Diplomvorprüfung in Statistik II – Sommersemester 2005 – 22. Juli 2005

Aufgabe 1 **[46 Punkte]**

Ein Marktforschungsunternehmen wurde beauftragt, die Einflussfaktoren auf die tägliche Fernsehdauer zu untersuchen. Dazu wurden T=100 Personen zufällig ausgewählt und zu folgenden Aspekten befragt:

TV:	durchschnittliche Fernsehdauer pro Tag in Stunden
EK:	monatliches Nettoeinkommen in €
Alter:	Alter der befragten Person in Jahren
Abi:	Person hat Abitur (ja = 1, nein = 0)
G:	Geschlecht (weiblich = 1, männlich = 0)
Allein:	Person lebt alleine (ja = 1, nein = 0)

Die Marktforscher unterstellen folgendes Modell:

$$TV_t = \beta_1 + \beta_2 \cdot EK_t + \beta_3 \cdot Alter_t + \beta_4 \cdot Abi_t + \beta_5 \cdot G_t + \beta_6 \cdot Allein_t + e_t$$

Die Auswertung der Daten mit R ergab folgenden Output:

```
Call:
lm(formula = TV ~ EK + Alter + Abi + G + Allein)

Coefficients:
               Estimate      Std. Error    t value    Pr(>|t|)

(Intercept)    4.0020803     0.6649798     6.018      3.37e-08 ***
EK            -0.0010280     0.0002175       ?        8.01e-06 ***
Alter          0.0333995        ?          3.817      0.000242 ***
Abi           -0.2571470     0.2308090    -1.114      0.268072
G             -0.1293437     0.2011797    -0.643      0.521837
Allein        -0.5251771     0.2122891    -2.474      0.015160 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9752 on 94 degrees of freedom
Multiple R-Squared: ?,     Adjusted R-squared: 0.5526
F-statistic: 25.46 on 5 and 94 DF,  p-value: 3.768e-16
```

a) Berechnen Sie unter Angabe des Rechenwegs die im Output fehlenden Werte für (5)
   - a1) den t-Wert für b_2,
   - a2) den Standardfehler von b_3,
   - a3) die Quadratsumme der Fehlerterme (SSE),
   - a4) das Bestimmtheitsmaß!

b) Betrachten Sie die Koeffizienten b_2, b_3 und b_4. (4)
   - b1) Beurteilen Sie die Signifikanz dieser Koeffizienten!
   - b2) Um wie viele Minuten ändert sich die Fernsehdauer, wenn das Einkommen um 100,- € steigt?
   - b3) Wie unterscheidet sich die Fernsehdauer in Minuten für Personen mit und ohne Abitur?

c) Die Marktforscher vermuten, dass der Einkommenseffekt für Männer und Frauen gleich ist. Wie gehen Sie vor, um diese Hypothese zu testen? Unter welchen Umständen verwerfen Sie sie? Beschreiben Sie detailliert Ihre Vorgehensweise! (4)

d) Die Marktforscher wollen wissen, ob ihr Modell korrekt spezifiziert ist.

d1) Nennen Sie drei mögliche Arten von Fehlspezifikation! (3)

d2) Schreiben Sie das „künstliche“ Modell hin, das einem RESET-Test zu Grunde liegt, den man in R mit folgendem Befehl aufruft: *>reset(lm(TV ~ EK + Alter + Abi + G + Allein),2:3)*! (2)

d3) Ein RESET-Test in R liefert folgendes Ergebnis:

```
RESET test
data:  lm(TV ~ EK + Alter + Abi + G + Allein)
RESET = 0.1156, df1 = 1, df2 = 93, p-value = 0.7346
```

Interpretieren Sie dieses Ergebnis! (2)

e) Ein häufig auftretendes Problem bei der Schätzung von Regressionsmodellen ist Heteroskedastie.

e1) Erläutern Sie verbal, was man unter Heteroskedastie versteht! (2)

e2) Nennen Sie zwei Konsequenzen von Heteroskedastie für KQ-Schätzer! (2)

e3) Sie sollen nun für die Variable „Alter“ testen, ob Heteroskedastie vorliegt. Führen Sie den entsprechenden einseitigen Test (α=5%) mit Hilfe der folgenden ANOVA-Tabellen für die beiden Teilstichproben durch (Hinweis: Der Datensatz wurde nach dem Alter sortiert und anschließend in die 2 Gruppen aufgeteilt.)! Geben Sie dabei auch die Null- und Alternativhypothese, die Teststatistik, die Verteilung der Teststatistik und die Ablehnungsregion an! Interpretieren Sie das Ergebnis! (8)

