VRIJE UNIVERSITEIT AMSTERDAM

Exam Statistical Models, 19-02-2009

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. All exercise items are awarded with max. 2 points. Exam score equals min(10,10*(total score + 2)/(maximum score)). Final mark is computed as denoted on the website. **Nb.** The following formulas may be useful

(but not necessarily).

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

$\hat{v}_x = (\frac{df}{d\theta_1}(x, \hat{\theta}_1), \ldots, \frac{df}{d\theta_p}(x, \hat{\theta}_p))$ $\qquad$ $\hat{\theta}_j \pm \hat{\sigma}\sqrt{(\hat{V}^T\hat{V})^{-1}_{jj}}t_{(n-p);\alpha/2}$

1. A group of 100 men and 100 women is tested for the presence of the Human Papillomavirus (HPV). This virus may increase the risk for certain types of cancer. It spreads mainly via sexual intercourse. We want to know the potential risk factors for infection. The following covariates are used to determine those risk factors: age, gender, sexual orientation (homo- or heterosexual) and type of agglomeration: city or village (Ned: woont men in een stad of in een dorp?)

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

   (b) The covariate 'Age' is not used as a continuous covariate, but as a class-covariate instead: <18, 18-30, >30. Why is this a good idea for this type of data?

   (c) Below, you see the results of fitting two candidate models:

```
Analysis of Deviance Table

Model 1: y ~ age18 + age1830 + gender + orient + type
Model 2: y ~ age18 + age1830 + gender + orient
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        94     35.788
2        95     37.143 -1   -1.355     0.244,
```

   where age18 and age1830 are indicators for whether age is below 18 or between 18 and 30, respectively.
   Why have we not included a covariate for age larger than 30 (1 point)?

   (d) What hypothesis is tested (1 point)?

   (e) Can we conclude that type is not associated with the presence of HPV?

   (f) The odds-ratio for a covariate C with two levels A and B is defined as the ratio

$$\frac{\mathrm{P}(Y=1|C=A)/(1-\mathrm{P}(Y=1|C=A))}{\mathrm{P}(Y=1|C=B)/(1-\mathrm{P}(Y=1|C=B))}.$$

   The model we obtain for $p = \mathrm{P}(Y=1)$, with link function η (we use the usual canonical one) is $\eta(p) = 0.2 - 0.18 * \text{age18} - 0.02 * \text{age1830} + 0.2 * \text{gender} + 0.45 * \text{orient}$, where orient = 1 if the sexual orientation is homosexual, and 0 other wise, and gender = 1 for men and 0 for women. What is the estimated odds-ratio for orient=1 vs orient=0? [hint: what is $f(x)/(1-f(x))$, where $f(x) = e^x/(1+e^x)$?]

2. Right or wrong? Motivate your answer.

   (a) The graph in Figure 1 may represent autocorrelations of an MA-process.

   (b) $X_t = \alpha X_{t-1} + \epsilon_t$, with independent $\epsilon_t \sim N(0, \sigma^2)$, can be a stationary process when $\alpha \geq 1$.

   (c) The seasonal component for a weather-season related time series can be removed by differencing the series over $s = 3$ months (length of a season).

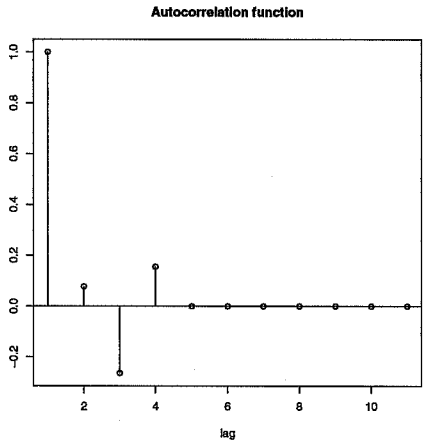

Figure 1: Autocorrelations over several lags for time series

3. We have a data set which is visualized in Figure 2 and which we want to describe by a nonlinear model:

$$Y_i = f(x_i; \theta) + \epsilon_i, i = 1, \ldots, n,$$

where the independent errors are distributed as $\epsilon_i \sim N(0, \sigma^2)$. Moreover,

$$f(x_i; \theta) = \theta_1 x + \log(\theta_2 + x).$$

   (a) Based on the data given in Figure 2, give a reasonable starting value for the vector θ and explain your choice.

