COMM 581 - Assignment #5
Multiple Linear Regression – Interactions

Name: _____Gurpal Bisra_____ **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

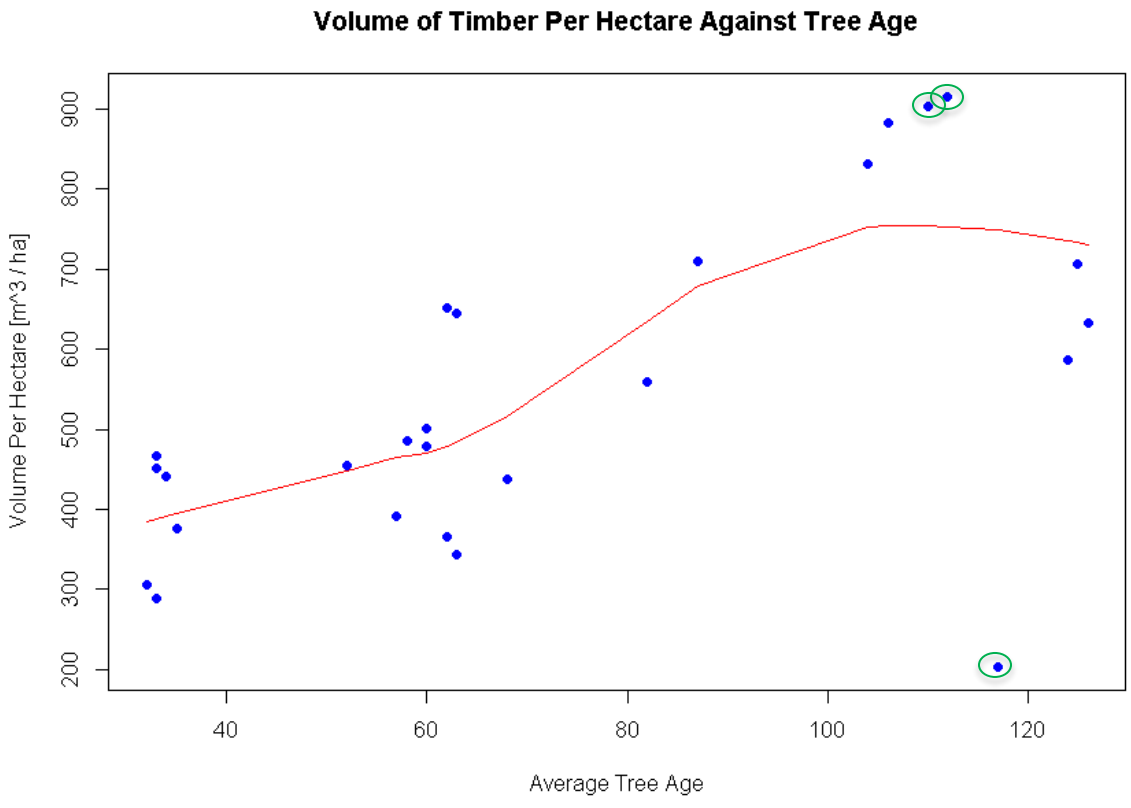
volha	age	baha	stemsha	topht	qdbh
441.6	34	36.2	3552	17.4	13.8
375.8	35	33.4	4368	15.6	12.2
451.4	33	35.4	2808	16.8	14.7
467	33	42	6096	16.4	12.2
306	32	27.4	3816	16.7	12.5
500.1	60	27.3	528	22.7	24.4
478.6	60	34	2160	19.4	9.9
652.2	62	42.5	1843	20.5	13.2
644.7	63	40.4	1431	21	16.1
559.3	82	32.8	1071	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.

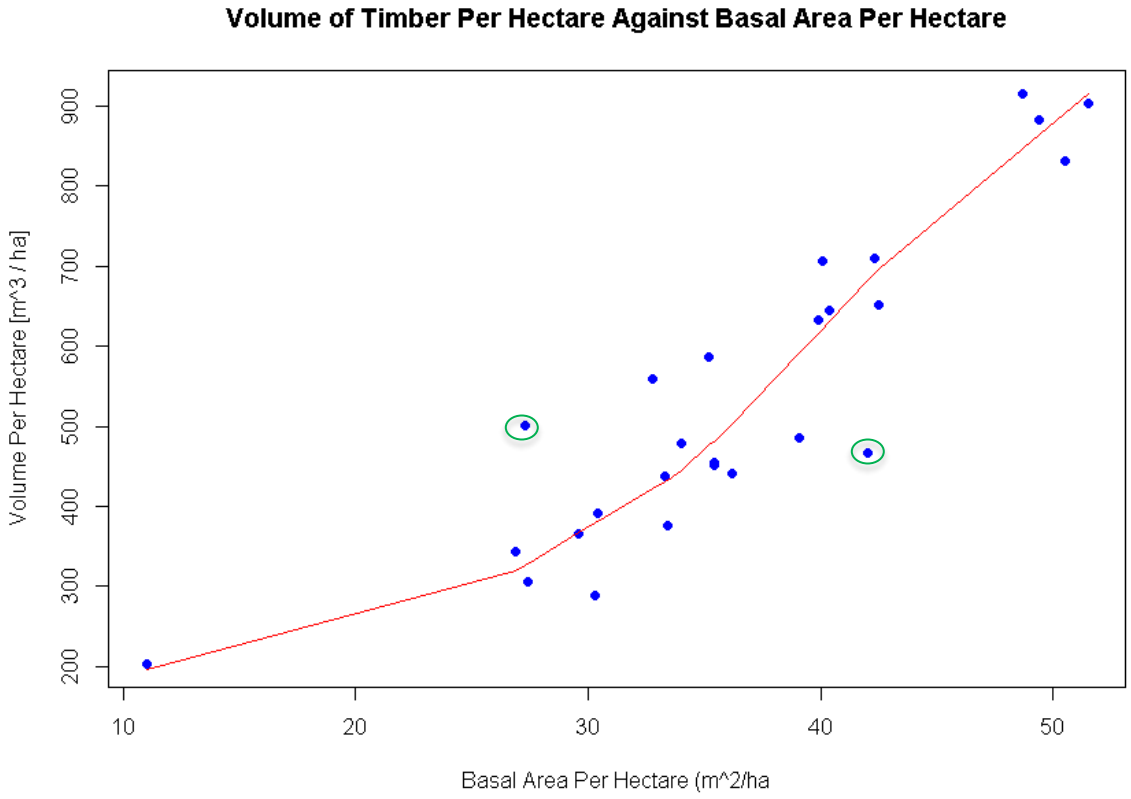
1. Graph the relationship between each of the potential explanatory variables and the response variable. Describe what you see on each scatterplot (form, direction, strength, outliers). For each variable, state if you would transform the explanatory variable, and if so, which transformation(s) you would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

My graphs of the response variable, volume of timber per hectare, as a function of each of the response variables are plotted below. In order to assist me determine any outliers, I added a smoothed red curve through the data with my delta parameter equaling 0.1 for each graph.

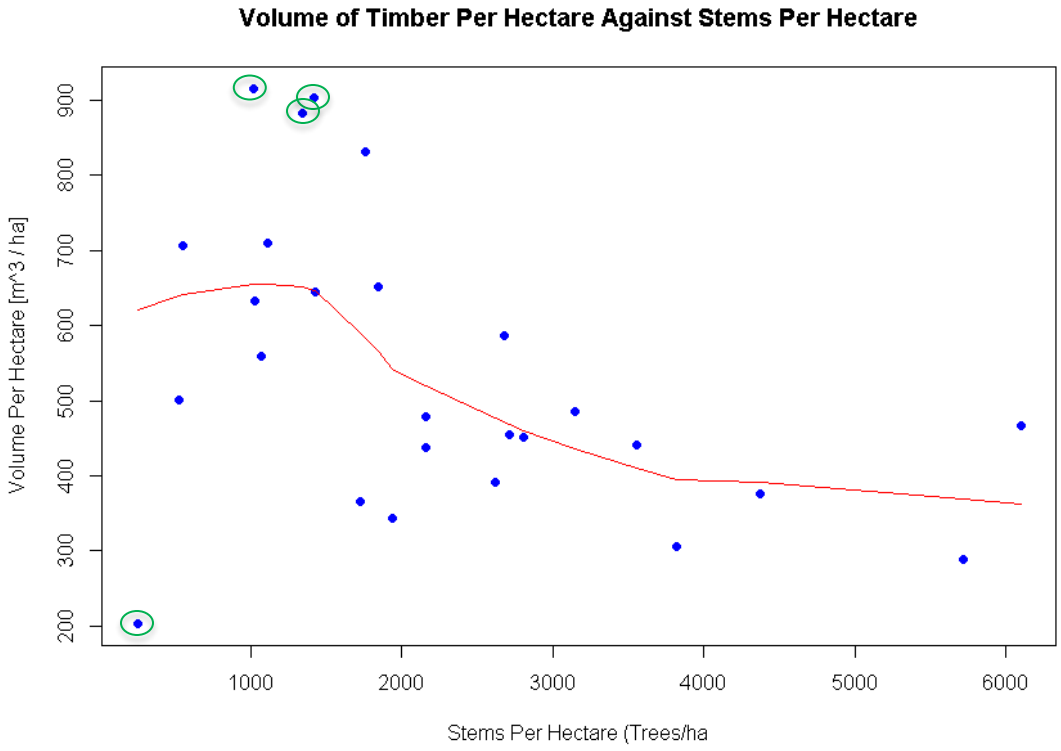
First, I saw a weak linear relationship in the positive direction, with some outliers which I have circled, when I plotted the response variable against the average tree age. Therefore, I would try to transform the average tree age with a $\sqrt{x}$ transformation to see if that models the relationship more linearly. My plot is shown below.



