Lehrstuhl für Statistik und emp. Wirtschaftsforschung, Prof. Regina T. Riphahn, Ph.D.

Musterlösung zur Diplomvorprüfung Statistik II – Einf. Ökonometrie im WS 06/07

- korrigierte Fassung v. 18.07.2007 -

Aufgabe 1: **[21 Punkte]**

Ein Forschungsinstitut hat den Auftrag bekommen, die individuellen monatlichen Ausgaben für Bioprodukte zu erklären. Es wird eine Kleinstquadrate Regression der Höhe der Ausgaben für Bioprodukte (*bio*, in €) für 744 Personen durchgeführt.

Erklärende Variablen sind zwei Entfernungsdummies, welche die Distanz zum nächsten Bioladen erfassen (*distance2*=höchstens 5 km, *distance3*=mehr als 5km, Referenz: *distance1*=höchstens 2 km), das Alter (*age*, in Jahren), Alter im Quadrat (*age2*) sowie das logarithmierte Einkommen (*log(income)*).

```
Call:
lm(formula = bio ~ distance2 + distance3 + age + age2 + log(income))

Coefficients:
             Estimate   Std. Error
(Intercept)   15.7831       4.3945
distance2     -0.0935       0.0314
distance3     -3.5792       1.1152
age            3.1854       0.9817
age2          -0.0235       0.0085
log(income)    0.2847       0.0947
---

Residual standard error: 0.8185 on 738 degrees of freedom
Multiple R-Squared: 0.6377,                          Adjusted R-squared: ?
F-statistic:  5497 on 5 and 738 DF,  p-value: < 1.7e-13
```

a) Berechnen Sie ***(9 Punkte)***

a1) den *t*-Wert für den Koeffizienten von *age2*. Führen Sie außerdem für *age2* einen Signifikanztest auf dem 1%-Niveau durch und geben Sie die Null- und Alternativhypothese sowie Ihre Schlusslogik.

- $H_0 : \beta_{age2} = 0 \qquad H_1 : \beta_{age2} \neq 0$
- $t_{age2} = \dfrac{-0.0235 - 0}{0.0085} = -2.764$
- $t_c = t_{1\%,738} = \pm 2.576$
- $\left|t_{age2}\right| > \left|t_c\right|$
- Daher kann die Nullhypothese auf dem 1%-Signifikanzniveau verworfen werden.

a2) den Wert des korrigierten Bestimmtheitsmaßes.

- $R^2 = 1 - \dfrac{SSE}{SST} \Rightarrow SST = \dfrac{SSE}{\left(1 - R^2\right)}$ mit $SSE = \hat{\sigma}^2 \cdot (T-6)$, wobei $\hat{\sigma}^2 = \left(Res.\,std.\,error\right)^2 = 0.8185^2$
- $\bar{R}^2 = 1 - \dfrac{SSE/(T-6)}{SST/(T-1)} = 1 - \dfrac{494.417/738}{1364.663/743} = 0.6354$

a3) ein 90%-Konfidenzintervall um den Koeffizienten von *log(income)*. Interpretieren Sie das Ergebnis.

- $t_c = t_{10\%,738} = \pm 1.645$
- $b_{\log(inc)} \pm t_c \cdot se\left(b_{\log(inc)}\right) = 0.2847 \pm 1.645 \cdot 0.0947 = \left(0.1289\,;\,0.4404\right)$

- Interpretation:

  Der Intervallschätzer sagt aus, dass bei wiederholter Anwendung der Berechnungsmethode für verschiedene Stichproben in 90% der Fälle die Auswirkung einer Erhöhung des Einkommens um 1% zu einer Erhöhung der Ausgaben für Bioprodukte führt, die innerhalb der berechneten Intervallgrenzen von 0.1289/100 (€) und 0.4404/100 (€) liegt.

b) Testen Sie die Aussage „Eine Steigerung des Einkommens um 1% erhöht die Höhe der Ausgaben für Bioprodukte um mehr als 30 Cent" auf dem 1% Signifikanzniveau. ***(2 Punkte)***

- $H_0 : \frac{\beta_{\log(income)}}{100} \leq 0.3 \qquad H_1 : \frac{\beta_{\log(income)}}{100} > 0.3$
- $t = \frac{0.2847/100 - 0.3}{0.0947/100} = -313.7835$

