Prof. Dr. U. Stadtmüller
C. Hering

Sommersemester 2010
02.08.2010

Prüfungsklausur Angewandte Statistik für Biometrie

Bearbeitungszeit: 90 Minuten,
Hilfsmittel: Taschenrechner, ein beidseitig beschriebenes Din A4-Blatt,
erreichbare Punkte: 103, 100 Punkte entsprechen 100%

1. Gegeben seien die folgenden Modelle:

   (a) $X_t = \mu + \beta_0 \cos y_t + \beta_1 \sin y_t + \beta_0 e^{\beta_1} y_t + \varepsilon_t,\ t = 1, \ldots, n$

   (b) $X_t = \mu + \beta_0 e^{y_t - \bar{y}} + \beta_1 y_t \sin y_t + \varepsilon_t,\ t = 1, \ldots, n$

   Hierbei ist $\varepsilon_t \sim N(0, \sigma^2)$ iid. Entscheiden Sie, ob es sich hierbei um lineare Modelle handelt. Geben Sie ggf. die Designmatrix an. Begründen Sie Ihre Aussagen.

   (8 + 8 Punkte)

2. Gegeben sei das lineare Modell

$$\vec{X} = A\vec{\beta} + \vec{\varepsilon},\ \text{mit } A \in \mathbb{R}^{n \times p}, \vec{\beta} \in \mathbb{R}^p, \vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I_n).$$

   Als Schätzer für die Varianz kann $\hat{\sigma}^2 = \frac{1}{n-r} ||\vec{X} - A(A^T A)^- A^T \vec{X}||^2$ verwendet werden. Hierbei sei r der Rang von A.

   (a) Berechnen Sie $\mathbb{E}\left(\hat{\sigma}^2\right)$.

   (b) Bestimmen Sie die Verteilung von $\frac{\hat{\sigma}^2}{\sigma^2}(n-r)$ mittels einer quadratischen Form.

   Hinweis: Verwenden Sie, dass hier gilt $\mathrm{rg}(P_{\mathrm{Im}(A)^\perp}) = \mathrm{spur}(P_{\mathrm{Im}(A)^\perp}) = n - r$.

   (6 + 7 Punkte)

3. Wir betrachten das Modell der einfachen Varianzanalyse, d.h.

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \text{mit } 1 \le i \le k, 1 \le j \le n_i, \alpha_i \in \mathbb{R} \text{ und } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ iid.}$$

   Zeigen Sie mit Hilfe einer quadratischen Form, dass

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 \sim \chi^2_{d,\lambda},$$

   und konkretisieren Sie die Freiheitsgrade d. Geben Sie außerdem an, wann der Nichtzentralitätsparameter $\lambda = 0$ ist.

   (12 Punkte)

4. Betrachtet wird das Kovarianzanalysemodell:

$$X_{ij} = \rho_0 + \rho_i + \gamma(y_{ij} - \bar{y}_{i.}) + \varepsilon_{ij},\ \text{mit } 1 \le i \le k, 1 \le j \le n_i,\ \rho_0, \rho_i \in \mathbb{R} \text{ und } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ iid.}$$

   (a) Geben Sie die Designmatrix an.

   (b) Leiten Sie die Schätzer für $\rho_0, \rho_i, i = 1, \ldots, k$ und γ her. Die Schätzer für die übliche Varianzanalyse bzw. lineare Regression müssen nicht explizit hergeleitet werden.

   (4 + 8 Punkte)

5. Gegeben sei die folgende Zeitreihe: $X_t = 5e^t\eta_t$, $t \in \mathbb{Z}$. Hierbei gelte für $Y_t = \log \eta_t$:

$$Y_t - \tfrac{1}{2}Y_{t-1} = \varepsilon_t \text{ und } \varepsilon_t \sim N\big(0, \sigma^2\big) \text{ iid.}$$

   (a) Verifizieren Sie, dass $Y_t \sim N\big(0, \frac{4}{3}\sigma^2\big)$.

   (b) Berechnen Sie $\mathbb{E}(\eta_t)$.

   (c) Ist X_t stationär?

   (d) Transformieren Sie das Modell so, dass es sich als Summe eines Regressionsanteils und einer stationären Zeitreihe schreiben lässt.