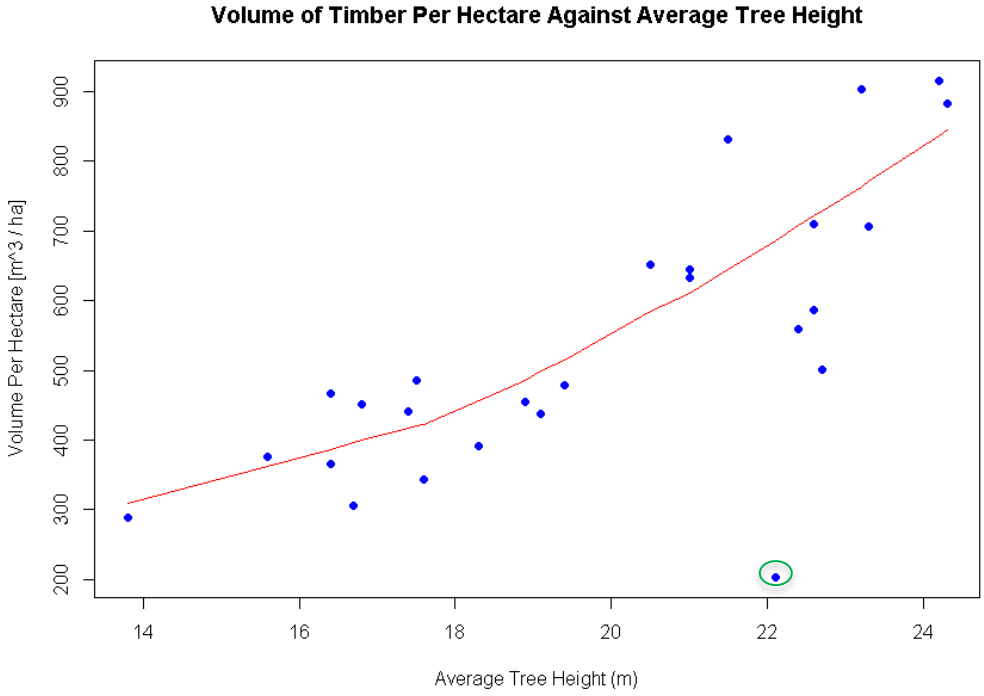
When I plotted the response variable against the basal area per hectare, I did not see a linear relationship in the positive direction, and the data was stronger at higher basal area per hectare values apart from some outliers. I would transform this explanatory variable with a log(x) or x^2 transformation to see if that models the relationship more linearly. My plot is shown below.



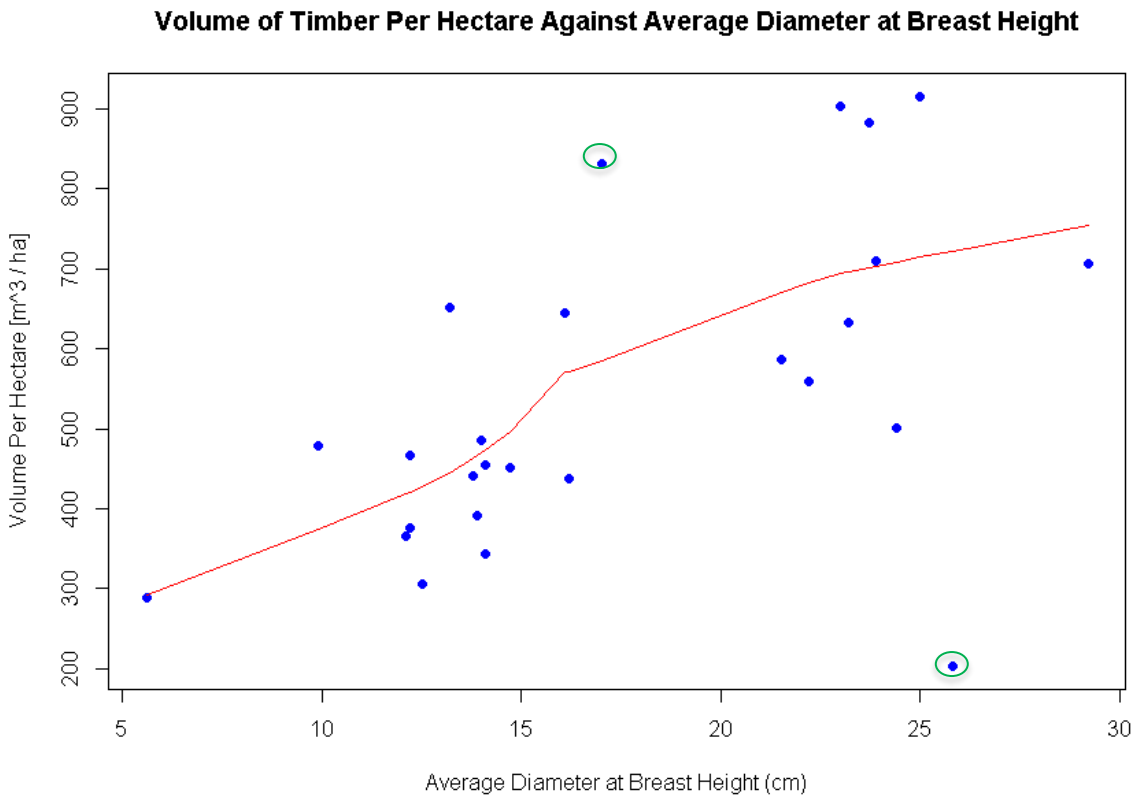
When I plotted the response variable against the stems per hectare, I saw a weak linear relationship in the negative direction with some outliers which I have denoted. Therefore, I would try to use a x^2 transformation to see if that models the relationship more linearly. My plot is shown below.



When I plotted the response variable against the average tree height, I saw a strong linear relationship in the positive direction with some outliers which I have denoted. I would still consider transforming this explanatory variable with a log-transformation. My plot is shown below.



Finally, I plotted the response variable against the average diameter at breast height, I saw a weak linear relationship in the positive direction with some outliers which I have denoted. Therefore, I would try to transform the average tree age with a $\sqrt{x}$ transformation to see if that models the relationship more linearly. My plot is shown below.



INTERACTION MODEL

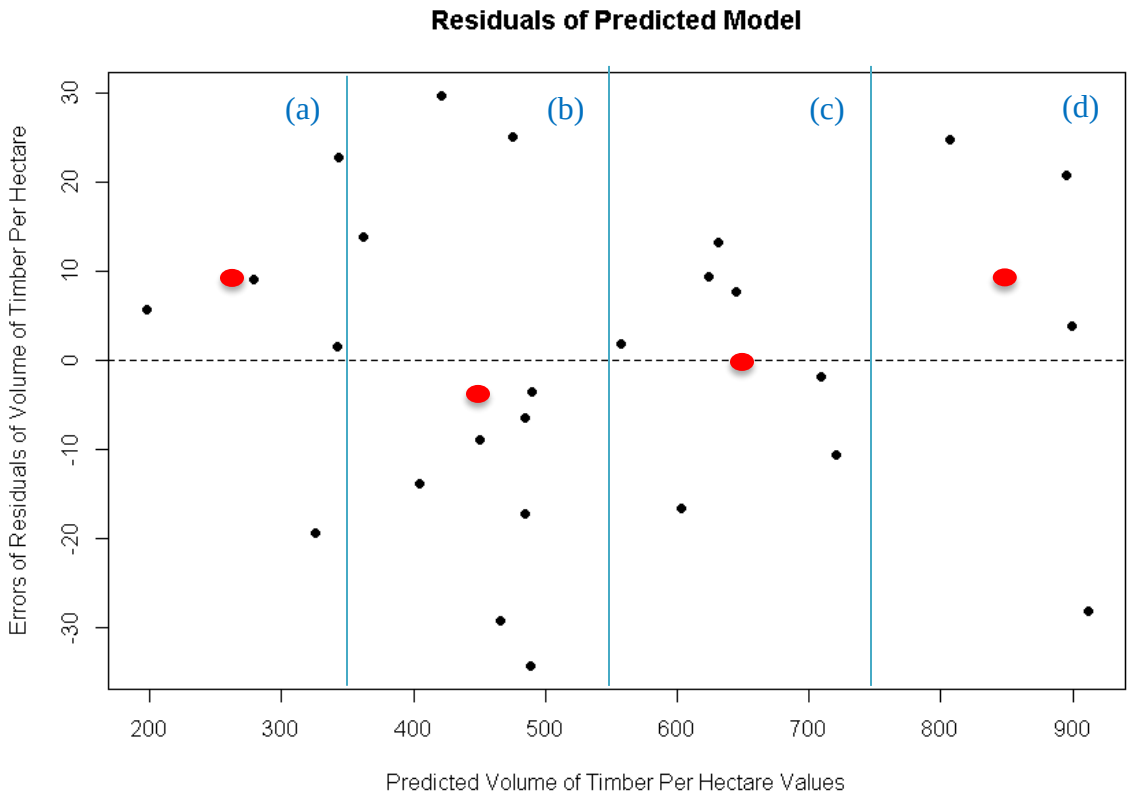
2. Create a linear model that includes baha, topht and the interaction between the two (**Interaction Model**). What does the interaction term do in the model? Is this an additive or multiplicative model? (**1 mark**)

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

The interaction term in the model accounts for fact that the value of our response variable depends greatly on one explanatory variable depending on the value of another explanatory variable. For example, the effect x1 has on y can very greatly depending on the value of x2; and vice versa. Our model would more likely be multiplicative because we are determine the volume per hectare of trees and our explanatory variables can fit into a volume equation as terms which are multiplied.

