Aufgabe 1: ***(23 Punkte)***

Ihnen liegt ein Datensatz über Angestellte eines ländlichen, mittelständischen Unternehmens vor. Darin sind u.a. folgende Daten enthalten:
SICK → Krankheitstage; *AGE* → Alter in Jahren; *SEX* → weiblich = 1, männlich = 0; *EDUC* → Hochschulabschluss = 1, kein Hochschulabschluss = 0; *SPORT* → der/ die Befragte ist Mitglied in einem Sportverein = 1; ist nicht Mitglied = 0; *DIST* → Entfernung vom Wohnort zum nächsten Sportverein in Kilometern

Sie haben den Auftrag erhalten, die Anzahl der Krankheitstage pro Jahr im Unternehmen zu untersuchen und entscheiden sich für das folgende lineare Modell:

$$\text{SICK} = \beta_1 + \beta_2\,\text{AGE} + \beta_3\,\text{SEX} + \beta_4\,(\text{AGE*SEX}) + \beta_5\,\text{EDUC} + \beta_6\,\text{SPORT} + e$$

```
Call:
lm(formula = SICK ~ AGE + SEX + AGE*SEX + EDUC + SPORT)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3675 -0.3575 -0.1346  0.4579  2.2426

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58834    0.48408   1.215   0.2281
AGE          ..???..    0.00980   4.466 2.81e-05 ***
SEX          0.72924    0.51790   1.408   0.1633
AGE*SEX     -0.02246    0.01308  -1.717   0.0901 .
EDUC         0.01440    0.15804   0.091   0.9277
SPORT       -1.65159    0.20856  -7.919 1.84e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6929 on 74 degrees of freedom
Multiple R-Squared: ???,      Adjusted R-squared: 0.715
F-statistic: 40.63 on 5 and 74 DF,  p-value: < 2.2e-16
```

a) Berechnen Sie den fehlenden Wert für b_2 und das multiple Bestimmtheitsmaß (R^2). *(5 Punkte)*

- *b2=se(b2)*t = 0,04377*

  (0,6929^2)(80-6)=SSE=35,5282*
  0,715 = 1- SSE:74 / SST:79
  SSE:74 / SST:79 = 1 – 0,715 = 0,285
  *SSE/SST = 0,285*74/79 = 0,267*
  Multiple R-squared = 1 – 0,267 = 0,733

b) Interpretieren Sie die geschätzten Parameter (b_1 bis b_6) inhaltlich und statistisch. *(7,5 Punkte)*

- *b_1: auf 10% Signifikanzniveau nicht von 0 verschieden, die Anzahl der Krankheitstage entspricht b_1, wenn alle erklärenden Variablen den Wert 0 annehmen*

- *b_2 AGE: auf 1% Niveau signifikant von 0 verschieden, nur im Zusammenhang mit AGE * SEX interpretierbar: für Männer steigt die Anzahl der Krankheitstage im Durchschnitt um 0,04377 pro Lebensjahr*

- *b_3 SEX: auf 10% Signifikanzniveau nicht von 0 verschieden, nur im Zusammenhang mit AGE * SEX interpretierbar, Frauen haben im Alter von 0 Jahren im Durchschnitt 0,729 Krankheitstage mehr als Männer*

- *b_4 AGE*SEX: auf 10% Signifikanzniveau signifikant von 0 verschieden, nur im Zusammenhang mit SEX und AGE interpretierbar: die im Vergleich zu den Männern höhere Anzahl Krankheitstage der Frauen nimmt mit zunehmendem Alter pro Jahr um 0,02246 ab (oder: ab einem Alter von ca. 36,5 Jahren sind Frauen seltener krank als Männer)*

- *b_5 EDUC: bei Hochschulabschluss ist die Krankheitshäufigkeit um durchschnittlich 0,01440 Tage höher; dieser Zusammenhang ist nicht signifikant*

