbf{y}\right)\right)$, wobei die Erwartung über die wahre Verteilung der Daten gebildet wird. Laut der Definition des Erwartungswertes

$$\begin{aligned} E\left(\frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right) &= \underbrace{\int \cdots \int}_{n\ Integrale} \frac{\mathcal{L}\left(\theta;y_1,\ldots,y_n\right)}{\mathcal{L}\left(\theta_0;y_1,\ldots,y_n\right)} \mathcal{L}\left(\theta_0;y_1,\ldots,y_n\right)\, dy_1 \ldots dy_n \\ &= \int_{\mathbf{y}} \frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)} \mathcal{L}\left(\theta_0;\mathbf{y}\right) d\mathbf{y} \\ &= \int_{\mathbf{y}} \mathcal{L}\left(\theta;\mathbf{y}\right) d\mathbf{y} \\ &= 1. \end{aligned}$$

wobei die letzte gleichung ergibt sich dadurch, daß $\mathcal{L}\left(\theta;\mathbf{y}\right)$ eine Dichte ist. Somit: $E\left(\mathcal{L}\left(\theta;\mathbf{y}\right)/\mathcal{L}\left(\theta_0;\mathbf{y}\right)\right) = 1$. Gleichzeitig aber, laut Jensens Ungleichung,

$$E\left(\ln \frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right) < \ln E\left(\frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right),$$

denn $\ln\left(\cdot\right)$ ist eine konkave Funktion. Daraus folgt

$$\ln E\left(\frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right) > E\left(\ln \frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right) = E\left(\ln \mathcal{L}\left(\theta;\mathbf{y}\right)\right) - E\left(\ln \mathcal{L}\left(\theta_0;\mathbf{y}\right)\right).$$

Mit

$$\ln E\left(\frac{\mathcal{L}\left(\theta;\mathbf{y}\right)}{\mathcal{L}\left(\theta_0;\mathbf{y}\right)}\right) = \ln 1 = 0$$

auf der rechten Seite der Ungleichung bekommen wir schließlich

$$E\left(\ln \mathcal{L}\left(\theta;\mathbf{y}\right)\right) < E\left(\ln \mathcal{L}\left(\theta_0;\mathbf{y}\right)\right).$$

Dies zeigt uns, daß der Erwartungswert der Likelihoodfunktion ist nur dann maximiert, wenn der Argument dem wahren Parametervektor gleich ist.

Nun, wenn der Parametervektor die Likelihoodfunktion in einer Stichprobe maximiert, maximiert er auch deren Erwartungswert. In der kleinen Stichprobe ist der maximierte empirische Erwartungswert dem wahren nicht unbedingt gleich. Allerdings, in einer großen Stichprobe greift das Gesetz der großen Zahlen und somit für $\hat{\theta} = \underset{\theta \in \mathbf{\Theta}}{\arg\max}\, \mathcal{L}\left(\theta;\mathbf{y}\right)$ bekommen wir $E\left(\ln \mathcal{L}\left(\hat{\theta};\mathbf{y}\right)\right) \overset{n\to\infty}{\to} E\left(\ln \mathcal{L}\left(\theta_0;\mathbf{y}\right)\right)$ was schließlich zu $\hat{\theta} \overset{n\to\infty}{\to} \theta_0$ führt.

2.3 Eigenschaften des Maximum Likelihood Schätzers

Lassen wir weiterhin θ_0 den wahren Parametervektor bezeichnen. Der ML Schätzer hat die folgende Eigenschaften

1. Konsistenz (Erwartungstreue)

   $$\lim_{n\to\infty} P(|\hat{\theta} - \theta_0| > \varepsilon) = 0$$

   für eine beliebig kleine $\varepsilon > 0$. [äquivalent: $\text{plim}(\hat{\theta}) = \theta_0$, $\hat{\theta} \xrightarrow{p} \theta_0$].

2. Asymptotische Normalität

   $$\hat{\theta} \xrightarrow{d} N\left(\theta_0, \mathbf{I}\left(\theta_0\right)^{-1}\right),$$

   wobei $\mathbf{I}\left(\theta\right) = -E\left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta\partial\theta'}\right)$ ist die Informationsmatrix.

3. Effizienz

   $$n \to \infty: \quad Var(\hat{\theta}) \leq Var(\tilde{\theta}),$$

   wobei $\tilde{\theta}$ ein anderer erwartungstreue und asymptotisch normale Schätzer ist.

Die asymptotische Varianz des Schätzers

$$aVar(\hat{\theta}) = \mathbf{I}(\hat{\theta})^{-1} = \left[-E\left(\left.\frac{\partial^2 \ln \mathcal{L}\left(\theta\right)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}}\right)\right]^{-1}$$

erreicht geleichzeitig die sogenannte Cramér-Rao untere Grenze für die Varianz, die zu $\mathbf{I}\left(\theta\right)^{-1}$ gleich ist. Genau das impliziert, daß der ML-Schätzer die kleinstmögliche Varianz besitzt.

Da der Erwartungswert der Hesse-Matrix nicht immer so einfach zu ermitteln ist wie im Bernoulli Beispiel, weil die zweiten Ableitungen zum Beispiel keine linearen Funktionen der y_i sind, nimmt man die Hesse-Matrix direkt, um $aVar(\hat{\boldsymbol{\theta}})$ zu schätzen:

$$\widehat{aVar(\hat{\theta})} = -\left[\left.\frac{\partial^2 \ln \mathcal{L}\left(\theta\right)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}}\right]^{-1} = \left[-\sum_{i=1}^{n}\left.\frac{\partial^2 \ln \ell_i\left(\theta\right)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}}\right]^{-1}$$

wo $\ell_i\left(\theta\right)$ eine individuellen Beitrag zu der Likelihoodfunktion bezeichnet.

Veranschaulichung Konsistenz (Beispiel Normalverteilung)

$\hat{\mu} = \frac{1}{n}\sum y_i$	In diesem Fall $\hat{\mu} = \overline{y}$ und $\text{plim}(\overline{y}) = \mu$.
$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \overline{y})^2$	In diesem Fall $\lim_{n\to\infty} \hat{\sigma}^2 = \lim_{n\to\infty} \frac{n-1}{n} s^2 = s^2$ und $\text{plim}(s^2) = \sigma^2$.

Veranschaulichung asymptotischer Normalität (Beispiel Normalverteilung)

$\hat{\mu} = \frac{1}{n}\sum y_i$	In diesem Fall ist y_i NV, $\Rightarrow$ $\sum y_i$ ist auch NV
$\hat{\sigma}^2 = \frac{1}{n}\underbrace{\sum(y_i - \overline{y})^2}_{\text{Summe i.i.d.}}$	$(y_i - \overline{y})^2$ ist hier χ^2–verteilt, aber zentraler Grenzwertsatz greift: Summe i.i.d. Zufallsvariable (geeignet normiert) strebt gegen eine Normalverteilung)

Veranschaulichung Effizienz (Beispiel Bernoulli Verteilung)

$$\begin{aligned}
Var(\hat{p}) &= Var\left(\frac{n_1}{n}\right) \\
&= Var\left(\frac{1}{n}\sum y_i\right) \\
&= \frac{1}{n^2} Var\left(\sum y_i\right) \\
&= \frac{1}{n^2}\sum Var(y_i) \\
&= \frac{1}{n^2} np(1-p) \\
&= \frac{1}{n} p(1-p)
\end{aligned}$$

Cramér-Rao untere Grenze:

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}}{\partial p^2} &= \sum_i \left(\frac{-y_i}{p^2} + \frac{-(1-y_i)}{(1-p)^2}(-1)(-1)\right) \quad \Rightarrow \\
-E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial p^2}\right) &= -E\left(\sum_i \frac{-y_i}{p^2}\right) - E\left(\frac{\sum_i -(1-y_i)}{(1-p)^2}\right) \\
&= \frac{\sum_i E(y_i)}{p^2} + \frac{n - \sum_i E(y_i)}{(1-p)^2} \\
&= \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} \\
&= \frac{n(1-p) + np}{p(1-p)} \\
&= \frac{n}{p(1-p)}
\end{aligned}$$

Demnach ist die kleinstmögliche Varianz $\left[-E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial p^2}\right)\right]^{-1} = \frac{p(1-p)}{n}$, die der $Var(\hat{p})$ entspricht. D.h. $\hat{p}$ ist ein effizienter Schätzer.

2.4 Informationsgleichheit

Die Informationsgleichheit bietet uns einen äquivalenten ausdruck für $Var(\hat{\theta})$. Dies wird verwendet für die Herleitung ausgewählten Teststatistiken, sowie für Schätzung der Kovarianzmatrix in den fehlspezifizierten Modellen.

Die Likelihoodfunktion ist eine multivariate Dichtefunktion. Weil für Dichtefunktion gilt, dass die Fläche (im univariaten Fall) sich zu eins integriert, kann man schreiben:

$$\int_{\mathbf{y}} \mathcal{L}(\theta; \mathbf{y})\, d\mathbf{y} = 1$$

Leitet man beide Seiten dieser Gleichung nach θ ab, so erhält man

$$\frac{\partial}{\partial \theta}\left(\int_{\mathbf{y}} \mathcal{L}(\theta; \mathbf{y})\, d\mathbf{y}\right) = \frac{\partial 1}{\partial \theta} = \mathbf{0} \quad \Leftrightarrow \quad \int_{\mathbf{y}} \frac{\partial \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} d\mathbf{y} = \mathbf{0} \tag{1}$$

Da wir üblicherweise die Log-Likelihood betrachten, soll diese nach θ abgeleitet werden:

$$\frac{\partial \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{\mathcal{L}(\theta; \mathbf{y})} \frac{\partial \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} \quad \Leftrightarrow \quad \frac{\partial \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} = \frac{\partial \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} \mathcal{L}(\theta; \mathbf{y}) \tag{2}$$

Setzen wir dies in Gleichung (1) ein, erhalten wir

$$\begin{aligned} &\int_{\mathbf{y}} \frac{\partial \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} \mathcal{L}(\theta; \mathbf{y})\, d\mathbf{y} = \mathbf{0} \qquad (3) \\ \Leftrightarrow \quad &E\left(\frac{\partial \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta}\right) = \mathbf{0} \end{aligned}$$

Beim Übergang zur letzten Gleichung haben wir die aus der univariaten Statistik bekannte Beziehung $E(h(x)) = \int h(x) \cdot f(x) dx$ verwendet.

Nochmaliges Ableiten der Gleichung (3) nach θ' ergibt

$$\int_{\mathbf{y}} \frac{\partial^2 \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta \partial \theta'} \mathcal{L}(\theta; \mathbf{y})\, d\mathbf{y} + \int_{\mathbf{y}} \frac{\partial \ln \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} \frac{\partial \mathcal{L}(\theta; \mathbf{y})}{\partial \theta'} d\mathbf{y} = \mathbf{0}$$

In diesem Ausdruck setzen wir erneut die Beziehung (1) ein und verwenden wieder die Erwartungswertschreibweise:

$$\int_{\mathbf{y}} \frac{\partial^2 \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'} \mathcal{L}(\theta;\mathbf{y})\, d\mathbf{y} + \int_{\mathbf{y}} \frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta} \frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta'} \mathcal{L}(\theta;\mathbf{y})\, d\mathbf{y} = \mathbf{0}$$

$$\Leftrightarrow$$

$$\underbrace{-E\left(\frac{\partial^2 \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right)}_{=\mathbf{I}(\theta)} = E\left(\frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta} \frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta'}\right)$$

Die Informationsgleichheit zeigt uns, daß wenn das Varianz des ML-Schätzers auf zwei verschieden Arten dargestellt werden kann. Dies hält, allerdings nur wenn das Modell richtig spezifiziert ist, d.h. die wahre Dichte $\mathcal{L}(\theta;\mathbf{y})$ ist bekannt. Diese Tatsache bildet auch eine Grundlage für die Spezifikationstest: Werden die beide Matrizen statistisch voneinander unterscheiden, dann ist das Modell fehlspezifiziert.

2.5 Bedingungen für die Schätzbarkeit

Abgesehen von der zentralen Annahme der unabhängigen und identisch verteilten Zufallsvariablen (*i.i.d.*), sollen weitere Bedingungen gelten, damit der ML-Schätzer erwartungstreu, effizient und asymptotisch normalverteilt bleibt.

- *Identifizierbarkeit*

Definition 2 *Der Parametervektor $\theta \in \mathbf{\Theta}$ ist aus der Dateninformation $Y_1, \ldots, Y_n$ identifizierbar wenn es kein andere Parametervektor $\theta^* \in \mathbf{\Theta}$, $\theta^* \neq \theta$, existiert, so daß $\mathcal{L}(\theta^*;\mathbf{y}) = \mathcal{L}(\theta;\mathbf{y})$ ist.*

Diese Definition sagt, daß die Schätzung nur dann möglich ist, wenn die vorhandene Dateninformation ausreichend ist um alle Parameter eindeutig zu bestimmen.

Theorem 1 (Rothenberg, 1971) *Der Parametervektor $\theta_0 \in \mathbf{\Theta}$ ist identifizierbar wenn die Matrix*

$$\mathbf{B}(\theta) = E\left(\frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta} \frac{\partial \ln \mathcal{L}(\theta;\mathbf{y})}{\partial\theta'}\right)$$

einen vollen Rang hat.

Über Informationsgleichheit impliziert Theorem 1 weiterhin, daß die Informationsmatrix auch einen vollen Rang haben muss. Dadurch, im identifizierbaren Modell hat der ML Schätzer immer eine endliche Varianz.

Theorem 1 ist auch eine Verallgemeinerung der bekannten Bedingung der KQ-Schätzung, wo die Matrix $(\mathbf{X}'\mathbf{X})$ invertierbar sein muss um den Schätzer berechnen zu können.

- *Regularitätsbedingungen*

1. (White, 1982): Das wahre Modell, d.h. die parametrische Form der wahren Dichtefunktion, bzw. Likelihoodfunktion, ist bekannt.

Sei es nicht der Fall, gehen Effizienz und u.U. die Konsistenz verloren. Immerhin, bleibt in diesem Fall die asymptotische Normalität erhalten. Falls Konsistenz dazu erhalten bleibt,

$$\hat{\theta} \xrightarrow{d} N\left(\theta_0, \left[\mathbf{A}\left(\theta_0\right)\right]^{-1}\mathbf{B}\left(\theta_0\right)\left[\mathbf{A}\left(\theta_0\right)\right]^{-1}\right), \quad \text{wobei:}$$

$$\mathbf{A}\left(\theta\right) = -E\left(\frac{\partial^2 \ln \mathcal{L}\left(\theta;\mathbf{y}\right)}{\partial\theta\partial\theta'}\right) \quad \text{und} \quad \mathbf{B}\left(\theta\right) = E\left(\frac{\partial \ln \mathcal{L}\left(\theta;\mathbf{y}\right)}{\partial\theta}\frac{\partial \ln \mathcal{L}\left(\theta;\mathbf{y}\right)}{\partial\theta'}\right).$$

Die Informationsgleichcheit gilt nicht mehr.

2. Der wahre Parametervektor θ_0 liegt im inneren des Parameterraums $\mathbf{\Theta}$.

Sei es nicht der Fall, ist der Schätzer immer noch Konsistent. Verliert jedoch Effizienz und asymptotische Normalität.

3. Die Gernzen der Datenverteilung hängen von dem Parametervektor θ nicht ab.

Sei es nicht der Fall, gehen alle Eigenschaften verloren.

Literatur

◦ Cameron, C., and P., Trivedi, "Microeconometrics: Methods and applications", (Cambridge University Press: 2005), Ch.5.6., p.139-146.

◦ Greene, W., "Econometric analysis", (Prentice Hall: 2003), 4th Ed., Ch.17.1-17.4., p.468-483.

∗ Rothenberg, T., "Identification in parametric models", *Econometrica*, 1971, Vol. 39(3), p.577-591.

∗ White, H., "Maximum likelihood estimation of misspecified models", *Econometrica*, 1982, Vol. 50(1), p.1-25.

3 Testen von Hypothesen

Im wesentlichen werden wir drei asymptotisch äquivalente Testverfahren zur Überprüfung von Parameterrestriktionen verwenden:

- Likelihood-Quotienten-Test
- Wald-Test
- Lagrange-Multiplikator-Test

Die drei Tests sollen zunächst in ihrer Grundidee dargestellt werden und formal hergeleitet werden.

Likelihood Ratio(LR) - Test

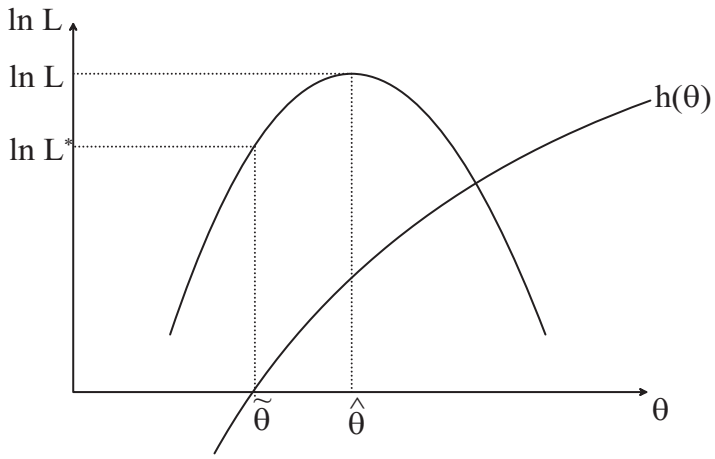


Abbildung 1: Likelihood Ratio-Test

Das Maximum der $\ln \mathcal{L}$-Funktion wird im Punkt $\hat{\theta}$ erreicht. Kommt eine Nebenbedingung der Form $h(\theta) = 0$ hinzu, muss $\ln \mathcal{L}$ unter dieser Nebenbedingung maximiert werden (Lagrangeverfahren). Der Wert $\tilde{\theta}$ maximiert die Lagrangefunktion, d.h. ist der maximale Wert der $\ln \mathcal{L}$ unter Berücksichtigung der Nebenbedingung. Der Quotient $\frac{\mathcal{L}^*}{\mathcal{L}}$ ist somit kleiner als 1 und damit $\ln \frac{\mathcal{L}^*}{\mathcal{L}} < 0$:

$$\begin{aligned} &\Leftrightarrow \quad \ln \mathcal{L}^* - \ln \mathcal{L} < 0 \quad | \cdot (-2) \\ &\Leftrightarrow \quad -2(\ln \mathcal{L}^* - \ln \mathcal{L}) > 0 \end{aligned}$$

Man benötigt hierzu den unrestringierten Schätzer $\hat{\theta}$ und den restringierten Schätzer $\tilde{\theta}$. Getestet wird, ob der Quotient $\frac{\mathcal{L}^*}{\mathcal{L}}$ signifikant kleiner als 1 ist, bzw. ob die

Differenz $-2(\ln \mathcal{L}^* - \ln \mathcal{L})$ signifikant größer Null ist.

