Statistics for Engineering, 4C3/6C3, 2012

Assignment 6

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 12 February 2012, at 16:00**

Question 1 [3]

In a previous assignment you used an ordinary (*unpaired*) test of differences when a large sample of water was split in 22 portions, and the biochemical oxygen demand (BOD) was measured by the dilution method (11 times) and the manometric method (11 times). Here are the values again:

Dilution method	Manometric method
11	25
26	3
18	27
16	30
20	33
12	16
8	28
26	27
12	12
17	32
14	16

The confidence interval for the average difference of the two methods was calculated as: $-0.768 < \mu_{\text{M}} - \mu_{\text{D}} < 13.3$, showing that statistically there is no difference between the two methods. Though there is a practical difference, and the statistical lack of difference is due to an outlier.

The purpose of this question is for you to show that you can obtain the same result using a confidence interval from this least squares model:

$$y = b_0 + \gamma g_i$$

where $g_i = 0$ when the dilution method was used and $g_i = 1$ when the manometric method was used.

You may use any software to calculate the model and standard error for you, but you must show the confidence interval calculation for γ by hand.

Question 2 [5]

In class we showed, given a x_{new} value and the linear model $y = b_0 + b_1 x$, that the prediction interval for $\hat{y}_{\text{new}}$ is:

$$\hat{y}_{\text{new}} \pm c_t \sqrt{V\{\hat{y}_{\text{new}}\}}$$

where c_t is the critical t-value, for example at the 95% confidence level.

Use the distillation column data set and with y as `VapourPressure` (units are kPa) and x as `TempC2` (units of degrees Farenheit) fit a linear model. Calculate and plot the prediction interval at the 95% confidence level, for vapour pressure at these temperatures: 410, 480, 530 °F. At each temperature compare the 95% prediction interval to the more conservative $2S_E$ prediction interval. Are these results expected?

Question 3 [12]

The percentage yield from bioreactor was investigated for a research project. Two variables were adjusted in the experiments: lactose purity and glucose purity.

Lactose	Glucose	Yield
90	85	51
94	85	79
90	90	72
94	90	94
88	87.5	50
96	87.5	99
92	82.5	49
92	92.5	89
92	87.5	73
92	87.5	74
92	87.5	64
92	87.5	65

1. Mean center the two x-variable vectors and combine them in matrix $\mathbf{X}$. Repeat the same for vector $\mathbf{y}$.
2. Now calculate $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$. Give a clear interpretation for each of these.
3. Let $\mathbf{b} = [b_L, b_G]$, then calculate $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ *by hand*. Interpret each of these slope coefficients in $\mathbf{b}$ in the context of the linear model, $y = b_L x_L + b_G x_G$; where y is yield, x_L is the lactose concentration, and x_G is the glucose concentration.
4. Build the linear regression model, $y = b_0 + b_L x_L + b_G x_G$ (note the intercept) in R, or other statistical packages.
   (a) What is the intercept value?
   (b) Calculate this same intercept value by hand, from the values given in the answer to part 1 of this question.
   (c) Are the slope terms calculated by the software in agreement with yours (from part 3)?
5. Give confidence intervals for each of the slope coefficients at the 95% level.
6. Are the model's residuals normally distributed?

Question 4 [15]

Data on the course website were collected to predict the blending efficiency of an industrial mixer from 4 variables: particle size (x_P), mixer diameter (x_D), mixer rotational speed (x_R) and blending time (x_T): $y = b_0 + b_P x_P + b_D x_D + b_R x_R + b_T x_T$.

1. Calculate the variance-covariance matrix when the four x-variables are combined into an $\mathbf{X}$ matrix. What does this matrix tell you about this data set?
2. Fit the above linear model and determine which coefficients are significant at the 90% or higher level.
3. Interpret the slope coefficient for each variable. If these blending experiments were very expensive to conduct, based on the results from this part and the previous part 2, which variables could you disregard in the future?
4. Are the residuals normally distributed? If you identify any outliers, list the predicted value of y and the actual value of y for each outlier.
5. Omit any outliers and rebuild the model. Do any of the previous answers change? Are the residuals more normally distributed now?

Use the software tutorial on the course website to help you with this question.