   (b) Describe the steps of the algorithm that may be used to estimate θ.

   (c) Give a two-sided 95% confidence interval for the value of f at $x = 20$, knowing that $n = 16$, $t_{15;0.05} = 1.75, t_{15;0.025} = 2.13, t_{14;0.05} = 1.76, t_{14;0.025} = 2.14$, and

$$\hat{\sigma}^2(\hat{V}^T\hat{V})^{-1} = \begin{pmatrix} 0.00004 & -0.00206 \\ -0.00206 & 0.31595 \end{pmatrix}.$$

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.

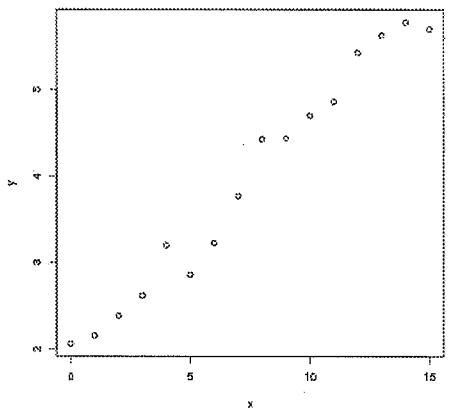

Figure 2: y vs x

(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) Look at the diagnostic plots in Figure 4. Suppose the plot on the left is a plot of the residuals against the fitted values for the ANOVA model used to answer (b). And suppose the plot on the right is a normal Q-Q plot of the same residuals. What do these plots tell you about the validity of the assumptions made in (a) ? Comment on what this means for the analysis in (b).

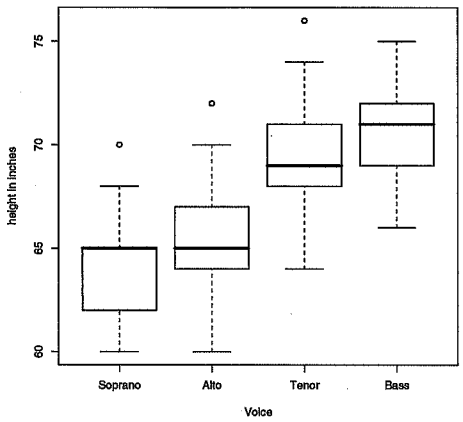

Figure 3: Heights of singers in the New York Choral Society

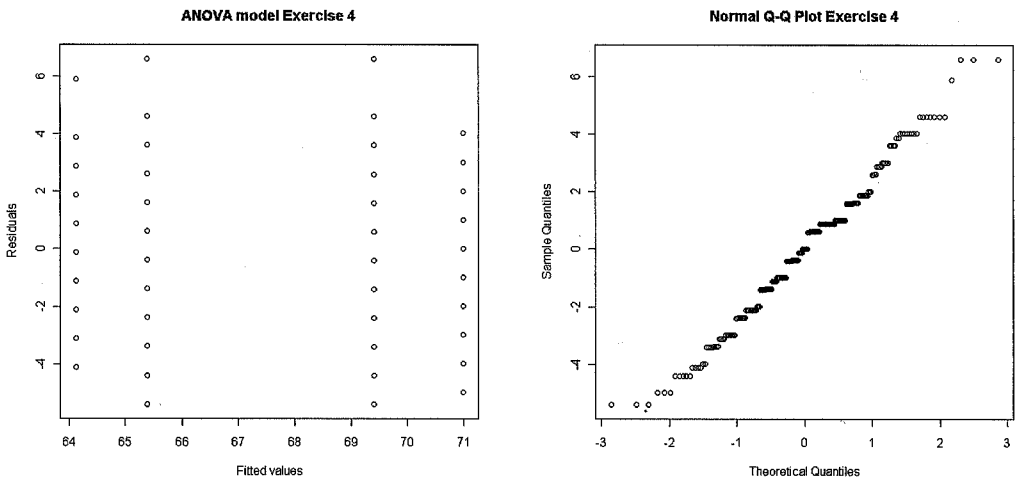

Figure 4: Diagnostic plots, Exercise 4