UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

Statistik I
Prof. Dr. Rainer Dahlhaus
Wintersemester 2020/2021

2. Übungsblatt

Aufgabe 5	Aufgabe 6	Aufgabe 7	Aufgabe 8	Summe:

Übungsgruppe: Tutor(in):

Namen:

Aufgabe 5 (F-Test im Regressionsmodell, 4 = 2 + 2 Punkte).

In seinem Modell geht der Biologe K. Fischer davon aus, dass zwischen der Größe G (in Meter) und dem Alter A (in Jahren) von jungen Karpfen ein quadratischer Zusammenhang besteht,

$$G = c_1 \cdot A + c_2 \cdot A^2$$

für Parameter $c_1, c_2 \in \mathbb{R}$. Er hat den Verdacht, dass sogar $c_2 = 0$ gelten sollte. Dazu fängt er $n = 10$ junge Karpfen und bestimmt deren Alter und Größe:

A_i (in Jahren)	1.1	2.8	3.7	4.4	5.3	5.8	6.9	7.5	8.3	9.5
G_i (in Meter)	0.30	0.32	0.67	0.56	0.67	0.96	0.75	1.06	0.9	1.03

Überprüfen Sie den Verdacht von K. Fischer zum Niveau $\alpha = 0.05$, indem Sie wie folgt vorgehen:

(a) Bestimmen Sie den Kleinste-Quadrate-Schätzer für $\beta := (c_1, c_2)'$ im Modell

$$G_i = c_1 \cdot A_i + c_2 \cdot A_i^2 + \varepsilon_i, \quad i = 1, ..., n,$$

wobei ε_i unabhängige und normalverteilte Fehler mit $\mathbb{E}\varepsilon_i = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ sind. Bestimmen Sie nun den KQ-Schätzer für $\tilde{\beta} := c_1$ im reduzierten Modell mit $c_2 = 0$.

(b) Geben Sie die Formel für die F-Statistik in dem Test an, so dass alle auftretenden Optimierungsprobleme gelöst sind.
Bemerkung: Sie müssen die Terme aber nicht weiter vereinfachen.
Berechnen Sie anhand der gegebenen Daten weiter die Werte der KQ-Schätzer und entsprechend den Wert der F-Statistik. Werten Sie dann den zugehörigen F-Test aus.
Hinweis: Einige α-Quantile $F_{\ell,n-r,\alpha}$-Verteilung lauten
$F_{1,8,0.05} = 0.0,\ F_{1,8,0.95} = 5.32,\ F_{2,9,0.05} = 0.05,\ F_{2,9,0.95} = 4.26.$

Aufgabe 6 (Varianzanalyse mit einem Faktor, 4 = 0.5 + 2 + 0.5 + 1 Punkte).
Gegeben seien unabhängige Zufallsvariablen

$$\begin{aligned} Y_{11}, Y_{12}, ..., Y_{1r} &\sim \mathcal{N}(\mu_1, \sigma^2) \\ Y_{21}, Y_{22}, ..., Y_{2s} &\sim \mathcal{N}(\mu_2, \sigma^2), \end{aligned}$$

wobei $r, s > 0$.

(a) Formulieren Sie damit das lineare Modell $Y = X\beta + \varepsilon$, indem Sie die Objekte X, Y und ε definieren.

(b) Wir wollen in diesem Modell untersuchen, ob $\mu_1 = \mu_2$ gilt. Formulieren Sie das entsprechende Testproblem und finden Sie einen geschlossenen, so weit wie möglich vereinfachten Ausdruck für die entsprechende F-Statistik.

(c) Geben Sie die Verteilung der F-Statistik aus (b) an.

(d) Formulieren Sie kurz eine praktische Situation, bei welchem der F-Test aus (b) benutzt werden könnte.

Aufgabe 7 (Testen auf Normalverteilung, 4 = 1.5 + 2.5 Punkte).
Gegeben sei ein einfaches lineares Modell

$$Y_i = \mu + \varepsilon_i, \quad i = 1, ..., n,$$

mit Parameter $\mu \in \mathbb{R}$ und iid Zufallsvariablen ε_i, wobei $\mathbb{E}\varepsilon_1 = 0$ und $\mathrm{Var}(\varepsilon_1) = \sigma^2 > 0$. In der Vorlesung wurde die Theorie des F-Tests hergeleitet für den Fall, dass ε_i normalverteilt ist. Mit dem Jarque-Bera-Test kann vorher überprüft werden, ob diese Annahme gerechtfertigt ist. Dieser Test lehnt die Nullhypothese $H_0 : \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ zum Niveau α ab, falls

$$J_n := n \cdot \left(\frac{1}{6} \hat{W}^2 + \frac{1}{24} (\hat{K} - 3)^2 \right) > \chi^2_{2,1-\alpha},$$

wobei $\chi^2_{2,1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2_2-Verteilung bezeichne, und

$$\hat{W} := \frac{1}{\hat{\sigma}^3} \frac{1}{n} \sum_{i=1}^n (Y_i - \overline{Y}_n)^3, \qquad \hat{K} := \frac{1}{\hat{\sigma}^4} \frac{1}{n} \sum_{i=1}^n (Y_i - \overline{Y}_n)^4,$$

die normierten dritten und vierten Momente $S := \frac{E[(Y_1-\mu)^3]}{\sigma^3}$, $K := \frac{\mathbb{E}[(Y_1-\mu)^4]}{\sigma^4}$ schätzen. Ferner ist $\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ und $\overline{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$.
Im Folgenden soll verifiziert werden, dass der Jarque-Bera-Test asymptotisch ein Test zum Niveau α ist.

