VRIJE UNIVERSITEIT AMSTERDAM

Exam Statistical Models

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 + 2)/3. 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. 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,

   $$f(x_i; \theta) = \theta_1 x + \theta_2 \exp(-\theta_3 x).$$

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

   (b) Give a 90% confidence interval for θ_1, knowing that $\hat{\theta}_1 = 1.503$. Note that $n = 31$; $t_{30;0.9} = 1.697, t_{28;0.9} = 1.701, t_{30;0.95} = 1.310, t_{28;0.95} = 1.312$ and

   $$\hat{\sigma}^2(\hat{V}^T\hat{V})^{-1} = \begin{pmatrix} 1.89 * 10^{-5} & 0.000267 & 6.12 * 10^{-6} \\ 2.67 * 10^{-4} & 0.023100 & 1.88 * 10^{-4} \\ 6.12 * 10^{-6} & 0.000188 & 3.12 * 10^{-6} \end{pmatrix}.$$

   (c) Formulate the hypotheses for testing whether f is in fact linear. What statistic would you use with how many degrees of freedom?

   (d) For a new value of x, $x_0 = 2.5$, we would like to have an upper bound of which we are 95% confident that the response (Y) will be below this value. What is this upper bound? Note that $\hat{\theta} = (1.5, 20, 0.1)$.

2. Consider the following stationary AR(2) model. $X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + Z_t$.

   (a) Derive the Yule-Walker equations for this model.

   (b) Argue how these may be used to estimate β_1 and β_2.

   (c) A quadratic trend, $m(t) = a_0 + a_1 t + a_2 t^2$, is added to the series. Show that when we difference the series twice, the result is a stationary series.

3. We have data on the number of accidents on a certain crossroads (Ned: 'kruispunt'). Over 9 years, 3 different traffic light settings have been used, setting 1 from 2000-2002, 2 from 2003-2005 and 3 from 2006-2008. We would like to know whether 'setting' is relevant with respect to the number of accidents.

(a) What model would you use for this problem? Write it down in mathematical terms.

(b) Explicitly state the relevant hypothesis [in terms of the parameter(s)] for testing whether 'setting' is associated with the number of accidents.

(c) What is wrong with this 'experimental set-up' with respect to answering the question posed?

4. Plant researchers conducted a study on how certain genetic features affect the growth of rice plants under different conditions. Two varieties of rice plants, a wild-type strain and a strain of genetically modified rice plants, were grown under three different fertilization (Ned: 'bemesting') strategies. In each condition a total number of 24 plants, 12 from each variety, were grown under equal conditions for a predetermined amount of time and then harvested. The total number of observations is thus $12 \times 3 \times 2 = 72$. The ShootDryMass of all plants, which is a measure of plant growth, was determined on a ratio scale. The specific research question of interest was whether the growth of modified plants is significantly differently affected by (at least one of) the conditions *compared to* the wild-type, i.e. whether the effect of a certain growth condition on the growth of rice plants is *different* for the two varieties of plants considered.

(a) Write down a statistical model that can be used to answer the research question of interest and mention the basic underlying assumptions.

(b) State the research question as a formal statistical test regarding parameter(s) of the model in (a). Mention the null hypothesis, the test statistic and an appropriate critical value for the test at significance level α.

(c) Use the following summary ANOVA table to perform the test. Use $\alpha = 0.05$. You may use (one of the) following quantiles from the F-distribution : $F_{1,66;0.95} = 3.99, F_{2,66;0.95} = 3.14, F_{5,66;0.95} = 2.35, F_{1,71;0.95} = 3.98$. What is your conclusion ?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fert	?	7018.78	?	?	< 0.001
variety	?	22684.50	?	?	< 0.001
fert*variety	?	38622.33	?	?	?
Residuals	66	24562.17	372.15		

(d) The three different fertilization strategies (the factor **fert** in the preceding table), are labeled "A","B" and "C" and the plants (factor **variety**) as "wt" and "mutant". We estimate the parameters corresponding to the ANOVA model in (c) using the basic sum to zero constraints on the parameters. Figure 2, displays box-plots of the observed ShootDryMass values (12 values/plants in each plot). Each plot corresponds to one of the 6 possible combinations of plant variety and fertilization strategy. Given the following coefficient estimates, can you identify the two plots corresponding to fertilization strategy "A" ?

