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

Musterlösung zur Diplomvorprüfung Statistik II – Einf. Ökonometrie im WS 05/06

- korrigierte Fassung v. 27.06.2007 -

Aufgabe 1: **[45 Punkte]**

Eine Mitarbeiterin eines Nürnberger Statistik-Lehrstuhls möchte herausfinden, welche Faktoren die Diplomnote von Studierenden beeinflussen. Sie befragt dazu T = 100 zufällig ausgewählte Absolventen, die gerade ihr Diplom bekommen haben, zu folgenden Aspekten:

DNote: Diplomnote (Wertebereich von 1,0 bis 4,0)
VDNote: Note im Vordiplom (Wertebereich von 1,0 bis 4,0)
Sem: Zahl der Fachsemester bis zum Erhalt des Diploms
Alter: Alter der befragten Person in Jahren bei Erhalt des Diploms
Sex: Geschlecht (weiblich = 1, männlich = 0)
BY: Person stammt aus Bayern (ja = 1, nein = 0)

Die Mitarbeiterin unterstellt, dass eine Note ein stetiges, metrisches (= quantitatives) Merkmal ist, das hier in 0,1-Schritten gemessen wurde, und formuliert folgendes Modell:

$$DNote_t = \beta_0 + \beta_1 \cdot VDNote_t + \beta_2 \cdot \ln\left(Sem_t\right) + \beta_3 \cdot Alter_t + \beta_4 \cdot Sex + \beta_5 \cdot BY + e_t$$

Die Auswertung der Daten mit R ergibt folgenden Output:

```
Call:
lm(formula = DNote ~ VDNote + log(Sem) + Alter + Sex + BY)

Coefficients:
             Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)  -2.0289613     1.2423869     -1.633     0.1058
VDNote        0.6257252     0.0754450        ?       7.7e-13 ***
log(Sem)      1.4027298     0.5366281      2.614     0.0104 *
Alter        -0.0006449     0.0473704     -0.014     0.9892
Sex          -0.0132849         ?         -0.099     0.9215
BY           -0.2538084     0.1210013     -2.098     0.0386 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5945 on 94 degrees of freedom
Multiple R-Squared: ?,     Adjusted R-squared: 0.5146
F-statistic: 21.99 on ? and ? DF,  p-value: 1.582e-14
```

a) Berechnen Sie unter Angabe des Rechenwegs die im Output fehlenden Werte für **(5 Punkte)**

a1) den t-Wert für b_1;

- $t_{b_2} = \dfrac{b_1}{se\left(b_1\right)} = \dfrac{0.6257}{0.0754} = 8.2984$

a2) den Standardfehler für b_4;

- $t_{b_4} = \dfrac{b_4}{se\left(b_4\right)} \Rightarrow se\left(b_4\right) = \dfrac{b_4}{t_{b_4}} = \dfrac{-0.01328}{-0.099} = 0.1341$

a3) Die Zahl der Freiheitsgrade für den totalen F-Test (df1 und df2);

- df1 = K-1 = 6-1 = 5
- df2 = T-K = 100-6 = 94

a4) die geschätzte Fehlertermvarianz;

- $\hat{\sigma}^2 = \left(\hat{\sigma}\right)^2 = 0.5945^2 = 0.3534$

a5) das Bestimmtheitsmaß.

- $\overline{R}^2 = 1 - \frac{SSE/T-K}{SST/T-1}$
- $\overline{R}^2 = 0,5146$
- T = 100
- K = 6
- $\overline{R}^2 = 0,5146 = 1 - \frac{SSE/T-K}{SST/T-1} \Rightarrow \frac{SSE \cdot (T-1)}{SST \cdot (T-K)} = 0,4854$
  $\Rightarrow \frac{SSE}{SST} = 0,4854 \cdot \frac{T-K}{T-1} = 0,4854 \cdot \frac{94}{99} = 0,4609$
- $R^2 = 1 - \frac{SSE}{SST} = 1 - 0,4609 = 0,5391$

b) Welche der unabhängigen Variablen haben einen signifikanten Einfluss auf die Diplomnote und welche nicht (α=5%)? Begründen Sie Ihre Entscheidung. (2,5 Punkte)

- Bei einem Signifikanzniveau von 5% haben die Vordiplomsnote, der Logarithmus der Semesterzahl sowie der Dummy Bayern einen signifikanten Einfluss, da ihre p-Werte jeweils kleiner als 5%=0,05 sind. Bei Alter und Geschlecht trifft dies nicht zu, sie haben also keinen signifikanten Einfluss.

c) Betrachten Sie die Koeffizienten b_1, b_2 und b_5. (6 Punkte)

c1) Angenommen Student A war im Vordiplom um 5 Noteneinheiten, d.h. eine halbe Note besser als Student B. Was erwarten Sie, um wie viele Noteneinheiten wird A im Diplom besser als B sein, wenn A und B ansonsten in allen Merkmalen übereinstimmen? Begründen Sie. *(2 Punkte)*

