Grading: 400-level students: 35 + 5 extra points for Q3; 600-level students: 40 + 5 extra points for Q3

Question 1 [20]

Consider the bioreactor data set that you saw in the quiz and class 07B, http://yint.org/bioreactor

Load the data set, look at the raw data, and also use the **`summary(...)`** command:

```
bio <- read.csv('http://openmv.net/file/bioreactor-yields.csv')
summary(bio)
```

1. What type of variable is the "**`bio$baffles`**" variable?
2. Look at the raw data in R (**`> bio`**). Can you tell if baffles increases or decreases the yield? Use this command to confirm your answer: **`plot(bio$baffles, bio$yield)`**. What plot is that?
3. So what sign would you expect for coefficient g in a regression model $y_i = b_0 + gd_i$ where d_i represents whether baffles are present or not.
4. Verify that now in R: **`mod.1 <- lm(bio$yield ~ bio$baffles)`** and interpret the value of g **and** also interpret the confidence interval for that categorical variable:
5. What is the S_E of the regression model where you only use **`baffles`** to predict yield?
6. Now add the **`temperature`** variable to the model. What is the S_E value now?

   `mod.2 <- lm(bio$yield ~ bio$baffles + bio$temperature)`
7. Next, add all input variables to the model and report the standard error.
   `mod.4 <- lm(bio$yield ~ bio$baffles + bio$temperature + bio$duration + bio$speed)`
8. Comment on the standard error values in the models used in parts 5, 6 and 7 of this question.
9. What unusual feature do you notice in the **`summary(mod.4)`** and **`confint(mod.4)`** output. What has happened, and why?

Hint: read step 20, and step 21 of the R software tutorial.

Solution (Solution courtesy of: M. Lewandowski, M. Arshad and N. Rasanayagam)

1. Baffle is a categorical variable (Yes or No), which takes limited number of values.

2. Type of plot is a Box Plot. According to the boxplot it seems that when baffles are present the median is lower than when no baffles are present. This means when there are baffles, this decreases yield.

3. The coefficient will be negative since the yield decreases with the addition of baffles.

4. The value of g is -16.7, this indicates that 16.7% decrease in yield, on average, for having baffles while holding rest of the variables including temperature, duration and speed constant.

   The confidence interval for a categorical variable is -25.40651 bio$baffles(Yes) -8.01016. confidence interval for g does not span zero, therefore the baffle effect on yield (outcome variable) is statistically significant.

5. Residual standard error: 7.392 on 12 degrees of freedom

6. Residual standard error: 5.499 on 11 degrees of freedom

7. Residual standard error: 4.651 on 10 degrees of freedom

8. The standard error decreased (R squared increased) when more variables are added to the model. Each variable added allows for an improved fit (this is always true: adding more variables will keep the fit the same, or improve it). Every variable that is added though decreases the degrees of freedom by 1.

9. The `summary(..)` command gives more results including the predicted coefficients (intercepts and slopes) and the errors associated with the coefficients, whereas `confint(...)` command outputs the confidence interval for each variables and the intercept.

   The unusual feature for `summary(mod.4)` and `confint(mod.4)` shows that duration to run the bioreactor did not affect the yield in the least square models where the outputs of these command provided N/A for variable “`duration`” . Duration does not have an effect on the yield, because throughout the experiment (for all 14 experiments) it is kept constant. This is can be seen in the results below. Since there is no variation in duration, no slope, error, confidence interval can be calculated for the the batch duration.

Question 2 [15]
For a distillation column, it is well known that the column temperature directly influences the purity of the product, and this is used in fact for feedback control, to achieve the desired product purity. Use the distillation data set and build a least squares model that predicts **`VapourPressure`** from the temperature measurement, **`TempC2`**. Report the following values:

1. the slope coefficient, and describe what it means in terms of your objective to control the process with a feedback loop
2. the interquartile range and median of the model's residuals
3. the model's standard error
4. a confidence interval for the slope coefficient, and its interpretation.

Use any computer package to build the model and read these values off the computer output. (no code is required in your answer).

Solution courtesy of: K.Pennings, D.Bowie, B.Huzevka

1. The slope coefficient is -0.33 atm/F. With regards to process control, this means that in order to decrease the vapour pressure, the temperature at C2 should be increased. More information about the confidence interval for this slope is in part 4 below.
2. The IQR = Interquartile Range = 4.38 and the median = 0.0667.
3. S_E = 2.99 on 251 degrees of freedom
4. A 95% confidence interval is determined for this slope coefficient as -0.35 < slope < -0.31.

   This shows, at the 95% confidence level, the range within which we can expect to find the true slope coefficient. This range is remarkably narrow; i.e. our feedback controller gain is unlikely to change on either extreme. So we can likely design our control loop at the center point, and be sure it will work over the entire range of expected operation

Question 3: Solution courtesy of: Felipe Gomes. Other solutions ideas are possible.

For this problem we want to predict if we want to run the Mode 1 (all day heating) or Mode 2 (only night heating) for a garage based on the forecast. The threshold is that the garage temperature cannot go below 5℃.
In order to understand how the system works in these two operation modes, a multivariable linear model will constructed based on the data provided from observations during operation in both modes.

