# Linear Models. List 3

1. a) Use R to find the critical value that you would use for a two-tailed t signicance test with $\alpha = 0.05$ and 10 degrees of freedom. Call this value tc.

   b) Use R to find the critical value that you would use for an F significance test with $\alpha = 0.05$, one degree of freedom in the numerator and 10 degrees of freedom in the denominator. Call this value Fc.

   c) Verify that the square of tc is Fc.

2. Here you have the part of ANOVA table:

| | $df$ | $SS$ |
|---|---|---|
| $Model$ | 1 | 100 |
| $Error$ | 20 | 400 |

   a) How many observations do you have in your file ?

   b) Calculate the estimate of $\sigma$.

   c) Test if $\beta_1$ is equal to zero. (Give the test statistic with the numbers of degrees of freedom and the conclusion).

   d) What proportion of the variation of the response variable is explained by your model ?

   e) What is the sample correlation coefficient between your response and explanatory variables ?

3. For this and the next problem you will use the data set table1_6.txt, containing the grade point average (GPA) [second column], score on a standard IQ test [third column], gender and a score on the Piers-Harris Childrens Self-Concept Scale (a psychological test, fifth column) for 78 seventh-grade students.

   a) Use a simple regression model to describe the dependence of gpa on the results of iq test. Report the fitted regression equation and $R^2$. Test the hypothesis that gpa is not correlated with iq : give the test statistic, p-value and the conclusion in words.

   b) Predict gpa for a student whose iq is equal to 100. Report 90% prediction interval.

   c) Draw a band for 95% prediction intervals (i.e. join the limits of the prediction intervals with the smooth line). How many observations fall outside this band ?

4. a) Use a simple regression model to describe the dependence of gpa on the results of Piers-Harris test. Report the fitted regression equation and $R^2$.

   b) Test the hypothesis that gpa is not correlated with Piers-Harris score : give the test statistic, p-value and the conclusion in words.

   c) Predict gpa for a student whose Piers-Harris score is equal to 60. Report 90% prediction interval .

   d) Draw a band for 95% prediction intervals. How many observations fall outside this band ?

   e) Which of the two variables : result of iq test or result of Piers-Harris test, is a better predictor of gpa ?

5. For the next two problems you will use the copier maintenance data,ch01pr20.txt, discussed in class.

a) Verify that the sum of the residuals is zero.

b) Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.

c) Plot the residuals versus the order in which the data appear in the data file and briefly describe the plot noting any unusual patterns or points.

d) Examine the distribution of the residuals by getting a histogram and a normal probability plot. What do you conclude ?

6. Change the data set by changing the value of service time for the first observation from 20 to 2000.

   a) Run the regression with changed data and make a table comparing the results of this analysis with the results of the analysis of the original data. Include in the table the following: fitted equation, t-test for the slope with P-value, $R^2$, and the estimate of $\sigma^2$. Briefly summarize the differences.

   b) Repeat points (b), (c) and (d) from problem 5 above on the modified data set and show the unusual observation on each of these plots.

   For next six problems you will use the solution concentration data ch03pr15.txt. The first column gives the values of the solution concentration and the second column gives the time.

7. Run the linear regression with time as the explanatory variable and the solution concentration as the response variable. Summarize the regression results by giving the fitted regression equation, the value of $R^2$ and the results of the significance test for the null hypothesis that the solution concentration does not depend on time (formulate the statistical model, give null and alternative hypotheses in terms of the model parameters, test statistic with degrees of freedom, P-value, and brief conclusion in words).

8. Plot the solution concentration versus time. Add a fitted regression line and a band for 95% prediction intervals. What do you conclude ? Calculate the correlation coefficient between the observed and predicted value of the solution concentration.

9. Use the Box-Cox procedure to find an appropriate transformation for the solution concentration.

10. Construct a new response variable by taking the log of the solution concentration (define $logy = \log(Y)$). Repeat points 7 and 8 of this homework with logy as the response variable (and time as the explanatory variable). Summarize your results.

11. Plot the solution concentration versus time. Add a regression curve and a band for 95% prediction intervals based on the results obtained in point 10. Compare to the graph obtained in point 8. Calculate the correlation coefficient between the observed solution concentration and the predictions based on the model from point 10.

12. Construct a new explanatory variable $t1 = time^{-1/2}$. Repeat points 10 and 11 of this exercise using the regression model with the solution concentration as the response variable and $t1$ as the explanatory variable. Summarize your results. Which model seems to be the best ?

Małgorzata Bogdan