Tutorium

Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

14. Termin: (12) Multiple Regression

Sean Mulready

Selbstkontrollfragen

1. Erläutern Sie das Anwendungsszenario und die Ziele der multiplen Regression.
2. Definieren Sie das Modell der multiplen Regression.
3. Erläutern Sie die Begriffe Regressor, Prädiktor, Kovariate und Feature im Rahmen der multiplen Regression.
4. Erläutern Sie, warum $\hat{\beta} \approx \text{Regressorkovariabilität}^{-1} \text{Regressordatenkovariabilität}$ gilt.
5. Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und Korrelationen in einem multiplen Regressionsmodell mit Interzeptprädiktor und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = \frac{r_{\upsilon,x_1} - r_{\upsilon,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \frac{s_\upsilon}{s_{x_1}}. \tag{1}$$

6. Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und partieller Korrelation in einem multiplen Regressionmodell mit Interzeptprädiktor und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = r_{\upsilon,x_1 \setminus x_2} \sqrt{\frac{1 - r^2_{\upsilon,x_2}}{1 - r^2_{x_1,x_2}}} \frac{s_\upsilon}{s_{x_1}}. \tag{2}$$

7. $X \in \mathbb{R}^{n \times 2}$ sei die Designmatrix eines multiplen Regressionsmodells mit zwei Prädiktoren und Betaparametervektor $\beta := (\beta_1, \beta_2)^2$. Geben Sie den Kontrastgewichtsvektor an, um die Nullhypothese $H_0 : \beta_1 = \beta_2$ mithilfe der T-Statistik zu testen.

1.Erläutern Sie das Anwendungsszenario und die Ziele der multiplen Regression.

Anwendungsszenario

- Generalisierung der einfachen linearen Regression zu mehr als einer unabhängigen Variable.
- Eine univariate abhängige Variable bestimmt an randomisierten experimentellen Einheiten.
- Zwei oder mehr "kontinuierliche" unabhängige Variablen.
- Die unabhängigen Variablen heißen Regressoren, Prädiktoren, Kovariaten oder Features.

Ziele

- Quantifizierung des Erklärungspotentials der Variation der AV durch die Variation der UVs.
- Quantifizierung des Einflusses einzelner UVs auf die AV im Kontext anderer UVs.
- Prädiktion von AV Werten aus UV Werten nach Parameterschätzung.

Anwendungsbeispiel

- BDI Differenzwerte in Abhängigkeit von Therapiedauer und Alter

2.Definieren Sie das Modell der multiplen Regression.

Definition (Modell der multiplen Regression)

υ_i mit $i = 1, \ldots, n$ sei die Zufallsvariable, die den $i$ten Wert einer abhängigen Variable modelliert. Dann hat das *Modell der multiplen Regression* die strukturelle Form

$$\upsilon_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \ldots, n \text{ und } \sigma^2 > 0, \tag{3}$$

wobei $x_{ij} \in \mathbb{R}$ mit $1 \leq i \leq n$ und $1 \leq j \leq p$ den $i$ten Wert der $j$te unabhängigen Variable bezeichnet. Die unabhängigen Variablen werden auch *Regressoren, Prädiktoren, Kovariaten* oder *Features* genannt. Mit

$$x_i := (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p \text{ und } \beta := (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p \tag{4}$$

hat das Modell der multiplen Regression die Datenverteilungsform

$$\upsilon_i \sim N(\mu_i, \sigma^2) \text{ u.i.v. für } i = 1, \ldots, n, \text{ wobei } \mu_i := x_i^T\beta. \tag{5}$$

In diesem Zusammenhang wird $x_i \in \mathbb{R}^p$ auch als $i$ter *Featurevektor* bezeichnet. Die Designmatrixform des Modells der multiplen Regression schließlich ist gegeben durch

$$\upsilon = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{6}$$

mit

$$\upsilon := (\upsilon_1, \ldots, \upsilon_n)^T, X := (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \in \mathbb{R}^{n \times p}, \beta := (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p \text{ und } \sigma^2 > 0. \tag{7}$$

3.Erläutern Sie die Begriffe Regressor, Prädiktor, Kovariate und Feature im Rahmen der multiplen Regression.

