Statistische Modellierung III

-Wiederholung: Lineare Modelle-

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2019/2020

Das klassische lineare Regressionsmodell

Klassisches lin. Modell - Setup

- Unabhängige Beobachtungen $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots n$ einer Zielvariable Y_i und eines k-dim. Kovariablenvektors $\mathbf{x}_i$
- Annahme: $Y_i \sim N(\mathbf{x}_i\beta, \sigma^2)$ für $\beta \in \mathbb{R}^k$ und $\sigma^2 > 0$
- Die Daten folgen somit dem linearen Regressionsmodell

$$Y_i = \mathbf{x}_i\beta + U_i \text{ mit } U_i \sim N(0, \sigma^2), \; i = 1, \ldots, n$$

Modell in Matrixschreibweise

- Definiere

$$\mathbf{Y} = (Y_1, \dots Y_n)^T = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \mathbf{U} = (U_1, \dots, U_n)^T = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$$

- Dann ist $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, wobei $\mathbf{U} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ mit $\mathbf{0} = (0, \dots 0)^T \in \mathbb{R}^n$ und $\mathbf{I}_n$ die n-dimensionale Einheitsmatrix ist.
- Insgesamt folgt also: $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, wobei $N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ die mutlivariate Normalverteilung mit Erwartungswertvektor $\mathbf{X}\beta$ und Kovarianzmatrix $\sigma^2 \mathbf{I}_n$ ist.
- $\mathbf{X}$ wird Designmatrix genannt

Der KQ-Schätzer und seine Eigenschaften

Der kleinste Quadratschätzer (KQ-Schätzer)

- Zum schätzen der Regressionskoeffizienten wird in der Regel die KQ-Methode eingesetzt
- Ziel: Finde $\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i \beta)^2$
- Dieses Minimierungsproblem führt via Differentiation zum Normalgleichungssystem:
$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{Y}$$
- Hat $\mathbf{X}$ vollen Rang, dann ist $\mathbf{X}^T \mathbf{X}$ invertierbar und das Normalgleichungssystem hat die Lösung
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Eigenschaften des KQ-Schätzers

- Allgemein gilt:
  - Erwartungstreue: $E(\hat{\beta}) = \beta$,
  - Kovarianzmatrix: $Cov(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
- Ist zudem $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$, so gilt
  - $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
  - $(\hat{\beta} - \beta)^T\,\mathbf{X}^T\mathbf{X}\,(\hat{\beta} - \beta)/\sigma^2 \sim \chi^2_k$
  - $\hat{\beta}$ ist MLE für β

Eigenschaften des KQ-Schätzers

Gauß-Markov-Theorem

Gilt das lineare Modell $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, mit $E(\mathbf{U}) = \mathbf{0}$ und $Cov(\mathbf{U}) = \sigma^2 \mathbf{I}_n$ und hat die Designmatrix $\mathbf{X}$ vollen Rang, dann ist der KQ Schätzer $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ der beste lineare unverzerrte Schätzer (engl. **B**est **L**inear **U**nbiased **E**stimator, BLUE). Das heißt für jeden anderen linearen und erwartungstreuen Schätzer $\tilde{\beta}$ von β gilt:

$$Var(\mathbf{c}^T\hat{\beta}) \leq Var(\mathbf{c}^T\tilde{\beta}), \quad \forall \mathbf{c} \in \mathbb{R}^k$$

und $\exists \mathbf{c} \in \mathbb{R}^k$ mit $Var(\mathbf{c}^T\hat{\beta}) < Var(\mathbf{c}^T\tilde{\beta})$.

Asymptotische Eigenschaften des KQ-Schätzers

- Betrachten nun eine unendliche Folge von fixen (nicht-zufälligen) Kovariablenvektoren $\mathbf{x}_i$ und stochastisch unabhängigen Beobachtungen Y_i, $i = 1, 2 \ldots$, mit $Y_i = \mathbf{x}_i \beta + U_i$, wobei $E(U_i) = 0$ und $Var(U_i) = \sigma^2$ für alle i
- Nehmen nicht mehr an, dass Y_i normalverteilt sind
- Betrachten für jedes endliche n die KQ-Schätzer der ersten n Beoachtungen,

$$\hat{\beta}_n = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n$$

wobei $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ und $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^T$ die Designmatrix und der Zielvariablenvektor der ersten n Beobachtungen sind