Wald-Test

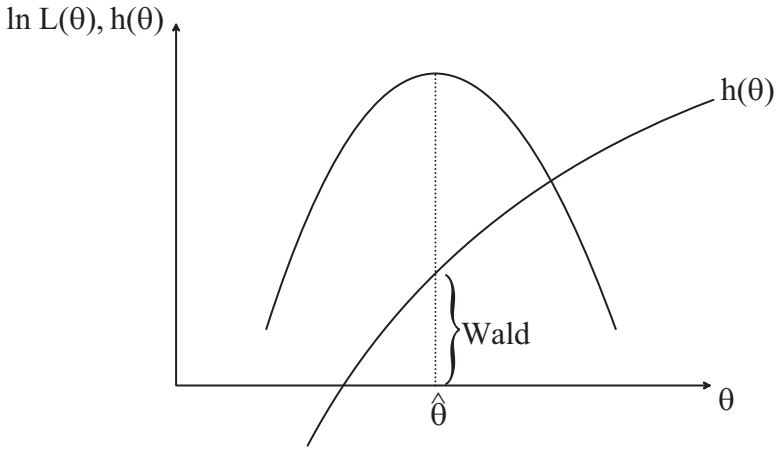


Abbildung 2: Wald-Test

Die Idee hier ist, den Wert $\hat{\theta}$ der unrestringierten Schätzung in die Nebenbedingung einzusetzen und zu bewerten, ob $h(\hat{\theta})$ signifikant von Null verschieden ist. Man benötigt hier nur die unrestringierte Schätzung $\hat{\theta}$.

Lagrange-Multiplikator-Test(LM)

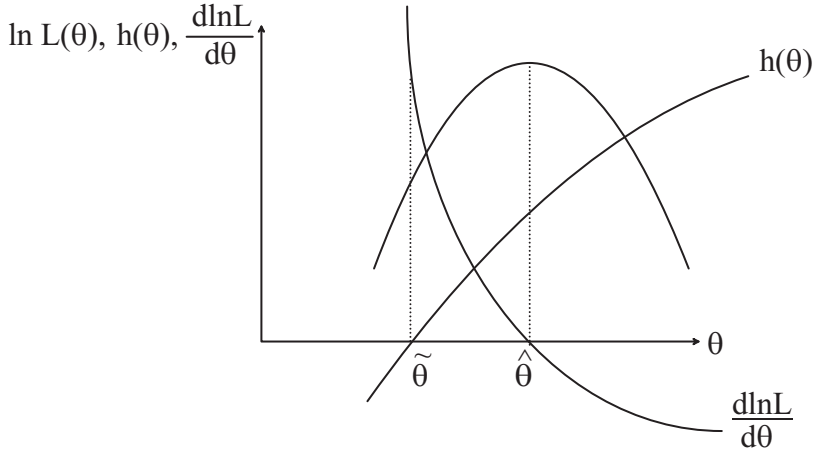


Abbildung 3: Lagrange-Multiplikator-Test

Der unrestringierte Schätzer $\hat{\theta}$ erfüllt die Bedingung erster Ordnung $\frac{d \ln \mathcal{L}}{d\theta} = 0$. Deshalb wird hier der restringierte Schätzer $\tilde{\theta}$ in die erste Ableitungsfunktion eingesetzt und getestet, ob $\frac{d \ln \mathcal{L}}{d\theta}\big|_{\tilde{\theta}}$ signifikant von Null verschieden ist.

Um die Teststatistiken und deren Verteilung zu verstehen, sollen hier zwei folgende Resultate der mathematischen Statistik ausgeführt werden.

Theorem 2 *Sei* $\underset{(k\times 1)}{\mathbf{z}} \sim N(\underset{(k\times 1)}{\mu}, \underset{(k\times k)}{\mathbf{\Sigma}})$. *Dann ist* $(\mathbf{z}-\mu)' \mathbf{\Sigma}^{-1} (\mathbf{z}-\mu) \sim \chi^2_k$.

Theorem 3 (Delta Theorem) *Ist* $\underset{(k\times 1)}{(\mathbf{z}-\mu)} \sim N(\mathbf{0}, \mathbf{\Sigma})$ *und* $\underset{(r\times 1)}{\mathbf{h}} = h(\mathbf{z})$, $r < k$, *dann ist*

$$(h(\mathbf{z}) - h(\mu)) \overset{a}{\sim} N(\mathbf{0}, \mathbf{D\Sigma D}'),$$

wobei

$$\mathbf{D} = \nabla h(\mathbf{z}) = \begin{bmatrix} \frac{\partial h_1}{\partial z_1} & \cdots & \frac{\partial h_1}{\partial z_k} \\ \cdots & & \cdots \\ \frac{\partial h_r}{\partial z_1} & \cdots & \frac{\partial h_r}{\partial z_k} \end{bmatrix}, \quad D_{ij} = \frac{\partial h_i}{\partial z_j}.$$

Diese Resultate verwenden wir zur Herleitung der Wald- und LM-Tests, sowie deren asymptotischer Testverteilungen.

- Herleitung des Wald-Tests

Weil $\underset{(k\times 1)}{\hat{\theta}} \overset{a}{\sim} N\left(\theta, \mathbf{I}(\theta)^{-1}\right)$, ist $\underset{(r\times 1)}{h(\hat{\theta})} \sim N\left(h(\theta), \mathbf{DI}(\theta)^{-1}\mathbf{D}'\right)$, wobei $\underset{(r\times k)}{\mathbf{D}} = \nabla h(\theta)$.

Für $h(\hat{\theta})$ verwenden wir Theorem 2 um zu sehen, daß

$$(h(\hat{\theta}) - h(\theta))' \left[\mathbf{DI}(\theta)^{-1}\mathbf{D}'\right]^{-1} (h(\hat{\theta}) - h(\theta)) \overset{a}{\sim} \chi^2_r$$

ist. Unter Nullhypothese $H_0 : h(\theta) = \mathbf{0}$ gilt somit

$$W = h(\hat{\theta})' \left[\mathbf{DI}(\theta)^{-1}\mathbf{D}'\right]^{-1} h(\hat{\theta}) \overset{a}{\sim} \chi^2_r.$$

Da die Informationsmatrix, bzw. die $\mathbf{D}$-Matrix unbekannt sind und müssen geschätzt werden, die Testgröße der Waldstatistik unter H_0 ist letztlich

$$W = h(\hat{\theta})' \left[\mathbf{D}(\hat{\theta})\mathbf{I}(\hat{\theta})^{-1}\mathbf{D}(\hat{\theta})'\right]^{-1} h(\hat{\theta}) \overset{a}{\sim} \chi^2_r.$$

- Herleitung des Lagrange-Multiplikator-Tests

Für den restringierten Schätzer gilt:

$$\max_{\theta,\lambda} \left\{ \ln \mathcal{L}\left(\theta; \mathbf{y}\right) + \lambda' h\left(\theta\right) \right\}$$

B.E.O.:

$$\begin{aligned} a) \quad &: \quad \left. \frac{\partial \ln \mathcal{L}\left(\theta; \mathbf{y}\right)}{\partial \theta} \right|_{\tilde{\theta}} + \left[\nabla h\left(\theta\right)|_{\tilde{\theta}} \right]' \lambda = \mathbf{0} \\ b) \quad &: \quad h(\tilde{\theta}) = \mathbf{0}. \end{aligned}$$

Als Nullhypothese beim LM-Test formuliert man $H_0 : \lambda = \mathbf{0}$. In diesem Fall ergibt sich aus der B.E.O., daß unter H_0 auch

$$\underbrace{\left. \frac{\partial \ln \mathcal{L}\left(\theta; \mathbf{y}\right)}{\partial \theta} \right|_{\tilde{\theta}}}_{\equiv S(\tilde{\theta})} = \mathbf{0}$$

gelten muss, wobei $S(\tilde{\theta})$ eine "Score-Vektor" ist. Dadurch ergeben sich zwei Wege um den Teststatistik hinzubekommen

a) Den Multiplikator λ direkt zu betrachten $\Rightarrow$ $\lambda' \left[Var\left(\lambda\right) \right]^{-1} \lambda \sim \chi^2_r$ [der Ausdruck für $Var\left(\lambda\right)$ ist ziemlich komplex].

b) Die Verteilung des Score-Vektors zu berücksichtigen [der einfachere Weg]

$$\text{Aus B.E.O.:} \quad S(\tilde{\theta}) = -\left[\nabla h\left(\theta\right)|_{\tilde{\theta}} \right]' \lambda \quad \Rightarrow \quad E\left(S(\tilde{\theta})\right) = E\left(-\frac{\partial \left[\lambda' h\left(\theta\right) \right]}{\partial \theta} \right) \overset{H_0:\lambda=\mathbf{0}}{=} \mathbf{0},$$

Somit, unter Nullhypothese $E\left(S(\tilde{\theta})\right) = \mathbf{0}$. Wie groß ist dann die Varianz des Scorevektors?

$$\begin{aligned} Var\left(S(\tilde{\theta})\right) &= E\left(S(\tilde{\theta})S(\tilde{\theta})'\right) - E\underbrace{\left(S(\tilde{\theta})\right)}_{=\mathbf{0}} E(\underbrace{S(\tilde{\theta})'}_{=\mathbf{0}'}) = E\left(S(\tilde{\theta})S(\tilde{\theta})'\right) \\ &= E\left(\left. \frac{\partial \ln \mathcal{L}\left(\theta; \mathbf{y}\right)}{\partial \theta} \right|_{\tilde{\theta}} \left. \frac{\partial \ln \mathcal{L}\left(\theta; \mathbf{y}\right)}{\partial \theta'} \right|_{\tilde{\theta}} \right) \overset{\text{Inf.Gleich.}}{=} -E\left(\left. \frac{\partial^2 \ln \mathcal{L}\left(\theta; \mathbf{y}\right)}{\partial \theta \partial \theta'} \right|_{\tilde{\theta}} \right) \\ &= \mathbf{I}(\tilde{\theta}). \end{aligned}$$

Somit, $E\left(S(\tilde{\theta})\right) = \mathbf{0}$ und $Var\left(S(\tilde{\theta})\right) = \mathbf{I}(\tilde{\theta})$. Da $\tilde{\theta} \overset{a}{\sim} N\left(\theta, \mathbf{I}(\theta)^{-1}\right)$ ist, zeigt die Anwendung der Theorem 3, daß $S(\tilde{\theta})$ wieder Normalverteilt ist, nämlich

$S(\tilde{\theta}) \overset{a}{\sim} N\left(E\left(S(\tilde{\theta})\right), Var\left(S(\tilde{\theta})\right)\right)$, was unter Nullhypothese zu $S(\tilde{\theta}) \overset{a}{\sim} N\left(\mathbf{0}, \mathbf{I}(\tilde{\theta})\right)$ führt. Mit diesem Ergebnis und Theorem 2 haben wir schließlich

$$LM = S(\tilde{\theta})' \mathbf{I}(\tilde{\theta})^{-1} S(\tilde{\theta}) \sim \chi^2_r$$

Die Anzahl der Freiheitsgrade dabei ergibt sich über die Überlegung, dass unter $H_0 : S(\tilde{\theta}) = -\left[\nabla h\left(\theta\right)|_{\tilde{\theta}}\right]' \lambda$ die Matrix $\nabla h\left(\theta\right)$ die Dimension $(r \times k)$ hat (dies ist zu Theorem 3 unterschiedlich).

- Verteilung der LR-Statistik

Wie schon angesprochen, die LR-Statistik lautet $-2(\ln \mathcal{L}^* - \ln \mathcal{L})$. Diese Statistik folgt asymptotisch eine χ^2 Verteilung

$$LR = -2\left(\ln \mathcal{L}^* - \ln \mathcal{L}\right) \sim \chi^2_r,$$

wobei Anzahl der Freiheitsgraden in dieser Verteilung der Anzahl der Restriktionen gleich ist. Dies kommt ohne Beweis (der Beweis existiert, natürlich).

Literatur

- Cameron, C., and P., Trivedi, "Microeconometrics: Methods and applications", (Cambridge University Press: 2005), Ch.7.3., p.233-239.
- Greene, W., "Econometric analysis", (Prentice Hall: 2003), 4th Ed., Ch.17.5., p.484-490.

Teil II

Modelle und Anwendungen

4 Modelle für qualitativ abhängige Variablen

Im Beispiel 1, Abschnitt 1 haben wir bereits im Ansatz die Grundzüge für solche y_i gelegt, die nur die Werte 0 und 1 annehmen können. Wir nennen eine solche Variable *dichotom*. Wir werden weiterhin auch die *polytome* Variablen betrachten, d.h. y_i kann r verschiedene Ausprägungen annehmen, wobei diese Ausprägungen beispielsweise mit $1, 2, ..., r$ kodiert sein können; andere Kodierungen sind ebenfalls denkbar. Im polytomen Fall muss man unterscheiden, ob die r Kategorien geordnet oder ungeordnet sind. Ein Beispiel für geordnete Kategorien sind Klausurnoten $y = 1, y = 2, ..., y = 5$; die Kategorien haben eine Ordnung: 1 ist besser als 2, 2 ist besser als 3 usw. Ein Beispiel für ungeordnete Kategorien sind Automarken. Die r verschiedenen Automarken, z.B. "Volkswagen", "Opel", "Audi", "BMW" etc. lassen sich zwar auch zu '1', '2', '3', '4' etc. kodieren, aber hier gibt es keine natürliche Ordnung. Bei dichotomen Variablen besteht dieser Unterschied zwar auch, aber die Modellierungen sind gleich.

4.1 Binäre Modelle

4.1.1 Darstellung des binären Modells

Ein binäres Modell beschreibt das Vorkommnis eines Ereignisses, was auch immer dieses Ereignis sein kann (Beschäftigt / Arbeitslos, Uni-Abschluß / Kein Uni-Abschluß, Auto kaufen / Fahrrad anstatt dessen usw.). Die abhängige Variable y_i hat somit nur zwei Ausprägungen

$$y_i = \begin{cases} 1 & \text{wenn Ereignis tritt ein} \\ 0 & \text{wenn Ereignis tritt nicht ein} \end{cases}.$$

Die Ausprägungen y_i stimmen aus einer Bernoulli Verteilung, nämlich

$$y_i = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } \ p \\ 0 & \text{mit Wahrscheinlichkeit } \ 1-p \end{cases}.$$

Somit, die Wahrscheinlichkeitsdichte für die beobachtete Ausprägung ist

$$g(p) = p^{y_i} (1-p)^{1-y_i},$$

wobei $0 < p < 1$, weil p eine Wahrscheinlichkeit ist. Demzufolge die Likelihoodfunktion für die Strichprobe mit n Beobachtungen ist

$$\mathcal{L}(p) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}.$$

Dieses Modell ist "unbedingt", d.h. hängt von der individuellen Charakteristiken nicht ab. Da wir uns für den Einfluß einer oder mehreren Charakteristiken auf die Wahrscheinlichkeit des Ereignisses interessieren, schlagen wir weiter vor, daß p sich von Individuum zu Individuum unterscheidet, wobei die unterschiede durch die individuellen Charakteristiken verursacht werden, d.h.

$$p_i = P(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\beta),$$

wobei $F(\cdot)$ die kumulative Dichtefunktion einer beliebigen Verteilung bezeichnet. Somit bekommen wir endgültig das bedingte binäre Modell

$$\mathcal{L} = \prod_{i=1}^{n} [F(\mathbf{x}_i'\beta)]^{y_i} [1 - F(\mathbf{x}_i'\beta)]^{1-y_i}.$$

- Ökonomischer Inhalt des Modells

Es gibt unterschiedliche Wege das binäre Modell zu interpretieren.

a) Allgemeiner Zusammenhang

Die zuerklärende Variable y_i hängt auf einer unbekannten Art und Weise von den individuellen Merkmalen $\mathbf{x}_i$ ab.

Beispiel: Entscheidung während des Schulalters das Abitur zu machen [soziale + ökonomische Faktoren].

b) Latente Variable und Indexfunktion

Man geht von einer kontinuierlichen latenten Variablen y_i^* aus, deren bedingter Erwartungswert linear modelliert werden soll

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i,$$

wobei $E(\varepsilon_i) = 0$ und ε_i eine bekannte symmetrische Verteilung $F(\varepsilon)$ folgt. Die latente Variable ist unbeobachtbar. Anstatt dessen beobachten wir nur y_i , und zwar: wenn die latente Variable y_i^* eine bestimmte Schwelle überschreitet, beobachten wir $y_i = 1$; ansonsten beobachten wir $y_i = 0$. In diesem Fall nennt man y_i auch die "Indexfunktion".

Beispiel: Unbeobachtbare Produktivität des Arbeitnehmers und Einstellungsentscheidung der Firma während der Probezeit.

Da $\mathbf{x}_i$ in sich auch einen Absolutglied beinhaltet, ohne Verlust der Allgemeinheit, setzt man die Schwelle zu Null gleich und somit

$$y_i = \begin{cases} 1 & \text{wenn } y_i^* > 0 \\ 0 & \text{wenn } y_i^* \leq 0 \end{cases}$$

$$\begin{aligned} \text{Dann:} \quad P(y_i = 1|\mathbf{x}_i) &= P(y_i^* > 0|\mathbf{x}_i) \\ &= P(\mathbf{x}_i'\beta + \varepsilon_i > 0) \\ &= P(\varepsilon_i > -\mathbf{x}_i'\beta) \\ &= 1 - P(\varepsilon_i \leq -\mathbf{x}_i'\beta) \\ &= 1 - F(-\mathbf{x}_i'\beta) \\ &= F(\mathbf{x}_i'\beta), \end{aligned}$$

wobei der letzte Schritt ist wegen der Symmetrie der Verteilung F.

Somit: $P(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\beta)$, was zu $P(y_i = 0|\mathbf{x}_i) = 1 - F(\mathbf{x}_i'\beta)$ führt, und die Likelihoodfunktion für die beobachtbare Indexvariable y_i ist nichts anderes als

$$\mathcal{L} = \prod_{i=1}^{n} \left[F(\mathbf{x}_i'\beta)\right]^{y_i} \left[1 - F(\mathbf{x}_i'\beta)\right]^{1-y_i}.$$

c) Unbeobachtbare Nutzenunterschiede ("random utility model")

Es gibt zwei Produkte: A und B. Der Nutzen U_A und U_B sind nicht direkt beobachtbar. Wir wissen, jedoch, daß man A nur dann den Vorzug gibt, wenn $U_A > U_B$ (und umgekehrt). Wir beobachten $y_i = 1$ wenn A gewählt wird, und $y_i = 0$, wenn B gewählt wird.

