Lineare Regression

Rep-FS21 - Aufgabe 7:

Sie wollen mit nachfolgendem R-Ausdruck eine lineare Regression machen, b durch a erklären:

```
> a <- c(0,1,2,3,4)
> mean(a)
[1] 2
> var(a)
[1] 2.5
> b <- 2*a+rnorm(5)
> b
[1] -1.255575 3.206705 4.049252 8.149300 7.108355
> mean(b)
[1] 4.251607
> var(b)
[1] 13.70371
> cor(a,b)
[1] 0.9255906
> cov(a,b)
[1] 5.417614
```

Hinweis: Es kann gut sein, dass Sie nicht alle obigen Angaben brauchen, um untere Fragen zu beantworten.

a) Wie sind die "wahren" $\beta_0, \beta_1, \sigma^2$ (bei Simulationen gibt es "wahre" Parameter)?

b) Wie sind $\hat{\beta_0}, \hat{\beta_1}, \hat{\sigma^2}$, wenn man mit OLS schätzt?

c) Welchen Wert erhalten Sie für die Teststatistik, wenn Sie testen ob $\beta_1 = 0$ oder nicht?

d) Wie entscheiden Sie in c)? Bitte kritischen Wert und Entscheid angeben, wenn zweiseitig und $\alpha = 0.05$.

Lösung:

a) $\beta_0 = 0, \beta_1 = 2, \sigma^2 = 1$

b) $\hat{\beta_0} = -0.08248,\ \hat{\beta_1} = 2.1671,\ \hat{\sigma^2} = 2.618$

c) 4.235 d) $4.235 > 3.182 \rightarrow \mathcal{H}_0$ ablehnen

FS21 - Aufgabe 7B:

Sie machen eine lineare Regression, um den Zusammenhang zwischen der Körpergrösse von Müttern und Töchtern zu untersuchen. Dabei erhalten Sie folgende Werte $x = (167, 181, 182, 186)$ der Mütter und $y = (180, 180, 188, 191)$ der Töchter. Wir stellen Ihnen wegen der Prüfungssituation folgende Grössen zur Verfügung: $SS_{xx} = 206, SS_{yy} = 94.75, SS_{xy} = 101, \bar{x} = 179, \bar{y} = 184.75$.

a) Berechnen Sie $\hat{\beta}_0$ und $\hat{\beta}_1$.

b) Nehmen Sie ein $\alpha = 0.05$. Wo ist bei einem Test ob $\beta_1 = 0$ der Annahmebereich der Nullhypothese, wenn wir zweiseitig testen wollen, ob es überhaupt einen linearen Zusammenhang gibt? Wo ist bei einem Test ob $\beta_1 = 0$ der Annahmebereich der Nullhypothese, wenn wir einseitig testen wollen, ob grössere Töchter tendenziell grössere Mütter haben?

c) Berechnen Sie die Korrelation zwischen x und y.

d) Wenn (in einem anderen Datensatz) alle Datenpunkte auf einer Gerade der Art $y = a - \frac{1}{2}x$ liegen; wie gross ist dann der Korrelationskoeffizient?

e) Was muss gelten: Die Stichprobengrösse n muss viel grösser sein als die Anzahl erklärende Variablen k oder umgekehrt?

Lösung:

a) $\hat{\beta}_1 = 0.4903,\ \hat{\beta}_0 = 96.988$

b) zweiseitig: $[-4.303, 4.303]$, einseitig: $[-\infty, 2.920]$

c) $r_{xy} = 0.72293$ d) -1 e) $n > k$

FS21 - Aufgabe 7A:

Sie machen eine lineare Regression, um den Zusammenhang zwischen der Körpergrösse von Müttern und Töchtern zu untersuchen. Dabei erhalten Sie folgende Werte $x = (177, 182, 173, 188)$ der Mütter und $y = (182, 182, 171, 190)$ der Töchter. Wir stellen Ihnen wegen der Prüfungssituation folgende Grössen zur Verfügung: $SS_{xx} = 126, SS_{yy} = 182.75, SS_{xy} = 141, \bar{x} = 180, \bar{y} = 181.25$.

a) Berechnen Sie $\hat{\beta}_0$ und $\hat{\beta}_1$.