Asymptotische Eigenschaften des KQ-Schätzers

- Unter der Annahme, dass $\frac{1}{n}\mathbf{X}_n^T\mathbf{X}_n$ für n gegen unendlich gegen eine positiv definite Matrix $\mathbf{V}$ konvergiert und weiteren Regularitätsannahmen (siehe z.B. Brannath, 2015), gilt
$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{V}^{-1}), \qquad (1)$$
wobei "$\xrightarrow{d}$" Konvergenz in Verteilung bedeutet.

- Hieraus folgt insbesondere die Konsistenz von $\hat{\beta}_n$, d.h. $\hat{\beta}_n \xrightarrow{p} \beta$, wobei "$\xrightarrow{p}$" Konvergenz in Wahrscheinlichkeit heißt.

- Aus (1) folgt, dass $\hat{\beta}_n$ bei hinreichend großem n approximativ $N(\beta, \sigma^2(\mathbf{X}_n^T\mathbf{X}_n)^{-1})$ verteilt ist. Wir schreiben dafür auch: $\hat{\beta}_n \overset{a}{\sim} N(\beta, \sigma^2(\mathbf{X}_n^T\mathbf{X}_n)^{-1})$.

Prognosen und Residuen

Modellraum

- Unter den Annahmen des linearen Modells ($\mu_i = E(Y_i) = \mathbf{x}_i\beta$) liegt der Erwartungswert $\mu = E(\mathbf{Y}) \in \mathbb{R}^n$ im folgenden Teilraum des $\mathbb{R}^n$:

$$\mathcal{M} = \{\, \tilde{\mu} = \mathbf{X}\beta \,|\, \beta \in \mathbb{R}^k \,\}$$

- $\mathcal{M}$ wird als *Modellraum* bezeichnet
- Wir betrachten auch das orthogonale Komplement von $\mathcal{M}$:

$$\mathcal{M}^\perp = \{\, \tilde{\mathbf{e}} \in \mathbb{R}^n \,|\, \tilde{\mathbf{e}}^T \tilde{\mu} = 0 \text{ für alle } \tilde{\mu} \in \mathcal{M} \,\}$$

Orthogonale Zerlegung

- Jedes $\mathbf{Y} \in \mathbb{R}^n$ lässt sich eindeutig als Summe eines Vektors $\hat{\mathbf{y}}$ in $\mathcal{M}$ und eines Vektors $\hat{\mathbf{e}}$ in $\mathcal{M}^\perp$ zerlegen, d.h.

$$\mathbf{y} = \hat{\mathbf{y}} \oplus \hat{\mathbf{e}} \qquad \text{mit eindeutigen } \hat{\mathbf{y}} \in \mathcal{M} \text{ und } \hat{\mathbf{e}} \in \mathcal{M}^\perp.$$

- Das Symbol "$\oplus$" deutet darauf hin, dass die beiden Vektoren $\hat{\mathbf{y}}$ und $\hat{\mathbf{e}}$ othogonal zueinander sind, d.h. das innere Produkt $\hat{\mathbf{y}}^T\hat{\mathbf{e}} = \hat{\mathbf{y}}_1\hat{\mathbf{e}}_1 + \cdots + \hat{\mathbf{y}}_n\hat{\mathbf{e}}_n$ ist gleich 0.

- Geometrisch gesehen, stehen $\hat{\mathbf{y}}$ und $\hat{\mathbf{e}}$ senkrecht aufeinander.

Projektion auf den Modellraum

- $\hat{\mathbf{Y}} = \text{argmin}_{\tilde{\mu} \in \mathcal{M}} ||\mathbf{Y} - \tilde{\mu}||^2$,
- D.h. $\hat{\mathbf{Y}}$ aus der orthogonalen Zerlegung ist jenes $\tilde{\mu} \in \mathcal{M}$ mit dem kleinsten Euklid-Abstand zu $\mathbf{Y}$.
- $\tilde{\mu} \in \mathcal{M}$ ist gleichbedeutend mit $\tilde{\mu} = \mathbf{X}\tilde{\beta}$ für ein $\tilde{\beta} \in \mathbb{R}^k$.
- Es gilt hier: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, wobei $\hat{\beta}$ der KQ-Schätzer von β ist.
- $\hat{\mathbf{Y}} = \mathbf{P}_{\mathcal{M}}\mathbf{Y}$ ist (per Definition) die orthogonale Projektion von $\mathbf{Y}$ auf $\mathcal{M}$. Sie ist auch der Vektor der Prognosen $\hat{\mathbf{Y}}_i = \mathbf{x}_i\hat{\beta}$.
- $\hat{\mathbf{Y}}$ wird *Prognose* genannt

