Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2014

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | Institut für Statistik | STABLAB Statistisches Beratungs Labor

Probleme bei der Regression und Diagnose

Gegeben sei das multiple Regressionsmodell (2.1) und (2.5):

$$Y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

Es geht darum herauszufinden, ob das Modell zur Analyse der jeweiligen Daten geeignet ist.

Da sich die Modellannahmen auf die Störterme beziehen, werden typischerweise die Residuen betrachtet.

Beachte, dass sich die Annahmen nicht auf die Randverteilung von Y beziehen.

Verschiedene Typen von Residuen

Beachte: Die Residuen ergeben sich aus den (unbekannten) Störtermen ε durch

$$\hat{\varepsilon} = Q\varepsilon = (I - X(X'X)^{-1}X')\varepsilon = (I - P)\varepsilon$$

Daher gilt:

$$Var(\hat{\varepsilon}_i) = q_{ii}\sigma^2 \tag{6.1}$$

Beispiel:

$$X := \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$$

$$Q := \begin{bmatrix} 2/5 & -2/5 & -1/5 & 0 & 1/5 \\ -2/5 & \frac{7}{10} & -1/5 & -1/10 & 0 \\ -1/5 & -1/5 & 4/5 & -1/5 & -1/5 \\ 0 & -1/10 & -1/5 & \frac{7}{10} & -2/5 \\ 1/5 & 0 & -1/5 & -2/5 & 2/5 \end{bmatrix}$$

Standardisierte Residuen:

$$r_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{q_{ii}}} \quad i = 1, \ldots, n \tag{6.2}$$

$q_{ii} = 1 - h_{ii}$: Elemente der Residualmatrix Q

Es gilt:

$$\begin{array}{ll} \frac{1}{n} \leq h_{ii} \leq 1 & \sum h_{ii} = p' \\ \Rightarrow \quad 0 \leq q_{ii} \leq 1 - \frac{1}{n} & \sum q_{ii} = n - p' \end{array}$$

Bemerkung:
Nach (6.2) sind die Residuen in der Mitte der Daten geringer.
Dies wird durch die Standardisierung ausgeglichen.

Studentisierte Residuen:

Problem: Bei der Schätzung von σ geht das Residuum mit ein.
Dies kann insbesondere bei kleinen Stichproben ein Problem sein.

Daher definiert man:

$$r_i^* := \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{q_{ii}}} \quad i = 1, \ldots, n \tag{6.3}$$

$\hat{\sigma}_{(i)}$:= Schätzung von σ **ohne** die Beobachtung i.

Kreuzvalidierungs - Residuen:

$$e_{(i)} := y_i - x_i' \hat{\beta}_{(i)} \tag{6.4}$$

$\hat{\beta}_{(i)}$: Schätzung von β ohne Beobachtung i

$$PRESS := \sum_{i=1}^{n} e_{(i)}^2$$

PRESS: (Predicted Residual Sum of Squares)

Es gilt:

$$\begin{aligned} e_{(i)} &= \hat{\epsilon}_i / q_{ii} \\ Var(e_{(i)}) &= \frac{1}{q_{ii}}\sigma^2 > \sigma^2 \end{aligned}$$

Rekursive Residuen:

Bei Zeitreihen verwendet man häufig:

$$\omega_i := \frac{y_i - x_i' \hat{\beta}_{[i-1]}}{\sqrt{1 - x_i'(X_{[i-1]}'X_{[i-1]})^{-1}x_i}} \quad i = p' + 1, \ldots, n \tag{6.5}$$

$\hat{\beta}_{[i-1]}$: Schätzung von β aus den ersten $i-1$ Beobachtungen

$X_{[i-1]}$: X-Matrix der ersten $i-1$ Beobachtungen

Die Störterme ε_i sind nicht normalverteilt

Ursachen: Die Y-Variable stellt eine Zählgröße, eine Überlebenszeit, oder einen Anteil dar. Y ist nicht-negativ etc.

Folgen: Der KQ-Schätzer $\hat{\beta}$ ist immer noch erwartungstreuer Schätzer mit kleinster Varianz. Der F-Test ist i. a. robust. Problematisch sind insbesondere bei kleinen Stichprobenumfängen die Konfidenzintervalle der Parameter. Außerdem sind die Prognoseintervalle nicht mehr gültig, da hierbei die NV-Annahme besonders eingeht.

