Aufgabe 1:

Als Ergebnis einer Regressionsanalyse (mit Absolutglied β_0) seien folgende Matrizen gegeben:

$$X'X = \begin{pmatrix} 9 & 136 & 269 & 260 \\ 136 & 2114 & 4176 & 3583 \\ 269 & 4176 & 8257 & 7104 \\ 260 & 3583 & 7104 & 12276 \end{pmatrix}, \quad X'Y = \begin{pmatrix} 45 \\ 648 \\ 1283 \\ 1821 \end{pmatrix},$$

$$(X'X)^{-1} = \begin{pmatrix} 9.610932 & 0.0085878 & -0.2791475 & -0.0445217 \\ 0.0085878 & 0.5099641 & -0.2588636 & 0.0007765 \\ -0.2791475 & -0.2588636 & 0.1395 & 0.0007369 \\ -0.0445217 & 0.0007765 & 0.0007396 & 0.0003698 \end{pmatrix},$$

$$(X'X)^{-1}X'Y = \begin{pmatrix} -1.163461 \\ 0.135270 \\ 0.019950 \\ 0.121954 \end{pmatrix}, \quad Y'Y = 285$$

(a) Berechnen Sie die Teststatistik des Overall-F-Tests und stellen Sie die Zwischenschritte in einer Tabelle (sogenannte Varianzanalysetafel) dar. Interpretieren Sie das Ergebnis.

(b) Berechnen Sie $\hat{\beta}$ und die Diagonalelemente von $\hat{V}(\hat{\beta})$ und testen Sie jeweils die Hypothese, dass die Regressionskoeffizienten gleich Null sind.

(c) Wie lautet die Matrix A zum Testen der Hypothese H_0: $\beta_0 = 0, \beta_1 = \beta_3, \beta_2 = 0$?
Wie viele Freiheitsgrade hat der damit verbundene Test?

(d) Wie lautet die geschätzte Modellgleichung für das reduzierte Modell aus (c)?

Aufgabe 2:

Unter `http://www.statistik.lmu.de/~kneib/regressionsbuch/download/sambia92.raw`
finden Sie einen Datensatz zum Ernährungszustand von Kindern in Sambia[1]. Zielgröße ist der sogenannte Z-Score, der eine Maßzahl für chronische Unterernährung darstellt. Betrachten Sie für alle folgenden Fragestellungen nur diejenige Region Sambias, für die die meisten Beobachtungen vorliegen.

(a) Schätzen Sie ein multiples Regressionsmodell mit den Einflussgrößen „Geschlecht des Kindes“ (`k_geschl`), „Stilldauer in Monaten“ (`k_still`), „Alter des Kindes in Monaten“ (`k_alter`), „Alter der Mutter bei der Geburt in Jahren“ (`m_alterg`), „Größe der Mutter in cm“ (`m_groesse`) und „Body-Mass-Index der Mutter“ (`m_bmi`). Für dieses Modell sollen im Folgenden verschiedene Hypothesen getestet werden.

(b) Testen Sie die Signifikanz der Variable „Stilldauer“.

(c) Testen Sie die Signifikanz der Variablen „Geschlecht“ und „Alter der Mutter bei der Geburt“ zusammen.

[1] Vgl. hierzu Fahrmeir/Kneib/Lang: „Regression – Modelle, Methoden und Anwendungen“, Springer-Verlag 2007, S. 5

(d) Testen Sie, ob die Größe der Mutter denselben Einfluss hat wie ihr BMI.

(e) Testen Sie, ob sich der Z-Score mit einem Anstieg des Alters um ein Monat um 2 verringert.

Aufgabe 3:

Es soll das Tippverhalten kanadischer Lottospieler untersucht werden. Dazu enthält Datensatz 3[2] auf der Übungshomepage folgende Information:

- Absolute Ziehungshäufigkeit der Zahlen 1 bis 49 als Gewinnzahlen bis zum Juni 1993 (2. Spalte).
- Relative Tipphäufigkeit dieser Zahlen bei den Ziehungen im Juli 1993 (in Prozent, 3. Spalte).

(a) Gehen Sie der Frage nach, ob kanadische Lottospieler im Juli 1993 diejenigen Zahlen häufiger ankreuzten, die in der Vergangenheit vermehrt Gewinnzahlen waren. Untersuchen Sie in einem zweiten Modell, ob ein (linearer) „Trend“ dazu besteht, niedrigere Zahlen häufiger anzukreuzen.

(b) Verwenden Sie nun die multiple Regression, um beide Fragestellungen in einem Modell zu klären. Wie verändern sich die Ergebnisse? Betrachten Sie den Zusammenhang auch graphisch.

(c) Sie haben die Hypothese, dass sich die kanadischen Lottospieler nicht von früheren Ziehungen beeinflussen lassen. Welche zwei Möglichkeiten gibt es, diese Hypothese im multiplen Modell zu testen und wie ist der Zusammenhang? Bestimmen Sie A, a und c. Interpretieren Sie das Ergebnis.

(d) Diskutieren Sie verschiedene Modelle und zeigen Sie Alternativen auf.

Aufgabe 4:

Betrachten Sie folgende Zerlegung des multiplen linearen Regressionsmodells

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{1}$$

sowie das von X_2 bereinigte Modell

$$\mathbf{Y}^\star = \mathbf{X}_1^\star\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}^\star \tag{2}$$

mit $\mathbf{Y}^\star = \mathbf{Q}_2\mathbf{Y}$, $\mathbf{X}_1^\star = \mathbf{Q}_2\mathbf{X}_1$ und $\mathbf{Q}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. Man kann zeigen, dass die KQ-Schätzer $\hat{\boldsymbol{\beta}}_1$ für Modelle (1) und (2) gleich sind.

(a) Interpretieren Sie dies und erklären Sie, weshalb in einer linearen Regression von $\mathbf{Y}$ auf $\mathbf{X}_1$ und $\mathbf{X}_2$ der KQ-Schätzer $\hat{\boldsymbol{\beta}}_1$ ermittelt werden kann, indem die Residuen einer Regression von $\mathbf{Y}$ (nur) auf $\mathbf{X}_2$ auf die Residuen regressiert werden, die man erhält, wenn man jede Spalte von $\mathbf{X}_1$ auf $\mathbf{X}_2$ regressiert.

(b) Überprüfen Sie dieses Verfahren anhand von Datensatz 3 aus Aufgabe 3 mit

- Y - die relative Tipphäufigkeit im Juli 1993
- X_1 - die Zahl
- X_2 - die Ziehungshäufigkeit bis Juni 1993.

Stellen Sie Modell (2) graphisch mit Hilfe eines Streudiagramms dar (Eine solche Graphik nennt man auch *Partial Leverage Plot*).

[2] Aus Riedwyl, Hans: „Lineare Regression und Verwandtes“, Birkhäuser Verlag: 1997