

# Countdata, Postestimation und Modellvergleich

## Statistik II

Wiederholung/Fortsetzung  
Count Data  
Postestimation und Modellvergleich  
Zusammenfassung

- ① Wiederholung/Fortsetzung  
Literatur
- ② Count Data  
Einführung  
Das Poisson-Modell
- ③ Postestimation und Modellvergleich  
Postestimation  
Modellvergleich
- ④ Zusammenfassung

## Zum Nachlesen

- Für heute: Scott/Freese ch. 8
- Für nächste Woche: Wooldridge Kapitel 10.1 und 10.2 (im Reader)

## Was sind Count Data?

- Viele interessante Variablen sind *Zählungen*
- Was ist daran besonders?
  - Immer größer null
  - Nur ganzzahlige Werte
- Beispiele:
  - IB: Bewaffnete Konflikte pro Beobachtungszeitraum, Zahl der Staatsstreiche in Afrika, Zahl anderer Ereignisse
  - American Politics: Zahl der präsidentiellen Vetos pro Jahr, Zahl der Mitarbeiter von Abgeordneten, Zahl der verlorenen Sitze bei midterm elections, Zahl der Teilnehmer bei Kabinettsitzungen, Zahl der Parteiwechsler im Kongreß pro Jahr

## Warum soll das problematisch sein?

- Zählvariablen machen keine Probleme als unabhängige Variablen (intervallskaliert)
- Aber als abhängige Variablen:
  - Werte  $< 0$  nicht sinnvoll
  - Rationale Werte (z. B. 2.5) nicht sinnvoll
  - Fehlervarianz nicht konstant, sondern vom Niveau von  $y$  abhängig (mehr Varianz für größere  $y$ )
  - Für kleine  $y$  Fehler nach unten beschränkt ( $y$  kann nicht kleiner 0 werden)
- „Normale“ Regression (OLS): ineffiziente, inkonsistente, verzerrte Schätzungen; problematische Standardfehler
- Lösung: spezielle Modelle, z. B. Poisson-Modell

## Was ist/wozu braucht man das Poisson-Modell?



- Siméon Denis Poisson (1781-1840), französischer Mathematiker und Astronom
- Poisson-Verteilung:
  - „Ereignis“ tritt innerhalb gegebenen Zeitraums mit bestimmter Rate ein
  - Zufällige Abweichungen nach oben und nach unten
  - Diskrete Verteilung
- Wie wahrscheinlich ist eine bestimmte Häufigkeit bei gegebener Rate?

## Was beschreibt die Poisson-Verteilung?

### Poisson-Verteilung

$$\Pr(y|\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

- $\Pr(y|\mu)$ : Konditionale (d. h. auf bestimmtes  $\mu$  bezogene Wahrscheinlichkeit eines  $y$ -Wertes)
- $\mu$ : „Rate“, d. h. „erwartete“ oder „wahre“ Häufigkeit der Größe, die gezählt wird (z. B. Zahl präsidentieller Vetos)
- $y = 0, 1, 2, \dots$ : beobachtbare, ganzzahlige Häufigkeit
- $y!$ :  $y$ -Fakultät, d. h.  $y \times (y - 1) \times (y - 2) \times \dots \times 1$  mit
  - $1! = 1$
  - $0! = 1$

## Ein Beispiel

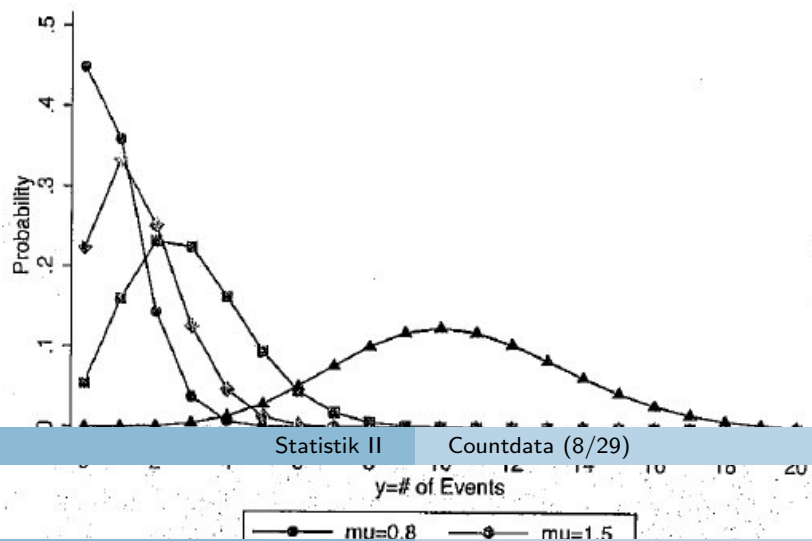
- In den letzten 20 Jahren (Bush, Clinton, Bush) gab es 4.65 Vetos per Jahr ( $\mu$ )
- Obama: bisher nur ein Veto pro Jahr
- Wenn 4.65 die wahre Rate ist, wie wahrscheinlich ist  $y = 1$  ?