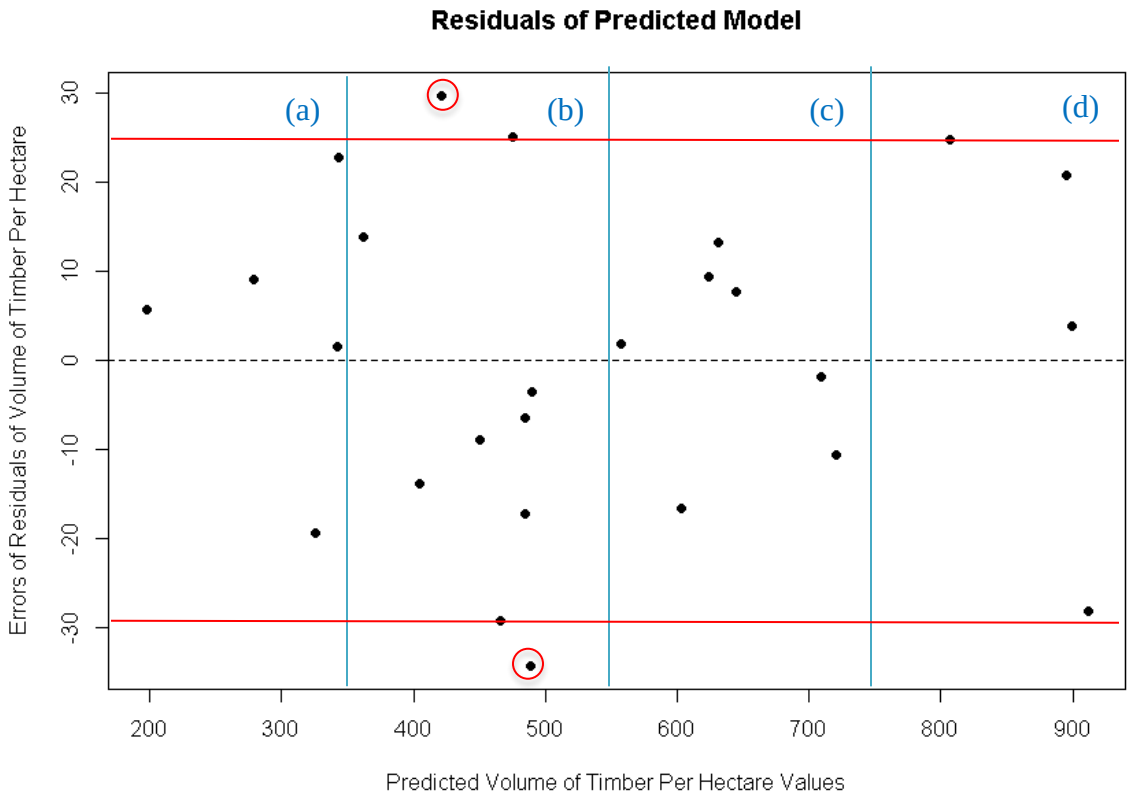
3. Using the residual plot, assess the assumptions of linearity and equal variance. Divide the residual plot into 3-4 segments to assess these assumptions. State any concerns you have and their consequences. (**0.5 marks**)

The assumptions of linearity and equal variance are required to be met in order for one to fit a multiple linear regression line into the data well. To test these assumptions, I plotted the residuals of my multiple linear model against the predicted average volume per hectare values below. Next, I divided my plot of the residuals into 4 segments labelled a, b, c, and d as seen below.



I am trying to detect whether enough points appear evenly spread above and below the line of zero residuals. First, I visually predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. There does appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate close to the line of zero residuals. Hence, I conclude that the assumption of linearity has been met. This means the regression line would fit into my data well and the estimates of my coefficients and standard errors would not be biased.

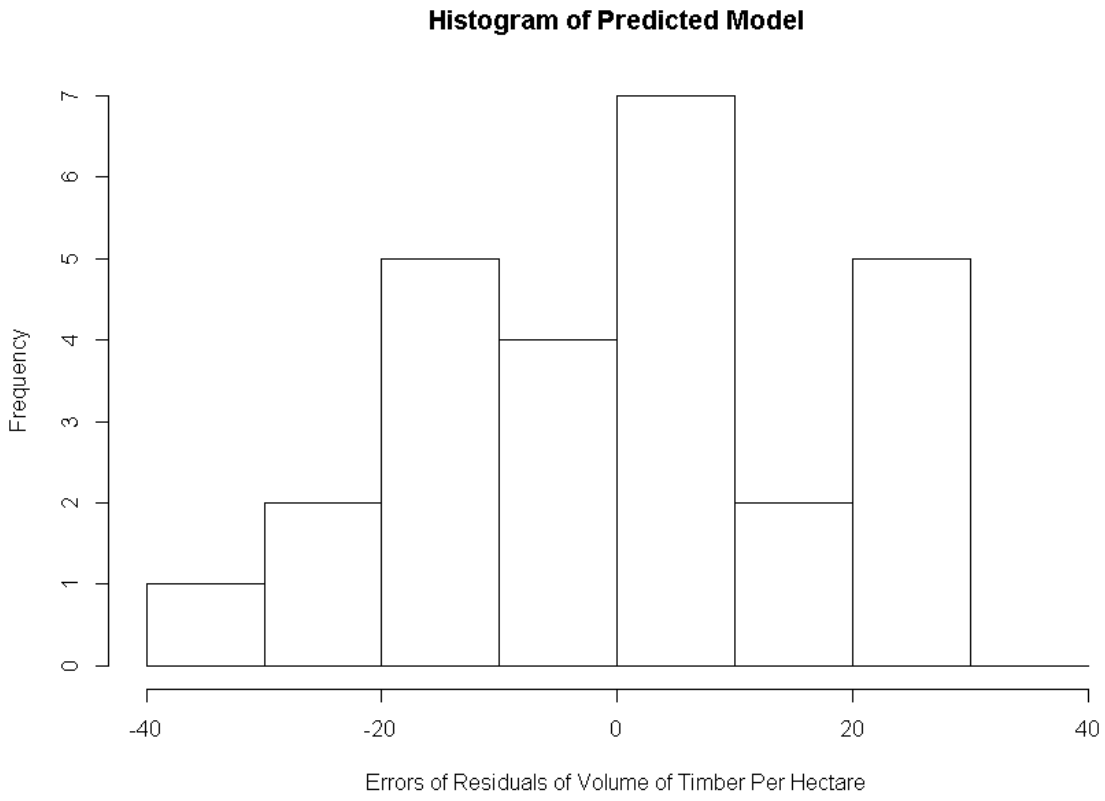
The assumption of equal variance, or the measure of spread, can be verified if the residuals fall above and below zero residuals evenly. I used the following plot below to test this assumption.



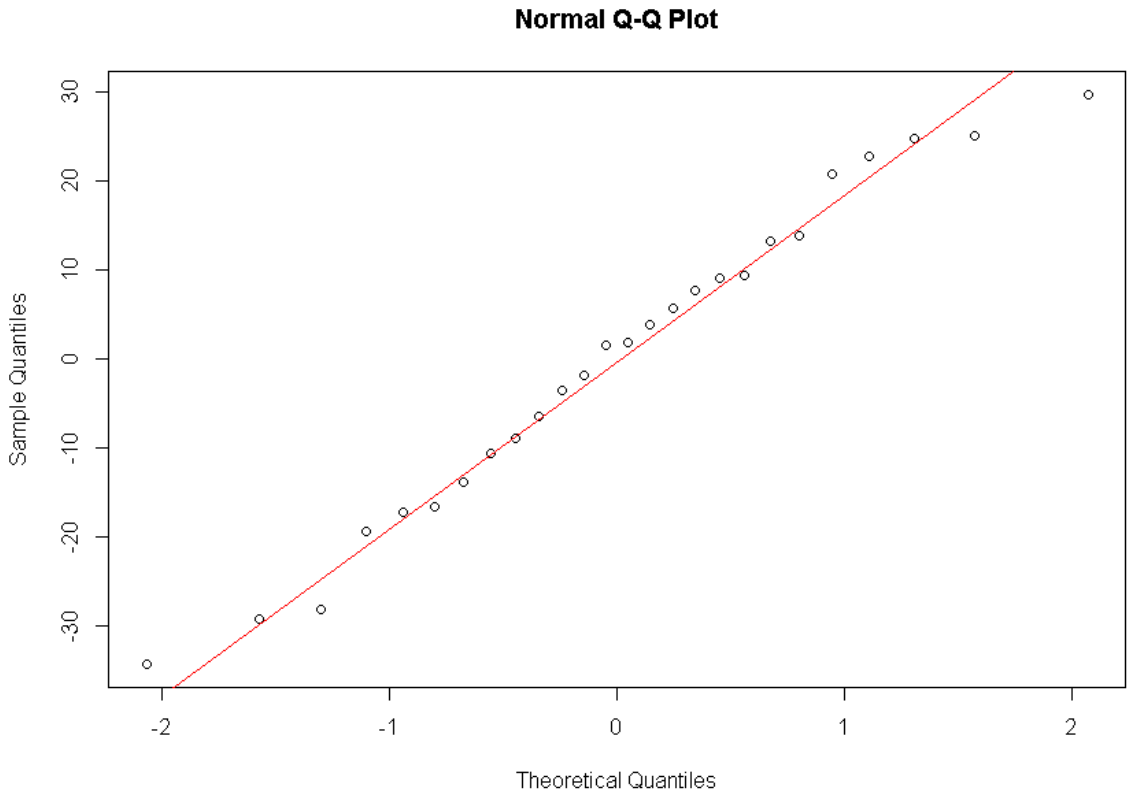
I drew 2 straight red lines in the graph to illustrate approximate equal variance. In addition, I have circled the points which do not contribute to determining whether the assumption of equal variance is met. Given my analysis above, I concluded the assumption of equal variance can be met. This means I can calculate the confidence intervals (CI's) and test the significance of the explanatory variable. In addition, the co-efficients of my regression and estimates of standard errors of co-efficients should not be biased. I used the following plot below to test this assumption.

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**)

In order to fulfill the assumption of normality of errors, the errors must be normally distributed. I plotted a histogram of the residual errors from my predicted model, as seen below, and it appears approximately normally distributed upon visual assessment.



Next, I plotted the Q-Q plot of my log-log data as seen below. Since the standardized residuals change linearly by the theoretical quantile, there is further evidence that the residuals errors are normally distributed for the explanatory variables. In particular, the Q-Q plot exhibits light-tails.



Furthermore, I performed four normality tests whose results are summarized below. My hypothesis is:

H0: Errors of predicted model are normally distributed.

H1: Errors of my predicted are not normally distributed.

