Universität Duisburg-Essen
Fakultät für Wirtschaftswissenschaften
Dr. Yannick Hoga

WS 2018/19

Hauptklausur
Einführung in die Ökonometrie

Bitte bearbeiten Sie **3 der 4** *Aufgaben. Falls Sie mehr als 3 Aufgaben bearbeiten, geben Sie bitte deutlich an, welche 3 Aufgaben bewertet werden sollen. In jeder Aufgabe sind 20 Punkte zu erreichen.*

Die Bearbeitungszeit beträgt 60 Minuten (zzgl. 5 Minuten Einlesezeit). Sie können maximal 60 Punkte erreichen.

Zugelassene Hilfsmittel: RStudio und die angehängte Formelsammlung

Wenn Sie keinen vollständigen Lösungsweg angeben können, skizzieren Sie bitte zumindest die Idee einer Lösung.

Viel Erfolg!

1. Betrachten Sie das einfache Regressionsmodell

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i, \ \ i = 1, \dots, n.. \tag{1}$$

Auf dem Klausurbogen zu lösen:

(a) Nennen und erläutern Sie *kurz* die drei Kleinste-Quadrate-Annahmen für das einfache Regressionsmodell (1).
[6 *Punkte*]

(b) Angenommen es gibt eine weitere Variable X_2, die Y beeinflusst und mit X_1 korreliert.

(1) Welche KQ-Annahme ist dann in Modell (1) verletzt?
[4 *Punkte*]

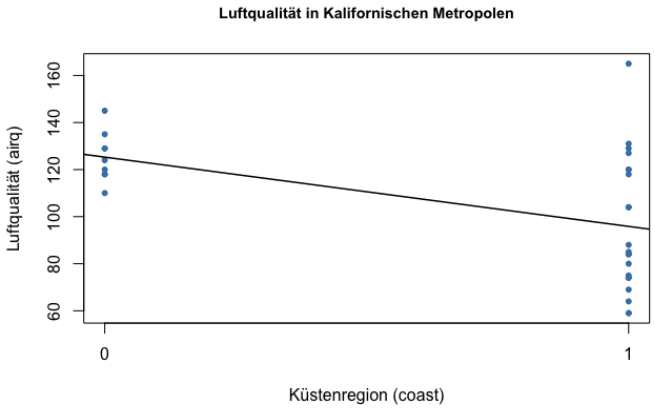
(2) Erläutern Sie anhand einer Formel, warum eine KQ-Schätzung von β_1 anhand von Modell (1) problematisch ist.
[4 *Punkte*]

Betrachten Sie nun das multiple Regressionsmodell

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \ \ i = 1, \dots, n. \tag{2}$$

(c) Wie müssen die drei Annahmen aus (a) hier angepasst werden? Welche weitere Annahme ist nötig, damit die Schätzung durchgeführt werden kann? Erklären Sie diese Annahme *kurz*!
[4 *Punkte*]

(d) Wie lautet hier das Minimierungsproblem zur Berechnung der KQ-Schätzer von β_0, β_1 und β_2? Erläutern Sie *kurz*!
[2 *Punkte*]



Variable	Beschreibung
`airq`	Indikator für Luftqualität
`vala`	Wertschöpfung ansässiger Unternehmen (in tsd. USD)
`rain`	Niederschlagsmenge (in cm)
`coast`	Dummy für Küstenregion
`dens`	Populationsdichte (Einw./Quadratmeile)
`medi`	durchschn. Pro-Kopf-Einkommen (in USD)

Abbildung 1: `Airq` — Luftqualität in Kalifornien

2. Es soll untersucht werden, ob die Luftqualität an der Pazifikküste gelegener Metropolregionen in Kalifornien besser ist als im Inland. Hierfür wurde zunächst die Luftqualität *airq* (kleine Werte sind besser) auf die Dummyvariable *coast* (1=Region an der Küste, 0=Region im Inland) regressiert. Die in Abbildung 1 dargestellte Schätzung eines einfachen Regressionsmodells lautet

$$\widehat{airq} = \underset{(8.29)}{125.33} - \underset{(9.91)}{29.48} \times coast, \qquad R^2 = 0.24. \tag{3}$$

Die bei Homoskedastie gültigen Standardfehler sind in (3) in Klammern angegeben.

