Übung zu Kapitel 3: Multiple lineare Regression

Aufgabe 1

Sei $\mathbf{z} = (v, w)'$, worin v und w Zufallsvariablen sind. Zeige, dass

$$\Sigma(z) = \left[\begin{array}{cc} Var(v) & Cov(v, w) \\ Cov(w, v) & Var(w) \end{array} \right]$$

ist.

Zeige, unter welchen Bedingungen $\Sigma(z)$ positiv definit ist.

Aufgabe 2

a) Sei $\mathbf{x}$ der Spaltenvektor $(x_1, \ldots, x_n)'$. Welche Dimension hat $\mathbf{x}'\mathbf{x}$?

b) Gebe einen Summenausdruck für $\mathbf{x}'\mathbf{x}$ an.

c) Welche Dimension hat $\mathbf{x}\mathbf{x}'$?

d) Ermittle beide vorigen Ausdrücke für $\mathbf{x}' = (1, 2, 3)$.

e) Betachte die $m \times n$ Matrix A mit typischem Element a_{ij}. Unter welcher Bedingung bezeichnet man A als symmetrische Matrix?

f) Zeige, dass für eine $n \times k + 1$ Matrix X gilt, dass $X'X$ existiert und symmetrisch ist.

g) Definiere folgende Begriffe:
   1. lineare Unabhängigkeit von Vektoren
   2. Rang einer $n \times k+1$ Matrix mit $k < n$
   3. Nichtsingularität einer quadratischen $k+1 \times k+1$ Matrix
   4. Einheitsmatrix
   5. Inverse einer quadratischen Matrix

h) Unter welcher Bedingung existiert die Inverse eine quadratischen Matrix?

i) Ermittle $(A'B)^{-1}$ für

$$A = \begin{pmatrix} 1 & 7 \\ 0 & 1 \end{pmatrix} \text{ und } B = \begin{pmatrix} -1 & 1 \\ 1 & -0.5 \end{pmatrix}^{-1}$$

Aufgabe 3

Benutze für diese Aufgabe den Datensatz bwght.xls.

Betrachte das ökonometrische Modell

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

- Welches Vorzeichen vermutest du für β_2?
- Glaubst du, dass *cigs* und *faminc* korreliert sind? Erkläre, warum die Korrelation positiv oder negativ sein könnte.
- Schätze die Gleichung mit und ohne *faminc* und vergleiche die Resultate.

Aufgabe 4

Beschreibe jeweils, ob und ggf. wie du die folgenden Modelle als multiples lineares Regressionsmodell schätzen kannst, und gebe jeweils an, wie du die Elastizität von y_i bezüglich x_{i1} ermitteln würdest.

a) $y_i = \beta_0 x_{i1}^{\beta_1} x_{i2}^{\beta_2} \exp(u_i)$

b) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + u_i$

c) $y_i = (\beta_0 + \beta_1 x_{i1})^{\beta_2} + u_i$

d) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_2 x_{i1} x_{i2} + u_i$

e) $\ln(y_i) = \beta_0 + \ln(x_{i1}) + \beta_1 x_{i2} + u_i$

Aufgabe 5

Du willst den Einfluss der Unternehmensgröße g_i auf die Managergehälter m_i untersuchen. Außerdem vermutest du, dass die Branche des Unternehmens eine Rolle für die Vergütung des Managements spielt. In deinen Daten sind alle Unternehmen nach zwei Branchen, produzierendes Gewerbe p_i und Dienstleistungssektor d_i, klassifiziert. Es ist also die Variable $p_i = 1$; wenn das i-te Unternehmen im produzierenden Gewerbe tätig ist, und $p_i = 0$, wenn es im Dienstleistungssektor tätig ist. Entsprechend ist $d_i = 1$, wenn das i-te Unternehmen im Dienstleistungssektor tätig ist, und null, wenn es im produzierenden Gewerbe tätig ist.

a) Du stellst die Spezifikation $m_i = \beta_0 + \beta_1 g_i + \beta_2 p_i + \beta_3 d_i + u_i$ auf. Welches Problem kann hier bestehen?

b) Schlage eine alternative Spezifikation vor, und erläutere die Interpretation von deren Parametern.