- Eine Noteneinheit entspricht 0,1. Ein um eine Note besseres Vordiplom führt im Durchschnitt zu einem um 0,6257·1 = 0,6257 besseren Diplom. War das Vordiplom um eine halbe Note (0,5) besser, erwartet man also eine Diplomnote, die um 0,5·0,6257=0,31285 Einheiten besser ist (also bspw. statt einer 2,0 eine 1,7)

c2) Angenommen Student C und Student D stimmen in allen Merkmalen überein, außer dass C aus Bayern stammt und D nicht. Wenn D als Diplomnote eine 2,0 hat, welche Diplomnote erwarten Sie dann für C? Begründen Sie. *(2 Punkte)*

- Der marginale Effekt von *BY* beträgt -0,25, also ein Viertel einer Notenstufe. Das heißt ein bayerisch-stämmiger Student ist im Durchschnitt um 0,25 Noteneinheiten besser als ein nicht aus Bayern stammender. Man würde also erwarten, dass C als Diplomnote eine 2,0 – 0,25 = 1,75 hat.

c3) Was gibt der Koeffizient b_2 an? Interpretieren Sie.

- Der Koeffizient b_2 ist ein Schätzwert für die Semi-Elastizität der Semesterzahl auf die Diplomnote (auch okay: b_2 gibt die Semi-Elastizität der Semesterzahl auf die Diplomnote an – dass es ein Schätzwert ist, muss nicht unbedingt erwähnt werden). Das heißt, b_2 gibt an, um wie viele Einheiten sich die Diplomno-

te ändert, wenn die Semesterzahl um 1 Prozent steigt. Wenn also die Semesterzahl um 1% zunimmt, verschlechtert sich im Durchschnitt die Diplomnote um 0,014 Einheiten.

d) Mit einem Interaktionsterm kann der gemeinsame Einfluss von zwei Merkmalen untersucht werden. Nennen Sie zwei sinnvolle Interaktionsterme, die die Mitarbeiterin in das obige Modell aufnehmen könnte und interpretieren Sie diese inhaltlich. (4 Punkte)

- einige denkbare Interaktionsterme:
  - VDNote*Sex: Ist der Einfluss der Vordiplomsnote auf die Diplomnote für Männer und Frauen gleich?
  - VDNote*BY: Unterscheidet sich der Einfluss der Vordiplomnote auf die Diplomnote zwischen bayerisch-stämmigen und nicht bayerisch-stämmigen Studenten?
  - Alter*Sex: Ist der Alterseffekt für Männer und Frauen gleich groß?
  - Sex*BY: Aus Bayern stammende Studierende sind im Durchschnitt im Diplom etwas besser als nicht aus Bayern stammende Studierende. Ist dieser Effekt für Männer und Frauen gleich?
  - usw.

e) Eine häufig genannte Annahme bei Regressionsschätzungen ist die der Normalverteilung. Der Jarque-Bera-Test bietet eine Möglichkeit, diese zu überprüfen. (7 Punkte)

e1) Beschreiben Sie kurz, was die Normalverteilungsannahme aussagt und warum sie notwendig ist.

- Die Normalverteilungsannahme besagt, dass die Residuen einer KQ-Schätzung normalverteilt sind mit Mittelwert Null und konstanter Varianz. Sollen die Koeffizienten auf Signifikanz getestet werden und Konfidenzintervalle berechnet werden, so sind Annahmen über die Verteilung der Schätzfunktionen nötig. Diese Annahmen sind nur dann korrekt, wenn die Residuen (und damit auch y) einer Normalverteilung folgen.

e2) Mit welchem R-Befehl rufen Sie den Jarque-Bera-Test für das obige Modell auf?

> jarque.bera.test(resid(lm(DNote ~ VDNote + log(Sem) + Alter + Sex + BY)))

Erklärung dazu: Es soll getestet werden, ob die Residuen des Modells normalverteilt sind. Daher *jarque.bera.test(resid(…))*, wobei mit *>resid(Modell)* die Residuen des Modells angesprochen werden.

e3) Der Jarque-Bera-Test aus e2) liefert folgenden Output:

```
Jarque Bera Test
Chi-squared = 0.8799, df1 = 2, p-value = 0.644
```

Begründen Sie kurz, ob die Normalverteilungsannahme hier als erfüllt betrachtet werden kann oder nicht (α=5%).

- Der p-Wert dieses Tests liegt bei 0.644=64.4% und ist damit deutlich größer als 5%. Das heißt, der JB-Test liefert kein signifikantes Ergebnis, die Annahme (und Nullhypothese) der Normalverteilung der Residuen kann also empirisch nicht verworfen werden.

f) Die Mitarbeiterin vermutet, dass es für die Diplomnote auch eine Rolle spielt, ob ein Studierender Statistik als Schwerpunktfach im Hauptstudium gewählt hatte (Variable „Stat“, ja=1, nein=0). Sie schätzt dazu das obige Modell zusätzlich separat für Statistiker und Nicht-Statistiker. Die jeweiligen ANOVA-Tabellen sind im Folgenden angegeben (Tabelle 1: nur Statistiker, Tabelle 2: nur Nicht-Statistiker, Tabelle 3: Statistiker und Nicht-Statistiker zusammen). Führen Sie einen Chow-Test auf dem 5%-Niveau durch. Geben Sie dabei auch die Null- und Alternativhypothese, die Teststatistik, die Verteilung der Teststatistik und die Ablehnungsregion an. Interpretieren Sie das Ergebnis. (9 Punkte)

