Statistics for Engineers, 4C3/6C3

Assignment 6

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 09 March 2011**

Assignment objectives

- To become more comfortable using R to fit, interpret and manipulate least squares models.
- The questions in this assignment are typical of the exploratory/learning type questions that will be in the take-home midterm.

Question 1 [1.5]

Use the mature cheddar cheese data set for this question.

1. Choose any x-variable, either `Acetic` acid concentration (already log-transformed), `H2S` concentration (already log-transformed), or `Lactic` acid concentration (in original units) and use this to predict the `Taste` variable in the data set. The `Taste` is a subjective measurement, presumably measured by a panel of tasters.

   Prove that you get the same linear model coefficients, R^2, S_E and confidence intervals whether or not you first mean center the x and y variables.

2. What is the level of correlation between each of the x-variables. Also show a scatterplot matrix to learn what this level of correlation looks like visually.

   - Report your correlations as a 3×3 matrix, where there should be 1.0's on the diagonal, and values between -1 and $+1$ on the off-diagonals.

3. Build a linear regression that uses all three x-variables to predict y.

   - Report the slope coefficient and confidence interval for each x-variable
   - Report the model's standard error. Has it decreased from the model in part 1?
   - Report the model's R^2 value. Has it decreased?

Question 2 [2.5]

In this question we will revisit the bioreactor yield data set and fit a linear model with all x-variables to predict the yield.

1. Provide the interpretation for each coefficient in the model, and also comment on its confidence interval when interpreting it.
2. Compare the 3 slope coefficient values you just calculated, to those from the last assignment:

   - $\hat{y} = 102.5 - 0.69T$, where T is tank temperature
   - $\hat{y} = -20.3 + 0.016S$, where S is impeller speed
   - $\hat{y} = 54.9 - 16.7B$, where B is 1 if baffles are present and $B = 0$ with no baffles

   Explain why your coefficients do not match.

3. Are the residuals from the multiple linear regression model normally distributed?

4. In this part we are investigating the variance-covariance matrices used to calculate the linear model.
   (a) First center the x-variables and the y-variable that you used in the model.

   Note: feel free to use MATLAB, or any other tool to answer this question. If you are using R, then you will benefit from this page in the R tutorial. Also, read the help for the `model.matrix(...)` function to get the $\mathbf{X}$-matrix. Then read the help for the `sweep(...)` function, or more simply use the `scale(...)` function to do the mean-centering.

   (b) Show your calculated $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ variance-covariance matrices from the centered data.

   (c) Explain why the interpretation of covariances in $\mathbf{X}^T\mathbf{y}$ match the results from the full MLR model you calculated in part 1 of this question.

   (d) Calculate $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and show that it agrees with the estimates that R calculated (even though R fits an intercept term, while your $\mathbf{b}$ does not).

5. What would be the predicted yield for an experiment run without baffles, at 4000 rpm impeller speed, run at a reactor temperature of 90 °C?

Question 3 [3]

In this question we will use the LDPE data which is data from a high-fidelity simulation of a low-density polyethylene reactor. LDPE reactors are very long, thin tubes. In this particular case the tube is divided in 2 zones, since the feed enters at the start of the tube, and some point further down the tube (start of the second zone). There is a temperature profile along the tube, with a certain maximum temperature somewhere along the length. The maximum temperature in zone 1, `Tmax1` is reached some fraction `z1` along the length; similarly in zone 2 with the `Tmax2` and `z2` variables.

We will build a linear model to predict the `SCB` variable, the short chain branching (per 1000 carbon atoms) which is an important quality variable for this product. Note that the last 4 rows of data are known to be from abnormal process operation, when the process started to experience a problem. However, we will pretend we didn't know that when building the model, so keep them in for now.

1. Use only the following subset of x-variables: `Tmax1`, `Tmax2`, `z1` and `z2` and the y variable = `SCB`. Show the relationship between these 5 variables in a scatter plot matrix.

   Use this code to get you started (make sure you understand what it is doing):

   ```
   LDPE <- read.csv('http://datasets.connectmv.com/file/ldpe.csv')
   subdata <- data.frame(cbind(LDPE$Tmax1, LDPE$Tmax2, LDPE$z1, LDPE$z2, LDPE$SCB))
   colnames(subdata) <- c("Tmax1", "Tmax2", "z1", "z2", "SCB")
   ```

   Using bullet points, describe the nature of relationships between the 5 variables, and particularly the relationship to the y-variable.

2. Let's start with a linear model between `z2` and `SCB`. We will call this the `z2` model. Let's examine its residuals:
   (a) Are the residuals normally distributed?

   (b) What is the standard error of this model?

   (c) Are there any time-based trends in the residuals (the rows in the data are already in time-order)?

   (d) Use any other relevant plots of the predicted values, the residuals, the x-variable, as described in class, and diagnose the problem with this linear model.

   (e) What can be done to fix the problem? (You don't need to implement the fix yet).

3. Show a plot of the hat-values (leverage) from the `z2` model.
   (a) Add suitable horizontal cut-off lines to your hat-value plot.

   (b) Identify on your plot the observations that have large leverage on the model

(c) Remove the high-leverage outliers and refit the model. Call this the `z2.updated` model

(d) Show the updated hat-values and verify whether the problem has mostly gone away

Note: see the R tutorial on how to rebuild a model by removing points

4. Use the `influenceIndexPlot(...)` function in the `car` library on both the `z2` model and the `z2.updated` model. Interpret what each plot is showing for the two models. You may ignore the *Bonferroni p-values* subplot.

END