STAT2008/STAT2014/STAT6014

Tutorial 4

Question 1. The data file **Lubricant.csv** (available on Wattle) contains 53 measurements of the viscosity of a particular lubricating agent at various temperatures and pressures. The names of the three variables in the data are **viscos, pressure** and **tempC**. In tutorial 2, we have already used this data to fit a simple linear regression with viscosity as the response and pressure as the predictor variable. Based on this fitted model, please answer the following questions.

(a) Find a 90% confidence interval of the mean of viscosity at a pressure of 1,000.

(b) Find a 90% prediction interval of viscosity at a pressure of 1,000.

Question 2. The data file **auscars.csv** contains data gathered by the NRMA on 62 different makes and models of automobiles selling in Australia in 1991. Nine different variables were measured for each type of car, and the data set contains the names of these variables.

(a) Download and view the data file. Use the *read.csv()* command to read the data into a data frame called auscars in R. Type the name of this new data frame auscars in R and check that the data has been read in correctly.

(b) Fit a simple linear regression with the variable L.100k (which measures the fuel efficiency of the car in litres needed to travel 100 kilometres) as the response and Weight (which measures the unladen weight of the car in kilograms) as the predictor variable. What is the equation of the fitted regression line? What are the standard errors of the parameter estimates? Plot the data and superimpose the regression line, making sure that the limits on the y-axis range from 5 up to 20.

(c) Check the model assumptions and find if there are any unusual data points.

(d) Do you think that there is a significant relationship between the weight of a car and its fuel efficiency? If so, what is the nature of that relationship?

(e) Calculate the coefficient of determination, R^2. What is its interpretation?

(f) Calculate a 95% confidence interval for the parameter β_0. Is this a useful interval?

(g) Calculate a 95% confidence interval for the expected fuel efficiency of cars weighing 1800 kilograms. Calculate a corresponding 95% prediction interval for a single such car.

(h) Create a vector containing values spanning the range of Weight in the dataset using seq(min(Weight), max(Weight), 10) and calculate the predicted values of fuel efficiency and their standard errors for each of the values in this vector. Create two vectors one containing all the upper endpoints of 95% confidence intervals for each of these expected responses and the other containing all the lower endpoints. Draw both of these vectors as lines on the scatterplot; in other words, connect all the upper endpoints together and all the lower endpoints together. Repeat this procedure for 95% prediction intervals. What do you notice about these curves?

Question 3. The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns were counted on election night it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was carried by the democrat nominee, Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that the Republican nominee George W. Bush, had carried Florida and won the presidency. Gore called Bush to concede. While on route to his concession speech, though, the Florida count changed rapidly in his favour. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1738 and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush's lead to be less than 400. Meanwhile, angry Democrat voters in Palm Beach County complained that a confusing "butterfly" lay-out ballet caused them to accidentally vote for the Reform Party candidate Pat Buchanan instead of Gore. The ballot, shown below, listed presidential candidates on both a left-hand and right-hand page. Voters were to register their vote by punching the circle corresponding to their choice, from the column of circles between the pages. It was suspected that since Bush's name was listed first on the left-hand page, Bush voters likely selected the first circle. Since Gore's name was listed second on the left-hand page, many voters – who already new who they wished to vote for – did not bother examining the right-hand side of the ballet and consequently selected the second circle in the column; the one actually corresponding to Buchanan. Two pieces of evidence supported this claim: Buchanan had an unusually high percentage of the vote in that county, and an unusually large number of ballots (19000) were discarded because voters had marked two circles (possibly inadvertently voting for Buchanan and then trying to correct the mistake by then voting for Gore.)

The file *vote.csv* contains data on the number of votes for Buchanan and Bush in all 67 counties in Florida.

(a) Produce a plot of the number of Buchanan votes versus the number of Bush votes and another plot for the log of these two variables. Does the log-log transformation appear better for performing a simple linear regression?

(b) Analyse the data without the Palm Beach County results to obtain an equation for predicting Buchanan votes from Bush votes.

(c) Use the residual and QQ-plot from the regression in (b) to check the adequacy of the model.

(d) Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach – assuming the relationship is the same in this county as the others.

(e) Comment on the result in (d) given that Buchanan actually received 3407 votes in Palm Beach County.

DISPLAY 8.22 Confusing ballot in Palm Beach County, Florida