Auf dem Klausurbogen zu lösen:

(a) Interpretieren Sie die geschätzten Koeffizienten sowie R^2.
[3 *Punkte*]

(b) Ist es sinnvoll, dass Regressionsergebnis aus (3) als Regressionsgerade wie in Abbildung 1 darzustellen? Begründen Sie!
[2 *Punkte*]

(c) Prüfen Sie, ob der Koeffizient von *coast* zum Signifikanzniveau 5% von 0 verschieden ist. Warum könnte ein t-Test im Hinblick auf die Informationen in (3) und Abbildung 1 ungültig sein? Könnten Sie ggf. stattdessen ein 95%-Konfidenzintervall für den Koeffizienten von *coast* zum Testen der Hypothese verwenden?
Hinweis: Für den Test benötigen Sie das 97.5%-Quantil der N(0,1)-Verteilung.
[5 *Punkte*]

In R zu lösen:

(d) Verschaffen Sie sich einen Überblick über den Datensatz `Airq` und reproduzieren Sie die Regression aus (3) sowie den Plot in Abbildung 1.
[3 *Punkte*]

(e) Schätzen Sie das multiple Regressionsmodell

$$airq = \beta_0 + \beta_1 coast + \beta_2 vala + \beta_3 rain + \beta_4 dens + \beta_5 medi + u \tag{4}$$

und erstellen Sie eine bei Heteroskedastie gültige statistische Zusammenfassung der geschätzten Koeffizienten.
[3 *Punkte*]

(f) Berechnen Sie eine bei Heteroskedastie gültige Schätzung der Varianz-Kovarianz-Matrix der KQ-Schätzer für die Koeffizienten im Modell aus (b). Erstellen Sie ein robustes 95%-Konfidenzintervall für den Koeffizienten von *coast*.
Hinweis: Nutzen Sie die Funktion `vcovHC()`.
[4 *Punkte*]

3. Betrachten Sie das Regressionsmodell

$$Y_i = \beta_1 X_i + u_i, \qquad i = 1, \ldots, n. \tag{5}$$

Auf dem Klausurbogen zu lösen:

(a) Zeigen Sie, dass $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$ der KQ-Schätzer von β_1 ist.
[4 *Punkte*]

(b) Es sei $\mathrm{E}(u|X) = 0$. Zeigen Sie, dass $\mathrm{E}(\widehat{\beta}_1) = \beta_1$.
[6 *Punkte*]

(c) Was bedeutet die Aussage $\widehat{\beta}_1 \xrightarrow{p} \beta_1$? Welche Bedingungen müssen in Modell (5) erfüllt sein, damit diese Aussage gültig ist?
[4 *Punkte*]

Nehmen Sie für die nachfolgenden Teilaufgaben an, dass X_i eine Dummyvariable ist, d.h. X_i kann nur die Werte 0 oder 1 annehmen.

(d) Welche Interpretation hat β_1 im Modell (5)?
Hinweis: Betrachten Sie $\mathrm{E}(Y_i|X_i = 1) = \ldots$
[3 *Punkte*]

(e) Zeigen Sie, dass $\widehat{\beta}_1 = \frac{1}{n_1} \sum_{n_1} Y_i$, wobei n_1 die Anzahl der Beobachtungen mit $X_i = 1$ ist und $\sum_{n_1}$ die Summe über diese Beobachtungen meint.
[3 *Punkte*]

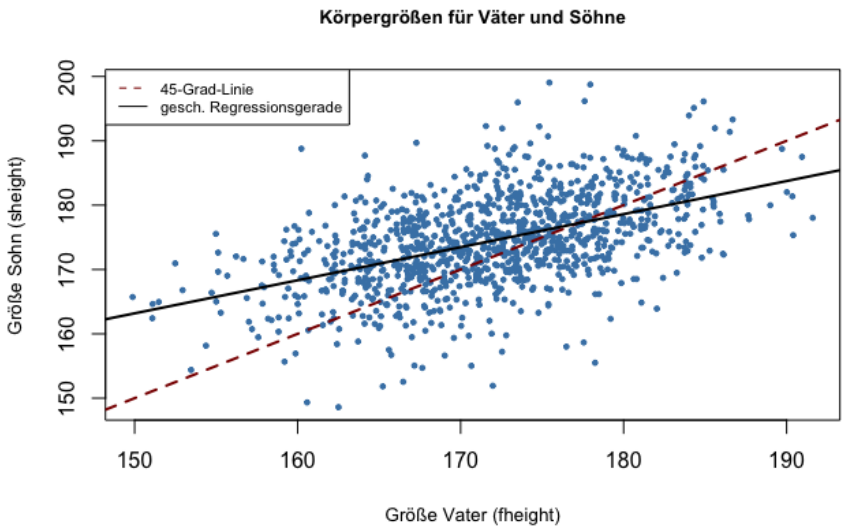


Abbildung 2: Daten aus Pearson and Lee (1903)

4. Diese Aufgabe befasst sich mit dem Datensatz `father.son`, welcher 1078 Beobachtungen der Körpergröße von Vätern und ihren (ausgewachsenen) Söhnen enthält. Die Daten stammen aus einer berühmten Studie von Karl Pearson.