Tabelle 1: **Analysis of Variance Table** Response: DNote[Stat==1]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VDNote[Stat==1]	1	8.5878	8.5878	22.0595	6.881e-05 ***
log(Sem)[Stat==1]	1	2.4646	2.4646	6.3310	0.01811 *
Alter[Stat==1]	1	0.4689	0.4689	1.2044	0.28213
Sex[Stat==1]	1	0.0066	0.0066	0.0171	0.89701
BY[Stat==1]	1	1.0397	1.0397	2.6708	0.11382
Residuals	27	10.5111	0.3893		

Tabelle 2: **Analysis of Variance Table** Response: DNote[Stat==0]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VDNote[Stat==0]	1	25.5642	25.5642	73.3512	4.727e-12 ***
log(Sem)[Stat==0]	1	1.0556	1.0556	3.0287	0.08684
Alter[Stat==0]	1	0.4328	0.4328	1.2418	0.26949
Sex[Stat==0]	1	0.0818	0.0818	0.2346	0.62989
BY[Stat==0]	1	0.4903	0.4903	1.4068	0.24018
Residuals	61	21.2596	0.3485		

Tabelle 3: **Analysis of Variance Table** Response: DNote

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VDNote	1	34.261	34.261	96.9548	3.894e-16 ***
log(Sem)	1	3.014	3.014	8.5301	0.004373 **
Alter	1	0.019	0.019	0.0539	0.816887
Sex	1	0.007	0.007	0.0195	0.889226
BY	1	1.555	1.555	4.3998	0.038627 *
Residuals	94	33.217	0.353		

- $DNote_t = Ursprungsmodell + \delta_1 \cdot Stat_t + \delta_2 \cdot VDNote_t \cdot Stat_t + \delta_3 \cdot \log(Sem_t) \cdot Stat_t + \\ + \ldots + \delta_6 \cdot BY_t \cdot Stat_t + e_t$

- $H_0: \delta_1 = \delta_2 = \ldots = \delta_6 = 0$
  H_1: mindestens ein $\delta_i \neq 0$

- $$F = \frac{\frac{SSE_R - SSE_U}{J}}{\frac{SSE_U}{T-K}} \sim F_{m_1 = J;\, m_2 = T-K}$$

- J = 6, T = 100, K = 2·J = Anzahl der Parameter im unrestringierten Modell = 12
  $\Rightarrow m_1 = 6$, $m_2 = 100 - 12 = 88$

- $F_{0,95;\,6;\,88} \approx F_{0,95;\,6;\,60} = 2{,}25$
  (Für $m_2 = 88$ sind keine F-Werte tabelliert. In der Klausur hatten wir deshalb darauf hingewiesen, in einem solchen Fall den nächstgelegenen Wert zu verwenden.)

  $\Rightarrow$ Ablehnungsregrion = [2,25; ∞[

- $SSE_U = 21{,}2596 + 10{,}5111 = 31{,}7707$

- $SSE_R = 33{,}217$

$$\Rightarrow F = \frac{\frac{(33,217 - 31,7707)}{6}}{\frac{31,7707}{88}} = 0,6677 \notin \text{AR} \Rightarrow H_0 \text{ nicht ablehnen.}$$

- Interpretation: Es bestehen keine signifikanten Unterschiede zwischen den geschätzten Parameterwerten für Statistiker und Nicht-Statistiker, die Daten können also gepoolt werden.

g) Ein bei der Schätzung von Regressionsmodellen häufig auftretendes Phänomen ist die sog. Autokorrelation. (11,5 Punkte)

g1) Erläutern Sie kurz verbal und formal, was man unter Autokorrelation versteht.

- Bei Autokorrelation liegen Korrelationsmuster bei zeitlich aufeinander folgenden Fehlertermen vor. Die Annahme, dass $cov(e_t,e_s) = 0$ gilt dann also nicht mehr.

g2) Nennen Sie zwei Konsequenzen von Autokorrelation für KQ-Schätzer.

- Mgl. Auswahl:
  - KQ-Schätzer weiterhin unverzerrt, aber nicht mehr BLUE
  - Standardfehler und damit Präzision der Parameterschätzwerte werden falsch ausgewiesen
  - Hypothesentests und Konfidenzintervalle sind nicht mehr verlässlich

g3) Die Mitarbeiterin unterstellt die Gültigkeit des AR(1)-Modells. Sie ruft in R mit `> dwtest(lm(DNote ~ VDNote + log(Sem) + Alter + Sex + BY, alternative = c("greater"))` einen Durbin-Watson-Test auf und erhält dabei folgendes Ergebnis:

```
Durbin-Watson test
data:  lm(DNote ~ VDNote + log(Sem) + Alter + Sex + BY)
DW = 1.2432, p-value = 5.57e-05
alternative hypothesis: true autocorrelation is greater than 0
```

Geben Sie für diesen Test die Null- und Alternativhypothese an und begründen Sie, ob Autokorrelation vorliegt oder nicht.

