**Statistical analysis in R**

**Exercise (Topic 5): General Linear Models**

**(Based on class material from UBC FRST 530)**

Data for 30 aspen trees (*Populus tremuloides*) were obtained (volume30.xlsx and volume30.crv), showing the crown class (**crownclass:** 1=dominant; 2=codominant; 3=intermediate; 4=suppressed), **age** (years), stem **dbh** (diameter outside bark measured at 1.3 m above ground, cm), **height** (m) from ground to tree tip, and stem **volume** per tree (m3). The stem volume (also called bole volume) is a very difficult variable to measure, often measured by cutting the tree and either: 1) measuring the volume of water displaced using a xylometer, or 2) by cutting the stem into pieces, measuring the diameters and lengths of each piece, and then assuming a volume shape (i.e., paraboloid frustum) for each piece. Therefore, the objective is to predict volume using these other more easily measured variables. Of the predictor variables that are available, **crownclass** and **dbh** are the cheapest (i.e., easiest) to measure, followed by **height**, and then finally, **age** is the most expensive.

**Part I:**

Initially, you will select a model to predict volume without using **crownclass**. To do this, you will follow these steps:

1. Graph **volume** versus **dbh**, **height**, and **age.**
2. Using the graphs as a guide, add some transformations of variables to the data.
3. Decide on two models that you will fit using **volume** or a transformation of volume as the y-variable, and then using **dbh**, **height** and/or **age**, or transformations of these variables as the x-variables.
4. Fit each model. You will need: i) the diagnostic plots and tests (residual plot, homoscedasticity test, normality (Q-Q plot), normality tests, fitted line plot); ii) analysis of variance table; iii) table of estimated parameters with standard errors, etc.); and iv) fit statistics.
5. Using the diagnostic plots first, select two models that meet the assumptions of OLS, that have normally distributed, iid error terms.
6. For the selected models, do a comparison of the models and select one as your volume model.
7. Using your selected model, calculate the estimated volume and a 95% confidence interval for a tree that has: dbh=50 cm; height=25 m; and age=80 years.
8. Using your selected model, calculate the estimated volume and a 95% prediction interval for a tree that has: dbh=50 cm; height=25 m; and age=80 years.

**Part II:**

Using your selected model from Part II, you will add in **crownclass** and decide if **crownclass** improves the model. To do this, you will use the following steps:

1. Graph your **predicted volumes** versus volume using **crownclass** to label the points. Based on the graph, will crown class improve the model?
2. Change the model you selected in Part I to include **crownclass** to modify all parameters (intercepts and slopes).
3. Fit the model. You will need: i) the diagnostic plots (residual plot, normality (Q-Q plot), normality tests, fitted line plot); ii) analysis of variance table; iii) table of estimated parameters with standard errors, etc.); and iv) fit statistics.
4. Using the diagnostic plots, determine if the assumptions of OLS were met.
5. Use a partial F-test to test whether **crownclass** improves the model or not.
6. If **crownclass** improves the model, determine if some terms can be dropped to simplify the model, and chose your final model.
7. Using your final model, calculate the estimated volume and a 95% confidence interval for a tree that has: dbh=50 cm; height=25 m; age=80 years; and **crownclass**=1.
8. Using your final model, calculate the estimated volume and a 95% prediction interval for a tree that has: dbh=50 cm; height=25 m; age=80 years; and **crownclass**=1.

**Sample Code:**

Sample code has been provided for Part I and Part II using R (**05\_EX\_General\_linear\_models.R**). For each of these, run the code in parts, modifying the code as you go along, to obtain the outputs you need.