(4 + 4 + 3 + 3 Punkte)

6. Der Datensatz `rats` enthält die Überlebensdauern von 48 Ratten in Zehntel-Stunden. Die Ratten wurden zunächst in 3 Gruppen von 16 eingeteilt, wobei jede dieser Gruppen ein anderes Gift (I,II oder III) verabreicht bekam. Anschließend wurde jede dieser Gruppen weiter in 4 Einheiten unterteilt. Jede dieser Einheiten wurde dann mit einem Gegenmittel (A,B,C oder D) behandelt.

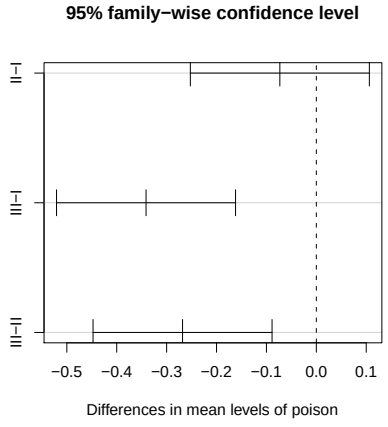
   (a) Gegeben sei der folgende Ausschnitt aus `R`. Interpretieren Sie die Ausgabe.

```
> anova(lm(time~poison*treat,data=rats))
Analysis of Variance Table

Response: time
             Df  Sum Sq Mean Sq F value    Pr(>F)
poison        2 1.03301 0.51651 23.2217 3.331e-07 ***
treat         3 0.92121 0.30707 13.8056 3.777e-06 ***
poison:treat  6 0.25014 0.04169  1.8743    0.1123
Residuals    36 0.80072 0.02224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   (b) Setzen Sie ein neues sinnvolles Modell für die Daten auf und geben Sie den entsprechenden `R` Befehl an.

   (c) Berücksichtigt man nur die Variable `poison`, so ergibt ein Tukey-Test die folgende Ausgabe. Erläutern Sie die Ergebnisse.



(4 + 6 + 4 Punkte)

7. Gegeben sei der Datensatz `oring.dat` aus der Vorlesung. Er enthält die Temperatur an bestimmten Tagen, sowie die Angabe, ob ein bestimmtes Bauteil des Space Shuttles ausfällt (1) oder nicht (0).

   (a) Die ersten Zeilen des Datensatzes sehen wie folgt aus:

   ```
   > oring
      temp defekt
   1    66      0
   2    70      1
   3    69      0
   ```

   Geben Sie den `R` Code an, um ein logit-Modell an die Daten anzupassen.

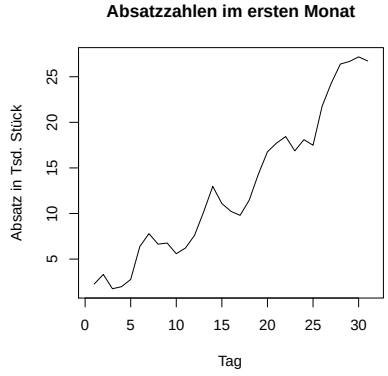
   (b) Der `summary` Befehl des angepassten Modells ergibt:

   ```
   > summary(oring.logit)
   ...
   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  15.0429     7.3786   2.039   0.0415 *
   temp         -0.2322     0.1082  -2.145   0.0320 *
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   ...
   ```

   Berechnen Sie die geschätzte Wahrscheinlichkeit $\hat{p}$ eines Ausfalls bei einer Temperatur von 45°F. Wie lautet der `R` Befehl zum Durchführen dieser Schätzung?

(5 + 7 Punkte)

8. Wir nehmen an, dass der Datensatz `zeitung.dat` Absatzzahlen (in Tsd. Stück) einer neuen regionalen täglich erscheinenden Zeitung enthält. Dabei wurde der erste Monat beobachtet. Der folgende plot veranschaulicht die Daten:



Wir nehmen an, dass $X_t = f(t) + \varepsilon_t$ ein passendes Modell für die Daten ist. Hierbei seien $t = 1, \ldots, 31$ feste Tage, $\varepsilon_t \sim N(0, \sigma^2)$ iid und X_t beschreibe die Absatzzahlen. Beschreiben Sie, wie Sie vorgehen würden, um die Funktion f zu schätzen, wenn vorausgesetzt wird, dass $f(t) = p(t) + \gamma \cos(\delta t)$ ist. Welche Wahl von δ erscheint hier sinnvoll?

(12 Punkte)