Analysis of Variance Table
Response: TV[1:50]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
EK[1:50]	1	197899	19.7899	31.5932	1.214e-06 ***
Alter[1:50]	1	24.3355	24.3355	38.8499	1.533e-07 ***
Abi[1:50]	1	0.5054	0.5054	0.8068	0.373962
G[1:50]	1	0.8491	0.8491	1.3556	0.250576
Allein[1:50]	1	7.6950	7.6950	12.2846	0.001063 **
Residuals	44	27.5615	0.6264		

Analysis of Variance Table
Response: TV[51:100]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
EK[51:100]	1	17.527	17.527	14.3133	0.0004639 ***
Alter[51:100]	1	6.811	6.811	5.5624	0.0228572 *
Abi[51:100]	1	4.649	4.649	3.7965	0.0577545
G[51:100]	1	0.001	0.001	0.0008	0.9771001
Allein[51:100]	1	1.392	1.392	1.1369	0.2921332
Residuals	44	53.880	1.225		

e4) Aus früheren Studien ist bekannt, dass $\text{var}(e_t) = \sigma^2 \cdot \text{Alter}_t^2$. Formulieren Sie ein GLS-Modell, das das Heteroskedastie-Problem löst, und zeigen Sie, warum im transformierten Modell keine Heteroskedastie mehr vorliegt! (5)

f) Einer der Marktforscher vermutet, dass als wichtiges Kriterium für die Fernsehdauer auch berücksichtigt werden muss, ob jemand arbeitslos ist (Variable „al“, ja=1, nein=0). Dazu wird das obige Modell separat für Arbeitslose und Nicht-Arbeitslose geschätzt. Die jeweiligen ANOVA-Tabellen sind im Folgenden angegeben (Tabelle 1: nur Arbeitslose, Tabelle 2: nur Nicht-Arbeitslose). 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)

Tabelle 1: Analysis of Variance Table Response: TV[al==0]					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
EK[al==0]	1	15.992	15.992	15.4978	0.0003050 ***
Alter[al==0]	1	30.729	30.729	29.7801	2.387e-06 ***
Abi[al==0]	1	8.088	8.088	7.8381	0.0076964 **
G[al==0]	1	0.002	0.002	0.0024	0.9612339
Allein[al==0]	1	2.432	2.432	2.3565	0.1322614
Residuals	42	43.339	1.032		

Tabelle 2: Analysis of Variance Table Response: TV[al==1]					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
EK[al==1]	1	27.894	27.894	41.6959	5.957e-08 ***
Alter[al==1]	1	36.032	36.032	53.8592	2.834e-09 ***
Abi[al==1]	1	1.120	1.120	1.6737	0.202218
G[al==1]	1	0.241	0.241	0.3608	0.551005
Allein[al==1]	1	7.043	7.043	10.5282	0.002194 **
Residuals	46	30.774	0.669		

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.	Was macht R, wenn Sie den Befehl „*>abline(5,10)*“ eingeben?	
	❑	Es wird eine Linie erzeugt, die vom x-Wert 5 zum y-Wert 10 verläuft.
	❑	Es wird eine Linie mit der Steigung 5 und dem Achsenabschnitt 10 erzeugt.
	❑	Es wird eine Linie mit dem Achsenabschnitt 5 und der Steigung 10 erzeugt.

2.	Sie wollen sich die Häufigkeitsverteilung des Merkmals X anzeigen lassen. Welcher R-Befehl ist dazu nicht geeignet?	
	❑	*> table(X)*
	❑	*> hist(X)*
	❑	*> quant(X)*

3.	Mit welchem der folgenden Parameter der Funktion „*>plot()*“ ändern Sie die Skalierung der y-Achse?	
	❑	ylim
	❑	ylab
	❑	yscale

4.	Mit welchem der folgenden Befehle schätzen Sie ein Regressionsmodell ohne Konstante?	
	❑	*> lm(y ~ -const+x1+x2+x3)*
	❑	*> lm(y ~ x1+x2+x3,const=F)*
	❑	*> lm(y ~ 0+x1+x2+x3)*