Test	Statistic	p Value		Accept or Reject H0
Shapiro-Wilk normality test	W = 0.97307	p < W	p = 0.7039	Fail to Reject H0
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.07533	p > D	p = 0.9657	Fail to Reject H0
Cramer-von Mises normality test	W = 0.021308	p > W	p = 0.9515	Fail to Reject H0
Anderson-Darling normality test	A = 0.18264	p < A	p = 0.9018	Fail to Reject H0

In my testing, I am using an alpha value of 0.05. Given that 100% of the tests fail to reject the null hypothesis (i.e. p > 0.05), more evidence is given that the errors of the predicted model are normally distributed. Therefore, I will continue with my model.

Most importantly, statistical tests do not always work. If it were the case that the tests only passed because I had data that mimicked a normal distribution, then it would mean I could not calculate the CI's or test the significance of the explanatory variable, log of the number of employees, because I wouldn't know what probabilities to use. Hence, then the estimated coefficients would then no longer equal to the maximum likelihood solutions.

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**)

The assumption of independence of observations cannot be verified because I am not provided any information on when or where the data was collected. For instance, if I wanted to prove my data depended on another explanatory variable, I would need to either look at my data at 2 different time points or plot my residuals against what I believe is causing the dependency. In this case, if I observed some relationship between my plotted residuals against time, or space, then the assumption of independence might be broken.

Firstly, it might be possible that the harvesting output is affected by location. For instance, the timber harvesting may yield different volumes of timber per hectare depending on where the harvesting takes place. This may occur because different locations have different weather, soil nutrients or pH, and access to sunlight throughout the year. For example, the same species of trees grown in New Zealand generally grow faster than trees in British Columbia due to the aforementioned factors. In addition, different locations may contain measuring instruments with different precision. This may allow some workers to get one value for volume which could be different from what another worker would measure. Furthermore, some hectares might be overlapping. This would break the assumption of independence because some stems (i.e. stemsha) would be double-counted.

Secondly, the data might have been collected at different points in time. For example, different trees might be harvested after a number of years apart. In this case, the outside temperature, or humidity, might affect the average volume per hectare depending on what time gap between measurements even further.

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

Step 1: Hypothesis for Multiple Linear Regression with Interactions

Ho: the regression is not significant

(B0, B1, B2, and B3 = 0)

H1: the regression is significant

(not all slopes B0, B1, B2, and B3 ≠ 0)

Step 2: Determine my Global F and p values.

Using the summary command, whose output is shown below, I determined:

p-value = 2.2 * 10^{-16}; Global F_{value} = 886.2

```
> summary(z1)

Call:
lm(formula = volha ~ baha + topht + baha * topht, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-34.24 -13.00   1.74  12.30  29.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.1176   151.7813  -0.238 0.814116
baha         -2.1824     4.3492  -0.502 0.620790
topht         2.5438     7.0644   0.360 0.722212
baha:topht    0.8278     0.1987   4.165 0.000403 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.99 on 22 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9907
F-statistic: 886.2 on 3 and 22 DF,  p-value: < 2.2e-16
```

Step 3: Compare my Global F and p values to their critical values.

I calculated my F-critical value in R, since $F_{m, n-m-1, 1-a} = F_{3, 22, 0.95}$ by:

```
> qf(0.95, 3, 22)
[1] 3.049125
```

($F_{critical}$ = 3.049125) < (F_{value} = 125.76)

(p_{value} = 9.106 * 10^{-7}) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, not all the slopes are equal to 0.

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**)

Step 1: Hypothesis for Multiple Linear Regression with Interactions

Ho: the interaction is not significant given the other x-variables in the model (B3 = 0)

H1: the interaction is significant given the other x-variables in the model (B3 ≠ 0)

Step 2: Determine my Partial F and p values.

I used the drop1() command to determine the partial F-values, by dropping one variable at a time. Now, I am comparing:

Full Model: volha ~ baha + topht + baha*topht

Reduced Model: volha ~ baha + topht (i.e. dropped interaction)

I read my Partial F and p-values from the output of my command below:

```
> drop1(z1, test="F")
Single term deletions

Model:
volha ~ baha + topht + baha * topht
           Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                  7935.6 156.75
baha:topht  1    6257.7 14193.3 169.86  17.348 0.0004031 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 0.0004031

F-partial-value = 17.348

Step 3: Compare my Global F and p values to their critical values.

I calculated my F-critical value in R, since $F_{r,\ n\text{-}m\text{-}1,1\text{-}a} = F_{1,\ 22,\ 0.95}$ by:

```
> qf(0.95, 1, 22)
[1] 4.30095
```

($F_{critical}$ = 4.30095) < (F_{value} = 125.76)

(p_{value} = 0.0004031) < (alpha = 0.5)

Step 4: We reject the null hypothesis, therefore the interaction is significant. Therefore, I will keep my B3 term in my multiple linear regression model since the interaction predicts the response variable better when it is included in the model.

Calculation of the partial F statistic based on the sums of squares:

```
> anova(z1)
Analysis of Variance Table

Response: volha
           Df Sum Sq Mean Sq  F value    Pr(>F)
baha        1 765380  765380 2121.868 < 2.2e-16 ***
topht       1 187368  187368  519.442 < 2.2e-16 ***
baha:topht  1   6258    6258   17.348 0.0004031 ***
Residuals  22   7936     361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSreg_{full}$= 765380 + 187368 + 6258 = 959006

SSE_{full}= 7936

$SSreg_{partial}$= 765380 + 187368 = 952748

$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{r*(SSE_{full})} = \frac{(26-3-1)*(959006-6257.7)}{(1)*(952748)}$ = 22.0000 ~ 17.348

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**)

First, I calculated SSy in R as follows:

```
> SSY <- sum((mydata$volha - mean(mydata$volha))^2)
> SSY
[1] 966941
```

The estimates of R^2 value and standard error of the estimate (root MSE) were calculated as follows:

$$R^2 = \frac{SSReg}{SSy} = \frac{959006}{966941} = \mathbf{0.99179}$$

$$SE_E = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{7936}{26-3-1}} = \mathbf{18.99282}$$

In R, I verified these values were correct using the summary ( ) command as shown below.

```
> summary(z1)

Call:
lm(formula = volha ~ baha + topht + baha * topht, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-34.24 -13.00   1.74  12.30  29.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.1176   151.7813  -0.238 0.814116
baha         -2.1824     4.3492  -0.502 0.620790
topht         2.5438     7.0644   0.360 0.722212
baha:topht    0.8278     0.1987   4.165 0.000403 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.99 on 22 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9907
F-statistic: 886.2 on 3 and 22 DF,  p-value: < 2.2e-16
```

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

$\hat{y}_i = B_0 + B_1 x_{1i} + B_2 x_{2i} + B_3 x_{1i} x_{2i}$ becomes

$$\hat{y}_i = -36.1176 - 2.1824 * x_{1i} + 2.5438 * x_{2i} + 0.8278 * x_{1i} * x_{2i}$$

Where:

x_{1i} = the ith value of baha
x_{2i} = the ith value of topht
B_0 has units $\frac{m^3}{ha}$
B_1 has units m
B_2 has no units

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**)

For Area A: baha = 30, topht = 24.

$$\hat{y}_i = -36.1176 - 2.1824 * x_{1i} + 2.5438 * x_{2i} + 0.8278 * x_{1i} * x_{2i}$$

$$\hat{y}_i = -36.1176 - 2.1824 * (30)[\frac{m^2}{ha}] + 2.5438 * (24)[m] + 0.8278 * (30) * (24)[\frac{m^3}{ha}]$$

$$\boldsymbol{\hat{y}_i = 555.4776 \frac{m^3}{ha}}$$

In R, I calculated the prediction intervals using the predict ( ) command as shown below:

```
> predict(z1, data.frame(baha = 30, topht=24, fit=555.4776),interval
= "prediction", level = 0.95)
      fit      lwr      upr
1 555.464 511.9101 599.0178
```

$\hat{y}_i|_{x1=30, x2=24}$ prediction interval = [511.9101, 599.0178]

For Area B: baha = 37, topht = 20.

$$\hat{y}_i = -36.1176 - 2.1824 * (37)[\frac{m^2}{ha}] + 2.5438 * (20)[m] + 0.8278 * (37) * (20)[\frac{m^3}{ha}]$$

$$\boldsymbol{\hat{y}_i = 546.5816 \frac{m^3}{ha}}$$

In R, I calculated the prediction intervals using the predict ( ) command as shown below:

```
> predict(z1, data.frame(baha = 37, topht=20, fit=546.5816),interval
= "prediction", level = 0.95)
       fit     lwr      upr
1 546.5673 506.199 586.9355
```

$\hat{y}_i|_{x1 = 37, x2 = 20}$ prediction interval = [506.199, 586.9355]

I predict Area A to produce a greater volume than Area B.

