An agricultural problem

Some agricultural researchers collected data in order to develop a method of estimating the total production (biomass) of mesquite leaves using easily measured parameters of the plant, before actual harvesting takes place. Two separate sets of measurements were taken, one on a group of 26 mesquite bushes and the other on a different group of 20 mesquite bushes measured at a different time of year. All the data were obtained in the same geographical location (ranch), but neither constituted a strictly random sample.

The outcome variable is the total weight (in grams) of photosynthetic material (i.e. leaves!) as derived from actual harvesting of the bush. The input variables are:

diam1:	diameter of the canopy (the leafy area of the bush) in meters, measured along the longer axis of the bush
diam2:	canopy diameter measured along the shorter axis
canopy.height:	height of the canopy
total.height:	total height of the bush
density:	plant unit density (# of primary stems per plant unit)
group:	group of measurements (0 for the first group, 1 for the second group)

You will be turning this in as part of HW05. Please use whatever tools are convenient for you (R, Rmarkdown, MS Word, LaTeX, etc.).

1. Download the file mesquite.dat from Canvas (week06 in the Files area) and read it into R. The variable names aren't the same as in the description above. Make sure you understand the correspondence between the two sets of variable names.
2. Regress the total leaf weight on the other input variables. Call this model `lm.1`. Summarize any important conclusions you can make from this regression. Indicate any weaknesses you can find with this model.
3. Now transform all of the numerical variables in the model by taking their logarithms. Refit the model in problem 2 using these new log variables, and the group factor variable. Call this model `lm.2`. Summarize any important conclusions you can make from this regression. Indicate any weaknesses you can find in this model.
4. Consider the coefficient on canopy height in each model.
   (a) What is the interpretation of this coefficient in lm.1?
   (b) What is the interpretation of this coefficient in lm.2?
5. We would like to compare the predictive performance of `lm.1` vs the predictive performance of `lm.2`. Since there is so little data, we will use in-sample mean squared error between y and $\hat{y}$.
   (a) Calculate the mean-squared error[1] between total leaf weight and the fitted values from `lm.1`.
   (b) Calculate the mean-squared error between total leaf weight and the fitted values from `lm.2`. *Note: We need to transform[2] the results of* `lm.2` *back to the total leaf weight scale, before doing this calculation. Why?*

[1] You can also compare this with 5- or 10-fold cross-validation. The minimal code to do this in R is `library(boot); myfit <- glm(y ~ x, data=mydata); cv.glm(mydata,myfit,K=5)$delta[1]`. Note use of `glm` instead of `lm`!

[2] Because of the transformation, the simple cross-validation code above won't work.

(c) I am *not* suggesting that you compare `lm.1` with `lm.2` using AIC or BIC. Why not??

6. Go back to the original data. Find the best model you can, by trying transformations, interactions or two or more terms, and variable selection. In this case "best" means: a good compromise between the competing goals of:

   - Best reflects the science
   - Best satisfies modeling assumptions
   - Is most clearly indicated by the data
   - Can be explained to someone who is more interested in mesquite plants than in mathematics and statistics.

   Explain why the final model you found is a good compromise between these four criteria.

   For trying interactions, it might be useful to remember that

   - *You specify interactions with a colon:* `X1:X2` *includes the term* $\beta_{12}X1 \cdot X2$ *in the model, and* `X1:X2:X3` *includes the term* $\beta_{123}X1 \cdot X2 \cdot X3$ *in the model.*
   - *You can use* $\star$ *as a shorthand to specify an interaction and all of its lower-order terms:*
     - `X1*X2` *includes the terms* `X1 + X2 + X1:X2` *in the model*
     - `X1*X2*X3` *includes the terms* `X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3 + X1:X2:X3` *in the model, etc.*

     Usually one wants to use $\star$ *rather than* `:` *to specify interactions, so that lower-order terms are automatically included in the model.*

7. Summarize any important conclusions you can make from your final model in problem 6. Indicate any weaknesses you can find in this model.
8. Go back to the original data again. Can you think of one or more transformations or combinations[3] of the variables that might be plausibly related to the total leaf weight? Write down whatever transformation(s) or combination(s) you might think of, and explain why you think they would be related to leaf weight (note: this is about your understanding of how plants grow & produce leaves (i.e. the science), rather than your understanding of statistics!).
9. Fit a model that uses the new variables you created in problem 8 (your model might or might not include other variables as well... you decide!). Summarize any important conclusions you can make from your final model. Indicate any weaknesses you can find in this model.
10. Review your work in all the previous problems. Which is the model that you would want to use to explain to the agricultural researcher how to best predict total leaf weight? *N.b.: This may require some more statistical or numerical comparisons between models. If so, explain what you did & why.*

Note: Problem #5 is about prediction. So mean-squared error and cross-validation make sense. Problems #6 through #9 are about finding the "right" model. Because we are working in-sample, none of the statistics (F, t, etc.) have their theoretical distributions. If there were enough data, we could split the data in half, find the best model in one half, and then test it (with valid statistics) in the other half. Finding the best model in-sample can be a reasonable way to suggest hypotheses for future work (when we get more data!).

[3] e.g. multiplying variables together, dividing one variable by another, etc.