5.	Sie wollen die Werte der folgenden Variablen in einem R-Objekt zusammenfassen: Vorname, Alter, Geschlecht und Einkommen. Welchen R-Objekttyp müssen Sie dafür wählen?	
	❑	Vektor
	❑	Dataframe
	❑	Matrix

6.	Welchen Wert berechnen Sie mit folgender Formel (y ist die abhängige Variable einer linearen Regression): *>sum((y – predict(y))^2)* ?	
	❑	SSE
	❑	SSR
	❑	SST

7.	Mit welchem R-Befehl bestimmen Sie den kritischen Wert einer t-Verteilung mit 15 Freiheitsgraden bei einem Signifikanzniveau von 10%?	
	❑	*> pt(0.1,15)*
	❑	*> qt(0.9,15)*
	❑	*> pt(0.9,15)*

8.	Bei welchem der folgenden Befehle bekommen Sie eine Fehlermeldung?	
	❑	*> pf(0.214,12,2)*
	❑	*> df(50,1,1)*
	❑	*> qf(49.99,1,1)*

9.	Sie wollen die Wahrscheinlichkeit dafür berechnen, dass eine standardnormalverteilte Zufallsvariable zwischen 0.8 und 1.5 liegt. Mit welchem R-Befehl erhalten Sie das richtige Ergebnis?	
	❑	*> dnorm(1.5,0,1) – dnorm(0.8,0,1)*
	❑	*> pnorm(1.5,0,1) – pnorm(0.8,0,1)*
	❑	*> pnorm(0.7,0,1)*

10.	Der t-Wert für den Koeffizienten b_2 beträgt 5.25 (Zahl der Freiheitsgrade = 15, Signifikanzniveau = 5%). Wie berechnen Sie mit R den entsprechenden p-Wert, der auch im R-Output erscheinen würde?	
	❑	*> 2*(1-pt(5.25,15))*
	❑	*> 1-(2*pt(5.25,15))*
	❑	*> 1-(2*qt(0.05,15))*

Aufgabe 3 **[6 Punkte]**

Wo sind die Fehler? Die folgende Funktion enthält 6 Fehler. Markieren Sie diese deutlich auf dem Aufgabenblatt und schreiben Sie die korrekte Formulierung darüber! 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 wurde oder
- der Fehler zwar richtig erkannt wurde, die Verbesserung aber falsch ist.

Die Gesamtpunktzahl kann nicht negativ werden.

Die folgende Funktion soll einen Vorhersagewert für y berechnen, wenn x_{1t} den Wert x_0 annimmt. Es gilt das folgende Regressionsmodell: $y_t = \beta_0 + \beta_1 \cdot x_{1t} + \beta_2 \cdot x_{1t}^2 + e_t$

Es werden zunächst die Koeffizienten geschätzt und anschließend der y-Wert berechnet, der sich ergibt, wenn man für x_{1t} einen beliebigen Wert x_0 einsetzt:

```
> Vorhersage == function(y,x1,x0)

  {

  reg = lm(y~x1+x1^2)

  neu = dataframe(x1=c(X0))

  Vorhersage = Predict(neu,reg)

  return(Vorhersage)

  }
```

Aufgabe 4 **[6 Punkte]**

Beantworten Sie die folgenden Fragen zu R. Für jede korrekte Antwort bekommen Sie 2 Punkte. Für falsche Antworten wird nichts abgezogen.

a) Mit welchem Befehl weisen Sie dem Vektor x die natürlichen Logarithmen der ganzen Zahlen 20 bis 120 zu? Welches Ergebnis würde im Anschluss daran der Befehl *>length(x)* liefern?

b) Der Vektor x enthält die ganzen Zahlen von 5 bis 8. Welches Ergebnis liefert der Befehl *> sum(x[1:3]*x[2:4])*?

c) Mit welchem Befehl erhalten Sie die Fläche unter der Dichtefunktion der Standardnormalverteilung im Bereich von $-\infty$ bis 0.8?

Aufgabe 5 **[21 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.

