Regression Modelling

(STAT2008/STAT4038/STAT6038)

Tutorial 1 – Simple Linear Regression

Question One

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**.

(a) Use **lm()** to perform a simple linear regression with viscosity as the response and pressure as the predictor variable. What are the least-squares estimates of the slope and intercept? What are their standard errors? Check the standard error of the slope coefficient by using **R** to calculate SS_{Error} and S_{xx} and then put these values into the appropriate formula. Give full details of a *t*-test on the estimated slope coefficient. Find a *p-value* for the test statistic.

(b) Plot viscosity against pressure and use **abline()** to superimpose the estimated regression line. Use the estimated coefficients of the regression line to predict what the viscosity of the lubricant would be at a pressure of 1,000? [Hint: one way to do this is to create a suitable vector of new *x* values and use the vector multiplication operator (**%*%**) in **R** to multiply the vector of estimated coefficients by this new vector.] Also predict what the viscosity of the lubricant would be at a pressure of 10,000? Locate these predictions on your plot and comment on whether or not they appear to be sensible predictions.

(c) Use **R** to find the means of both pressure and viscosity and check that together the two means form a point (called the centroid of the data) which is located on the estimated regression line.

(d) Use **anova()** to produce the Analysis of Variance table for the regression. What is the MS_{Error} for this regression? How many degrees of freedom are associated with this MS_{Error}? Have a look at the *F* statistic in the ANOVA table. Is the *p-value* the same as for the t-test in part (a)? Square the test statistic you got for the *t*-test in part (a) – is it the same as the *F* statistic in the ANOVA table?

(e) Use the **residuals()** and **fitted()** functions to produce a plot of the residuals versus the fitted values for this regression. Examine this plot closely. Do you think that the estimated regression model satisfies the underlying assumptions?

(f) Now for a more advanced question which will take a lot of R coding. Again plot viscosity against pressure, but this time use a different plotting symbol to indicate which value of temperature is associated with each data point. [Hint: use **type="n"** to start with a blank **plot** and then use the **points()** function and the **pch** option to add the points for the different levels of **tempC** – you will probably need to do some searching through the help files associated with the graphical parameters **help(par)**.] What do you notice from your plot? Fit separate simple linear regression lines for each level of temperature and add these to your plot (and possibly include a suitable **legend**). Do the slopes of these separate models appear to be the same or different?

Question Two

I (Ian McDermid) have been having problems with my hearing for over 20 years, but have only been wearing hearing aids for the last decade. A couple of years ago, I purchased some brand new hearing aids and decided to also complete an on-line program designed to improve my listening skills (at the suggestion of my audiologist, who arranged free access to the program).

I have only recently managed to finally complete the 11 daily sessions of the full program and here are my daily scores:

Day	Cumulative **Score**
0	0
1	195
2	351
3	503
4	683
5	847
6	1011
7	1193
8	1378
9	1561
10	1743
11	1925

Information provided about interpreting these scores is presented in the file **LACE_Results.pdf**, which is available on Wattle.

Can I reasonably conclude that I have done significantly better in the program than someone in the typical range (“Many people training with LACE get between 60 and 100 points per day”)?

Analyse the above data in **R** and conduct a test of some appropriately chosen hypotheses.

Question Three (optional extra – there will be no question like this one on the final exam)

Recall that any matrix, A, is called a projection if it satisfies the identities: $A^T = A$ and $A^2 = A$ (see page 9 of the lecture notes). Also, recall that the hat matrix is defined as $H = X(X^TX)^{-1}X^T$.

(a) Show that the hat matrix, H, and the matrix $I - H$ are projections.

(b) The data frame **protpreg** (from the Protein in Pregnancy example in lectures) has two columns, the first relating to the amount of a certain protein and the second to the number of weeks of gestation for a group of pregnant women. Create a vector named **gest** containing the gestation data. Now, create the design matrix, X by attaching an initial column of ones to gest using the command:
X <− cbind(rep(1, length(gest)), gest)

(c) Now, use **R**'s matrix multiplication capabilities to construct the hat matrix H. Multiply H by itself and construct its transpose to check that it is indeed a projection. Also, use matrix calculations to find the least-squares regression estimate $b = (X^TX)^{-1}X^TY$, where Y is the vector of response values in the first column of **protpreg**. Check the results against the values given for this problem in the example discussed in lectures.

Question Four (optional extra – there will be no question like this one on the final exam)

Recall (from page 10 of the lecture notes) the breakdown of the total sum of squares SS_{Total}, into the sum of $SS_{Regression}$ and SS_{Error}:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

(a) In demonstrating this breakdown, we used the fact that: $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) = 0$.
Show that this fact is indeed true.
[Hint: Recall that the residuals, $e_i = Y_i - \hat{Y}_i$, from a least-squares regression satisfy: $\sum_{i=1}^{n} e_i = 0$; $\sum_{i=1}^{n} x_i e_i = 0$.]

(b) Recall that the coefficient of determination, R^2, and the sample correlation coefficient, r, are defined as:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{1}{SS_{Total}}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SS_{Total}}}$$

Show that $R^2 = r^2$.

Question Five (optional extra – there will be no question like this one on the final exam)

Refer to page 12 of the lecture notes. Show that the expectation of the mean square for the regression is:

$$E\left(MS_{Regression}\right) = E\left\{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2\right\} = \sigma^2 + \beta_1^2 S_{xx}$$

[Hint: Recall that $\hat{Y} = b_0 + b_1 x_i$; $b_0 = \bar{Y} - b_1\bar{x}$; and for any random variable Z, $E(Z^2) = Var(Z) + \{E(Z)\}^2$.]