Predicting Lumber Volume From Tree Diameter: A Comparison of Two Simple Linear Regression Models

The Lumber Company Inc.

Yang Wu

Math Department

Kenyon College

March 2020

Contents

1 Introduction

This report summarizes all of the statistical modeling and analysis results associated with the study of the statistical relation between tree diameters and the volume of lumber obtained from the tree after processing(henceforth, lumber volume or volume). The purpose of this report is to document both the two competing linear regression models and all corresponding inferences during the subsequent statistical analyses. We understand that a useful predictive model is conducive to the success of your company, and we are confident that the following results and recommendations address your need.

The remainder of this report is organized as follows. Section 2 describes the data in detail, including descriptive statistics and graphical exploration of the variables. Section 3 presents the competing models in equation forms and assesses the appropriateness of the model assumptions, entertaining potential remedial measures. Section 4 presents regression results, including a sensitivity analysis. Next, section 5 takes up inferences concerning the regression parameters and constructs interval estimates of means responses and predictions. Lastly, section 6 concludes with our recommendations on the two competing models based on the results of our analyses.

2 Data

The sample of data is collected by our client, The Lumber Company Inc. In this study, tree volume, measured in cubic feet, is a *random* output for the input of tree diameter, measured in inches. Table 1 summarizes the descriptive statistics of the variables.

2.1 Descriptive Statistics

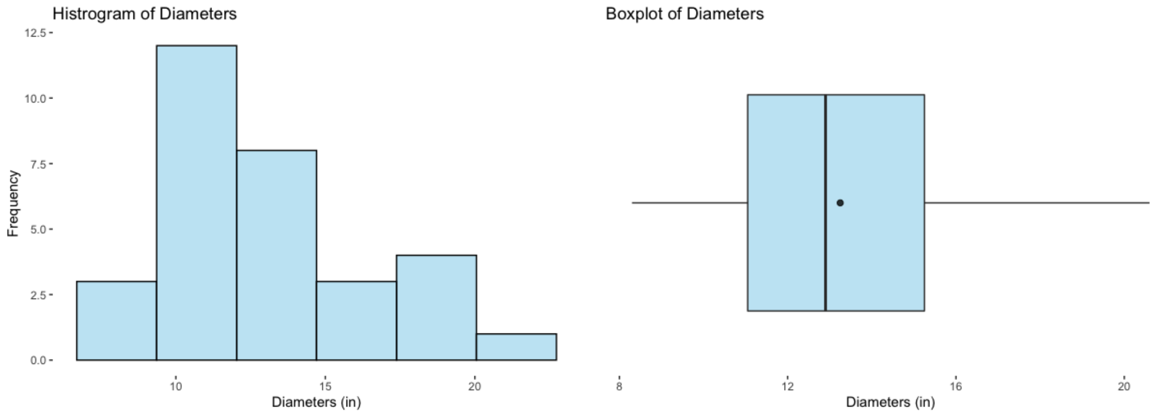
Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Diameter	31	13.248	3.138	8.300	11.050	12.900	15.250	20.600
Volume	31	30.171	16.438	10.200	19.400	24.200	37.300	77.000

The sample contains $N = 31$ observations on the two variables— Volume and Diameter. The medians of both variables are smaller than their means, which indicate that the distributions of diameter and volume are right-skewed. The middle 50% of the diameters in this sample range from 11.050 to 15.250 inches. For volume, the middle 50% ranges from 19.400 and 37.300 cubic feet.[1]

[1] Descriptive Statistics and graphical explorations for the log-transformed diameter and volume are carried out. There is

2.2 Graphical Exploration

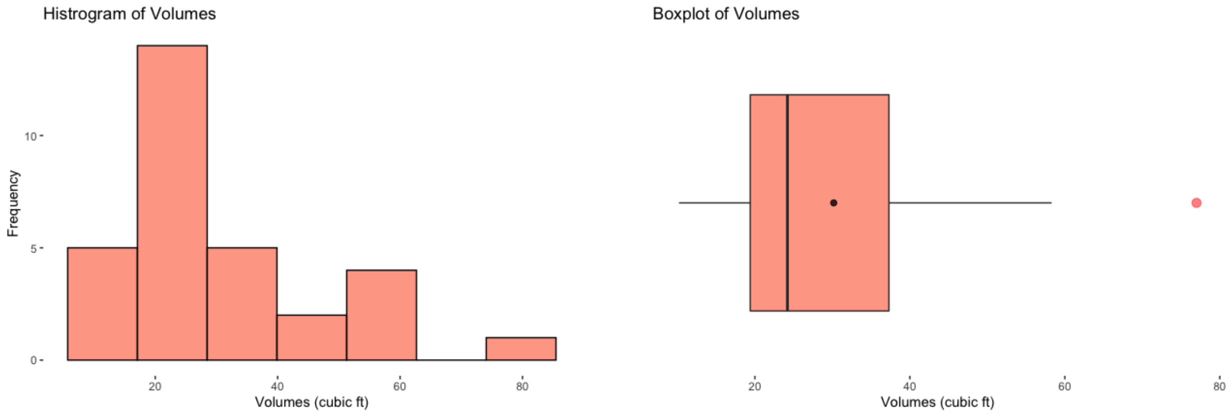


(a) Histogram

(b) Box plot

Figure 1: Diameters

Figure 1 shows that the distribution of diameters is indeed right-skewed.[2] In Figure 1 (a), the range of validity for the regression analysis is shown to be the interval between the minimum and maximum diameters in the sample from Table 1. The box plot in Figure 1 (b) includes the median (black line) and the mean (black point), confirming that there is slight asymmetry; however, there is no diameter that is far outlying.



(a) Histogram

(b) Box plot

Figure 2: Volumes

no presence of outlying value in either transformed variables. Both distributions of the log-transformed diameter and volume appear normal and symmetric. We do not report these plots and tables but include them in the R script that accompanies this report.

[2] The width of the bins in the histograms is selected using the Freedman-Diaconis rule:

$$\text{Bin width} = 2\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

where IQR(x) is the interquartile range of data and n is the number of observations in the sample x.

