2. Übungsblatt *506.715 Angewandte Statistik, WS 2009/2010*

Univ.-Prof. DI Dr. Ernst Stadlober

1.) [T] Freiheitsgrad der Statistik T_W des Welch–Tests.

Die Teststatistik

$$T_W = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \overset{as}{\sim} t_\nu \quad \text{falls} \quad \mu_D = 0\,,$$

des Welch–Tests hat den Freiheitsgrad $\nu = \dfrac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2}$

Man zeige, dass

$$\nu \leqslant n + m - 2\,.$$

2.) Simulation von Stichproben, Transformation zur Normalverteilung; `[R 2.9]`.

(a) Erzeugen Sie jeweils $n = 64$ bzw. $n = 128$ Stichproben aus der Gamma-Verteilung mit den Shape-Parametern $a = 2, 5, 10$ und der Rate $\lambda = 1$ ($f(x) = x^{a-1}e^{-x}$, $x > 0$), sowie aus der Standard-Normal $N(0,1)$–Verteilung (Aufruf für n Gamma-verteilte Zufallszahlen: `rgamma(n,shape=a, rate=1)`). Speichern Sie diese 4 Stichprobenvektoren auf die Datenfiles `simgam64` und `simgam128` ab. Stellen Sie jedes Merkmal mittels *Boxplot*, *Steam-Leaf-Display* und *Histogramm* dar. Berechnen Sie statistische Kenngrößen, führen Sie Tests auf Normalverteilung durch und stellen Sie die Situation durch Q-Q-Plots mit der $N(0,1)$-Verteilung als Referenz dar.

(b) Erstellen Sie für die Gamma–verteilten Stichproben die Q-Q-Plots bzgl. der Gamma(a,1)–Verteilungen mit dem Befehl `qqmath()`.

(c) Transformieren Sie die Stichproben $(x_1, \ldots, x_n)$ aus der Gamma-Verteilung nach der

i. Fisher–Transformation $y_i = \sqrt{4x_i} - \sqrt{4a-1}$,

ii. Wilson–Hilferty–Transformation $w_i = \left(\left(\frac{x_i}{a}\right)^{1/3} - \mu\right)/\sigma$ mit $\mu = 1 - \frac{1}{9a}$, $\sigma = \sqrt{\frac{1}{9a}}$

zu annähernd $N(0,1)$–verteilten Stichproben und erweitern Sie die Datenfiles um diese vier transformierten Vektoren. Analysieren Sie die Verteilung der transformierten Merkmale wie in (a).

(d) Fassen Sie Ihre Ergebnisse und Interpretationen in Form eines `pdf`–Dokuments (max. 4 Seiten) zusammen.

3.) [T] Einfache lineare Regression.

Sei $Y_i \overset{i}{\sim} N(\mu_i, \sigma)$, $i = 1, \ldots, n$, mit

$$\mu_i = \beta_1 + \beta_2 x_i = \beta_1 + \beta_2 \bar{x} + \beta_2(x_i - \bar{x}) = \alpha + \beta_2 t_i.$$

Man löse **zwei** der folgenden Aufgaben.

(a) Man berechne explizit die Hat–Matrix $H = (h_{ij}) = X(X^TX)^{-1}X^T$.

(b) Man zeige, dass folgendes gilt:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma}{\sqrt{n}}\right), \quad \hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma}{\sqrt{S_t}}\right)$$

und

$$\hat{\alpha} \quad \text{und} \quad \hat{\beta}_2 \quad \text{sind unabhängige Zufallsvariable}.$$

Für $\hat{\beta}_1 = \hat{\alpha} - \hat{\beta}_2\bar{x}$ gilt

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_t}}\right) \quad \text{und} \quad \rho(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{x}}{\sqrt{\frac{S_t}{n} + \bar{x}^2}}.$$

(c) Sei $R_i = Y_i - \hat{\mu}_i = Y_i - \hat{\alpha} - \hat{\beta}_2 t_i$ das i–te Residuum. Man zeige, dass

$$\begin{aligned} E(R_i) &= 0, \quad Cov(R_i, \hat{\alpha}) = Cov(R_i, \hat{\beta}_2) = 0, \\ R_i &\sim N\left(0, \sigma\sqrt{1 - \frac{1}{n} - \frac{t_i^2}{S_t}}\right) = N\left(0, \sigma\sqrt{1 - h_{ii}}\right), \\ \rho(R_i, R_j) &= -\frac{h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}. \end{aligned}$$

4.) Lineare Regressionsanalyse der Baum–Daten `baum.txt`; [R 2.9].

