Assignments
Statistical Modelling in the Sciences III - STAT300

Printed at the University of New England, May 24, 2004
Approximate submission dates.

The dates given are Mondays on which to post the assignment. They should be taken as a guide in order that you maintain the pace. There is no penalty for late submission other than that all assignments are required before end of semester, Friday, November 5, 2004.

<table>
<thead>
<tr>
<th>Assignment number</th>
<th>Posting Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 Aug</td>
</tr>
<tr>
<td>2</td>
<td>9 Aug</td>
</tr>
<tr>
<td>3</td>
<td>16 Aug</td>
</tr>
<tr>
<td>4</td>
<td>23 Aug</td>
</tr>
<tr>
<td>5</td>
<td>30 Aug</td>
</tr>
<tr>
<td>6</td>
<td>6 Sept</td>
</tr>
<tr>
<td>7</td>
<td>13 Sept</td>
</tr>
<tr>
<td>8</td>
<td>4 Oct</td>
</tr>
<tr>
<td>9</td>
<td>11 Oct</td>
</tr>
<tr>
<td>10</td>
<td>18 Oct</td>
</tr>
<tr>
<td>11</td>
<td>25 Oct</td>
</tr>
<tr>
<td>12</td>
<td>1 Nov</td>
</tr>
</tbody>
</table>

Data files and R programs for the assignments are stored at:

http://mcs.une.edu.au/~stat300/
Assignment 1: Linear regression for 2 groups, using R, (12 marks)

The data compares the rate of gas consumption with decreasing outside temperatures for a house before and after it was insulated. The data are filed in gas.txt. Also located there is a skeleton R program for analysing the data.

1. Plot Gas against Temp for each of the Insulate categories. 2 marks
2. State, using mathematical symbols, a statistical model that would explain the observed differences. 2 marks
3. Fit the model for the data and explain the information conveyed by the AOV. 2 marks
4. List the prediction equations, including standard errors of the coefficients, for each category. 2 marks
5. Is the rate of change of Gas with Temp different between Insulate categories? 2 marks
6. What is the distribution of residuals? 2 marks
Assignment 2: Graphics (12 marks)

1. A production plant cost-control engineer is responsible for cost reduction. One of the costly items in his plant is the amount of water used by the production facilities each month. He decided to investigate water usage by collecting seventeen observations on his plant’s water usage and other variables.

Data are filed in water.txt

(a) Produce a "pairs" plot for exploring the data. (cf. 3.1.1) 2 marks
(b) Plot Water as a function of Production and include a loess fit to the data on the plot. (cf. Example 3.1) 2 marks

2. A split plot experiment measured the yield of 3 varieties of oats at 4 levels of nitrogen. The data are obtained by

library(nlme)
data(Oats)

Use the xyplot function in the lattice library to plot yield as a function of nitro for each Block with superposed line traces for each Variety. As a guide, refer to the dotplot at section 3.1.8, using the xyplot function rather than dotplot and panel.superpose is the only panel function required. 4 marks

3. The data set warpbreaks lists the numbers of breaks in fabrics from 2 types of wool at 3 tensions.

Generate a dot plot to show the interaction of tension and wool on the number of breaks. 4 marks

data(warpbreaks)
library(lattice)

dp <- dotplot(tension ~ breaks, data=warpbreaks, panel=panel.superpose, groups=wool)
Assignment 3: Contrasts (14 marks)

1. The weights of chicks fed 4 diets are given in Table 1.

Table 1: Chick weights from 4 diets

<table>
<thead>
<tr>
<th>chick</th>
<th>diet</th>
<th>wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>53</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>56</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>61</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>53</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>57</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>67</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>76</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>69</td>
</tr>
<tr>
<td>15</td>
<td>C</td>
<td>65</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>66</td>
</tr>
<tr>
<td>17</td>
<td>D</td>
<td>63</td>
</tr>
<tr>
<td>18</td>
<td>D</td>
<td>71</td>
</tr>
<tr>
<td>19</td>
<td>D</td>
<td>68</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>76</td>
</tr>
</tbody>
</table>

(a) Produce a boxplot to suggest a statistical model for these data. 2 marks
(b) Propose a set of contrasts for comparing the diets. 2 marks
(c) Fit a linear model and present the results. Indicate which contrasts are significant. 4 marks

2. An experiment measured tensile strength of fibre with different percentages of cotton. The data are listed in Table 2.

(a) Plot the data in a way that suggests a suitable model. 2 marks
(b) Use these contrasts to analyse the effect of Cotton % on strength. Table the coefficients of the model. 4 marks
Table 2: tensile strengths for different % cotton.

<table>
<thead>
<tr>
<th>Cotton (%)</th>
<th>experiment units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>35</td>
<td>7</td>
</tr>
</tbody>
</table>

Cmat <- matrix(c(-3,-1, 1,3, 0, 
1,-1,-1,1, 0, 
0, 0, 0,1,-1 ), nrow=5,ncol=3) 
dimnames(Cmat)[2] <- list(c("lin","quad","30-35"))
cotton.mod2 <- lm(strength ~ C(wt,Cmat),data=cotton) 
print(summary(cotton.mod2)$coefficients)
Assignment 4: Subset selection (12 marks)

Fat and muscle depths were measured on pork carcasses in order to predict the saleable yield of meat.

