VRIJE UNIVERSITEIT AMSTERDAM

Exam Statistical Models on 10-2-2010

You may only use a calculator. The answers to the exercises need to be unambiguous ('no double answers') and clearly, but preferably concisely (Ned: 'bondig'), motivated. Exercise items are awarded with max. **2** points . Exam score is min(10,(total score+2)/3). Final mark is computed as denoted on the web site. **All statistical tests are to be performed using $\alpha = 0.05$, unless stated otherwise.**

Nb. The following formulas may be useful (but not necessarily).

$f(x_0,\hat\theta) \pm \hat\sigma\sqrt{\hat v_{x_0}^T(\hat V^T\hat V)^{-1}\hat v_{x_0}}\,t_{(n-p);\alpha/2}$ $\qquad f(x,\hat\theta) \pm \hat\sigma\sqrt{\hat v_x^T(\hat V^T\hat V)^{-1}\hat v_x}\sqrt{pF_{p,(n-p);\alpha}}$

$\hat v_x = (\frac{df}{d\theta_1}(x,\hat\theta_1),\ldots,\frac{df}{d\theta_p}(x,\hat\theta_p))$ $\qquad \hat\theta_j \pm \hat\sigma\sqrt{(\hat V^T\hat V)^{-1}_{jj}}\,t_{(n-p);\alpha/2}$

$\hat\mu_1 - \hat\mu_2 \pm \hat\sigma t_{n+m-2;\alpha/2}\sqrt{1/m+1/n}$

1. Figure 1 shows a series of measurements from the following model:

$$Y_i = \exp(\theta x_i) + \varepsilon_i,$$

where $\varepsilon_1,\ldots,\varepsilon_{20}$ are independent measurement errors which are assumed to be normally distributed with mean 0 and variance σ^2. The values of θ and σ^2 are unknown.

(a) Because of the exponential on the right-hand side, one might be tempted to use linear regression on $\log(Y_i)$ vs x_i. Why is this problematic?

(b) Find a starting value for θ, using $x = 1$ and using $x = 20$. Which starting value do you think is more correct?

(c) Given a starting value for θ what would be a good method to generate a starting value for estimating σ^2?

(d) Suppose $\hat\theta = 0.099$ and $\hat\sigma^2(\hat V^T\hat V)^{-1} = 1.237700 * 10^{-6}$. Give a 95% confidence interval for $\exp(12\theta)$, knowing that $t_{0.025,19} = 2.09, t_{0.025,18} = 2.10, t_{0.05,19} = 1.73, t_{0.05,18} = 1.73$.

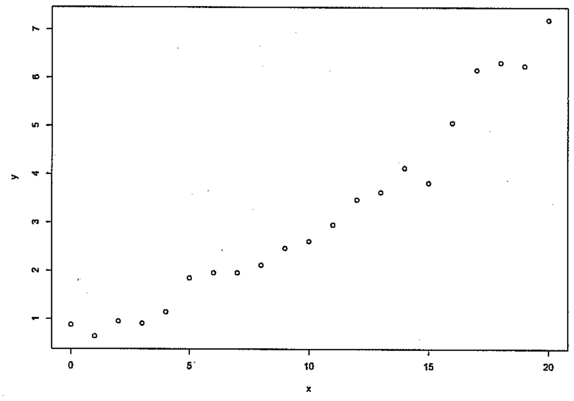

Figure 1: y vs x

2. The number of accidents on a crossroad (Ned: "kruispunt") is modeled as a function of the following covariates: nr of cars crossing per time unit (cars); average speed on roads to the crossroad (speed); type of crossroads (type: traffic lights, roundabout (Ned: "rotonde") or other), Data from 100 crossroads is available.

   (a) Specify the model you would use for such data.

   (b) Below, you see the results of testing the significance of speed in two settings (A and B):

```
A.

Analysis of Deviance Table

Model 1: y ~ speed
Model 2: y ~ 1  (intercept alone)
  Resid. Df        Resid.  Dev Df    Deviance    P(>|Chi|)
1          99      33.788
2         100      37.443  -1        3.655       0.036


B.

Analysis of Deviance Table

Model 1: y ~ cars + type1 + type2 + speed
Model 2: y ~ cars + type1 + type2
  Resid. Df      Resid.  Dev Df     Deviance P(>|Chi|)
1         96     31.788
2         97     33.212  -1         1.424    0.164
```

   What hypotheses are tested? What could be the reason(s) for the results to differ considerably between settings A and B?