$$\begin{aligned}\Pr(y = 1|\mu = 4.65) &= \frac{e^{-4.65} \times 4.65^1}{1!} \\ &\approx 0.0095616 \times 4.65 \\ &\approx 0.04446 \\ &\approx 4.4\%\end{aligned}$$

- Wenn die Rate konstant bei 4.65 Vetos pro Jahr liegt und
- Wenn die konkrete Zahl der Vetos in einem Jahr zufällig um

## Poisson-Verteilungen

- Poisson-Verteilung hat einen einzigen Parameter  $\mu$
- Diskret, d. h. Wahrscheinlichkeiten für  $y = 0, 1, 2, \dots$
- Equidispersion, d. h.  $\mu =$  Mittelwert und Varianz der Verteilung
- $\mu$  bestimmt Form und Lage der Verteilung
- Für große  $\mu$  Annäherung an Normalverteilung



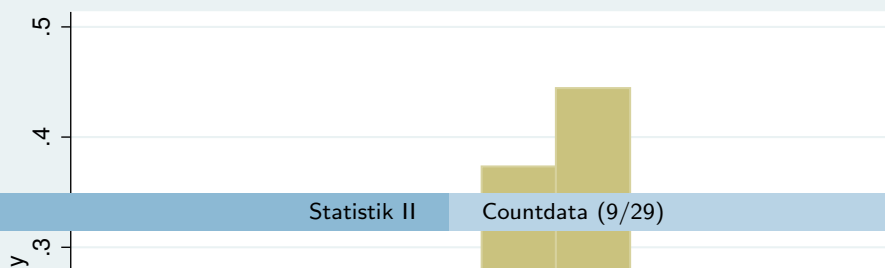
## Noch ein Beispiel

- 75 Kursteilnehmer haben Länder der EU zugeordnet
- Wieviele richtige Zurechnungen (ohne negative scoring)?