Figure 2 (a) shows that the distribution of volumes is right-skewed. From figure 2 (b), we see that there is an outlying observation (red point) that lies beyond $Q3 + 1.5 * IQR$. If this observation is not due a measurement error, it behooves us to investigate its effect on our analysis.

3 Models

3.1 Two Competing Models

We develop the two competing models:

$$\text{Model 1:} \quad V_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{1}$$

where

- β_0 and β_1 are the parameters
- D_i are the tree diameters
- V_i are the volumes of lumber obtained from the trees after processing
- ε_i are independent $N(0, \sigma^2)$

$$\text{Model 2:} \quad V_i \approx \alpha D_i^{\beta}$$

To linearize model 2, we take the natural log of both sides of equation above:

$$\begin{aligned} \ln(V_i) &\approx \ln(\alpha D_i^{\beta}) \\ &\approx \ln(\alpha) + \ln(D_i^{\beta}) \\ &\approx \ln(\alpha) + \beta \ln(D_i) \end{aligned}$$

which now can be treated as a linear regression model:

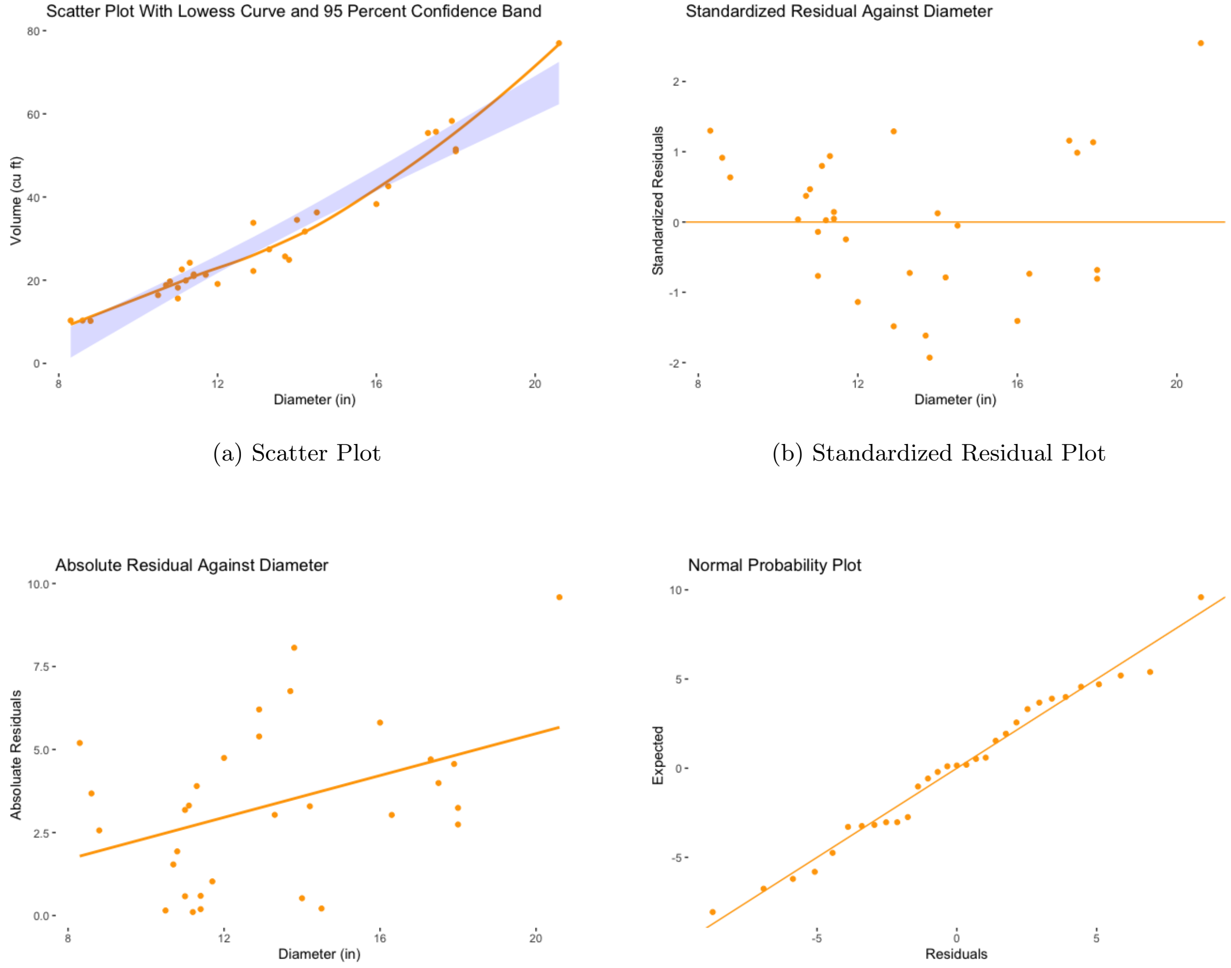
$$\text{Model 2 (Linearized):} \quad V_i^{'} = \alpha^{'} + \beta D_i^{'} + \varepsilon_i \tag{2}$$

where

- $\alpha^{'} = \ln(\alpha)$ and β are the parameters
- $D_i^{'} = \ln(D_i)$ are the log-transformed tree diameters
- $V_i^{'} = \ln(V_i)$ are the log-transformed volumes of lumber obtained from the trees after processing
- ε_i are independent $N(0, \sigma^2)$

3.2 Graphical Assessment of Model Assumptions

To examine the aptness of model 1, we present the following panel of plots.



