4.Tutorium Generalisierte Regression

- Tests/Devianz -

Nicole Schüller:

30.11.2015 und 07.12.2015

Hannah Busen:

03.12.2015 und 10.12.2015

Institut für Statistik, LMU München

Gliederung

Gliederung

Mögliche Problemstellungen und Hypothesen

Häufige Problemstellungen:

(a) Sind die Modellparameter überhaupt aussagekräftig, oder würde ein reines Interceptmodell genügen?

(b) Kann man Einflussgrößen aus dem Modell herausnehmen, ohne die Güte des Fits dabei zu stark zu beeinträchtigen?

Zugehörige Hypothesen:

zu (a)
H_0: $\beta_1 = \beta_2 = ... = 0$; β_0 beliebiger Intercept
H_1: $\beta_j \neq 0$ für mind. ein $j \in \{1, \cdots, p\}$

zu (b)
H_0: $\boldsymbol{\beta} \in \Omega_0$
H_1: $\boldsymbol{\beta} \in \Omega_1$; mit $\Omega_0 \subseteq \Omega_1$

Leicht zu erkennen: (a) ist Spezialfall von (b)

Darstellung linearer Hypothesen

- Lineare Hypothesen (C mit vollem Spaltenrang!)

$$\begin{aligned} H_0 :& \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\delta} \\ H_1 :& \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\delta} \end{aligned}$$

- Beispiel:

$$\begin{aligned} & H_0 : \ \beta_1 = \beta_3 \\ & \Leftrightarrow \underbrace{\begin{pmatrix} 0 & 1 & 0 & -1 & 0 & \dots \end{pmatrix}}_{\mathbf{C}} \boldsymbol{\beta} = 0 \\ & \Leftrightarrow \beta_1 - \beta_3 = 0 \end{aligned}$$

- Der Rang von **C** bestimmt die Anzahl an Hypothesen!

Gliederung

Form des LR-Tests

- Hypothesen:

  H_0: $\boldsymbol{\beta} \in \Omega_0$
  H_1: $\boldsymbol{\beta} \in \Omega_1$; mit $\Omega_0 \subseteq \Omega_1$

- Likelihood-Quotient(LQ)/Likelihood-Ratio(LR):

$$\Lambda = \frac{\max L(\boldsymbol{\beta}|\Omega_0)}{\max L(\boldsymbol{\beta}|\Omega_1)}$$

- Falls $\boldsymbol{\beta} \in \Omega_0 \;\Leftrightarrow\; \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\delta}$ (als lineare Hypothese) $\Rightarrow \Lambda \approx 1$!

Form des LR-Tests

- Teststatistik:

$$\lambda = -2\log(\Lambda) = -2[\ell(\tilde{\beta}) - \ell(\hat{\beta})]$$

  $\tilde{\boldsymbol{\beta}}$ ist ML-Schätzer unter H_0
  $\hat{\boldsymbol{\beta}}$ ist ML-Schätzer unter H_1

- Ablehnungsbereich:

$$\lambda > \chi^2_{1-\alpha}(rg(\mathbf{C}))$$

Gliederung

Wald-Test

- Prinzip: Betrachte die Differenz $\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}$
- Teststatistik:

$$w = (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\delta})^T (\mathbf{C}\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\delta})$$

- Falls H_0 nicht wahr $\Rightarrow \mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}$ groß!
- Spezialfall: $H_0 : \beta_r = 0$

$$w = \left(\frac{\hat{\beta}_r}{\sqrt{a_{rr}}}\right)^2 \overset{H_0}{\sim} \chi^2(1)$$

(a_{rr} ist r-tes Diagonalelement von $\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}$)

Scoretest

- Prinzip: Betrachte Scorefunktion

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{x}\mathbf{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})]$$

- Teststatistik:

$$u = \mathbf{s}(\tilde{\boldsymbol{\beta}})^T \mathbf{F}(\tilde{\boldsymbol{\beta}})^{-1}\mathbf{s}(\tilde{\boldsymbol{\beta}})$$

- Falls H_0 wahr, sollte $\mathbf{s}(\tilde{\boldsymbol{\beta}})$ ähnlich sein zu $\mathbf{s}(\hat{\boldsymbol{\beta}})$, also $\mathbf{0}$!

Vergleich von LR-, Wald- und Score-Test:

- Gemeinsamkeit: $\lambda,\ w,\ u \overset{H_0}{\sim} \chi^2(rg(\mathbf{C}))$
- Berechnung von w ausschließlich auf Basis von $\hat{\boldsymbol{\beta}}$
- Berechnung von u ausschließlich auf Basis von $\tilde{\boldsymbol{\beta}}$
- λ am kompliziertesten zu berechnen, da $\hat{\boldsymbol{\beta}}$ und $\tilde{\boldsymbol{\beta}}$ berechnet werden müssen. Aber dafür exakteste Teststatistik
- λ durch w und u approximierbar
- $\lambda,\ w$ und u asymptotisch effizient

Gliederung

Definition

- Maß für die Anpassungsgüte des Modells an die Daten
- Devianz

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -\phi 2 \left(\underbrace{\ell(\mathbf{y}, \hat{\boldsymbol{\mu}})}_{\text{Likelihood des Modells}} - \underbrace{\ell(\mathbf{y}, \mathbf{y})}_{\text{Likelihood des saturierten Modells}} \right)$$

- Betrachte Form der Exponentialfamilie

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} (y_i(\theta(y_i) - \theta(\hat{\mu}_i)) - (b(\theta(y_i)) - b(\theta(\hat{\mu}_i))))$$

- Beispiel: Normalverteilung

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

LR-Test und Devianz

Die Teststatistik des LR-Tests ist über die Devianzen der zu vergleichenden Modelle berechenbar:

$$
\begin{aligned}
\lambda &= -2\log(\Lambda) \\
&= -2[\ell(\mathbf{y}, \hat{\boldsymbol{\mu}}|\Omega_0) - \ell(\mathbf{y}, \hat{\boldsymbol{\mu}}|\Omega_1)] \\
&= -2[\ell(\mathbf{y}, \hat{\boldsymbol{\mu}}|\Omega_0) - \ell(\mathbf{y}, \mathbf{y}) - \ell(\mathbf{y}, \hat{\boldsymbol{\mu}}|\Omega_1) + \ell(\mathbf{y}, \mathbf{y})] \\
&= \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}}|\Omega_0) - D(\mathbf{y}, \hat{\mu}|\Omega_1)}{\phi}
\end{aligned}
$$

Gliederung

LR-Test mit R

[vgl. Problemstellungen auf Folie 4]

zu (a)
- Berechnung über Summary:

$$\frac{\text{Null Deviance - Residual Deviance}}{\text{Dispersion Parameter}} \overset{H_0}{\sim} \chi^2(rg(\mathbf{C}))$$

- Berechnung über `lrtest`(*Modell*) aus Package `lmtest`

zu (b)
- Berechnung über Summaries:

$$\frac{\text{Residual } D_{\Omega_0} - \text{Residual } D_{\Omega_1}}{\text{Dispersion Parameter}} \overset{H_0}{\sim} \chi^2(rg(\mathbf{C}))$$

- Berechnung über `lrtest`($Modell|\Omega_0$, $Modell|\Omega_1$)

[⇒Identische Dispersionsparameter ϕ für beide Modelle nötig!]