- *b_6 SPORT: bei Mitgliedschaft in einem Sportverein ist die Anzahl der Krankheitstage im Durchschnitt um 1,65 niedriger; dieser Zusammenhang ist auf dem 1% Niveau signifikant*

c) Welche Wahrscheinlichkeit gibt der p-Wert der Koeffizientenschätzer im Regressionsoutput an (rechte Spalte)? *(2,5 Punkte)*

- *der p-Wert gibt hier die Wahrscheinlichkeit dafür an, dass eine t-verteilte Zufallsvariable eine Ausprägung annimmt, deren Betrag größer ist als der Betrag des t-Wertes des geschätzten Regressionsparameters.*

d) Eine Kollegin zweifelt ihr Modell und Ihre Schätzergebnisse an. Sie argumentiert, dass von Hause aus gesunde Menschen häufiger Mitglied in einem Sportverein sind als kränkliche.

di) Worauf zielt der Einwand Ihrer Kollegin ab und was wäre die Konsequenz für Ihre KQ Schätzung, wenn sie Recht hat? Erläutern Sie. *(3 Punkte)*

- *SPORT könnte endogen sein, da vorwiegend gesunde Menschen Sport treiben und gesunde Menschen weniger Krankheitstage haben als kränkliche*
- *der KQ Schätzer von Sport ist inkonsistent*

dii) Ein anderer Kollege schlägt vor, eine Instrumentvariablenschätzung durchzuführen. Gegeben den vorliegenden Datensatz, welche Möglichkeit haben Sie, den Zusammenhang der Variable SPORT mit den Krankheitstagen verlässlich zu schätzen? Erläutern Sie kurz. Zählen Sie die Anforderungen auf, die allgemein an ein Instrument gestellt werden und diskutieren Sie, inwieweit diese Anforderungen im vorliegenden Fall erfüllt sind. *(5 Punkte)*

Hinweis: R liefert Ihnen folgende Korrelationskoeffizienten.

	SPORT	DIST
SPORT	1	-0,499659
DIST	-0,499659	1

- *DIST als Instrument verwenden*
- *soll unkorreliert mit dem Störterm und korreliert mit SPORT sein*
- *dass Entfernung vom Wohnort zum nächsten Sportverein mit dem Störterm unkorreliert ist, ist plausibel – gesunde Menschen wohnen nicht näher an Sportvereinen*
- *DIST ist mit SPORT negativ korreliert, d.h. zunehmende Entfernung zum Sportverein verringert die Wahrscheinlichkeit in diesen einzutreten.*
- *Es ist plausibel anzunehmen, dass die Anforderungen an ein Instrument im vorliegenden Fall erfüllt sind*

Aufgabe 2: ***(24 Punkte)***

Ihnen liegt das Ergebnis einer Untersuchung mit 50 Beobachtungen vor, in der das Jahreseinkommen (in 1000 €) als lineare Funktion von Geschlecht (SEX), Ausbildung in Jahren (EDUC), Alter (AGE) und Alter zum Quadrat (AGE2) geschätzt wurde. Das Modell lautet:

$EINK = \beta_1 + \beta_2\ SEX + \beta_3\ EDUC + \beta_4\ AGE + \beta_5\ AGE2 + e$

```
Call:
lm(formula = EINK ~ SEX + EDUC + AGE + AGE2)

Coefficients:
            Estimate Std. Error t value
Intercept  16.27950   15.13120   1.076
SEX         4.72105    1.22198   3.863
EDUC        0.03725    0.00960   3.880
AGE         0.96152    0.01308  73.511
AGE2       -0.00920    0.21615  -0.043
```

a) Berechnen Sie den marginalen Effekt des Alters auf das Jahreseinkommen und interpretieren sie ihn. Wann ist er Null? Wie können Sie die Signifikanz des Alterseffekts testen? Geben Sie die Nullhypothese und die Alternativhypothese an. *(5 Punkte)*