  Erläuterung zum Nenner-Term: $se(b) = \sqrt{Var(b)}$. Für b/100 folgt sodann:

  $$\sqrt{Var\left(\frac{b}{100}\right)} = \sqrt{\frac{Var(b)}{100^2}} = \frac{1}{100} \cdot se(b)$$

- $t_c = 2.326$
- $t < t_c$
- Daher kann die H_0 auf einem Signifikanzniveau von 1% nicht verworfen werden.

c) Testen Sie auf dem 5% Signifikanzniveau, ob der marginale Effekt des Alters *∂bio/∂age* für 20-Jährige gleich 2,7 ist. (Hinweis: cov(b_{age},b_{age2}) = -0.00154) ***(5 Punkte)***

- $\frac{\partial bio}{\partial age} = \beta_{age} + 2 \cdot \beta_{age2} \cdot age$
- $H_0 : \beta_{age} + 2 \cdot \beta_{age2} \cdot age = 2.7 \qquad H_1 : \beta_{age} + 2 \cdot \beta_{age2} \cdot age \neq 2.7$
- $t = \frac{(b_{age} + 40 \cdot b_{age2}) - 2.7}{se(b_{age} + 40 \cdot b_{age2})}$
- $(b_{age} + 40 \cdot b_{age2}) = 3.1854 + 40 \cdot (-0.0235) = 2.245$
- $se(b_{age} + 40 \cdot b_{age2}) = \sqrt{\text{var}(b_{age} + 40 \cdot b_{age2})} =$

  $= \sqrt{\text{var}(b_{age}) + 1600 \cdot \text{var}(b_{age2}) + 80 \cdot \text{cov}(b_{age}, b_{age2})} =$

  $= \sqrt{0.956} = 0.978$
- $t = \frac{2.245 - 2.7}{0.987} = -0.465$
- $t_c = t_{5\%,738} = \pm 1.96$
- $|t| < |t_c|$
- Daher kann die H_0 nicht verworfen werden.

d) **Spielt die Entfernung zum nächsten Bioladen eine signifikante Rolle für die Ausgaben? Testen Sie auf einem Signifikanzniveau von 5%. Geben Sie die Null- und Alternativhypothese an und beschreiben Sie kurz allgemein die Vorgehensweise des Tests.** ***(3 Punkte)***

- $H_0 : \beta_{Dist2} = \beta_{Dist3} = 0 \qquad H_1 : \beta_{Dist2} \neq 0 \vee \beta_{Dist3} \neq 0$

- $F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$

- $F_c = F_{5\%,2,738} = 3.0$

- Wenn $F > F_{5\%,2,738} = 3.0$, kann die H_0 auf dem 5% Signifikanzniveau verworfen werden.

e) Wie ändert sich die durch die Regression vorhergesagte Höhe der Ausgaben für Bioprodukte, wenn ***(2 Punkte)***

e1) die Distanz zum nächsten Bioladen von einem auf vier km steigt?

- Ändert sich die Entfernung zum nächsten Bioladen von 1 auf 4 km, sinken *c.p.* die Ausgaben für Bioprodukte im Vergleich zur Referenzkategorie um 9.35 Cent (= $b_{\text{distance2}}$).

e2) das Einkommen um 5% steigt?

- Steigt das Einkommen um 5%, so steigen *c.p.* die Ausgaben für Bioprodukte um ca. 0.014 € (= $5 \cdot b_{\text{log(income)}} \cdot 0.01$).

Aufgabe 2: **[13 Punkte]**

Sie sind an der Frage interessiert, was die bestimmenden Einflussfaktoren für die Höhe von Löhnen *(log(einkommen))* sind. Dazu führen Sie für 1482 Personen eine Kleinstquadrate Schätzung durch, in der Sie als erklärende Variablen das Bildungsniveau (*bildg*, gemessen in Ausbildungsjahren), Erfahrung (*erfahrung*, Alter minus Ausbildungsdauer minus 6 Jahre), sowie Erfahrung im Quadrat (*erfahrung2*) berücksichtigen.

```
Call:
lm(formula = log(einkommen) ~ bildg + erfahrung + erfahrung2)


Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  6.205846   2.947777    2.105    0.039
bildg        0.107294   0.041789    2.567 4.15e-10
erfahrung    0.085791   0.015913    5.391 2.97e-12
erfahrung2  -0.001285   0.000615   -2.089 4.84e-04
---

Residual standard error: 0.04825 on 1478 degrees of freedom
Multiple R-Squared: 0.2856,    Adjusted R-squared: 0.2714
F-statistic: 29.35 on 3 and 1478 DF,  p-value: < 2.1e-15
```

```
2SLS Estimates

Model Formula: log(einkommen) ~ bildg + erfahrung + erfahrung2

Instruments: ~ IQ + berfahrung + berfahrung2


            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.759874   3.559742    1.618    0.472
bildg        0.438741   0.158963    2.760 5.36e-11
erfahrung    0.067489   0.012552    5.376 4.97e-06
erfahrung2  -0.002691   0.002179   -3.171 4.28e-04

Residual standard error: 0.9955 on 1478 degrees of freedom
```

