Vrije Universiteit Amsterdam, Department of Mathematics

Exam Statistical Models on 14-02-2008

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 is the (total number of points [max. 32] plus 3) divided by 3.5. Final mark is computed as denoted on the website.

1. In order to investigate the question whether caffeine affects the capacity of memory we choose 18 students who have to study a certain material for 8 hours. Every 60 minutes there is a break so that they can have a drink containing caffeine. The drinks vary in the amount of caffeine (low, middle, high) and the way they are sweetened (only sugar, half sugar half sweetener, only sweetener). Two students are randomly assigned to the same drink and at every break they get the same drink (same caffeine-sweetener-combination). No other drinks or food is allowed.
   At the end the students have to take a test and the number of mistakes are reported in the table below:

	low	middle	high
only sugar	12 10	7 12	19 21
half sugar, half sweetener	6 3	10 16	21 29
only sweetener	15 12	8 6	24 13

   (a) Write down an appropriate model.
   (b) Could you still use the model you chose in (a) if
      (i) the sweetening would be described by: 10 grams of sugar, 5 grams of sugar and 2ml sweetener, 4 ml sweetener or
      (ii) each group of the two students could decide on their own how many grams of sugar they want?
   (c) For the moment don't distinguish between different ways of sweetening and only consider the factor caffeine. Let α_1 be the parameter that corresponds to the effect *low*. Choose one constraint and derive the formula for the estimator of α_1. (If you don't know how to derive it then at least state the formula.)
   (d) Going back to the full model, define the term *additive model*. Explain how to test whether we have such a model by mentioning the hypothesis, stating the test statistic, explain the distribution of the test statistic and finally conclude by using 0.1003385 as the corresponding p-value. Is your conclusion supported by the interaction plot in Figure 1?
   (e) Assuming a model without the sweeteners, how would you answer the initial question if the corresponding F-value of the test is $\mathcal{F}_A = 11.1902$ using the quantiles $F_{3,15,0.99} = 5.416965, F_{2,16,0.9999} = 17.29822, F_{2,16,0.99} = 6.226235$ and/or $F_{16,2,0.99} = 99.4$?

2. We have a data set which is visualized in Figure 1 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,
$$f(x_i;\theta) = \theta_1 + \theta_2 \frac{\exp(x-\theta_3)}{1+\exp(x-\theta_3)}.$$

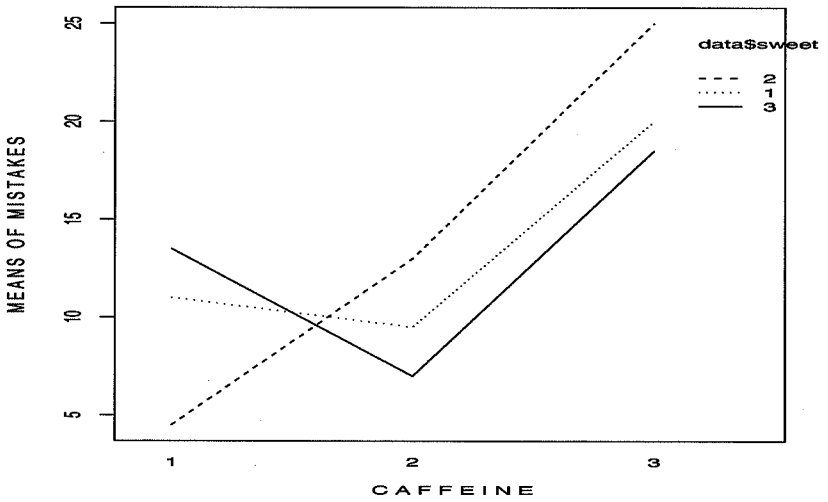

Figure 1: Interaction plot

(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 2, give a reasonable starting value for the vector θ and explain your choice.