Projektion auf den Modellraum

- Ist $\text{Rang}(\mathbf{X}) = k$ so kann man zeigen, dass

$$\mathbf{P}_{\mathcal{M}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbb{H}$$

- $\mathbb{H}$ wird häufig als „Hat-Matrix“ bezeichnet
- Es gilt:

$$E(\hat{\mathbf{Y}}) = E(\mathbf{P}_{\mathcal{M}}\mathbf{Y}) = \mathbf{P}_{\mathcal{M}}E(\mathbf{Y}) = \mathbf{P}_{\mathcal{M}}\mu = \mu$$

Residuen

- Der Vektor $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$ besteht aus den Residuen $\hat{e}_i = Y_i - \mathbf{x}_i\hat{\beta}$ der linearen Regression
- Es gilt
$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbb{H}\mathbf{Y} = (\mathbf{I}_n - \mathbb{H})\mathbf{Y}$$
- $\mathbf{I}_n - \mathbb{H} = \mathbf{I}_n - \mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}^\perp}$ ist die Projektionsmatrix auf $\mathcal{M}^\perp$

Eigenschaften der Residuen

- Es gilt
$$E(\hat{\mathbf{e}}) = E(\mathbf{Y}) - E(\hat{\mathbf{Y}}) = \mu - \mu = \mathbf{0}$$

- Zudem ist
$$Cov(\hat{\mathbf{e}}) = Cov(\mathbf{P}_{\mathcal{M}^\perp}\mathbf{Y}) = \mathbf{P}_{\mathcal{M}^\perp} Cov(\mathbf{Y})\mathbf{P}_{\mathcal{M}^\perp}^T = \sigma^2 \mathbf{P}_{\mathcal{M}^\perp}\mathbf{P}_{\mathcal{M}^\perp}^T = \sigma^2 \mathbf{P}_{\mathcal{M}^\perp} = \sigma^2(\mathbf{I}_n - \mathbb{H})$$

- Insbesondere gilt $Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ für das i-te Diagonalelement h_{ii} von $\mathbb{H}$.
- Während $Var(U_i) = \sigma^2$ unabhängig von i ist, ist $Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ im Allgemeinen abhängig von i. Während $Cov(U_i, U_j) = 0$ für $i \neq j$, ist $Cov(\hat{e}_i, \hat{e}_j) = \sigma^2(1 - h_{ij})$ im Allgemeinen $\neq 0$.

Verteilung von Prognosen und Residuen

- Wenn $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, dann ist $\mathbb{H}(\mathbf{I} - \mathbb{H}) = \mathbb{H} - \mathbb{H} = \mathbf{0}$.
- Es folgt

$$\begin{pmatrix} \hat{\mathbf{Y}} \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} \mathbb{H} \\ (\mathbf{I}_n - \mathbb{H}) \end{pmatrix} \mathbf{Y} \sim N\left(\begin{pmatrix} \mathbf{X}\beta \\ \mathbf{0} \end{pmatrix}, \Sigma \right)$$

mit

$$\Sigma = \sigma^2 \begin{pmatrix} \mathbb{H} \\ \mathbf{I}_n - \mathbb{H} \end{pmatrix} \mathbf{I}_n \left(\mathbb{H}, \mathbf{I}_n - \mathbb{H} \right) = \sigma^2 \begin{pmatrix} \mathbb{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n - \mathbb{H} \end{pmatrix}.$$