- *marg. Effekt:* $\partial EINK/\partial AGE = b_4 + 2{*}b_5\ AGE = 0{,}96152 - 0.0184\ AGE$
- $0 = 0{,}96152 - 0.0184\ AGE;\ AGE = 52{,}26$
- *der Effekt des Alters auf das Einkommen ist nicht linear, sondern nimmt mit zunehmendem Alter ab*
- *er ist bis zum Alter von ca. 52 Jahren positiv, danach negativ*

- *die Signifikanz des Alterseffekts kann durch einen F Test abgeschätzt werden*
- *die gemeinsame Nullhypothese lautet H0:* $\beta 4 = \beta 5 = 0$ *die Alternativhypothese besagt H1: mindestens ein Parameter ist ungleich 0*

b) Was ist der Effekt *(2 Punkte)*

bi) der Logarithmierung des Jahreseinkommens auf die Interpretation von β_3?

- *bei gleichen Parameterwerten ließe sich der Koeffizient wie folgt interpretieren: ein Anstieg des Regressors um eine Einheit erhöht das Jahreseinkommen um 100* β3 %, also um 3,72 %*

bii) auf die Ausprägung des Koeffizienten β_3 wenn das Jahreseinkommen in € statt in 1.000 € gemessen wurde?

- *der Koeffizient erhöht sich um den Faktor 1000, also auf 37,2*

c) Betrachten Sie nun den Ausbildungseffekt. *(9 Punkte)*

ci) Unterscheidet sich im vorliegenden Schätzmodell der Ausbildungseffekt für Männer und Frauen? Begründen Sie kurz.

- *nein, es werden lediglich Geschlecht und Ausbildung losgelöst voneinander untersucht*

cii) Wie könnte man prüfen, ob es einen statistisch signifikanten Geschlechterunterschied im Ausbildungseffekt gibt?

- *ein Interaktionsterm SEX*EDUC einfügen*
- *prüfen, ob Interaktionsterm signifikant von Null verschieden ist, t – Test*

ciii) Welche Hypothese testet in diesem Zusammenhang der Chow Test? Beschreiben Sie am obigen Beispiel die Vorgehensweise des Tests. Geben Sie die Teststatistik, die Freiheitsgrade, die Null- und Alternativhypothese sowie die Interpretation möglicher Testergebnisse an.

- *der Test prüft, ob sich die Parameter des Modells signifikant für beide Teilgruppen unterscheiden. Das Modell wird dazu vollständig mit dem Geschlechtsdummy interagiert, sodass folgende Gleichung geschätzt wird:*

$(EINK = \beta_1 + \beta_2\, SEX + \beta_3\, EDUC + \beta_4\, AGE + \beta_5\, AGE2 + \beta_6\, SEX{*}EDUC + \beta_7\, SEX{*}AGE + \beta_8\, SEX{*}AGE2 + e)$

- *die Nullhypothese ist also H0: $\beta 2 = \beta 6 = \beta 7 = \beta 8 = 0$*
  *β6 ist der Parameter von SEX*EDUC, β7 ist der Parameter von SEX*AGE und β8 ist der Parameter von SEX*AGE2*
- *die Alternativhypothese besagt, dass mindestens einer der Parameter β2, β6, β7, β8 einen signifikanten Einfluss auf das Einkommen hat*
- *F-Test , Statistik $F=\{(SSEr – SSEu)/J\}/\{SSEu/T\text{-}K\}$*
- *Freiheitsgrade, J=4, T-K = 50-5 = 45*
- *wird H0 abgelehnt, unterscheidet sich das Modell bzgl. der Steigungsparameter und der Regressionskonstante für die Geschlechter*
- *wird H0 nicht abgelehnt, sind die Modellparameter für Männer und Frauen nicht signifikant verschieden*

d) Wie würden Sie vorgehen, um die Hypothese zu testen, dass sich die Varianz des Störterms für Männer und Frauen unterscheidet? Beschreiben Sie Ihre Vorgehensweise, die Nullhypothese und die genaue Teststatistik. *(4 Punkte)*