- Hypothesen:
  - H_0: $\rho=0$ (oder H_0: $\rho<=0$)
  - H_1: $\rho>0$ (positive Autokorrelation – eine andere H_1 ist falsch!)
- Begründung:

  Der p-Wert ist mit $5{,}57 \cdot 10^{-5}$ deutlich kleiner als 5%, der Test liefert also ein signifikantes Ergebnis. Das heißt, die Nullhypothese kann verworfen werden, positive Autokorrelation liegt vor.
- (nicht erwartet) Teststatistik und Ergebnis:

$$d = DW = \frac{\sum_{t=2}^{T=100} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T=100} \hat{e}_t^2} = 1,2432 \text{ (auch } d \approx 2 \cdot (1-\rho)\text{), bei } e_t = \rho e_{t-1} + \nu_t$$

g4) Hilft Ihnen die Kenntnis des vorliegenden AR(1)-Parameters bei der Vorhersage der abhängigen Variablen für die nächste Beobachtung, T = 101. Begründen Sie Ihre Antwort.

- Kenntnis des AR(1)-Parameters ist hilfreich bei der Vorhersage von T = 101. Die Information über den Störwert in t-1 (in t) kann herangezogen werden, um die Vorhersage für e_t (für e_{t+1}) zu verbessern:

  $y_{101} = \beta_0 + \beta_1 \cdot x_{101} + e_{101}$,

  mit $e_{101} = \rho \cdot e_{100} + \upsilon_{101}$

  folgt $y_{101} = \beta_0 + \beta_1 \cdot x_{101} + \rho \cdot e_{100} + \upsilon_{101}$.

- Eine Vorhersage für $\beta_0 + \beta_1 \cdot x_{101}$ ergibt sich als $\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{101}$, ρ kann mit $\hat{\rho}$ geschätzt werden und eine Vorhersage für e_{100} erhält man über das geschätzte Residuum $\hat{e}_{100} = y_{100} - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{100}$. Die Vorhersage für υ_{101} ist Null (da die Größe nicht mit verzögerten Werten korreliert und ν einen Erwartungswert von Null hat.

- Es ergibt sich also: $y_{101} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{101} + \hat{\rho}\left(y_{100} - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{100}\right)$.

Aufgabe 2: **[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.	Ein Dataframe unterscheidet sich von einer Matrix dadurch, dass	
	❑	eine Matrix nicht mehrere Variablen enthalten kann.
	☒	eine Matrix nur Vektoren des gleichen Datentyps enthalten darf.
	❑	eine Matrix überhaupt kein R-Objekt ist.

2.	Welche Aussage ist korrekt?	
	❑	In R ist das Dezimaltrennzeichen das Komma.
	☒	R unterscheidet zwischen Groß- und Kleinschreibung.
	❑	In R werden Befehlszeilen mit einem Punkt abgeschlossen.

3.	Mit welchem der folgenden Befehle berechnet man in R den arithmetischen Mittelwert der Elemente des Vektors x?	
	☒	`> mean(x)`
	❑	`> mw(x)`
	❑	`> middle(x)`

4.	Mit welchem der folgenden R-Befehle kontrolliert man für den quadrierten Wert der Variable x?	
	☒	`> lm(y~w+I(x^2)+z)`
	❑	`> lm(y~w+x^2+z)`
	❑	`> lm(y~w+x*x+z)`

5.	Welche Kennzahl der Häufigkeitsverteilung der Elemente des Vektors x berechnet man mit dem R-Befehl `> sum(x)/length(x)` ?	
	❑	Standardabweichung
	❑	Median
	☒	arithmetischen Mittelwert

6.	Welchen Wert berechnet man mit folgender Formel (y ist die abhängige Variable einer linearen Regression): `> sum((y – mean(y))^2)`?	
	❑	SSE
	❑	SSR
	☒	SST

7.	Mit welchem R-Befehl bestimmt man den kritischen Wert einer F-Verteilung mit 8 und 21 Freiheitsgraden bei einem Signifikanzniveau von 5%?	
	❑	`> pf(0.05,8,21)`
	❑	`> qf(0.05,8,21)`
	☒	`> qf(0.95,8,21)`

8.	Welche Schreibweise des Befehls `> read.table()` ist korrekt, um den Datensatz "Daten.txt" einzulesen?	
	❑	`> read.table("C:\Studien\Daten.txt",header=T)`
	☒	`> read.table("C:/Studien/Daten.txt",header=T)`
	❑	`> read.table("C:/Studien/Daten",header=T)`

9.	Sie wollen die Wahrscheinlichkeit dafür berechnen, dass eine mit 15 Freiheitsgraden t-verteilte Zufallsvariable höchstens den Wert 0,6 annimmt. Mit welchem R-Befehl erhalten Sie das richtige Ergebnis?	
	❑	`> dt(0.6,15)`
	☒	`> pt(0.6,15)`
	❑	`> pt(0.6,15) – pt(0,15)`

10.	Mit welchem der folgenden Parameter des Befehls `> plot()` ändern Sie die Beschriftung der y-Achse?	
	☒	`ylab`
	❑	`ylim`
	❑	`yaxis`

Aufgabe 3: **[12 Punkte]**

In R wurde folgende Funktion programmiert:

```
> auswertung = function(x,y)
      {
            r = cor(x,y)
            reg=lm(y~x)
            plot(x,y)
            abline(reg)
            return(r)
      }
```

Der Datensatz, auf den diese Funktion angewendet werden soll, enthält die beiden Variablen x und y, die die folgenden Ausprägungen haben:

t	x_t	y_t
1	3	-3
2	4	-4
3	5	-5

a) Mit welchen Befehlen geben Sie diesen Datensatz in R ein?