Diagnose: Schiefe und Kurtosis der Verteilung der Residuen, Normal-Plots der Residuen

Therapie: Transformationen der Y-Variablen, Verwendung von generalisierten linearen Modellen

Heterogene Varianzen

Die Varianz der Störterme ε_i ist von i abhängig.

Ursachen: Y Zähldaten, Anteile. Gruppierte Daten führen zu verschieden Residualvarianzen innerhalb der Gruppen. Multiplikative Fehlerstruktur, d. h. σ_i ist abhängig von der Größe von Y_i.

Folgen: Schätzer für β ist erwartungstreu, aber er hat nicht mehr die kleinste Varianz. Konfidenzintervalle und Tests für β nicht mehr korrekt.

Diagnose: Residualplot der $\hat{\varepsilon}_i$ auf $\hat{Y}_i$, Berechnung der Residualvarianzen in den einzelnen Gruppen (bei gruppierten Daten).

Therapie: Transformation der Y-Variable, Gewichtete KQ-Schätzung

Korrelation zwischen den Störtermen

Es gilt $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$ für einige $i \neq j$.

Ursachen: Zeitreihenstruktur oder räumliche Struktur der Daten führen zu positiver Korrelation von aufeinander folgenden (bzw. nahen) Beobachtungen. Residuen bei gruppierten Beobachtungen, bei denen die Gruppenzugehörigkeit nicht zusätzlich modelliert wird, sind häufig positiv korreliert.

Folgen: Schätzer von β erwartungstreu, aber nicht mit geringster Varianz. Bias bei der Varianzschätzung führt zu fehlerhaften Konfidenzintervallen und zu Problemen bei den F-Tests.

Diagnose: Analyse der Zeitreihenstruktur der Residuen, z.B. mit Durbin–Watson–Test, Plots der Residuen gegen die Zeit, Plots von $\hat{\varepsilon}_i$ gegen $\hat{\varepsilon}_{i-1}$.

Therapie: Verwendung von Zeitreihenmethoden, Einbeziehung von Trend und Saison, Gewichtete KQ-Methode

Durbin-Watson-Test

Um zu testen, ob die Störterme ε_i und ε_{i-1} korreliert sind benutzt man folgende Testgröße:

$$d := \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} \approx 2(1 - \hat{\rho})$$

$\hat{\rho}$: Korrelation zwischen $\hat{\varepsilon}_i$ und $\hat{\varepsilon}_{i-1}$.

H_0: $\rho = 0$
Lehne H_0 ab, falls $d > d_1$ oder $d < d_2$.
(d_1, d_2 sind von p und n abhängige feste Werte).

Kleine Werte von d: positives ρ
Große Werte von d: negatives ρ

$d \approx 2 \longrightarrow$ keine Autokorrelation

Ausreißer und Punkte mit starkem Einfluss

Einflussreiche Beobachtungen (high leverage points) sind in den X-Werten weit vom Zentrum der Daten entfernt.
Ausreißer haben dem Betrag nach sehr große Störterme.

Ursachen: Falsche Erhebung, Beobachtung gehört nicht zur Grundgesamtheit, Besonderheiten bei einzelner Untersuchungseinheit

Folgen: Einflussreiche Beobachtungen wirken stark auf die Schätzung von β. Ausreißer können zu erheblicher Verzerrung der Schätzung von β führen. Dies gilt besonders für Ausreißer, die gleichzeitig high leverage points sind.

Diagnose: Analyse der Diagonalelemente der Hat-Matrix P zum Auffinden von high leverage points, verschiedene Residuenplots zur Ausreißeranalyse, Influence-Statistiken

Therapie: fehlerhafte Daten weglassen, robuste Regression, gewichtete Regression

Wichtige Einflussmaße: Leverage

Das i-te Diagonalelement der Hat-Matrix P

$$h_{ii} := x_i'(X'X)^{-1}x_i \tag{6.6}$$

heißt **Leverage** der Beobachtung x_i.

