ETH zürich

# Varianzkomponentenschätzung


Peter von Rohr

Qualitas AG

# Multiple Lineare Regression

## Annahmen

- Modell

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- Varianz der Fehler $e_i$ ist konstant, d.h. $Var(e_i) = \sigma^2$ für alle $i$
- Fehler $e_i$ sind nicht korreliert, d.h. $Cov(e_i, e_j) = 0$ für $i \neq j$
- Fehler $\mathbf{e}$ ist multivariat normal verteilt

## Momente

- Erwartungswerte: $E\left[\mathbf{e}\right] = \mathbf{0}$, $E\left[\mathbf{y} = \mathbf{X}\mathbf{b}\right]$
- Varianzen: $Var(y_i) = \sigma^2$, $Var(e_i) = \sigma^2$ für alle $i$

# Parameterschätzung

- Im linearen Model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ sind $\mathbf{b}$ und $\sigma^2$ unbekannte Parameter
- Parameter sollen als Funktion der Daten ($\mathbf{y}$) geschätzt werden
- Ein möglicher Schätzer $\hat{b}_{LS}$ für $\mathbf{b}$ kann mit **Least Squares** (kleinste Quadrate) berechnet werden

$$\hat{\mathbf{b}}_{LS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- **Aber** Least Squares gibt keine Schätzung für $\sigma^2$

# Schätzung der Varianzkomonente $\sigma^2$

Eine Beobachtung betreffend der Eigenschaft der **Erwartungstreue** (Unbiasedness) eines Schätzers führt zu folgendem Ergebnis

- Erwartungstreue (Unbiasedness) ist eine Eigenschaft eines Schätzers $\hat{\theta}$ für einen Parameter $\theta$, sofern gilt, dass $E\left[\hat{\theta}\right] = \theta$
- Erwartungswert der Summe der quadrierten Residuen ($r_i$)

$$E\left[\sum_{i=1}^{n} r_i^2\right] = \sum_{i=1}^{n} E\left[r_i^2\right] = \sum_{i=1}^{n} Var(r_i) = \sigma^2 \sum_{i=1}^{n}(1 - P_{ii})$$

$$= \sigma^2(n - tr(P)) = \sigma^2(n - p)$$

wobei: $r_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^T \hat{b}$ und Matrix $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$,
$n$: Anzahl Beobachtungen und $p$: Anzahl Parameter

# Schätzung der Varianzkomonente $\sigma^2$ II

- Somit gilt Erwartungstreue für den Schätzer

$$\hat{\sigma^2} = \frac{1}{(n-p)} \sum_{i=1}^{n} r_i^2$$

- da für die Erwartungswerte gilt

$$E\left[\hat{\sigma^2}\right] = E\left[\frac{1}{(n-p)} \sum_{i=1}^{n} r_i^2\right] = \frac{1}{(n-p)} E\left[\sum_{i=1}^{n} r_i^2\right]$$

$$= \frac{1}{(n-p)} \left[\sigma^2 (n-p)\right] = \sigma^2$$

- Dieser Schätzer wird oft als LS-Schätzer bezeichnet

# Varianzanalyse - ANOVA

- Effekte oder Faktoren im Modell (bis jetzt $x$ Variable) sind kategorisch oder qualitativ
- Ziel ist es herauszufinden, wie viel von der totalen empirischen Varianz in den Beobachtungen durch die Faktoren erklärt werden kann.
- Folgendes Modell wird angenommen:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

  wobei: Faktor $\alpha$ $i = 1, ..., I$ Stufen aufweist und $j = 1, ..., J_i$ Beobachtungen pro Stufe vorliegen.
- Effekte sind nur unter gewissen Restriktionen schätzbar. Häufig wird die Summe aller Effekte $\sum_i J_i \alpha_i = 0$ gesetzt

# Parameterschätzung ANOVA

- Mit Restriktion $\sum_i J_i \alpha_i = 0$ folgen Schätzer

$$\hat{\mu} = \bar{y} \text{ und } \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}$$

  wobei: $\bar{y}$ der Mittelwert über alle Beobachtungen und $\bar{y}_{i.}$ der Mittelwert über alle Beobachtungen auf Stufe $i$ ist.
- Die Varianzanteile werden aufgrund der Summenquadrate bestimmt