(a) Scatter Plot

(b) Standardized Residual Plot

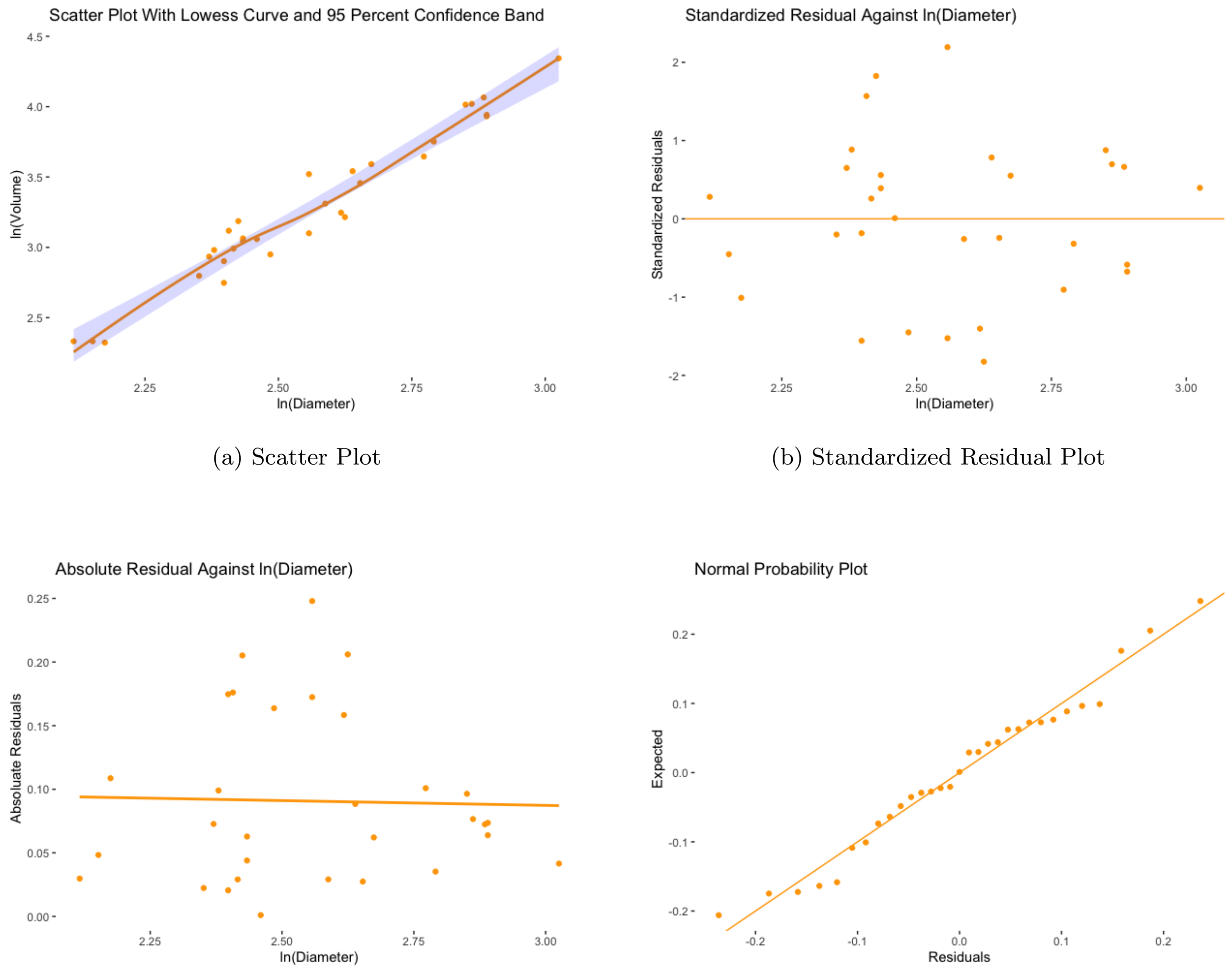
(c) Absolute Residual Plot

(d) Normal Probability Plot

Figure 3: Diagnostic Plots for Model 1

Figure 3 (a) plots the the smoothed curve obtained using the lowess method together with the Working-Hotelling 95% confidence band. As can be seen, the shape of the lowess curve has some curvature and the curve fails to fall within the 95% confidence band entirely, which suggests some non-linearity in the regression relation. The outlying observation observed from figure 2 (b), which is in the top right corner of figure 3 (a), will certainly have an effect on parameter estimates. Nonetheless, the general trend of the points in plot (a) is fairly linear. Figure 3 (b) shows no presence of outliers as there are no standardized residuals with absolute value of four or more. The standardized residuals in figure 3 (b) appear to be varying systematically with tree diameters; that is, variability appears to increase with tree diameter. Figure 3 (c) shows more clearly that the residuals tend to be larger in absolute value for wider trees. We will conduct formal tests on the constant variance condition. Lastly, Figure 3 (d)[3] shows no *substantial* departure from normality.

For model 2, we present the following panel of plots:



(a) Scatter Plot

(b) Standardized Residual Plot

(c) Absolute Residual Plot

(d) Normal Probability Plot

Figure 4: Diagnostic Plots for Model 2

Figure 4 (a) shows that the smoothed curve falls entirely within the 95% confidence band and thereby supports the appropriateness of model 2. In figure 4 (b), there is no presence of outliers, i.e. $|e_{standardized}| \geq 4$, and the standardized residuals are randomly scattered below and above the estimated regression line. In addition, the standardized residuals form a fairly constant band. Figure 4 (c) confirms that error term variance is constant as the absolute residuals do not appear to increase or decrease with the log-transformed diameter. Finally, figure 4 (d) shows no *substantial* departure from normality.

[3] We use the ratio recommended by Blom (1958), $\frac{3}{8}$, to find the expected values under normality.

3.3 Assessment of Model Assumptions Using Formal Tests

Table 2: Tests For Model 1

Test	**Hypotheses**	α	**Test Statistic**	**Criteria**	**Conclusion**
Correlation Test	H_0 : Normality H_a : Non-normality	0.1	Pearson's r = 0.9904	critical-value = 0.972	Conclude H_0
Anderson-Darling	H_0 : Normality H_a : Non-normality	0.1	A = 0.2703	p-value = 0.653	Conclude H_0
Lilliefors	H_0 : Normality H_a : Non-normality	0.1	D = 0.0989	p-value = 0.615	Conclude H_0
Breusch-Pagan	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BP = 4.099	p-value = 0.0429	Conclude H_a
Brown-Forsythe	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BF = 1.5726	p-value = 0.2198	Conclude H_0
We choose $\alpha = 0.1$ to be more conservative and not hasten to conclude H_0 in support of the model conditions. For the Brown-Forsythe test, we choose *diameter* ≥ 13 (the median in Table 1) as the threshold for grouping.					