Beispiel: Apfel oder Banane? (kann auch für mehrere Wahlmöglichkeiten erweitert werden)

Wir nehmen an, das die Nutzen von beiden Produkten sich durch die folgende Gleichungen beschreiben lassen

$$\begin{aligned} U_{A,i} &= \mathbf{x}_i'\beta_A + \varepsilon_{A,i}, \qquad E(\varepsilon_{A,i}) = 0,\ \varepsilon_{A,i} \sim F_A \\ U_{B,i} &= \mathbf{x}_i'\beta_B + \varepsilon_{B,i}, \qquad E(\varepsilon_{B,i}) = 0,\ \varepsilon_{B,i} \sim F_B \end{aligned}$$

wobei F_A und F_B die bekannte symmetrische Verteilungen sind. Die Wahrscheinlichkeit A über B zu wählen ist dann

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(U_{A,i} > U_{B,i}|\mathbf{x}_i) = P(U_{A,i} > U_{B,i}|\mathbf{x}_i) \\ &= P(\mathbf{x}_i'\beta_A + \varepsilon_{A,i} > \mathbf{x}_i'\beta_B + \varepsilon_{B,i}) = P(\varepsilon_{A,i} - \varepsilon_{B,i} > -\mathbf{x}_i'\beta_A + \mathbf{x}_i'\beta_B) \\ &= 1 - P\left(\underbrace{\varepsilon_{A,i} - \varepsilon_{B,i}}_{\equiv\varepsilon_i} \leq -\mathbf{x}_i'\underbrace{(\beta_A - \beta_B)}_{\equiv\beta}\right) = 1 - P(\varepsilon_i \leq -\mathbf{x}_i'\beta), \end{aligned}$$

und, wenn wir die Verteilung des Unterschiedes $\varepsilon_i = \varepsilon_{A,i} - \varepsilon_{B,i}$ als F definieren,

$$P(y_i = 1|\mathbf{x}_i) = 1 - P(\varepsilon_i \leq -\mathbf{x}_i'\beta) = 1 - F(-\mathbf{x}_i'\beta) \overset{\text{symm.}}{=} F(\mathbf{x}_i'\beta).$$

Da wir wieder die $P(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\beta)$ haben, ist die Wahrscheinlichkeitsmodell für die Auswahl von A und B wieder

$$\mathcal{L} = \prod_{i=1}^{n} \left[F(\mathbf{x}_i'\beta)\right]^{y_i} \left[1 - F(\mathbf{x}_i'\beta)\right]^{1-y_i}.$$

4.1.2 Binäres Logit-Modell

Für $F(\cdot)$ bietet sich an, eine Verteilungsfunktion zu nehmen. In diesem Abschnitt verwenden wir die *logistische* Verteilung, was sich auch im Namen des Modells wiederspiegelt. Ganz allgemein ist eine Zufallsvariable logistisch verteilt, wenn sie, jeweils, folgenden Dichtefunktion und kumulative Dichtefunktion besitzt

$$\lambda(z) = \frac{1}{\gamma} \frac{\exp\left\{-\frac{z-\delta}{\gamma}\right\}}{[1 + \exp\left\{-\frac{z-\delta}{\gamma}\right\}]^2}, \qquad \Lambda(z) = \int_{-\infty}^{z} \lambda(x)dx = \frac{1}{1 + \exp\left\{-\frac{z-\delta}{\gamma}\right\}}.$$

Die Zufallsvariable Z hat $E(Z) = \delta$ und $Var(Z) = \frac{\gamma^2\pi^2}{3}$. Wir setzen für unsere Zwecke $\delta = 0$ und $\gamma = 1$. Somit

$$\lambda(z) = \frac{\exp\{-z\}}{[1 + \exp\{-z\}]^2} \qquad \text{und} \qquad \Lambda(z) = \frac{1}{1 + \exp\{-z\}}.$$

- Die Likelihoodfunktion

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left[\Lambda\left(\mathbf{x}_i'\beta\right)\right]^{y_i} \left[1-\Lambda\left(\mathbf{x}_i'\beta\right)\right]^{1-y_i} \quad \Leftrightarrow \\
\mathcal{L} &= \prod_{i=1}^{n} \left[\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right]^{y_i} \left[1-\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right]^{1-y_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right]^{y_i} \left[\frac{\exp\left\{-\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right]^{1-y_i} \\
&= \prod_{i=1}^{n} \frac{1}{\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)^{y_i}} \frac{\left(\exp\left\{-\mathbf{x}_i'\beta\right\}\right)^{1-y_i}}{\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)^{1-y_i}} \\
&= \prod_{i=1}^{n} \frac{\exp\left\{-\left(1-y_i\right)\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}} = \frac{\prod_{i=1}^{n} \exp\left\{-\left(1-y_i\right)\mathbf{x}_i'\beta\right\}}{\prod_{i=1}^{n}\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)}
\end{aligned}$$

und, mit $e^a e^b = e^{a+b}$, schließlich: $\mathcal{L} = \dfrac{\exp\left\{-\sum_{i=1}^{n}\left(1-y_i\right)\mathbf{x}_i'\beta\right\}}{\prod_{i=1}^{n}\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)}$.

- Die log-Likelihoodfunktion und B.E.O.

$$\ln\mathcal{L} = -\sum_{i=1}^{n}\left(1-y_i\right)\mathbf{x}_i'\beta - \ln\left(\prod_{i=1}^{n}\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)\right),$$

und schließlich

$$\ln\mathcal{L} = -\sum_{i=1}^{n}\left[\left(1-y_i\right)\mathbf{x}_i'\beta + \ln\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)\right].$$

Daraus ergeben sich die B.E.O.

$$\begin{aligned}
\frac{\partial\ln\mathcal{L}}{\partial\beta} &= -\sum_{i=1}^{n}\left[\left(1-y_i\right)\frac{\partial}{\partial\beta}\left(\mathbf{x}_i'\beta\right) + \frac{\partial}{\partial\beta}\ln\left(1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right)\right] \\
&= -\sum_{i=1}^{n}\left[\left(1-y_i\right)\mathbf{x}_i + \frac{\exp\left\{-\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\left(-\mathbf{x}_i\right)\right] \\
&= -\sum_{i=1}^{n}\left[1-y_i-\frac{\exp\left\{-\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right]\mathbf{x}_i = -\sum_{i=1}^{n}\left[\underbrace{\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}}_{=\Lambda_i} - y_i\right]\mathbf{x}_i \\
&= \sum_{i=1}^{n}\left[y_i-\Lambda_i\right]\mathbf{x}_i,
\end{aligned}$$

wobei wir die Λ_i anstatt $\Lambda\left(\mathbf{x}_i'\beta\right)$ schreiben um die Darstellung zu vereinfachen.

Die Hesse-Matrix dann ist

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta'}\left(\sum_{i=1}^{n}\left[y_i-\Lambda_i\right]\mathbf{x}_i\right) \\
&= -\sum_{i=1}^{n}\frac{\partial\Lambda_i}{\partial\beta'}\mathbf{x}_i = -\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}\left(\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}\right)\mathbf{x}_i \\
&= -\sum_{i=1}^{n}\left(-\frac{\exp\left\{-\mathbf{x}_i'\beta\right\}(-1)}{\left[1+\exp\left\{-\mathbf{x}_i'\beta\right\}\right]^2}\right)\mathbf{x}_i\mathbf{x}_i' \\
&= -\sum_{i=1}^{n}\underbrace{\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}}_{=\Lambda_i}\underbrace{\frac{\exp\left\{-\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}}_{=1-\Lambda_i}\mathbf{x}_i\mathbf{x}_i' \\
&= -\sum_{i=1}^{n}\Lambda_i\left(1-\Lambda_i\right)\mathbf{x}_i\mathbf{x}_i'.
\end{aligned}$$

Da die Hesse-Matrix in sich keine y_i beinhaltet, die Informationsmatrix

$$\mathbf{I}\left(\beta\right) = -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'}\right) = -E\left(-\sum_{i=1}^{n}\Lambda_i\left(1-\Lambda_i\right)\mathbf{x}_i\mathbf{x}_i'\right) = \sum_{i=1}^{n}\Lambda_i\left(1-\Lambda_i\right)\mathbf{x}_i\mathbf{x}_i'$$

ist der negativen Hessematrix gleich.

- Eindeutigkeit des Schätzers und Identifizierbarkeitsrestriktionen

Im Matrix Form lässt sich die Hesse-Matrix als

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}\Lambda_i\left(1-\Lambda_i\right)\mathbf{x}_i\mathbf{x}_i' = -\mathbf{X}'\mathbf{V}\mathbf{X}$$

darstellen, wobei $\mathbf{X}_{(n\times k)} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}$ und $\mathbf{V}_{(n\times n)} = \begin{bmatrix} \Lambda_1\left(1-\Lambda_1\right) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Lambda_n\left(1-\Lambda_n\right) \end{bmatrix}$

sind.

Da $\mathbf{V}$ eine Diagonalmatrix ist und $\Lambda_i\left(1-\Lambda_i\right) > 0\ \forall\, i$, ist $\mathbf{V}$ positiv definit. Jede positiv definite Matrix lässt eine eindeutige Zerlegung $\mathbf{V} = \mathbf{\Gamma}\mathbf{\Gamma}'$ zu (in unserem Fall $\mathbf{\Gamma}$ ist wieder eine Diagonalmatrix mit $\sqrt{\Lambda_1\left(1-\Lambda_1\right)}$ an der Hauptdiagonal). Dadurch

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'} = -\mathbf{X}'\mathbf{V}\mathbf{X} = -\mathbf{X}'\left(\mathbf{\Gamma}\mathbf{\Gamma}'\right)\mathbf{X} = -(\underbrace{\mathbf{\Gamma}'\mathbf{X}}_{\equiv\mathbf{Z}})'(\underbrace{\mathbf{\Gamma}'\mathbf{X}}_{\equiv\mathbf{Z}}) = -\mathbf{Z}'\mathbf{Z}.$$

Solange $\mathbf{X}'\mathbf{X}$ nicht singulär ist, ist $\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'} = -\mathbf{Z}'\mathbf{Z}$ negativ definit, d.h. die log-Likelihood *global konkav* in α und β. Dies bedeutet, daß es <u>nur</u> ein Maximum gibt. Da Definitheit auch einen vollen Rang impliziert, ist die Aussage oben der Aussage der Theorem 1 äquivalent.

Ist binäres Logit-Modell immer identifizierbar?

Wir haben angenommen, daß bei der logistischen Verteilungsfunktion $\delta = 0$ und $\gamma = 1$ sind. Hätten wir einen Parameter davon, z.B. δ, nicht restringiert, dann bekämen wir

$$F(z_i) = \frac{1}{1 + \exp\{-\mathbf{x}_i'\beta - \delta\}} = \frac{1}{1 + \exp\left\{-\left[(\beta_0 + \delta) + \underline{\mathbf{x}}_i'\underline{\beta}\right]\right\}},$$

wobei $\underline{\mathbf{x}}_i$ ein Vektor der erklärenden Variablen ohne Absolutglied ist, und $\underline{\beta}$ ist der entsprechende Parametervektor. Somit gäbe es unendlich viele Kombinationen von β_0 und δ die zu dem gleichen Wert der $F(z_i) \Leftrightarrow \ln \mathcal{L}$ führen. D.h., ohne Beschränkung $\delta = 0$ ist die Schätzung des Parameter β_0 nicht möglich. Das gleiche gilt für γ

$$F(z_i) = \frac{1}{1 + \exp\{-\mathbf{x}_i'\beta/\gamma\}},$$

wo unendlich viele Kombinationen von β/γ zu dem gleichen Wert der $F(z_i) \Leftrightarrow \ln \mathcal{L}$ führen. Demzufolge kann β getrennt von γ nicht geschätzt werden. Die Normierung $\delta = 0$ und $\gamma = 1$ ist, also, eine implizite Identifizierbarkeitsrestriktion. Es kann auch gezeigt werden, daß wenn entweder δ oder γ zum gegeben Wert nicht restringiert ist, hat die Hesse-Matrix kein vollen Rang mehr. Am ökonomischen Inhalt des Modells ändern die obige Identifizierbarkeitsrestriktionen nichts.

- Interprätation der Schätzergebnisse

Partielle Effekte im Logit-Modell erhält man über

$$\frac{\partial E(y_i|\mathbf{x}_i)}{\partial x_j} = \frac{\partial P\left(y_i = 1|\mathbf{x}_i\right)}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{1}{1 + \exp\{-\mathbf{x}_i'\beta\}}\right) = \lambda\left(\mathbf{x}_i'\beta\right)\beta_j.$$

Logit bietet auch eine günstige Möglichkeit die relative Wahrscheinlichkeiten zu untersuchen an

$$\frac{P\left(y_i = 1|\mathbf{x}_i\right)}{P\left(y_i = 0|\mathbf{x}_i\right)} = \frac{\frac{1}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}}{\frac{\exp\left\{-\mathbf{x}_i'\beta\right\}}{1+\exp\left\{-\mathbf{x}_i'\beta\right\}}} = \exp\left\{\mathbf{x}_i'\beta\right\} \quad \Rightarrow \quad \frac{\partial}{\partial x_j}\left(\frac{P\left(y_i = 1|\mathbf{x}_i\right)}{P\left(y_i = 0|\mathbf{x}_i\right)}\right) = \beta_j \exp\left\{\mathbf{x}_i'\beta\right\},$$

und schließlich: $$\frac{\partial}{\partial x_j}\left(\frac{P\left(y_i = 1|\mathbf{x}_i\right)}{P\left(y_i = 0|\mathbf{x}_i\right)}\right) = \beta_j \frac{P\left(y_i = 1|\mathbf{x}_i\right)}{P\left(y_i = 0|\mathbf{x}_i\right)}.$$

4.1.3 Binäres Probit-Modell

Probit-Modell ist das binäre Modell, die eine *Normalverteilung* für die Spezifikation der Wahrscheinlichkeit des Ereignisses verwendet. Die Dichtefunktion und die kumulative Dichte der Normalverteilung sind gegeben durch

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}((z-\mu)/\sigma)^2} \equiv \phi\left(\frac{z-\mu}{\sigma}\right), \text{ und } P(Z \leq z) = \int_{-\infty}^{z} f(x)dx = \Phi\left(\frac{z-\mu}{\sigma}\right).$$

Aus dem schon bekannten Identifizierbarkeitsgrund setzen wir $\mu = 0$ und $\sigma = 1$. Dann sind

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \equiv \phi(z), \quad \text{und} \quad P(Z \leq z) = \int_{-\infty}^{z} f(x)dx = \Phi(z),$$

und somit ist die Likelihoodfunktion

$$\mathcal{L} = \prod_{i=1}^{n} \left[\Phi(\mathbf{x}_i'\beta)\right]^{y_i} \left[1 - \Phi(\mathbf{x}_i'\beta)\right]^{1-y_i}.$$

Ansonsten, entsteht Probit-Modell auf einer natürlichen Weise aus der latenten Variable / Indexfunktion Darstellung, da in der linearen Regressionen üblicherweise die Normalität des Störterms angenommen wird. Gegeben

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

und $y_i = 1$ wenn $y_i^* > 0$ ($y_i = 0$ sonst) die bedingte Wahrscheinlichkeit $y_i = 1$ zu beobachten ist

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(y_i^* > 0|\mathbf{x}_i) = P(\mathbf{x}_i'\beta + \varepsilon_i > 0) \\ &= P(\varepsilon_i > -\mathbf{x}_i'\beta) = 1 - P(\varepsilon_i \leq -\mathbf{x}_i'\beta) \\ &= 1 - \Phi\left(\frac{-\mathbf{x}_i'\beta - 0}{\sigma}\right) = 1 - \left[1 - \Phi(\mathbf{x}_i'\beta/\sigma)\right] \\ &= \Phi(\mathbf{x}_i'\beta/\sigma), \end{aligned}$$

Darüber hinaus, wegen der Unmöglichkeit β und σ getrennt zu schätzen wird $\sigma = 1$ angenommen (die andere Restriktion $\mu = 0$ ist implizit durch $E(\varepsilon_i) = 0$ gegeben). Somit $P(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i'\beta)$ und die Likelihoodfunktion ist wie oben.

- Die Likelihoodfunktion
$$\mathcal{L} = \prod_{i=1}^{n} \left[\Phi(\mathbf{x}_i'\beta)\right]^{y_i} \left[1 - \Phi(\mathbf{x}_i'\beta)\right]^{1-y_i}.$$

- Die log-Likelihoodfunktion und B.E.O.
$$\ln \mathcal{L} = \sum_{i=1}^{n} \left[y_i \ln \Phi\left(\mathbf{x}_i'\beta\right) + (1 - y_i) \ln\left(1 - \Phi\left(\mathbf{x}_i'\beta\right)\right)\right].$$

Wir schalgen weiter vor die Abkürzungen $\Phi_i = \Phi\left(\mathbf{x}_i'\beta\right)$ und $\phi_i = \phi\left(\mathbf{x}_i'\beta\right)$ zu verwenden. Für die partielle Ableitung nach einem β_j erhalten wir dementsprechend

$$\frac{\partial \Phi_i}{\partial \beta_j} = \phi_i \frac{\partial \mathbf{x}_i'\beta}{\partial \beta_j} = \phi_i x_{ij} \quad \Rightarrow \quad \frac{\partial \Phi_i}{\partial \beta} = \phi_i \mathbf{x}_i.$$

Daraus resultieren die B.E.O.

