Robuste Schätzung von Standardfehlern bei longitudinalen Daten

Julia Braun

Abteilung Biostatistik

3. April 2008

Übersicht

1. Einleitung
2. Modell
3. Robuster Ansatz
4. Beispiel 1: Ratten
5. Beispiel 2: Haarausfall
6. Zusammenfassung

Modellierung longitudinaler Daten

Zur Modellierung longitudinaler Daten: Annahmen über Korrelationsstruktur benötigt, z. B.

- konstante Korrelation
- exponentiell abfallende Korrelation

Problem: Annahmen oft inkorrekt

$\Longrightarrow$ Verwendung von Methoden, die möglichst robust gegenüber einer Fehlspezifikation der Kovarianzstruktur sind.

Allgemeines lineares Modell

- Balancierte Daten

$$Y_{ij} = \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij}$$

  mit
  $i = 1, ..., m$ Individuen,
  $j = 1, ..., n$ Zeitpunkten sowie
  $k = 1, ..., p$ erklärenden Variablen.
- ϵ_{ij} sind innerhalb eines Individuums korreliert.
- Modellannahme:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$$

- $\mathbf{V}$ muss block-diagonal sein mit $n \times n$ Einträgen $\mathbf{V_0}$, die den Beobachtungen eines Individuums entsprechen.

Gewichtete KQ-Schätzung

Minimierung der quadratischen Form

$$(\mathbf{y} - \mathbf{X}\beta)'\mathbf{W}(\mathbf{y} - \mathbf{X}\beta)$$

ergibt Schätzer

$$\hat{\beta}_{\mathbf{W}} = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}$$

mit Kovarianzmatrix

$$Cov(\hat{\beta}_{\mathbf{W}}) = \sigma^2\{(\mathbf{X'WX})^{-1}\mathbf{X'W}\}\mathbf{V}\{\mathbf{WX}(\mathbf{X'WX})^{-1}\}.$$

Eigenschaften

- Schätzung ist für beliebiges $\mathbf{W}$ unverzerrt.
- Bei Wahl $\mathbf{W} = \mathbf{V}^{-1}$ ist

$$\hat{\beta} = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{y}$$

  der ML-Schätzer unter Normalverteilungsannahme.
- Dieser hat die Kovarianzmatrix $Cov(\hat{\beta}) = \sigma^2(\mathbf{X'V}^{-1}\mathbf{X})^{-1}$.
- Problem: V ist unbekannt!

Schätzung der Kovarianzmatrix

Möglichkeit 1: Maximum Likelihood

- Für festes $\mathbf{V}_0$ erhält man ML-Schätzer $\hat{\boldsymbol{\beta}}(\mathbf{V}_0)$ und $\hat{\sigma}^2(\mathbf{V}_0)$.
- Einsetzen von $\hat{\boldsymbol{\beta}}(\mathbf{V}_0)$ und $\hat{\sigma}^2(\mathbf{V}_0)$ in die Likelihood
- Maximierung der reduzierten Log-Likelihood liefert ML-Schätzer $\hat{\mathbf{V}}_0$.
- Einsetzen von $\hat{\mathbf{V}}_0$ liefert $\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}}_0)$ und $\hat{\sigma}^2(\hat{\mathbf{V}}_0)$.

Nachteile:

- Meist numerische Verfahren zur Maximierung nötig
- ML-Schätzung von σ^2 und $\mathbf{V}_0$ ist in der Regel verzerrt.
- Besonders problematisch, wenn viele Einflussgrössen

Schätzung der Kovarianzmatrix

Möglichkeit 2: Restricted Maximum Likelihood (REML)

- Schätzung der Varianzkomponenten σ^2 und $\mathbf{V}_0$ in einem GLM
- Basiert auf dem Prinzip der marginalen Likelihood
- Vorgehen sehr ähnlich wie bei ML-Schätzer
- Ein deutlicher Unterschied zwischen ML und REML ergibt sich nur, wenn p gross ist.