Download the program carc.r and the data carc.txt from the web page.

The data consist of fat and muscle depth measurements, lean meat %, Wt and sex (1CM,F,M) from carcasses. The data were collected to calibrate relationships between the measurements taken at 4 positions on the carcass (P2, 3/4, H1, H2) and lean meat %.

(a) Run the program to do backwards elimination using the option k=AIC in the step() function.  
State the final model after non-significant terms have been deleted. (Hint: try formula(step.model); anova(step.model) etc).  
2 marks

(b) Change that option to k=BIC and rerun.  
2 marks

(c) Examine the R help files to understand the difference between AIC and BIC and then comment on the different fits obtained from (a) and (b). Type help(step) in the R console.  
2 marks

(d) How does the option trace=0 affect the step() function?  
2 marks

(e) Plot observed against fitted values for either (1) or (2).  
2 marks

Your report may be a brief explanation of the results with the regression coefficients from (1) and (2) attached as an appendix.

---

1 castrated males
Assignment 5: Lattice graphics, Nonparametric regression, (10 marks)

The data in abundance.txt are the measures of abundances of a forb\(^2\) and a weed along 4 transects through a paddock at 4 sites. The variable Distance refers to the distance from the edge of the paddock and the researcher is interested in modelling the change in abundance due to distance from the edge.

1. Use the lattice library to produce an exploratory data plot which would indicate how to model the changes in forb and weed abundances due to distance. You are free to use the R code accompanying the data on the web. \(2\) marks

2. Model the weed abundance from Site9 as a function of Distance. List the AOV and table of regression coefficients. \(4\) marks

3. Plot the fitted curve and the 95% CI for the fit. \(4\) marks

\(^2\)A forb is any herbaceous plant that is not a grass.
Assignment 6: Linear Mixed Model, (10 marks)

A balanced cross-over trial was conducted on 16 subjects to compare 2 treatments for asthma. The first group of 8 subjects received the treatment A in period 1 and after a rest, received B in period 2. The second group received the treatments in order BA. The measurement was Forced Expired Volume and the data are saved in `fev.txt`.

1. For exploratory data analyses, plot the observed differences between periods for each subject within each group using the `lattice` graphics. 2 marks

2. Fit a linear mixed model for interpreting FEV that has Trt and Per as fixed main effects (do not fit the interaction) and Subj as a random effect. 4 marks

3. List the AOV and treatment effects and state conclusions. 2 marks

4. Plot the cumulative distribution function of the random effects. 2 marks
Assignment 7: Incomplete Blocks (10 marks)

An engineer is studying the mileage characteristics of five types of fuel additives. He uses cars as blocks in an incomplete block design.

1. Boxplot the data to compare sample means.  
   2 marks

2. Calculate the sample means for additives.
   (Hint: Use `tapply(mileage, additive, mean)`)  
   2 marks

3. Analyse the data from this experiment as a mixed model with random blocks.  
   2 marks

4. List the predicted means and standard errors from the model.  
   2 marks

5. Compare the observed (or sample) means and the estimated additive means.  
   2 marks

<table>
<thead>
<tr>
<th>Additive</th>
<th>Car 1</th>
<th>Car 2</th>
<th>Car 3</th>
<th>Car 4</th>
<th>Car 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
Assignment 8: Power (14 marks)

A researcher wishes to conduct experiments on white clover lines and as part of his planning, he wants to know how many leaves he should sample. His supervisor has data on the variables of interest which will be used to gauge sample sizes \( a \text{ priori } \). The data are filed in a Lines.txt. The variables in the file are separated by commas and are read by 
\[ \text{read.csv("Lines.txt",header=T)} \]

For the variable leafLmm (leaf length in mm),

1. Calculate tables of means and standard deviations, eg. 
\[ \text{sd.tab <- sqrt(tapply(clover$LeafLmm,INDEX=clover$Line,FUN=var))} \quad 2 \text{ marks} \]

2. Produce the results of the R command 
\[ \text{plot(mean.tab, sd.tab)} \quad 2 \text{ marks} \]

3. Analyse the data as a linear model and print the coefficients. These are Line effects and are a useful guide to detectable differences. \( 4 \text{ marks} \)

4. Nominate 3 standard deviations (L,M,H) that might cover the variability in the future experimental effects.
   Note: s.d. \( \approx \) s.e. \( \times \sqrt{5} \) in this case. Use \texttt{power.t.test} to derive the detectable differences for sample sizes 3 to 10 when the power is set at 0.8, \( \text{crit}=0.05 \) and the standard deviation is each of the 3 nominated values. Plot the curves. \( 4 \text{ marks} \)

5. Make a recommendation. \( 2 \text{ marks} \)

A brief R program called \texttt{clover.r} is listed on the web page for assistance.
Assignment 9: Longitudinal Data (10 marks)

Download a9.r. The data are part of the nlme library and plotting functions may depend on the latest releases of R and the library lattice. However, if the advanced plotting is not possible, you can list coefficients to answer the questions.

