# STAC32 Assignment 4

Due Tuesday November 25 4:00pm on Blackboard, or by e-mail to the instructor by the same deadline

1. A baseball is thrown upwards off the observation platform of a tall building. Every 0.2 seconds, the height of the ball above ground is recorded (until just before the ball hits the ground). The data are in `http://www.utsc.utoronto.ca/~butler/c32/fall.txt`. The first column is the time in seconds and the second column is the height of the ball above the ground in metres.
   (a) (2 marks) In SAS, read in the data and make a scatterplot of height against time.
   (b) (3 marks) Describe the relationship you see on the scatterplot. You can use the STAB22 ideas of "direction, form, strength" to guide you: direction is "up" or "down" or "both up and down", form is "a line" or "a curve", strength is anything from "strong relationship" to "very weak relationship".
   (c) (3 marks) I assert that a straight line fits very badly. Demonstrate this using a suitable residual plot.
   (d) (2 marks) Add a time-squared term to the regression and show the slope estimates. (In SAS, `**` raises a number to a power, or you can multiply time by itself.)
   (e) (2 marks) Is your squared term significant? Does its significance (or non-significance) surprise you? Explain briefly.
   (f) (3 marks) One of Newton's equations of motion is $s = ut + \frac{1}{2}at^2$, where $s$ is distance from the starting point, $u$ is initial velocity and $a$ is acceleration. Compare that with your regression output from (d). How high is the observation deck of the tower? With what velocity upwards was the ball thrown? What is the acceleration (due to gravity, here)?

2. Mesquite is a plant that grows in Mexico and the southwestern US. It has a hard wood that is good for furniture and implements. The wood also burns slowly and very hot, which makes it ideal for barbecues. See, for example, `http://en.wikipedia.org/wiki/Mesquite`.

   Data were collected on 46 mesquite bushes. The aim of the study was to predict the total weight of photosynthetic material (`LeafWt`), as derived from the actual harvesting of the bush, from other variables that are more easily measured while the bush is growing. Before we get to those other variables, note that the "canopy" is the leafy area of the bush, and that the canopy is typically shaped (when viewed from above) like an ellipse, with a longer axis and a shorter axis.

   The other variables are:

   **Obs:** Just the observation number, which you can ignore from here on.

   **Group:** measurements in different groups were taken at different times of year (but in the same location).

   **Diam1:** the diameter of the canopy measured along the longer axis (metres)

   **Diam2:** canopy diameter measured along shorter axis (metres)

   **TotHt:** total height of the bush (metres)

   **CanHt:** canopy height (metres) from top to bottom

   **Dens:** number of primary stems

   The data are in `http://www.utsc.utoronto.ca/~butler/c32/mesquite.txt`.

(a) (2 marks) Read in the data, and obtain a matrix of scatterplots between all the variables. (This is sometimes called a "pairs plot".)

(b) (3 marks) In your pairs plot, do you see any relationships, especially with `LeafWt`? Explain briefly (as briefly as seems reasonable).

(c) (2 marks) The function `quantile` obtains a five-number summary for the variable fed into it. Obtain a five-number summary for all the variables except for the first two in the data frame. Use only *one* line of R code. (If you need, you can take one more line for setup, but the actual finding of the quantiles must take only one line of code.)

(d) (2 marks) Obtain boxplots for all the explanatory variables in the data frame (this is columns 3 through 7). These variables are all on similar scales, so you can obtain a side-by-side boxplot of them all by feeding the appropriate columns of the data frame into `boxplot`.

(e) (1 mark) What do you notice about all the boxplots in (d)?

(f) (5 marks) Do the "leverage" thing to detect any outliers. Which observation(s) are detected as outliers? What seems to be unusual about them? (Include your response variable among the variables to detect an outlier in.)

(g) (2 marks) Looking at your boxplots of (d) (in fact, the response variable `LeafWt` has a similar shape) and the nature of any outliers that you found in (f), do you think it would be a good idea to transform any or all of your variables? Explain briefly.