Robuster Ansatz

Gewichtete KQ-Schätzung mit Arbeitskorrelationsmatrix

Vorteile

- Benötigt keine parametrischen Annahmen
- Unempfindlich gegenüber Missspezifikation der Kovarianzstruktur
- $\hat{\beta}_W$ ist immer erwartungstreu.

Sinnvoll, wenn

- kein Vorwissen über Kovarianzstruktur
- relativ kurze und komplette Sequenzen
- primäres Ziel: Inferenz für mittleren Response

Gewichtete KQ-Schätzung mit Arbeitskorrelationsmatrix

- Verwendung des gewichteten KQ-Schätzers

$$\hat{\beta}_{\mathbf{W}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

  mit Arbeitskorrelationsmatrix $\mathbf{W}$
- $\mathbf{W}$ muss nur bis auf einen Proportionalitätsfaktor gewählt werden.
- Geschätzte robuste Kovarianzmatrix:

$$\hat{\mathbf{R}}_{\mathbf{W}} = \sigma^2\{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\}\hat{\mathbf{V}}\{\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\}$$

- Schätzung von $\mathbf{V}_0$ und σ^2, z. B. mit REML
- Unterschied zum parametrischen Ansatz: Konfidenzintervalle und Hypothesentests sind asymptotisch korrekt, unabhängig von der wahren Form von $\mathbf{V}$.

Situation

- Geplante Versuche zu gleichen Zeitpunkten
- Vollständige Daten
- Saturiertes Modell ohne Kovariablen:

$$E(Y_{ij}) = \mu_j$$

- Wahl für Arbeitskorrelationsmatrix $\mathbf{W}^{-1} = \mathrm{I}$
- Designmatrix:

$$\mathbf{X}' = [\mathrm{I}, \mathrm{I}, ..., \mathrm{I}]$$

Schätzungen

- Als Schätzer für die μ_j dienen die jeweiligen Stichprobenmittelwerte (entspricht KQ-Methode)

$$\hat{\mu}_j = \frac{1}{m} \sum_{i=1}^{m} y_{ij}$$

- REML-Schätzung für $\sigma^2 \mathbf{V}_0$:

$$\hat{\sigma}^2 \hat{\mathbf{V}}_0 = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'$$

- Die benötigte Matrix $\hat{\sigma}^2 \hat{\mathbf{V}}$ ist dann block-diagonal mit den Einträgen $\hat{\sigma}^2 \hat{\mathbf{V}}_0$.
- Einsetzen ergibt die Schätzer $\hat{\boldsymbol{\beta}}_{\mathbf{W}}$ und $\hat{\mathbf{R}}_{\mathbf{W}}$.

Testen von Hypothesen

- Nullhypothese:

$$\mathbf{Q}\boldsymbol{\beta} = 0,$$

  wobei $\mathbf{Q}$ von vollem Rang mit Dimension $q \times p$
- Es gilt:

$$\mathbf{Q}\hat{\beta}_{\mathbf{W}} \sim MVN(\mathbf{Q}\beta, \mathbf{Q}\hat{\mathbf{R}}_{\mathbf{W}}\mathbf{Q}').$$

- Teststatistik:

$$T = \hat{\beta}'_{\mathbf{W}}\mathbf{Q}'(\mathbf{Q}\hat{\mathbf{R}}_{\mathbf{W}}\mathbf{Q}')^{-1}\mathbf{Q}\hat{\beta}_{\mathbf{W}}.$$

