Statistics for Engineering, 4C3/6C3

Assignment 5

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 12 March 2014**

Note: Once again, I strongly recommend you submit this assignment electronically (see instructions on the course website), so that you can practice using the electronic system for the course project.

Note: For this assignment you will benefit from studying the R tutorial on vectors and matrices.

Question 1 [12]

A company has 3 reactors that are identical. Typical production schedules split the raw material equally between the 3 reactors. Data on the website contain the brittleness values of the product produced from the three reactors for the past few days.

1. Compare the brittleness values between reactors `TK104` and `TK107`, using a regular test for differences we learned about earlier. Feel free to use the `t.test(...)` function, but make sure you can get the same results by hand.
2. What is the interpretation of your confidence interval?
3. Next, build a least squares model where the brittleness values are predicted using a single integer variable, d, which is a coded as 0 for `TK104`, and coded as 1 for `TK107`. *Hint* use the `c(...)` function in R to combine vectors, and use the `numeric(..)` function to create vectors

   Report the R^2 and standard error values for the model.
4. Calculate the slope coefficient for variable d and report a confidence interval for it.

   What is your interpretation of the confidence interval?

Solution

1. Loading and showing the 23 values for each vector, we see that there are some missing entries in the `TK104` column. These can be ignored, so there are 20 values for `TK104` and 23 values for `TK107`. Note that this is not a paired test, so there can be unequal number of points in each group:

   ```
   colMeans(brit, na.rm=T)
      TK104    TK105    TK107
   421.0000 472.1905 470.0870
   ```

   shows that `TK104` average brittleness is 421 units, compared to 470 for `TK107`.

   ```
   brit <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')
   t.test(brit$TK107, brit$TK104, var.equal=TRUE)
   ```

   The output reads:

   ```
   data:  brit$TK107 and brit$TK104
   t = 1.4056, df = 41, p-value = 0.1674
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval:
    -21.43835 119.61226
   ```

   The confidence interval calculated by hand should be identical to this. Do not use short-cut functions in R, such as `t.test(...)`, without trying them out manually at least once or twice.

2. The interval indicates there is no statistical difference between the average brittleness values coming from `TK104` or from `TK107`. However there is a bias to one side, given the asymmetry of the interval from -21.4 to 119.6.

3. The least squares model is calculated using:

```
y <- c(brit$TK104, brit$TK107)
x <- c(numeric(23), numeric(23)+1)  # read the help for the "numeric()" function
```

   where the $R^2 = 0.05$ and the standard error is 114 brittleness unit.

   This is a poor model, indicating that we are not able to predict the brittleness well, when only told which reactor the material was processed in.

4. The slope coefficient for `d` is found from `summary(lm(y ~ x))` is 49.0, but it has a standard error of 35 units, very high when compared to the coefficient itself. The confidence interval:

```
confint(lm(y~x))
```

   and the output reads:

```
                2.5 %   97.5 %
(Intercept) 369.42079 472.5792
x           -21.43835 119.6123
```

   is the same range as before, from -21.4 to 119.6 (you might have the signs flipped: that is OK, simply swap "A" and "B" samples around).

   This indicates we get the same result from the least squares model, as we get from the regular confidence interval. *Note*: the purpose of this model was not to make a prediction of brittleness, the purpose was to see if there is a *statistically significant* effect of reactor on the brittleness.

   The model conclusively shows the reactors are statistically identical, though we might have some reservations from a practical perspective.

Question 2 [12]

A factorial experiment is used to investigate settings to *minimize* the production of an unwanted side product. Two factors being investigated are called **A** and **B** for simplicity, but are:

- **A** = reaction temperature: low level was 440 K, and high level was 450 K
- **B** = amount of surfactant: low level was 8 kg, high level was 12 kg

A full factorial experiment was run, randomly, on the same batch of raw materials, in the same reactor. The recorded amount, in grams, of the side product was:

Experiment	Run order	A	B	Side product formed
1	2	440 K	8 kg	89 g
2	1	450 K	8 kg	268 g
3	3	440 K	12 kg	179 g
4	4	450 K	12 kg	448 g

1. Write out a least squares model that will predict the amount of side product formed given the settings for **A**, **B** and the **AB** interaction.

2. Write out the $\mathbf{X}$ matrix and $\mathbf{y}$ vector that can be used to estimate the model coefficients using the equation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

3. Solve for the coefficients of your linear model, by using $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ directly.

   Show your calculations that you've done by hand.

   Feel free though to compare your answer to R, Minitab, Excel, or other software.

4. Give a clear interpretation of the slope coefficient of **A** and the slope coefficient for **B**.
5. What happens when you try to calculate confidence intervals? Explain clearly.