	Relative Konzentrationsmaße beschreiben, auf wie viele Merkmalsträger ein bestimmter Anteil der Merkmalssumme entfällt.
	Mit dem Lagrange-Multiplier Test kann getestet werden, ob der Störterm eines Modells durch einen autoregressiven Prozess zweiter Ordnung determiniert wird.
	Die Summe quadrierter χ^2-verteilter Zufallsvariablen ist F-verteilt, wenn diese Zufallsvariablen statistisch unabhängig voneinander sind.
	Zur Durchführung eines Instrumentvariablenschätzverfahrens benötigt man eine erklärende Variable, die mit keiner anderen erklärenden Variablen korreliert ist.
	Der Whiteschätzer wird auf Modelle mit transformierten Variablen angewendet.
	Der Koeffizient einer Dummyvariable gibt an, ob sich die Steigungsparameter eines Modells für Teilgruppen unterscheiden.
	Der Herfindahlindex wird berechnet als die Summe der quadrierten Anteile der Merkmalsträger an der Merkmalssumme.
	Multikollinearitätsprobleme lassen sich über eine Erhöhung der Beobachtungszahl reduzieren.
	Bei heteroskedastischen Störtermen bleiben die Kleinstquadrateschätzer der Steigungsparameter unverzerrt.
	Bei am Niveau α statistisch insignifikanten Koeffizienten liegt der Wert „Null“ innerhalb ihres $(1-\alpha)\cdot 100\%$-Konfidenzintervalls.
	Der Goldfeldt-Quandt Test kann sowohl als einseitiger als auch als zweiseitiger Test durchgeführt werden.
	Je höher der Ginikoeffizient ist, umso ungleicher ist die unterliegende Verteilung.
	Nach dem Method of Moments Schätzverfahren, lassen sich Bevölkerungsparameter durch analoge Parameter der Stichprobe schätzen.
	Es ist für die Qualität der Schätzergebnisse günstiger, zu viele als zu wenige erklärende Variablen im Modell zu berücksichtigen.
	Der Paascheindex ist kommensurabel.
	Ein Schätzer ist konsistent, wenn bei steigender Stichprobengrösse die Wahrscheinlichkeit gegen 1 konvergiert, dass der Schätzer in der Nähe des wahren Wertes liegt.
	Saisonkoeffizienten beschreiben die saisonale Abweichung eines Zeitreihenmittelwerts von $\bar{\bar{s}}$.
	Nicht alle Parameter des logistischen Trendmodells können im linearen Regressionsmodell geschätzt werden.
	Im Rahmen des Kleinstquadrateschätzers können ausschließlich lineare Beziehungen zwischen erklärenden und abhängigen Variablen geschätzt werden.
	Der Jarque-Bera Test nutzt eine χ^2-verteilte Teststatistik.
	Im einfachen linearen Modell führt eine Multiplikation der erklärenden Variable mit 100 zu einem um den Faktor 100 größeren Steigungsparameter.

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 welcher Wahrscheinlichkeit fällt eine mit 3 Freiheitsgraden t-verteilte Zufallsvariable in das Intervall [1.638 ; 3.182]?	
	❑	0.075
	❑	0.15
	❑	0.925

2.	Um das lineare Modell $y_t = \beta_0 + \beta_1 \cdot x_{1t} + \beta_2 \cdot x_{2t} + \beta_3 \cdot x_{3t} + e_t$ unter der Restriktion zu schätzen, dass die Summe der Steigungsparameter 5 beträgt, nutzt man folgendes Modell:	
	❑	$y_t - 5 \cdot x_{1t} = \beta_0 \cdot (1 - x_0) + \beta_2 \cdot (x_{2t} - x_{1t}) + \beta_3 \cdot (x_{3t} - x_{1t}) + e_t$
	❑	$y_t - 5 \cdot x_{1t} = \beta_0 + \beta_2 \cdot (x_{2t} - x_{1t}) + \beta_3 \cdot (x_{3t} - x_{1t}) + e_t$
	❑	$y_t - 5 \cdot x_{1t} - \beta_0 = \beta_2 \cdot (x_{2t} - x_{1t}) + \beta_3 \cdot (x_{3t} - x_{1t}) + e_t$

3.	Welche Interpretation gilt für β_1 im Modell für die Einfachregression $\ln y = \beta_0 + \beta_1 \cdot \ln x + e$?	
	❑	Marginaler Effekt: bei einer Änderung von x um eine Einheit ändert sich y um eine Einheit.
	❑	Semielastizität: bei einer Änderung von x um eine Einheit ändert sich y um ein Prozent.
	❑	Elastizität: bei einer Änderung von x um ein Prozent ändert sich y um ein Prozent.