- $\hat{\mathbf{Y}}$ und $\hat{\mathbf{e}}$ sind damit stochastisch unabhängig mit

$$\hat{\mathbf{Y}} \sim N(\mathbf{X}\beta, \mathbb{H}\sigma^2) \quad \text{und} \quad \hat{\mathbf{e}} \sim N(\mathbf{0}, (\mathbf{I}_n - \mathbb{H})\sigma^2).$$

Weitere Eigenschaften von Prognosen und Residuen

(i) $\hat{\mathbf{Y}}$ und $\hat{\mathbf{e}}$ sind immer orthogonal, d.h. $\hat{\mathbf{Y}} \perp \hat{\mathbf{e}}$ (denn $\hat{\mathbf{Y}} \in \mathcal{M}$ und $\hat{\mathbf{e}} \in \mathcal{M}^{\perp}$),

(ii) $\hat{\mathbf{e}} \perp \mathbf{x}^j$ für alle Spalten $\mathbf{x}^j$ von $\mathbf{X}$, denn $\mathbf{x}^j = \mathbf{X}\mathbf{e}_j \in \mathcal{M}$ für den j-ten Einheitsvektor $\mathbf{e}_j$.

(iii) Falls die erste Spalte von $\mathbf{X}$ der Einsvektor $\mathbf{1} = (1, \ldots, 1)^T$ ist, dann ist $\sum_{i=1}^n \hat{e}_i = \mathbf{e}^T\mathbf{1} = 0$. Damit hat $\hat{\mathbf{e}}$ den Mittelwert $\mathbf{0}$.

(iv) Aus (iii) folgt $\bar{Y} = \frac{1}{n}\sum_i Y_i = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \bar{\hat{Y}}$, weil $Y_i = \hat{Y}_i + e_i$,

(v) Aus (iii) folgt auch $\bar{Y} = \beta_1\bar{x}_1 + \ldots + \beta_k\bar{x}_k = \beta_1 + \beta_2\bar{x}_2 + \ldots + \beta_k\bar{x}_k$ wobei $\bar{x}_j$ Mittelwert von $\mathbf{x}^j = (x_{j1}, \ldots, x_{jn})^T$ ist. D.h. die lineare Regressionsfunktion geht durch den Schwerpunkt $(\bar{Y}, \bar{x}_1, \ldots, \bar{x}_k)$.

Schätzung der Residualvarianz

Schätzung der Residualvarianz

- Wir schätzen die Residualvarianz σ^2 durch den Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\hat{\beta})^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n-k}\hat{\mathbf{e}}^T\hat{\mathbf{e}}$$

- Unter der Normalverteilungsannahme $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$ und wenn $\text{Rang}(\mathbf{X}) = k$ gilt

$$(n-k)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$$

- Da $\chi^2_\nu \sim \sum_{i=1}^{\nu} Z_i^2$ für $Z_i \overset{i.i.d.}{\sim} N(0,1)$ gilt $E(\chi^2_\nu) = \nu$
- Daraus folgt $E(\hat{\sigma}^2) = \sigma^2$

Standardisierte und studentisierte Residuen

Standardisierte Residuen

- Man verwendet die Residuen auch zur Überprüfung der Modellannahmen

  $$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \quad \text{mit} \quad Cov(\mathbf{U}) = \sigma^2 \mathbf{I}_n,$$

  d.h. der Annahme homoskedastischer und unkorrelierten U_i.

- **Problem**: $Cov(\hat{\mathbf{e}}) = (\mathbf{I}_n - \mathbb{H})\sigma^2$. D.h. im Allgemeinen sind die Residuen weder homoskedastisch noch unkorreliert.

- Korrelation typischerweise klein. Heteroskedasdizität ist größeres Problem.

- Verwende daher *standardisierte Residuen*

  $$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

  wobei h_{ii} das i-te Diagonalelement von $\mathbb{H}$.