- Dependent Variable (y): Garage Temperature (in ℃)
- Independent Variable (x1): Corrected Outside Temperature (in ℃)

- Independent Variable (d): Operation Mode (Yes - if the heating is on for all day operation, No - if the heating operates only at night)

The resultant linear model is presented as:

$$y' = 13.48 + 0.329*x1 + 0.694*d$$

To build this model the values of x_1 were considered continuous and for d were used integer values (0=only night heating-Mode 2, 1=all day heating-Mode1). With a standard error of 0.981 and a R^2 of 0.688.

It is possible now to use this model to predict the garage temperature over a range of negative outside temperature. The Figure 6 shows this correlation for both operating conditions.

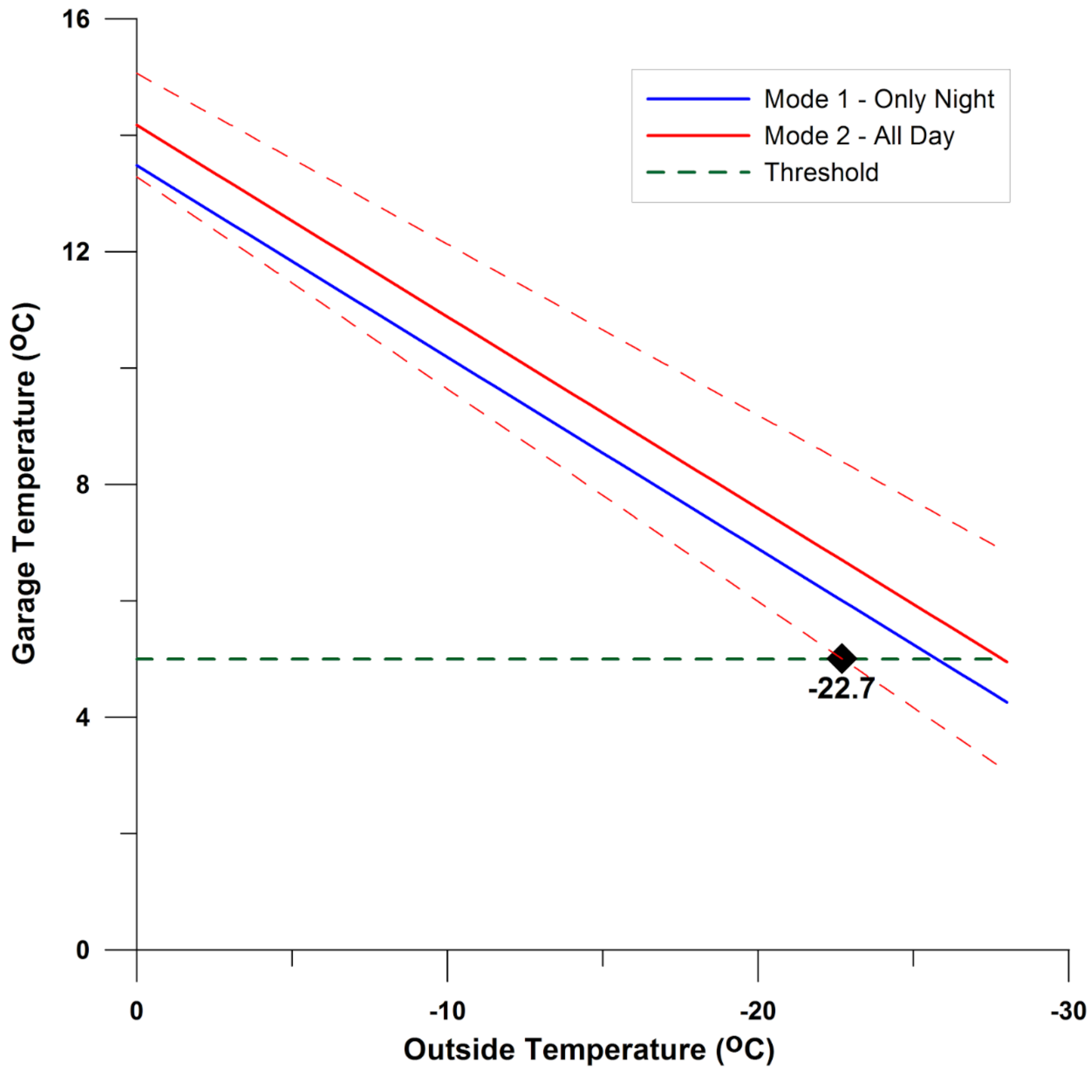

Fig. 6 - Fitted values of linear model for two operation modes (red=Mode 1 and blue=Mode 2) and Lower and Upper Bounds for the predicted model with a confidence interval of 95% (red dashed lines). Garage temperature threshold (dashed green line=5℃).

Considering the threshold of 5℃ as limit of the garage temperature to avoid freezing the water tubes, from the plot is possible to notice that this only would happen if the outside temperature was below -22.7 ℃, with 95% of confidence at the lower bound. As this condition is not predicted for the next 10-days weather forecast, the suggestion would be keeping the heating on **Mode 2, or only operating at night**.