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial \beta} &= \sum_{i=1}^{n} \left[\frac{y_i}{\Phi_i}\phi_i \mathbf{x}_i + \frac{1 - y_i}{1 - \Phi_i}\left(-\phi_i\right)\mathbf{x}_i\right] \\
&= \sum_{i=1}^{n} \frac{y_i\left(1 - \Phi_i\right) - \left(1 - y_i\right)\Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\phi_i \mathbf{x}_i \\
&= \sum_{i=1}^{n} \frac{y_i - \Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\phi_i \mathbf{x}_i \stackrel{!}{=} \mathbf{0}
\end{aligned}$$

Die Hesse-Matrix

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} &= \frac{\partial}{\partial \beta'}\left(\sum_{i=1}^{n} \frac{y_i - \Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\phi_i \mathbf{x}_i\right) \\
&= \sum_{i=1}^{n} \left[\phi_i \mathbf{x}_i \frac{-\phi_i \mathbf{x}_i'\left(\Phi_i\left[1 - \Phi_i\right]\right) - \left(y_i - \Phi_i\right)\left(\phi_i - 2\Phi_i\phi_i\right)\mathbf{x}_i'}{\left[\Phi_i(1 - \Phi_i)\right]^2} + \frac{y_i - \Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\mathbf{x}_i \frac{\partial \phi_i}{\partial \beta'}\right]
\end{aligned}$$

Nun: $\dfrac{\partial \phi_i}{\partial \beta'} = \dfrac{\partial}{\partial \beta'}\left(\dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\mathbf{x}_i'\beta\right)^2}\right) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\mathbf{x}_i'\beta\right)^2} \dfrac{\partial}{\partial \beta'}\left(-\dfrac{1}{2}\left(\mathbf{x}_i'\beta\right)^2\right) = -\phi_i\left(\mathbf{x}_i'\beta\right)\mathbf{x}_i'.$

Dann

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} &= \sum_{i=1}^{n} \left[\phi_i \mathbf{x}_i \mathbf{x}_i' \frac{-\phi_i \Phi_i + \phi_i \Phi_i^2 - \left[y_i \phi_i - \phi_i \Phi_i - 2\Phi_i \phi_i y_i + 2\Phi_i^2 \phi_i\right]}{\left[\Phi_i(1 - \Phi_i)\right]^2}\right. \\
&\qquad \left. - \frac{y_i - \Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\phi_i\left(\mathbf{x}_i'\beta\right)\mathbf{x}_i \mathbf{x}_i'\right] \\
&= \sum_{i=1}^{n} \left[\phi_i \mathbf{x}_i \mathbf{x}_i' \frac{-y_i \phi_i + 2\Phi_i \phi_i y_i - \Phi_i^2 \phi_i}{\left[\Phi_i(1 - \Phi_i)\right]^2} - \frac{y_i - \Phi_i}{\Phi_i\left[1 - \Phi_i\right]}\phi_i\left(\mathbf{x}_i'\beta\right)\mathbf{x}_i \mathbf{x}_i'\right] \\
&= \sum_{i=1}^{n} \frac{-\left(y_i - 2\Phi_i y_i + \Phi_i^2\right)\phi_i - \left(y_i - \Phi_i\right)\Phi_i\left[1 - \Phi_i\right]\left(\mathbf{x}_i'\beta\right)}{\left[\Phi_i(1 - \Phi_i)\right]^2}\phi_i \mathbf{x}_i \mathbf{x}_i'
\end{aligned}$$

$$= -\sum_{i=1}^{n} \frac{\psi_i}{[\Phi_i(1-\Phi_i)]^2}\phi_i \mathbf{x}_i \mathbf{x}_i',$$

wobei $\psi_i \equiv (y_i - 2\Phi_i y_i + \Phi_i^2)\,\phi_i + (y_i - \Phi_i)\,\Phi_i\,[1-\Phi_i]\,(\mathbf{x}_i'\beta)$.

Weil ψ_i von y_i abhängt, sind im binären Probit Hesse-Matrix und Informationsmatrix unterschiedlich. Es gilt: $E(y_i) = \Phi_i$ und damit

$$\begin{aligned} E(\psi_i) &= \left(E(y_i) - 2\Phi_i E(y_i) + \Phi_i^2\right)\phi_i + (E(y_i) - \Phi_i)\,\Phi_i\,[1-\Phi_i]\,(\mathbf{x}_i'\beta) \\ &= \left(\Phi_i - 2\Phi_i^2 + \Phi_i^2\right)\phi_i + \underbrace{(\Phi_i - \Phi_i)}_{=0}\Phi_i\,[1-\Phi_i]\,(\mathbf{x}_i'\beta) \\ &= \left(\Phi_i - \Phi_i^2\right)\phi_i = \Phi_i\,(1-\Phi_i)\,\phi_i \end{aligned}$$

Dies führt zu der Informationsmatrix

$$\mathbf{I}(\beta) = -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'}\right) = \sum_{i=1}^{n} \frac{E(\psi_i)}{[\Phi_i(1-\Phi_i)]^2}\phi_i \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} \frac{\phi_i^2}{\Phi_i(1-\Phi_i)}\mathbf{x}_i \mathbf{x}_i'.$$

- Eindeutigkeit des Schätzers und Identifizierbarkeitsrestriktionen

Wie im Fall Logit lässt sich die Informationsmatrix als $\mathbf{I}(\beta) = \mathbf{X}'\mathbf{V}\mathbf{X}$ darstellen, wobei $\mathbf{V}$ eine Daigonalmatrix mit $\phi_i^2/\Phi_i(1-\Phi_i)$ an der Hauptdiagonal ist. Da $\phi_i^2/\Phi_i(1-\Phi_i) > 0\ \forall\, i$, ist $\mathbf{V}$ positiv definit. Dadurch ist auch $\mathbf{I}(\beta) = \mathbf{X}'\mathbf{V}\mathbf{X}$ positiv definit, solange $\mathbf{X}'\mathbf{X}$ nicht singulär ist. Dies impliziert, daß die Informationsmatrix einen vollen Rang hat. Anwendung der Informationsgleichheit und Theorem 1 zeigt, daß Probit Modell identifiziert ist, d.h. ML Schätzer ist eindeutig.

- Interprätation der Schätzergebnisse

Partielle Effekte im Probit-Modell sind

$$\frac{\partial E(y_i|\mathbf{x}_i)}{\partial x_j} = \frac{\partial P(y_i = 1|\mathbf{x}_i)}{\partial x_j} = \frac{\partial \Phi(\mathbf{x}_i'\beta)}{\partial x_j} = \phi(\mathbf{x}_i'\beta)\,\beta_j.$$

4.2 Multinomiale Modelle für geordnete Kategorien

4.2.1 Darstellung des geordneten Modells

Geordnetes Model stellt eine Erweiterung des Modells mit der latenten Variable und binären Indexfunktion dar. Wir bedienen uns der Vorstellung, daß die kategoriale

Variable y mit mehreren Ausprägungen $(j = 1, 2, ..., r)$ mit der latenten Variablen y^* über das Messmodell

$$y = j \quad \Leftrightarrow \quad \gamma_{j-1} < y^* \leq \gamma_j, \qquad \text{mit} \quad \gamma_0 = -\infty \quad \text{und} \quad \gamma_r = \infty$$

zusammen hängt. Da nun alle r Kategorien geordnet sind, deutet eine höhere Ausprägung $y_i = j + 1$ gegenüber einer niedrigeren Ausprägung $y_k = j$ an, daß die Realisation der latenten Variablen für Individuum i größer ist als die von Individuum k: $y_i^* > y_k^*$.

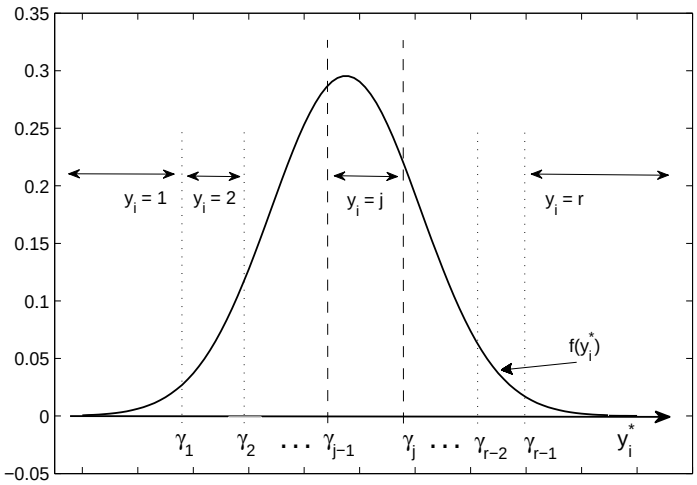
Die beobachtbare Ausprägungen der kategorialen Indexvariable werden also folgendermaßen dargestellt

$$y_i = \begin{cases} r, & \text{wenn} \quad \gamma_{r-1} < y_i^* < \infty \\ ... & \\ 2, & \text{wenn} \quad \gamma_1 < y_i^* \leq \gamma_2 \\ 1, & \text{wenn} \quad -\infty < y_i^* \leq \gamma_1 \end{cases},$$

$-\infty = \gamma_0 < \gamma_1 < \gamma_2 < ... < \gamma_{r-1} < \gamma_r = +\infty$. Die latente Variable wird, wie im Fall der binären Variablen y, mit einem Erwartungswert modelliert, der linear in den Einflussgrößen und ebenso linear in den Parametern ist, also

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i,$$

wobei $E(\varepsilon_i)$ und ε_i kommt aus einer bekannten Verteilung F. Die Annahme der linearität der latenten regression ist, allerdings, nicht erforderlich (siehe unten). Das folgende Schaubild erläutert das Modell:



Die Wahrscheinlichkeit für das Ereignis $y = j$ ergibt sich daraus als

$$\begin{aligned} P(y_i = j|\mathbf{x}_i) &= P(\gamma_{j-1} < y_i^* \leq \gamma_j|\mathbf{x}_i) \\ &= P(y_i^* \leq \gamma_j|\mathbf{x}_i) - P(y_i^* \leq \gamma_{j-1}|\mathbf{x}_i) \\ &= P(\mathbf{x}_i'\beta + \varepsilon_i \leq \gamma_j) - P(\mathbf{x}_i'\beta + \varepsilon_i \leq \gamma_{j-1}) \\ &= P(\varepsilon_i \leq \gamma_j - \mathbf{x}_i'\beta) - P(\varepsilon_i \leq \gamma_{j-1} - \mathbf{x}_i'\beta) \\ &= F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta) \end{aligned}$$

Definieren wir die Dummyvariable y_{ij}, so daß

$$y_{ij} = \begin{cases} 1, & \text{wenn } y_i = j \\ 0, & \text{sonst} \end{cases} .$$

Dann die multinomiale Dichte, d.h. die Wahrscheinlichkeit $y_i = j$ für ein beliebigen $j = 1, ..., r$ zu beobachten ist

$$g(y_i) = \prod_{j=1}^{r} \left[F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)\right]^{y_{ij}} .$$

Die Likelihoodfunktion für das geordnete Modell ist somit

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)\right]^{y_{ij}} .$$

- Ökonomischer Inhalt des Modells

Das Modell bildet den Inhalt eines Rankings ab. Dabei ist es vollkommen egal was gerankt wird. Das Modell schätzt den wahren Zusamenhang zwischen den beobachtabren Charakteristika und unbeobachtbaren Nutzen.

Beispiel: Evaluation der Zufriedenheit mit einem Produkt. Typische Information, die wir von einem Kunden erhalten sieht wie: "nicht zufrieden" (1), "eher unzufrieden als zufrieden" (2), "hmm .. keine Ahnung" (3), "eher zufrieden" (4), "alles wunderbar" (5) aus. Diese Ranking wird aufgrund der Höhe der Nutzenfunktion des Kunden erstellt. Wir beobachten keinen Nutzen. Darüber hinaus können wir auch nicht wissen ob der Nutzenunterschied zwischen "nicht zufrieden" und "eher unzufrieden als zufrieden" dem Nutzenunterschied zwischen "eher zufrieden" und "alles wunderbar" gleich ist. Geordnetes Modell schätzt die Schwellenparameter $\{\gamma_j\}_{j=1}^{r-1}$ die diese Nutzenunterschiede genau abbildet. Richtige Schätzung der Schwellenparameter versichert daß die restliche Parameter β durch eine arbiträre Festlegung der Nutzenunterschiede nicht verzerrt sind.

4.2.2 Geordnete Logit- und Probit-Modelle

Wie im binären Fall beide Modelle ergeben sich durch die Verteilungsannahmen für den Störterm in der latenten Regression. Setzt man

$$F(z) = \Lambda(z; 0, 1) = \frac{1}{1 + \exp\{-z\}},$$

erhält man ein geordnetes Logit-Modell. Setzt man

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx,$$

erhält man ein geordnetes Probit-Modell.

Aus dem gleichen Identifizierbarkeitsgrund wie früher normiert man die beide Verteilungen, um den doppelten Absolutglied bzw. die Skalierung des gesamten Parametervektors zu vermeiden. Ferner ist im Fall, dass $\mathbf{x}_i$ ein Absolutglied beinhaltet, nur die Differenz $(\gamma_j - \beta_0)$ identifiziert. D.h. ein Schwellenwert oder das Absolutglied müssen gleich null gesetzt werden. Üblich ist die Wahl $\gamma_1 = 0$ oder eben $\beta_0 = 0$. Wird $\gamma_1 = 0$ gesetzt und $\hat{\beta}_0$ geschätzt, kann sofort gefolgert werden, dass die alternative Restringierung $\beta_0 = 0$ einen Schätzwert von $\tilde{\gamma}_1 = -\hat{\beta}_0$ ergibt und $\tilde{\gamma}_j = \hat{\gamma}_j - \hat{\beta}_0$, wobei die Tilde die Schätzwerte der zweiten Spezifikation und das Dach die Schätzwerte der ersten Spezifikation bezeichnen. Wir werden weiterhin die Annahme $\beta_0 = 0$ aufrechterhalten. Somit sind alle $\{\gamma_j\}_{j=1}^{r-1}$ zu schätzen.

Da eine bestimmte Paramterische Annahme für $F(\cdot)$, ob Logit oder Probit, nichts an der Herleitung der B.E.O. ändert, betrachten wir wieder das allgeneine Modell.

- Die Likelihoodfunktion
  $$\mathcal{L} = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)\right]^{y_{ij}}.$$

- Die log-Likelihoodfunktion und B.E.O.
  $$\ln \mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \ln\left(F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)\right).$$

Definieren wir $p_{ij} \equiv F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)$ um die schreibweise an den günstigen Stellen einfacher darzustellen. Somit, die B.E.O.

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial \beta} &= \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \frac{f(\gamma_j - \mathbf{x}_i'\beta) \cdot (-\mathbf{x}_i) - f(\gamma_{j-1} - \mathbf{x}_i'\beta) \cdot (-\mathbf{x}_i)}{F(\gamma_j - \mathbf{x}_i'\beta) - F(\gamma_{j-1} - \mathbf{x}_i'\beta)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{y_{ij}}{p_{ij}} \left[f(\gamma_{j-1} - \mathbf{x}_i'\beta) - f(\gamma_j - \mathbf{x}_i'\beta)\right] \mathbf{x}_i \stackrel{!}{=} \mathbf{0}
\end{aligned}$$

Bei der partiellen Ableitungen nach den Schwellenwerten tauchen in jedem p_{ji} nur die Schwellen γ_j und γ_{j-1} auf, so daß

$$\frac{\partial p_{ji}}{\partial \gamma_j} \;=\; f(\gamma_j - \mathbf{x}_i'\beta) \qquad \text{und} \qquad \frac{\partial p_{ji}}{\partial \gamma_{j-1}} \;=\; -f(\gamma_{j-1} - \mathbf{x}_i'\beta)$$

und alle anderen partiellen Ableitungen gleich null sind. Deshalb können unter Zuhilfenahme einer zusätzlichen Indikatorvariablen

$$\delta_{j,k} \begin{cases} 1, & \text{falls } \; j = k \\ 0, & \text{sonst} \end{cases}$$

die Ableitungen nach γ_k für alle $k = 1, ..., r-1$ als

$$\frac{\partial \ln L}{\partial \gamma_k} \;= \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{y_{ji}}{p_{ji}} \left[\delta_{j,k} f(\gamma_j - \mathbf{x}_i'\beta) \;- \delta_{j-1,\,k} f(\gamma_{j-1} - \mathbf{x}_i'\beta)\right] \;\stackrel{!}{=}\; 0$$

geschrieben werden.

Einsetzung der relevanten Dichte- bzw. kumulativen Dichtefunktion ergibt die B.E.O. für geordnetes Logit (oder Probit).

- Eindeutigkeit des Schätzers und Identifizierbarkeitsrestriktionen

Unter Identifizierbarkeitsrestriktionen oben, nämlich unter der Normierung der Wahrscheinlichkeitsfunktion und Annahme $\beta_0 = 0$ (alternativ: $\gamma_1 = 0$) sind die geordnete Logit / Probit identifizierbar, d.h. Der ML Schätzer in deisen Modelle ist eindeutig.

- Interprätation der Schätzergebnisse

Interpretation bei der latenten Regression

1. Partielle Effekte, analog zum linearen Regressionsmodell, sind $\partial E\left(y^*|\mathbf{x}_i\right)/\partial x_k = \beta_k$. Allerdings, wegen der Normierung, wie $\sigma = 1$ im geordneten Probit, zum Beispiel, entspricht der Schätzwert $\hat{\beta}_k$ dem wahren Einfluß auf die latente Variable, sagen wir β_k^*, nicht, denn $\hat{\beta}_k = \beta_k^*/\sigma^*$. Immerhin bleibt der relative wahre Einfluß auf die latente Variable nachvollziehbar

$$\frac{\partial E\left(y^*|\mathbf{x}_i\right)/\partial x_k}{\partial E\left(y^*|\mathbf{x}_i\right)/\partial x_l} = \frac{\hat{\beta}_k}{\hat{\beta}_l} = \frac{\beta_k^*/\sigma^*}{\beta_l^*/\sigma^*} = \frac{\beta_k^*}{\beta_l^*}.$$

2. Da y_i^* latent ist, gibt es eine Beobachtungsäquivalenz zu streng-monotonen Transformationen, nämlich

$$y_i = j \;\Leftrightarrow\; \gamma_{j-1} < y_i^* \leq \gamma_j \;\Leftrightarrow\; h(\gamma_{j-1}) < h(y_i^*) \leq h(\gamma_j),$$

wobei $h(\cdot)$ eine beliebige streng monoton steigende Funktion ist. Dies bedeutet, daß die ursprüngliche latente Regression nicht unbedingt linear sein muss, um zum geordneten Modell zu führen. Zum Beispiel, eine typische Lohnregression ist

$$w_i^* = \exp\left\{\mathbf{x}_i'\beta\right\}\epsilon_i, \quad E\left(\epsilon_i\right) = 1,$$

da die Löhne positiv sind. Die Umformung und anschließende streng monotone Transformation zeigen, jedoch, daß

$$\begin{aligned} w_i^* = \exp\left\{\mathbf{x}_i'\beta\right\}\epsilon_i \quad &\Leftrightarrow \quad w_i^* = \exp\left\{\mathbf{x}_i'\beta\right\}e^{\varepsilon_i} \quad \Leftrightarrow \quad w_i^* = \exp\left\{\mathbf{x}_i'\beta+\varepsilon_i\right\} \\ \Rightarrow \quad h(\cdot) = \ln(\cdot) \quad &\Rightarrow \qquad \ln\left(w_i^*\right) = \mathbf{x}_i'\beta+\varepsilon_i, \quad E\left(\varepsilon_i\right) = 0. \end{aligned}$$

Aus dieser Hinsicht ist die Spezifikation $E\left(y^*|\mathbf{x}_i\right) = \mathbf{x}_i'\beta$ wenig restriktiv, da die ursprüngliche latente Variable häufig geeignet transformiert werden kann um schließlich einen linearen Erwartungswert zu ergeben.