- *Unterteilen der Stichprobe in je eine Teilstichprobe für jedes Geschlecht, Schätzen der Fehlertermvarianz für jede Teilstichprobe*
- *H0: σ_M^2 = σ_W ^2 d.h. gleiche Varianz in beiden Teilstichproben*
- *Teststatistik σ_M^2 / σ_W ^2, wobei die größere Varianz im Zähler steht*
- *F-Test mit (TM-K) und (TW-K) FG, bzw. umgekehrte Reihenfolge, wenn σW^2 / σM ^2*

e) Berechnen Sie das 95% Konfidenzintervall für den Effekt der Ausbildungsjahre und interpretieren Sie es. *(4 Punkte)*

- *P(b3-se(b3)*tc < β3 < b3+se(b3)*tc) = 0,95 oder b3±se(b3)*tc*
- *tc = 2,009*
- *untere Grenze: 0,03725-(0,00960*2,009)=0,01796*
- *obere Grenze: 0,05654*
- *Interpretation: "wird die Methode zur Schätzung des Konfidenzintervalls unendlich oft durchgeführt, so enthält das Konfidenzintervall in 95% aller Fälle den wahren Wert β3"*

Aufgabe 3: ***(22 Punkte)***

Da Sie sich mit dem Gedanken tragen, in Nürnberg ein Eiscafé zu eröffnen, interessieren Sie sich für Einflussfaktoren, die den örtlichen Eiskonsum bestimmen. Sie haben Daten für 30 Beobachtungen erhoben, wobei jede Beobachtung einem Zeitraum von 4 Wochen entspricht. Der letzte Beobachtungszeitraum ging gestern zu Ende.

Die abgefragten Variablen sind:

Q: Eiskonsum pro Kopf (in Litern)
P: Preis pro Liter Eis (in Euro)
E: Durchschnittliches wöchentliches Haushaltseinkommen (in Euro)
T: Durchschnittliche Temperatur (in Grad Celsius)

Die Regressionsgleichung lautet: $Q_t = \beta_1 + \beta_2 \cdot P_t + \beta_3 \cdot E_t + \beta_4 \cdot T_t + e_t$

Sie führen in R eine Kleinstquadrateschätzung (KQ) durch und erhalten folgenden Output:

```
Call:
lm(formula = Q ~ P + I + F)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065302 -0.011873  0.002737  0.015953  0.078986

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
P           -1.0444140  0.8343573  -1.252  0.22180
E            0.0033078  0.0011714   2.824  0.00899 **
T            0.0034584  0.0004455   7.762  3.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-Squared: 0.719,      Adjusted R-squared: 0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

a) Sie zeigen die Ergebnisse einem Kommilitonen, der sich die KQ-Residuen in einem Plot genauer ansieht und Sie auf das Problem möglicher Autokorrelation hinweist. Er kritisiert Ihre Ergebnisse und sagt, die KQ-Schätzer seien verzerrt und die Standardfehler zu groß. Nehmen Sie zu der Kritik Stellung. *(2 Punkte)*

Beide Aussagen sind falsch:

- *Die KQ-Schätzer bleiben unverzerrt, wenn die KQ-Residuen autokorreliert sind.*
- *Die Standardfehler werden im Fall von autokorrelierten KQ-Residuen falsch ausgewiesen. Sie müssen nicht zwangsläufig zu groß sein.*

b) Erläutern Sie allgemein das Problem Autokorrelation verbal. Geben Sie im Anschluss für den allgemeinen Fall eine formale Darstellung eines AR(1) Prozesses, in der Sie die einzelnen Formelelemente kurz beschreiben. *(4 Punkte)*

- *Autokorrelation tritt dann auf, wenn der Fehlerterm für unterschiedliche Beobachtungen eine Kovarianz ungleich Null aufweist. Insbesondere enger zeitlich benachbarte Residuen kovariieren stärker. Ein im Residuum einer Periode aufgefangener Einfluss oder Schock wirkt in der darauf folgenden Periode nach und ist im Residuum der Folgeperioden immer noch wirksam.*

- $e_t = \rho \cdot e_{t-1} + \upsilon_t$

  ρ *mißt die lineare Korrelation zwischen e_t, dem Residuum der Periode t und e_{t-1}, dem Residuum der Vorperiode.*

  υ_t *sind unkorrelierte Zufallsvariablen, welche die Anforderungen an die Residuen eines nach dem KQ-Prinzip geschätzten Modells erfüllen.*

c) Um zu überprüfen, ob Ihr Kommilitone mit seiner Autokorrelationsvermutung Recht hat, führen Sie in R einen Durbin-Watson Test auf positive Autokorrelation erster Ordnung auf dem 5% Signifikanzniveau durch. Sie erhalten folgenden R-Output: *(4 Punkte)*