- Da es sich nur um zwei sehr kurze Vektoren handelt, können die Daten händisch eingegeben werden:

  > *x=c(3,4,5)* (statt "=" geht auch "<-")

  > *y=c(-3,-4,-5)*

- Der Laufindex t wird nicht als eigene Variable eingegeben, da R die Elemente in Vektoren automatisch durchnummeriert.

b) Welche Ergebnisse gibt R bei folgenden Befehlen aus?

b1) `> x[2]`

- Das zweite Element des Vektors x, also 4.

b2) `> x[y==-4]`

- Das Element von x, dessen entsprechendes Element von y den Wert -4 hat, also 4.

b3) `> x[y<6]`

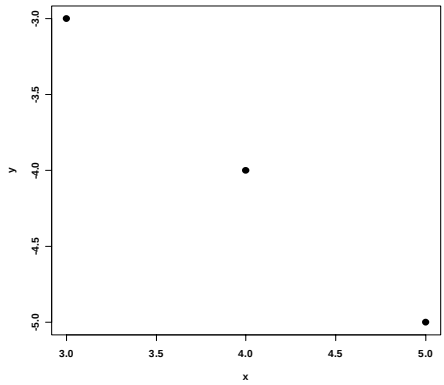
- Die Elemente von x, deren entsprechende Elemente von y kleiner als 6 sind, also 3, 4, 5.

c) Welchen Befehl müssen Sie in R eingeben, um die Funktion für den gegebenen Datensatz auszuführen?

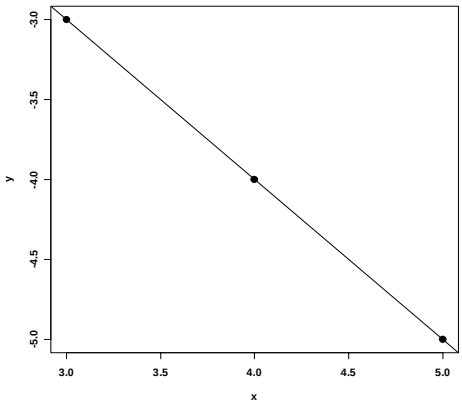
> *auswertung(x,y)*

d) Geben Sie alle Ausgaben an, die mit dieser Funktion für den Datensatz erzeugt werden.

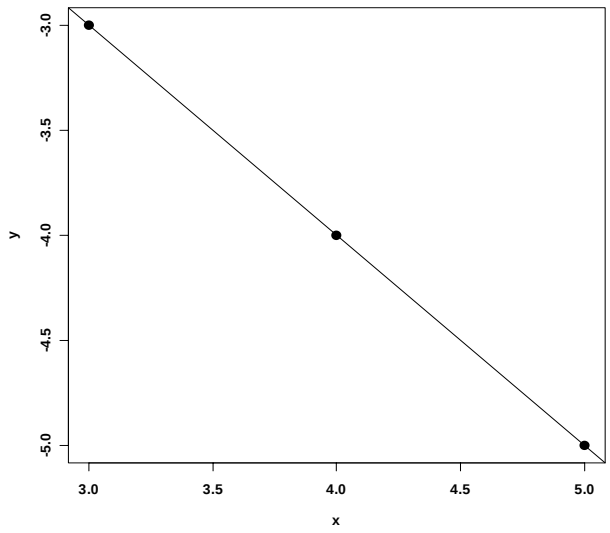
- Zunächst wird der Variablen r der Wert des Korrelationskoeffizienten zwischen x und y zugewiesen. Da die Beträge der Elemente von x und y jeweils gleich, nur die Vorzeichen entgegengesetzt sind, ist der Korrelationskoeffizient zwischen x und y gleich -1. Eine Ausgabe erfolgt an dieser Stelle aber noch nicht.
- Als nächstes wird das Regressionsmodell $y_t = \beta_0 + \beta_1 \cdot x_t + e_t$ geschätzt. Eine Ausgabe findet hier auch noch nicht statt.
- Der Befehl >*plot(x,y)* führt zur ersten Ausgabe. Es wird folgende Grafik erzeugt:



- Die Grafik in den Klausuren sollte zumindest so ähnlich aussehen. **Wichtig ist, dass überhaupt eine Grafik gezeichnet wird** (dass also überhaupt erkannt wird, dass eine Grafik ausgegeben wird) und dass die Punkte auf einer Linie liegen. Im Idealfall stimmt auch die Achsenbeschriftung.
- Der nächste Befehl ist *>abline(reg)*. Damit wird eine Gerade in die bestehende Punktwolke eingezeichnet, die als Achsenabschnitt und Steigungsparameter den Achsenabschnitt und Steigungsparameter der Regression "reg" aufweist. Da hier eine perfekte Korrelation vorliegt, geht die Regressionsgerade durch alle drei Punkte:



- Schließlich wird mit dem Befehl *>return(r)* der Inhalt der Variable r ausgegeben. Dies ist der Wert des am Anfang berechneten Korrelationskoeffizienten zwischen x und y, also -1.
- Das heißt, die Musterlösung sieht so aus:



-1

Aufgabe 4: **[25 Punkte]**

Wahr oder falsch? Tragen Sie für jede der folgenden Aussagen ein „w“ für „wahr“ oder ein „f“ für „falsch“ ein. Für jede richtige Antwort gibt es 1 Punkt, für jede falsche Antwort wird 1 Punkt abgezogen. Die Gesamtpunktzahl kann nicht negativ werden.