Lehrstuhl für Statistik und emp. Wirtschaftsforschung, Prof. Regina T. Riphahn, Ph.D.

Musterlösung zur Diplomvorprüfung Statistik II – Einf. Ökonometrie im WS 06/07

- korrigierte Fassung v. 18.07.2007 -

---

a) Sie vermuten, dass es sich bei der Variable Bildung um eine endogene Variable handelt. ***(2 Punkte)***

a1) Welche Annahme des linearen Regressionsmodells wird bei Endogenität verletzt?

- Die Annahme, dass die Kovarianz zwischen einer erklärenden Variable *x* und dem Störterm Null ist: $cov(x,e)=0$.

a2) Welche Auswirkungen hat dies auf die Schätzung?

- Der KQ-Schätzer ist nicht mehr konsistent. Das bedeutet, im Grenzfall von $T \to \infty$ konvergiert die Varianz des KQ-Schätzers zwar gegen Null, die Wahrscheinlichkeitsverteilung des KQ-Schätzers kollabiert jedoch nicht auf dem wahren Parameterwert.

b) Sie wollen die Variable IQ, die den Intelligenzquotienten misst (*IQ* sei metrisch skaliert), als Instrumentvariable für *bildg* verwenden. Unter welchen Umständen ist dies sinnvoll? ***(2 Punkte)***

- Die Kovarianz zwischen IQ und Bildung sollte möglichst hoch sein und IQ selbst sollte nicht mit dem Störterm der ursprünglich geschätzten Gleichung für *log(einkommen)* korreliert sein.

c) Gehen Sie davon aus, dass die Variable IQ die Voraussetzungen erfüllt. Sie benutzen IQ als Instrument und erhalten den oben angegebenen R-Output einer two-stage least-squares Regression. Wie ist Ihre Vermutung bezüglich der Korrelationsrichtung der Bildungsvariable und des Störterms? Begründen Sie. ***(3 Punkte)***

- $b_{bildg}^{KQ} < b_{bildg}^{2sls}$
- Weil $\hat{\beta} = \frac{\text{cov}(x,y)}{\text{var}(x)} - \frac{\text{cov}(x,e)}{\text{var}(x)}$ folgt:
- Der KQ-Schätzer überschätzt den wahren Parameter, wenn *cov(x,e) > 0* ;
- Der KQ-Schätzer unterschätzt den wahren Parameter, wenn *cov(x,e) < 0* .
- Da $b_{bildg}^{KQ} < b_{bildg}^{2sls}$, kann man auf *cov(x,e) < 0* schließen.

d) Erläutern Sie präzise, wie man sich mit Hilfe eines Hausman-Tests zwischen der Kleinstquadrate Schätzung und der Instrumentvariablenschätzung entscheiden kann. Nennen Sie die Hypothesen und beschreiben Sie die Vorgehensweise zur Durchführung des Tests. Wie lautet das Testergebnis? ***(6 Punkte)***

```
 Hausman Test

Model Formula: log(einkommen) ~ bildg + erfahrung + erfahrung2 + uhat

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.758874   3.649132    1.578    0.412
bildg        0.338279   0.117254    3.020 5.71e-12
erfahrung    0.182536   0.030376    6.010 3.62e-05
erfahrung2  -0.002381   0.000960   -2.480 4.31e-04
uhat        -0.036455   0.011524   -3.163 4.14e-03
```

Vorgehensweise:

- Schätzung der Hilfsregression: $bildg_t = a_0 + a_1 \cdot IQ_t + \upsilon_t$
- Schätzung der Regression: $\log(einkommen)_t = \beta_0 + \beta_1 \cdot bildg_t + \beta_2 \cdot erfahrung + \beta_3 \cdot erfahrung^2 + \delta\hat{\upsilon}_t + e_t$
- Die Hypothesen lauten $H_0 : \delta = 0$ und $H_1 : \delta \neq 0$

- *t*-Test der $H_0 : \delta = 0$ gegen $H_1 : \delta \neq 0$;
- führt hier zum Verwerfen der H_0, da `uhat` statistisch signifikant ist. Das Ergebnis des Tests ist, dass der IV-Schätzer dem KQ-Schätzer vorzuziehen ist.