Table 2 summarizes the formal test results for model 1. For model 1, the normality of the error terms is supported by the formal tests. For constancy of error term variance, the Breusch-Pagan test and the Brown-Forsythe test lead to conflicting conclusions. Specifically, the Breusch-Pagan test leads to our concluding that error variance is not constant while the Brown-Forsythe test leads to our concluding the opposite. However, we must note that the threshold for grouping the residuals in the Brown-Forsythe test is *arbitrarily* selected. Different thresholds tend to lead to different conclusions regarding the constancy of error variance. For instance, the Brown-Forsythe test using *diameter* ≥ 12 as the grouping threshold yields a p-value of 0.05065, which would lead to our concluding that error variance is not constant at the 10% significance level. For this particular sample, the evidence support the normality assumption but generally do not *conclusively* support the constancy of error variance condition. Non-constancy of error variance leads to less efficient estimates and invalid error variance estimates. The estimators of the model parameters in equation 1 remain unbiased and consistent but will no longer be efficient, i.e., BLUE (Best Linear Unbiased Estimator). A direct consequence of heteroskedasticity is that invalid error variance estimates lead to invalid or biased inferences, including confidence intervals and hypotheses tests. In the next subsection, we entertain some possible solutions to this problem.

Table 3 presents test results for model 2. As can be seen, all the model assumptions are satisfied. The normality of error term is not severely damaged. The Breusch-Pagan test now leads to our concluding that error variance is constant. We note further that the Brown-Forsythe test result is robust to different grouping thresholds. For model 2, we have strong evidence that error term variance is constant. More generally, the formal test results in Table 3 confirm our graphical assessment from section 3.2 that model 2 is appropriate.

Table 3: Tests For Model 2

Test	**Hypotheses**	α	**Test Statistic**	**Criteria**	**Conclusion**
Correlation Test	H_0 : Normality H_a : Non-normality	0.1	Pearson's r = 0.989	critical-value = 0.972	Conclude H_0
Anderson-Darling	H_0 : Normality H_a : Non-normality	0.1	A = 0.28851	p-value = 0.5937	Conclude H_0
Lilliefors	H_0 : Normality H_a : Non-normality	0.1	D = 0.093733	p-value = 0.6966	Conclude H_0
Breusch-Pagan	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BP = 0.14203	p-value = 0.7063	Conclude H_0
Brown-Forsythe	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BF = 0.64778	p-value = 0.4275	Conclude H_0
For the Brown-Forsythe test, we choose $\ln(diameter) \geq 2.56$ (the median) as the threshold for grouping.					

3.4 Heteroskedasticity and Remedial Measures for Model 1

Now that we have identified a problem of model 1, we turn to some possible solutions. Before we proceed, it behooves us to explain briefly the reason for needing to address the problem of non-constancy of error variance. The least squares estimators for the parameters in equation 1 will remain unbiased, consistent, and asymptotically normal despite heteroskedasticity. The cause for concern is due to the *inefficiency* of the estimators since we need to draw inferences and test hypotheses on them, using the standard errors. Recall that the standard errors are the standard deviations of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$. We can represent the sampling variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ in matrix expression as follows:

$$\sigma^2\{\hat{\boldsymbol{\beta}}\} = \sigma^2(\mathbf{X}^{'}\mathbf{X})^{-1} \tag{3}$$

$$= \begin{bmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{\sum(X_i-\bar{X})^2} & \frac{-\bar{X}\sigma^2}{\sum(X_i-\bar{X})^2} \\ \frac{-\bar{X}\sigma^2}{\sum(X_i-\bar{X})^2} & \frac{\sigma^2}{\sum(X_i-\bar{X})^2} \end{bmatrix} \tag{4}$$

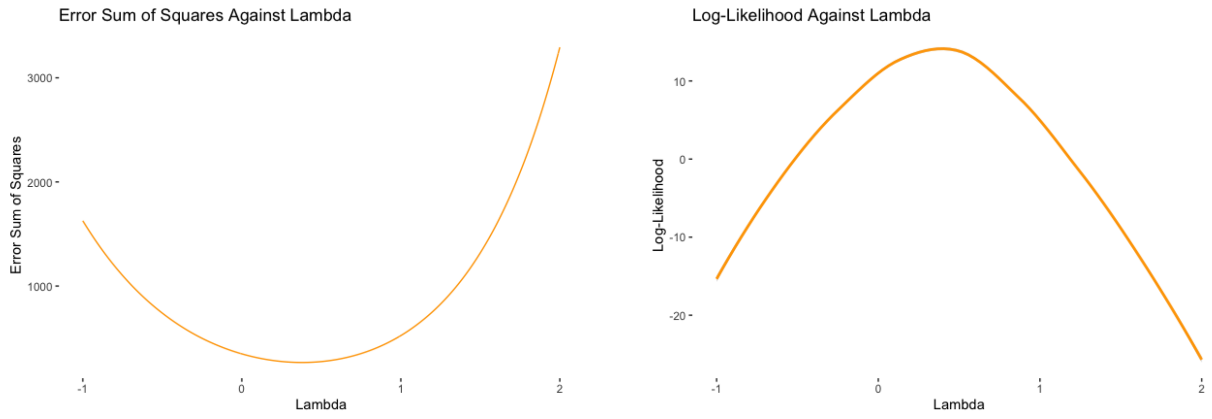
where the diagonal entries of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ are the sampling variances of $\hat{\beta}_1$ and $\hat{\beta}_0$. However, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is derived from

$$\sigma^2\{\hat{\boldsymbol{\beta}}\} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\sigma^2\{\mathbf{Y}\}\mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}$$

where $\sigma^2\{\mathbf{Y}\}$ is assumed to be $\sigma^2\mathbf{I}$. Under non-constant error variance, the assumption that the errors are homoskedastic here is implausible. Therefore, the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ by substituting MSE for σ^2 in equation 3 will be biased. The estimated sample variances, and, in turn, the estimated standard deviations for the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ will be biased. If we proceed to use the biased standard errors to calculate inferential statistics or construct interval estimates, our results may be misleading.

Transformation on V

One possible measure to stabilize non-constant error variance is through a transformation on the response variable V. The Box-Cox procedure[4] is used to identify an appropriate transformation from the family of power transformations.