Variable	Beschreibung
`fheight`	Körpergröße des Vaters (in cm)
`sheight`	Körpergröße des Sohnes (in cm)

In R zu lösen:

(a) Erstellen Sie die R-Funktionen `Schiefe()` und `Woelbung()`, welche die empirische Schiefe bzw. die empirische Wölbung für einen numerischen Vektor berechnen. Es sei

$$\text{Emp. Schiefe}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)^3, \quad \text{Emp. Wölbung}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)^4,$$

wobei $\overline{X}$ das arithmetische Mittel der X_i ist und s_X die Stichprobenstandardabweichung für die X_i meint.
[5 *Punkte*]

(b) Nutzen Sie die Funktionen aus (a), um die empirische Schiefe und die empirische Wölbung von `fheight` und `sheight` zu berechnen. Eine Normalverteilung hat eine Schiefe von 0 und eine Wölbung von 3. Scheinen beide Variablen einer Normalverteilung zu folgen?
[3 *Punkte*]

(c) Stellen Sie die Häufigkeitsverteilungen von `fheight` und `sheight` mit Histogrammen dar. Was fällt auf?
[4 *Punkte*]

(d) Regressieren Sie die Körpergröße der Väter auf die Körpergröße der Söhne und eine Konstante. Stellen Sie die Beobachtungen beider Variablen und die geschätzte Regressionsgerade grafisch dar.
[3 *Punkte*]

(e) Nutzen Sie eine geeignete R-Funktion, um mit dem geschätzten Modell aus (c) die erwartete Größe von Söhnen für Väter mit den Größen 150, 170 und 190cm zu bestimmen. Interpretieren Sie die Ergebnisse im Hinblick auf den Plot aus Abbildung 2. Erläutern Sie in diesem Zusammenhang *kurz*, was man unter *regression to the mean* (Regression zur Mitte) versteht.
[5 *Punkte*]

Formelanhang

Erwartungswert $E(Y) = \mu_Y$

Varianz $E(Y - \mu_Y)^2 = \sigma_Y^2$

Stichprobenvarianz $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

Standardabweichung / Stichprobenstandardabweichung $\sigma_Y = \sqrt{Varianz}$ bzw. $s_Y = \sqrt{Stichprobenvarianz}$

Kovarianz $Cov(X, Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$

Stichprobenkovarianz $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$

Korrelationskoeffizient $r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$

KQ-Schätzer $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

$\widehat{\beta}_0 = \overline{Y} - \overline{X}\widehat{\beta}_1$

Standardfehler der Regression $SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i^2}$

Root Mean Squared Error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}$

Bestimmtheitsmaß $R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$

$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$

Standardfehler von $\widehat{\beta}_1$ $SE(\widehat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}}$, wobei $\widehat{v}_i = (X_i - \overline{X})\widehat{u}_i$

$Var(\widehat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{Var[(X_i - \mu_X)u_i]}{\mathrm{Var}^2(X_i)}$

Bei Homoskedastie gültiger Standardfehler von $\widehat{\beta}_1$ $SE(\widehat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}$.

F-Statistik für β_1 und β_2 $F = \frac{1}{2}\left[\frac{t_1^2 + t_2^2 - 2\widehat{\rho}_{t_1,t_2}t_1t_2}{1 - \widehat{\rho}_{t_1,t_2}^2}\right]$

F-Statistik bei Homoskedastie $F = \frac{(R^2_{unrestringiert} - R^2_{restringiert})/q}{(1 - R^2_{unrestringiert})/(n - k_{unrestringiert} - 1)}$

Formel für Omitted Variable Bias: $\widehat{\beta}_1 - \beta_1 \xrightarrow{p} \left(\frac{\sigma_u}{\sigma_x}\right)\rho_{xu}$

95%-Konfidenzintervall für β_k: $\left[\widehat{\beta}_k - 1.96 \cdot SE(\widehat{\beta}_k), \widehat{\beta}_k + 1.96 \cdot SE(\widehat{\beta}_k)\right]$