Die Datei `baum.txt` enthält 3 Messungen an $n = 31$ Kirschbäumen aus dem *Allegheny National Forest, Pennsylvania*. Die erste Spalte gibt den Durchmesser `d` in *Inches =(0.0254 Meter)*, gemessen in einer Höhe von 1.37 Meter, die zweite die Höhe `h` in *Feet (=0.3048 Meter)* und die dritte das Volumen `v` in *cubic feet* an. Auf Grund der Messung von Höhe und Durchmesser möchte man das Volumen eines Baumes vorhersagen.

(a) Vergeben Sie Labels und rechnen Sie die Einheiten in Meter (Kubikmeter) um. Analysieren Sie die Daten mit geeigneten graphischen Verfahren.

(b) Stellen Sie ein (lineares) Regressionsmodell für `v` in Abhängigkeit von `d` und `h` auf. Erstellen Sie Residuenplots und beurteilen Sie die Resultate.

(c) Der geometrische Zusammenhang zwischen den Variablen `d`, `h` und `v` ist durch

$$v = \frac{\pi}{12} d^2 h \tag{1}$$

gegeben (unter der Annahme der Baum habe eine konische Form). Welches lineare Modell wäre geeignet diesen Zusammenhang zu beschreiben?

Hinweis: Man logarithmiere die Gleichung (1).

5.) [T] Lineare und quadratische Formen von normalverteilten Größen

Sei $Y_i \overset{iid}{\sim} N(\mu, \sigma)$, $i = 1, \ldots, n$.

Zeigen Sie

$$\sqrt{n}(\overline{Y} - \mu) \quad \text{und} \quad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad \text{sind unabhängige Zufallsvariable}$$

und

$$\sqrt{n}(\overline{Y} - \mu) \sim N(0, \sigma)\,, \quad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2_{n-1}\,.$$

Man benutze dazu **Satz 3.3.1** und **Satz 3.3.2** aus dem Skriptum.

6.) Fallbeispiel Luftschadstoffdaten (2. Teil) `grazluft`; [R 2.9].

(a) Erstellen Sie ein Regressionsmodell für `pm10` in Abhängigkeit von `no`, `no2` und dem Faktor `periode`.

(b) Analysieren Sie die standardisierten Residuen mittels Histogramm, Q–Q-Plot und Scatterplot `stdres` gegen `vorhersage`. Erstellen Sie eine Graphik (4 Plots) zur Beurteilung der Residuen mit dem Befehl `plot(lm(pm10~no+no2+periode))`.

(c) Welches Bestimmtheitsmaß r^2_{adj} und welche Streuung $\tilde{\sigma}$ erreicht man für das Modell? Wo tritt das größte negative (positive) Residuum auf? Gibt es Ausreißer? Ist die Periode von Bedeutung?

Herunter laden der Daten über die HomePage des Instituts: `www.statistics.tugraz.at`
Speichern Sie die **gesamten Übungen in einem pdf-File** mit folgendem Namen ab:
`Angstat_Nachname1*` z.B. `Angstat_Schiefer2.pdf`
und übermitteln Sie **einen File pro Gruppe** mit *Subject:* **Angstat** an die e-mail-Adresse `statistik@tugraz.at`.

Transfer der Files bis spätestens: Fr. 20. 11. 2009, 16.00 Uhr

Besprechungstermin: Mo. 23. 11. 2009, 9.00–10.45, SR Statistik