(a) Error Sum of Squares

(b) Log-Likelihood

Figure 5: Box-Cox

From figure 5, the λ value for which the error sum of squares is a minimum and for which the log-likelihood value is a maximum falls between 0 and 0.5, which are associated with the natural log transformation and the square root transformation, respectively. In fact, model 2 is the log-log specification and so we will instead explore the square root transformation. [5]

Table 4 reports the test results for model 1 after the transformation. As can be seen, the Breusch-Pagan and Brown-Forsythe tests now yields conclusions that are consistent with the normal error model assumptions. This is accomplished without damaging normality as indicated by the p-values of the Anderson-Darling and Lilliefors tests.

While the transformation is simple to implement, it does come at the cost of complicating the model. The strength of model 1 lies in its ease of interpretation; that is, the level-level specification allows for

[4] A numerical search of potential λ values ranging from -1 to 2 is carried out in R.

[5] The exact λ value is 0.38. Note that this would further suggest that model 2 would be more appropriate, compared to model 1, for this particular sample of data.

Table 4: Tests For Model 1 After Transformation

Test	**Hypotheses**	α	**Test Statistic**	**Criteria**	**Conclusion**
Correlation Test	H_0 : Normality H_a : Non-normality	0.1	Pearson's r = 0.9945	critical-value = 0.972	Conclude H_0
Anderson-Darling	H_0 : Normality H_a : Non-normality	0.1	A = 0.17402	p-value = 0.9186	Conclude H_0
Lilliefors	H_0 : Normality H_a : Non-normality	0.1	D = 0.070287	p-value = 0.9605	Conclude H_0
Breusch-Pagan	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BP = 0.20141	p-value = 0.6536	Conclude H_a
Brown-Forsythe	H_0 : Constant Error Variance H_a : Non-constant Error Variance	0.1	BF = 0.22844	p-value = 0.6363	Conclude H_0
We choose $\alpha = 0.1$ to be more conservative and not hasten to conclude H_0 in support of the model conditions. For the Brown-Forsythe test, we choose *diameter* ≥ 13 (the median in Table 1) as the threshold for grouping.					

straightforward interpretations of the estimated coefficients and interval estimates. Any transformation, however simple, will alter the nature of the regression relation. It may be desirable to entertain another remedial procedure that addresses the issue of non-constant error variance without complicating an otherwise simple model.

Robust Standard Error

The heteroscedasticity consistent covariance matrix (HCCM), called the HC estimator following the terminology used by MacKinnon and White (1985),[6] may be used to estimate robust standard errors for the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. Recall the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\sigma^2\{\hat{\boldsymbol{\beta}}\} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\Omega\mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}$$

The HC estimator[7] addresses the problem of heteroskedasticity by entertaining estimates other than $\sigma^2\mathbf{I}$ for $\Omega = diag(\omega_1, ..., \omega_n)$ that is consistent in the presence of heteroskedasticity. Different HC estimators have been suggested in the literature; these estimators vary in their choice of the ω_i (for instance, under the normal error model, $\omega_i = \sigma^2$). Long & Ervin (2000)[8] uses a Monte Carlo simulation study of HC estimators (HC0

[6] MacKinnon, James G., and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." Journal of Econometrics 29 (3): 305–25. http://dx.doi.org/10.1016/0304-4076(85)90158-7

[7] Zeileis, Achim. "Econometric Computing with HC and HAC Covariance Matrix Estimators." Journal of Statistical Software, vol. 11, no. 10, 2004, doi:10.18637/jss.v011.i10.

[8] Long, J. Scott, and Laurie H. Ervin. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." The American Statistician, vol. 54, no. 3, 2000, p. 217., doi:10.2307/2685594.

to HC3, again, following the terminologies used by MacKinnon and White (1985)) in the linear regression model, recommending the use of HC3, which performs consistently for sample size as small as $N = 25$. For the HC3 estimator, the ω_i take the form:

$$\omega_i = \frac{e_i^2}{(1 - h_i)^2} \tag{5}$$

where e_i are the residuals and $h_i = H_{ii}$ are the diagonal elements of the hat matrix, defined as follows:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}$$

The diagonal elements of the hat matrix are the leverages, which describe the influence each response value has on the fitted value for that same observation.

Implementation of this correction using the HC3 estimator is simple in R.[9] We argue that this method may be more preferable for model 1, since it does not come at the cost of complicating the model. In the remaining sections, hypothesis testing for model 1 will be based solely on the test statistic computed using the robust standard error. Interval estimates for the parameter β_1 and the mean responses at different levels of tree diameter will be provided using both the robust standard errors and the non-corrected standard errors for comparison, if and when possible.

4 Regression Analysis

4.1 Estimated Regression Function and Regression Table

The estimated regression function for model 1:

$$\hat{V} = -36.9435 + 5.0659D \tag{6}$$

The estimated regression function for model 2:

$$\hat{V}' = -2.3533 + 2.2000D^{'} \tag{7}$$

For model 2, the estimated regression function in the original units:

$$\hat{V} = 0.0951D^{2.2000} \tag{8}$$

Table 5 summarizes the regression outputs for model 1 and model 2. A few important findings:

- For both model 1 and 2, the coefficients on the predictor variables of interest— diameter and ln(*diameter*) are statistically significant at the 1% significance level. Note that, for model 1, the t-tests are based on the robust standard errors (reported in the parentheses). In other words, we have evidence that

[9] The **sandwich** package implements the HCCM procedure in R. For more details https://cran.r-project.org/web/packages/sandwich/sandwich.pdf

there is a linear association between tree diameter and lumber volume and between the log-transformed diameter and the log-transformed volume. We caution that the linear associations here do not establish *cause-and-effect* relations between the predictors and the responses, since the data you provide are not obtained from a randomized controlled experiment. In the next subsection, we interpret these coefficients and provide a formula for estimating the change in the response variable for a prescribed change in the predictor variable.