# Tabelle der Varianzanalyse - ANOVA Table

| Quelle | DF | SQ | MSQ | F | P |
|---|---|---|---|---|---|
| Faktor $\alpha$ | $i-1$ | SQZ | MSQZ | MSQZ/MSQE | |
| Fehler $e$ | $n-i$ | SQE | MSQE | | |
| Total | $n-1$ | SQT | | | |

wobei

$$SQT = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y})^2$$

$$SQZ = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{y}_{i.} - \bar{y})^2;\ MSQZ = SQZ/(i-1)$$

$$SQE = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2;\ MSQE = SQE/(n-i)$$

# Beispieldatensatz

Koagulationsdatensatz von
https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

```
> if (!require(faraway)) {
+   install.packages("faraway");require(faraway)}
> data("coagulation")
> head(coagulation, n=5)
```

```
  coag diet
1   62    A
2   60    A
3   63    A
4   59    A
5   63    B
```

# Beispieldatensatz II

```
> summary(coagulation)
      coag          diet
 Min.   :56.00   A:4
 1st Qu.:61.75   B:6
 Median :63.50   C:6
 Mean   :64.00   D:8
 3rd Qu.:67.00
 Max.   :71.00
```

# Beispieldatensatz Anova

```
> lmCoDi <- lm(coag ~ diet, data = coagulation)
> anova(lmCoDi)

Analysis of Variance Table

Response: coag
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05
```

# Maximum Likelihood

- Definition der Likelihood als gemeinsame Dichteverteilung der Daten ($\mathbf{y}$) gegeben die Parameter ($\mathbf{b}$, $\sigma^2$)

$$Pr\left(\mathbf{y}|\mathbf{b}, \sigma^2\right)$$

- Unter der Annahme einer Normalverteilung und nach Beobachtung der Daten als Funktion der Parameter aufgefasst, folgt

$$L\left(\mathbf{b}, \sigma^2; \mathbf{y}\right) = Pr\left(\mathbf{y}|\mathbf{b}, \sigma^2\right) = (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} exp\left[-\frac{(y_i - \mathbf{x}_i^T\mathbf{b})^2}{2\sigma^2}\right]$$

# Maximierung der Likelihood Funktion

- Maximierung der Likelihood Funktion $L\left(\mathbf{b}, \sigma^2; \mathbf{y}\right)$ ist analog zur Maximierung von

$$l\left(\mathbf{b}, \sigma^2; \mathbf{y}\right) = \log\left(L\left(\mathbf{b}, \sigma^2; \mathbf{y}\right)\right)$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{b})^2$$

- Schätzungen für $\mathbf{b}$ und $\sigma$ erhalten wir durch Nullsetzen der partiellen Ableitungen von $l\left(\mathbf{b}, \sigma^2; \mathbf{y}\right)$ nach $\mathbf{b}$ und $\sigma$

# Partielle Ableitung nach b

- Partielle Ableitung von $l\left(\mathbf{b}, \sigma^2; \mathbf{y}\right)$ nach $\mathbf{b}$

$$\frac{\partial l}{\partial \mathbf{b}} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\mathbf{x}_i(y_i - \mathbf{x}_i^T\mathbf{b})$$

- Extremum ist erreicht, falls partielle Ableitung $= 0$, somit ist

$$\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\hat{\mathbf{b}} = \sum_{i=1}^{n} \mathbf{x}_i y_i$$

$$\hat{\mathbf{b}}_{ML} = \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

- Beobachtung: $\hat{\mathbf{b}}_{LS} = \hat{\mathbf{b}}_{ML}$

# Schätzer für $\sigma^2$

- Analoges Vorgehen, wie für $\mathbf{b}$
- Partielle Ableitung von $l\left(\mathbf{b}, \sigma^2; \mathbf{y}\right)$ nach $\sigma^2$ null setzen
- Für $\mathbf{b}$ setzen wir $\hat{\mathbf{b}}_{ML}$ ein

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{b})^2$$

$$n\hat{\sigma^2}_{ML} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\mathbf{b}}_{ML})^2$$

$$\hat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\mathbf{b}}_{ML})^2$$

- $\hat{\sigma^2}_{ML}$ ist nicht erwartungstreu, d.h. $E\left[\hat{\sigma^2}_{ML}\right] \neq \sigma^2$

# REML

- Restricted oder Residual Maximum Likelihood
- Basiert auf der Log-likelihood Funktion

$$l_{REML}(\sigma^2; \mathbf{y}) = -\frac{n-p}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i^2$$

wobei:

$$r_i = y_i - \mathbf{x}_i^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

- $l_{REML}$ ist unabhängig von $\mathbf{b}$
- Maximierung von $l_{REML}$ führt zum Schätzer

$$\hat{\sigma^2}_{REML} = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2$$