Regressor, **Prädiktor**, **Kovariate** und **Feature** sind synonyme Bezeichnungen für die unabhängige Variable im Modell der multiplen Regression.

4.Erläutern Sie, warum $\hat{\beta} \approx$ Regressorkovariabilität^{-1}Regressordatenkovariabilität gilt.

Der Betaparameterschätzer hat bekanntlich die Form

$$\hat{\beta} := (X^T X)^{-1} X^T \upsilon$$

Dabei quantifizieren in sehr grober Auflösung

- $X^T \upsilon \in \mathbb{R}^p$ die Kovariabilität der Regressoren (Spalten der Designmatrix) mit den Daten υ und
- $X^T X \in \mathbb{R}^{p \times p}$ die Kovariabilität der Regressoren untereinander.

Damit ergibt sich für die Betaparameterschätzer also eine Interpretation als "regressorkovariabilitätsnormalisierte Regressordatenkovariabilität".

5.Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und Korrelationen in einem multiplen Regressionsmodell mit Interzeptprädiktor und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = \frac{r_{\upsilon,x_1} - r_{\upsilon,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \frac{s_\upsilon}{s_{x_1}}. \tag{8}$$

- Nur im Fall $r_{x_1,x_2} = 0$ und $s_\upsilon = s_{x_1}$ gilt $\hat{\beta}_1 = r_{\upsilon,x_1}$.
- Im Fall $r_{x_1,x_2} = \pm 1$ ist $\hat{\beta}_1$ nicht definiert
- Je größer $|r_{x_1,x_2}|$, desto größer der von r_{υ,x_1} subtrahierte Term $r_{\upsilon,x_2} r_{x_1,x_2}$
- Je größer $|r_{\upsilon,x_2}|$, desto größer der von r_{υ,x_1} subtrahierte Term $r_{\upsilon,x_2} r_{x_1,x_2}$
- Bei identischen Korrelationen und gleich bleibender Regressorstandardabweichung steigt $\hat{\beta}_1$ mit s_υ

6.Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und partieller Korrelation in einem multiplen Regressionmodell mit Interzept-prädiktor und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = r_{\upsilon,x_1\setminus x_2}\sqrt{\frac{1-r^2_{\upsilon,x_2}}{1-r^2_{x_1,x_2}}}\frac{s_\upsilon}{s_{x_1}}.$$

- Im Allgemeinen gilt für $1 \leq i,l \leq k$, dass $\hat{\beta}_k \neq r_{\upsilon,x_k\setminus x_l}$.
- Betaparameterschätzer sind also im Allgemeinen keine partiellen Stichprobenkorrelationen.
- $\hat{\beta}_k = r_{\upsilon,x_k\setminus x_l}$ für $1 \leq i,l \leq k$ gilt genau dann, wenn $s_\upsilon = s_{x_1} = s_{x_2}$ und zudem
  - $r_{\upsilon,x_l} = r_{x_k,x_l} = 0$, wenn also die Stichprobenkorrelationen der Daten und der Werte des zweiten Regressors, sowie die Stichprobenkorrelation der Werte der beiden Regressoren gleich Null sind. Dies kann der Fall sein, wenn einer der Regressoren die Daten "sehr gut erklärt" und der andere Regressor von dem ersten "sehr verschieden" ist.
  - $|r_{\upsilon,x_l}| = |r_{x_k,x_l}|$, wenn also die obige Stichprobenkorrelationen dem Betrage nach gleich sind. Dies ist vermutlich selten der Fall.

7.$X \in \mathbb{R}^{n \times 2}$ sei die Designmatrix eines multiplen Regressionsmodells mit zwei Prädiktoren und Betaparametervektor $\beta := (\beta_1, \beta_2)^2$. Geben Sie den Kontrastgewichtsvektor an, um die Nullhypothese $H_0 : \beta_1 = \beta_2$ mithilfe der T-Statistik zu testen.

$c = (1, -1)^T$