Regressionsanalyse – Übungen: Blatt 3

1. Zeige explizit ohne Verwendung von Matrizen, dass für ein multiples lineares Modell, MLR, die folgende Identität hält:

$$\sum_{i=1}^{n} r_i \hat{\mu}_i = 0$$

mit Residuen $r_i = y_i - \hat{\mu}_i$.

2. Ein lineares Regressionsmodell $y = X\beta + \epsilon$ mit r unabhängigen (erklärenden) Variablen wurde angepasst. Angenommen, das wahre Modell beinhaltet weitere s unabhängige Variablen, die in Z enthalten sind, also

$$y = X\beta + Z\gamma + \epsilon\,, \qquad \gamma \neq 0\,.$$

Finde $\mathrm{E}(\hat{\beta})$ und zeige, dass im Allgemeinen $\hat{\beta}$ verzerrter Schätzer für β ist. Unter welchen Bedingungen ist $\hat{\beta}$ unverzerrt?

3. Eine Responsevariable y ($n = 20$) hängt von 3 erklärenden Variablen und dem Intercept ab, die in X enthalten sind. Folgende Informationen wurden berechnet:

$$X'X = \begin{pmatrix} 20 & 0 & 0 & 0 \\ 0 & 250 & 401 & 0 \\ 0 & 401 & 1013 & 0 \\ 0 & 0 & 0 & 128 \end{pmatrix} \qquad X'y = \begin{pmatrix} 1900.00 \\ 970.45 \\ 1674.41 \\ -396.80 \end{pmatrix} \qquad y'y = 185883$$

   (a) Berechne $\hat{\beta}$ und schreibe die Regressionsgleichung auf.

   (b) Schätze σ^2 und berechne die Standardfehler der Regressionskoeffizienten (Hinweis: Betrachte dazu die Identität $y = \hat{\mu} + (y - \hat{\mu}) = \hat{\mu} + r$).

   (c) Schätze die Kovarianz zwischen $\hat{\beta}_1$ und $\hat{\beta}_2$ sowie jene zwischen $\hat{\beta}_1$ und $\hat{\beta}_3$.

4. Beobachtungen (x_i, Y_i), $i = 1, \ldots, n$, wurden unter dem Modell

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

gemacht, wobei $x_1, \ldots, x_n$ feste Konstanten und $\epsilon_1, \ldots, \epsilon_n$ iid aus $N(0, \sigma^2)$ sind. Dieses Modell wird jetzt reparametrisiert zu

$$Y_i = \alpha' + \beta'(x_i - \bar{x}) + \epsilon_i\,.$$

Seien $\hat{\alpha}$ und $\hat{\beta}$ die MLE's von α und β und $\hat{\alpha}'$ und $\hat{\beta}'$ die MLE's von α' und β'.

   (a) Zeige, dass $\hat{\beta}' = \hat{\beta}$.

   (b) Zeige, dass $\hat{\alpha}' \neq \hat{\alpha}$ und weiters $\hat{\alpha}' = \bar{Y}$ gilt. Finde die Verteilung von $\hat{\alpha}'$.

   (c) Zeige, dass $\hat{\alpha}'$ und $\hat{\beta}'$ unkorreliert und daher unter Normalverteilung unabhängig sind.

5. Ein Ökologe verwendet Daten (x_i, Y_i), $i = 1, \ldots, n$, wobei x_i die Größe eines Gebietes und Y_i die Anzahl von Moosgewächsen in diesem Gebiet bezeichnen. Wir modellieren diese Daten durch unabhängige Responsevariablen $Y_i \sim \mathrm{Poisson}(\mu_i)$ mit $\mu_i = \theta x_i$.

   (a) Zeige, dass der Kleinste Quadrate Schätzer von θ gleich $\sum x_i Y_i / \sum x_i^2$ ist und dass dieser Schätzer Varianz $\theta \sum x_i^3 / (\sum x_i^2)^2$ hat. Berechne auch seinen Bias.

(b) Zeige, dass als MLE von θ der Schätzer $\sum Y_i / \sum x_i$ resultiert und dass dieser Varianz $\theta / \sum x_i$ hat. Berechne auch den Bias des MLE's.

(c) Finde den besten unverzerrten Schätzer für θ und zeige, dass seine Varianz die Cramér-Rao Schranke erreicht.

6. Verwende den Datensatz `aimu` und betrachte ein Regressionsmodell für `VC` in Abhängigkeit von `Groesse` und `Gewicht`. Halte die Werte von `Groesse` und `Gewicht` in dem Datensatz fest und generiere $R = 1000$ mal dazu den simulierten Vektor der Responsevariablen unter dem Modell

$$y_i = -770 + 7.2\texttt{Groesse}_i + 0.67\texttt{Gewicht}_i + \epsilon_i$$

mit $\epsilon_i \overset{iid}{\sim} N(0, 56^2)$. Passe ein entsprechendes MLR an jeden simulierten Datensatz an.

(a) Teste für jedes Modell die Hypothese, dass `Groesse` zusätzlich zu `Gewicht` im Modell irrelevant ist. Betrachte dazu die Monte Carlo Verteilung der Teststatistik und des p-Wertes. Interpretiere beide Histogramme.

(b) Teste für jedes Modell, dass beide Prädiktoren im Modell irrelevant sind. Betrachte auch dazu die Monte Carlo Verteilung der entsprechenden Teststatistik und ihres p-Wertes und interpretiere diese Histogramme.

(c) Generiere nun $R = 1000$ neue Responses unter dem einfacheren Modell

$$y_i = -810 + 7.7\text{Groesse}_i + \epsilon_i$$

mit $\epsilon_i \overset{iid}{\sim} N(0, 56^2)$. Schätze aber wiederum die beiden Parameter im multiplen Modell und teste für jedes Modell die Hypothese, dass `Gewicht` zusätzlich zu `Groesse` im Modell irrelevant ist. Betrachte dazu die Monte Carlo Verteilung der Teststatistik und des p-Wertes. Interpretiere beide Histogramme.