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**)

I created a linear model which includes log baha, log tropht, and log volha (i.e. log volha ~ log(baha) + log(tropht). My proof, a series of R outputs, are shown below. Creating a log model, which excludes the interaction term, makes sense based on the data because the average volume of a tree can be modeled as a cone. For example, the volume of a cone is:

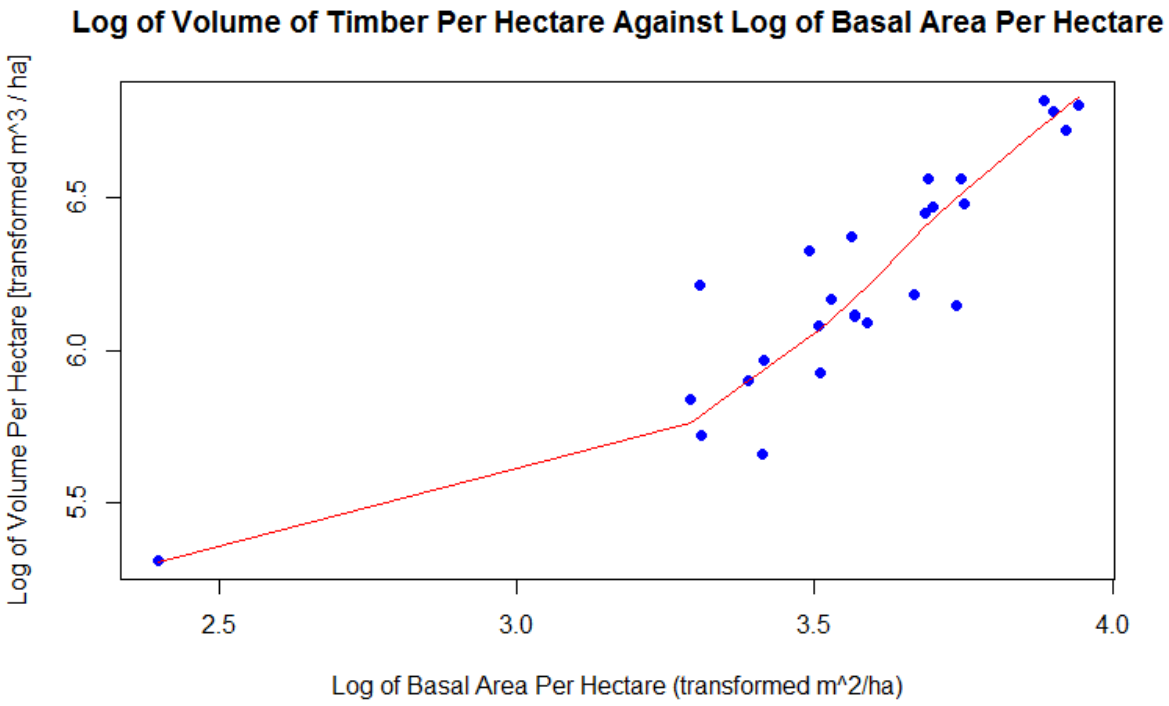
$$V = \frac{1}{3} * \pi * r^2 * h$$

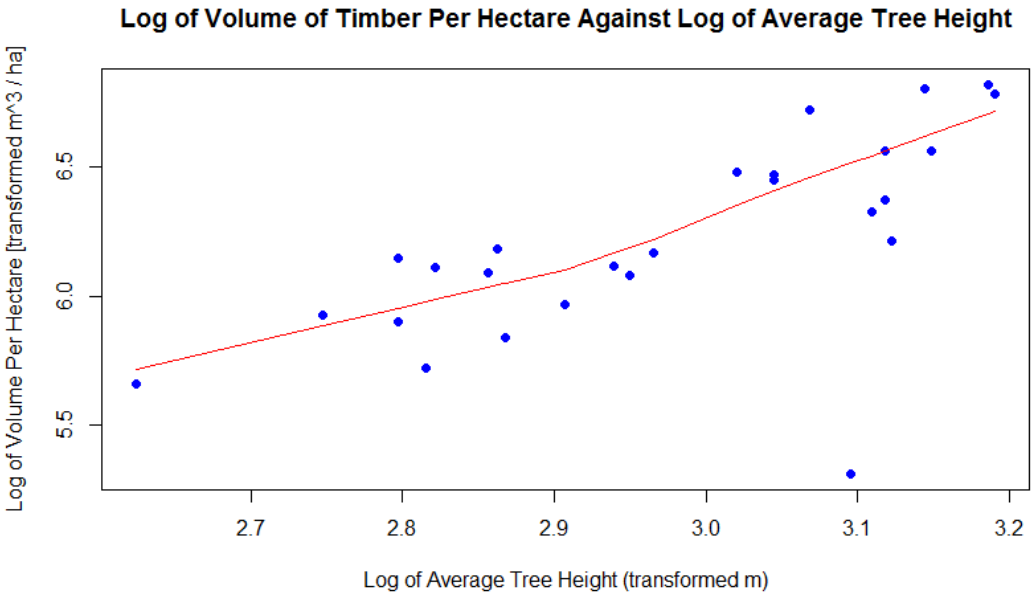
Here, the area is $\pi * r^2$. When comparing this to our dataset, V is described by volha, A is described by baha, while h is depicted by the variable topht. Hence, the log-transformation of this function then becomes:

$$\log(V) = \log\left(\frac{1}{3}\right) + \log(baha) + \log(topht)$$

Therefore, a log model does make sense based on our data. In particular, the log model removes terms that are r multiplied by h. Consequently, the interaction term (i.e. r*h) can be removed from the log-transformed model of volha.

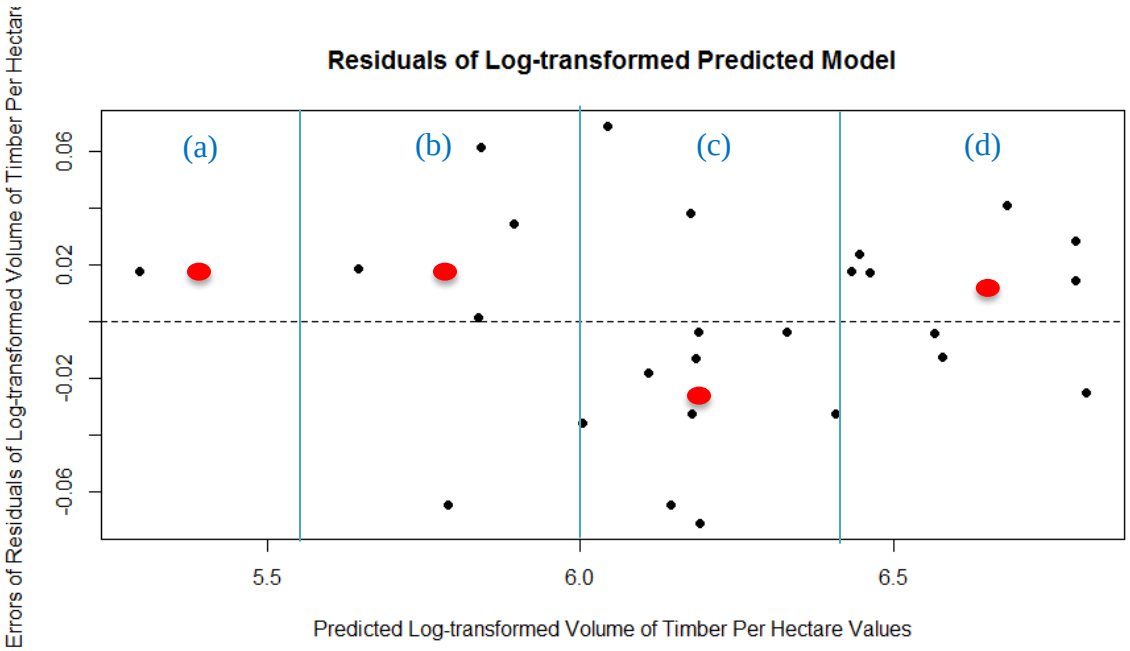
```
z2 <- lm(volha.log ~ baha.log + topht.log, data = mydata)
> str(mydata)
'data.frame':   26 obs. of  11 variables:
 $ volha    : num  442 376 451 467 306 ...
 $ age      : int  34 35 33 33 32 60 60 62 63 82 ...
 $ baha     : num  36.2 33.4 35.4 42 27.4 27.3 34 42.5 40.4 32.8 ...
 $ stemsha  : int  3552 4368 2808 6096 3816 528 2160 1843 1431 1071 ...
 $ topht    : num  17.4 15.6 16.8 16.4 16.7 22.7 19.4 20.5 21 22.4 ...
 $ dbh      : num  13.8 12.2 14.7 12.2 12.5 24.4 9.9 13.2 16.1 22.2 ...
 $ predict1 : num  451 362 422 484 325 ...
 $ resid1   : num  -8.94 13.82 29.74 -17.11 -19.34 ...
 $ baha.log : num  3.59 3.51 3.57 3.74 3.31 ...
 $ topht.log: num  2.86 2.75 2.82 2.8 2.82 ...
 $ volha.log: num  6.09 5.93 6.11 6.15 5.72 ...
```


Log of Volume of Timber Per Hectare Against Log of Basal Area Per Hectare
Log of Volume Per Hectare [transformed m^3 / ha]
5.5
6.0
6.5
2.5
3.0
3.5
4.0
Log of Basal Area Per Hectare (transformed m^2/ha)


Log of Volume of Timber Per Hectare Against Log of Average Tree Height
Log of Volume Per Hectare [transformed m^3 / ha]
5.5
6.0
6.5
2.7
2.8
2.9
3.0
3.1
3.2
Log of Average Tree Height (transformed m)

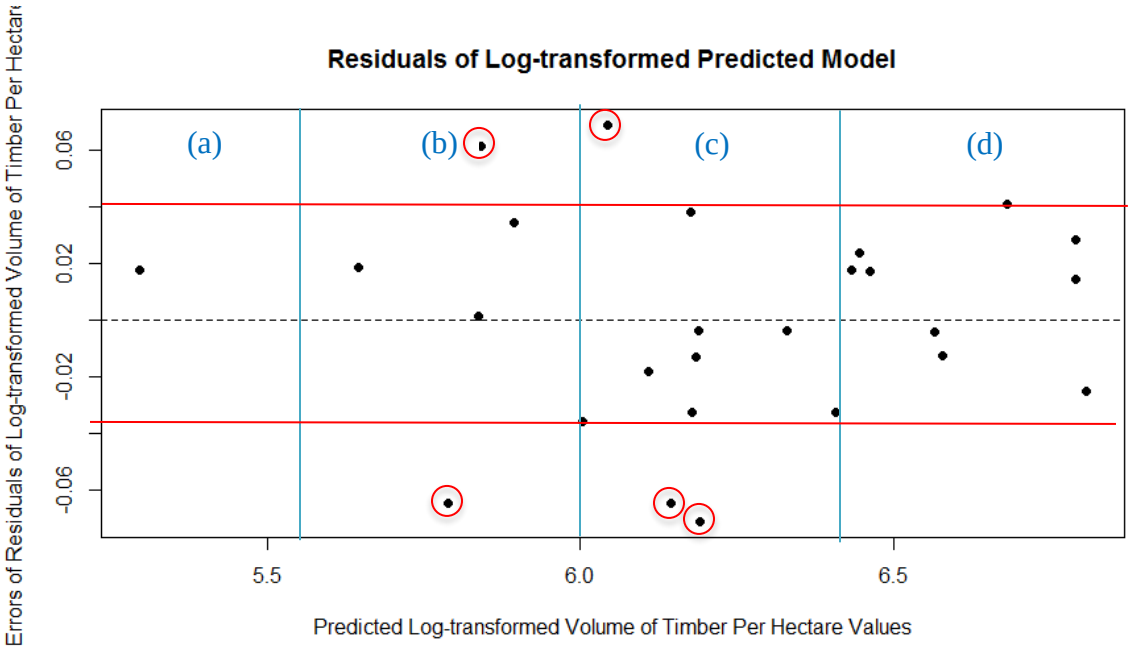
12. Assess the assumptions of linearity, equal variance and normality of errors. (Your conclusions regarding the assumption of independence should be the same for this model because both models are based on the same data). (**1 mark**)

The assumptions of linearity and equal variance are required to be met in order for one to fit a multiple linear regression line into the data well. To test these assumptions, I plotted the residuals of my multiple linear model against the predicted log-transformed average volume per hectare values below. Next, I divided my plot of the residuals into 4 segments labelled a, b, c, and d as seen below.