   (c) How would you test the significance of the factor "type"? State the models.

3. Motivate your answer to the following questions.

   (a) Right or wrong? The graph in Figure 2 may represent autocorrelations of an MA-process.

   (b) It is known that a time series $X_t = \alpha X_{t-1} + \epsilon_t$, with independent $\epsilon_t \sim N(0, \sigma^2), \sigma^2 > 0$, can only be stationary for $\alpha < 1$. Show that if such a time series would be stationary for $\alpha \geq 1$ this would imply that $\text{Var}(X_t) \leq 0$ (and hence, indeed, such a time series cannot be stationary).

   (c) Explain why filtering may not be adequate for removing an exponential trend.

4. In 1979, the heights in inches of the singers in the New York Choral Society were recorded. The data may be grouped according to the voice parts of the singers in the choir. The voice parts, which differ in vocal range, are ordered from highest to lowest: "Soprano", "Alto", "Tenor" and "Bass". Hence, the data collected consist of a ratio scale variable, called "height" with the height in inches and a categorical variable "voice" with four levels. The total number of observations (singers) is 235. Figure 3 contains a boxplot of the observed distribution of the heights of the singers for each of the 4 voice parts. We want to investigate if there are significant differences in the average heights of singers in the different voice parts using ANOVA.

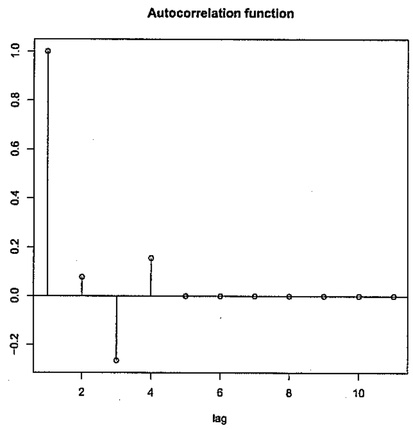

Figure 2: Autocorrelations over several lags for time series

(a) Write down an ANOVA model that may be used to analyze differences in average heights of singers in different voice parts. What assumptions do you make ?

(b) Perform a statistical test and conclude whether there are (any) significant differences in the average heights of singers in different voice parts. State the null hypothesis, the test statistic and perform the test using the data from the ANOVA table below. Test using a significance level $\alpha = 0.05$. In your conclusion, you may use (one of) the following quantiles of the F-distribution : $F_{1,235,0.05} = 3.88$,$F_{1,231,0.05} = 3.88$, $F_{3,235,0.05} = 2.64$,$F_{3,231,0.05} = 2.64$.

	Df	Sum Sq	Mean Sq	F value
voice	3	1962.31	654.10	?
Residuals	231	1460.84	6.32	

Table 1: Analysis of Variance Table

(c) Give a 95% confidence interval for the difference in average heights between singers that sing a "Soprano" (Sop) and singers that sing an "Alto" (Alt) part, based on a two-sample t-statistic. You may assume the variances in the two samples are equal. Calculate the confidence interval using the following data : sample averages : $\hat{\mu}_{Sop} = 64.12$, $\hat{\mu}_{Alt} = 65.39$, sample sizes : $N_{Sop} = 66$,$N_{Alt} = 62$, sample variances : $\hat{s}^2_{Sop} = 4.75$,$\hat{s}^2_{Alt} = 7.03$,t quantiles : $t_{126,0.05} = 1.66$,$t_{126,0.025} = 1.98$. Recall that for two indendent samples X (of size n) and Y of (size m) from a normal distribution with equal variance, the *pooled* estimate of the variance is given by

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

(d) Consider the plot in Figure 3. Suppose we would be able to measure voice parts on a more continuous scale, what would have been a more appropriate analysis of this data and why?

(e) Suppose we introduce a second factor "gender (male/female)" in the model. What do you expect to happen with the significance level of the factor voice part?

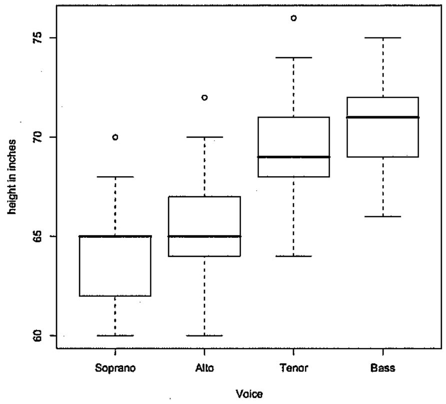

Figure 3: Heights of singers in the New York Choral Society