W	Die Grundidee des F-Tests besteht darin, den Erklärungsgehalt unterschiedlicher Modelle zu vergleichen.
W	Immer wenn entweder kategorische erklärende Variablen oder Dummyvariablen im linearen Modell betrachtet werden, muss eine Referenzgruppe gebildet werden.
F	Nimmt eine Zufallsvariable einen beliebigen Wert (z.B. x = 2) an, so ist der an dieser Stelle berechnete Wert der Wahrscheinlichkeitsfunktion größer als der der Wahrscheinlichkeitsdichtefunktion.
F	Beim Laspeyres-Index entspricht das Produkt von Preis- und Mengenindex der Umsatzmesszahl.
F	Der Ausdruck $\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$ gilt nur für stetige Zufallsvariablen.
W	Eine Vorhersage auf Basis eines linearen Modells ist genau dann unverzerrt, wenn der Erwartungswert des Vorhersagefehlers 0 beträgt.
F	Der KQ Schätzer ist inkonsistent, wenn cov(x,e) = 0.
W	Der Herfindahl-Index ist ein absolutes Konzentrationsmaß.
W	Empirisches Arbeiten lässt sich mit den Forderungen des kritischen Rationalismus nach Falsifikation von Theorien begründen.
W	Der KQ-Schätzer ist (asymptotisch) normalverteilt.
F	Erwartungswert und Varianz der Chi^2-Verteilung sind identisch.

W	Das Gauss-Markov-Theorem macht keine Aussage zu nichtlinearen Schätzverfahren.
W	Der Chow-Test benutzt die F-Verteilung.
F	Im linearen Modell gibt die Regressionskonstante den Mittelwert der abhängigen Variablen an.
W	Auf der Hauptdiagonale der Varianz-Kovarianz Matrix der geschätzten Parameter befinden sich ausschließlich die Varianzen der einzelnen Parameter.
W	Die gemeinsame Dichtefunktion f(X,Y) zweier unabhängiger Zufallsvariablen X und Y unterscheidet sich von der gemeinsamen Dichtefunktion zweier korrelierter Zufallsvariablen.
W	Bei Autokorrelation erster Ordnung in den Fehlertermen eines linearen Modells lässt sich ein effizienter KQ Schätzer gewinnen, wenn die Daten vor der Schätzung transformiert werden.
F	Grundidee des KQ Schätzers ist, eine Linie so durch eine Punktwolke zu legen, dass die Summe der quadrierten horizontalen Abweichungen der beobachteten Werte von der Linie minimiert wird.
W	Sobald heteroskedastische Fehlerterme vorliegen, ist der KQ Schätzer ineffizient.
W	Unverzerrte Schätzer der Störtermvarianz erhält man nur, wenn die Freiheitsgrade als T – K berücksichtigt werden, wobei T die Anzahl der Beobachtungen und K die Anzahl der geschätzten Parameter (inklusive der Konstanten) ist.
W	Zu den Methoden statistischer Inferenz gehören das Schätzen, das Testen und das Vorhersagen.
W	Um zu prüfen, ob eine erklärende Variable signifikant ist, die als Polynom dritter Ordnung in der Regression berücksichtigt wurde, sollte der F-Test genutzt werden.
W	Bei linearen Regressionen wird die Schätzgüte mit dem Wert des R^2 gemessen.
F	Je größer der Gini-Koeffizient, umso gleichmäßiger die Verteilung.
W	Wenn a eine Konstante ist und Y eine Zufallsvariable, dann gilt Var(Y-a) = Var(Y).

Aufgabe 5: **[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.	Wenn c eine Konstante ist und X und Y Zufallsvariablen sind, dann ist die Varianz von (cX – Y)	
	❑	$c^2 \cdot Var(X) – Var(Y)$
	❑	$c \cdot Var(X – Y)$
	☒	$c^2 \cdot Var(X) + Var(Y) – 2 \cdot c \cdot Cov(X,Y)$

2.	Um mit Hilfe eines linearen KQ Schätzers Koeffizienten zu gewinnen, die als Elastizitäten von Y hinsichtlich X interpretiert werden können, muss man	
	❑	die erklärende Variable X als Polynom zweiter Ordnung schätzen.
	❑	die abhängige Variable Y logarithmiert betrachten.
	☒	abhängige und erklärende Variable (Y und X) in logarithmierter Form betrachten.

3.	KQ Parameterschätzer sind Zufallsvariablen, weil	
	☒	sie als gewichtete Summe von Zufallsvariablen beschrieben werden können.
	❑	das Schätzverfahren keine exakten Werte ergibt.
	❑	Intervallschätzer keine präzise Interpretation zulassen.