Aufgabe 3: **[17 Punkte]**

Sie möchten herausfinden, wie sich die Diplomnote auf die Produktivität von wissenschaftlichen Assistenten auswirkt. Anhand der geschriebenen Artikelseiten im vergangenen Jahr (pages) von 1084 wissenschaftlichen Assistenten in Deutschland betrachten Sie den Einfluss folgender Faktoren auf die Produktivität:
***DA* = Alter zum Zeitpunkt des Diploms (in Jahren), *DN* = Diplomnote (zwischen 1,0 und 3,0), *F* = Geschlecht (Frau=1, Mann=0), *Ki* = Anzahl der Kinder unter 6 Jahren.**

Sie formulieren folgendes Modell:

$$\mathbf{pages_t = \beta_1 + \beta_2 * DA_t + \beta_3 * DN_t + \beta_4 * F_t + \beta_5 * (F_t * Ki_t) + \beta_6 * Ki_t + e_t}$$

Die Auswertung der Daten mit R ergibt folgenden Output:

```
Call:
lm(formula = pages ~ DA + DN + F + F*Ki + Ki)

Residuals:
      Min        1Q    Median        3Q       Max
-2.383135 -0.281114  0.008058  0.281124  2.021414

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.10490  0.97331    5.245 1.88e-07 ***
DA             1.3209   0.1075    12.285  < 2e-16 ***
DN             0.64052  0.06618    9.679  < 2e-16 ***
F              0.81412  1.26354    0.644   0.5195
I(F*Ki)       -0.21708  0.09657    2.248   0.0248 *
Ki            -2.95842  0.26417  -11.199  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4269 on 1078 degrees of freedom
Multiple R-Squared: 0.3844,     Adjusted R-squared: 0.3816
F-statistic: 134.6 on 5 and 1078 DF,  p-value: < 2.2e-16
```

a) Sie vermuten, dass Sie entscheidende Variablen ausgelassen haben. Deshalb schätzen Sie Ihr Modell erneut, diesmal zusätzlich mit den prognostizierten Werten der abhängigen Variablen in quadrierter Form als erklärender Variable. ***(7 Punkte)***

a1) Nennen Sie den Test, den Sie mit dieser Schätzung durchführen und beschreiben Sie kurz seine Idee.

- RESET-Test: Der geschätzte Parameter von $\hat{y}^2$ ist signifikant, wenn die ausgelassene Variable mit *x* korreliert, weil dann deren Einfluss über den Parameter von $\hat{y}^2$ aufgefangen würde.

a2) Die nun berechnete Varianz des Störterms beträgt 0,1752. Testen Sie, ob Sie entscheidende Variablen ausgelassen haben. Geben Sie dabei die Null- und Alternativhypothesen, den Wert der Teststatistik, die kritischen Werte und die Schlussfolgerung an.

- H_0: = Der Parameter von $\hat{y}^2$, $b_{\hat{y}^2,} = 0 \rightarrow$ Modell korrekt spezifiziert,

- H_1: = $b_{\hat{y}^2} \neq 0 \rightarrow$ Modell nicht korrekt spezifiziert.

  $SSE_R = \sigma_R^2 (T-K) = 0{,}1822 \cdot (1084-6) = 196{,}4586$

  $SSE_U = \sigma_U^2 (T-K) = 0{,}1752 \cdot (1084-7) = 188{,}6904$

  $$F = \frac{(196{,}458 - 188{,}6904)/1}{188{,}6904/1077} = \frac{7{,}7682}{0{,}1752} = 44{,}339$$

  Kritischer Wert für $F_{1,1077}$ mit α=0,05: $F_c = 3{,}84$, also: $F > F_c \rightarrow H_0$ ablehnen.
- Das Modell ist nicht korrekt spezifiziert.