(b) Based on 50 observations, the parameter estimate resulting from the Gauss-Newton procedure is given by $\hat{\theta} = (1.05, 1.98, 10.15)$. The matrix $\hat{\sigma}^2(\hat{V}^T V)^{-1}$ that estimates the covariance matrix of the maximum likelihood estimator looks the follow- ing:

$$\begin{pmatrix} 0.002610116 & 0.002509865 & 0.003893631 \\ 0.002509865 & 0.003715858 & 0.001865147 \\ 0.003893631 & 0.001865147 & 0.029469388 \end{pmatrix}$$

Construct an approximate 95%-confidence interval for θ_3 based on classical theory. If needed use $t_{47,0.95} = 1.6779$ and $t_{47,0.975} = 2.0117$

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

(d) What is the advantage of the bootstrap estimator with respect to the classical one?

3. Consider the following MA model:

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q}$$

where $Z_i \sim N(0, \sigma^2)$.

(a) We assume a stationary series. What does this mean?

(b) Find the expectation EX_t, variance VX_t and lag h covariance $\gamma_X(h)$ for this series.

(c) Now add an AR component to an MA(2) model so $X_t = \alpha X_{t-1} + \beta_0 Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2}$. What is the VX_T under the assumption of stationarity?

(d) Now suppose there is a trend and seasonality present in the data. Discuss two methods to remove these. Make clear why these methods work.

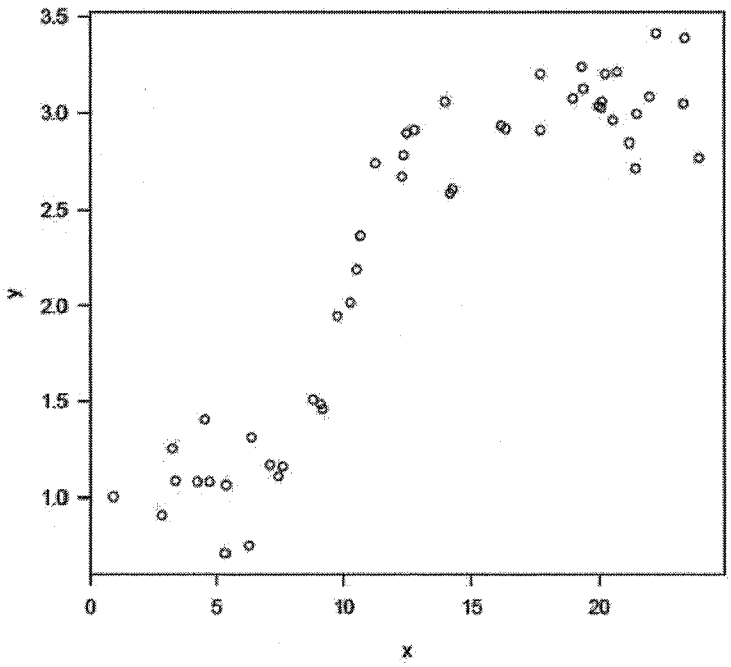

Figure 2: y vs x

4. The remission (Ned: 'terugkeer') of a disease is modeled using a logistic regression model. The scientists believe this event may depend on only two covariates: age and success of treatment.

   (a) Write down the model, including the assumptions.

   (b) The general form of the exponential family is $f(y,\theta) = \exp(\frac{y\theta - b(\theta)}{\phi/A} + c(y,\phi/A))$. The binomial density is $f_2(y,\mu) = \binom{n}{y}\mu^y(1-\mu)^{n-y}$. Use this to show that the binomial density is a member of the exponential family.

   (c) We obtain the following analysis of deviance table. Test at $\alpha = 0.05$ (including formulation of the hypotheses), whether Age is relevant and whether the full model (both covariates present) fits better than an empty model. Note that $\chi_{1;0.95} = 3.84, \chi_{2;0.95} = 5.99$ and $\chi_{3;0.95} = 7.81$.

   Response: Remission

Terms	Resid Df	Resid Dev	Test	Df	Deviance
Age+Success	50	40.45			
Success	51	42.34	Age	1	-1.89
(Intercept)	52	47.49	Success	1	-5.15