4.	Bei einem Hypothesentest ist die Ablehnungsregion	
	❑	umso größer, je niedriger das Signifikanzniveau.
	❑	unabhängig von der Typ-I Fehlerwahrscheinlichkeit.
	☒	abhängig von der Anzahl der Beobachtungen.

5.	Eine Division der erklärenden Variable X_k durch den Faktor a führt zu	
	❑	einem um den Faktor a reduzierten Parameterschätzwert für β_k.
	☒	einem um den Faktor a erhöhten Parameterschätzwert für β_k.
	❑	um den Faktor a erhöhten Schätzwerten für alle Steigungsparameter des Modells.

6.	Die Präzision der Schätzung eines Steigungsparameters ist umso höher,	
	❑	je weniger Beobachtungen vorliegen.
	❑	je mehr Parameter geschätzt werden.
	☒	je größer die Streuung der erklärenden Variable.

7.	Ausgelassene relevante erklärende Variable führen dann nicht zu verzerrten Parameterschätzern,	
	❑	wenn die ausgelassene Variable mit dem Störterm korreliert ist.
	❑	wenn die Standardfehler heteroskedastisch sind.
	☒	wenn die ausgelassene Variable mit den berücksichtigten Variablen nicht korreliert ist.

8.	Der Goldfeld-Quandt Test	
	❑	überprüft, ob die Störterme des Modells miteinander korreliert sind.
	❑	hat T-K Freiheitsgrade.
	☒	wird in der Regel als einseitiger Test durchgeführt.

9.	Bei gegen unendlich konvergierender Stichprobengröße	
	❑	konvergiert der Intervallschätzer der Steigungsparameter gegen das Signifikanzniveau.
	☒	konvergiert die Varianz des KQ Schätzers gegen Null.
	❑	konvergiert das R^2 gegen 1.

10.	Ein RESET Test mit quadrierten und kubischen vorhergesagten Werten ($\hat{y}^2$ und $\hat{y}^3$) der abhängigen Variable ergibt eine Teststatistik von 4,8 mit einem p-Wert von 0,067. Dies bedeutet:	
	❑	Das Modell sollte in logarithmierter Form geschätzt werden.
	❑	Am Signifikanzniveau von 10% wird H_0 nicht verworfen.
	☒	Am Signifikanzniveau von 5% ist das Modell nicht fehlspezifiziert.

Aufgabe 6: **[18 Punkte]**

Brada und Graves argumentieren in ihrem 1998 erschienen Artikel "The slowdown in Soviet Defense Expenditures" (*Southern Economic Journal*, 969-984), dass die sowjetischen Ausgaben für Verteidigung eine Funktion des sowjetischen Bruttosozialprodukts und der Verteidigungsausgaben der USA seien. Weniger sicher waren sie sich über den Einfluss der Anzahl sowjetischer Nuklearsprengköpfe im Vergleich zur Anzahl US-amerikanischer Nuklearsprengköpfe. Um ihre Hypothesen zu überprüfen, verwenden sie Jahresdaten von 1960 bis 1984 und schätzen zwei Modellspezifikationen:

$$\ln(SDH_t) = \beta_0 + \beta_1 \ln(USD_t) + \beta_2 \ln(SY_t) + e_1 \quad (1)$$

$$\ln(SDH_t) = \beta_0 + \beta_1 \ln(USD_t) + \beta_2 \ln(SY_t) + \beta_3 \ln(SP_t) + e_2 \quad (2)$$

wobei SDH_t: Sowjetische Verteidigungsausgaben im Jahr t (in 1970 Mill. Rubel, Schätzung durch die CIA)
USD_t: US-amerikanische Verteidigungsausgaben im Jahr t (in 1980 Mill. US$)
SY_t: sowjetisches Bruttosozialprodukt im Jahr t (in 1970 Mill. Rubel)
SP_t: Verhältnis der Anzahl sowjetischer Nuklearsprengköpfe zur Anzahl US-amerikanischer Nuklearsprengköpfe

Die Schätzung von Spezifikation (1) mit R liefert folgende Ergebnisse:

```
Call:
lm(formula = log(SDH) ~ log(USD) + log(SY))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.88121    0.53367  -5.399 2.02e-05 ***
log(USD)     0.10462    0.07256   1.442    0.163
log(SY)      1.06611    0.03796  28.086  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04704 on 22 degrees of freedom
Multiple R-Squared: 0.9787,     Adjusted R-squared: 0.9767
F-statistic: 505.1 on 2 and 22 DF,  p-value: < 2.2e-16
```

Die Schätzung von Spezifikation (2) mit R liefert sodann folgenden Output:

```
Call:
lm(formula = log(SDH) ~ log(USD) + log(SY) + log(SP))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.99213    0.70841  -2.812   0.0104 *
log(USD)     0.05619    0.07416   0.758   0.4570
log(SY)      0.96941    0.06471  14.981 1.10e-12 ***
log(SP)      0.05731    0.03181   1.802   0.0859 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04481 on 21 degrees of freedom
Multiple R-Squared: 0.9815,     Adjusted R-squared: 0.9789
F-statistic: 372.2 on 3 and 21 DF,  p-value: < 2.2e-16
```

a) Welche der beiden Spezifikationen würden Sie bevorzugen? Begründen Sie kurz. (2 Punkte)

