Statistics for Engineering, 4C3/6C3

Assignment 5

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 08 March 2013**

Note: Assignment objectives

- Build least squares models in R.
- Extract useful information about the model outputs.
- Investigate and understand multiple linear regression (MLR) models.

Question 1 [12]

No need to use software. Question from the final exam, 2011.

Some data were collected from tests where the compressive strength, x, used to form concrete was measured, as well as the intrinsic permeability of the product, y. There were 16 data points collected. The mean x-value was $\overline{x} = 3.1$ and the variance of the x-values was 1.52. The average y-value was 40.9. The estimated covariance between x and y was -5.5.

The least squares estimate of the slope and intercept was: $y = 52.1 - 3.6x$.

1. What is the expected permeability when the compressive strength is at 5.8 units?
2. Calculate the 95% confidence interval for the slope if the standard error from the model was 4.5 units. Is the slope coefficient statistically significant?
3. Provide a rough estimate of the 95% prediction interval when the compressive strength is at 5.8 units (same level as for part 1). What assumptions did you make to provide this estimate?
4. Now provide a more accurate, calculated 95% prediction confidence interval for the previous part.

Question 2 [10]

Use the gas furnace data from the website to answer these questions. The data represent the gas flow rate (centered) from a process and the corresponding CO_2 measurement.

1. Make a scatter plot of the data to visualize the relationship between the variables. How would you characterize the relationship?
2. Calculate the variance for both variables, the covariance between the two variables, and the correlation between them, $r(x, y)$. Interpret the correlation value; i.e. do you consider this a strong correlation?
3. Now calculate a least squares model relating the gas flow rate as the x variable to the CO_2 measurement as the y-variable. Report the intercept and slope from this model.
4. Report the R^2 from the regression model. Compare the squared value of $r(x, y)$ to R^2. What do you notice? Now reinterpret what the correlation value means (i.e. compare this interpretation to your answer in part 2).
5. Switch x and y around and rebuild your least squares model. Compare the new R^2 to the previous model's R^2. Is this result surprising? How do interpret this?

Question 3 [15]

In this question we consider the bioreactor yield data set and fit a linear model using all x-variables simultaneously to predict the yield.

1. Provide the interpretation for each coefficient in the model, and also comment on each one's confidence interval when interpreting it.
2. Compare the 3 slope coefficient values the case when you regress yield onto each x-variable on its own.:
   - $\hat{y} = 102.5 - 0.69T$, 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 the R tutorial on the course website. 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 4 [8]

The grades from a recent midterm exam are available, as well as the time taken by the student to write the exam. It was an "infinite" time midterm, so there was no time pressure to finish within the allocated period.

The data are available on the data set website.

1. Use the data to confirm whether or not the amount of time used to write the test has an influence on the grade value obtained.
2. Do the regular least squares assumptions apply in this instance? Assess the assumptions individually.

END