Es gilt:

$$\frac{1}{n} \leq h_{ii} \leq 1$$

Normalwert: $h_{ii} = \frac{p'}{n}$
großer Wert: $h_{ii} > \frac{2p'}{n}$

Cook's Distanz:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\sigma}^2 p'} \tag{6.7}$$

$\hat{\beta}_{(i)}$: Schätzung von β ohne Beobachtung i

Es gilt:

$$D_i = \underbrace{\frac{r_i^2}{p'}}_{\text{Residuen}} \cdot \underbrace{\frac{h_{ii}}{1-h_{ii}}}_{\text{Leverage}} \tag{6.8}$$

$$D_i = \frac{\left(\hat{Y}_{(i)} - \hat{Y}\right)'\left(\hat{Y}_{(i)} - \hat{Y}\right)}{p'\hat{\sigma}^2} \tag{6.9}$$

Cooks Distanz wird im Vergleich mit den Beobachtungen interpretiert.

Regressionsgleichung ist nicht korrekt

Die Gleichung $y = X\beta + \varepsilon$ ist fehlerhaft.

Ursachen: Variablen wurden weggelassen oder überflüssigerweise in das Modell einbezogen. Der Zusammenhang ist nicht linear, Interaktionen werden nicht in das Modell einbezogen

Folgen: Systematische Fehler bei der Schätzung der Modellparameter und bei der Prognose, aber Modellschätzung liefert häufig brauchbare Näherung

Diagnose: Residuenplots $\hat{\varepsilon}_i$ gegen $\hat{y}_i$, F-Tests auf Einfluss von weiteren Variablen, Interaktionen, Polynomterme höherer Ordnung etc.

Therapie: Modellerweiterung, Transformationen der Einflussgrößen, Variablenselektionsverfahren.

Partial Leverage Plot:

y^* auf x_k^* mit

$y^* := Q_{(k)}y$
$x_k^* := Q_{(k)}x_k$

$Q_{(k)}$: Q-Matrix der Einflussgrößen **ohne** Variable k

$\Longrightarrow$ Darstellung des Zusammenhangs zwischen y und der Einflussgröße x_k unter Berücksichtigung der übrigen Einflussgrößen.

Kollinearität

Die Spalten von X sind (annähernd) linear abhängig.

Ursachen Hohe Korrelation zwischen den Einflussgrößen, Ungünstiges Versuchs-Design, Codierung von diskreten Variablen

Folgen: Ungenaue Schätzung von β, häufig sogar falsches Vorzeichen.
Aber: Konfidenzintervalle korrekt und damit entsprechend groß.

Diagnose: Analyse der Matrix $(X'X)$ und der Korrelationsmatrix der metrischen Einflussgrößen.

Therapie: Zusammenfassen bzw. Weglassen von Einflussgrößen, Verwendung von anderen Schätzmethoden, z.B.: Ridge-Regression

Kollinearitätsdiagnostik

a) **Konditionszahl**

$$K(X) := \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{6.10}$$

$\lambda_{min}, \lambda_{max}$: minimaler, bzw. maximaler Eigenwert von X'X

b) **Varianz Inflationsfaktor**

$$\text{VIF}_j := \frac{1}{1 - R_j^2} \tag{6.11}$$

R_j^2: Bestimmheitsmaß der Regression von x_j auf die übrigen x.
Es gilt für die Varianz von β_j:

$$\sigma_{\hat{\beta}_j} = \frac{\sigma^2}{(x_j - \bar{x})'(x_j - \bar{x})}\text{VIF}_j \tag{6.12}$$

Beachte: VIF ist auch für kategoriale Variablen geeignet.

Messfehler

Ursachen Messfehler im engeren Sinne (Messgerät) und im weiteren Sinne (z.B. falsche Beantwortung von Fragen).

Folgen Meist systematische betragsmäßige Unterschätzung der zu den fehlerhaft gemessenen Größen gehörigen β_i. Geringere Power der entsprechenden F-Tests.

Diagnose Mehrfach-Messungen der entsprechenden Größen

Therapie Verwendung von Korrektur-Verfahren $\longrightarrow$ Theorie der Fehler-in-den-Variablen-Modelle