STAT2008/6038 Regression Modelling Assignment One

Research School of Finance, Actuarial Studies and Applied Statistics

Semester 1, 2013

This assignment is worth 20% of your overall course mark. Solutions should be placed in the STAT2008/6038 assignment box labelled with your tutor's name by:

28 March, 2013, 12 Noon

The boxes are located adjacent to the office of the Research School of Finance, Actuarial Studies and Applied Statistics, on level 4 of Building 26C. The assignment MUST be completed individually. This is NOT a group assignment.

Assignment format: Please use both sides of the page if possible to save paper. You can submit hand written or typed solutions. If you submit hand written solutions they must be neat and legible. Please present your assignment in a single stack of pages held together by a staple in the upper left hand corner. (That is, do not use folders, plastic sheaths or other devices to join pages together.) On a cover sheet, a copy of which is available on the Wattle site for the course, clearly state the details required, including your name and student number and your tutorial group details (tutorial group number, day, time and tutor's name).You will be assessed on your understanding of statistics and your ability to use **R**, not on your typing and word-processing expertise.

Late Assignments: Each assignment will be due just before the first class that discusses its solutions, so as a general rule, NO LATE ASSIGNMENTS WILL BE ACCEPTED. If you are unable to hand in your assignment by the due date please see me at least one working day before the due date (with your medical certificate) to arrange alternative assessment. No exceptions. This is School Policy.

R should be used to produce the relevant graphics and statistics, unless the question directs you otherwise. Answers should be presented as an edited Microsoft Word file. You should include all

relevant graphs and values generated in **R** (e.g. if a question directs you to create and discuss a graph, you should include the graph itself as well as your discussion in your submitted assignment). For each question, the data can be found on Wattle.

All numeric answers should be rounded to 4 decimal places as appropriate, clearly stating where rounding has been used. You should show all relevant calculations and working for questions which you are instructed to do "by hand". Marks will be deducted for failure to show working or calculations. Additionally, you should be careful to define all notation used. Marks may be deducted for unexplained notation.

1 Question One

The attendance and temperature for 100 rugby games were recorded in the .csv file `rugby.csv`. The variables of interest in this question are the Temperature (in degrees celcius) and Attendance (in thousands) for each rugby game. Analyse the data by answering the following questions.

1. (5 marks) Fit a simple linear regression with Attendance as the response and Temperature as the explanatory variable. What is the equation of the fitted regression line? Calculate the coefficient of determination, R^2.

2. (3 marks) Plot the data and superimpose the fitted regression line.

3. (10 marks) Using the `cor.test()` function and an analysis of variance table constructed in **R** test whether the regression is significant at the $\alpha = 0.10$ level. Use both the F distribution and the T distribution. (you will need to present THREE full hypothesis tests. You are permitted to submit hand written solutions for this question). You may only use the ANOVA table and the output from the cor.test() function to answer this question.

4. (5 marks) Is it reasonable to believe that any relationship between these two variables should pass through the origin? Explain. Test this proposition for the model fit in (1).

5. (7 marks) Consider the regression model fit in part (1). Construct a plot of residuals versus fitted values, a Q-Q plot of residuals and a barplot of leverages for this model, comment on the model assumptions and identify any unusual data points. Comment on any unusual data points.

2 Question Two

The data set `ass1q2.csv` contains information about two variables X and Y. Unfortunately, we don't know what the two variables are but we have been given the task of trying to find a suitable model to explain the relationship between the two variables.

1. (8 marks) Fit a simple linear regression with Y as the response and X as the predictor. Write down the fitted model. Plot the data and superimpose the regression line. Construct a residual plot for this model and comment on the validity of the usual model assumptions.

2. (8 marks) Now fit a simple linear regression with the logarithm of Y as the response and X as the predictor. Write down the fitted model in the form "Y = ..." (that is, in terms of the original response, not the transformed response). Plot the transformed data and superimpose

the regression line. Construct a residual plot for this model and comment on the validity of the usual model assumptions.

3. (8 marks) Now fit a simple linear regression model with Y as the response and X squared as the predictor. Write down the fitted model in the form "Y = ..." (that is, in terms of the original response, not the transformed response). Plot the transformed data and superimpose the regression line. Construct a residual plot for this model and comment on the validity of the usual model assumptions.

4. (6 marks) A new data point has X value 2.30. What would you predict the Y of this X to be using each of the models fit above. Find a 99% interval for the Y of this X using each of the models fit above. Comment on which prediction is likely to be the most appropriate.

(60 total marks)