b) Nehmen Sie ein $\alpha = 0.05$. Wo ist bei einem Test ob $\beta_1 = 0$ der Annahmebereich der Nullhypothese, wenn wir zweiseitig testen wollen, ob es überhaupt einen linearen Zusammenhang gibt? Wo ist bei einem Test ob $\beta_1 = 0$ der Annahmebereich der Nullhypothese, wenn wir einseitig testen wollen, ob grössere Töchter tendenziell grössere Mütter haben?

c) Berechnen Sie die Korrelation zwischen x und y.

d) Wenn (in einem anderen Datensatz) alle Datenpunkte auf einer Gerade der Art $y = a - 2x$ liegen; wie gross ist dann der Korrelationskoeffizient?

e) Was muss gelten: Die Stichprobengrösse n muss viel grösser sein als die Anzahl erklärende Variablen k oder umgekehrt?

Lösung:

a) $\hat{\beta}_1 = 1.1190,\ \hat{\beta}_0 = -20.179$

b) zweiseitig: $[-4.303, 4.303]$, einseitig: $[-\infty, 2.920]$

c) $r_{xy} = 0.92919$ d) -1 e) $n > k$

FS18 - Aufgabe 8:

Unten haben Sie einen R-Ausdruck.

```
> a <- c(1, 2, 3, 4, 5, 6)
> b <- c(1.1, 1.8, 2.7, 4.7, 5.0, 6.8)
> var(a)
[1] 3.5
> var(b)
[1] 4.733667
> d<-lm(b ~ a)
> summary(d)
.....
Coefficients:
Estimate Std. Error t value Pr(>|t|)
Coefficients:
            Estimate    Std. Error    t value    Pr(>|t|)
(Intercept) -0.32667       0.38854     -0.841    0.447810
a            1.14571       0.09977     11.484    0.000328 ***
...
```

Beantworten Sie dazu bitte folgende Fragen.

a) Berechnen Sie den empirischen Korrelationskoeffizienten zwischen a und b, ohne bei Ihrem Taschenrechner eine voreingebaute Funktion zu benutzen. Geben Sie dazu die Formel aus dem Skript an und geben Sie an, welche Zahl Sie wofür einsetzen.

b) Tragen Sie die Punkte und die Regressionsgerade in einer Skizze möglichst genau ein.
Alle Punkte müssen insbesondere auf der richtigen Seite der Regressionsgerade sein, sonst gibt es keinen Punkt. Beschreiben Sie anhand eines der sechs Punkte, mit welcher Rechnung Sie dies erreichen, wenn Sie in Ihrem Taschenrechner keine Graphik anschauen können.

c) Beschreiben Sie mit Worten den Test, für den der p-Wert 0.000328 lautet. Wie lauten $\mathcal{H}_0$ und $\mathcal{H}_1$?

Lösung:

a) $r_{ab} = r_{xy} = \frac{SS_{xy}}{SS_{xx}SS_{yy}} = \hat{\beta}_1 \cdot \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}} = 1.1457 \cdot \frac{\sqrt{5 \cdot 3.5}}{\sqrt{5 \cdot 4.7337}} \approx 0.98516$

b) siehe ML c) 2-seitiger t-Test ob Steigung 0 mit $\mathcal{H}_0: \ \beta_1 = 0, \ \mathcal{H}_1: \ \beta_1 \neq 0$

FS17 - Aufgabe 8:

Unten haben Sie einen R-Ausdruck. Beantworten Sie dazu bitte die folgenden Fragen.

a) Welches ist die erklärende Variable und welches die "response"-Variable?

b) Tragen Sie die Punkte und die Regressionsgerade in einer Skizze möglichst genau ein.

Alle Punkte müssen insbesondere auf der richtigen Seite der Regressionsgeraden sein, sonst gibt es keinen Punkt.