```
        Durbin-Watson test
data:  kq
DW = 1.0212, p-value = 0.0003024
```

Geben Sie die getesteten Hypothesen, die kritischen Werte für den Durbin-Watson Test sowie die darauf basierende Testentscheidung an. Erläutern Sie, ob der p-Wert Ihre Testentscheidung unterstützt. Zu welcher Schlussfolgerung kämen Sie, wenn DW=1,307 wäre?

- $H_0 : \rho = 0$ *oder* $H_0 : \rho \leq 0$, $H_1 : \rho > 0$
- *Unterer kritischer Wert:* $d_{Lc} = 1{,}214$
- *Oberer kritischer Wert:* $d_{Uc} = 1{,}650$
- *Testentscheidung:* $DW < d_{Lc}$

  ⇨ *Die* $H_0 : \rho = 0$ *muß auf dem 5%-Signifikanzniveau verworfen werden.*
- $p < \alpha = 0{,}05$*: Die Interpretation des p-Werts liefert das gleiche Ergebnis.*
- *Für DW=1,307: Der Test liefert kein Ergebnis auf dem 5%-Signifikanzniveau*

d) Sie möchten Ihr Ergebnis aus c) überprüfen und führen in R einen Lagrange-Multiplier Test auf dem 5% Signifikanzniveau durch. Sie erhalten folgenden R-Output: *(4,5 Punkte)*

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2588680  0.2571651   1.007  0.32376
P           -1.1920552  0.7918621  -1.505  0.14476
E            0.0031920  0.0011085   2.879  0.00805
T            0.0032551  0.0004328   7.520 7.12e-08
e            0.4282815  0.2112149   2.028  0.05338
---

Residual standard error: 0.03481 on 25 degrees of freedom
Multiple R-Squared: 0.7587,     Adjusted R-squared: 0.7201
F-statistic: 19.65 on 4 and 25 DF,  p-value: 2.009e-07
```

Geben Sie die geschätzte Gleichung für das Beispiel der Aufgabe an und erläutern Sie allgemein die Testidee knapp und präzise. Wird das Ergebnis des ersten Durbin-Watson Tests (aus c)) bestätigt? Warum könnten sich die Testergebnisse unterscheiden?

- $Q_t = \beta_1 + \beta_2 \cdot P_t + \beta_3 \cdot E_t + \beta_4 \cdot T_t + \rho \cdot e_{t-1} + \upsilon_t$
- *Die Testidee beruht darauf, die geschätzten verzögerten Residuen* $\hat{e}_{t-1}$ *als erklärende Variablen in die Schätzgleichung mit einzubeziehen. Anhand eines t- oder F-Tests wird die Signifikanz des* ρ*-Parameters getestet.*

- $p = 0{,}0534 > \alpha = 0{,}05$ *: Da der p-Wert größer als das gewählte Signifikanzniveau ist, kann die* $H_0 : \rho = 0$ *nicht verworfen werden.*

