COMM 581 - Assignment #5
Multiple Linear Regression – Interactions

Name: _____ **Total: 20 marks**
Due date: Tuesday Oct. 11, 2015 (1pm)

Background:
You are helping a timber harvest company to predict the volume of timber (which is directly related to profit) that can be harvested from different areas. Volume per hectare is a difficult quantity to measure and can only be determined after the area has been harvested (destructively). It is easy to get information on some others variables, so your goal is to develop a model to predict volume per hectare from these "easy to obtain" variables. You are able to use data from areas that have previously been harvested, and for which you have the volume per hectare measurement (n = 26).

The variables you have information about are the following:

Volume of timber per hectare (response variable) in m^3 / ha – **volha**

Average height of trees (m) - **topht**
Average diameter at breast height (cm) – **dbh**
Stems per hectare (number of trees per hectare) – **stemsha**
Basal area per hectare (obtained from dbh and stems per hectare) m^2 / ha – **baha**
Average age of trees – **age**

Data

volha	age	baha	stemsha	topht	qdbh
441.6	34	36.2	3552	17.4	13.8
375.8	35	33.4	4368	15.6	12.2
451.4	33	35.4	2808	16.8	14.7
467	33	42	6096	16.4	12.2
306	32	27.4	3816	16.7	12.5
500.1	60	27.3	528	22.7	24.4
478.6	60	34	2160	19.4	9.9
652.2	62	42.5	1843	20.5	13.2
644.7	63	40.4	1431	21	16.1
559.3	82	32.8	1071	22.4	22.2
831.9	104	50.5	1764	21.5	17
365.7	62	29.6	1728	16.4	12.1
454.3	52	35.4	2712	18.9	14.1
486	58	39.1	3144	17.5	14
288.1	33	30.3	5712	13.8	5.6
437.1	68	33.3	2160	19.1	16.2
633.2	126	39.9	1026	21	23.2
707.2	125	40.1	552	23.3	29.2
203	117	11	252	22.1	25.8
915.6	112	48.7	1017	24.2	25
903.5	110	51.5	1416	23.2	23
883.4	106	49.4	1341	24.3	23.7
586.5	124	35.2	2680	22.6	21.5
343.5	63	26.9	1935	17.6	14.1
390.8	57	30.4	2616	18.3	13.9
709.8	87	42.3	1116	22.6	23.9

All graphs should be created using R, and all graphs discussed in your assignment should be included with your submission. You can use sum of squares and standard errors from the R output. Show calculations for test statistics, confidence intervals and measures of goodness of fit based on sums of squares.

1. Graph the relationship between each of the potential explanatory variables and the response variable. Describe what you see on each scatterplot (form, direction, strength, outliers). For each variable, state if you would transform the explanatory variable, and if so, which transformation(s) you would try. (**1.5 marks**)

INTERACTION MODEL

2. Create a linear model that includes baha, topht and the interaction between the two (**Interaction Model**). What does the interaction term do in the model? Is this an additive or multiplicative model? (**1 mark**)

```
z1 <- lm(volha ~ baha + topht + baha*topht, data=mydata)
```

3. Using the residual plot, assess the assumptions of linearity and equal variance. Divide the residual plot into 3-4 segments to assess these assumptions. State any concerns you have and their consequences. (**0.5 marks**)

4. Check the normality assumption using a histogram, normality plot, and normality tests. Adjust the bin width for the histogram to be informative. State any concerns you have and their consequences. (**0.5 marks**)

5. Discuss whether or not you think the assumption of independence is met. Describe one way that the assumption of independence could be violated for this dataset. (**0.5 marks**)

6. Test the significance of the regression. (**0.5 marks**)

7. Test the significance of the interaction term using a partial F-test – show the calculation of the partial F statistic based on the sums of squares. What does this result mean for the model? (**1 mark**)

8. Calculate the co-efficient of multiple determination (R^2) and the standard error of the estimate (root MSE). Check your results with the R output. (**0.5 marks**)

9. Write the equation for the model including the co-efficients from the R output. (**0.5 marks**)

10. Using this equation, calculate the predicted volume per hectare for **Area A** with baha = 30 and topht = 24. Calculate the predicted volume per hectare for **Area B** with baha = 37 and topht = 20. Use R to obtain the prediction intervals for these point estimates. Which area do you predict will produce higher volume? (**1.5 marks**)

LOG MODEL

11. Create a linear model that includes log baha and log topht, and uses log volha as the response variable (**Log model**). Why does this model make sense based on the type of data? (**1 mark**)

12. Assess the assumptions of linearity, equal variance and normality of errors. (Your conclusions regarding the assumption of independence should be the same for this model because both models are based on the same data). (**1 mark**)

13. Test the significance of the regression. (**0.5 marks**)

14. Test each of the explanatory variables using a partial F-test. Show the calculations for the partial F statistic based on the sums of squares. (**2 marks**)

15. Using R, convert the predicted values into their original units. Calculate the SSY, SSreg and SSE in original units. Calculate the pseudo-R^2 and the standard error of the estimate (root MSE) based on the SSE in original units. (**1 mark**)

16. In log units, calculate the confidence intervals for the slopes. What would these co-efficients be (in their correct units) if tree volume could be modeled perfectly with the equation for volume of a cone? Do the confidence intervals include these values? (**1.5 marks**)

17. Write the equation for the model including the co-efficients from the R output. (**0.5 marks**)

18. Using this equation, calculate the predicted volume per hectare for **Area A** with baha = 30 and topht = 24. Calculate the predicted volume per hectare for **Area B** with baha = 37 and topht = 20. Use R to obtain the prediction intervals for these point estimates. Remember to backtransform these values into the original units. Which area do you predict will produce higher volume? (**1.5 marks**)

COMPARING THE MODELS

19. Compare the **Interaction Model** and the **Log Model** in terms of how well they met assumptions. Does one meet the assumptions better than the other? (**0.5 marks**)

20. Compare the co-efficient of determination and standard error of the estimate for the **Interaction Model** and the **Log Model**. Which model has a better fit? (**0.5 marks**)

21. Discuss how your predictions for Area A and B differed or were similar between the **Interaction Model** and the **Log Model**. (**0.5 marks**)

22. Which model would you recommend the timber harvest company use for making predictions? Why? (**1.5 marks**)