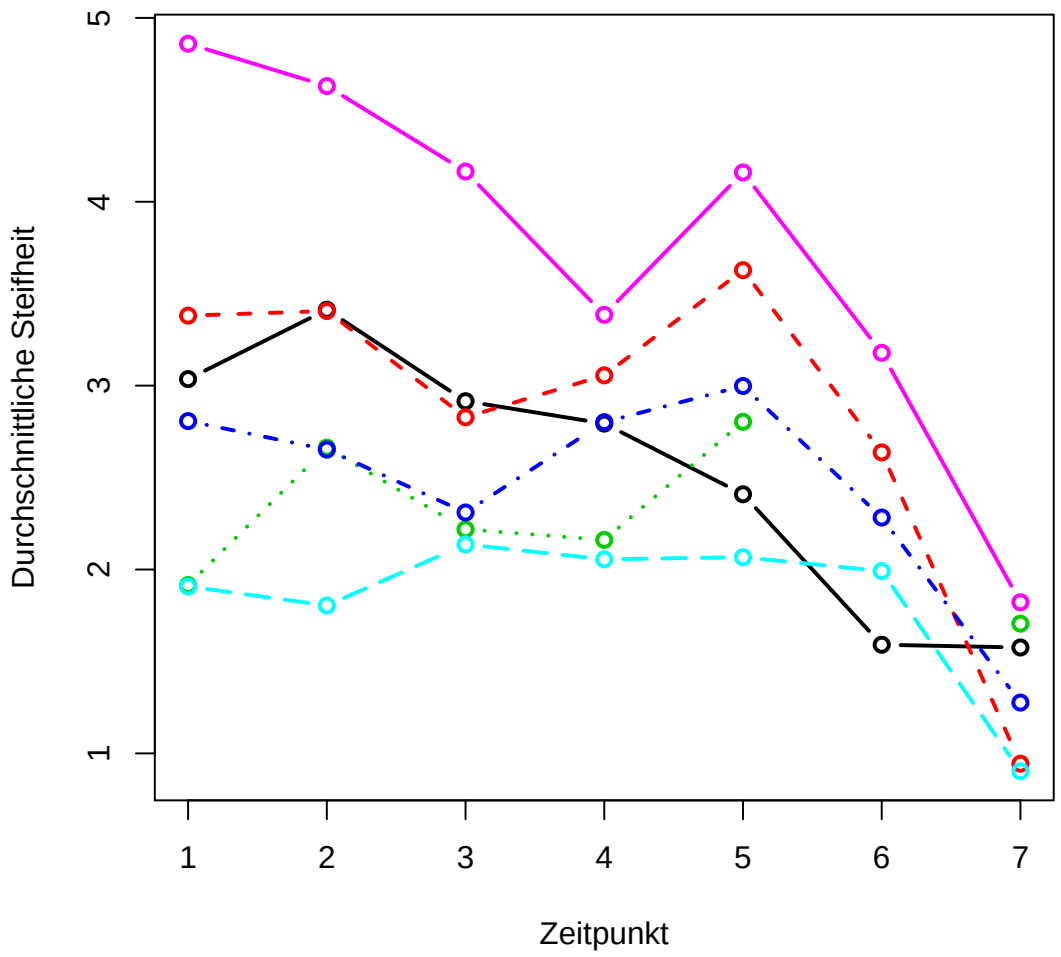
- T ist unter H_0 χ^2-verteilt mit q Freiheitsgraden.

Beispiel 1: Ratten

Fragestellung

- Neu entwickelte Maschine zur Überprüfung der Beweglichkeit des Schultergelenks
- Daten von 6 Rattenschultern
- Ratten werden nach Betäubung in Maschine eingespannt
- Messung von maximaler Steifheit, durchschnittlicher Steifheit, Energie, Flexion, Bewegungsradius
- Danach jeweils Durchtrennung einer Sehne $\longrightarrow$ 7 Situationen
- Ziel: Nachweis von Unterschieden in der Beweglichkeit zwischen den 7 Zeitpunkten