- *Das Ergebnis des Durbin-Watson Tests wird nicht bestätigt. Dies kann daran liegen, daß die Stichprobe mit 30 Beobachtungen relativ klein ist, der Lagrange-Multiplier Test jedoch nur für große Stichproben exakt ist.*

e) Sie führen eine Generalised Least Squares (GLS-) Schätzung in R durch und erhalten folgende Schätzergebnisse: *(7,5 Punkte)*

```
Generalized least squares fit by REML
  Model: Q ~ P + I + F

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept)  0.1569894 0.2896017  0.542087  0.4725
P           -0.8922715 0.8108406 -1.100428  0.2215
E            0.0032041 0.0015456  2.073046  0.0090
T            0.0035585 0.0005545  6.417493  0.0000
```

Machen Sie auf Basis der Ergebnisse der GLS-Schätzung eine möglichst präzise Vorhersage für die Höhe des Nürnberger pro Kopf Eiskonsums in den kommenden vier Wochen unter Ausnutzung aller Informationen. Gehen Sie davon aus, dass

$Q_{30} = 0{,}763$

$P_{30} = 4{,}1 \qquad P_{31} = 3{,}8$

$E_{30} = 300 \qquad E_{31} = 310$

$T_{30} = 18{,}6 \qquad T_{31} = 19{,}2$

$\sum_{t=2}^{T} \hat{e}_{t-1} \cdot \hat{e}_t = 0{,}072$

$\sum_{t=2}^{T} \hat{e}_{t-1}^2 = 0{,}081$

und geben Sie Ihren Rechenweg an.

- $Q_{31} = \hat{\beta}_1 + \hat{\beta}_2 \cdot P_{31} + \hat{\beta}_3 \cdot E_{31} + \hat{\beta}_4 \cdot T_{31} + \hat{\rho} \cdot \tilde{e}_{30} + \upsilon_{31}$

- *Berechnung des Korrelationskoeffizienten:*

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{e}_{t-1} \cdot \hat{e}_t}{\sum_{t=2}^{T} \hat{e}_{t-1}^2} = \frac{0{,}072}{0{,}081} = 0{,}\overline{8}$$

- *Berechnung* $\tilde{e}_{30}$ *:*

$$\tilde{e}_{30} = Q_{30} - \hat{\beta}_1 - \hat{\beta}_2 \cdot P_{30} - \hat{\beta}_3 \cdot E_{30} - \hat{\beta}_4 \cdot T_{30} = 3{,}289$$

- *Berechnung des Vorhersagewerts:*
  $Q_{31} = 0{,}698$

Aufgabe 4: *(10 Punkte)*

In R wurde folgende Funktion programmiert:

```
> my.prog <- function(x,y)
   {
    s1 <- sum(x)/length(x)
    s2 <- sum(y[-5])/length(y[-5])
    plot(x,y)
    points(s1,s2)
    c <- c(s1,s2)
    return(c)
  }
```

Die Funktion soll auf zwei Vektoren `v` und `z` angewendet werden, welche die Ziffern von 1 bis 5 bzw. von 2 bis 6 enthalten.

a) Geben Sie einen möglichen R-Befehl an, mit dem man den Vektor `z` generieren kann.

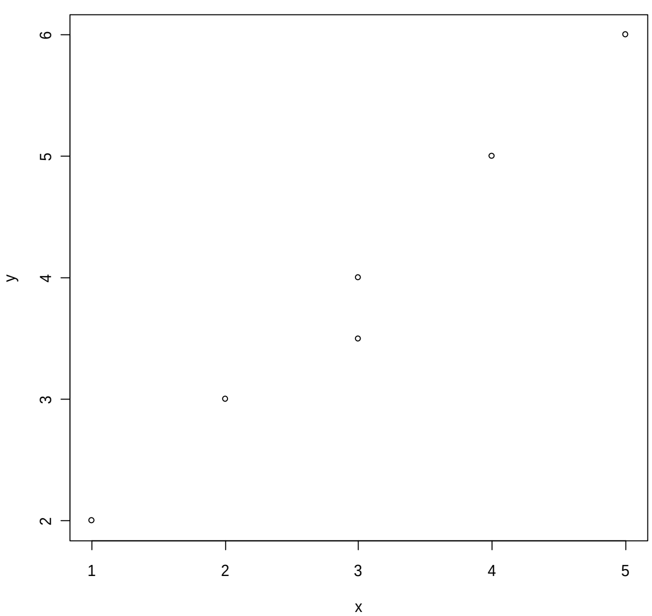
Es gibt mehrere Möglichkeiten (statt `<-` kann zudem auch = verwendet werden):

```
> z <- c(2,3,4,5,6)
> z <- c(2:6)
> z <- seq(2:6)
```