- For model 1, the coefficient of determination is 0.935. We say that the variation in lumber volume is reduced by 93.5% when tree diameter is included in the model. For model 2, the R^2 value is 0.954. Due to the outlying observation identified in figure 2 (b), the R^2 for model 1 may be somewhat inflated. Nonetheless, the R^2 value may be interpreted as a descriptive measure for the *degree of linear association* between tree diameter and lumber volume and between the log-transformed diameter and the log-transformed volume. A caveat is that the R^2 alone is not an adequate measure for the usefulness of the models.[10] The next section will provide more inferences based on these results.

Table 5: Regression Results

	Response variable:	
	Volume	$\ln(Volume)$
	(1)	(2)
Diameter	5.066^{***}	
	(0.3707)	
$\ln(Diameter)$		2.200^{***}
		(0.090)
Constant	-36.943^{***}	-2.353^{***}
	(4.5916)	(0.231)
Observations	31	31
R^2	0.935	0.954
Adjusted R^2	0.933	0.952
Note:	* p $<$ 0.1; ** p $<$ 0.05; *** p $<$ 0.01	

[10] The scope of the regression model is restricted to the range of the predictor variable, represented by the max and min values of Diameter reported in Table 1. This is important to keep in mind when examining the regression results.

4.2 Sensitivity Analysis

The box plot of lumber volume in figure 2 (b) reveals an outlying observation, $V = 77$ (henceforth, case 31). To the extent that case 31 is a legitimate observation, we may wish to investigate further the manner in which it affects our regression results. Below we obtain the estimated regression function for model 1 based on the sample of 30 observations save case 31.

The new estimated regression function for model 1:

$$\hat{V} = -33.3104 + 4.7619D \tag{9}$$

Comparing equation 9 to equation 6, it is evident that $\hat{\beta}_0$ is greater ($-33.3104 > -36.9435$) and $\hat{\beta}_1$ is smaller ($4.7619 < 5.0659$) for the new estimated regression function based on the remaining 30 observations; in other words, the estimated regression line is pulled disproportionately towards the outlying value when case 31 is included in the analysis. The new R^2 value is slightly lower, 0.9303 compared to 0.9353 as reported in Table 5. Furthermore, a 99% prediction interval for a new V observation at $D = 20.6$ inches

$$52.88951 \leq V_{h(new)} \leq 76.67943$$

fails to contain case 31, $V = 77$. We therefore conclude that case 31 is an influential observation. This influential point will be brought along in our analysis as we do not have a valid reason to exclude it. However, we note that there is now a caveat to interpreting the results of our analysis on model 1. Simply put, the positive effect of tree diameter on volume represented by the coefficient from Table 5 may be subject to an upward bias.

5 Interval Estimation

5.1 Model 1: Variance-Covariance Matrix & HC3 Estimator Matrix

For comparison, we provide the variance-covariance matrix and the HC3 estimator. As mentioned earlier, we will use both to construct interval estimates for model 1.

$$s^2\{\hat{\boldsymbol{\beta}}\}_{original} = \begin{bmatrix} 11.3242 & -0.8107 \\ -0.8107 & 0.0612 \end{bmatrix} \quad \& \quad s^2\{\hat{\boldsymbol{\beta}}\}_{HC3} = \begin{bmatrix} 21.0823 & -1.6810 \\ -1.6810 & 0.1374 \end{bmatrix}$$

5.2 Model 1: Confidence Interval for β_1

We provide an estimation of, and a 95% confidence interval for, the increase or decrease in average lumber volume, $\Delta\hat{V}$, for a prescribed increase or decrease in tree diameter, ΔD. For model 1, we have the following equation

$$\Delta\hat{V} = 5.0659 \cdot \Delta D. \tag{10}$$

In words, equation 10 says that a ΔD inch(es) change in tree diameter is associated with a $5.0659 \cdot \Delta D$ cubic feet change in average volume of the lumber obtained from the tree after processing. The interpretation here can be modified for any prescribed increase or decrease in tree diameter. The 95% confidence interval for the change in mean lumber volume for a prescribed change in diameter then amounts to constructing the 95% confidence interval for β_1. For model 1, our confidence interval is specified as:

$$(\hat{\beta}_1 \; - \; t(.975 \; ; \; df = 29) \; s\{\hat{\beta}_1\} \; \leq \beta_1 \; \leq \; \hat{\beta}_1 \; + \; t(.975 \; ; \; df = 29) \; s\{\hat{\beta}_1\})$$

$$(4.5599 \; \leq \beta_1 \; \leq \; 5.5718)$$

With confidence coefficient 0.95, we estimate that mean lumber volume changes by somewhere between $4.5599 \cdot \Delta D$ and $5.5718 \cdot \Delta D$ cubic feet for a ΔD inch(es) change in tree diameter. **Caution**: the confidence coefficient is interpreted to mean that if repeated independent samples are taken with the same levels of X (the tree diameters) and a 95% confidence interval is constructed for each sample, 95% of the intervals will contain the true value of β_1. The confidence coefficient 0.95 here refers to the estimation procedure. For any given sample, however, we cannot say conclusively that the the true value of β_1 is contained in the interval. Furthermore, since the standard error may be biased under heteroskedasticity, the probability that this interval contains the true value of β_1 may not be 95%. We will proceed with the construction of subsequent intervals for model 1 with this fact in mind. For comparison, the same 95% confidence interval for β_1 constructed using the robust standard error is provided[11]:

$$(\hat{\beta}_1 \; - \; t(.975 \; ; \; df = 29) \; s\{\hat{\beta}_{robust}\} \; \leq \beta_1 \; \leq \; \hat{\beta}_1 \; + \; t(.975 \; ; \; df = 29) \; s\{\hat{\beta}_{robust}\})$$

$$(4.3076 \; \leq \beta_1 \; \leq \; 5.8241)$$

This interval is also interpreted as stating that we are 95% confident that the true value of the ΔD multiplier β_1 (from equation 10) lies within this interval; we use a method that produces intervals containing β_1 in 95% of the random samples. In addition, this interval is noticeably wider; however, this interval may be more robust in the presence of heteroskedasticity. We provide both intervals to improve model 1's usefulness in different situations. Since decisions related to scheduling tree-cutting and mill processing will have different requirements, depending on the situation and the level of precision required, either or both intervals may prove to be useful in the decision-making process.