Durchschnittliche Steifheit



Ergebnisse

- Bei einer Schulter fehlt Messung 6 $\rightarrow$ Zeitpunkt 6 weggelassen
- Verwendung der R-Funktion gee oder "händische" Programmierung
- Globaler Test auf Unterschiede
- Danach Einzeltests zum Vergleich aller Zeitpunkte mit Zeitpunkt 1 mit Bonferroni-Korrektur

Globaler Test: $p < 0.001$
Einzeltests:

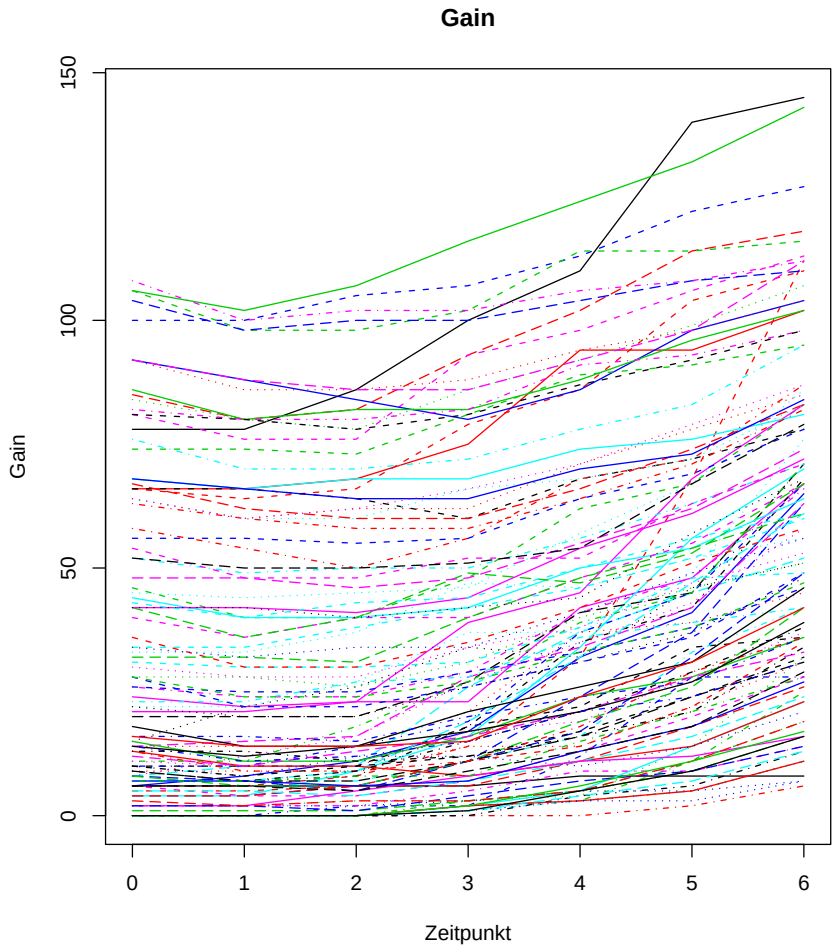
	Zeit 2	Zeit 3	Zeit 4	Zeit 5	Zeit 7
Zeit 1	0.476	0.199	0.281	0.914	< 0.001

