

Slide 1

BRM mit Stata
Ulrich Kohler, WZB
6. Mai 2005

Slide 2

1 Das Binäre Regressionsmodell (BRM)

Gegeben

$$\Pr(y = 1|x) = \Pr(y^* > 0|x) \quad (1)$$

mit $y^* = \mathbf{x}_i\beta + \epsilon_i$. Dann gilt:

$$\begin{aligned} \Pr(y = 1|x) &= \Pr(\mathbf{x}_i\beta + \epsilon_i > 0|x) \\ &= \Pr(\epsilon > -(\mathbf{x}_i\beta)) \end{aligned}$$

und, solange die Fehler symmetrisch verteilt sind

$$= \Pr(\epsilon \leq \mathbf{x}_i\beta) \quad (2)$$

Slide 3

2 cdf

Gesucht ist also die Wahrscheinlichkeit von Werten von ϵ kleiner $\mathbf{x}_i\beta$. Diese lässt sich ermitteln mit Hilfe kumulativer Wahrscheinlichkeitsdichtefunktionen (cdf). Eine cdf gibt an, wie wahrscheinlich es ist, dass ein Wert kleiner oder gleich einem bestimmten Wert ist. Daraus folgt:

$$\Pr(\epsilon \leq \mathbf{x}_i\beta) = \text{cdf}(\mathbf{x}_i\beta) \quad (3)$$

Slide 4

2.1 Die cdf der Fehler

Die cdf erhält man über spezifische Annahmen über die Verteilung der Fehler. Wenn $e \sim \text{normal}(0,1)$ so erhält man die cdf der Standardnormalverteilung:

$$\phi(\epsilon) = \int_{-\infty}^{\epsilon} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt \quad (4)$$

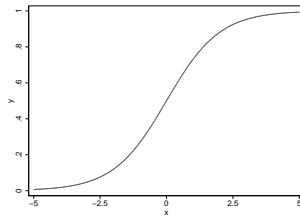
Wenn ($e \sim \text{logistic}(0, \pi^2/3)$) so erhält man die cdf der logistischen Verteilung:

$$\Lambda(\epsilon) = \frac{\exp(\epsilon)}{1 + \exp(\epsilon)} \quad (5)$$

Slide 5

2.2 Graphik der cdf

```
. tw function y=exp(x)/(1+exp(x)), range(-5 5) xlabel(-5(2.5)5)
```



Gegeben

$$y_L^* = \mathbf{X}\beta_L + \epsilon_L \quad \text{and} \quad y_P^* = \mathbf{X}\beta_P + \epsilon_P \quad (9)$$

wobei L ein Logit-Modell und P ein Probit-Modell indiziert. In beiden Modellen wird die Varianz von y^* durch Annahmen über Varianz der Fehler identifiziert, wobei gilt:

$$\text{Var}(\epsilon_L|\mathbf{x}_i) = \frac{\pi^2}{3} \text{Var}(\epsilon_P|\mathbf{x}_i) \quad (10)$$

und daher $\epsilon_L \approx (\pi/\sqrt{3})\epsilon_P$, d.h. der Fehler im Logit-Modell entspricht (ungefähr) dem Fehler im Probit-Modell, multipliziert mit einer Konstanten. Entsprechend gilt wie in Gleichung (8):

$$\beta_L = \sqrt{\text{Var}(\epsilon_L|\mathbf{x})}\beta_P \approx \sqrt{\frac{\pi^2}{3}}\beta_P \approx 1.81\beta_P \quad (11)$$

Slide 7

Slide 6

2.3 Konsequenzen der Wahl der Fehlerverteilung I

Gegeben

$$y = \mathbf{x}_i\beta_y + \epsilon_y \quad (6)$$

Erzeugt man eine neue Variable $w = \delta y$ mit δ einer Konstanten so entspricht die Varianz von w

$$\text{Var}(w) = \text{Var}(\delta y) = \delta^2 \text{Var}(y) \quad (7)$$

Außerdem gilt dann:

$$w = \delta(\mathbf{x}_i\beta_y + \epsilon_y) = \mathbf{x}_i\delta\beta_y + \delta\epsilon_y \quad (8)$$

d.h. $\beta_w = \delta\beta_y$

▷ Beispiel:

```
. use http://www.stata.com/datenanalyse/data1, clear
(SOEP'97 (Kohler/Kreuter))
. gen eigent=wohnst==1
. replace hhein=hhein/1000
hhein was int now float
(3201 real changes made)
. quietly logit eigent hhein gebjahr hhgr
. estimates store Logit
. quietly probit eigent hhein gebjahr hhgr
. estimates store Probit
```