Interpretation bei der Regression auf die beobachtbare werte y_i

3. Partielle Effekte auf die Wahrscheinlichkeit in die Kategorie j zu gelangen sind

$$\frac{\partial p_{ji}}{\partial x_l} = \left[f\left(\gamma_{j-1} - \mathbf{x}_i'\beta\right) - f\left(\gamma_j - \mathbf{x}_i'\beta\right)\right]\beta_l.$$

Dies, wie im binären Modell, sowie im allgemeinen nichtlinearen Modell, hängt von $\mathbf{x}_i$ ab. Abgesehen davon, daß man den partiellen Effekt für eine bestimmte $\mathbf{x}_i$ berechen kann, gibt es auch zwei weitere Möglichkeiten einen durchschnittlichen Effekt zu bekommen. Definiert man $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k]' = \left[\frac{1}{n}\sum_{i=1}^n x_{1,i}, ..., \frac{1}{n}\sum_{i=1}^n x_{k,i}\right]'$, dann:

$$\begin{aligned} a) \quad : \quad & \left.\frac{\partial p_j}{\partial x_l}\right|_{\mathbf{x}_i=\bar{\mathbf{x}}} = \left[f\left(\gamma_{j-1} - \bar{\mathbf{x}}'\beta\right) - f\left(\gamma_j - \bar{\mathbf{x}}'\beta\right)\right]\beta_l, \\ b) \quad : \quad & \frac{\partial p_j}{\partial x_l} = \frac{1}{n}\sum_{i=1}^n \left[f\left(\gamma_{j-1} - \mathbf{x}_i'\beta\right) - f\left(\gamma_j - \mathbf{x}_i'\beta\right)\right]\beta_l. \end{aligned}$$

Dabei ist a) der partielle Effekt für die durchschnittliche Beobachtung und b) ist der durchschnittliche partielle Effekt für alle Beobachtungen. Die Beide sind nie gleich. Der Unterschied ist die folge der Jensens Ungleichung.

4.3 Gütemaße und Spezifikationstests

Nach der Schätzung soll nun die Güte des Modells, d.h. die Beurteilung, wie gut das Modell die Daten beschreibt, sowie Tests, ob wichtige Aspekte des Modells vernachlässigt wurden, im Vordergrund stehen.

4.3.1 Gütemaße

- Pseudo-R^2

Im linearen Regressionsmodell wird die quadrierte Abweichung der beobachteten y_i-Werte von den erklärten Werten $\hat{y}_i = \mathbf{x}_i'\hat{\beta}$ als Gütemaß herangezogen:

$$s_\varepsilon^2 = \frac{1}{n}\sum (y_i - \widehat{E(y_i|\mathbf{x}_i)})^2 = \frac{1}{n}\sum (y_i - \mathbf{x}_i'\hat{\beta})^2 \qquad \Rightarrow \quad R^2 = 1 - \frac{s_\varepsilon^2}{s_y^2}.$$

Allerdings ist die Interpretation des R^2 nur für das lineare Regressionsmodell gültig. Im nichtlinearen Modell, der Nichtlinearität wegen, ist das nicht mehr der Anteil erklärten Variation an der gesamten variation der zureklärenden Variable. Deshalb wurden die Alternativen vorgeschalgen. Ein erster Vorschlag ist ein Pseudo-R^2-Maß nach McFadden

$$R^2_{MF} = 1 - \frac{\ln \mathcal{L}}{\ln \mathcal{L}_0},$$

wobei $\mathcal{L}_0$ das Modell ausschließlich mit Absolutglied darstellt. $\ln \mathcal{L}_0$ ist kleiner als 0, weil $\mathcal{L}$ als gemeinsame Wahrscheinlichkeit zu sehen ist, und ebenso kleiner als $\ln \mathcal{L}$, weil $\mathcal{L}_0$ ein gegenüber $\mathcal{L}$ restringiertes Modell repräsentiert. Demnach liegt $\ln \mathcal{L}/\ln \mathcal{L}_0$ zwischen 0 und 1 und R^2_{MF} ebenso zwischen 0 und 1. Es kann nicht den Wert 1 annehmen, da gemäß Modellierung $0 < P(\mathbf{x}_i) < 1$. Für $\mathcal{L} = 1$ bzw. $\ln \mathcal{L} = 0$ muss aber $P(\mathbf{x}_i) = 1$ werden. Ein alternatives Maß, das etwas bessere Eigenschaften als R^2_{MF} hat, ist R^2_{AN} nach Aldrich und Nelson

$$R^2_{AN} = \frac{2(\ln \mathcal{L} - \ln \mathcal{L}_0)}{2(\ln \mathcal{L} - \ln \mathcal{L}_0) + n} = \frac{LR}{LR + n}.$$

Mehrere findet man im Greene (2003), Ch.21.4.5. In der Tat, ist in einem nichtlinearen Modell ein Pseudo-R^2 eher ein Vergleichsmaß als ein Maß der Anpassungsgüte.

- Tests für die Anpassungsgüte

Diese Tests, ähnlich zu der Idee eines R^2 im linearen Modell, stehen die durch ein Modell vorausgesagte Werte den tatsächlich beobachteten Werten gegenüber.

Ausschließlich für binäre Modelle steht der Hosmer-Lemeshow Test zur Verfügung (Hosmer und Lemeshow, 2000). Man teilt die Daten auf J Gruppen auf (z.B. Mann/Frau & Ost/West ergeben schon vier Gruppen). Dann für jede $j = 1, ..., J$, definiert man

$$y(j) = \sum_{i \in \{j\}} y_i \quad \text{und} \quad \hat{p}(j) = \sum_{i \in \{j\}} F(\mathbf{x}_i'\hat{\beta})/n_j,$$

wobei n_j ein Anzahl der Beobachtungen in der Gruppe j ist. Somit $y(j)/n_j$ ist nichts anderes als die beobachtete Häufigkeit des $y_i = 1$ Ereignisses in Gruppe j, und $\hat{p}(j)$ ist die vorausgesagte Häufigkeit des $y_i = 1$ Ereignisses in Gruppe j. Hosmer-Lemeshow Statistik lautet

$$HL_J = \sum_{j=1}^{J} \frac{(y(j) - n_j \hat{p}(j))^2}{n_j \hat{p}(j)[1 - \hat{p}(j)]} \simeq \chi^2_{J-2}.$$

Die Statistik betrachtet die Summe der quadrierten Abweichungen zwischen den Anzahl an beobachteten und "vorausgesagten" Ereignissen in jeder Gruppe, normiert jeweils auf deren Varianz. Die asymptotische Verteilung ist unbekannt, aber zahlreiche Simulationen zeigen, daß die zu χ^2 mit $J-2$ Freiheitsgraden neigt. Die übliche Anzahl der Gruppen ist 10. Die Nullhypothese, wie bei allen Tests für anpassungsgüte ist, daß das Modell die beobachtbare Daten präzise abbildet.

4.3.2 Spezifikationstests

Die im Teil I, Abschnitt 3, behandelten Testverfahren sollen hier auf zwei Situationen: gemeinsame Signifikanz der Variablen und Heteroskedastizität, angewendet werden. Exemplarisch sollen sie für binäres Probit-Modell gezeigt werden.

- Wald Test für gemeinsame Signifikanz

Für die latente abhängige Variable y_i^* gelte es der folgende wahre Zusammenhang

$$E(y_i^* | \underset{(k\times 1)}{\mathbf{x}_i}, \underset{(m\times 1)}{\mathbf{z}_i}) = \mathbf{x}_i'\beta + \mathbf{z}_i'\gamma.$$

Kompakter lässt sich dies schreiben durch

$$\mathbf{w}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{pmatrix} \quad \text{und} \quad \theta = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}, \quad \text{und somit:} \quad E(y_i | \mathbf{x}_i, \mathbf{z}_i) = \mathbf{w}_i'\theta.$$

Ein Test, ob die Variablen in $\mathbf{z}$ vernachlässigt werden dürfen, lautet dann $H_0 : \gamma = \mathbf{0}$:

$$\underset{m\times[k+m]}{\mathbf{R}} \cdot \underset{[k+m]\times 1}{\theta} = \mathbf{0}, \quad \text{wobei} \quad \mathbf{R} = [\underset{m\times k}{\mathbf{0}}, \underset{m\times m}{\mathbf{I}}] = \begin{bmatrix} 0 & \dots & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & 0 & & 1 \end{bmatrix}$$

$\mathbf{R}\theta = \mathbf{0}$ entspricht $h(\boldsymbol{\theta}) = \mathbf{0}$ aus Teil I, Abschnitt 3. Die Informationsmatrix für unser wahres Modell ist

$$\mathbf{I}(\hat{\theta}) = \sum_{i=1}^{n} \frac{\left[\phi(\mathbf{w}_i'\hat{\theta})\right]^2}{\Phi(\mathbf{w}_i'\hat{\theta})\left[1 - \Phi(\mathbf{w}_i'\hat{\theta})\right]} \mathbf{w}_i\mathbf{w}_i'.$$

Die Inversion von $\mathbf{I}(\hat{\theta})$ ergibt die asymptotische Varianz-Kovarianz-Matrix von $\hat{\theta}$, die wir wie folgt partitionieren:

$$\mathbf{Var}(\hat{\theta}) = \mathbf{I}(\hat{\theta})^{-1} = \begin{bmatrix} \boldsymbol{V}_{\hat{\beta}\hat{\beta}} & \boldsymbol{V}_{\hat{\beta}\hat{\gamma}} \\ \boldsymbol{V}_{\hat{\beta}\hat{\gamma}} & \boldsymbol{V}_{\hat{\gamma}\hat{\gamma}} \end{bmatrix}.$$

Die Wald Test-Statistik $W = (h(\hat{\theta}) - h(\theta))'[\mathbf{DI}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{D}']^{-1}(h(\hat{\boldsymbol{\theta}}) - h(\theta))$ lautet in unserem Fall mit $h(\theta) = \mathbf{0} = \mathbf{R}\theta$ und $\mathbf{D} = \nabla h(\theta) = \mathbf{R}$ wie

$$W = (\mathbf{R}\hat{\theta})'(\mathbf{R}\mathbf{Var}(\hat{\theta})\mathbf{R}')^{-1}(\mathbf{R}\hat{\theta}) \qquad \Rightarrow \qquad W = \hat{\gamma}'\boldsymbol{V}_{\hat{\gamma}\hat{\gamma}}^{-1}\hat{\gamma} \sim \chi_m^2.$$

Wie bereits in Teil I, Abschnitt 3, erwähnt, braucht man für den Wald-Test lediglich den unrestringierten Schätzer $\hat{\theta}$.

- LM Test für Heteroskedastizität

Der bedingte Mittelwert soll nun wieder die übliche Form $E(y_i^*|\mathbf{x}_i) = \mathbf{x}_i'\beta$ haben. Wir nehmen allerdings an, daß in der Regressinsmodell für die latente Variable y_i^*

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$$

die Varianz $Var(\varepsilon_i) = \sigma_i \neq \sigma$, d.h. ungleich über die Individuen ist. Unter $\sigma = 1$, schlagen wir

$$Var(\varepsilon_i) = \left(\exp\{\mathbf{z}_i'\gamma\}\right)^2 = Var(y_i^*|\mathbf{x}_i)$$

vor. Demnach gilt

$$P(y_i = 1|\mathbf{x}_i) = \Phi\left(\frac{\mathbf{x}_i'\beta}{\exp\{\mathbf{z}_i'\gamma\}}\right) = \Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta).$$

Dies eingesetzt in die Likelihoodfunktion ergibt nun eine Maximierung bezüglich β und γ. Wir setzen wieder

$$\theta = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \quad \text{und} \quad \tilde{\mathbf{w}}_i = \begin{bmatrix} \mathbf{x}_i \\ -(\mathbf{x}_i'\beta)\,\mathbf{z}_i \end{bmatrix}, \quad \text{und erhalten:}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \sum_{i=1}^{n} \frac{y_i - \Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)}{\Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)\left[1 - \Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)\right]} \phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta) e^{-\mathbf{z}_i'\gamma}\tilde{\mathbf{w}}_i$$

und

$$\mathbf{I}(\theta) = \sum_{i=1}^{n} \frac{[\phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)]^2}{\Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)\left[1 - \Phi(e^{-\mathbf{z}_i'\gamma}\mathbf{x}_i'\beta)\right]} (e^{-\mathbf{z}_i'\gamma})^2 \tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i'.$$

Da die Schätzung komplizierter ist als im Fall $\gamma = \mathbf{0}$, ist der LM-Test einfacher, weil wir nur $\hat{\boldsymbol{\theta}}^{(r)} = [\hat{\beta}^{(r)}; \mathbf{0}]$ benötigen, und $\hat{\beta}$ erhalten wir aus einer Standardsoftware. Unter $H_0 : \gamma = \mathbf{0}$ bekomment somit

$$S(\hat{\boldsymbol{\theta}}^{(r)}) = \sum_{i=1}^{n} \frac{[y_i - \Phi(\mathbf{x}_i'\hat{\beta}^{(r)})]\phi(\mathbf{x}_i'\hat{\beta}^{(r)})}{\Phi(\mathbf{x}_i'\hat{\beta}^{(r)})\left[1 - \Phi(\mathbf{x}_i'\hat{\beta}^{(r)})\right]} \tilde{\mathbf{w}}_i$$

und

$$\mathbf{I}(\hat{\boldsymbol{\theta}}^{(r)}) = \sum_{i=1}^{n} \frac{[\phi(\mathbf{x}_i'\hat{\beta}^{(r)})]^2}{\Phi(\mathbf{x}_i'\hat{\beta}^{(r)})\left[1 - \Phi(\mathbf{x}_i'\hat{\beta}^{(r)})\right]} \tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i',$$

wobei der Unterschied zum nichtrestringierten Modell nur darin besteht, daß wir $\tilde{\mathbf{w}}_i = [\mathbf{x}_i; -(\mathbf{x}_i'\beta)\,\mathbf{z}_i]$ anstatt des einfachen Vektor $\mathbf{x}_i$ haben. Die LM-Test Statistik ist dann, wie früher

$$LM = S(\hat{\boldsymbol{\theta}}^{(r)})' \mathbf{I}(\hat{\boldsymbol{\theta}}^{(r)})^{-1} S(\hat{\boldsymbol{\theta}}^{(r)}) \sim \chi^2_m.$$

Die beiden behandelten Situationen haben eine gewisse Verwandtschaft. Nehmen wir an, das wahre Modell sei

$$y_i^* = \mathbf{x}_i'\beta + \mathbf{z}_i'\gamma + \varepsilon_i, \qquad E(\varepsilon_i) = 0, \;\; Var(\varepsilon_i) = \sigma^2.$$

Wenn wir nun $\mathbf{z}_i$ in der Schätzung vernachlässigen, dann betrachten wir faktisch

$$y_i^* = \mathbf{x}_i'\beta + \xi_i, \quad \text{wobei} \quad \xi_i = \varepsilon_i - \mathbf{z}_i'\gamma.$$

Damit $Var(\xi_i) = \sigma_i^2 \neq \sigma^2$ für nicht-deterministische $\mathbf{Z}$. Im Gegensatz zu dem linearen Modell, allerdings, wo der Erwartungswert des KQ Schätzers $E(\hat{\beta}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma = \beta$, d.h. immer noch unverzerrt bleibt, gegeben daß $\mathbf{X}$ orthogonal zu $\mathbf{Z}$ ist, schon im einfachen binären Modell ist das nicht mehr der Fall. Das heißt, sogar unter Orthogonalität von $\mathbf{X}$ und $\mathbf{Z}$ ist der Schätzer verzerrt. Dies bedeutet, daß Ablehnung der Nullhypothese, daß keine Heteroskedastizität vorliegt, eine nichtkonsistente ML-Schätzung bzw. falsche Kovarianzmatrix impliziert.

Literatur

- Cameron, C., and P., Trivedi, "Microeconometrics: Methods and applications", (Cambridge University Press: 2005), Ch.14.2-14.4, p.464-477.
- Greene, W., "Econometric analysis", (Prentice Hall: 2003), 4th Ed., Ch.21.3, Ch.21.4.1-5, (p.665-686); Ch.21.8, (p.736-740).
- Hosmer, D., and S., Lemeshow, "Applied logistic regression", (Wiley: 2000), 2th Ed., Ch.5.2.2, p.147-149.
- Ronning, G., "Mikroökonometrie", (Springer: 1991), Kap.2.1.1-2, (S.29-38), Kap.2.2.1-2, (S.44-51), Kap.2.3-2.4, (S.55-61).

5 Modelle für quantitativ und begrenzt abhängige Variablen

5.1 Modelle für Zähldaten

5.1.1 Poisson-Modell

Die abhängige Variable Y in einem Modell für Zähldaten kann nur die natürlichen Zahlen 0,1,2,3,... annehmen.

Zur Beschreibung der Daten eignen sich mehrere Verteilungen. Die bekannteste, sowie die einfachste, ist die Poissonverteilung

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0, 1, 2, ...$$

Dies ist eine Einparameterverteilung mit $\lambda > 0$. Die Poissonverteilung hat der Erwartungswert $E(Y) = \lambda$ und die Varianz $Var(Y) = \lambda$.

Da der Parameter λ, der gleichzeitig dem Erwartungswert gleich ist, immer positiv sein muss, erfolgt das übliche Bedingen auf die beobachtete Charakteristiken mit Hilfe einer exponentialen Funktion

$$E(y_i|\mathbf{x}_i) = \lambda_i = \exp\{\mathbf{x}_i'\beta\}.$$

Somit die bedingte Dichte ist

$$g(y_i|\mathbf{x}_i) = \frac{e^{-\exp\{\mathbf{x}_i'\beta\}}(\exp\{\mathbf{x}_i'\beta\})^{y_i}}{y_i!}.$$

Das ergibt die Likelihoodfunktion

$$\mathcal{L} = \prod_{i=1}^{n} \frac{e^{-\exp\{\mathbf{x}_i'\beta\}}\exp\{y_i(\mathbf{x}_i'\beta)\}}{y_i!}.$$

- Ökonomischer Inhalt des Modells

Das Modell beschreibt die Anzahl an Ereignisse, was auch immer das Ereignis sein mag. Der offensichtliche Anwendung deswegen ist bei der Analyse der Nachfrage nach einem bestimmten Gut oder einer bestimmten Dienstleistung – einmalig oder während der gegebenen Zeitperiode. Einmalig: Anzahl an gekauften Flaschen Wein

in Abhängigkeit von der Promo-Aktionen des Weinladens (z.B., wie wirkt das Angebot: "Fürs Kauf über EUR 50 - ein Gutschein für EUR 5!", wenn überhaupt?). Während der gegebenen Zeitperiode: Anzahl an Arztbesuche (z.B., Signifikanz individuellen Gesundheitsvorstellungen gegeben die gleiche Gesundheitsversicherung).

Das Modell hat auch eine "Zeitinterpretation", denn die Dauer zwischen zwei nacheinanderfolgenden Ereignissen eines Poisson Prozesses ist eine Zufallsvariable, sagen wir S_i, die eine *Exponentialverteilung*

$$f(s) = \lambda_i e^{-\lambda_i s} \quad \text{bzw.} \quad F(s) = \int_0^s \lambda_i e^{-\lambda_i x} dx = 1 - e^{-\lambda_i s},$$

mit dem gleichen Poissonparameter λ_i folgt. Dies ist für die Anwendungen innerhalb einer Zeitperiode relevant. Der Erwartungswert der Wartezeit bis zum nächsten Ereignis ist $E(S) = 1/\lambda_i$. Somit, wenn man λ_i weiß, dann weiß man auch wie lange muss man im Schnitt warten bis das Ereignis wieder eintritt.