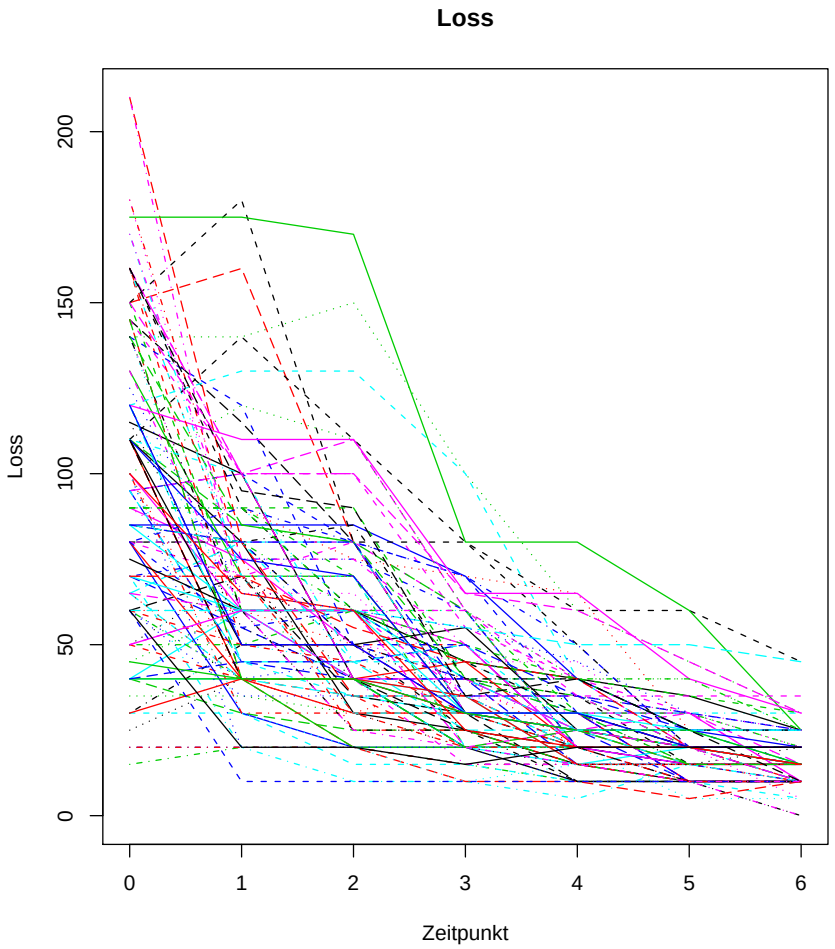
Beispiel 2: Haarausfall

Fragestellung

- 121 Probanden
- Therapie mit Medikament gegen Haarausfall
- Ausmass des Haarausfalls (Loss) und neuer Haarwuchs (Gain) gemessen
- 1 Messung vor Behandlung, danach 6 Messungen in monatlichen Abständen
- Ziel: Unterschiede zwischen den Zeitpunkten nachweisen

Gain und Loss





Ergebnisse

- Gleiches Vorgehen wie bei Ratten
- Globaler Test signifikant für beide Variablen
- Einzelne Vergleiche mit Zeitpunkt 0 signifikant ab Zeitpunkt 4 bzw. Zeitpunkt 3

Gain:

	Zeit 1	Zeit 2	Zeit 3	Zeit 4	Zeit 5	Zeit 6
Zeit 0	0.087	0.532	0.156	0.002	< 0.001	< 0.001

Loss:

	Zeit 1	Zeit 2	Zeit 3	Zeit 4	Zeit 5	Zeit 6
Zeit 0	0.043	0.010	< 0.001	< 0.001	< 0.001	< 0.001

Zusammenfassung

Vor- und Nachteile:

- Kein Vorwissen über Korrelationsstruktur nötig
- Gut geeignet, wenn primär Vergleich der Zeitpunkte gewünscht
- Schnell und einfach zu berechnen
- Evtl. numerisch instabil (z.B. bei n klein)
- Relativ starke Anforderungen: Balancierte Daten, keine fehlenden Werte,...
- Bei zeitlich variierenden Einflussgrössen: REML-Schätzung für $\mathbf{V}_0$ nur noch numerisch möglich

Zusammenfassung

Fazit:

- Gut geeignet bei wenig Vorwissen und bei klarer Fragestellung
- Wenn Vorwissen oder Datensatz mit vielen Zeitpunkten: parametrischer Ansatz
- Besseres Vorgehen als Bonferroni-Korrektur?

Literatur:

Diggle, P. J., Heagerty, P. J., Liang, K.-L., Zeger, S. L. (2002). Analysis of Longitudinal Data, Oxford University Press, 54-80.