# Statistics for Engineers, 4C3/6C3

# Assignment 5

**Kevin Dunn, dunnkg@mcmaster.ca** **Due date: 16 February 2011**

---

**Assignment objective: explore and understand least squares models a bit more.**

## Question 1 [2]

Use the bioreactor data, which shows the percentage yield from the reactor when running various experiments where temperature was varied, impeller speed and the presence/absence of baffles were adjusted.

1. Build a linear model that uses the reactor temperature to predict the yield. Interpret the slope and intercept term.
2. Build a linear model that uses the impeller speed to predict yield. Interpret the slope and intercept term.
3. Build a linear model that uses the presence (represent it as 1) or absence (represent it as 0) of baffles to predict yield. Interpret the slope and intercept term.

   *Note*: if you use R it will automatically convert the `baffles` variable to 1's and 0's for you. If you wanted to make the conversion yourself, to verify what R does behind the scenes, try this:

   ```
   # Read in the data frame
   bio <- read.csv('http://datasets.connectmv.com/file/bioreactor-yields.csv')

   # Force the baffles variables to 0's and 1's
   bio$baffles <- as.numeric(bio$baffles) - 1
   ```

4. Which variable(s) would you change to boost the batch yield, at the lowest cost of implementation?
5. Use the `plot(bio)` function in R, where `bio` is the data frame you loaded using the `read.csv(...)` function. R notices that `bio` is not a single variable, but a group of variables, i.e. a data frame, so it plots what is called a *scatterplot matrix* instead. Describe how the scatterplot matrix agrees with your interpretation of the slopes in parts 1, 2 and 3 of this question.

## Question 2 [2]

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. **600-level**: 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 [1.5]

A new type of thermocouple is being investigated by your group. These devices produce an *almost* linear voltage (millivolt) response at different temperatures. In practice though it is used the other way around: use the millivolt reading to predict the temperature. The process of fitting this linear model is called *calibration*.

1. Use the following data to calibrate a linear model:

| Temperature [K] | 273 | 293 | 313 | 333 | 353 | 373 | 393 | 413 | 433 | 453 |
|---|---|---|---|---|---|---|---|---|---|---|
| Reading [mV] | 0.01 | 0.12 | 0.24 | 0.38 | 0.51 | 0.67 | 0.84 | 1.01 | 1.15 | 1.31 |

   Show the linear model and provide the predicted temperature when reading 1.00 mV.

2. Are you satisfied with this model, based on the coefficient of determination ($R^2$) value?

3. What is the model's standard error? Now, are you satisfied with the model's prediction ability, given that temperatures can usually be recorded to an accuracy of $\pm 0.5$ K with most inexpensive thermocouples.

4. What is your (revised) conclusion now about the usefulness of the $R^2$ value?

**Note**: This example explains why I don't use the terminology of *independent* and *dependent* variables in this course. Here the temperature truly is the independent variable, because it causes the voltage difference that we measure. But the voltage reading is the independent variable in the least squares model. The word *independent* is being used in two different senses (its English meaning *vs* its mathematical meaning), and this can be misleading.

## Question 4 [1]

1. Use the linear model you derived in Question 2, where you used the gas flow rate to predict the $CO_2$ measurement, and construct the analysis of variance table (ANOVA) for the dataset. Use your ANOVA table to reproduce the residual standard error, $S_E$ value, that you get from the R software output.

   Go through the R tutorial to learn how to efficiently obtain the residuals and predicted values from a linear model object.

2. Also for linear model in Question 2, verify whether the residuals are normally distributed.

3. Use the linear model you derived in Question 3, where you used the voltage measurement to predict the temperature, and construct the analysis of variance table (ANOVA) for that dataset. Use your ANOVA table to reproduce the residual standard error, $S_E$ value, that you get from the R software output.

---

END