Statistik IV für Nebenfachstudierende

3 Multivariate lineare Regression

3.1. Grundkonzepte der linearen Regression

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Wiederholung Lineare Modelle

Ein lineares Regressionsmodell hat die Form

$$y_i = \underline{x}_i^\top \underline{\beta} + \epsilon_i, \quad i = 1, \ldots, n$$

bzw. in Matrixschreibweise

$$\underline{y} = \underline{\underline{X}}\, \underline{\beta} + \boldsymbol{\epsilon}$$

mit

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \underline{\underline{X}} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \underline{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

wobei:

- $\underline{\mathbf{y}}$: Zufallsvektor der Zielgröße
- $\underline{\underline{X}}$: feste Design-Matrix (Matrix der Kovariablen)
- $\underline{\boldsymbol{\epsilon}}$: Zufallsvektor der Fehlerterme
- $\underline{\beta}$: Vektor der Regressionsparameter der Länge $p + 1$

Allgemeine Annahmen:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\epsilon}) &= \mathbf{0} \\ \operatorname{cov}(\boldsymbol{\epsilon}) &= \sigma^2 \mathbf{I} \end{aligned}$$

Spezielle Verteilungsannahme:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- Aus dem KQ-Ansatz

$$(\underline{y} - \underline{\underline{X}}\,\underline{\beta})^\top (\underline{y} - \underline{\underline{X}}\,\underline{\beta}) \to \min_{\underline{\beta}}$$

ergibt sich durch Nullsetzen der ersten Ableitung nach $\underline{\beta}$ und, falls $\underline{\underline{X}}^\top \underline{\underline{X}}$ invertierbar ist, anschließendem Lösen des resultierenden Gleichungssystems der KQ-Schätzer

$$\hat{\underline{\beta}} = (\underline{\underline{X}}^\top \underline{\underline{X}})^{-1} \underline{\underline{X}}^\top \underline{y}$$

- Varianz der Schätzer $\hat{\beta}_j$, $j = 0, 1, \ldots, p$:

  Mit den Diagonalelementen v_j aus $(\underline{\underline{\mathbf{X}}}^\top \underline{\underline{\mathbf{X}}})^{-1}$ erhält man für bekanntes σ^2 als Varianz von $\hat{\beta}_j$

$$\sigma_j^2 = \mathsf{Var}(\hat{\beta}_j) = \sigma^2 v_j$$

  bzw. für unbekanntes σ^2

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 v_j.$$

- Zusammenfassende Darstellung in Vektornotation

$$\mathrm{cov}(\underline{\hat{\beta}}) = \sigma^2 (\underline{\underline{X}}^\top \underline{\underline{X}})^{-1} \quad \text{bzw.} \quad \widehat{\mathrm{cov}}(\underline{\hat{\beta}}) = \hat{\sigma}^2 (\underline{\underline{X}}^\top \underline{\underline{X}})^{-1}$$

- Der Vektor der geschätzten Residuen ist

$$\hat{\boldsymbol{\epsilon}} = \underline{y} - \underline{\underline{X}}\, \underline{\hat{\beta}}.$$

Der KQ-Schätzer besitzt folgende Eigenschaften:

1. Ist $\underline{\underline{X}}^\top \underline{\underline{X}}$ invertierbar, so gilt

   $\hat{\underline{\beta}}$ existiert und $\hat{\underline{\beta}}$ ist eindeutig.

2. Der KQ-Schätzer ist erwartungstreu, d.h. es gilt

   $$E(\hat{\underline{\beta}}) = \underline{\beta}.$$

3. Unter der speziellen Verteilungsannahme gilt

   $$\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2(\underline{\underline{X}}^\top \underline{\underline{X}})^{-1}).$$

- Unter der allgemeinen Annahme ist

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\underline{\epsilon}}^\top \hat{\underline{\epsilon}} = \frac{1}{n-p-1}\sum_{i=1}^{n} \hat{\epsilon}_i^2$$

  ein erwartungstreuer Schätzer für σ^2.

- Prognose:

$$\hat{y}_0 = \underline{x}_0^\top \hat{\underline{\beta}}$$

  $(1-\alpha)$-Prognoseintervall für $\hat{y}_0$ unter der speziellen Verteilungsannahme:

$$\left[\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}\,(n-p-1)\,\hat{\sigma}\sqrt{\underline{x}_0^\top \left(\underline{\underline{\mathbf{X}}}^\top \underline{\underline{\mathbf{X}}}\right)^{-1} \underline{x}_0 + 1}\,\right].$$

Gegeben sei ein lineares Modell mit der Designmatrix $\underline{\underline{X}}$, die vollen Rang hat. Dann gilt

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$$

bzw. in Matrixnotation

$$\underbrace{(\underline{y} - \underline{\bar{y}})^\top (\underline{y} - \underline{\bar{y}})}_{SST} = \underbrace{(\underline{y} - \underline{\hat{y}})^\top (\underline{y} - \underline{\hat{y}})}_{SSE} + \underbrace{(\underline{\hat{y}} - \underline{\bar{y}})^\top (\underline{\hat{y}} - \underline{\bar{y}})}_{SSM}$$

wobei $\hat{y}_i = \underline{x}_i^\top \underline{\hat{\beta}}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ und $\underline{\bar{y}} = (\bar{y}, \ldots, \bar{y})$.

Interpretation:

SST : Gesamt-Streuung, Gesamt-Quadratsumme

SSE : Fehler-Quadratsumme, Residuen-Quadratsumme

SSM : Modell-Quadratsumme