Aufgabe 6

Betrachte die multiple Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

für $i = 1, \ldots, n$.

a) Erstelle in Gretl einen Datensatz mit den zufällig generierten Werten x_{i1}, x_{i2}, u_i und mit von dir ausgesuchten Parameterwerten $\beta_0, \beta_1, \beta_2$. Suche dir für β_1 eine außergewöhnliche Zahl aus.
b) Schätze die Parameter im Modell $y_i = \delta_0 + \delta_1 x_{i2} + u_i$
c) Berechne die Residuen $r_{yi} = y_i - \hat{\delta}_0 - \hat{\delta}_1 x_{i2}$
d) Schätze die Parameter im Modell $x_{i1} = \gamma_0 + \gamma_1 x_{i2} + u_i$
e) Berechne die Residuen $r_{1i} = x_{i1} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{i2}$
f) Schätze die Parameter im Modell $r_{yi} = \alpha_0 + \alpha_1 r_{1i} + u_i$
g) Nimmt $\hat{\alpha}_1$ einen ähnlichen Wert an wie β_1?

Aufgabe 7

Benutze den Datensatz hprice1.xls für diese Aufgabe.

a) Schätze das Modell

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

b) Um wie viel steigt der geschätzte Preis eines Hauses, wenn es bei gleicher Fläche (sqrft) ein Schlafzimmer (bdrms) mehr hat? Gibt es mit dieser Frage ein Problem?

c) Wie viel Prozent der Variation im Preis wird von der Regression erklärt?

d) Das erste Haus in der Stichprobe wurde für 300.000$ verkauft (price=300). Berechne das Residuum für dieses Haus. Hat der Käufer/die Käuferin zuviel bezahlt?

Aufgabe 8

Du willst eine lineare Regression $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$ schätzen. Allerdings ist die abhängige Variable $\mathbf{y}$ unbeobachtbar, weil die Daten nur mit Messfehlern erhoben werden. Die gemessenen Daten $\mathbf{z}$ hängen mit der tatsächlichen Variablen $\mathbf{y}$ zusammen gemäß $\mathbf{z} = \mathbf{y} + \mathbf{v}$, wobei $\mathbf{v}$ ein Vektor von zufälligen Messfehlern mit $E[\mathbf{v}] = \mathbf{0}$ ist. Die Zufallsfehler $\mathbf{v}$ und $\mathbf{u}$ sind unabhängig voneinander und von X. Bis auf die Messfehler erfüllt das Modell alle klassischen Annahmen, insbesondere ist die Spezifikation korrekt und es gilt $E[\mathbf{u}|X] = \mathbf{0}$.

a) Du schätzt das Modell mit $\mathbf{z}$ als der abhängigen Variablen mit OLS, und ignorierst dabei zunächst die Messfehler. Gebe die Formel für den Schätzer $\hat{\boldsymbol{\beta}}$ an.

b) Ermittle, ob der unter a) genannte Schätzer erwartungstreu ist.

Aufgabe 9

Benutze die Daten aus *ceosal2.xls* für diese Aufgabe.

1. Schätze ein Modell, welches Jahresgehälter durch Firmenumsätze und Marktwert erklärt. Erstelle hierbei ein Modell mit konstanten Elastizitäten für beide erklärenden Variablen.
2. Füge dem Modell Gewinne (*profits*) hinzu. Warum kann diese Variable nicht in der logarithmischen Form inkludiert werden? Würdest du sagen, dass diese Leistungsgröße der Firma den größten Anteil an der Variation der Jahresgehälter erklärt?
3. Füge dem Modell aus (2.) die Beschäftigungsdauer (*ceoten*) hinzu. Was ist der prozentuale ceteris paribus Effekt eines zusätzlichen Jahres Beschäftigungsdauer auf das Jahresgehalt?
4. Finde die Stichprobenkorrelation zwischen den Variablen ln(*mtkval*) und *profits*. Sind diese Varibalen stark korreliert? Was sagt dies über die OLS-Schätzer aus?

Aufgabe 10

Benutze die Daten aus *wage1.xls* für diese Aufgabe und betrachte die Schätzung

$$\widehat{\ln(wage)} = .284 + 0.92 educ + 0.0041 exper + 0.022 tenure$$

Bestätige den Schätzwert des Parameters für *educ* durch eine partitionierte Regression.

- Erkläre *educ* durch *exper* und *tenure*, schätze das Modell und berechne die Residuen r_1.
- Erkläre $\ln(wage)$ durch r_1. Was fällt dir auf?

Aufgabe 11

Benutze die Daten aus *wage2.xls* für diese Aufgabe.

- Führe eine einfache Regressionsanalyse durch, in welcher du *IQ* durch *educ* erklärst. Nenne den Steigungskoeffizienten $\tilde{\delta}_1$.
- Führe eine einfache Regressionsanalyse durch, in welcher du $\ln(wage)$ durch *educ* erklärst. Nenne den Steigungskoeffizienten $\hat{\alpha}_1$.
- Führe eine multiple Regressionsanalyse durch, in welcher du $\ln(wage)$ durch *educ* und *IQ* erklärst. Nenne die Steigungskoeffizienten $\hat{\beta}_1$ und $\hat{\beta}_2$.
- Bestätige, dass $\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.
- Erkläre das Resultat durch eine geeignete formale Darstellung.