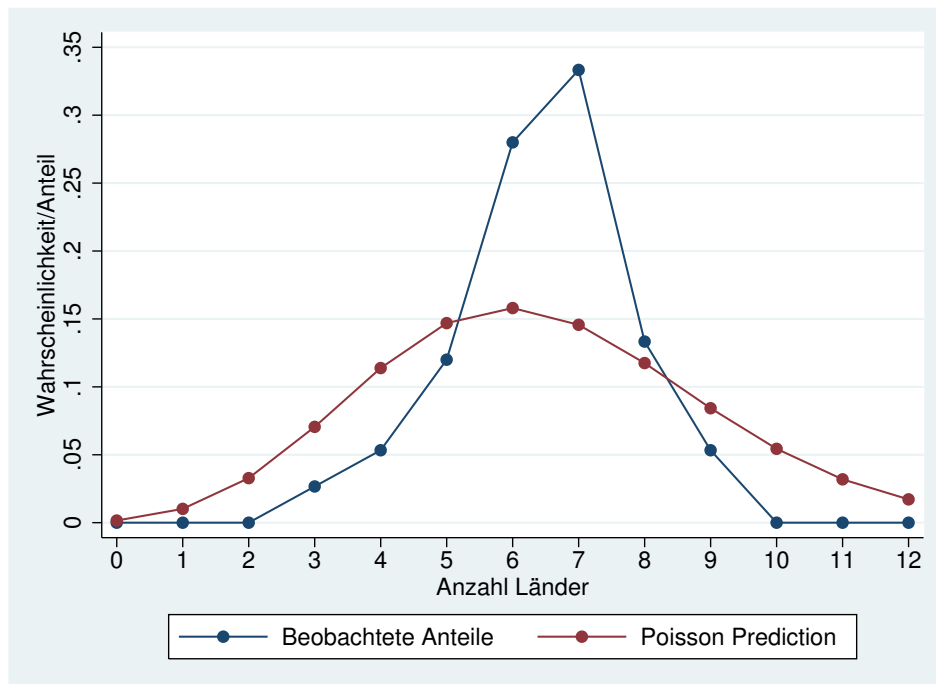
```
. summ eulaender ,det
```

Anzahl richtig zugeordneter Länder

Percentiles		Smallest		
1%	3	3		
5%	4	3		
10%	5	4	Obs	75
25%	6	4	Sum of Wgt.	75
50%	7		Mean	6.453333
		Largest	Std. Dev.	1.318202
75%	7	9	Variance	1.737658
90%	8	9	Skewness	-.3793161
95%	9	9	Kurtosis	3.166489
99%	9	9		



## Vergleich empirische/Poisson-Verteilung



nicht besonders gut

## Wie sieht das Poisson-Modell aus?

- Individuelle Rate für jeden Befragten
- Bzw. Rate hängt ab von unabhängigen Variablen
- Modell enthält Exponentialfunktion
  - Rate immer positiv
  - Nicht-lineare Effekte

### Poisson-Modell

$$\begin{aligned}\mu &= \exp(x_0 + \beta_1 x_1 + \beta_2 x_2 + \dots) \\ &= e^{x_0 + \beta_1 x_1 + \beta_2 x_2 + \dots}\end{aligned}$$

```
. poisson eulaender vielfernsehen leser dailynet
```

```
Iteration 0: log likelihood = -148.61946
```

```
Iteration 1: log likelihood = -148.61946
```

```
Poisson regression
```

```
Number of obs = 75
LR chi2(3) = 1.69
Prob > chi2 = 0.6389
Pseudo R2 = 0.0057
```

```
Log likelihood = -148.61946
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
vielfernse-n	.0478098	.0917573	0.52	0.602	-.1320312	.2276508
leser	-.0027584	.0450933	-0.06	0.951	-.0911395	.0856228
dailynet	-.1250795	.1052377	-1.19	0.235	-.3313417	.0811827
_cons	1.931206	.1204311	16.04	0.000	1.695166	2.167247

```
. poisson eulaender vielfernsehen leser dailynet
```

```
Iteration 0: log likelihood = -148.61946
```

```
Iteration 1: log likelihood = -148.61946
```

```
Poisson regression
```

```
Number of obs = 75
LR chi2(3) = 1.69
Prob > chi2 = 0.6389
Pseudo R2 = 0.0057
```

```
Log likelihood = -148.61946
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
vielfernse-n	.0478098	.0917573	0.52	0.602	-.1320312	.2276508
leser	-.0027584	.0450933	-0.06	0.951	-.0911395	.0856228
dailynet	-.1250795	.1052377	-1.19	0.235	-.3313417	.0811827
_cons	1.931206	.1204311	16.04	0.000	1.695166	2.167247

- Richtung
- Signifikanz
- Mehr dazu gleich

```
. poisson eulaender vielfernsehen leser dailynet male wiesbaden magister finish
> ed
```

```
Iteration 0: log likelihood = -147.81184
```

```
Iteration 1: log likelihood = -147.81184
```

```
Poisson regression
```

```
Number of obs = 75
LR chi2(7) = 3.31
Prob > chi2 = 0.8553
Pseudo R2 = 0.0111
```

```
Log likelihood = -147.81184
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
vielfernse-n	.0600494	.094572	0.63	0.525	-.1253084	.2454072
leser	.0022951	.0454195	0.05	0.960	-.0867254	.0913156
dailynet	-.1436569	.1170262	-1.23	0.220	-.3730239	.0857102
male	.0242565	.1087385	0.22	0.823	-.1888669	.23738
wiesbaden	-.0409788	.1641448	-0.25	0.803	-.3626967	.2807392
magister	.1947908	.1531499	1.27	0.203	-.1053775	.4949592
finished	.0002446	.0147941	0.02	0.987	-.0287514	.0292406
_cons	1.899616	.1398603	13.58	0.000	1.625495	2.173737

## Wie interpretiert man das?

- Interpretation über die erwartete Rate und deren Veränderung
- Wie kommt man vom Modell zur Rate?
- (Drittes Modell)

## Dritter Versuch