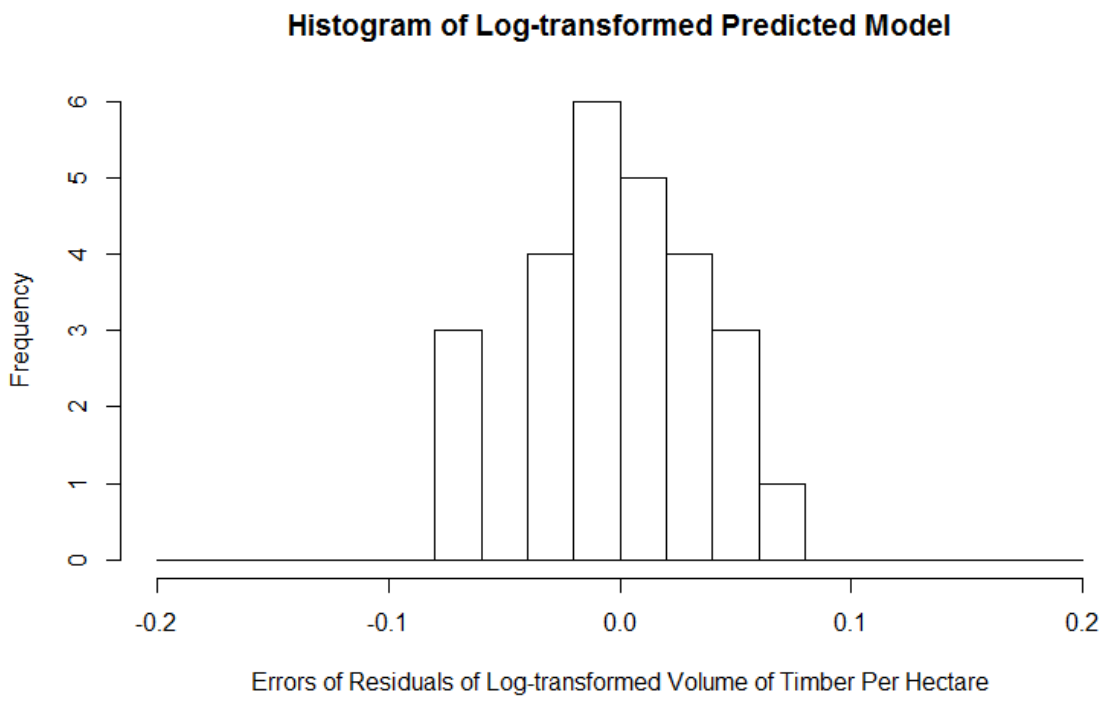
I am trying to detect whether enough points appear evenly spread above and below the line of zero residuals. First, I visually predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. There does appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate close to the line of zero residuals. Hence, I conclude that the assumption of linearity has been met. This means the regression line would fit into my data well and the estimates of my coefficients and standard errors would not be biased.

The assumption of equal variance, or the measure of spread, can be verified if the residuals fall above and below zero residuals evenly. I used the following plot below to test this assumption.

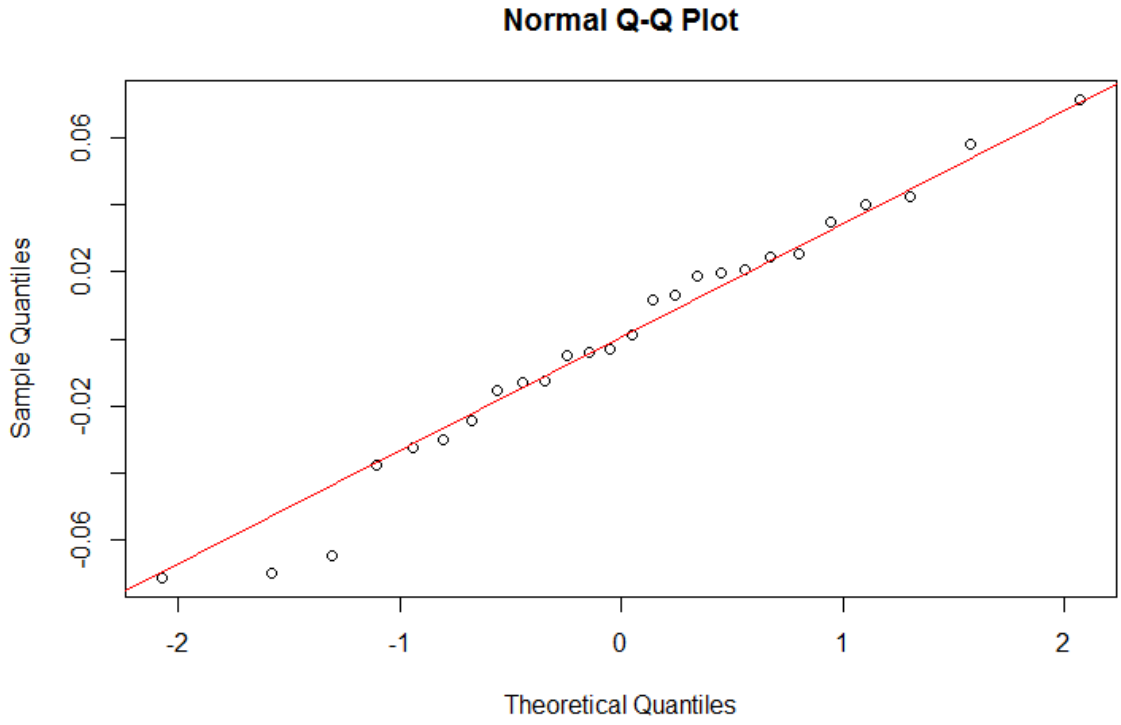


I drew 2 straight red lines in the graph to illustrate approximate equal variance. In addition, I have circled the points which do not contribute to determining whether the assumption of equal variance is met. Given my analysis above, I concluded the assumption of equal variance can be met. This means I can calculate the confidence intervals (CI's) and test the significance of the explanatory variable. In addition, the co-efficients of my regression and estimates of standard errors of co-efficients should not be biased. I used the following plot below to test this assumption.

In order to fulfill the assumption of normality of errors, the errors must be normally distributed. I plotted a histogram of the residual errors from my predicted model, as seen below, and it appears approximately normally distributed upon visual assessment.



Next, I plotted the Q-Q plot of my log-log data as seen below. Since the standardized residuals change linearly by the theoretical quantile, there is further evidence that the residuals errors are normally distributed for the explanatory variables. In particular, the Q-Q plot appears to exhibit light-tails.



Furthermore, I performed four normality tests whose results are summarized below. My hypothesis is:

H0: Errors of predicted model are normally distributed.

H1: Errors of my predicted are not normally distributed.

Test	Statistic	p Value		Accept or Reject H0
Shapiro-Wilk normality test	W = 0.97533	p < W	p = 0.7628	Fail to Reject H0
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.085055	p > D	p = 0.8988	Fail to Reject H0
Cramer-von Mises normality test	W = 0.026813	p > W	p = 0.8808	Fail to Reject H0
Anderson-Darling normality test	A = 0.21217	p < A	p = 0.838	Fail to Reject H0

In my testing, I am using an alpha value of 0.05. Given that 100% of the tests fail to reject the null hypothesis (i.e. p > 0.05), more evidence is given that the errors of the predicted model are normally distributed. Therefore, I will continue with my model.

Most importantly, statistical tests do not always work. If it were the case that the tests only passed because I had data that mimicked a normal distribution, then it would mean I could not calculate the CI's or test the significance of the explanatory variable, log of the number of employees, because I wouldn't know what probabilities to use. Hence, then the estimated coefficients would then no longer equal to the maximum likelihood solutions.

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

Step 1: Hypothesis for Multiple Linear Regression with Interactions

Ho: the regression is not significant (B0, B1, B2 = 0)

H1: the regression is significant (not all slopes B0, B1, and B2 ≠ 0)

Step 2: Determine my Global F and p values.

Using the summary command, whose output is shown below, I determined:

p-value = 2.2 * 10^{-16}; Global F_{value} = 1183

```
> summary(z2)

Call:
lm(formula = volha.log ~ baha.log + topht.log, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.070922 -0.022110 -0.000806  0.023470  0.071203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.81561    0.16045  -5.083  3.8e-05 ***
baha.log     0.93369    0.02613  35.734  < 2e-16 ***
topht.log    1.24995    0.05126  24.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03864 on 23 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9895
F-statistic:  1183 on 2 and 23 DF,  p-value: < 2.2e-16
```

Step 3: Compare my Global F and p values to their critical values.

I calculated my F-critical value in R, since $F_{m, n-m-1,1-a} = F_{2, 26-2-1, 0.95)}$ by:

```
> qf(0.95, 2, 23)
[1] 3.422132
```

($F_{critical}$ = 3.422132) < (F_{value} = 1183)

(p_{value} = 2.2 * 10^{-16}) < (alpha = 0.5)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, not all the slopes are equal to 0.