(a) Zeigen Sie unter der Annahme $\mathbb{E}[\varepsilon_1^4] < \infty$,

$$n^{-1/2} \sum_{i=1}^n (Y_i - \overline{Y}_n)^3 = n^{-1/2} \sum_{i=1}^n (\varepsilon_i^3 - 3\sigma^2 \varepsilon_i) + R_n^1,$$

wobei $R_n^1 \xrightarrow{p} 0$. *Es muss hierbei nicht notwendig H_0 gelten.*
Hinweis: Nach dem Zentralen Grenzwertsatz gilt $n^{1/2}\overline{\varepsilon_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ und eine ähnliche Aussage für den Mittelwert $\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$. Rekapitulieren Sie dabei die Zusammenhänge von schwacher und stochastischer Konvergenz von Zufallsvariablen.

(b) Nehmen Sie nun H_0 an. Benutzen Sie im Folgenden ohne Beweis, dass

$$n^{-1/2} \sum_{i=1}^{n} \left\{ (Y_i - \overline{Y}_n)^4 - 3\hat{\sigma}^4 \right\} = n^{-1/2} \sum_{i=1}^{n} (\varepsilon_i^4 - 6\sigma^2 \varepsilon_i^2 + 3\sigma^4) + R_n^2,$$

wobei $R_n^2 \xrightarrow{p} 0$. Zeigen Sie mit dem multivariaten zentralen Grenzwertsatz und dem Continuous-Mapping-Theorem, dass $J_n \xrightarrow{\mathcal{D}} \chi_2^2$.
Hinweis: Nutzen Sie ohne Beweis, dass für $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[\varepsilon_1^k] = 0$, falls k ungerade ist, und $\mathbb{E}[\varepsilon_1^k] = (k-1) \cdot (k-3) \cdot ... \cdot 3 \cdot 1 \cdot \sigma^k$, falls k gerade ist, gilt. Das Continuous-Mapping-Theorem begsagt: Gilt $X_n \xrightarrow{\mathcal{D}} Z$ für Zufallsvektoren X_n, $Z : \Omega \to \mathbb{R}^k$, und ist $g : \mathbb{R}^k \to \mathbb{R}^l$ stetig, so folgt $g(X_n) \xrightarrow{\mathcal{D}} g(Z)$.

Aufgabe 8 (Multiples Testen mit der Bonferroni-Methode, 4 = 2 + 2 Punkte).
In dieser Aufgabe betrachten wir ein allgemeines lineares Modell der Form

$$Y = X\beta + \varepsilon,$$

wobei $X \in \mathbb{R}^{n \times k}$ deterministisch und bekannt ist, $\beta = (\beta_1, ..., \beta_k)' \in \mathbb{R}^k$ und $\varepsilon = (\varepsilon_1 ..., \varepsilon_n)'$ mit iid $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. In Fragestellungen, bei welchen nicht klar ist, welche Regressoren Auswirkungen auf das Ergebnis Y haben, werden häufig sehr viele Regressoren (d.h. Spalten in X) und entsprechend viele Parameter gewählt und dann eine lineare Regression durchgeführt. Es tritt die Frage nach der Signifikanz der Parameter auf. Ein Parameter β_i heißt signifikant, falls $\beta_i \neq 0$; er also über X einen Einfluss auf das Ergebnis Y ausübt.

(a) Sei $\hat{\beta}$ der KQ-Schätzer für β. Zeigen Sie, dass für ein festes $j \in \{1, ..., k\}$,

$$\phi_j(Y) = \begin{cases} 1, & T_n := \frac{1}{\hat{\sigma}} \cdot \frac{\hat{\beta}_j}{\sqrt{((X'X)^{-1})_{jj}}} > t_{n-k, 1-\alpha_j} \\ 0, & T_n \leq t_{n-k, 1-\alpha_j} \end{cases}$$

ein Test für die Hypothesen $H_0 : \beta_j = 0$ gegen $H_1 : \beta_j \neq 0$ zum Niveau α_j ist. Hierbei ist $\hat{\sigma}^2 := \frac{1}{n-k}(Y - X\hat{\beta})'(Y - X\hat{\beta})$.
Hinweis: Nutzen Sie Aufgabe 4 auf dem 1. Übungsblatt.

(b) Angenommen, Sie führen den Test ϕ_j aus (a) nun für jedes $j = 1, ..., k$ durch. Zeigen Sie, dass die Wahl $\alpha_1 = \alpha_2 = ... = \alpha_k = \frac{\alpha}{k}$ dafür sorgt, dass der Fehler 1. Art (also einer der Parameter $\beta_1, ..., \beta_k$ wird als signifikant erkannt, obwohl er es nicht ist) weiterhin durch ein vorgegebenes α beschränkt ist.
Hinweis: Sie dafür annehmen, dass die Nullhypothese nun $H_0' : \beta_1 = ... = \beta_k = 0$ lautet und der Test durch

$$\phi(Y) = \begin{cases} 1, & \phi_j = 1 \text{ für ein } j \in \{1, ..., k\}, \\ 0, & \phi_j = 0 \text{ für alle } j \in \{1 ..., k\} \end{cases}$$

gegeben ist. Nutzen Sie die Bonferroni-Ungleichung $\mathbb{P}(\bigcup_{i=1}^{s} A_i) \leq \sum_{i=1}^{s} \mathbb{P}(A_i)$ für Ereignisse $A_1, ..., A_s$, um den Fehler 1. Art von ϕ abzuschätzen.

Abgabe:
In Zweiergruppen, bis spätestens Donnerstag, den **19.11.2020, 11:00 Uhr**.
Bitte geben Sie Ihre Lösungsvorschläge im PDF-Format ab. Es reicht, wenn einer der beiden Personen das Dokument hochlädt.

Homepage der Vorlesung:
`https://ssp.math.uni-heidelberg.de/stat-ws2020/index.html`