

Slide 1

## Fitmaße für Maximum-Likelihood Modelle

Ulrich Kohler, WZB

7. Juni 2005

Slide 2

### Überblick

- Likelihood basierte Maßzahlen
  - Likelihood-Ratio-Chi-Quadrat
  - Devianz
  - Pseudo- $r^2$ -Werte
  - Informations-Kennziffern
- Maßzahlen auf der Basis vorhergesagter Werte
  - Efron's  $r^2$
  - McKelvey and Zavoina's  $r^2$
  - Count  $r^2$
  - Adjusted Count  $r^2$

Literatur: Long (1997, 93–97,102–113) sowie die darin zitierte Literatur

### Devianz

$$D = 2 \times \ln L(M_{\text{Full}}) \quad (2)$$

Slide 5

$D$  vergleicht die (logarithmierte) Wahrscheinlichkeit der beobachteten Daten eines Modells mit einem Parameter für jede Beobachtung ( $\ln L(M_{\text{Saturate}=0})$ ) mit der entsprechenden Wahrscheinlichkeit für die gefundenen Werte der Parameter ( $\ln L(M_{\text{Full}})$ ).

$D$  folgt nicht der Chi-Quadrat Verteilung.

Im Standard-Output von Likelihood-Modellen wird die Devianz nicht ausgewiesen.

Im Output von `fitstat` wird die Devianz mit „D(#)“ bezeichnet.

### Pseudo- $r^2$ -Werte

Likelihood-Ratio-Chi-Quadrat und Devianz

- haben keinen vorgegebenen Wertebereich
- sinken mit der Fallzahl
- steigen mit der Anzahl der Parameter der Modelle

Slide 6

Mit Pseudo- $r^2$ -Werten wird versucht, eine Maßzahl mit einem festen Wertebereich zu entwickeln, welche unabhängig von der Fallzahl ist.

Von einigen Pseudo- $r^2$ -Werten existieren „Adjusted“-Versionen. In diese wird zudem die Anzahl der Parameter der Modelle eingerechnet.

### Maximum Likelihood $r^2$

Slide 9

$$r_{\text{ML}}^2 = 1 - \left( \frac{L(M_{\text{Intercept}})}{L(M_{\text{Full}})} \right)^{2/n} = 1 - \exp(-G^2/n) \quad (5)$$

Der Wertebereich von  $r_{\text{ML}}^2$  ist  $[0, 1 - L(M_{\text{Intercept}})^{2/n}]$

$r_{\text{ML}}^2$  findet sich nicht im Standard-Output. In `fitstat` wird der Wert als „Maximum Likelihood R<sup>2</sup>“ bezeichnet.

▷ Beispiel:

Slide 10

```
. use titanic, clear
. logit survived sex age

Iteration 0:  log likelihood = -1384.7284
Iteration 1:  log likelihood = -1165.9473
Iteration 2:  log likelihood = -1164.5484
Iteration 3:  log likelihood = -1164.5475

Logit estimates          Number of obs   =          2201
snip >>

. di 1 - exp(-2 * (-1164.5475 - (-1384.7284)))/2201
.18132943
```

## Akaike Information Criteria

$$\text{AIC} = \frac{-2 \ln \widehat{L}(M_k) + 2P}{n} \quad (7)$$

Slide 13

mit  $\widehat{L}(M_k)$  der Likelihood des Modells und  $P$  der Anzahl der Parameter im Modell (einschl. der Konstante).

Information-Kriterien dienen zum Vergleich zwischen Modellen aus verschiedenen Stichproben bzw. zum Vergleich nicht „genesteter“ Daten.

*AIC* findet sich nicht im Standard-Output. In *fitstat* wird der Wert als „AIC“ bezeichnet.

▷ Beispiel:

```
. use titanic, clear
. logit survived sex age

Iteration 0:  log likelihood = -1384.7284
Iteration 1:  log likelihood = -1165.9473
Iteration 2:  log likelihood = -1164.5484
Iteration 3:  log likelihood = -1164.5475

Logit estimates                Number of obs   =          2201
snip >&

. display (-2 * (-1164.5475) + 2 * 3) / 2201
1.0609246
```

Slide 14

## Maßzahlen auf der Basis vorhergesagter Werte

Bei den vorhergesagten Werte sind zu unterscheiden:

- der lineare Prädiktor

$$y_i^* = \mathbf{x}_i \beta + \epsilon \quad (10)$$

Slide 17

- die vorhergesagte Wahrscheinlichkeit für einen bestimmten Wert der abhängigen Variable

$$\hat{\pi}_i^{(Y=y)} = \widehat{\Pr}(Y = y | \mathbf{x}_i) \quad (11)$$

- der wahrscheinlichste Wert der abhängigen Variable

$$\hat{y}_i = y_{\max \pi_i^{(Y=y)}} \quad (12)$$

## Efron's $r^2$

$$r_{\text{Efron}}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

Slide 18

$r_{\text{Efron}}^2$  ist nur für binäre abhängige Variablen definiert

$r_{\text{Efron}}^2$  vergleicht die vom Modell vorhergesagten Wahrscheinlichkeiten eines „Erfolgs“ mit der Erfolgswahrscheinlichkeit aus der Randverteilung.

Der Wertebereich von  $r_{\text{Efron}}^2$  ist  $[0, 1]$

$r_{\text{Efron}}^2$  findet sich nicht im Standard-Output. In `fitstat` wird der Wert als „Efron's R2“ bezeichnet.

Slide 21

### Count $r^2$

$$r_{\text{count}}^2 = \frac{1}{n} \sum_{i=1}^n C_i \quad (15)$$

$$C_i = \begin{cases} 0 & \text{wenn } \hat{y}_i \neq y_i \\ 1 & \text{wenn } \hat{y}_i = y_i \end{cases} \quad (16)$$

$r_{\text{count}}^2$  ermittelt den Anteil korrekt klassifizierte Beobachtungen. Da allein auf Grund der Randverteilung bereits hohe Anteile von Beobachtungen klassifiziert werden können empfiehlt sich die Anwendung der „Adjusted“-Version der Maßzahl:

$$r_{\text{count}}^2 = \frac{\sum_{i=1}^n C_i - \max(n_{r+})}{n - \max(n_{r+})} \quad (17)$$

Slide 22

▷ Beispiel:

```
. use titanic, clear
. logit survived sex age
. predict phat
. gen yhat=phat > .5
. gen c = sum(yhat==survived)
. display c[_N]/2201
.7760109

. count if survived==0
. local n0 = r(N)
. count if survived==1
. local nmax = max('n0',r(N))
. display (c[_N] - 'nmax')/(2201 - 'nmax')
.30661041
```