Solution

1. The model would have the form:

$$y = b_0 + b_A x_A + b_B x_B + b_{AB} x_{AB} + e$$

2. The matrices and vectors to solves this least squares model are:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & +1 \\ 1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & +1 & +1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_A \\ b_B \\ b_{AB} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

$$\begin{bmatrix} 89 \\ 268 \\ 179 \\ 448 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & +1 \\ 1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & +1 & +1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_A \\ b_B \\ b_{AB} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

3. Using the above matrices we can calculate $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, even by hand!

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} +89 + 268 + 179 + 448 \\ -89 + 268 - 179 + 448 \\ -89 - 268 + 179 + 448 \\ +89 - 268 - 179 + 448 \end{bmatrix} = \begin{bmatrix} 984 \\ 448 \\ 270 \\ 90 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 246 \\ 112 \\ 67.5 \\ 22.5 \end{bmatrix}$$

Please note, you must answer this question by hand, so you understand why the shortcut methods work.

4. The coefficient for **A** is $b_A = 112$, indicates more side product is produced, about 112 g on average, for a one unit change in x_A. A unit change in x_A represents a 5 K increase (recall a 2 unit change in x_A is from -1 to +1, or from 440K to 450K).

   The coefficient for **B** is $b_A = 67.5$, indicates more side product is produced, about 67.5 g on average, for a one unit change in x_B. A unit change in x_B represents a 2kg increase (recall a 2 unit change in x_B is from -1 to +1, or from 8kg to 12kg surfactant).

5. Confidence intervals cannot be found, as there are no degrees of freedom.

Question 3 [12]

We considered data from a lab-scale bioreactor, y, earlier in the course. In class, we looked at an example where the reactor temperature, batch duration, impeller speed and reactor type (one with with baffles and one without) were used to judge the effect on yield, y.

Here are the data once again, and on the website:

Temp = T **[°C]**	**Duration =** d **[minutes]**	**Speed =** s **[RPM]**	**Baffles =** b **[Yes/No]**	**Yield =** y **[g]**
82	260	4300	No	51
90	260	3700	Yes	30
88	260	4200	Yes	40
86	260	3300	Yes	28
80	260	4300	No	49
78	260	4300	Yes	49
82	260	3900	Yes	44
83	260	4300	No	59
64	260	4300	No	60
73	260	4400	No	59
60	260	4400	No	57
60	260	4400	No	62
101	260	4400	No	42
92	260	4900	Yes	38

1. Demonstrate that you get the same regression slope when building the following two models:

   (a) a model using only temperature to predict yield;

   (b) as in part (a) above, but first mean center the temperature vector;

   (c) as in part (b) above, but also mean center the yield vector.

2. Next, build a linear model to predict the yield from all remaining variables. See the R tutorial for help to build and interpret linear models containing integer variables.

   Show your model, and interpret each variable in the model. If you are using R, then the `confint(...)` function will be helpful as well.

3. What is the predicted yield for a new batch, operating at 95°C for 260 minutes, at a speed of 4000 rpm in a tank with no baffles?

Solution

1. Using the following code, for example, you will find all slope coefficients are the same:

   ```
   bio <- read.csv('http://datasets.connectmv.com/file/bioreactor-yields.csv')
   summary(lm(bio$yield ~ bio$temperature))    # -0.685

   t.center <- bio$temperature - mean(bio$temperature)
   summary(lm(bio$yield ~ t.center))           # -0.685

   y.center <- bio$yield - mean(bio$yield)
   summary(lm(y.center ~ t.center))            # -0.685
   ```

   The R^2 and standard error values are also identical.

2. After importing the data, just make sure the `baffles` variable is imported as a factor. Then build the model as usual. The computer output below shows the linear model's coefficients.

   ```
   bio <- read.csv('http://datasets.connectmv.com/file/bioreactor-yields.csv')
   attach(bio)
   summary(bio)
   is.factor(baffles)
   # [1] TRUE

   model <- lm(yield ~ speed + baffles + temperature )
   summary(model)

   # Call:
   ```

```
# lm(formula = yield ~ speed + baffles + temperature)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -5.5521 -3.2543 -0.4356  2.2953  8.1519
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 52.483652  18.421511   2.849  0.01728 *
# speed        0.008711   0.003757   2.319  0.04288 *
# bafflesYes  -9.090700   3.048811  -2.982  0.01377 *
# temperature -0.470997   0.119242  -3.950  0.00273 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.651 on 10 degrees of freedom
# Multiple R-squared: 0.8659,   Adjusted R-squared: 0.8256
# F-statistic: 21.52 on 3 and 10 DF,  p-value: 0.0001108
```

The confidence intervals for each variable is significant at the 95% level. The duration variable must be omitted from the model, because it has no variation. While it might affect the yield, there is no variability in this data set to assess that.

- $0.00034 \leq b_{\text{speed}} \leq 0.017$: a 100rpm increase in impeller speed serves to increase yield by 0.87g on average, keeping all other variables constant
- $-15.9 \leq b_{\text{baffles}} \leq -2.30$: the use of baffles decreases yield, on average, by 9.1g, keeping all other variables constant
- $-0.74 \leq b_{\text{temp}} \leq -0.21$: each one degree increase in temperature lowers yield by 0.47g on average, keeping all other variables constant
- We cannot say anything about the effect of batch duration

3. Using the above model output, the predicted yield at these conditions would be

$$y = 52.5 + 0.0087(4000) - 9.09(0) - 0.471(95) = 42.5\,\text{g}$$