b) Mit welchem R-Befehl rufen Sie die Funktion für die beiden Vektoren auf?

```
my.prog(v,z)
```

c) Stellen Sie alle Ausgaben so dar, wie sie mit dieser Funktion für diese beiden Vektoren erzeugt werden.




```
[1,]  3 3.5
```

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.	Mit welchem R-Befehl erzeugen Sie einen Vektor *x*, der ungerade Zahlen zwischen 1 und 99 enthält?	
	☒	`> x <- seq(1,99, by=2)`
	❑	`> x <- seq(1,99, x[2]-1)`
	❑	`> x <- seq(1,99, odd=T)`

2.	Welchen Grafik-Bestandteil erzeugen Sie mit dem R-Befehl `> abline(h=0, lty=2)`?	
	❑	Abszisse in Fettdruck
	☒	Horizontale gestrichelte Linie mit Ordinatenabschnitt Null
	❑	Regressionsgerade eines vorab geschätzten KQ-Modells

3.	Mit welchem R-Befehl können Sie die Standardfehler eines als Objekt `mod.kq` vorliegenden Modelloutputs auslesen?	
	❑	`> mod.kq$coef[,1]`
	☒	`> mod.kq$coef[,2]`
	❑	`> mod.kq$coef[,3]`
4.	Bei welchem der folgenden R-Befehle erhalten Sie keine Fehlermeldung?	
	❑	`> pf(.95;12,2)`
	❑	`> pf(.95,12;2)`
	☒	`> pf(.95,12,2)`

5.	Welchen R-Befehl müssen Sie verwenden, um ein KQ-Modell zu schätzen, welches eine Variable enthält, die die Interaktion zwischen *x* und *z* abbildet?	
	❑	`> lm(y ~ x + z + Int{x*z})`
	❑	`> lm(y ~ x + z + IA[x^z])`
	☒	`> lm(y ~ x + z + I(x*z))`

6.	Mit welcher Option des R-Befehls `read.table` kann man einzulesenden Variablen Namen zuweisen?	
	☒	`> col.names=c("var1","var2","var3")`
	❑	`> var.names=c("var1","var2","var3")`
	❑	`> row.names=c("var1","var2","var3")`

7.	Welchen R-Befehl müssen Sie verwenden, um ein KQ-Modell mit der ersten Hälfte eines Datensatzes mit 50 Beobachtungen durchzuführen?	
	❑	`> lm(y ~ x, data.frame[1:25])`
	❑	`> lm(y ~ x, lower.half=T)`
	☒	`> lm(y[1:25] ~ x[1:25])`

8.	Welchen R-Befehl kann man verwenden, um einen Chow-Test durchzuführen?	
	❑	`> chow.test(mod1.kq, mod2.kq)`
	☒	`> anova(mod1.kq, mod2.kq)`
	❑	`> F.test(mod1.kq, mod2.kq)`

9.	Welche Kennzahl berechnet man mit dem Befehl `sum.mod$sigma^2` (`sum.mod` sei der Modelloutput)?	
	❑	Autokorrelationskoeffizient eines AR(1)-Fehlers
	❑	quadrierter Vorhersagewert bei Heteroskedastie
	☒	geschätzte Fehlertermvarianz des KQ-Modells

10.	Der *t*-Wert eines geschätzten Koeffizienten sei 0.67 (bei 48 Freiheitsgraden und einem Signifikanzniveau von 5%). Mit welchem R-Befehl kann man **nicht** den im Output ausgewiesenen *p*-Wert manuell berechnen?	
	❑	`> 2*(1-pt(0.67,df=48))`
	❑	`> 2*pt(0.67,df=48,lower.tail=F)`
	☒	`> 1-(2*pt(0.67,df=48,lower.tail=T))`

Aufgabe 6: ***(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.