- Die Likelihoodfunktion

$$\mathcal{L} = \prod_{i=1}^{n} \frac{e^{-\exp\{\mathbf{x}_i'\beta\}} \exp\{y_i (\mathbf{x}_i'\beta)\}}{y_i!}.$$

- Die log-Likelihoodfunktion und B.E.O.

$$\begin{aligned} \ln \mathcal{L} &= \sum_{i=1}^{n} \ln \left(\frac{e^{-\exp\{\mathbf{x}_i'\beta\}} \exp\{y_i (\mathbf{x}_i'\beta)\}}{y_i!} \right) \\ &= \sum_{i=1}^{n} \left[\ln \left(e^{-\exp\{\mathbf{x}_i'\beta\}} \exp\{y_i (\mathbf{x}_i'\beta)\} \right) - \ln (y_i!) \right] \end{aligned}$$

und schließlich

$$\ln \mathcal{L} = \sum_{i=1}^{n} \left[-e^{\mathbf{x}_i'\beta} + y_i (\mathbf{x}_i'\beta) - \ln (y_i!) \right].$$

Daraus resultieren die B.E.O.

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left[-e^{\mathbf{x}_i'\beta} + y_i (\mathbf{x}_i'\beta) \right] = \sum_{i=1}^{n} \left[-e^{\mathbf{x}_i'\beta} + y_i \right] \mathbf{x}_i \stackrel{!}{=} \mathbf{0}$$

Die Hesse-Matrix in diesem Modell hat eine besonders einfache Form

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta'} = \frac{\partial}{\partial \beta} \left(\sum_{i=1}^{n} \left[-e^{\mathbf{x}_i'\beta} + y_i \right] \mathbf{x}_i \right) = - \sum_{i=1}^{n} \underbrace{e^{\mathbf{x}_i'\beta}}_{=\lambda_i} \mathbf{x}_i \mathbf{x}_i' = - \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i'.$$

Da die Hesse-Matrix keine Abhängige Variable in sich beinhaltet, ist die Informationsmatrix der negativen Hesse-Matrix gleich

$$\mathbf{I}(\beta) = -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'}\right) = -E\left(-\sum_{i=1}^{n}\lambda_i \mathbf{x}_i\mathbf{x}_i'\right) = \sum_{i=1}^{n}\lambda_i \mathbf{x}_i\mathbf{x}_i'.$$

- Identifizierbarkeit / Eindeutigkeit des Schätzers

Genauso wie im Logit-Modell früher, können wir sehen, daß die Hesse-Matrix negativ definit ist, denn

$$\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}\lambda_i \mathbf{x}_i\mathbf{x}_i' = -\mathbf{X'VX},$$

wobei $\mathbf{V}$ eine Diagonalmatrix mit der positiven Hauptdiagonal ist. Dadurch, solange $\mathbf{X'X}$ nichtsingulär ist, ist die Likelihoodfunktion global konkav in β, und somit gibt es nur ein Maximum. D.h., der ML-Schätzer in diesem Modell ist eindeutig. Es gibt weiterhin keine Normierungen die die Identifizierbarkeit beeinträchtigen können.

- Interpretation der Schätzergebnisse

Partielle Effekte im Poisson-Modell, ähnlich zu den früheren Modellen, sind

$$\frac{\partial E(y_i|\mathbf{x}_i)}{\partial x_j} = \frac{\partial\left(e^{\mathbf{x}_i'\beta}\right)}{\partial x_j} = \beta_j e^{\mathbf{x}_i'\beta} = \beta_j E(y_i|\mathbf{x}_i).$$

Die relative Wahrscheinlichkeiten im Poisson-Modell lassen sich allgemein durch

$$\frac{P(Y=y+1)}{P(Y=y)} = \frac{e^{-\lambda}\lambda^{y+1}}{(y+1)!}\cdot\frac{y!}{e^{-\lambda}\lambda^{y}} = \frac{\lambda^{y}\lambda}{(y+1)\,y!}\cdot\frac{y!}{\lambda^{y}} = \frac{\lambda}{y+1}$$

ausdrücken. Somit, für die bedingte relative Wahrscheinlichkeiten gilt

$$\frac{P(Y_i=y_i+1|\mathbf{x}_i)}{P(Y_i=y_i|\mathbf{x}_i)} = \frac{e^{\mathbf{x}_i'\beta}}{y_i+1}.$$

Leitet man dies nach x_j, dann sieht man wie schnell steigt / wie langsam sinkt die Wahrscheinlichkeit des nächsten Ereignisses, wenn x_j sich ändert. Nämlich

$$\frac{\partial}{\partial x_j}\left(\frac{P(Y_i=y_i+1|\mathbf{x}_i)}{P(Y_i=y_i|\mathbf{x}_i)}\right) = \frac{\partial}{\partial x_j}\left(\frac{e^{\mathbf{x}_i'\beta}}{y_i+1}\right) = \beta_j\frac{e^{\mathbf{x}_i'\beta}}{y_i+1} = \beta_j\frac{P(Y_i=y_i+1|\mathbf{x}_i)}{P(Y_i=y_i|\mathbf{x}_i)}.$$

Schließlich, der partielle Effekt für die erwartete Dauer zwischen zwei Ereignissen ist

$$\frac{\partial E(s_i|\mathbf{x}_i)}{\partial x_j} = \frac{\partial(1/\lambda_i)}{\partial x_j} = \frac{\partial\left(e^{-\mathbf{x}_i'\beta}\right)}{\partial x_j} = -\beta_j e^{-\mathbf{x}_i'\beta} = -\frac{\beta_j}{E(y_i|\mathbf{x}_i)}.$$

5.1.2 Zero-Inflated und Hurdle Poisson-Modelle

Das sind zwei Erweiterungen des ursprünglichen Modells, die weitere Dateneigenschaften mitberücksichtigen.

- Zero-Inflated Poisson-Modell

Anzahl an gekauften Güter oder Dienstleistungen kann an sich Poissonverteilt sein. In der Stichprobe, jedoch, können wir zwei typen von Menschen haben: die, die das Gut eventuell kaufen und die, die das nie machen. Dadurch beobachtet man viel mehr "0"-Ausprägungen, als im üblichen Poisson sein sollte ("0" bei denen, die in der gegebenen Periode sich entschieden haben das Gut nicht zu kaufen, und "0" bei denen, die das nie machen). Das Zero-Inflated Poissonregression modelliert die überflüssige "0"-Ausprägungen explizit und damit vermeidet Verzerrungen durch die Fehlspezifikation.

Die Wahrscheinlichkeit, daß ein bestimmtes Ereignis für das Individuum ausgeschlossen ist, sei durch φ gegeben. Ansonsten, beleibe die Anzahl der Ereignisse Poissonverteilt. Dann, die Wahrscheinlichkeit "0" zu beobachten ist

$$P\left(Y=0|\mathbf{x}_i\right)=\varphi+\left(1-\varphi\right)\frac{e^{-\exp\{\mathbf{x}_i'\beta\}}\left(\exp\{\mathbf{x}_i'\beta\}\right)^0}{0!}=\varphi+\left(1-\varphi\right)e^{-\exp\{\mathbf{x}_i'\beta\}}.$$

Für die weitere Ausprägungen bleiben die Poissonwahrscheinlichkeiten unverändert

$$P\left(Y=y|\mathbf{x}_i\right)=\frac{e^{-\exp\{\mathbf{x}_i'\beta\}}\left(\exp\{\mathbf{x}_i'\beta\}\right)^{y_i}}{y_i!},\qquad y=1,2,...$$

Erwartungswert und Varianz in dieser Verteilung sind durch

$$E\left(Y\right)=e^{\mathbf{x}_i'\beta}\left(1-\varphi\right)\quad\text{und}\quad Var\left(Y\right)=\left(1-\varphi\right)e^{\mathbf{x}_i'\beta}\left[1+\varphi e^{\mathbf{x}_i'\beta}\right]$$

gegeben. Da $Var\left(Y\right)>E\left(Y\right)$, besitzt dieses Modell die restriktive Eigenschaft der "Gelichdispersion" $[E\left(Y\right)=Var\left(Y\right)]$ des Basismodells nicht mehr.

Definiert man die Indexfunktion $I_{y=0}$, so daß $I_{y=0}=1$ wenn $y=0$ und $I_{y=0}=0$ sonst, dann bekommt man die Likelihoodfunktion

$$\mathcal{L}=\prod_{i=1}^{n}\frac{\varphi^{I_{y=0}}+\left(1-\varphi\right)^{I_{y=0}}e^{-\exp\{\mathbf{x}_i'\beta\}}\exp\{y_i\left(\mathbf{x}_i'\beta\right)\}}{y_i!}$$

Die log-Likelihood, somit, ist

$$\ln\mathcal{L}=\sum_{i=1}^{n}\left[\ln\left(\varphi^{I_{y=0}}+\left(1-\varphi\right)^{I_{y=0}}e^{-\exp\{\mathbf{x}_i'\beta\}}\exp\{y_i\left(\mathbf{x}_i'\beta\right)\}\right)-\ln\left(y_i!\right)\right].$$

Der ML-Schätzer in diesem Modell ist ohne zusätzlichen Identifizierbarkeitsrestriktionen eindeutig. Letztlich, die Wahrscheinlichkeit φ kann selbst parametrisiert werden. Eine normierte logistische Wahrscheinlichkeitsfunktion ist die übliche Wahl.

- Hurdle Poisson-Modell

In dieser Erweitering des Basismodells wird angenommen, daß die Verteilung sich ab einem Bestimmten Anzahl der Ereignisse (hurdle) ändert. Definieren wir diese Schwellenanzahl an Ereignisse als y_H. Darüber hinaus nehmen wir an, daß für $y \leq y_H$ folgen die Ereignisse eine Poissonverteilung mit parameter $\lambda_{1,i}$, wobei für $y > y_H$ greift die Poissonverteilung mit parameter $\lambda_{2,i}$. Dann

$$P(Y=y) = \begin{cases} P(Y=y|\lambda_{1,i}), & \text{wenn } y \leq y_H. \\ [1-P(Y \leq y_H|\lambda_{1,i})] \frac{P(Y=y|\lambda_{2,i})}{1-P(Y \leq y_H|\lambda_{2,i})}, & \text{wenn } y > y_H. \end{cases}$$

Die $\lambda_{1,i}$ und $\lambda_{2,i}$ werden, wie früher, als $\lambda_{1,i} = \exp\{\mathbf{x}_i'\beta_1\}$ und $\lambda_{2,i} = \exp\{\mathbf{x}_i'\beta_2\}$ parametriesiert.

Üblicherweise nimmt man an, daß die Schwelle bei $y_H = 0$ liegt. Unter dieser Annahme hat das Modell eine Interpretation der Entscheidung ($y = 0$ vs. $y > 0$) und der Folgen dieser Entscheidung ($y = 1, 2, ...$). Ein klassisches Beispiel ist die Anzahl der Arztbesuche. Der Patient Entschiedet darüber ob er überhaupt zum Arzt gehen muss ($y = 0$ vs. $y > 0$). Dies ist nur seine eigene Entscheidung und somit wird vollständig durch $\lambda_{1,i}$ beschrieben. Falls es einschieden wurde den Arzt zu besuchen, die Anzahl der weiteren Besuchen wird nicht nur vom Patienten, sondern auch vom Arzt zusammenbestimmt. Deswegen die Verteilung der Ereignissen für $y > 0$ ist mit $\lambda_{2,i} \neq \lambda_{1,i}$ determiniert.

Unter der Annahme $y_H = 0$ ist das Hurdle Poisson-Modell:

$$P(Y=y) = \begin{cases} e^{-\exp\{\mathbf{x}_i'\beta_1\}}, & \text{wenn } y = 0. \\ \frac{1-\exp\{-\exp\{\mathbf{x}_i'\beta_1\}\}}{1-\exp\{-\exp\{\mathbf{x}_i'\beta_2\}\}} \left[e^{-\exp\{\mathbf{x}_i'\beta_2\}} \frac{\left(\exp\{\mathbf{x}_i'\beta_2\}\right)^{y_i}}{y_i!} \right], & \text{wenn } y > 0. \end{cases}$$

Die Likelihoodfunktion ist

$$\mathcal{L} = \prod_{i \in \{y_i=0\}} e^{-\exp\{\mathbf{x}_i'\beta_1\}} \prod_{i \in \{y_i>0\}} \frac{1-\exp\{-\exp\{\mathbf{x}_i'\beta_1\}\}}{1-\exp\{-\exp\{\mathbf{x}_i'\beta_2\}\}} \left[\frac{e^{-\exp\{\mathbf{x}_i'\beta_2\}} \exp\{y_i(\mathbf{x}_i'\beta_2)\}}{y_i!} \right]$$

Die log-Likelihoodfunktion ist dann

$$\begin{aligned}\ln \mathcal{L} &= \underbrace{-\sum_{i\in\{y_i=0\}} \exp\{\mathbf{x}_i'\beta_1\} + \sum_{i\in\{y_i>0\}} \ln\left(1-\exp\{-\exp\{\mathbf{x}_i'\beta_1\}\}\right)}_{\equiv \ln \mathcal{L}_1(\beta_1)} \\ &\quad \underbrace{-\sum_{i\in\{y_i>0\}} \ln\left(1-\exp\{-\exp\{\mathbf{x}_i'\beta_2\}\}\right) + \sum_{i\in\{y_i>0\}} \left[-e^{\mathbf{x}_i'\beta_2} + y_i\left(\mathbf{x}_i'\beta_2\right) - \ln\left(y_i!\right)\right]}_{\equiv \ln \mathcal{L}_2(\beta_2)} \\ &= \ln \mathcal{L}_1\left(\beta_1\right) + \ln \mathcal{L}_2\left(\beta_2\right)\end{aligned}$$

Wie bei allen Hurdle-Modellen, teilt sich der Parmeterraum $\mathbf{B}$ ($\beta \in \mathbf{B}$) in zwei Teilen ($\mathbf{B}_1, \mathbf{B}_2 : \mathbf{B}_1 \cap \mathbf{B}_2 = \varnothing$) auf. D.h., wenn man $\ln \mathcal{L}$ maximieren möchte, kann man $\ln \mathcal{L}_1\left(\beta_1\right)$ und $\ln \mathcal{L}_2\left(\beta_2\right)$ getrennt voneinander maximieren.

Der Erwartungswert im Hurdle Poisson-Modell mit $y_H = 0$ ist

$$E\left(Y\right) = P\left(y > 0; \beta_1\right) E\left(Y | y > 0; \beta_2\right) = \frac{1 - e^{-\exp\{\mathbf{x}_i'\beta_1\}}}{1 - e^{-\exp\{\mathbf{x}_i'\beta_2\}}} e^{\mathbf{x}_i'\beta_2}.$$

Die Varianz ist auch bekannt.

- Gütemaße

Für Poisson sowie Zero-Inflated und Hurdle Poisson-Modelle gelten die pseudo-R^2 , das sie sich ausschließlich auf die Likelihoodfunktionswerten basieren.

Ein zusätzliches Maß der Anpassungsgüte für Poisson-Modell, sowie seine Erweiterungen oben, dient die Pearson-Statistik. Im allgemeinen lässt sich diese Statistik folgender maßen darstellen. Definieren wir als $\hat{\mu}_i$ den vorhergesagten Mittelwert und als $\hat{\omega}_i$ die vorhergesagte Varianz. Dann ist die Pearson-Statistik

$$P = \sum_{j=1}^{n} \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\omega}_i} \simeq \chi^2_{n-k},$$

wobei n die Anzahl der Beobachtungen und k die Anzahl der Modellparameter bezeichnen. Die Nullhypothese dabei ist wie im Hosmer-Lemeshow Test. Außerdem, für $J \to n$ und n groß genug, ist die HL_J der Pearson-Statistik fast identisch. Die Verteilung der Pearson-Statistik ist wieder eine Approximation, denn n endlich sein muss, wobei χ^2 sich nur asymptotisch ergibt. Schließlich, um Pearson-Statistik berechnen zu können müssen wir immer die funktionale Form der $\hat{\mu}_i$ und $\hat{\omega}_i$ wissen. Im Fall eines einfaches Poisson Modells ist der Mittelwert immer der Varianz

gleich. Somit die Funktional formen sind jeweils $\hat{\mu}_i = \exp\{\mathbf{x}_i'\hat{\beta}\}$ und $\hat{\omega}_i = \exp\{\mathbf{x}_i'\hat{\beta}\}$. Für Zero-Inflated und Hurdle Poisson-Modelle sind die funktionale forme des Mittelwertes und der Varianz ebenso bekannt. Ihre Ausdrücke sind allerdings relativ kompliziert.

Ein χ^2-Test der Anpassungsgüte, der, ähnlich zum Hosmer-Lemeshow Test, den Unterschied zwischen den vorhergesagten und tatsächlichen relativen Häufigkeiten als Konstruktionsprinzip verwendet, steht auch zur Verfügung (Andrews, 1988). Der Test von Andrews (1988) bezieht sich auf ein allgemeines Modell und gilt somit für alle Modelle in diesem Skript ohne Ausnahme (und weit außerhalb dieses Skriptes).

5.2 Modelle für begrenzt abhängige Variable

5.2.1 Tobit Modell

Tobit-Konzept stellt ein Modell für die begrenzt beobachtbare abhängige Variable dar. Es kann mit dem folgenden Beispiel veranschaulicht werden. Nehmen wir an es gibt die übliche Regression der latenten Variable y_i^* auf x_i.

$$y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad E(\varepsilon_i) = 0,\ \ Var(\varepsilon_i) = \sigma^2.$$

Die latente Variable kann jedoch nur dann beobachtet werden, wenn sie eine Schwelle $y^* = C$ überschreitet. Ansonsten beobachtet man nur den Schwellenwert. Dieses Beobachtungsmechanismus heißt *Zensierung*. Es lässt sich als

$$y_i = \begin{cases} y_i^*, & \text{wenn } y_i^* > C \\ C, & \text{wenn } y_i^* \leq C \end{cases}$$

darstellen, wobei y_i eine beobachtete Variable ist. Für beispielsweise $C = 0$ bekommen wir das folgende Bild (siehe Abbildung 4).

Hierbei stellt die gestrichelte Gerade den wahren Regressionszusammenhang dar, während die durchgehende Gerade die geschätzte Regressionsgerade mittels KQ repräsentiert. Man erkennt, daß KQ eine inkonsistente Schätzung von β_1 und β_2 impliziert, wenn die wahre Struktur der latenten Variable vernachlässigt wird. Tobit-Modelle beseitigen dieser Nachteil.