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**)

First, I will test whether the term log.topht should be included in my log-transformed model. I will use the anova () command to determine my partial F and p-values.

Full Model: log.volha ~ log.baha + log.topht

Reduced Model: log.volha ~ log.baha

Step 1: Hypothesis for Multiple Linear Regression with Interactions

Ho: The log.topht term is not significant (B2 = 0)

H1: The log.topht term is significant (B2 ≠ 0)

Step 2: Determine my Partial F and p values.

I used the anova command to determine the partial F-values. I read my Partial F and p-values from the output of my command below:

```
> z.baha.log <- lm(volha.log ~ baha.log, data=mydata)
>     # topht.log is removed so that's the variable I am checking
> anova(z2, z.baha.log)
Analysis of Variance Table

Model 1: volha.log ~ baha.log + topht.log
Model 2: volha.log ~ baha.log
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     23 0.03435
2     24 0.92241 -1  -0.88806 594.71 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = $2.2 * 10^{-16}$

F-partial-value = 594.71

Step 3: Compare my partial F and p-values to their critical values.

I calculated my F-critical value in R, since $F_{r,\ n\text{-}m\text{-}1,1\text{-}a} = F_{1,\ 26\text{-}2\text{-}1,\ 0.95}$ by:

```
> qf(0.95, 1, 23)
[1] 4.279344
```

($F_{critical}$ = 4.279344) < (F_{value} = 594.71)

(p_{value} = $2.2 * 10^{-16}$) > (alpha = 0.5)

Step 4: I reject the null hypothesis, therefore the log.topht term is significant (i.e. B2 ≠ 0. Therefore, I will keep my B2 term in my multiple linear regression model since the log.topht data predicts the response variable better when it is included in the model.

Calculation of the partial F statistic for interaction based on the sums of squares:

```
> anova(z2)
Analysis of Variance Table

Response: volha.log
           Df  Sum Sq Mean Sq F value    Pr(>F)
baha.log    1 2.64572 2.64572 1771.78 < 2.2e-16 ***
topht.log   1 0.88806 0.88806  594.71 < 2.2e-16 ***
Residuals  23 0.03435 0.00149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSreg_{full}$= 2.64572 + 0.88806 + 0.03435 = 3.56813

SSE_{full}= 0.03435

$SSreg_{partial}$= 2.64572 + 0.03435 = 2.68007

$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{(r)*(SSE_{full})} = \frac{(26-2-1)*(3.56813-2.68007)}{(1)*(0.03435)}$ = 594.6253275

First, I will test whether the term log.baha should be included in my log-transformed model. I will use the anova () command to determine my partial F and p-values.

Full Model: log.volha ~ log.baha + log.topht

Reduced Model: log.volha ~ log.topht

Step 1: Hypothesis for Multiple Linear Regression with Interactions

Ho: The log.baha term is not significant (B1 = 0)

H1: The log.baha term is significant (B1 ≠ 0)

Step 2: Determine my Partial F and p values.

I used the anova command to determine the partial F-values. I read my Partial F and p-values from the output of my command below:

```
> z.topht.log <- lm(volha.log ~ topht.log, data=mydata)
>     # baha.log is removed so that's the variable I am checking
> anova(z2, z.topht.log)
Analysis of Variance Table

Model 1: volha.log ~ baha.log + topht.log
Model 2: volha.log ~ topht.log
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     23 0.03435
2     24 1.94114 -1   -1.9068 1276.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 2.2 * 10^{-16}

F-partial-value = 1276.9

Step 3: Compare my partial F and p-values to their critical values.

I calculated my F-critical value in R, since $F_{r,\ n-m-1,1-a} = F_{1,\ 26-2-1,\ 0.95}$ by:

```
> qf(0.95, 1, 23)
[1] 4.279344
```

($F_{critical}$ = 4.279344) < (F_{value} = 1276.9)

(p_{value} = 2.2 * 10^{-16}) > (alpha = 0.5)

Step 4: I reject the null hypothesis, therefore the log.baha term is significant (i.e. B1 ≠ 0. Therefore, I will keep my B1 term in my multiple linear regression model since the log.baha data predicts the response variable better when it is included in the model.

Calculation of the partial F statistic for interaction based on the sums of squares:

```
> anova(z2)
Analysis of Variance Table

Response: volha.log
           Df  Sum Sq Mean Sq F value    Pr(>F)
baha.log    1 2.64572 2.64572 1771.78 < 2.2e-16 ***
topht.log   1 0.88806 0.88806  594.71 < 2.2e-16 ***
Residuals  23 0.03435 0.00149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSreg_{full}$= 2.64572 + 0.88806 + 0.03435 = 3.56813

SSE_{full}= 0.03435

$SSreg_{partial}$= 0. 88806 + 0.03435 = 0.92241

$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{(r)*(SSE_{full})} = \frac{(26-2-1)*(3.56813-0.92241)}{(1)*(0.03435)}$ = 1771.515575

---

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**)

First, I converted my predicted values to their original units using the following commands:

```
# backtransform the predicted values into the original units
yhat.original <- exp(predict2)
yhat.original

yhat.original <- as.data.frame(yhat.original)
    #make this into a data frame so that you will be able to add it to your
   original data frame

mydata.2 <- cbind(mydata, yhat.original) # since these data frames have the same
number of rows, they should line up nicely
```

Using the back-transformed values, I calculated the errors in original units as follows:

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ = 8386.284
- $SSR = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2$ = 939407
- $SSY = \sum_{i=1}^{n}(y_i - \bar{y})^2$ = 966941

- The commands I used in R are shown below:

```
> SSE <- sum((mydata.2$volha - mydata.2$yhat.original)^2)
> SSE
[1] 8386.284
> SSR <- sum((mean(mydata.2$volha) - mydata.2$yhat.original)^2)
> SSR
[1] 939407
> SSY <- sum((mydata.2$volha - mean(mydata.2$volha))^2)
> SSY
[1] 966941
```

Then, using the back-transformed values, I calculated my pseudo-r^2 value as shown below.

$$pseudo - r^2 = I^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{8386.284}{966941}$$

$$pseudo - r^2 = 0.99132699513 \qquad \text{(original units)}$$

The standard error of the estimate (SE_E) is a measure of the variation not accounted for in the model. I calculated SE_E', as shown below:

$$SE_E' = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{8386.284}{26-2}} = 18.6930067137 \qquad \text{(original units)}$$

16. In log units, calculate the confidence intervals for the slopes. What would these coefficients 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**)

First, I determined my estimates of the slopes B0, B1, and B2 using the summary command. My results are as follows:

```
> summary(z2)

Call:
lm(formula = volha.log ~ baha.log + topht.log, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.070922 -0.022110 -0.000806  0.023470  0.071203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.81561    0.16045  -5.083  3.8e-05 ***
baha.log     0.93369    0.02613  35.734  < 2e-16 ***
topht.log    1.24995    0.05126  24.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03864 on 23 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9895
F-statistic:  1183 on 2 and 23 DF,  p-value: < 2.2e-16
```

The value for $(t_{1-\frac{\alpha}{2},n-2})$ = $(t_{0.975,24})$ since n = 25 was calculated in R as follows:

```
> qt(0.975, 24, lower.tail=TRUE )
[1] 2.063899
```

In log units, I calculated the 95% confidence intervals below using the output values I obtained from the summary () command:

$For\ B_0 = b_0 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_0}$ = -0.81561 ± (2.064)*(0.16045)

$For\ B_1 = b_1 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_1}$ = 0.93369 ± (2.064)*(0.02613)

$For\ 2 = b_2 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_2}$ = 1.24995 ± (2.064)*(0.05126)

Hence, the 95% confidence intervals are as follows:

b_0	[-1.1467788, -0.4844412]	(Log-transformed units)
b_1	[0.87975768, 0.98762232]	(Log-transformed units)
b_2	[1.14414936, 1.35575064]	(Log-transformed units)

The 95% confidence intervals were calculated, in original units, by back-transforming with the exponential function as follows:

b_0	[0.31765836423, 0.6160413445]	(Original units)
b_1	[2.41031556798, 2.68484317921]	(Original units)
b_2	[3.13976940712, 3.87967210095]	(Original units)

In addition, I verified my results in R using the command confint(z2, level=0.95). My result was as follows:

```
> confint(z2, level=0.95) # this is based on the log model, so everything will be in lo
g units
                2.5 %     97.5 %
(Intercept) -1.147536 -0.4836843
baha.log     0.879635  0.9877375
topht.log    1.143917  1.3559752
```

If the tree volume could be modeled perfectly with the equation for the volume of a cone, as I showed in question 11, then the log-transformed equation would appear as:

$$\log(V) = \log\left(\frac{1}{3}\right) + \log(baha) + \log(topht)$$

Therefore:

Bo = log(1/3) = -0.47712125472

B1 = 1

B2 = 1

Interestingly, the confidence intervals do not include these values. Although, the value for B1 is very close.

My estimates of the Bo, B1, and B2 coefficients are -0.81561, 0.93369, and 1.24995, respectively, in log-transformed units. I obtained each value, in original units, by doing the following calculations: (1) Bo.original = $e^{-0.81561}$ = 0.4423693997; (2) B1.original = $e^{0.93369}$ = 2.54387879279; (3) B2.original = $e^{1.24995}$ = 3.49016844468.

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

$\log(\hat{y}_i)$ = -0.81561 + 0.93369*log(x_{1i}) + 1.24995*log(x_{2i})

where: x1 refers to baha, and x2 refers to topht

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**)

For Area A: baha = 30, topht = 24.