Slide 8

Slide 9

```
. estimates table Logit Probit, b(%4.3f)
```

Variable	Logit	Probit
hhein	0.301	0.179
gebjahr	-0.022	-0.013
hhgr	0.180	0.109
_cons	40.647	24.565

Slide 11

▷ Beispiel:

```
. quietly logit eigent hhein gebjahr hhgr
. predict Phat1
(option p assumed; Pr(eigent))
(139 missing values generated)
. quietly probit eigent hhein gebjahr hhgr
. predict Phat2
(option p assumed; Pr(eigent))
(139 missing values generated)
. sum Phat1 Phat2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Phat1	3201	.3601999	.1581886	.0679601	.9993523
Phat2	3201	.3591948	.1563447	.0571272	.9999935

```
. cor Phat1 Phat2
(obs=3201)
```

Slide 10

2.4 Konsequenzen der Wahl der Fehlerverteilung II

Die β -Koeffizienten hängen ab (1) von der Beziehung zwischen y_i^* und x und (2) von der Identifizierungsannahme. Sie können darum nicht unmittelbar interpretiert werden.

Anders verhält es sich mit $\Pr(y_i = 1|x_i)$. Dies ist eine sogenannte *estimable function*.

Slide 12

	Phat1	Phat2
Phat1	1.0000	
Phat2	0.9998	1.0000

3 ML-Estimation

Gegeben

$$p_i = \begin{cases} \Pr(y=1|x_i) & \text{wenn } y_i = 1 \text{ beobachtet wurde} \\ \Pr(y=0|x_i) & \text{wenn } y_i = 0 \text{ beobachtet wurde} \end{cases} \quad (12)$$

ist die Likelihood bei voneinander unabhängigen Beobachtungen

$$L(\beta|y, \mathbf{x}_i) = \prod_{i=1}^n p_i \quad (13)$$

wobei $\Pr(y = 1|x) = \text{cdf}(\mathbf{x}_i, \beta)$

Slide 13

Probit-Modell:

```
capture program drop myprobit
program define myprobit
    version 7.0
    args lnf theta
    quietly replace `lnf' = ln(norm(`theta')) if $ML_y1==1
    quietly replace `lnf' = ln(norm(-`theta')) if $ML_y1==0
end

ml model lf myprobit (eigent = hhein gebjahr hhgr)
ml maximize
```

Slide 15

3.1 Maximum-Likelihood-Estimation mit Stata

Logit-Modell:

```
capture program drop mylogit
program define mylogit
    version 7.0
    args lnf theta
    quietly replace `lnf' = ln(exp(`theta')/(1+exp(`theta')))) if $ML_y1==1
    quietly replace `lnf' = ln(exp(-`theta')/(1+exp(-`theta')))) if $ML_y1==0
end

ml model lf mylogit (eigent = hhein gebjahr hhgr)
ml maximize
```

Slide 14

4 Interpretation mit „Idealtypen“

Logit-Modell:

```
. quietly logit eigent hhein gebjahr hhgr
. prvalue, x(hhgr=1 gebjahr=1950) rest(mean) nobase

logit: Predictions for eigent
Pr(y=1|x):      0.2946   95% ci: (0.2685,0.3222)
Pr(y=0|x):      0.7054   95% ci: (0.6778,0.7315)

. prvalue, x(hhgr=3 gebjahr=1950) rest(mean) nobase

logit: Predictions for eigent
Pr(y=1|x):      0.3745   95% ci: (0.3559,0.3934)
Pr(y=0|x):      0.6255   95% ci: (0.6066,0.6441)
```

Slide 16

Slide 17

Probit-Modell

```
. quietly probit eigent hhein gebjahr hhgr
. prvalue, x(hhgr=1 gebjahr=1950) rest(mean) nobase

probit: Predictions for eigent
Pr(y=1|x):      0.2964  95% ci: (0.2704,0.3236)
Pr(y=0|x):      0.7036  95% ci: (0.6764,0.7296)

. prvalue, x(hhgr=3 gebjahr=1950) rest(mean) nobase

probit: Predictions for eigent
Pr(y=1|x):      0.3757  95% ci: (0.3574,0.3943)
Pr(y=0|x):      0.6243  95% ci: (0.6057,0.6426)
```

Slide 19

6 Conditional Effects Plot