# Bayes'sche Methoden

## Philosophie

- Welt der Statistik unterteilt sich in zwei Lager
  1. Frequentisten
  2. Bayesianer

## Frequentisten

- Parameterschätzungen basieren auf ML oder REML
- Trennung zwischen Daten und Parameter, fehlende Daten werden ignoriert
- Effekte in Modellen werden in fix und zufällig unterteilt
- Keine Berücksichtigung von a priori Information

# Bayes'sche Methoden II

## Bayesianer

- Unterteilung in bekannte und unbekannte Grössen, dies können Daten oder Parameter sein
- Schätzung von unbekannten Grössen basieren auf der a posteriori Verteilung der unbekannten Grössen gegeben die bekannten Grössen
- Fehlende Daten können berücksichtigt werden
- A priori Information wird berücksichtigt

# Beispiel für Bayes'sche Parameterschätzung

- Modell: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$
- Bekannte Grössen sind die beobachteten Daten $\mathbf{y}$
- Unbekannte Grössen sind die Parameter $\mathbf{b}$ und $\sigma$
- Schätzungen für Unbekannte $\mathbf{b}$ und $\sigma$ werden gemacht aufgrund der a posteriori Dichteverteilungen $f(\mathbf{b}|\mathbf{y}, \sigma^2)$ und $f(\sigma^2|\mathbf{y}, \mathbf{b})$
- A posteriori Dichteverteilungen werden aufgrund des Satzes von Bayes berechnet

# Satz von Bayes

$$f(\mathbf{b}|\mathbf{y}, \sigma^2) = \frac{f(\mathbf{y}|\mathbf{b}, \sigma^2) f(\mathbf{b}) f(\sigma^2)}{f(\mathbf{y}, \sigma^2)}$$

wobei

| | |
|---|---|
| $f(\mathbf{y}\|\mathbf{b}, \sigma^2)$ | Bayes'sche Likelihood |
| $f(\mathbf{b})$ | a priori Dichteverteilung von $\mathbf{b}$ |
| $f(\sigma^2)$ | a priori Dichteverteilung von $\sigma^2$ |
| $f(\mathbf{y}, \sigma^2)$ | Normalisierungskonstante |

# Bayes'sche Schätzung für b

- Für Schätzung werden nur die von $\mathbf{b}$ abhängigen Teile aus $f(\mathbf{b}|\mathbf{y}, \sigma^2)$ verwendet
- Konstante $f(\sigma^2)$ und $f(\mathbf{y}, \sigma^2)$ werden ignoriert
- Daraus folgt

$$f(\mathbf{b}|\mathbf{y}, \sigma^2) \propto f(\mathbf{y}|\mathbf{b}, \sigma^2) f(\mathbf{b})$$

- Ist a priori nichts über $\mathbf{b}$ bekannt, dann wird $f(\mathbf{b})$ als konstant angenommen (uninformativer Prior)
- Weitere Vereinfachung zu

$$f(\mathbf{b}|\mathbf{y}, \sigma^2) \propto f(\mathbf{y}|\mathbf{b}, \sigma^2)$$

# Bayes'sche Schätzung für b II

- Unter der Annahme der Normalverteilung folgt

$$f(\mathbf{y}|\mathbf{b}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})}{2\sigma^2}\right\}$$

- Damit die Dimensionen konsistent sind, machen wir folgende Transformation

$$f(\mathbf{b}|\mathbf{y}, \sigma^2) \propto exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})}{2\sigma^2}\right\}$$

$$\propto exp\left\{-\frac{(\mathbf{b} - \hat{\mathbf{b}})^T(\mathbf{X}^T\mathbf{X})(\mathbf{b} - \hat{\mathbf{b}})}{2\sigma^2}\right\}$$

wobei: $\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

# Bayes'sche Schätzung III

- Schätzung für $\sigma^2$ analog basierend auf $f(\sigma^2|\mathbf{y}, \mathbf{b})$
- Schätzwerte basieren auf Momenten (Erwartungswert und Varianz) der a posteriori Dichteverteilungen
- Entsprechen die a posteriori Dichteverteilungen keinen Standarddichteverteilungen (wie Normal, Student-t, Beta, Gamma, ...), dann werden Erwartungswert und Varianzen aufgrund von Zufallszahlen von diesen Dichten approximiert
- Zufallszahlen werden aufgrund von Markovketten generiert