Coefficient	Estimate
Intercept	108.33
Fertilization B	-58.08
Fertilization C	-35.00
mutant	-101.00
mutant*Fertilization B	97.33
mutant*Fertilization C	99.17

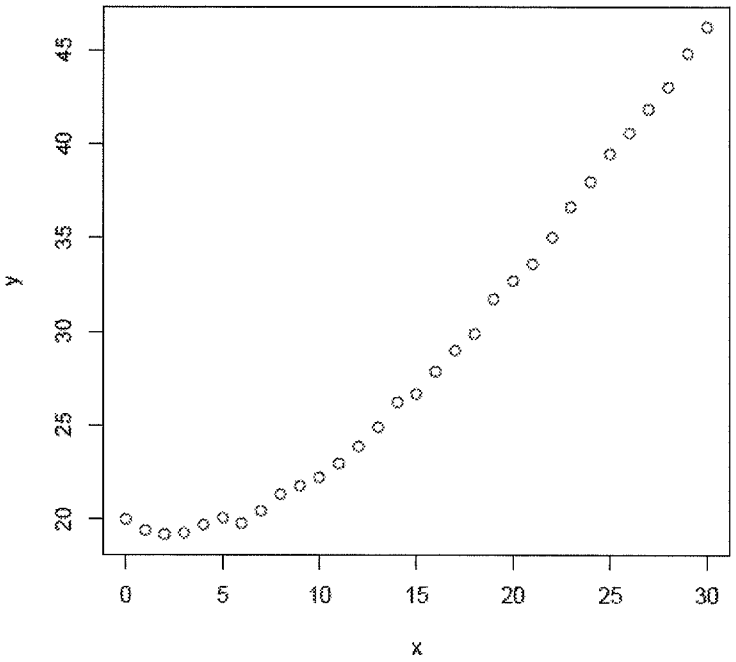

Figure 1: y vs x

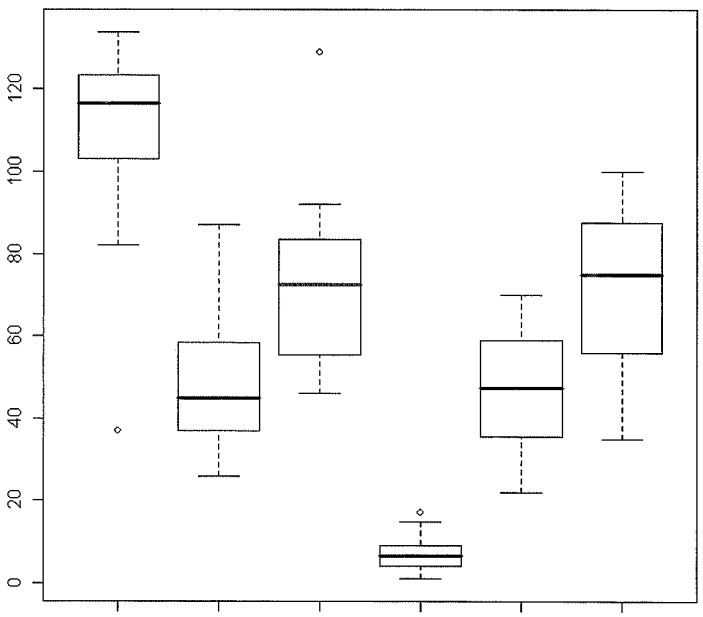

Figure 2: Box-plots of ShootDryMass of different plants under different growing conditions