b) Sie glauben, dass sich der Zusammenhang zwischen den erklärenden Variablen und der Produktivität für Männer und Frauen grundlegend unterscheidet. Schreiben Sie eine Schätzgleichung auf, mit der Sie diese Hypothese testen können und nennen Sie die Nullhypothese. ***(2 Punkte)***

- Idee: Zusätzliche Interaktionen zwischen *F* und *DA* sowie zwischen *F* und *DN* ins Modell aufnehmen:
- $pages_t = \beta_1 + \beta_2 DA_t + \beta_3 DN_t + \beta_4 F_t + \beta_5 (F_t \cdot K_{it}) + \beta_6 Ki_t + \beta_7 (DA_t \cdot F_t) + \beta_8 (DN_t \cdot F_t) + e_t$
- Teste H_0: $\beta_4 = \beta_5 = \beta_7 = \beta_8 = 0$.

c) Welche Produktivitätsunterschiede ergeben sich aus der Schätzung für ***(5 Punkte)***
c1) Männer und Frauen ohne Kinder?

- b_F: Frauen sind um 0,814 Seiten produktiver. [Beachte aber: b_F ist nicht statistisch signifikant.]

c2) Frauen mit 2 und ohne Kindern?

- $2 \cdot (b_{Ki} + b_{F \cdot Ki}) = 2 \cdot (-2{,}95842) + 2 \cdot (-0{,}21708) = -6{,}35096$
- Das heißt, Frauen mit 2 Kindern sind um etwa 6,35 Seiten weniger produktiv als Frauen ohne Kinder.

c3) Männer mit 2 Kindern und Frauen ohne Kindern?
Erläutern Sie kurz, welche der jeweiligen Teilgruppen um wie viel produktiver ist.

- Die Regressionsgleichung für Männer mit 2 Kindern lautet $pages_t = \ldots + 2 \cdot (-2{,}95842)$, die für Frauen ohne Kinder $pages_t = \ldots + 0{,}81412$. Die Differenz zwischen beiden Gruppen berechnet sich also als $2 \cdot (-2{,}95842) - 0{,}81412 = -6{,}73096$. Das heißt, Männer mit 2 Kindern sind um 6,73 Seiten weniger produktiv als Frauen ohne Kinder.

d) In Querschnittsregressionen findet man häufig Heteroskedastie. ***(3 Punkte)***
d1) Welche Annahme des linearen Regressionsmodells wird durch Heteroskedastie verletzt?

- Verletzung der Annahme einer konstanten Varianz des Störterms: $Var(e_t) = \sigma^2$.

d2) Welche Auswirkungen hat Heteroskedastie auf den Schätzer?

- Parameter werden weiterhin unverzerrt geschätzt, sind aber nicht mehr BLUE, da die Standardfehler falsch geschätzt werden. Damit sind Testverfahren und Intervallschätzung ungültig.

Aufgabe 4: **[8 Punkte]**

Unterstellen Sie das einfache Regressionsmodell

$$y_t = \beta_1 + \beta_2 x_t + e_t ,$$

wobei e_t unabhängige Fehlerterme sind, mit $E(e_t)=0$ und $var(e_t)=\sigma^2 h_t$. Unterstellen Sie die fünf Beobachtungen $y_t=(0,4,2,3,5)$ und $x_t=(1,9,2,4,6)$. Ermitteln Sie die Generalized Least Squares Schätzwerte b_1 und b_2 für $h_t=(1,1,4,4,4)$.

- $y^* = \frac{y}{\sqrt{n}}$; $x^* = \frac{x}{\sqrt{n}}$

	y	x	$\sqrt{n}$	y^*	x^*	x^{*2}	y^*x^*
	0	1	1	0	1	1	0
	4	9	1	4	9	81	36
	2	2	2	1	1	1	1
	3	4	2	1,5	2	4	3
	5	6	2	2,5	3	9	7,5
Summe				9	16	96	47,5

- $b_2 = \frac{T \cdot \sum x_t^* y_t^* - \sum x_t^* \sum y_t^*}{T \cdot \sum x_t^{*2} - \left(\sum x_t^*\right)^2} = \frac{5 \cdot 47,5 - 16 \cdot 9}{5 \cdot 96 - 16^2} = 0,4174$;

- $b_1 = \bar{y}^* - b_2 \bar{x}^* = \frac{9}{5} - 0,417 \cdot \frac{16}{5} \simeq 0,4643$.

