620-270 Applied Statistics

Assignment 2 (sem 2/02)

Due date: 4 pm, Monday 9th September.

Question 1 and Question 2(j) require you to use MINITAB. These may be done during Lab 5 next week. All other questions are to be done by hand.

- Your assignment should show all working. Marks will be given for method as well as for correct answers.
- Assignments count 25% of your assessment. Tutors are not permitted to help you directly with the assignment questions. However, they may give appropriate guidance. Use your imagination.
- This assignment must be placed in your tutor’s box in the north-south corridor on the ground floor of the Richard Berry Building, near the Masson Road entrance.
- Late assignments will only be accepted under exceptional circumstances and must be handed personally to the lecturer, Karen Baker (room G41), with a written explanation for submitting late and/or a medical certificate. A late penalty may be imposed.
- Label your assignment with the following information:
  Your name and student number, tutor’s name, tutorial group (day and time).
- Attach a signed and dated plagiarism statement (see attached sheet).

1. Below are characteristics of 69 cars in the 1979 model year as sold in the United States. The variables are:

   mpg = mileage in miles per gallon
   headroom = headroom in inches
   price = price in $US
   weight = weight in pounds
   tr space = trunk space in cubic feet
   length = length in inches
   engine = engine displacement in cubic inches
   where = where it was made

   The data look like:

   car name  mpg  price  headroom  tr space  weight  length  engine  where
   Amc.Concord  22  4099  2.5  11  2930  186  121  US
   Amc.Pacer  17  4749  3.0  11  3350  173  258  US
   Audi.5000  17  9690  3.0  15  2830  189  131  EU
   Audi.Fox  23  6295  2.5  11  2070  174  97  EU
   etc.

   The data are stored in a MINITAB worksheet called car_mileage.mtw which is located in the 620-270 Semester 2 folder in the Statistics folder in the Maths & Stats folder on the Mac desktop. Transfer the worksheet to the Student Folder so you can work on it.

   Suppose that for the car mileage data we had access to only the first four variables: mpg, price, headroom and tr space. We want to fit a model which predicts mpg from some or all of the three explanatory variables.
(a) Plot all four variables against each other using matrixplot.

(b) Calculate the correlation coefficients between all pairs of these four variables. Which explanatory variable has the strongest association with the response variable? Which pair of explanatory variables has the strongest association?

(c) Fit a simple linear regression model relating mpg to the variable with the strongest association. State the regression equation and give an interpretation.

(d) Find the standard error of the slope of the regression line, and a 95% confidence interval for the slope.

(e) We now want to fit a multiple linear regression model by adding the explanatory variable most likely to improve the current model. We could look at the $r$ values calculated earlier, but a better way is to look at the residuals from the current model, plotted against the other explanatory variables. Run the simple linear regression again, this time saving the residuals. Plot the residuals against the other two explanatory variables.

(f) From the residual plot, choose the variable with the strongest association, and fit a multiple regression model with this variable added. What is the new equation? Does this variable make a significant improvement? (i.e. test the hypothesis that its coefficient is zero).

(g) Using the list of “Unusual Observations” in the MINITAB output, underline the observation with the largest residual, and circle it on one of the plots in the matrixplot.

(h) Add the third explanatory variable to the model, and test whether it makes a significant improvement to the model.

(i) Using the most appropriate model, find the predicted mpg when the price is $6000, the headroom is 3 inches, and the trunk space is 15 cubic feet.

2. An experiment to investigate the relationship between the heat evolved during the hardening of cement and the composition of the cement gave rise to the following:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>26</td>
<td>60</td>
<td>78</td>
</tr>
<tr>
<td>11</td>
<td>52</td>
<td>20</td>
<td>104</td>
</tr>
<tr>
<td>11</td>
<td>55</td>
<td>22</td>
<td>109</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>6</td>
<td>102</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>44</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>22</td>
<td>93</td>
</tr>
<tr>
<td>21</td>
<td>47</td>
<td>26</td>
<td>115</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>34</td>
<td>83</td>
</tr>
<tr>
<td>11</td>
<td>66</td>
<td>12</td>
<td>113</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>14</td>
<td>109</td>
</tr>
</tbody>
</table>

where $x_1$, $x_2$, $x_3$ denote percentages by weight of three components, and $y$ denotes the heat evolved in calories per gram of cement.