```
. poisson eulaender dailynet alter magister
Iteration 0: log likelihood = -144.26713
Iteration 1: log likelihood = -144.26713
Poisson regression
Log likelihood = -144.26713
Number of obs = 73
LR chi2(3) = 3.01
Prob > chi2 = 0.3895
Pseudo R2 = 0.0103
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
dailynet	-.1333206	.0991716	-1.34	0.179	-.3276933 .061052
alter	-.0148306	.0292862	-0.51	0.613	-.0722305 .0425692
magister	.1983553	.1534169	1.29	0.196	-.1023363 .4990469
_cons	2.265212	.6519851	3.47	0.001	.9873446 3.543079

- Rate für 25 Jahre alten BA, der nicht täglich das Internet benutzt?
- $\exp(2.265 + 0 \times -0.133 + 25 \times -0.015 + 0 \times 0.198) = 6.62$
- Was erwarten wir für gleichalten Magister mit gleichen Gewohnheiten?
- $\exp(2.265 + 0 \times -0.133 + 25 \times -0.015 + 1 \times 0.198) = 8.07$

## Wie sieht das Poisson-Modell aus?

- Individuelle Rate für jeden Befragten
- Bzw. Rate hängt ab von unabhängigen Variablen
- Modell enthält Exponentialfunktion
  - Rate immer positiv
  - Nicht-lineare Effekte

### Poisson-Modell

$$\begin{aligned}\mu &= \exp(x_0 + \beta_1 x_1 + \beta_2 x_2 + \dots) \\ &= e^{x_0 + \beta_1 x_1 + \beta_2 x_2 + \dots}\end{aligned}$$

## Dritter Versuch

```
. poisson eulaender dailynet alter magister
Iteration 0:  log likelihood = -144.26713
Iteration 1:  log likelihood = -144.26713
Poisson regression              Number of obs   =          73
                                LR chi2(3)       =           3.01
                                Prob > chi2      =          0.3895
                                Pseudo R2       =          0.0103
Log likelihood = -144.26713
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
dailynet	-.1333206	.0991716	-1.34	0.179	-.3276933 .061052
alter	-.0148306	.0292862	-0.51	0.613	-.0722305 .0425692
magister	.1983553	.1534169	1.29	0.196	-.1023363 .4990469
_cons	2.265212	.6519851	3.47	0.001	.9873446 3.543079

- Rate für 25 Jahre alten BA, der nicht täglich das Internet benutzt?
- $\exp(2.265 + 0 \times -0.133 + 25 \times -0.015 + 0 \times 0.198) = 6.62$
- Was erwarten wir für gleichalten Magister mit gleichen Gewohnheiten?
- $\exp(2.265 + 0 \times -0.133 + 25 \times -0.015 + 1 \times 0.198) = 8.07$

## Generell: Interpretation als faktorielle Veränderung

- Veränderung in  $x_1$  führt zu additiver Veränderung des Exponenten in  $\exp(x_0 + \beta_1 x_1 \dots) = e^{x_0 + \beta_1 x_1 \dots}$  (in diesem Fall + 0.198)
- Alles andere bleibt konstant
- *Multiplikative* Veränderung der erwarteten Rate:  
 $e^{2+1} = e^2 \times e^1$
- D. h. bei  $\beta_1 x_1$  und Veränderung von  $x_1$  um eine Einheit verändert sich erwartete Rate um *Faktor*  $e^{\beta_1}$
- $e^{0.198} = 1.22$ ; entspricht Zunahme um 22%
- $6.62 \times 1.22 = 8.07$
- Magister kennen *ceteris paribus* 22% mehr Länder (nicht notwendigerweise 1.4 Länder!) mehr als BAs

## Für andere Variablen ...

- Für häufige Internetnutzung (Dummy):
  - Veränderung 0  $\rightarrow$  1 :  $\exp(-0.133) = 0.875$
  - D. h. Rückgang um ca. 12.5%
- Für das Alter (in Jahren):
  - Veränderung um ein Jahr:  $\exp(-0.0148) = 0.985$
  - Rückgang um ca. 1.5%
  - Veränderung um drei Jahre?
    - $\exp(-0.0148 \times 3) = 0.957$
    - Rückgang um rund 4%
- Veränderung in absoluten Werten (Länder) hängt vom Ausgangsniveau *aller* unabhängigen Variablen ab
- Weil Modell non-linear ist

## Was ist Postestimation?

- Generell: Alles, was nach der Schätzung der Koeffizienten passiert
  - Tests und Interpretation für Koeffizienten
    - Sind zwei Koeffizienten (Effekte) gleich stark?
    - *Wahrscheinlichkeiten* und *graphische Darstellungen* für interessante Szenarien
  - Fit
    - Auf der Ebene einzelner Beobachtungen („Ausreißer“ etc.)
    - Auf der Ebene des Gesamtmodells
  - In Stata sehr gut unterstützt, u. a. durch Kommandos zum Speichern von Schätzungen und durch Zusatzpakete
- `poisson eulaender ... est store zweiterversuch`

## Übersicht