Aufgabe 5: **[12 Punkte]**

In R wurde folgende fehler- und lückenhafte Funktion programmiert:

```
my.konfint <= function(koeff,se,niveau;T)
 (
  tc = qt((1-???)/2,T-2,lowertail=F]
  return(seq(??? - tc*??? ; ??? + tc*???))
 }
```

a) Die Funktion enthält 6 Fehler. Markieren Sie diese deutlich und schreiben Sie die korrekte Formulierung darüber. ***(6 Punkte)***
(Das Erkennen eines Fehlers gibt 0,5 Punkte, die richtige Verbesserung gibt ebenfalls 0,5 Punkte. Es werden 0,5 Punkte abgezogen, wenn entweder eine richtige Stelle als Fehler gekennzeichnet wird oder der Fehler zwar richtig erkannt wird, die Verbesserung aber falsch ist. Die Gesamtpunktzahl kann nicht negativ werden.)

b) Ergänzen Sie die Funktion an den mit Fragezeichen versehenen Stellen so, dass ein Konfidenzintervall ausgegeben wird (unter den Fragezeichen notieren) ***(4 Punkte)***

– Korrigierte und ergänzte Funktion:

```
my.konfint <- (oder =) function(koeff,se,niveau,T)
 {
  tc = qt((1-niveau)/2,T-2,lower.tail=F)
  return(c(koeff - tc*se , koeff + tc*se))
 }
```

c) Ihnen liegen folgende Werte vor: `koeff` = 3.1415, `se` = 1.2345, T = 100. Welchen Befehl müssen Sie eingeben, um mit der Funktion ein Konfidenzintervall auf dem 5%-Signifikanzniveau zu berechnen. ***(2 Punkte)***

– my.konfint(3.1415,1.2345,0.95,100)

Aufgabe 6: **[10 Punkte]**

Welche Antwort ist richtig? Bitte kreuzen Sie die zutreffende Antwort an. Zu jeder Frage gibt es nur eine richtige Antwort. Für jede korrekt angekreuzte Antwort gibt es 1 Punkt, für jede falsch angekreuzte Antwort wird 1 Punkt abgezogen. Die Gesamtpunktzahl kann nicht negativ werden.

1.	Mit welchem R-Befehl erzeugen Sie einen Vektor x, der die Zahlen von 1 bis 100 enthält?	
	☒	`> x <- seq(1:100)`
	❑	`> x <- vector(1:100)`
	❑	`> x <- seq(1 to 100)`

2.	Mit welchem R-Befehl erzeugen Sie einen Plot zweier Objekte X und Y?	
	❑	`> plot(x,y)`
	☒	`> plot(X,Y)`
	❑	`> plot(x;y)`

3.	Mit welchem R-Befehl können Sie eine Funktion "`my.function`" editieren?	
	❑	`> get(my.function)`
	❑	`> change(my.function)`
	☒	`> fix(my.function)`

4.	Welchen der folgenden R-Befehle können Sie **nicht** verwenden, um die vorhergesagten Werte eines linearen Modells zu generieren?	
	❑	`> predict(lm(y ~ x))`
	☒	`> summary(lm(y ~ x))`
	❑	`> fitted.values(lm(y ~ x))`

5.	Welchen R-Befehl müssen Sie verwenden, um ein log-lineares Modell zu schätzen?	
	❑	`> lm(log(y) ~ log(x))`
	❑	`> lm(y ~ x, model=log.lin)`
	☒	`> lm(log(y) ~ x)`

6.	Mit welchem der folgenden R-Befehle kann man den *i*-ten Koeffizienten eines vorher geschätzten Modells auslesen (`s` ist der Modelloutput `summary(...)`)?	
	☒	`> s$coef[i]`
	❑	`> s$beta[i]`
	❑	`> s$bhat[i]`

7.	Welchen R-Befehl müssen Sie verwenden, um den Datensatz `data.dat` einzulesen?	
	❑	`> get.data("C:/StatistikII/data.dat")`
	☒	`> read.table("C:/StatistikII/data.dat")`
	❑	`> read.data("C:/StatistikII/data.dat")`

8.	Welche Option des R-Befehls `plot(...)` verwendet man, um die Beschriftung der X-Achse zu ändern?	
	❑	`> xtxt`
	☒	`> xlab`
	❑	`> xaxis`