The model:

$$ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e, \quad \text{where } E \sim N(0, \sigma) $$

and fitted to the data. Incomplete MINITAB output is shown:
Regression Analysis: y versus x1, x2, x3

The regression equation is
\[ y = 73.4 + 1.52 x1 + 0.381 x2 - 0.268 x3 \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>73.40</td>
<td>14.69</td>
<td>5.00</td>
<td>0.002</td>
</tr>
<tr>
<td>x1</td>
<td>1.5162</td>
<td>0.1295</td>
<td>11.71</td>
<td>0.000</td>
</tr>
<tr>
<td>x2</td>
<td>0.3815</td>
<td>0.1941</td>
<td>1.97</td>
<td>0.097</td>
</tr>
<tr>
<td>x3</td>
<td>-0.2685</td>
<td>0.1841</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S = _______  R-Sq = _______

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>--</td>
<td>2001.54</td>
<td>667.18</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Residual Error</td>
<td>--</td>
<td>______</td>
<td>_____</td>
<td>_____</td>
<td>_____</td>
</tr>
<tr>
<td>Total</td>
<td>--</td>
<td>______</td>
<td>_____</td>
<td>_____</td>
<td>_____</td>
</tr>
</tbody>
</table>

Use Minitab only in part (j).

(a) Compute the missing values in the Minitab output, shown by horizontal lines. Show how each of your answers is obtained.

(b) Estimate \( \beta_0, \beta_1, \beta_2, \beta_3, \sigma^2 \).

(c) What is the meaning of the \( \beta_i \) estimates.

(d) Find the predicted value of the heat evolved when the percentages of the three components \( x_1 = 10, x_2 = 40, x_3 = 30 \).

(e) Given that the standard error of the predicted value in (d) is 1.632, find a 95% confidence interval for the average heat evolved when \( x_1 = 10, x_2 = 40, x_3 = 30 \).

(f) Find a 95% prediction interval for the heat evolved when \( x_1 = 10, x_2 = 40, x_3 = 30 \).

(g) Interpret and compare the intervals found in (e) and (f).

(h) The correlation between \( x_1 \) and \( x_3 \) is -0.196, while the correlation between \( x_2 \) and \( x_3 \) is -0.965. Compare and interpret these values.

(i) Test the hypothesis that \( \beta_3 = 0 \) at the 0.05 significance level. Based on the result, what is a simpler model which adequately describes the data? How can the result of this test be seen to be consistent with the values of the correlation coefficients given in (h)?

(j) Use Minitab to fit the simpler model and, using it, find 95% confidence intervals for \( \beta_1 \) and \( \sigma^2 \).

(k) Use an \( F \)-test to compare this simpler model with the null model, \( \beta_i = 0 \) for all \( i \).

3. Four features of good experimental design are replication, randomisation, blocking and balance. Explain what is meant by each of these terms and what their purpose is.
4. Four fertilizers (A, B, C and D) are to be tested for their effectiveness in increasing the yield of wheat. For each proposed design below:

   i. define the experimental unit, and say how many there are;
   ii. state what type of design is being used, describing any blocking;
   iii. identify any problems with the design.

(a) 20 plots of land are used. 5 plots are randomly chosen to get each fertiliser.
(b) 20 plots of land are used. Each plot is divided into four parts, and the four fertilisers are randomly allocated to the parts of each plot.
(c) 4 plots of land are used, and the four fertilisers are randomly allocated to the plots.
(d) 4 plots of land are used. Each plot is divided into four subplots, and fertiliser C (randomly chosen) is given to all four subplots in the first plot, A (randomly chosen) to all four subplots in the second plot, etc.
(e) The experimenters examine the district and find five fields that have been fertilised with fertiliser A, another five with B, etc.
(f) 20 plots of land are used. The five plots that have in the past given the worst yield are given the fertiliser that is thought to be most effective, since they would seem to need it. Similarly for the other fertilisers.
(g) 20 plots of land are used. The five plots that have in the past given the worst yield are randomly assigned one of the treatments each, then the next worst four, and so on.

---

620-270   Applied Statistics

Assignment 2 (sem 2/02)

PLAGIARISM is the presentation by a student of an assignment which has in fact been copied in whole or in part from another student’s work, or from any other source without due acknowledgement in the text.

TEAMWORK: It is the goal of all courses in the Department of Mathematics and Statistics to develop generic skills that will be useful in the workplace and at the same time they develop specific skills, knowledge and understanding in Mathematics and Statistics. A key generic skill is teamwork. Students are encouraged to work with each other, inside and outside tutorials, in order to develop their technical skills, to increase their knowledge and understanding of the curriculum and to develop the capacity for teamwork. In the case of marked assignments or other non-examination assessments in the Department of Mathematics and Statistics, such teamwork includes general discussion and sharing of ideas on the work. All written work must however (without specific authorisation to the contrary) be done by individual students. Students are neither permitted to copy any part of another student’s work nor permitted to allow their own work to be copied by other students.

DECLARATION: I declare that this assignment is my own work and does not involve plagiarism or teamwork other than that authorised in the general terms above or specifically authorised for this assignment.

Signed:_____________________________ Date:_______________