W	Eine mögliche Interpretation des F-Tests ist, dass er den Erklärungsgehalt unterschiedlicher Modelle vergleicht.
F	Um einen Chow Test durchzuführen, sind zwei Schätzungen erforderlich.
F	Beim einseitigen t-Test liegt die Ablehnungsregion im Bereich positiver t-Werte.
W	Zur Bestimmung des Kleinstquadrateschätzers ist die Annahme normalverteilter Fehlerterme nicht erforderlich.
W	Die Annahme, dass der Störterm im multivariaten Regressionsmodell normalverteilt ist, ist für Erwartungstreue und Varianz der Kleinstquadrate-Schätzer unerheblich.
W	Im multiplen Regressionsmodell steigt die Varianz eines geschätzten Steigungsparameters β_k, wenn die entsprechende erklärende Variable x_k stark mit anderen erklärenden Variablen im Modell korreliert ist.
W	Es ist möglich, den Parameter ρ bei autokorrelierten Störtermen erster Ordnung als Steigungsparameter in einer KQ Schätzung zu schätzen.
F	Der Goldfeld-Quandt Test ist nur für Situationen mit proportionaler Heteroskedastie geeignet.
W	Der Herfindahl-Index ist ein absolutes Konzentrationsmaß.
W	Die Varianz des Vorhersagefehlers im einfachen Regressionsmodell ist am Mittelwert der erklärenden Variablen am geringsten.

W	Ein Typ I Fehler wird wahrscheinlicher, wenn α steigt.
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	Konsistente Schätzer können verzerrt sein.
W	Die gemeinsame Dichtefunktion f(X,Y) zweier unabhängiger Zufallsvariablen X und Y unterscheidet sich von der gemeinsamen Dichtefunktion zweier korrelierter Zufallsvariablen.
F	Der White Schätzer korrigiert das Problem stochastischer Fehlertermvarianzen.
W	Der LM Test auf Autokorrelation erster Ordnung in den Störtermen besteht aus einem Signifikanztest für den geschätzten Koeffizienten des um eine Periode verzögerten Fehlerterms, der zusätzlich ins ursprüngliche Modell aufgenommen wird.
F	Unter Heteroskedastie können bessere Vorhersagen gemacht werden, als ohne Heteroskedastie.
W	Kategoriale erklärende Variablen werden typischerweise mit Bezug auf eine Referenzgruppe interpretiert.
W	Multikollinearitätsprobleme lassen sich über eine Erhöhung der Beobachtungszahl reduzieren.
F	Um eine saisonbereinigte Zeitreihe zu erstellen, werden lineare, exponentielle oder logistische Saisonmodelle genutzt.
F	Ein hoher Gini-Koeffizient lässt auf eine gleichmäßige Verteilung schließen.

Aufgabe 7: ***(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.	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.
2.	Der Typ II Fehler	
	☐	tritt auf, wenn die Nullhypothese verworfen wird, obwohl sie zutrifft.
	☐	ist umso wahrscheinlicher, je größer die Stichprobe ist.
	☒	wird unwahrscheinlicher, wenn der Typ I Fehler wahrscheinlicher wird.
3.	Der Two Stage Least Squares Schätzer	
	☐	schätzt das gleiche lineare Regressionsmodell zweimal
	☒	nutzt vorhergesagte Werte auf der zweiten Stufe.
	☐	berücksichtigt ein Polynom zweiter Ordnung der erklärenden Variable.

4.	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.
5.	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.
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.	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.
8.	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.
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.	Punktschätzer sind	
	❑	informativer als Intervallschätzer.
	❑	nicht auf Basis von Stichproben interpretierbar.
	☒	umso verlässlicher, je kleiner die geschätzte Fehlervarianz $\hat{\sigma}^2$.