9.	Mit welchem R-Befehl bestimmt man den kritischen Wert einer χ^2-Verteilung mit einem Freiheitsgrad bei einem Signifikanzniveau von 5%?	
	❑	`> pchisq(0.95, df=1)`
	☒	`> qchisq(0.95, df=1)`
	❑	`> dchisq(0.975, df=1,lower.tail=F)`

10.	Welche Kennzahl berechnet man mit dem Befehl `cov(x,y)/sqrt((var(x)*var(y)))`.	
	☒	Korrelationskoeffizient zwischen x und y
	❑	bedingte Kovarianz zwischen x und y
	❑	Randverteilung von x und y

Aufgabe 7: **[30 Punkte]**

Wahr oder falsch? …

F	Der Gini-Koeffizient beschreibt den Anteil der Fläche zwischen Gleichverteilungskurve und Lorenzkurve an der gesamten Fläche unter der Lorenzkurve.
W	Bei der Kleinstquadrate-Schätzung wird eine Gerade so durch die Punktwolke gelegt, dass die Summe der quadrierten vertikalen Abweichungen der beobachteten Werte von der Geraden minimiert
F	Die Fehlerquadratsumme eines restringierten Modells (SSE_R) entspricht der gesamten Variation des unrestringierten Modells (SST_U).
W	Nimmt die Durbin-Watson Teststatistik den Wert 0 an, so kann die Nullhypothese, dass es keine Autokorrelation gibt, verworfen werden.
W	Der beim *t*-Test ausgegebene *p*-Wert gibt das Signifikanzniveau an, bei dem der empirische *t*-Wert dem kritischen *t*-Wert entspricht.
F	Die Varianz des Vorhersagefehlers entspricht der geschätzten Stichprobenvarianz $\hat{\sigma}^2$.
W	Die Multiplikation der Varianz mit einer Konstanten modifiziert die Varianz um das Quadrat der Konstanten.
W	Irrelevante Variablen im Regressionsmodell führen zu niedrigeren t-Werten der relevanten erklärenden Variablen.
W	Wenn die Störterme nicht normalverteilt sind, dann ist in großen Stichproben dennoch der Kleinstquadrate-Schätzer approximativ normalverteilt.

W	Im Zufallsfehler *e* spiegeln sich alle die Faktoren, die die abhängige Variable beeinflussen und im Regressionsmodell nicht berücksichtigt wurden.
F	Wenn die Nullhypothese beim Chow-Test verworfen wird, dann können die Daten gepoolt werden.
W	Je größer die Streuung der erklärenden Variablen, desto geringer ist die Varianz des Steigungsparameters im einfachen linearen Modell.
W	Wenn eine relevante Variable fehlerhaft gemessen ist, ist sie mit dem Fehlerterm korreliert, und der KQ-Schätzer ist inkonsistent.
F	Wenn die Regressionsgerade horizontal verläuft, ist das Bestimmtheitsmaß 1.
W	Beim F-Test entspricht die Zahl der Freiheitsgrade im Zähler der Anzahl der Restriktionen.
F	White's Proxy Schätzer korrigiert für verzerrte Parameterschätzer bei Heteroskedastie.
F	Bei diskreten Zufallsvariablen entspricht die Fläche unter der Wahrscheinlichkeitsdichtefunktion der Wahrscheinlichkeit, die Variable zu beobachten.
F	Das korrigierte R^2 steigt, wenn weitere erklärende Variablen im Modell berücksichtigt werden.
F	Wenn alle k Preise um den gleichen Prozentsatz λ steigen, steigt der Paasche-Index um k*λ.
F	Die Werte der abhängigen Variable y müssen im linearen Modell für jeden Wert der erklärenden Variable x um ihren Mittelwert normalverteilt sein.
W	Eine Änderung der Skalierung von ausschließlich der erklärenden Variable *x* um den Faktor *k* bewirkt eine Änderung im zugehörigen Koeffizienten um den Faktor *(1/k)*.
W	Bei Heteroskedastie sind die für den KQ-Schätzer berechneten Standardfehler nicht korrekt.
W	Bei Autokorrelation ist der KQ-Schätzer unverzerrt.
F	Die geschätzte Fehlertermvarianz entspricht der Summe der quadrierten Fehler.
W	Durch eine Dummyvariable können Unterschiede in Steigungsparametern für Teilstichproben modelliert werden.
W	Der LM-Test auf Autokorrelation ist im Gegensatz zum Durbin-Watson Test nur approximativ gültig.
W	Der Schätzwert für die Konstante in einem Regressionsmodell, β_1, ist eine Zufallsvariable.
W	Das Modell mit autokorrelierten Störtermen unterstellt, dass Schocks über eine Periode hinaus wirken.
F	Das (1-α)% Konfidenzintervall für den Steigungsparameter β_2 besagt, dass der wahre Wert von β_2 mit einer Wahrscheinlichkeit von (1-α) im beschriebenen Intervall liegt.
W	Es gibt Modelle, die nichtlinear in Variablen sind und gleichzeitig linear in den unbekannten Parametern.