```
. est tab *
```

Variable	dritterv~h	zweiterv~h
dailynt	-.13332065	-.14365686
alter	-.01483061	
magister	.19835532	.19479085
vielfernse-n		.06004937
leser		.00229513
male		.02425653
wiesbaden		-.04097876
finished		.00024461
_cons	2.2652118	1.8996163

- Hebt negativer Effekt des Internets positiven Effekt des Altstudienganges (unter Kontrolle des Alters) auf?

## Koeffiziententest

```
. est restore dritterversuch
(results dritterversuch are active now)
. poisson
Poisson regression                               Number of obs   =       73
                                                LR chi2(3)      =        3.01
                                                Prob > chi2     =       0.3895
Log likelihood = -144.26713                    Pseudo R2      =       0.0103
```

eulaender	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dailynet	-.1333206	.0991716	-1.34	0.179	-.3276933	.061052
alter	-.0148306	.0292862	-0.51	0.613	-.0722305	.0425692
magister	.1983553	.1534169	1.29	0.196	-.1023363	.4990469
_cons	2.265212	.6519851	3.47	0.001	.9873446	3.543079

```
. test dailynet = -1*magister
( 1) [eulaender]dailynet + [eulaender]magister = 0
      chi2( 1) =    0.14
      Prob > chi2 =    0.7035
```

## Rate/Länderzahl für Gruppen?

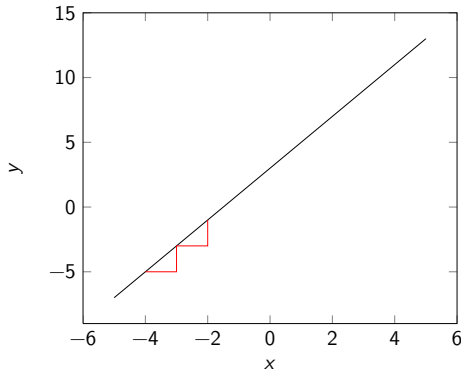
- Welche Rate erwarte ich für internetaffinen BA von 25 Jahren? Wieviele Länder werden mit welcher Wahrscheinlichkeit erkannt?
- Einsetzen – mühsam und fehleranfällig
- Befehl `prvalue` aus `spost` (funktioniert mit fast allen Regressionsmodellen)

```
. prvalue, x(dailynet=1 magister=0 alter=25) max(12)
poisson: Predictions for eulaender
Confidence intervals by delta method
```

	Rate:	5.819	95% Conf. Interval
Pr(y=0 x):	0.0030		[ 4.5688, 7.0692]
Pr(y=1 x):	0.0173		[-0.0007, 0.0067]
Pr(y=2 x):	0.0503		[-0.0006, 0.0352]
Pr(y=3 x):	0.0976		[ 0.0090, 0.0916]
Pr(y=4 x):	0.1419		[ 0.0385, 0.1566]
Pr(y=5 x):	0.1652		[ 0.0865, 0.1974]
Pr(y=6 x):	0.1602		[ 0.1361, 0.1942]
Pr(y=7 x):	0.1332		[ 0.1539, 0.1664]
Pr(y=8 x):	0.0969		[ 0.0994, 0.1669]
Pr(y=9 x):	0.0594		[ 0.0515, 0.1422]
Pr(y=10 x):	0.0304		[ 0.0007, 0.0007]
Pr(y=11 x):	0.0193		[-0.0022, 0.0407]

## Von was hängt die erwartete Veränderung ab?

- Lineares Modell: einfach, da Veränderung konstant
  - Unabhängig vom Niveau von  $x_1$
  - Unabhängig vom Niveau anderer  $x$ -Variablen (falls vorhanden)
- Z. B.  $y = 3 + 2x$



- $x - 4 \rightarrow -3$ :  $y + 2$
- $x - 3 \rightarrow -2$ :  $y + 2$

- Nicht-lineare Modelle: komplexer
- Wirkung Veränderung von  $x$  auf  $y$  nicht konstant

Statistik II Countdata (24/29)

- Hängt ab vom Ausgangsniveau von  $x_1$
- Hängt ab vom Niveau anderer  $x$ -Variablen (falls vorhanden)

prchange

