Exam Statistical Models Department of Mathematics
December 21, 2005 *vrije* Universiteit

1. a) Describe the three ingredients of a generalized linear model in general and clearly show that the loglinear model is of this type.

   b) A time series (X_t) is given. The model for this time series is that it consists of a trend m_t and a *stationary* stochastic process (Y_t). Describe a method to *detrend* the time series (X_t).

2. At an agricultural research center people are interested in the weight of certain types of melon having grown these for a specific period of time under controlled circumstances. There are four types of melon available, and for each type six were grown. The data frame `mdat` contains the eventually measured weights.

   a) Formally describe the ANOVA model that can be used to analyze the data in `mdat`.

   Below the in- and output of *R* is shown, where certain numbers are replaced by a question mark.

```
> m.aov<-aov(yield~type,data=mdat)
> summary(m.aov)
                Df    Sum Sq   Mean Sq   F value      Pr(>F)
type             ?   1291.48    430.49         ?   9.439e-07 ***
Residuals       20    367.65     18.38
```

   b) Describe a test that can be used to check whether the expected yield is the same for the various melon types. In any case describe the null hypothesis, the test statistic and the distribution of the test statistic under the null hypothesis. Moreover, also give the qualitative shape of the rejection region (and motivate your answer).

   c) Which numbers should be placed where the question marks are present? Your answer may contain sums of numbers, powers of numbers or ratios of numbers. Motivate your answer.

   d) Test the null hypothesis at level $\alpha = 0.05$ and concisely formulate the conclusion.

3. We have a data set $\{(x_i, y_i) : 1 \leq i \leq n,\ x_i \in I\!R, y_i \in I\!R\}$ and want to describe this by a nonlinear model:

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

where the ϵ_i's are i.i.d. normally distributed random variables with $E\epsilon_i = 0$ and $E\epsilon_i^2 = \sigma^2$, unknown. Moreover, the regression function is given by

$$f(x; \theta) = \theta_1 + \theta_2 e^{-\theta_3 x}$$

with $\theta = (\theta_1, \theta_2, \theta_3)^T \in I\!R^3$. Figure 1 shows a scatter plot of the data.

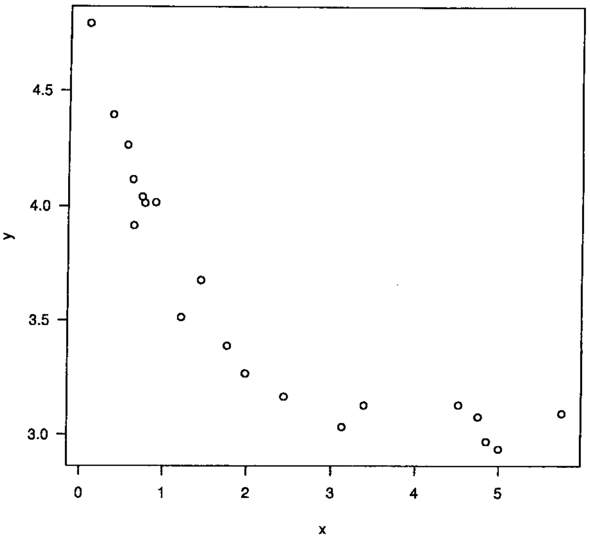

Figure 1: Scatter plot of the data

a) In order to estimate θ via maximum likelihood, one can employ the Gauss-Newton method. For this iterative numerical procedure, it is important to have reasonable starting values. Based on the data given in Figure 1, give a reasonable starting value for the vector θ and motivate your choice using Figure 1.

b) The matrix $\hat{\sigma}^2 \left(\hat{V}^T\hat{V}\right)^{-1}$ estimates the covariance matrix of the maximum likelihood estimator $\hat{\theta}$. For the given data set, this matrix is given by

$$\begin{pmatrix} 0.003955293 & -0.002125011 & 0.005465526 \\ -0.002125011 & 0.011406531 & 0.005120547 \\ 0.005465526 & 0.005120547 & 0.017765099 \end{pmatrix}$$

The parameter estimate resulting from the Gauss-Newton procedure is given by $\hat{\theta} = (3.05, 2.21, 1.16)^T$. Using this, construct a confidence interval for θ_3 of level approximately 95% based on classical theory. Also here, sums, ratios and powers of numbers may be used in the formula without explicitly evaluating these.

c) Describe concisely a procedure that could be followed for constructing a bootstrap confidence set for θ_3 of approximate level 95%.

4. Geneticists are studying the relationship between a disease and the genotype of a set of individuals and they hope to find an association between a certain genetic profile and the disease. Also the age of the individuals is recorded. Some individuals are affected by the disease and some of them are healthy. The genotype for each

individual consists of information on a series of SNPs, that are genes presenting only two possible variants. Every SNP, for every subject has two values measured. For example, there are two variants, A and a and for every SNP you can therefore be presented with the following combinations: AA, aa and Aa (here we omit the relevance of order so Aa and aA are considered to be equivalent).

a) What kind of variables are disease and SNP? How would you code disease and *one* SNP?

b) If you want to model the relationship between the response variable disease and the explanatory variables *age of the individual* and one specific *SNP*, what kind of statistical model would you use? Give a formal description of this model.

5. a) A collaborative study was conducted to study the precision of a method of determining the amount of niacin in cereal products. Homogenized samples of bread flakes were enriched with 0,2,4 mg of niacin per 100 g. Portions of the samples were sent to 12 labs, which were asked to carry out the specified procedures on each of three separate days. Which statistical method would you use to investigate whether the measurement procedures are equivalent, regardless of the lab and the amount of niacin added in the samples? Specify the underlying model.

b) A department store in Oxford Street is open every day from 9h to 17h. The managers want to plan their personnel efficiently in the future, so they want to employ the minimal number of people needed to achieve a certain service level. In order to do so, they need predictions for the number of customers that will actually visit the shop, depending on the day of the week (for morning and afternoon separately). For four years, data (per day, specific for morning and afternoon) are available on the number of customers that visited the shop. What kind of analysis would you use to advise the managers on this issue? Clearly describe the procedure to come to an appropriate model in this situation (indicate the steps you suggest).

Grading

1	2	3	4	5
a:5 b:3	a:3 b:3 c:2 d:2	a:3 b:3 c:4	a:4 b:5	a:3 b:5

Grade of written exam: 1+(number of points)/5

Final grade:

mean(weakly exercises)/3 + 2(grade of written exam)/3, provided the exam result is ≥5.5

Good luck with this exam!