Eine Schwelle kann nicht nur unten sondern auch oben liegen. In diesem Fall, für eine Regression für die latente Variable

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i, \qquad E(\varepsilon_i) = 0,\ \ Var(\varepsilon_i) = \sigma^2,$$

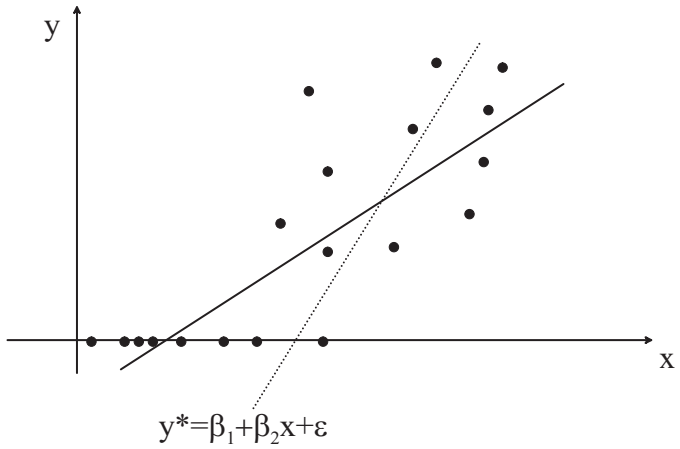


Abbildung 4: Tobit-Modell

verallgemeinert sich das Tobit-Modell zu

$$y_i = \begin{cases} C_2, & \text{wenn } y_i^* > C_2 \\ y_i^*, & \text{wenn } C_1 < y_i^* \leq C_2 \\ C_1, & \text{wenn } y_i^* \leq C_1 \end{cases}$$

Unterstellt man die Normalverteilung für den Störterm, $\varepsilon_i \sim N(0, \sigma^2)$, dann sind die bedingte Wahrscheinlichtkeiten $P(y_i^* \leq C_1)$ und $P(y_i^* > C_2)$

$$\begin{aligned} P(y_i^* \leq C_1|\mathbf{x}_i) &= P(\mathbf{x}_i'\beta + \varepsilon_i \leq C_1) = P(\varepsilon_i \leq C_1 - \mathbf{x}_i'\beta) = \Phi\left(\frac{C_1 - \mathbf{x}_i'\beta}{\sigma}\right), \\ P(y_i^* > C_2|\mathbf{x}_i) &= 1 - P(y_i^* \leq C_2|\mathbf{x}_i) = 1 - P(\varepsilon_i \leq C_2 - \mathbf{x}_i'\beta) \\ &= 1 - \Phi\left(\frac{C_2 - \mathbf{x}_i'\beta}{\sigma}\right) = \Phi\left(-\frac{C_2 - \mathbf{x}_i'\beta}{\sigma}\right). \end{aligned}$$

Mit Hilfe zwei Dummyvariablen $d_{1,i} = 1$, wenn $y_i^* \leq C_1$ ($d_{1,i} = 0$ sonst) und $d_{2,i} = 1$, wenn $y_i^* > C_2$ ($d_{2,i} = 0$ sonst) schreibt man die Likelihoodfunktion für dieses Modell als

$$\mathcal{L} = \prod_{i=1}^{n} \Phi\left(\frac{C_1 - \mathbf{x}_i'\beta}{\sigma}\right)^{d_{1,i}} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\beta}{\sigma}\right)\right]^{1-(d_{1,i}+d_{2,i})} \Phi\left(-\frac{C_2 - \mathbf{x}_i'\beta}{\sigma}\right)^{d_{2,i}}.$$

hin. Liegt keine Zensierung von unten oder oben vor, dann wird jeweils $d_{1,i} = 0$ oder $d_{2,i} = 0$ gesetzt. Liegt keine Zensierung vor, dann verschinden die $\Phi(\cdot)$-Ausdrücke und die Likelihoodfunktion verkleinert sich zu der Likelihoodfunktion der multiplen linearen Regression mit dem Normalverteilten Störterm. Schließlich, wird keine y_i^*

beobachtet, dann sind $C_1 = C_2 = C$ gleich. Für $C = 0$ ist das nichts anderes als ein übliches Probit-Modell. Das Tobit-Modell liegt somit zwischen der multiplen lienaren Regression, wo alle y_i^* beobachtet werden und Probit Regression, wo keine y_i^* beobachtet wird.

- Ökonomischer Inhalt des Modells

Das Modell wird für alle mögliche unvollständig (begrenzt) beobachtete abhängige Variablen verwendet. Alle Daten, die aus Administrativen Quellen stammen, werden in der Regel unvollständig beobachtet, denn die Erhebung dieser Daten wird immer von einem Gesetz und für einen bestimmten Zweck des Behörde durgeführt, im Gegensatz zu den Daten aus Umfragen, wo wir frei sind sie so auszuwahlen, damit sie für unsere Schätzmethoden am besten geeignet wären. Typischer Beispiel: Löhne unter Betragsbemessungsgrenze. Administrative Daten sind viel genauer und beinhalten häufig Information, die ansonsten nie vorhanden ist.

Auch in der Umfragen gibt es viel Daten unvollkommenheiten. Typischer Bespiel: Anzahl gearbeiteten Stunden in einer Arbeitsangebotregression (die Umfragen sagen, in der Regel: a) unter 10, b) genaue Zahl, c) über 45).

- Unterschied zwischen Zensierung und Stutzung

Für die Veranschaulichkeit der nachfolgenden Diskussion setzen wir $C_1 = 0$ und $d_{2,i} = 0$ um ein Modell nur mit Zensierung von unten bei der Schwellenwert $y_i^* = 0$ zu bekommen

$$\mathcal{L} = \prod_{i=1}^{n} \Phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)^{d_i} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\beta}{\sigma}\right)\right]^{1-d_i},$$

wobei d_i nun anstatt $d_{1,i}$ steht. Dies ist eine übliche Lehrbuchdarstellung des Tobit-Modells mit Zensierung.

Im zensierten Modell beobachten wir entweder $(0, \mathbf{x}_i)$ bei den Individuen, dessen y_i^* unter der Null-Schwelle liegt, oder $(y_i^*, \mathbf{x}_i)$ bei den Individuen, dessen y_i^* über die Null-Schwelle liegt. In diesem Modell, somit, geht nur die Information über y_i^* verloren.

Nun nehmen wir an, daß für alle Individuen, die $y_i = 0$ haben, können wir weder y_i^* noch $\mathbf{x}_i$ beobachten, dann soll der $\Phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)$ Teil im beitrag oben verschwinden, denn wir keine Information mehr über $\mathbf{x}_i$ haben. Allerdings, wir wissen immer noch,

daß alle beobachtete Werte der latenten Variable positiv sind, obwohl die latente variable an sich nicht immer positiv sein muss. Diese Beobachtungen kommen also, aus einer bedingten Dichte, wo die Bedingung $P(y^* > 0)$ ist. Für Normalverteilung schreibt man diese Dichte wie

$$f\left(y|y>0\right) = \frac{f\left(y\right)}{P(y>0)} = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mathbf{x}'\beta}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}'\beta}{\sigma}\right)}$$

hin. Sie heißt auch die gestutzte Dichte mit Stutzung am $y = 0$. Die Abbildung auf der nächsten Seite veranschaulicht den Unterschied zwischen der zensierten und gestutzten Dichtefunktionen.

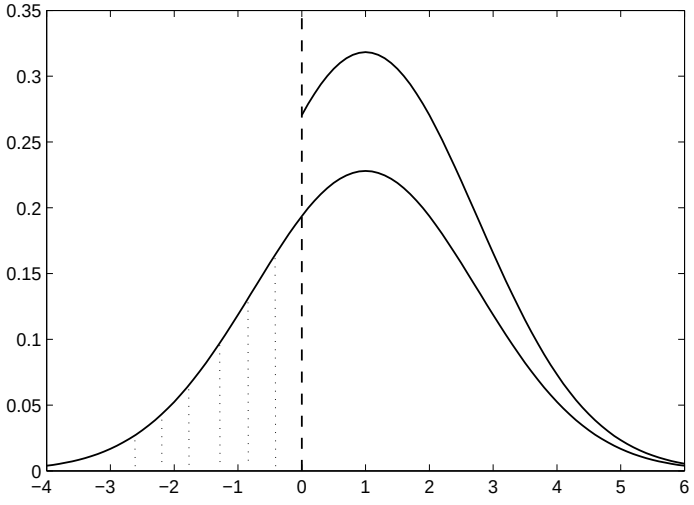
Somit, ist die Likelihoodfunktion unter Stutzung

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\beta}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)}$$

was der Likelihoodfunktion im zensierten Modell unterschiedlich ist.

Im allgemeinen Fall mit der Stutzung von unten und oben bei jeweils C_1 und C_2 bekommt man

$$P\left(C_1 < y^* \leq C_2\right) = P\left(C_2 \leq y^*\right) - P\left(C_1 \leq y^*\right) = \Phi\left(\frac{C_2 - \mathbf{x}_i'\beta}{\sigma}\right) - \Phi\left(\frac{C_1 - \mathbf{x}_i'\beta}{\sigma}\right)$$



Mit diesem Ergebnis ist die Likelihoodfunktion im allgemeineren Fall der Stutzung

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\beta}{\sigma}\right)}{\Phi\left(\frac{C_2-\mathbf{x}_i'\beta}{\sigma}\right) - \Phi\left(\frac{C_1-\mathbf{x}_i'\beta}{\sigma}\right)}.$$

- Identifizierbarkeit im Tobit-Modell

Sowohl im Fall Zensierung als auch im Fall Stutzung sehen die B.E.O. und die Ausdrücke für Informationsmatrix relativ unbequem aus. Rein intuitiv, allerdings, das Tobit-Modell mit Zensierung ist wegen der Identifizierbarkeit seinen zwei Bestandteilen - des Probit-Modells und der multiplen lineareren Regression mit Normailverteilten Störtermen - immer Identifizierbar. Man kann auch bemerken, daß man wegen des zweiten Bestandteils die Probit Normierungen $\mu = 0$ und $\sigma = 1$ nicht mehr braucht. Das gestutzte Modell ist immer identifizierbar, wenn das ursprüngliche Modell identifizierbar ist und der Stutzunkt ist keine Funktion der Modellparameter.

Somit ist ML Schätzer im Tobit Modell eindeutig.

- Partielle Effekte im Tobit-Modell

Partielle Effekt auf die latente Variable y_i^* ist nichts anderes als $\partial E(y_i^*|\mathbf{x}_i)/\partial x_j = \beta_j$. Außerdem, das ist der genaue Effekt (z.B. im Gegensatz zu Probit), denn die latente Regression wird nie mit σ Skaliert.

Partielle Effekt auf die Beobachtete Variable y_i unterscheidet sich zwischen Zensierung und Stutzung, ist mehr kompliziert und weniger interessant.

5.2.2 (Nicht)Konsistenz der KQ Schätzung im Tobit-Modell

- KQ-Schätzung

Im klassischen Regressionsmodell ist $E(y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta + 0$ und daraus

$$E(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{X}\beta + \varepsilon) = \beta + \mathbf{0} = \beta.$$

Im Tobit-Modell, wie oben gesehen, ist KQ-Schätzung mit allen Beobachtungen verzerrt.

Werden nun nur die y_i^* Beobachtungen berücksichtigt, die $y_i = 0$ Beobachtungen hingegen nicht, bleibt die Verzerrung trotzdem. Wir haben, nämlich

$$E(y_i|y_i^* > 0) = E(y_i^*|\ y_i^* > 0) = E(\mathbf{x}_i'\beta + \varepsilon_i|\ y_i^* > 0) = \mathbf{x}_i'\beta + \underbrace{E(\varepsilon_i|\ y_i^* > 0)}_{=???}$$

Um die Verzerrung zu sehen, nehmen wir, wie immer, an, daß $\varepsilon \sim N(0,\sigma^2)$ ist. Dann

$$\begin{aligned} E(\varepsilon_i|\ y^* &> 0) = E(\varepsilon_i|\mathbf{x}_i'\beta + \varepsilon_i > 0) = E(\varepsilon_i|\varepsilon_i > -\mathbf{x}_i'\beta) \\ &= \sigma E\left(\underbrace{\frac{\varepsilon_i}{\sigma}}_{\equiv u} \mid \underbrace{\frac{\varepsilon_i}{\sigma}}_{\equiv u} > \underbrace{-\frac{\mathbf{x}_i'\beta}{\sigma}}_{\equiv c}\right) = \sigma E\left(u|u>c\right), \end{aligned}$$

wobei $u \sim N(0,1)$ ist, denn $\varepsilon \sim N(0,\sigma^2)$. Nun, wie groß ist $E\left(u|u>c\right)$? Dies ist nichts anderes als der Erwartungswert der gestutzten standartnormalverteilten Variable mit dem Stustzungspunkt c. Laut der Definition des Erwartungswertes haben wir

$$\begin{aligned} E\left(u|u>c\right) &= \int_c^\infty u\frac{\phi\left(u\right)}{1-\Phi\left(c\right)}du \\ &= \frac{1}{1-\Phi\left(c\right)}\int_c^\infty u\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du \\ &= \frac{1}{1-\Phi\left(c\right)}\frac{1}{\sqrt{2\pi}}\int_{c^2/2}^\infty e^{-v}dv \\ &= \frac{1}{1-\Phi\left(c\right)}\frac{1}{\sqrt{2\pi}}\left[-e^{-\infty} - \left(-e^{-\frac{1}{2}c^2}\right)\right] \\ &= \frac{1}{1-\Phi\left(c\right)}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}c^2} \\ &= \frac{\phi\left(c\right)}{1-\Phi\left(c\right)}. \end{aligned}$$

Mit diesem Ergebnis,

$$E(\varepsilon_i|\ y^* > 0) = \sigma E\left(\frac{\varepsilon_i}{\sigma}|\frac{\varepsilon_i}{\sigma} > -\frac{\mathbf{x}_i'\beta}{\sigma}\right) = \sigma\frac{\phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)}{1-\Phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)} \neq 0,$$

woraus folgt, daß

$$E(y_i|y_i^* > 0) = \mathbf{x}_i'\beta + \underbrace{E(\varepsilon_i|\ y_i^* > 0)}_{\neq 0} = \mathbf{x}_i'\beta + \sigma\frac{\phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)}{1-\Phi\left(-\frac{\mathbf{x}_i'\beta}{\sigma}\right)} \neq \mathbf{x}_i'\beta.$$

Somit wird der einfache KQ Schätzer, auf positive Ausprägungen verwendet, auch verzerrt sein.

- Zwei-stufiges Schätzverfahren im Tobit-Modell mit Zensierung

Der Veranschaulichung der nichtkonsistenz des KQ Schätzung impliziert ein einfaches zweistufiges Schätzverfahren fürs Tobit-Modell mit Zensierung. Definieren wir einen neuen Parametervektor $\alpha = \beta/\sigma$. Der Erwartungswert der positiven Ausprägungen, wie oben gezeigt, ist

$$E(y_i|y_i^* > 0) = \mathbf{x}_i'\beta + \sigma \underbrace{\frac{\phi(-\mathbf{x}_i'\alpha)}{1 - \Phi(-\mathbf{x}_i'\alpha)}}_{=\lambda_i} = \mathbf{x}_i'\beta + \sigma\lambda_i.$$

In der Regression für ausschließlich positive Beobachtungen

$$y_i^* = \mathbf{x}_i'\beta + \sigma\lambda_i + \epsilon_i, \qquad y_i^* > 0,$$

haben wir nun tatsächlich $E(\epsilon_i) = 0$, denn $E(y_i|y_i^* > 0) = \mathbf{x}_i'\beta + \sigma\lambda_i$. Wir betrachten somit λ_i als eine zusätzliche erklärende Variable. Hätten wir λ_i gewusst, würden wie sie einfach in die obige Regression einsetzen und β konsistent Schätzen. Da $\lambda_i = \phi(-\mathbf{x}_i'\alpha)/[1 - \Phi(-\mathbf{x}_i'\alpha)]$, das einzige was wir wissen müssen sind die Parameter α, die die Wahrscheinlichkeit einer positiven Ausprägung bestimmen. Dadurch ergibt sich ein 2-Stufiges Verfahren

Stufe 1: Für alle Beobachtungen (zensierte und unzensierte) definieren wir

$$\tilde{y}_i = \begin{cases} 1, & y_i > 0 \\ 0, & y_i = 0 \end{cases}$$

und schätzen ein Probit-Modell mit der abhängigen Variable $\tilde{y}_i$ und erklärenden Variablen $\mathbf{x}_i$. Daraus bekommen wir die Schätzwerte $\hat{\alpha}$. Mit ihren Hilfe sagen wir $\hat{\lambda}_i = \phi(-\mathbf{x}_i'\hat{\alpha})/[1 - \Phi(-\mathbf{x}_i'\hat{\alpha})]$ vorher.

Stufe 2: Mit $\hat{\lambda}_i$ schätzen wir mittels KQ die Regressionsgleichung

$$y_i^* = \mathbf{x}_i'\beta + \sigma\hat{\lambda}_i + \epsilon_i.$$

Die Schätzwerte $\hat{\beta}$ und $\hat{\sigma}$ aus dieser Gleichung sind die konsistente Schätzungen der wahren β und σ.

Es kann noch weiter gezeigt werden, das der Störterm in der Schätzgleichung heteroskedastisch ist mit

$$Var(\epsilon_i) = \sigma^2 \left[1 - \lambda_i \left(\frac{\mathbf{x}_i'\beta}{\sigma} + \lambda_i\right)\right].$$

Außerdem, wird der vorausgesagte $\hat{\lambda}_i$ als eine fixe Größe betrachtet, wobei in Wirklichkeit hat diese Variable ihre eingene Varianz, die durch die Varianz von $\hat{\alpha}$ verursacht worden ist. Die Heteroskedastizität lässt sich durch Anwendung der verallgemeinerten KQ Schätzers beseitigen. Die Varianz von $\hat{\lambda}_i$ kann auch mitberücksichtigt werden, aber das ist ziemlich umständlich.[1] Aufgrund dieser zwei Komplikationen wird immer die ML Schätzung bevorzugt. Das 2-Stufige Verfahren kann allerdings in komplizierteren Modellen, wo ML Schätzung nicht möglich ist, eine gute Alternative darstellen.

5.3 Selktionsmodelle

5.3.1 Heckman-Modell

Bisher haben wir immer implizit angenommen, daß die Stichprobe, die die sämtliche Information über den Zusammenhang zwischen y_i und $\mathbf{x}_i$ in sich beinhaltet, repräsentativ ist. Dies bedeutet, daß der Zusammenhang zwischen y_i und $\mathbf{x}_i$, der mit Hilfe dieser Stichprobe quantifiziert werden kann, ist sofort als der Zusammenhang in der Gesamtbevölkerung interpretierbar. Nun, wenn wir plötzlich herausfinden, daß unsere Stichprobe nicht mehr auf einer zufällingen Art und Weise erhoben wurde, sondern ihre Erhebung von einem unbekannten Selektionsmechanismus beeinflusst worden war, wäre es trotzdem immer noch möglich den wahren Zusammenhang zu schätzen? Der Heckman-Ansatz setzt sich mit einem solchen Schätzproblem auseinander.