For this question, the R code is provided and answering the questions serves as a tutorial. Brief answers only are required as the exercise is for you to learn by doing.

The data are growth curve data which are repeated measurements from subjects over time. The response is the distance from the pituitary gland to the pterygomaxillary fissure taken every 2 years from 8 years of age until 14 years of age on a sample of 16 boys and 11 girls. The data were collected by orthodontists from X-rays.

Using the supplied R code,

1. Write the mathematical model, systematic and random parts, that is being fitted in lme1. \[2\text{ marks}\]
2. What are the average (over subjects) responses due to age for boys and girls? \[2\text{ marks}\]
3. What are the estimates of the distributions of the random effects? \[2\text{ marks}\]
4. Briefly suggest if the assumptions of the random effects are plausible. \[2\text{ marks}\]
5. In the comparison of fits from the naive model and the lme model, explain why the fits differ most for M13 and M04. \[2\text{ marks}\]
Assignment 10 : GLMs (10 marks)

The data give the incidence of non-melanoma skin cancer among women in Minneapolis-St Paul, Minnesota, and Dallas-Fort Worth, Texas.

(One would expect sun exposure to be greater in Texas than in Minnesota.)

The files for this are skin.r and skin.txt.

The aim of this assignment is to give you practice in fitting GLMs and in interpreting the results. The program suggests ways to do the analysis and serves as a tutorial where you pick up extra tips as well as an assignment.

1. Exploratory Data Analysis :
   Observe that the Populations are similar for both towns but weighted more toward younger people in Fort Worth.

   (a) Plot Cases against median Age for both Towns. \hspace{1cm} 1 mark
   (b) Plot Cases per 50,000 against median Age for both Towns. \hspace{1cm} 1 mark

2. Modelling Cases as a Poisson random variable with the offset log(Population).

   (a) Fit Cases as a response to medAge for each town. \hspace{1cm} 1 mark
   (b) Interpret the analysis of deviance table. \hspace{1cm} 2 marks
   (c) Plot observed values, fitted values and their 95\% confidence intervals. \hspace{1cm} 1 mark

3. Modelling Cases as a Binomial random variable, with Population as the number of trials.

   (a) Model the Cases/Population using family="binomial". \hspace{1cm} 2 marks
   (b) Plot observed, fitted values and confidence limits. \hspace{1cm} 2 marks
Assignment 11: Contingency table (10 marks)

The following data, which concern two species of lizard, GRAHAM and OPALINUS, were collected by observing occupied sites or perches and recording the appropriate description,

- species of lizard
- time of day - early, mid or late
- height of perch (feet)
- diameter of perch (inches)
- whether the site was sunny or shady

<table>
<thead>
<tr>
<th>Site</th>
<th>diam</th>
<th>height</th>
<th>Early</th>
<th>Mid</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>G  O</td>
<td>G  O</td>
<td>G  O</td>
</tr>
<tr>
<td>sunny</td>
<td>≤ 2</td>
<td>&lt; 5</td>
<td>20</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>≥ 5</td>
<td></td>
<td>13</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>&gt; 2</td>
<td>&lt; 5</td>
<td>8</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>≥ 5</td>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>shade</td>
<td>≤ 2</td>
<td>&lt; 5</td>
<td>34</td>
<td>11</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>≥ 5</td>
<td></td>
<td>31</td>
<td>5</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>&gt; 2</td>
<td>&lt; 5</td>
<td>17</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>≥ 5</td>
<td></td>
<td>12</td>
<td>1</td>
<td>21</td>
</tr>
</tbody>
</table>

These data are in a file lizards.txt and a skeleton script, lizards.r is provided to show new computing.

The model fitting is left to you as it has been covered in previous assignments.

1. Construct an appropriate model to analyse the multi-way contingency table. (Use step() in conjunction with glm().) 5 marks

2. Calculate the table of predicted values and standard errors for significant interactions. 3 marks

3. Use the dotplot function of the lattice graphics library to plot the interactions. 2 marks
**Assignment 12**: Multivariate Analysis of Variance (10 marks)

The optimum conditions for extruding plastic film have been examined using a new technique. In the course of this study, three responses Tear resistance, Gloss and Opacity were measured at two levels of factors, *rate of extrusion* and *amount of an additive*. The measurements were repeated 5 times at each combination of the factor levels. The data are given in the file `plastic.txt`.

The levels for the change in the rate of extrusion were Low (-10%) and High (+10%), while the levels for amount of additive were lo (1.0%) and hi (1.5%).

1. Use R to produce an analysis to test for effects due to the rate of extrusion and the amount of additive.  
2 marks

2. Choose an appropriate model using the output in 1., giving reasons for your choice.  
2 marks

3. Are the assumptions underlying the use of MANOVA tenable?  
Give evidence to back your claims.  
2 marks

4. Produce evidence to justify the use of MANOVA rather than separate AOVs for each response.  
2 marks

5. Present graphical evidence to support your choice of model in 2.  
2 marks