4.	Im linearen Modell ist die Varianz des Prognosefehlers umso kleiner,	
	❑	je größer der Wert der erklärenden Variablen für den Prognosezeitpunkt.
	❑	je kleiner die Streuung der geschätzten Parameter.
	❑	je stärker der Zusammenhang zwischen den geschätzten Parametern.

5.	Das R^2 beschreibt	
	❑	den Anteil der auf Basis des Modells korrekt vorhersagbaren Werte der abhängigen Variable.
	❑	den Korrelationskoeffizienten zwischen der abhängigen und den erklärenden Variablen.
	❑	den mit dem Modell erklärten Anteil der Variation der abhängigen Variablen.

6.	Interaktionseffekte zwischen erklärenden Variablen	
	❑	sind nötig, wenn die Effekte qualitativer erklärender Variablen geschätzt werden sollen.
	❑	können die Schätzgüte eines Modells reduzieren.
	❑	bieten die Möglichkeit, für Teilstichproben unterschiedliche Steigungsparameter zu schätzen.

7.	Der Kleinstquadratschätzer	
	❑	sollte nur bei normalverteilten abhängigen Variablen verwendet werden.
	❑	sollte nur bei normalverteilten Residuen verwendet werden.
	❑	kann auch bei nicht normalverteilten Koeffizienten das Gauss-Markov-Theorem erfüllen.

8.	Die Normalengleichungen des KQ-Schätzers	
	❑	ergeben sich bei Minimierung der Zielfunktion.
	❑	sind über das Method of Moments Verfahren nicht herleitbar.
	❑	können nur im einfachen KQ-Modell bestimmt werden.

9.	Die Varianz von in erster Ordnung autokorrelierten Störtermen (AR(1))	
	❑	ist immer heteroskedastisch.
	❑	ist nur definiert für $\rho \neq 1$.
	❑	ist umso größer, je länger die von der Stichprobe beschriebene Zeitspanne ist.

10.	Der Durbin-Watson Test	
	❑	unterstellt eine AR(1)-Verteilung der Fehlerterme.
	❑	kann genutzt werden, wenn sich im Modell die verzögerte abhängige Variablen unter den erklärenden Variablen befindet.
	❑	ist nur für positive Autokorrelation anwendbar.

Aufgabe 7 **[21 Punkte]**

Sie untersuchen im Rahmen eines einfachen linearen Regressionsmodells den Zusammenhang zwischen der Anzahl verkaufter Personenwagen eines Monats (X_t) und dem monatlichen Börsenschlusswert (Y_t) eines Automobilherstellers. Ihnen liegen Daten zu Verkaufszahlen und Börsenkursen für 60 Monate vor. Ihr Modell lautet

$$Y_t = \beta_0 + \beta_1 \cdot X_t + e_t$$

a) Sie sind besorgt, ob man den Verkaufszahlen trauen kann und vermuten, dass die X-Variable fehlerhaft gemessen ist. Erläutern Sie die Konsequenz für Ihren Schätzer! (3)

b) Außerdem sind Sie nicht sicher, ob die Annahme $cov(e_t,e_s) = 0$ in Ihrem Fall zutrifft. Beschreiben Sie die einzelnen Schritte des LM-Tests zur Überprüfung dieser Hypothese! (4)

c) Was bedeutet es für die Eigenschaften Ihres Schätzers, wenn die Annahme $cov(e_t,e_s) = 0$ nicht zutrifft? (2)

d) Ihr Kollege bezweifelt, dass die Annahme $cov(x_t,e_s) = 0$ erfüllt ist und schlägt vor, dass Sie dies mit einem Hausman-Test überprüfen. Beschreiben Sie die Grundidee des Tests und erläutern Sie die notwendigen Schritte, um den Test durchzuführen! (7)

e) Angenommen, die Schätzung ergibt: $\hat{\beta}_0 = 3$ und $\hat{\beta}_1 = 2$ und es gilt ein AR(1) Modell mit $\hat{\rho} = 0.8$. Sie erwarten, dass die Verkaufszahlen im Monat t = 61 mit $X_{61} = 100$ PKW genauso hoch sind, wie im letzten Beobachtungsmonat ($X_{60} = 100$), zu dessen Ende der Börsenkurs $Y_{60} = 205$ betragen hat. Wie lautet Ihre Prognose des Börsenschlusskurses für die Periode t = 61? Stellen Sie Ihren Rechenweg dar und erläutern Sie die notwendigen Schritte knapp in verbaler Form! (5)