Man nimmt, wie immer, an es existiere eine lineare Regression

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i, \quad y_i \in (-\infty, +\infty)\,.$$

Allerdings, die Ausprägungen y_i können nur dann beobachtet werden, wenn sie in die Stichprobe zugelassen sind. Ob eine Beobachtung i in die Stichprobe zugelassen ist, wird mit Hilfe einer Selektionsbedingung bestimmt. Es existiere also die zweite Regressionsgleichung

$$y_{is} = \mathbf{z}_i'\gamma + \varepsilon_{si}, \quad y_{is} \in (-\infty, +\infty)\,.$$

wobei, wenn y_{is} eine bestimmte Schwelle überschreitet, dann wird für das Individuum i die Ausprädung der ersten Variable, nämlich y_i, beobachtet. Sonst, wenn die

[1] Murphy, K. and R., Topel, "Estimation and inference in two step econometric model", *Journal of Business and Economic Statistics*, 1985, Vol. 3(4), p.370-379.

y_{is} eines Individuums unter dieser Schwelle liegt, haben wir keine Information über y_i. Gegeben daß $\mathbf{z}_i$ in sich einen Absolutglied beinhaltet, ist diese Schwelle, ohne Verlust der Allgemeinheit, dem Null gleich. Somit haben wir die folgende Stichprobeninformation

$$\begin{aligned} \{y_i, \mathbf{x}_i, \mathbf{z}_i\}, &\quad \text{wenn } y_{is} > 0. \quad i = 1, ..., n_1. \\ \{\varnothing, \varnothing, \mathbf{z}_i\}, &\quad \text{wenn } y_{is} \leq 0. \quad i = 1, ..., n_2. \end{aligned}$$

Die genaue Werte der Selektionsvariable ($y_{is}\ \forall\, i$) bleiben dabei immer nichtbeobachtbar.

Die beschriebene Abhängigkeit zwischen der Zielvariable y_i und Selektionsvariable y_{is} impliziert, daß diese beide Variablen von einer bivariaten Verteilung beschrieben sind. Daraus folgt, daß die Störterme $\{\varepsilon_i, \varepsilon_{is}\}$ auch einer bivariaten Verteilung folgen. Wir nehmen, wie immer, eine Normalverteilung an. Dann

$$\begin{bmatrix} \varepsilon_i \\ \varepsilon_{is} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma\sigma_s \\ \rho\sigma\sigma_s & \sigma_s^2 \end{bmatrix} \right),$$

wobei $-1 < \rho < 1$ ist der Korrelationskoeffizient der bivariaten Normalverteilung. Es ist bekannt, daß wenn zwei Zufallsvariablen X und Y der bivariaten Normalverteilung folgen, dann sind ihre bedingte Verteilungen, sowie die Randverteilungen auch normal, nämlich

$$Y|X \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left[x - \mu_X\right], \left(1-\rho^2\right)\sigma_Y^2\right) \quad \text{und} \quad Y \sim N\left(\mu_Y, \sigma_Y^2\right).$$

Mit diesen Ergebnissen die bedingte Dichtefunktion der abhängigen Variable in der erhobenen Stichprobe ist

$$f\left(y_i|y_{is}\right) = \frac{1}{\sigma\sqrt{2\pi\left(1-\rho^2\right)}} \exp\left\{ -\frac{1}{2} \frac{\left(y_i - \left(\mathbf{x}_i'\beta + \rho\frac{\sigma}{\sigma_s}\left[y_{is} - \mathbf{z}_i'\gamma\right]\right)\right)^2}{\left(1-\rho^2\right)\sigma^2} \right\}.$$

- KQ-Schätzung

Wenn die Erhebung der Ausprägungen y_i von den jeweiligen Werten y_{is} unabhängig ist, dann sind die Verteilungen von Y und Y_s auch unabhängig, was bedeutet, daß unter diesen Umständen ρ gleich Null ist. Wenn $\rho = 0$, dann verkleinert sich die bedingte Dichte $f\left(y_i|y_{is}\right)$ zur einfachen Dichte der Normalverteilung

$$f\left(y_i|y_{is}, \rho = 0\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\frac{\left(y_i - \mathbf{x}_i'\beta\right)^2}{\sigma^2} \right\} = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\beta}{\sigma}\right).$$

Dies impliziert, daß der Erwartungswert des Störtermes in der Regressionsgleichung $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ bleibt, wie früher,

$$E(\varepsilon_i) = E\left(y_i - \mathbf{x}_i'\beta\right) = E\left(y_i\right) - \mathbf{x}_i'\beta = \mathbf{x}_i'\beta - \mathbf{x}_i'\beta = 0,$$

und somit kann der Parametervektor β mittels KQ konsistent geschätzt werden.

Wenn, allerdings, die Erhebung der Ausprägungen y_i von den jeweiligen Werten y_{is} durch die Bedingung $y_{is} > 0$ beeinflusst wird, müssen wir den Erwartungswert des Störtermes in der Regressionsgleichung $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ unter Bedingung $y_{is} > 0$ betrachten. Dies ergibt sich durch

$$E(\varepsilon_i | y_{is} > 0) = E\left(y_i - \mathbf{x}_i'\beta | y_{is} > 0\right) = E\left(y_i | y_{is} > 0\right) - \mathbf{x}_i'\beta$$

Es kann gezeigt werden, daß der Erwartungswert $E\left(y_i | y_{is} > 0\right)$ ist zu

$$E\left(y_i | y_{is} > 0\right) = \mathbf{x}_i'\beta + \rho\sigma \frac{\phi\left(-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)}{1 - \Phi\left(-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)}$$

gleich ist.[2] Somit

$$E(\varepsilon_i | y_{is} > 0) = \rho\sigma \frac{\phi\left(-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)}{1 - \Phi\left(-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)} \neq 0,$$

was schließlich eine nicht-Konsistenz des KQ Schätzers des wahren Parametervektor β in der selektiv erhobenen Stichprobe bedeutet.

Da wir den Erwartunswert des Störtermes in der selektierten Stichprobe wissen, können wir, wie im Tobit-Analog, die Regressionsgleichung als

$$y_i = \mathbf{x}_i'\beta + \rho\sigma\lambda_i + \epsilon_i$$

umdefinieren, wo $\lambda_i \equiv \phi\left(-\mathbf{z}_i'\gamma/\sigma_s\right) / [1 - \Phi\left(-\mathbf{z}_i'\gamma/\sigma_s\right)]$ ist. In dieser neuen Regression ist es tatsächlich der Fall, daß $E\left(\epsilon_i\right) = 0$ ist, was die KQ Schätzung ermöglicht sobald man die zusätzliche "vernachlässigte" Variable λ_i hat. Da $1 - \Phi\left(-\mathbf{z}_i'\gamma/\sigma_s\right) =$

[2]Die Herleitung ist nicht schwierig, aber ziemlich umständlich, denn man eine bivariate Dichte über y_i und y_{is} integrieren muss. Der Ausgangspunkt ist

$$E\left(y_i | y_{is} > 0\right) = \int_{-\infty}^{+\infty} \int_{0}^{+\infty} y_i \frac{1}{\sigma\sqrt{1-\rho^2}} \phi\left(\frac{y_i - \left(\mathbf{x}_i'\beta + \rho\frac{\sigma}{\sigma_s}\left[y_{is} - \mathbf{z}_i'\gamma\right]\right)}{\sigma\sqrt{1-\rho^2}}\right) \frac{1}{\sigma_s}\phi\left(\frac{y_{is} - \mathbf{z}_i'\gamma}{\sigma_s}\right) dy_{is} dy_i$$

$\Phi\left(\mathbf{z}_i'\gamma/\sigma_s\right) = P\left(Y_{is} > 0\right)$, können wir diese Größe mit Hilfe einer Probit-Regression schätzen und mit den geschätzten Werten λ_i vorhersagen. Anwendung des Probit-Modells, wie üblich, impliziert eine Identifizierbarkeitsrestriktion. $\sigma_s = 1$. Für die Vorhersage der Fehlenden Variable λ_i spielt diese Restriktion, allerdings, keine Rolle, weil γ/σ_s immer zusammen vorkommen.

Daraus ergibt sich ein schon bekanntes 2-Stufiges Schätzverfahren

Stufe 1: Für den gesamten Datensatz definieren wir

$$\tilde{y}_i = \begin{cases} 1, & y_i \text{ - beobachtet } (\Leftrightarrow \quad y_{is} > 0) \\ 0, & y_i \text{ - nicht beobachtet } (\Leftrightarrow \quad y_{is} > 0) \end{cases}$$

und schätzen ein Probit-Modell mit der abhängigen Variable $\tilde{y}_i$ und erklärenden Variablen $\mathbf{z}_i$. Danach verwenden wir die geschätzte Werte $\hat{\gamma}$ (unter Normierung $\sigma_s = 1$) um $\hat{\lambda}_i = \phi\left(-\mathbf{x}_i'\hat{\gamma}\right)/\left[1 - \Phi\left(-\mathbf{x}_i'\hat{\gamma}\right)\right]$ vorherzusagen.

Stufe 2: Mit $\hat{\lambda}_i$ schätzen wir mittels KQ die Regressionsgleichung

$$y_i = \mathbf{x}_i'\beta + \xi\hat{\lambda}_i + \epsilon_i,$$

wobei $\xi \equiv \rho\sigma$. Somit $\hat{\beta}_{KQ}$ und $\hat{\xi}_{KQ}$ sind die konsistente Schätzer von β und $\rho\sigma$. Der Konsistente Schätzer für σ ergibt sich aus der Summe der quadrierten Residuen

$$\hat{\sigma}^2_{KQ} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\left(y_i - \mathbf{x}_i'\hat{\beta} - \hat{\xi}\hat{\lambda}_i\right)^2 + \hat{\xi}^2\hat{\lambda}_i\left(\mathbf{z}_i'\hat{\gamma} + \hat{\lambda}_i\right)\right]$$

Mit diesem Ergebnis, $\hat{\rho}_{KQ} = \hat{\xi}_{KQ}/\hat{\sigma}_{KQ}$.

Die Regression an der zweiten Stufe leidet unter den gleichen Problemen wie die in der 2-Stufigen Schätzung des Tobit-Modells, nämlich der Störterm in der Schätzgleichung ist heteroskedastisch mit

$$Var\left(\epsilon_i\right) = \sigma^2\left[1 - \rho^2\lambda_i\left(\frac{\mathbf{z}_i'\gamma}{\sigma_s} + \lambda_i\right)\right].$$

Außerdem hat $\hat{\lambda}_i$ ihre eigen Varianz dadurch, daß sie mit Hilfe von geschätzten Probit-Koeffizienten vorhergesagt wurde.

- Ökonomischer Inhalt des Modells

Der ökonomische inhalt ist wieder der Inhalt einer linearen Regression. Bei dieser Klasse der Modelle geht es darum, die Verzerungen, die durch eine selektive Stichprobe verursacht werden können, zu vermeiden. Klassisches Beispiel: Lohnregression.

5.3.2 ML Schätzung des Heckman-Modells

Unsere Stichprobe besteht aus zwei Teilen

$$\begin{aligned}
&\{y_i, \mathbf{x}_i, \mathbf{z}_i\}, \quad \text{wenn } y_{is} > 0. \quad i = 1, ..., n_1. \\
&\{\varnothing, \varnothing, \mathbf{z}_i\}, \quad \text{wenn } y_{is} \leq 0. \quad i = 1, ..., n_2.
\end{aligned}$$

Betrachten wir erst den ersten Teil $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}$, wenn $y_{is} > 0$. In dieser Teilstichprobe interessieren wir uns für die Wahrscheinlichkeit y_i zu beobachten wenn die entsprechende y_{is} positiv ist, was genau den Wahrscheinlichkeitsprozeß in der selektiven Stichprobe darstellt. Diese Wahrscheinlichkeit lautet

$$\varphi(y_i) = P(y_{is} > 0) f(y_i | y_{is} > 0),$$

d.h. die Wahrscheinlichkeit, daß die Selektionsvariable positiv ist [also: $P(y_{is} > 0)$] mal die Wahrscheinlichkeit y_i zu beobachten bedingt darauf, das die Selektionsvariable positiv ist [also: $f(y_i | y_{is} > 0)$]. Anwendung der Bayesianischen Regel

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

lässt uns schreiben

$$\begin{aligned}
&\varphi(y_i) = f(y_i | y_{is} > 0) P(y_{is} > 0) = P(y_{is} > 0 | y_i) f(y_i) \\
&\Rightarrow \quad \varphi(y_i) = P(y_{is} > 0 | y_i) f(y_i).
\end{aligned}$$

Somit besteht unsere Zielwahrscheinlichkeit $\varphi(y_i)$ aus zwei Komponenten: $P(y_{is} > 0 | y_i)$ und $f(y_i)$. Die erste Komponente bezieht sich auf die bedingte Dichte $f(y_{is} | y_i)$ der bivariaten Normalverteilung, die selbst auch normal ist, nämlich

$$f(y_{is} | y_i) = \frac{1}{\sigma_s \sqrt{2\pi (1 - \rho^2)}} \exp\left\{ -\frac{1}{2} \frac{\left(y_{is} - \left(\mathbf{z}_i'\gamma + \rho \frac{\sigma_s}{\sigma} [y_i - \mathbf{x}_i'\beta]\right)\right)^2}{(1 - \rho^2)\sigma_s^2} \right\}.$$

Somit

$$\begin{aligned}
P(y_{is} > 0 | y_i) &= 1 - P(y_{is} \leq 0 | y_i) \\
&= 1 - \int_{-\infty}^{0} f(y_{is} | y_i) \, dy_{is} \\
&= 1 - \Phi\left(\frac{0 - \left(\mathbf{z}_i'\gamma + \rho \frac{\sigma_s}{\sigma} [y_i - \mathbf{x}_i'\beta]\right)}{\sigma_s \sqrt{1 - \rho^2}} \right) \\
&= \Phi\left(\frac{\mathbf{z}_i'\gamma + \rho \frac{\sigma_s}{\sigma} [y_i - \mathbf{x}_i'\beta]}{\sigma_s \sqrt{1 - \rho^2}} \right) \\
&= \Phi\left(\frac{\frac{\mathbf{z}_i'\gamma}{\sigma_s} + \frac{\rho}{\sigma} [y_i - \mathbf{x}_i'\beta]}{\sqrt{1 - \rho^2}} \right).
\end{aligned}$$

Die zweite Komponente ist die Randdichte der bivariaten Normalverteilung, die auch normal ist, nämlich

$$f\left(y_i\right)=\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\beta}{\sigma}\right)$$

Somit haben wir schließlich

$$\varphi\left(y_i\right)=\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\beta}{\sigma}\right)\Phi\left(\frac{\frac{\mathbf{z}_i'\gamma}{\sigma_s}+\frac{\rho}{\sigma}\left[y_i-\mathbf{x}_i'\beta\right]}{\sqrt{1-\rho^2}}\right),$$

was eine Dichtefunktion der Ausprägung y_i in der Selektiven Teilstichprobe $\{y_i,\mathbf{x}_i,\mathbf{z}_i\}$ darstellt.

Die Wahrscheinlicht der anderen Teilstichprobe zu gehören, d.h. die Wahrscheinlichkeit nur $\{\varnothing,\varnothing,\mathbf{z}_i\}$ zu beobachten, denn $y_{is}\leq 0$ ist, ist einfach

$$\begin{aligned}P\left(y_{is}\leq 0\right) &= P\left(\mathbf{z}_i'\gamma+\varepsilon_{si}\leq 0\right)=P\left(\varepsilon_{si}\leq-\mathbf{z}_i'\gamma\right)=P\left(\frac{\varepsilon_{si}}{\sigma_s}\leq-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)\\ &= \Phi\left(-\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)=1-\Phi\left(\frac{\mathbf{z}_i'\gamma}{\sigma_s}\right)\end{aligned}$$

Das ist nichts anderes als eine Wahrscheinlichkeit einen "Null" in einer Probit-Regression zu beobachten.

Die Teilen $\varphi\left(y_i\right)$ und $P\left(y_{is}\leq 0\right)$ beschreiben die ganze zur Verfügung stehende Information. Da das Modell, ähnlich zu Probit, γ getrennt von σ_s nicht schätzen kann, greifen wir zu den gleichen Normierung $\sigma_s=1$ zurück. Definiert man eine Dummyvariable

$$d_i=\begin{cases}1, & \text{wenn } y_{is}>0\\ 0, & \text{wenn } y_{is}\leq 0\end{cases}$$

dann die Likelihoodfunktion des Heckman-Modells ist

$$\mathcal{L}=\prod_{i=1}^{n_1+n_2}\left[\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\beta}{\sigma}\right)\Phi\left(\frac{\mathbf{z}_i'\gamma+\frac{\rho}{\sigma}\left[y_i-\mathbf{x}_i'\beta\right]}{\sqrt{1-\rho^2}}\right)\right]^{d_i}\left[1-\Phi\left(\mathbf{z}_i'\gamma\right)\right]^{1-d_i}$$

Der gesamte Parametervektor θ in diesem Modell besteht aus $\theta=\{\beta,\gamma,\sigma,\rho\}$. Der gesamte Parametervektor wird in einer Stufe geschätzt.

- Identifizierbarkeit im Heckman-Modell

Unter der Normierung $\sigma_s=1$ ist das Modell identifizierbar und der ML Schätzer des Gesamtparametervektors θ ist eindeutig.

- Partielle Effekte im Heckman-Modell

Partieller Effekt auf y_i ist nichts anderes als $\partial E(y_i^*|\mathbf{x}_i)/\partial x_j = \beta_j$. Außerdem, das ist der genaue Effekt, denn β wird nie mit σ Skaliert. Partieller Effekt auf y_{is} ist dem in der Probit Modell identisch. Im Heckman-Rahmen wird der Effekt auf y_{is} in der Regel so gut wie nie angesprochen. Der Zweck dieser Regression ist konsistent β zu schätzen. Wenn man die Aussagen ausschließlich über γ machen will, ein einfaches Probit-Modell wäre dafür schon ausreichend.

Literatur

∗∗ Andrews, D., "Chi-square diagnostic tests for parametric models: Theory", *Econometrica*, 1988, Vol. 56(6), p.1419-1453.

◦ Cameron, C., and P., Trivedi, "Microeconometrics: Methods and applications", (Cambridge University Press: 2005), Ch. 20.2, (p.666-671); Ch. 20.4.5, (p.680-682); Ch. 16.2, 16.3.1-5, 16.5.1-5, (p.530-542, 546-551).

◦ Greene, W., "Econometric analysis", (Prentice Hall: 2003), 4th Ed., Ch. 21.9.1-2, (p.740-744); Ch. 21.9.6 (p.749-752); Ch. 22.2, Ch. 22.3.1-3, (p.756-768).

∗ Heckman, J., "Sample selection bias as a specification error", *Econometrica*, 1979, Vol. 47(1), pp.153-161.

◦ Ronning, G., "Mikroökonometrie", (Springer: 1991), Kap.3.1-3.2, (S.121-133).