Beschreiben Sie anhand eines der sechs Punkte, mit welcher Rechnung Sie dies erreichen, wenn Sie in Ihrem Taschenrechner keine Graphik anschauen können.

c) Berechnen Sie den Korrelationskoeffizienten zwischen a und b.

d) Welchen Wert hätte die Teststatistik für die Steigung, wenn Sie testen wollen, ob die Steigung -0.5 ist oder nicht?

```
> a<-c(2, 3,4,5,6,7)
> b<-c(4.2, 4.0, 2.8, 3.1, 2.6, 1.3)
> d<-lm(b~a)
> summary(d)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3657     0.4825  11.121 0.000372 ***
a            -0.5257     0.1002  -5.244 0.006322 **
...
```

Lösung:

a) a erklärende Vble, b Response-Vble b) Skizze siehe ML ($\hat{y}_i = 5.3657 - 0.5257 \cdot x_i$)

c) $r_{xy} = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = -0.934$ (von Hand) d) $(-0.5257 - (-0.5))/0.1002 \approx -0.2565$

R-Code selber generieren: (Daten anpassen, Copy & paste)

```
x <- c(2,4,6,8,10)
y <- c(3.1, 2.9, 5.4, 6.7, 5.5)
LinReg <- lm(y~x)
aov(LinReg)
summary(LinReg)
```

FS16 - Aufgabe 8:

Sie wollen wissen, ob der Steuerfuss einer Gemeinde und der Landpreis in m^2 einen Zusammenhang haben. Dazu wählen Sie 5 vergleichbar Gemeinden in Zürich aus; x gebe den Steuerfuss und y den Landpreis an. Es ergeben sich folgende Wertepaare:

$$(88, 2230),\ (93, 2210),\ (100, 1760),\ (85, 2030),\ (110, 1750).$$

Jetzt wollen Sie eine kliene Regressionsanalyse machen:

1.) Berechnen Sie aus den vorhandenen Daten die Steigung der Regressionsgeraden, welche mit der OLS-Methode geschätzt wird.

2.) Tragen Sie die Punkte und die Regressionsgerade in einer Skizze möglichst genau ein.

3.) Berechnen Sie den Korrelationskoeffizienten.

4.) Machen Sie einen zweiseitigen Test auf dem 5 %-Niveau, ob die Steigung gleich 0 ist oder nicht.

5.) Was ist mit diesem Modell der erwartete Landpreis in einer vergleichbaren Gemeinde mit Steuerfuss von 105?

Lösung: 1.) $\hat{\beta}_1 = \frac{SSxy}{SS_{xx}} = -18.06$ 2.) siehe ML ($\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \approx 3715.65$) 3.) $r_{xy} = \frac{SSxy}{\sqrt{SS_{xx}SSyy}} \approx -0.777$
4.) $\mathcal{H}_0 : \beta_1 = 0,\ \mathcal{H}_1 : \beta_1 \neq 0,\ t = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/SS_{xx}}} = -2.135,\ CV = t_{3,0.05} = 3.182,$
$|-2.135| < 3.182 \Rightarrow \mathcal{H}_0$ behalten 5.) $\hat{y}_i = 3715.65 - 18.06 \cdot 105 = 1819.35$

Rep-FS15 - Aufgabe 8:

Auf StudentenwohnenUZH.ch sind zur Zeit 5 Einzimmerwohnungen für immatrikuliere StudentInnen aufgelistet. Neben der Fläche x in Quadratmetern wird auch die monatliche Miete y in CHF angegeben. Es sind dies die folgenden 5 Wertepaare (jeweils (x, y)) : (22, 310), (28, 390), (34, 410), (39, 520), (42, 480).

a) Berechnen Sie mit der OLS-Methode die Regressionsgerade.

b) Tragen Sie diese Gerade und die Punkte in einem Graphen ein.

c) Wenn man die Gültigkeit des Modells unterstellt: welcher Wert wird dann für eine Wohnung von 50 Quadratmetern erwartet?

d) Machen Sie einen einseitigen Test zum Niveau 5 %, ob die Miete mit wachsender Grösse der Wohnung zunimmt oder nicht.

e) Formulieren Sie unter Annahme der Gültigkeit der Regressionsgerade den Zusammenhang zwischen Quadratmetern Wohnfläche und Mietzins in Worten.

Lösung: a) $\hat{y}_i = 110.75 + 9.432x_i$ c) 582.35 CHF d) $4.696 > 2.353 \Rightarrow \mathcal{H}_1$ annehmen
e) Die Miete steigt pro Quadratmeter um 9.43 CHF

FS15 - Aufgabe 8:

Sie haben bei einer Untersuchung 5 Datenpunkte in der x-y-Ebene erhalten. Es sind dies die folgenden 5 Punkte (2,8), (3,12), (4,17), (5,16), (6,19). Jetzt wollen Sie eine kleine Regressionsanalyse machen:

1.) Berechnen Sie aus den vorhandenen Daten die Steigung der Regressionsgeraden, welche mit der OLS-Methode geschätzt wird.