5.3 Model 1: Confidence Interval for Mean Response

To give a rough gauge of production throughout the distribution of tree diameters, we provide an estimate of, and the 95% confidence interval for, the average volume of lumber at three tree sizes: small

[11] Computations are included in the R scripts.

$(D = 10)$, medium $(D = 14)$ and large $(D = 18)$.

$$\text{Small} = 13.7151\text{ft}^3$$
$$\text{Medium} = 33.9785\text{ft}^3$$
$$\text{Large} = 54.2420\text{ft}^3$$

The 95% confidence interval for $E\{Y_h\}$ at each tree size X_h =10, 14, and 18:

$$(\hat{Y}_h \; - \; t(.975\;;\;df = 29)\; s\{\hat{Y}_h\} \;\leq\; E\{Y_h\} \;\leq\; \hat{Y}_h \;+\; t(.975\;;\;df = 29)\; s\{\hat{Y}_h\})$$
$$(11.4478 \;\; \text{ft}^3 \;\leq\; E\{Y_{10}\} \;\leq\; 15.9824 \;\; \text{ft}^3)$$
$$(32.3710 \;\; \text{ft}^3 \;\leq\; E\{Y_{14}\} \;\leq\; 35.5861 \;\; \text{ft}^3)$$
$$(51.3751 \;\; \text{ft}^3 \;\leq\; E\{Y_{18}\} \;\leq\; 57.1088 \;\; \text{ft}^3)$$

Each of these intervals are interpreted as saying that we are 95% confident that the true value of mean lumber volume at each tree size group lies within their respective interval; we use a method that produces intervals that contain the true mean volume of lumber at each tree size in 95% of the random samples. Again, we emphasize that the true confidence coefficient associated with these intervals may not be 0.95 since we have issues with the constancy of error variance. We provide the same intervals constructed using the heteroskedasticity-consistent variance-covariance matrix. Specifically, we use the following formula to compute the variance of $\hat{Y}_h$:

$$\begin{aligned} \sigma^2\{\hat{Y}_h\} &= \mathbf{X}_h^{'}\sigma^2\{\boldsymbol{\beta}\}_{HC}\mathbf{X}_h \\ &= \begin{bmatrix} 1 & X_h \end{bmatrix} \begin{bmatrix} \sigma^2\{\hat{\beta}_0\} & \sigma\{\hat{\beta}_0, \hat{\beta}_1\} \\ \sigma\{\hat{\beta}_1, \hat{\beta}_0\} & \sigma^2\{\hat{\beta}_1\} \end{bmatrix} \begin{bmatrix} 1 \\ X_h \end{bmatrix} \\ &= \sigma^2\{\hat{\beta}_0\} + 2X_h\sigma\{\hat{\beta}_0, \hat{\beta}_1\} + X_h^2\sigma^2\{\hat{\beta}_1\} \end{aligned}$$

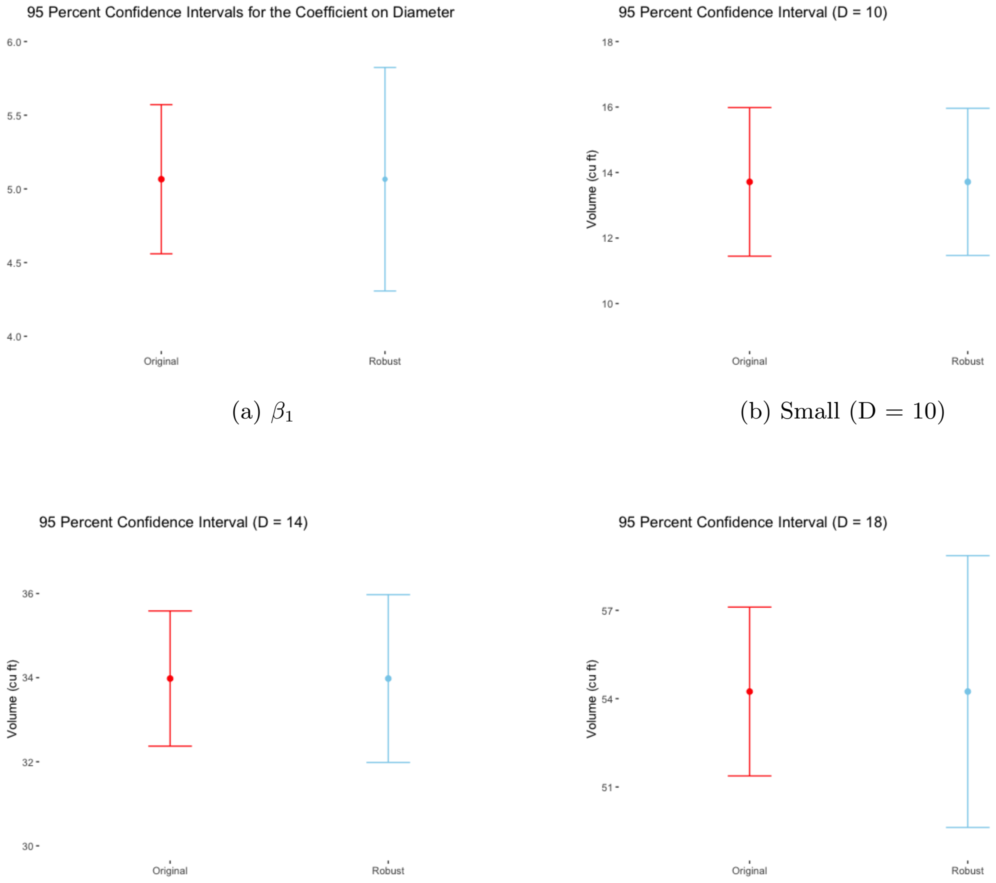
Taking the square-root of the variance of $\hat{Y}_h$, we obtain the estimated standard deviation $s\{\hat{Y}_h\}_{robust}$. The 95% confidence interval for the mean responses are constructed below.