Aufgabe 12

a) Sei X ein n-dimensionaler Zufallsvektor mit endlicher Varianz, $A \in \mathbb{R}^{m \times n}$ eine Matrix und $\mathbf{b} \in \mathbb{R}^m$ ein Vektor. Die Varianz-/Kovarianzmatrix ist definiert als $\Sigma(X) := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']$. Zeige:

$$\Sigma(AX + \mathbf{b}) = A\Sigma(X)A'.$$

b) Sei $X = (X_1, X_2)'$ ein bivariater Zufallsvektor mit

$$\Sigma(X) = \begin{pmatrix} 2 & 0.4 \\ 0.4 & 3 \end{pmatrix}.$$

Berechne die Varianz von $Y := AX + \mathbf{b}$ mit

$$A = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Aufgabe 12

c) Sei X eine n-dimensionale Zufallsvariable mit Erwartungswert μ und Kovarianzmatrix Σ. Berechne die Varianz von $Y = AX$ mit $A = \frac{1}{n}\boldsymbol{\iota}' \in \mathbb{R}^{1\times n}$, wobei $\boldsymbol{\iota}$ den Einsvektor bezeichnet.

d) Sei nun X ein n-dimensionaler Zufallsvektor mit paarweise unkorrelierten Komponenten. Ferner gelte $\mathbb{E}[X_i] = \mu$ und $Var(X_i) = \sigma^2$ für alle $i = 1, \ldots, n$. Bestimme mit Hilfe von Teilaufgabe c) die Varianz des arithmetischen Mittels

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Aufgabe 13

Betrachte die Streungszerlegung:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{=\mathrm{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}_{=\mathrm{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2}_{=\mathrm{SSR}}.$$

Falls das zugrunde liegende Regressionsmodell eine Konstante enthält, ist das (zentrierte) Bestimmtheitsmaß definiert durch

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

a) Zeige, dass

$$R^2 = \widehat{\mathrm{Corr}}(y, \hat{y})^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)},$$

also dass das R^2 den quadrierten empirischen Korrelationskoeffizienten zwischen y und $\hat{y}$ entspricht.

Aufgabe 13

b) Mit obiger Notation können wir das adjustierte Bestimmtheitsmaß definieren als

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\mathrm{SSR}}{\mathrm{SST}}.$$

Zeige, dass $\bar{R}^2 < R^2$.

c) Zeige, dass sich das R^2 bei einer affin-linearen Transformation von y, also durch Ersetzen von y durch $y^* = c_1 + c_2 y$ mit beliebigen reellen Zahlen c_1 und c_2, nicht ändert.

Aufgabe 14 (Datensatz: charity.xls)

a) Schätze die Gleichung

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

per OLS und berichte die Resultate in der üblichen Form, insbesondere auch die Stichprobengröße und R^2. Vergleiche das Bestimmtheitsmaß mit R^2 der einfachen Regression, welche auf *giftlast* und *propresp* verzichtet.

b) Interpretiere den Koeffizienten von *mailsyear*. Ist er größer oder kleiner als die entsprechende Größe aus der einfachen Regression?

c) Interpretiere den Koeffizienten von *propresp*. Beachte hier bei die Maßeinheiten von *propresp*.

d) Füge nun die Variable *avggift* zum Modell hinzu. Was passiert mit dem Koeffizienten von *mailsyear*?

e) In der Gleichung von *d*): Was ist mit dem Koeffizienten von *giftlast* passiert? Was geht hier vor?

Aufgabe 15

Betrachte das multiple Regressionsmodell mit drei unabhängigen Variablen unter den Annahmen MLR 1, MLR 3 und MLR 4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Du interessierst dich für die Summe der Parameter von x_1 und x_2; nenne dies $\theta_1 = \beta_1 + \beta_2$.

a) Zeige, dass $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ ein unverzerrter Schätzer für θ_1 ist.

b) Berechne $var(\hat{\theta}_1)$ in den Ausdrücken $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$ und $corr(\hat{\beta}_1, \hat{\beta}_2)$.

Aufgabe 16

Welche der folgenden Probleme kann verursachen, dass ein OLS-Schätzer verzerrt ist?

- Heteroskedastizität
- Weglassen einer wichtigen Variable
- Eine Stichprobenkorrelation von 0.95 zwischen zwei Regressoren, die beide im Modell enthalten sind
- Hinzufügen einer unwichtigen Variablen