2.) Tragen Sie die Punkte und die Regressionsgerade in einer Skizze möglichst genau ein.

3.) Berechnen Sie den Korrelationskoeffizienten.

4.) Machen Sie einen zweiseitigen Test auf dem 5%-Niveau, ob die Steigung gleich 0 ist oder nicht.

Lösung: 1.) $\hat{\beta}_1 = 2.6$ und $\hat{\beta}_0 = 4.0$ 2.) siehe ML 3.) $r_{xy} = 0.9358$ 4.) $t = 4.596 > 3.182 \Rightarrow \mathcal{H}_1$ annehmen

R-Output NEU für alte Aufgaben selber generieren!! (siehe R-Befehl Seite oberhalb)

```
Call:
   aov(formula = fs15)

Terms:
                    x Residuals
Sum of Squares  67.6       9.6
Deg. of Freedom    1         3

Residual standard error: 1.788854
Estimated effects may be unbalanced
> summary(fs15)

Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4    5
-1.2  0.2  2.6 -1.0 -0.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0000     2.4000   1.667   0.1942
x             2.6000     0.5657   4.596   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.789 on 3 degrees of freedom
Multiple R-squared:  0.8756,Adjusted R-squared:  0.8342
F-statistic: 21.13 on 1 and 3 DF,  p-value: 0.01936
```

Rep-FS14 - Aufgabe 8:

Theoretische Aufgaben zur Regression:

a) Beweisen Sie, dass die OLS-Schätzung der Steigung der Regressionsgeraden erwartungstreu ist.

b) Zeigen Sie, dass der Mittelpunkt der Daten $(\overline{x}, \overline{y})$ immer auf der OLS-Regressionsgeraden liegt.

Lösung: siehe Musterlösung

FS14 - Aufgabe 8:

Auf dem nächsten Blatt finden Sie einen lückenhaften R-Ausdruck. Lösen Sie bitte untenstehende Aufgabe dazu.

1. Berechnen Sie aus den vorhandenen Daten die Steigung der Regressionsgeraden.
2. Tragen Sie die Punkte und die Regressionsgerade in einer Skizze möglichst genau ein.
3. Berechnen Sie das SSR aus der Vorlesung mit Hilfe der angegebenen Daten.
4. Machen Sie einen zweiseitigen Test auf dem 5%-Niveau, ob die Steigung gleich 0 ist oder nicht.
5. Wenn Sie das Niveau α variieren, wo ist die Grenze in Fragestellung d) wo $\mathcal{H}_0$ bzw $\mathcal{H}_1$ gerade noch angenommen bzw. abgelehnt wird? Es geht uns hier nicht um die Frage $\geq, >, <, \leq$, sondern einfach, wo die Grenze selber ist - die richtige Zahl reicht als Lösung.

```
x <- c(2,4,6,8,10)
y <- c(3.1, 2.9, 5.4, 6.7, 5.5)
LinReg <- lm(y~x)
aov(LinReg)

Call:
   aov(formula = A8)

Terms:
                     x Residuals
Sum of Squares   fehlt     3.532
Deg. of Freedom      1         3

Residual standard error: 1.08505
Estimated effects may be unbalanced
> summary(A8)

Call:
lm(formula = y ~ x)

Residuals:
    1     2     3     4     5
 0.10 -0.96  0.68  1.12 -0.94

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1400     1.1380   1.880   0.1566
x             fehlt      0.1716   2.506   0.0872 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 3 degrees of freedom
Multiple R-squared:  0.6768,Adjusted R-squared:  0.5691
F-statistic: 6.282 on 1 and 3 DF,  p-value: 0.08721
```

Lösung: 1. $\widehat{\beta}_1 = 0.43$ 2. siehe Musterlösung 3. SSR = 7.396
4. p-Wert = 8.72% $> 5\% = \alpha \Rightarrow \mathcal{H}_0$ annehmen 5. bei $\alpha = 8.72\%$ = p-Wert