```
. prchange alter, x(magister=0 dailynet=0) fromto
poisson: Changes in Rate for eulaender
      from:      to:      dif:      from:      to:      dif:      from:
      x=min      x=max  min->max  x=0       x=1       0->1     x-1/2
alter   7.1606   6.1737  -0.9870   9.6332    9.4914   -0.1418   6.9932
      to:      dif:      from:      to:      dif:
alter   x+1/2    +1/2     x-1/2sd  x+1/2sd  +sd/2    MargEfct
alter   6.8902  -0.1029  7.0270   6.8570   -0.1700  -0.1029
exp(xb): 6.9415
      dailynet  alter  magister
      x=        0  22.0959  0
      sd_x=     .462028  1.6513  .296479
```

- Auf der Ebene der Fälle weniger gut entwickelt als bei linearer Regression
- Auf der Ebene des Gesamtmodells diverse Pseudo- $R^2$ 
  - Nicht als Prozent der erklärten Varianz oder linearer Zusammenhang interpretierbar
  - Verbesserung gegenüber einem Modell, das nur konstante enthält
  - Theoretisch zwischen 0 und 1
  - Praktisch meist niedrig
  - Modell mit hohem/höherem Pseudo- $R^2$  nicht unbedingt besser

## Wie vergleicht man Modelle?

- (Macht Modell theoretisch Sinn?)
  - Repräsentiert/testet theoretische Überlegungen
  - Keine wichtigen Variablen vergessen
  - Keine überflüssigen Variablen enthalten
- Vergleich zwischen theoretisch adäquaten Modellen
  - Guter Fit (Anpassung an die Daten)
  - Möglichst sparsam
- Information Criteria
- (Vorsicht: Overfitting)

## Was sind die Information Criteria?

- Maßzahlen, die Fit und Zahl der geschätzten Effekte berücksichtigen
- BIC = Bayesian Information Criterion, beliebt für Modellvergleiche
- Verschiedene Berechnungsvorschriften; Modell mit niedrigerem BIC bevorzugen

```
. fitstat, using(mod2) force
```

```
Measures of Fit for poisson of eulaender
```

	Current	Saved	Difference
Model:	poisson	poisson	
N:	73	75	-2
Log-Lik Intercept Only	-145.774	-149.465	3.691
Log-Lik Full Model	-144.267	-147.812	3.545
D	288.534(69)	295.624(67)	7.089(2)
LR	3.014(3)	3.306(7)	0.292(4)
Prob > LR	0.389	0.855	0.990
McFadden's R2	0.010	0.011	-0.001
McFadden's Adj R2	-0.017	-0.042	0.025
ML (Cox-Snell) R2	0.040	0.043	-0.003
Cragg-Uhler(Nagelkerke) R2	0.041	0.044	-0.003
AIC	4.062	4.155	-0.093

Statistik II      Countdata (28/29)

BIC	-7.507	6.352	-13.859
BIC'	9.857	26.916	-17.059
BIC used by Stata	305.696	330.164	-24.467

## Zusammenfassung

- Zählvariablen als abhängige Variablen problematisch → spezielle Modelle
- Poisson-Verteilung Modell für zufällige Verteilung von Zählvariablen
- Poisson-Regression: Rate der Verteilung ( $\mu$ ) als nicht-lineare Funktion der unabhängigen Variablen
- Postestimation: Qualität und Bedeutung eines Modells im Nachhinein analysieren
- Am besten graphisch/mit geschätzten Wahrscheinlichkeiten