(h) (3 marks) Calculate the logs of *all* the variables `Diam1` to `LeafWt`. Run a regression predicting the log `LeafWt` from the logs of the other variables that you calculated, plus `Group` (left as is, since there is no point taking the log of a categorical variable).

(i) (4 marks) Using backwards elimination, remove in sequence the one variable whose slope has the largest P-value, keeping going until all the variables remaining are significant at the $\alpha = 0.10$ level. (If you wish, you can investigate `update` for this, but you don't have to use this.)

(j) (2 marks) Compare the R-squared values for the initial model with all the explanatory variables and for the final model from the backward elimination. How does this comparison indicate that it was OK to remove any variables that you removed?

(k) (4 marks) A particular growing bush comes from group `ALS`. Its canopy has diameter 1.7 metres along the longer axis and 1.5 metres along the shorter axis. The total height of the bush is 1.6 and the height of the canopy is 1.4 metres. It has 1 primary stem. Predict leaf weight for this bush when it is harvested, using the final model from your backward elimination. (You can use R as a calculator, or use `predict`, but either way, show all the steps of your calculation. The group variable is called `Group`; your regressions will use `ALS` as a baseline group, and give you a coefficient `GroupMCD` which expresses how much being in group `MCD` gives a different predicted log-`LeafWt` than being in group `ALS` does.)

Tree 30 has almost the same values for all the explanatory variables. Is its actual leaf weight somewhere close to your prediction?

3. The data set in `http://www.utsc.utoronto.ca/~butler/c32/tamsales4.txt` contains real estate appraisal data for 675 homes in the Tampa, Florida area. These homes are in four neighbourhoods. The four columns of data in the file are the selling price, the value of the land, the value of improvements to the house (all in dollars), and a categorical variable naming the neighbourhood (this is an abbreviation of up to 8 letters).

(a) (3 marks) Read the data into SAS, and make a scatterplot of sale price against land value.

(b) (3 marks) A realtor is concerned that houses with land value greater than \$150,000 are different. Create a SAS data set with only those houses whose land value is less than \$150,000. Also, we are only interested in the relationship between land value and sale price, so we only need to retain those variables in our new data set. Obtain a new scatterplot of sale price against land value for your new data set.

(c) (3 marks) Do you see a group of houses that are out of line with the overall trend? Explain briefly.

(d) (1 mark) What other issue do you see on the scatterplot that would make you have concerns about doing a regression?

(e) (2 marks) How might you attempt to fix the problems of (c) and (d)? Your aim is to construct a regression that will satisfactorily predict sale price for houses on land valued at about $10,000 or more. (You will need to take two distinct actions.)

(f) (3 marks) Create another new SAS data set that implements your actions from (e). (If your proposed actions didn't say exactly what to do, make a guess at the best thing. The quality of your guess is not going to be evaluated.)

4. This uses the same data as the last question, but this time in R:

(a) (3 marks) Read the data into R. Note that this time, the data file has no header line. Give the variables suitable names.

(b) (2 marks) Make a scatterplot of sale price against land value. Use different colours for the different neighbourhoods by adding `col=` to your plotting command, and adding something suitable after the equals sign.

(c) (2 marks) Do the neighbourhoods look different from each other? If so, describe briefly how.

(d) (3 marks) Create a new data frame consisting of the houses with land value greater than $10,000. Use `subset` to drop the variable representing the value of improvements.

(e) (2 marks) Add variables containing the log of sale price and the log of land value to your data frame. Show the `head` of your data frame to demonstrate that those variables really did get added.

(f) (2 marks) Make a scatterplot of log sales against log land value, with the neighbourhoods identified by colour (as you did it before). Add a lowess curve.

(g) (1 mark) Does your scatterplot suggest that a linear regression would be appropriate here? Discuss briefly.