- Spezifikation (2) ist zu bevorzugen: Das korrigierte R^2 steigt (leicht) an ; SP ist statistisch signifikant auf dem 10%-Niveau.

b) Aus den Schätzergebnissen lassen sich weiterhin folgende Angaben ermitteln: (6 Punkte)

Spezifikation (1): $\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1} = 0.036$ und $\sum_{t=2}^{T} \hat{e}_{t-1}^2 = 0.048$;

Spezifikation (2): $\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1} = 0.029$ und $\sum_{t=2}^{T} \hat{e}_{t-1}^2 = 0.042$.

b1) Berechnen Sie die Autokorrelationskoeffizienten $\rho_{(1)}$ und $\rho_{(2)}$ für AR(1) Störtermprozesse e_1 und e_2.

- $\rho_{(1)} = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2} \approx \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2} = \frac{0.036}{0.048} = \mathbf{0.75}$;
- $\rho_{(2)} = \frac{0.029}{0.042} = \mathbf{0.69}$

b2) Berechnen Sie weiterhin die (approximativen) Durbin-Watson-Statistiken $d_{(1)}$ und $d_{(2)}$ und testen Sie auf dem 5%-Signifikanzniveau auf positive Autokorrelation erster Ordnung. Geben Sie hierzu auch die Nullhypothese, die Alternativhypothese sowie die Freiheitsgrade sowie die jeweiligen kritischen Werte an.

- $d_{(1)} \approx 2\left(1-\rho_{(1)}\right) = 2 - 2 \cdot 0.75 = \mathbf{0.5}$ *(exakt: 0.4273832)*
- $d_{(2)} \approx 2\left(1-\rho_{(2)}\right) = 2 - 2 \cdot 0.69 = \mathbf{0.62}$ *(exakt: 0.4864634)*
- Nullhypothese: $\rho_{(i)} \leq 0 \quad , \quad i = 1, 2$
- Alternativhypothese: $\rho_{(i)} > 0 \quad , \quad i = 1, 2$
- Kritische Werte 1 (T=25, K=3): $d_U = \mathbf{1.206}$; $d_O = \mathbf{1.550}$
- Kritische Werte 2 (T=25, K=4): $d_U = \mathbf{1.123}$; $d_O = \mathbf{1.654}$
- In beiden Fällen wird die Nullhypothese verworfen, es liegt also positive Autokorrelation erster Ordnung vor.

c) Modell-Spezifikation (2) wird erneut geschätzt unter Verwendung der verzögerten abhängigen Variablen (y_{t-1}) als zusätzliche erklärende Größe. Die Durbin-Watson Test-Statistik beträgt 1.85, ein Langrange Multiplier Test hat eine t-Test-Statistik von 0.1844. Liegt hier Autokorrelation erster Ordnung vor? Begründen Sie Ihre Antwort. Skizzieren Sie zudem kurz die Vorgehensweise des Lagrange Multiplier Tests. (6 Punkte)

- Der Durbin-Watson-Test darf nicht angewendet werden, wenn das Modell die verzögerte abhängige Variable als erklärende Größe enthält.
- Autokorrelation kann hier also nur anhand des Lagrange Multiplier Tests getestet werden. Bei LM=0.1844 erhält man (auf dem 5% Signifikanzniveau und N-K Freiheitsgraden) einen kritischen t-Wert von 1.721, die Nullhypothese, dass keine Autokorrelation vorliegt, kann nicht abgelehnt werden.

- Vorgehensweise: Ein Langrange Multiplier Test verwendet das geschätzte Residuum $\hat{e}_{t-1}$ als erklärende Größe des Modells. Mit einem t-Test (oder F-Test) kann sodann die Signifikanz des Parameters von $\hat{e}_{t-1}$ getestet werden. Ist dieser statistisch signifikant von Null verschieden, liegt Autokorrelation vor.

d) Zeigen Sie allgemein, dass e_t homoskedastisch ist, wenn ein AR(1) Störtermprozess der Form $e_t = \rho e_{t-1} + \upsilon_t$ gegeben ist, für den gilt: $E(e_t) = 0$; $E(\upsilon_t) = 0$; $Var(\upsilon_t) = \sigma_\upsilon^2$; $Cov(\upsilon_t, \upsilon_s) = 0$, für $t \neq s$. (4 Punkte)

- Homoskedastie liegt vor, wenn $Var(e_t) = \sigma^2$. Für den AR(1) Störtermprozess erhält man:

- $$Var(e_t) = Var(\rho e_{t-1} + \upsilon_t)$$
$$= Var(\rho e_{t-1}) + Var(\upsilon_t) + 2 \cdot \rho \cdot Cov(e_{t-1}\upsilon_t)$$

- $$\sigma_e^2 = \rho^2 \sigma_{e_{t-1}}^2 + \sigma_\upsilon^2 + 0$$

- Wenn $\sigma_{e_t}^2 = \sigma_{e_{t-1}}^2$
$$\sigma_e^2 = \frac{\sigma_{e_{t-1}}^2}{1-\rho^2}$$

- Da die Varianz σ_e^2 somit für alle Beobachtungen identisch ist, ist der Störterm homoskedastisch.