Aufgabe 8: **[11 Punkte]**

Welche Antwort ist richtig? …

1.	Ein Parameter β_2 wird umso präziser geschätzt	
	❑	je geringer die Streuung in x ist.
	❑	je kleiner der t-Wert von β_2 ist.
	☒	je geringer die Streuung von y ist.

2.	Der t-Wert für den Steigungsparameter einer einfachen Regressionsschätzung mit 25 Beobachtungen beträgt -1,9.	
	☒	H_0: $\beta_2 \geq 0$ gegen H_1: $\beta_2 < 0$ wird auf dem 5% - Niveau verworfen.
	❑	H_0: $\beta_2 = 0$ gegen H_1: $\beta_2 \neq 0$ wird auf dem 5% Niveau verworfen.
	❑	Die t-Verteilung ist von der Anzahl der geschätzten Parameter unabhängig.

3.	Wenn die Varianz des Störterms über die Beobachtungen hinweg nicht konstant ist,	
	❑	werden die Standardabweichungen der Parameter konsistent geschätzt.
	❑	können die Parameter nicht mehr unverzerrt geschätzt werden.
	☒	ist damit zu rechnen, dass die Teststatistik des Goldfeld-Quandt Tests größer als der kritische F-Wert ist.

4.	Punktschätzer sind	
	❑	informativer als Intervallschätzer.
	❑	nicht auf Basis von Stichproben interpretierbar.
	☒	umso verlässlicher, je kleiner die geschätzte Fehlervarianz $\hat{\sigma}^2$.

5.	Eine Division der abhängigen und unabhängigen Variablen durch 1000 führt zu	
	❑	einem um den Faktor 1000 erhöhten Achsenabschnittsparameter.
	☒	unveränderten Steigungsparametern.
	❑	um den Faktor 1000 reduzierten Werten für alle Achsen- und Steigungsparameter.

6.	Wenn die Residuen einer einfachen linearen Regression mit 15 Beobachtungen nicht normalverteilt sind,	
	❑	muss die Nullhypothese des Reset Tests für die Residuen verworfen werden.
	☒	ist der F-Test auf Gesamtsignifikanz des Modells nicht exakt gültig.
	❑	sind b_1 und b_2 nicht die „besten linearen und unverzerrten Schätzer“ von β_1 und β_2.

7.	Ausgelassene Variablen	
	❑	sollte man vermeiden, indem man möglichst viele Variablen in das Modell aufnimmt.
	☒	können zu verzerrten Parameterschätzern führen.
	❑	führen zu einer höheren Varianz der geschätzten Parameter.

8.	Multikollinearität	
	❑	führt dazu, dass das Modell nicht mehr geschätzt werden kann.
	☒	kann durch Berücksichtigung externer Information im Regressionsmodell reduziert werden.
	❑	führt zu hohen t-Werten.

9.	In der deskriptiven Zeitreihenanalyse wird der Trend	
	❑	linear modelliert, wenn sich die Zeitreihe mit einer konstanten Wachstumsrate ändert.
	☒	häufig aus den KQ-Koeffizienten einer linearen Schätzgleichung berechnet.
	❑	bei logistischen Trendmodellen immer positiv sein.

10.	Interaktionseffekte im Regressionsmodell	
	☒	können verwendet werden, um Parameter für zwei Teilstichproben zu vergleichen.
	❑	können mit dem RESET-Test auf ihre Signifikanz überprüft werden.
	❑	führen dazu, dass Koeffizienten nicht mehr interpretiert werden können.

11.	Bei einer einfachen linearen KQ-Schätzung mit logarithmierten Werten für x und y	
	☒	können die Parameter als Elastizitäten interpretiert werden.
	❑	ändert sich y um β_2, wenn sich x um 1 Prozent ändert.
	❑	ändert sich y um β_2%, wenn sich x um eine Einheit ändert.