Regressionsanalyse – Übungen: Blatt 4

1. Betrachte das Modell

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

mit der $n \times q$ Matrix $\mathbf{X}_1$ und der $n \times (p-q)$ Matrix $\mathbf{X}_2$. Sei $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Angenommen wir schätzen $\boldsymbol{\beta}_1$ und σ^2 nur unter Verwendung des Sub-Modells

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^* \,.$$

Also erhalten wir

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

und

$$\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{y}/(n-q) \,.$$

(a) Zeige, dass für eine quadratische Form mit fester $n \times n$ Matrix $\mathbf{A}$ gilt

$$\mathrm{E}(\mathbf{y}'\mathbf{A}\mathbf{y}) = \mathrm{trace}(\mathbf{A}\mathrm{var}(\mathbf{y})) + \mathrm{E}(\mathbf{y}')\mathbf{A}\mathrm{E}(\mathbf{y}) \,,$$

womit folgt, dass

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2 + \boldsymbol{\beta}_2'\mathbf{X}_2'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{X}_2\boldsymbol{\beta}_2/(n-q) \geq \sigma^2 \,.$$

Unter welchen Bedingungen ist $\hat{\sigma}^2$ unverzerrt für σ^2?

(b) Habe $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2)$ vollen Spaltenrang. Zeige, dass

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} + \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}' & -\mathbf{A}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{A}' & \mathbf{Q}^{-1} \end{bmatrix}$$

gilt mit $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ und $\mathbf{Q} = \mathbf{X}_2'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_1})\mathbf{X}_2$.
Verwende dazu folgende Eigenschaften:

$$\big(\mathbf{A} + \mathbf{BCD}\big)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\big(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\big)^{-1}\mathbf{DA}^{-1}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$$

$\mathbf{A}^{11} = \big(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\big)^{-1}$, $\mathbf{A}^{22} = \big(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\big)^{-1}$, $\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}$, $\mathbf{A}^{21} = -\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}$.

(c) Verwende obiges Resultat und zeige, dass die Kleinsten-Quadrate Schätzer der Elemente in $\boldsymbol{\beta}_1$ basierend auf dem Sub-Modell kleinere Varianzen als die entsprechenden Schätzer unter dem vollen Modell haben.

2. Zeige, dass

$$\mathrm{SSE}(\hat{\beta}_{(i)}) = \mathrm{SSE}(\hat{\beta}) - r_i^{*2}S^2$$

mit $S^2 = \mathrm{SSE}(\hat{\beta})/(n-p)$ und

$$r_i^* = \frac{y_i - \hat{\mu}_i}{S\sqrt{1-h_{ii}}}$$

gilt. Hierbei bezeichnet $\hat{\beta}_{(i)}$ den Kleinsten-Quadrate Schätzer ohne Verwendung der i-ten Beobachtung.

3. Ein Versuch mit den beiden zweistufigen (`hoch`, `niedrig`) Faktoren `A` und `B` wird derart durchgeführt, dass jede Kombination der Faktorstufen genau dreimal beobachtet wird. Wir betrachten ein Regressionsmodell mit diesen beiden Haupteffekten, also `y` $\sim$ `A + B`.

   (a) Kodiere die Stufen `hoch` und `niedrig` jeweils durch $+1$ und -1. Führe die dadurch resultierende Designmatrix an und berechne explizit die Kleinsten-Quadrate Schätzer sowie deren Varianz/Kovarianzmatrix.

   (b) Erweitere das Modell durch die Wechselwirkung `A:B`. Was ergibt sich jetzt explizit als Kleinster-Quadrate Schätzer und wie sieht die Varianz/Kovarianzmatrix aus?

4. Auf der Homepage zur Lehrveranstaltung findet man den Datensatz `houses.dat`. Finde für die Responsevariable `price` ein optimales lineares Regressionsmodell und kommentiere die durchgeführte Recherche. Überprüfe, ob die Größe `bed` bzw. `bath` dabei nicht eher als Faktor in das Modell eingehen sollten.

   Prüfe mittels geeigneter diagnostischer Methoden das gefundene Modell. Begründe und interpretiere die dabei erzielten Erkenntnisse.

   Ich bin daran interessiert, mein (altes) Haus in Gainesville zu verkaufen, welches auf einer Grundstücksfläche von 2300 square feet steht, sowie 3 Schlafzimmer und 2 Bäder hat. Welchen Preis werde ich wohl dafür mit hoher Wahrscheinlichkeit ($\alpha = 0.10$) bekommen?

   Ein Makler möchte 2 neue und 1 altes Haus kaufen. Alle stehen auf 1800 square feet Grundstücken und haben 3 Schlafzimmer. Ein neues Haus hat 3 Bäder die beiden anderen jeweils 2. Berechne Intervalle, die mit 95%-iger Wahrscheinlichkeit die zu erwartenden Kosten simultan überdecken.