Standardisierte Residuen

- Zur Überprüfung der Homoskedastizität von U_i trägt man die standardisierten Residuen r_i über die Prognosen Y_i auf. Man nennt diesen Plot den *Residuenplot*
- Obwohl $\hat{e}_i / \sqrt{\sigma^2(1-h_{ii})} \sim N(0,1)$ und $(n-k)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$ gilt i.A. nicht

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \sim t_{n-k}$$

- Grund: $\hat{\mathbf{e}}$ und $\hat{\sigma}^2$ sind offensichtlich nicht stochastisch unabhängig
- Lösung: Betrachte *studentisierte Residuen*

Studentisierte Residuen

- Die stochastische Unabhängigkeit kann durch Verwendung des „Leave-One-Out" Schätzers von $\hat{\sigma}^2$ erreicht werden
- Löschen die i-te Zeile von $\mathbf{Y}$ und $\mathbf{X}$, und erhalten so einen $(n-1)$-dimensionalen Vektor $\mathbf{Y}^{(i)}$ und eine $(n-1) \times k$ Matrix $\mathbf{X}^{(i)}$
- Bilden damit
$$\hat{\beta}^{(i)} = (\mathbf{X}^{(i)^T}\mathbf{X}^{(i)})^{-1}\mathbf{X}^{(i)^T}\mathbf{Y}^{(i)}$$
- D.h. wir führen die lineare Regression ohne die Beobachtung i durch.

Studentisierte Residuen

- Bilden die Prognose $\hat{Y}_{(i)} = \mathbf{x}_i \hat{\beta}^{(i)}$ für die i-te Beobachtung
- Damit berechnen wir schießlich das Residuum
$$\hat{e}_{(i)} = Y_i - \hat{Y}_{(i)} = Y_i - \mathbf{x}_i (\mathbf{X}^{(i)^T} \mathbf{X}^{(i)})^{-1} \mathbf{X}^{(i)} \mathbf{Y}^{(i)}.$$
- $\hat{e}_{(i)}$ wird als *Leave-One-Out-Residuumn* für die i-te Beobachtung bezeichnet
- Das Leave-One-Out-Residuum hat die Verteilung
$$\hat{e}_{(i)} \sim N(0, \sigma^2 (1 + \mathbf{x}_i (\mathbf{X}^{(i)^T} \mathbf{X}^{(i)})^{-1} \mathbf{x}_i^T))$$
- Also gilt:
$$\frac{\hat{e}_{(i)}}{\sqrt{\sigma^2 (1 + \mathbf{x}_i (\mathbf{X}^{(i)^T} \mathbf{X}^{(i)})^{-1} \mathbf{x}_i^T)}} \sim N(0, 1).$$

Studentisierte Residuen

- Bezeichne mit $\hat{\sigma}^2_{(i)}$ den Schätzer von σ^2 aus der linearen Regression ohne die i-te Beobachtung
- Man kann zeigen, dass $\hat{e}_{(i)}$ und $(n-k-1)\,\hat{\sigma}^2_{(i)}/\sigma^2 \sim \chi^2_{n-k-1}$ stochastisch unabhängig sind
- Damit folgt für das *studentisierte Residuum*:

$$\frac{\hat{e}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(1 + \mathbf{x}_i(\mathbf{X}^{(i)^T}\mathbf{X}^{(i)})^{-1}\mathbf{x}_i^T)}} \sim t_{n-k-1}.$$

- t-Verteilung kann z.B. verwendet werden, um Ausreißer zu diagnostizieren

Bestimmtheitsmaß

Streuungszerlegung

- Betrachte Modell mit Konstante (z.B. $\mathbf{x}^1 = \mathbb{1}$)
- Dann ist
$$\mathbf{Y} - \mathbb{1}\bar{Y} = (\hat{\mathbf{Y}} - \mathbb{1}\bar{Y}) \oplus \hat{\mathbf{e}}$$
- Aus dem Satz des Pythagoras folgt dann
$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = ||\mathbf{Y} - \mathbb{1}\bar{Y}||^2 = ||\hat{\mathbf{Y}} - \mathbb{1}\bar{Y}||^2 + ||\hat{\mathbf{e}}||^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{e}_i^2.$$
- Bilde das Bestimmtheitsmaß
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Eigenschaften des Bestimmtheitsmaßes

- R^2 ist der Anteil der Streuung von Y (um $\bar{Y}$) der durch das Modell erklärt werden kann
- Es gilt $0 \leq R^2 \leq 1$
- Liegt R^2 nahe bei 1 wird ein großer Teil der Streuung der Daten durch das Modell beschrieben. Dies spricht für eine Gute Anpassung
- Liegt R^2 nahe bei 0 spricht dies entsprechend für eine schlechte Anpassung der Daten