$\log(\hat{y}_i)$ = -0.81561 + 0.93369*log(x_{1i}) + 1.24995*log(x_{2i})

$$\log(\hat{y}_i) = -0.81561 + 0.93369 * \log(30) + 1.24995 * \log(24)$$

$$\boldsymbol{log(\hat{y}_i) = 6.3324623}$$ **(log-transformed units)**

Next, I back-transforming using the exponential function yields $\hat{y}_i$

$$\boldsymbol{\hat{y}_i = 562.54007}$$ **(back-transformed units)**

In R, I calculated the prediction intervals using the predict ( ) command as shown below:

```
> predict(z2, data.frame(baha.log = log(30), topht.log= log(24), fit=6.3324623),interva
l = "prediction", level = 0.95)
       fit      lwr      upr
1 6.332436 6.24731 6.417563
```

$\log(\hat{y}_i)|_{x1 = 30, x2 = 24}$ prediction interval = [6.24731, 6.417563] (log-transformed units)

If I take the exponential of my prediction interval, I get the following:

$\hat{y}_i|_{x1 = 30, x2 = 24}$ prediction interval = [516.621243, 612.508610] (back transformed units)

For Area B: baha = 37, topht = 20.

$\log(\hat{y}_i)$ = -0.81561 + 0.93369*log(x_{1i}) + 1.24995*log(x_{2i})

$$\log(\hat{y}_i) = -0.81561 + 0.93369 * \log(37) + 1.24995 * \log(20)$$

$$\boldsymbol{log(\hat{y}_i) = 6.3003835} \quad \textbf{(log-transformed units)}$$

Next, I back-transforming using the exponential function yields $\hat{y}_i$

$$\boldsymbol{\hat{y}_i = 544.780794} \quad \textbf{(back-transformed units)}$$

In R, I calculated the prediction intervals using the predict ( ) command as shown below:

```
> predict(z2, data.frame(baha.log = log(37), topht.log= log(20), fit=6.30038350119),int
erval = "prediction", level = 0.95)
       fit      lwr      upr
1 6.300357 6.218831 6.381884
```

log($\hat{y}_i$)|$_{x1 = 37, x2 = 20}$ prediction interval = [6.218831, 6.381884] (log-transformed units)

If I take the exponential of my prediction interval, I get the following:

$\hat{y}_i|_{x1 = 30, x2 = 24}$ prediction interval = [502.115915, 591.040179] (back transformed units)

I predict Area A to produce a greater volume than Area B.

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**)

A summary of the assumptions of both the Interaction and Log Model are summarized below:

Assumptions	Interaction Model	Log Model
Linearity	• assumption was met.	• assumption was met.
Equal Variance	• assumption was met, found 2 outliers.	• assumption was met, found 5 outliers.
Normality of Errors	• Q-Q Plot contained light-tails. • All 4 normality tests met assumption.	• Q-Q Plot contained light-tails. • All 4 normality tests met assumption.
Independence	• same for Log Model	• same for Log Model

Both models meet the assumptions as well as the other. The only exception is that the test of equal variance in the Log model indicates more outliers; however, both models pass the test for the assumption of equal variance.

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**)

A summary of my co-efficients of determination and standard error of the estimate for both the Interaction and Log Model are summarized below:

	Co-efficient of Determination (R^2)	Standard Error of the Estimate
Interaction Model	0.99179	18.99282
Log Model	0.99133 (i.e. Pseudo-R^2)	18.69301

A better fit is determined by a model which has a higher co-efficient of determination and smaller standard error of the estimate. However, both models are barely better on one aspect than the other. Hence, from just considering these two measures, one would guess that either regression model fit the data equally well.

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**)

A summary of my predictions for the area and prediction intervals for Area A and B, for both the Interaction and Log Model, are summarized below:

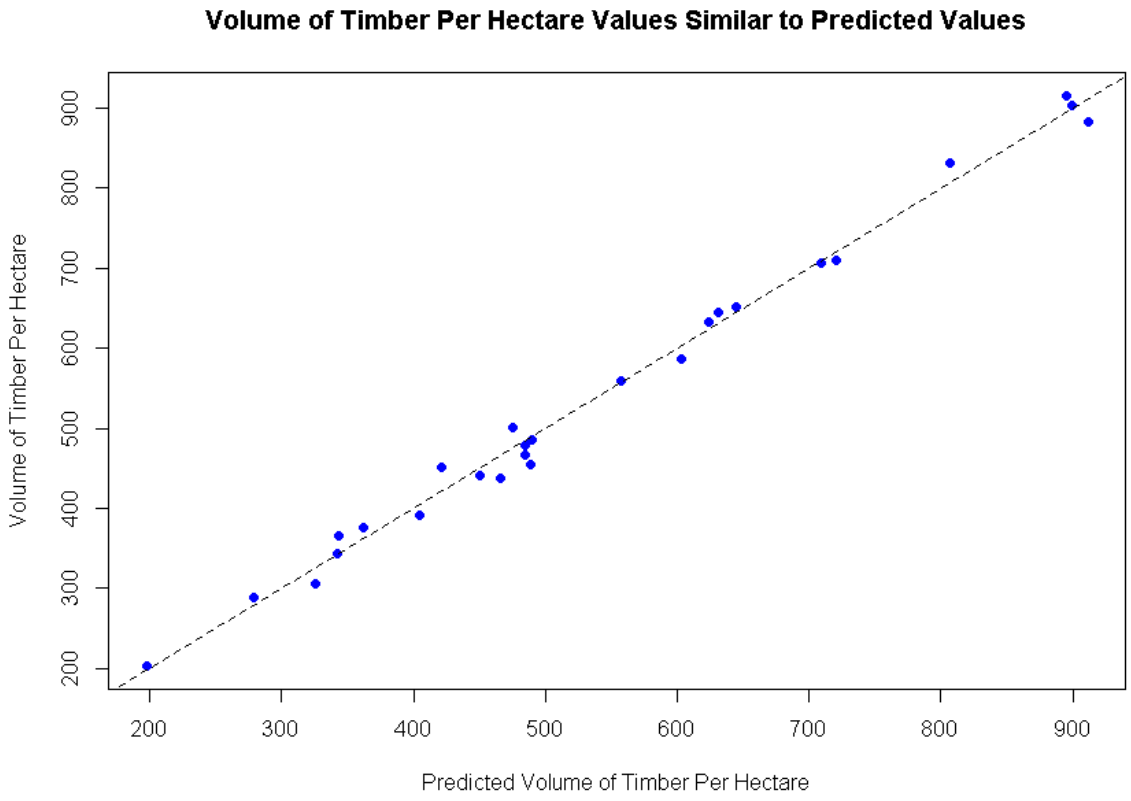
	Interaction Model	**Log Model**
Area A	$\hat{y}_i = 555.4776 \frac{m^3}{ha}$	$\hat{y}_i = 562.54007$ (back-transformed units)
	$[511.9101, 599.0178] \frac{m^3}{ha}$	[516.621243, 612.508610] (back transformed units)
Area B	$546.5816 \frac{m^3}{ha}$	$\hat{y}_i = 544.780794$ (back-transformed units)
	$[506.199, 586.9355] \frac{m^3}{ha}$	[502.115915, 591.040179] (back-transformed units)

Both models predict that area A will have a higher volume per hectare. In addition, the Log Model predicts that the area per hectare is higher in Area A, yet lower in Area B, when compared to the Interaction Model.

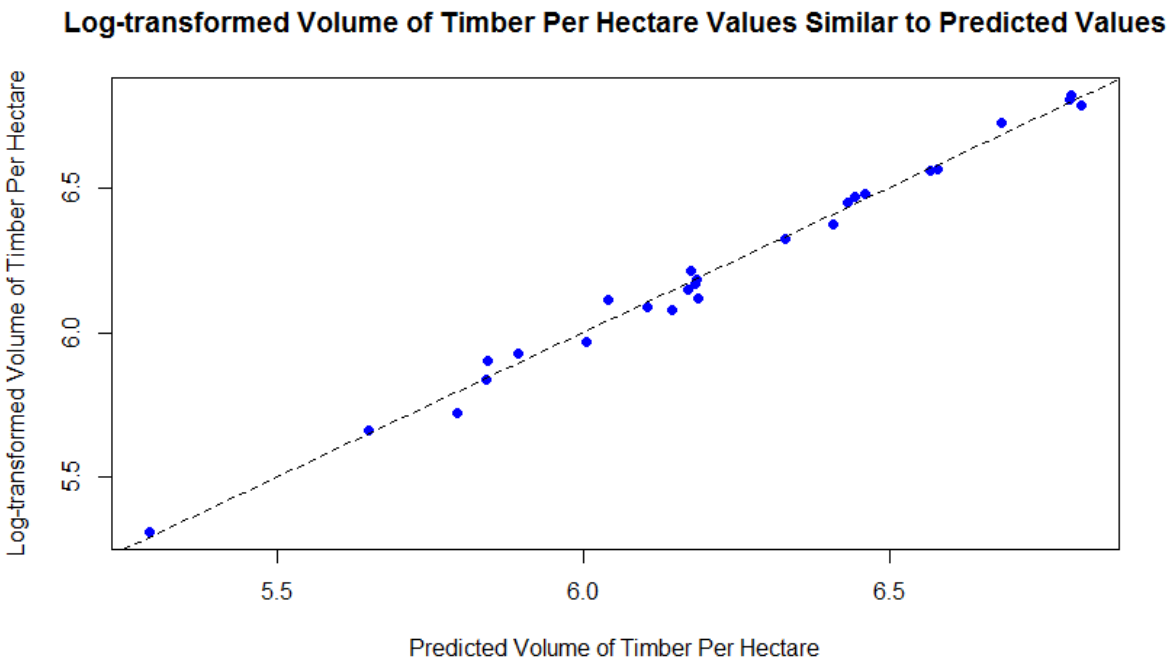
22. Which model would you recommend the timber harvest company use for making predictions? Why? (**1.5 marks**)

I would recommend the timber harvest company use the Log model for making their predictions. First, the Log model is simpler since it requires one to compute two explanatory variables instead of three (i.e. the interaction term). Furthermore, Log Models can take into account larger values better. For instance, if the harvest company obtains more data which has higher values, it is more likely able to be incorporated into the Log Model without seeming as an outlier.

I plotted the original yield values against my predicted yield values for the Interaction Model, obtained by using my model. The output graph, as shown below, further illustrates that my predicted model fits my original data well.



Finally, I plotted the original yield values against my predicted yield values, obtained by using my Log Model. The output graph, as shown below, further illustrates that my predicted model fits my original data well.



When I compare the 2 graphs above, both models appear to fit the data equally well.