```
. prgen hhein, from(0) to(6) gen(hgr1) x(hhgr=1) rest(mean) n(10)
. prgen hhein, from(0) to(6) gen(hgr2) x(hhgr=2) rest(mean) n(10)
. prgen hhein, from(0) to(6) gen(hgr3) x(hhgr=3) rest(mean) n(10)
. prgen hhein, from(0) to(6) gen(hgr4) x(hhgr=4) rest(mean) n(10)
.
. line hgr1p1 hgr2p1 hgr3p1 hgr4p1 hgr1x,      ///
> clpattern(solid dot dash dash_dot)      ///
> legend(lab(1 "hhgr=1") lab(2 "hhgr=2")      ///
>         lab(3 "hhgr=3") lab(4 "hhgr=4") )
```

Slide 18

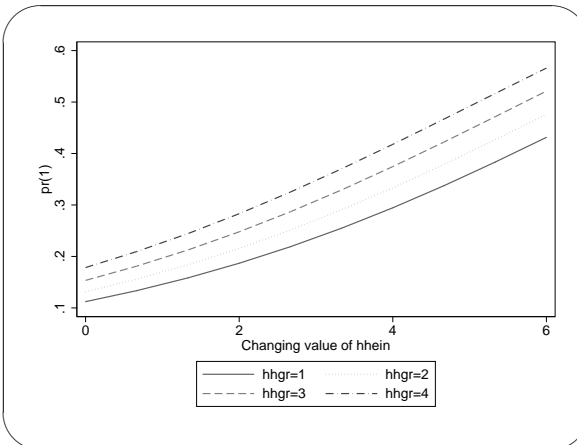
5 Tabellen mit \hat{P}

```
. quietly logit eigent hhein gebjahr hhgr sex
. prtab hhgr sex if hhgr < 7, rest(mean) nobase

logit: Predicted probabilities of positive outcome for eigent
```

Haushalts groesse	Geschlecht	
	Maenner	Frauen
1	0.3002	0.2724
2	0.3395	0.3096
3	0.3810	0.3495
4	0.4244	0.3916
5	0.4690	0.4353
6	0.5141	0.4801

Slide 20



7 Marginaleffekte, Discrete-Change

```
. prchange
logit: Changes in Predicted Probabilities for eigent
      min->max      0->1      -+1/2      -+sd/2      MargEfct
hhein      0.8540      0.0408      0.0677      0.1446      0.0677
gebjahr    -0.3991      0.0000      -0.0050      -0.0916      -0.0050
hhgr       0.4230      0.0354      0.0409      0.0526      0.0409
sex        -0.0309      -0.0321      -0.0309      -0.0154      -0.0309

      0      1
Pr(y|x) 0.6524 0.3476

      hhein gebjahr hhgr sex
x=      3.86193 1951.83 2.57919 1.52171
sd(x)=  2.14739 18.2245 1.28503 .499606
```

Slide 21

der sogenannten Odds-Ratio miteinander verglichen.

$$\begin{aligned} \frac{\Omega(\mathbf{x}, x_2 + 1)}{\Omega(\mathbf{x}, x_2)} &= \frac{\exp(\mathbf{x}b) \times \exp((x_2 + 1)b_2)}{\exp(\mathbf{x}b) \times \exp(x_2 b_2)} \\ &= \frac{\exp(\mathbf{x}b) \times \exp(x_2 b_2) \times \exp(b_2)}{\exp(\mathbf{x}b) \times \exp(x_2 b_2)} \\ &= \exp(b_2) \end{aligned} \quad (16)$$

Slide 23

d.h.: Eine Veränderung von x_2 um eine Einheit verändert die Chance auf $y = 1$ um das e^{b_2} -fache.

8 OR-Interpretation im Logit-Modell

Im Logit-Modell gilt

$$\mathbf{x}_i b = \ln \left(\frac{\Pr(y=1)}{1 - \Pr(y=1)} \right). \quad (14)$$

Slide 22

Der Term $\ln \left(\frac{\Pr(y=1)}{1 - \Pr(y=1)} \right)$ ist das so genannte "Logit", das logarithmierte Wahrscheinlichkeitsverhältnis ("Odd"). Das "Logit" steigt linear mit $\mathbf{x}_i b$.

Durch Exponentierung beider Seiten von (14) erhält man das Odd:

$$\frac{\Pr(y=1)}{1 - \Pr(y=1)} = \exp(\mathbf{x}_i b) = \Omega(\mathbf{x}) \quad (15)$$

Odds unterschiedlicher Gruppen werden typischerweise mit Hilfe

```
. listcoef
```

```
logit (N=3201): Factor Change in Odds
```

```
Odds of: 1 vs 0
```

eigent	b	z	P> z	e ^b	e ^b StdX	SDofX
hhein	0.29874	14.119	0.000	1.3482	1.8993	2.1474
gebjahr	-0.02221	-9.714	0.000	0.9780	0.6671	18.2245
hhgr	0.18056	5.535	0.000	1.1979	1.2612	1.2850
sex	-0.13623	-1.739	0.082	0.8726	0.9342	0.4996

Slide 24