$$(\hat{Y}_h \; - \; t(.975\;;\;df = 29)\; s\{\hat{Y}_h\}_{robust} \;\leq\; E\{Y_h\} \;\leq\; \hat{Y}_h \;+\; t(.975\;;\;df = 29)\; s\{\hat{Y}_h\}_{robust})$$
$$(11.4695 \;\; \text{ft}^3 \;\leq\; E\{Y_{10}\} \;\leq\; 15.9607 \;\; \text{ft}^3)$$
$$(31.9839 \;\; \text{ft}^3 \;\leq\; E\{Y_{14}\} \;\leq\; 35.9732 \;\; \text{ft}^3)$$
$$(49.6256 \;\; \text{ft}^3 \;\leq\; E\{Y_{18}\} \;\leq\; 58.8584 \;\; \text{ft}^3)$$

It may be surprising to find that the 95% confidence interval computed using the HC3 estimator (using the terminology detailed in section 3.4) is narrower for the small tree size group. This can be explained by the fact that the covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$ (highlighted in blue) is more negative in the HC3 estimator than in the original variance-covariance matrix under the constant-variance assumption (Section 5.1). For the medium and large tree size groups, the confidence intervals constructed using the HC3 estimator are both wider than the original intervals. For decision-making purposes, you may be pleased to find that, although the heteroskedasticity-consistent confidence intervals are wider, the differences in their widths compared to the original confidence intervals only become more noticeable for the large tree size group.

5.4 Model 1: Plot of Confidence Intervals

Visualization of the confidence intervals may contribute to the decision-making process in ways that numbers may not. In addition, a visual comparison of the confidence intervals calculated using both the corrected and non-corrected standard errors may be useful. We provide graphical representations of the confidence intervals for model 1.[12]



(a) β_1

(b) Small (D = 10)

(c) Medium (D = 14)

(d) Large (D = 18)

Figure 6: Confidence Interval Plots for Model 1

In figure 6, we present a panel of confidence interval plots. Panel (a) shows the the 95% confidence interval for β_1 calculated using both the uncorrected standard error and the robust standard error. The robust

[12] For model 2, graphical representation do not offer added benefits since the model does not suffer from heteroskedasticity and so interval estimates do not need to be computed twice using robust standard errors.

interval is wider than the original interval; however, the confidence coefficient for the robust interval may be more reliable given we have evidence of heteroskedasticity. Panels (b)-(d) display the 95% confidence interval for estimated mean responses at three tree size groups: small, medium, and large. The robust intervals are generally wider except for the small tree size group. As mentioned, the differences between the robust and original intervals become more pronounced as the trees under consideration become wider.

5.5 Model 1: Prediction Intervals for the Total Lumber Volume

Next, we estimate, and provide a 95% prediction interval for, the total lumber volume of a new pallet of sixteen trees at each tree size: small, medium, and large.

$$\begin{aligned}
\text{Small tree} &= 219.442 \text{ ft}^3 \\
\text{Medium tree} &= 543.6565 \text{ ft}^3 \\
\text{Large tree} &= 867.8713 \text{ ft}^3
\end{aligned}$$

The estimates reported above are obtained by multiplying the estimated mean volumes at each tree size $D=$ 10, 14, and 18 inches by 16. To find the 95% prediction intervals for the total lumber volume for a new pallet of sixteen trees at each tree size, we construct a 95% prediction interval for the mean volume of sixteen trees at each tree size, then multiply the limits by 16.

$$\begin{aligned}
&(169.1823 \text{ ft}^3 \le \text{Total Lumber Volume (16 small trees)} \le 269.7011 \text{ ft}^3) \\
&(500.3951 \text{ ft}^3 \le \text{Total Lumber Volume (16 medium trees)} \le 586.9179 \text{ ft}^3) \\
&(810.3034 \text{ ft}^3 \le \text{Total Lumber Volume (16 large trees)} \le 925.4392 \text{ ft}^3)
\end{aligned}$$

Caution: Unlike the confidence intervals, the prediction intervals are statements about the values to be taken by random variables. Thus, there is a great amount of uncertainty related to predictions. In addition, the confidence coefficient of these prediction intervals is not reliable due to non-constant error variance. Therefore, we recommend using these prediction intervals always with these facts in mind during the decision-making process.

5.6 Model 2: Confidence Interval for β

Model 2 is nonlinear, even though the regression relation between the log-transformed diameter and the log-transformed volume is linear. To see this

$$\begin{aligned}
\ln(V + \Delta V) - \ln(V) &= [\ln(\alpha) + \beta \ln(D + \Delta D)] - [\ln(\alpha) + \beta \ln(D)] \\
&= [\beta \ln(D + \Delta D) - \beta \ln(D)] \\
&= \beta[\ln(D + \Delta D) - \ln(D)]
\end{aligned}$$

Using natural logarithm rules and properties:

$$\ln\left(\frac{V+\Delta V}{V}\right) = \beta\ln\left(\frac{D+\Delta D}{D}\right)$$

$$e^{\ln(\frac{V+\Delta V}{V})} = e^{\beta\ln(\frac{D+\Delta D}{D})}$$

$$\cancel{e}^{\cancel{\ln}(\frac{V+\Delta V}{V})} = \cancel{e}^{\cancel{\ln}[(\frac{D+\Delta D}{D})^{\beta}]}$$

$$\frac{V+\Delta V}{V} = \left(\frac{D+\Delta D}{D}\right)^{\beta}$$

$$1+\frac{\Delta V}{V} = \left(1+\frac{\Delta D}{D}\right)^{\beta}$$

This provides us with the following formula:

$$\frac{\Delta\hat{V}}{V} = \left(1+\frac{\Delta D}{D}\right)^{\beta} - 1 \qquad (11)$$

Therefore, we say that a $\frac{\Delta D}{D}\cdot 100$ (that is, ΔD is a prescribed change in tree diameter and D is the original tree diameter) percent change in tree diameter is associated with a $[(1+\frac{\Delta D}{D})^{2.19997}-1]\cdot 100$ percent change in lumber volume. Furthermore, $\frac{\Delta\hat{V}}{V}$ is approximated by the following equation:

$$\frac{\Delta\hat{V}}{V} \cong 2.19997\frac{\Delta D}{D} \qquad (12)$$

which provides us with a more straight forward interpretation. Equation 12 states that a $(\frac{\Delta D}{D}\cdot 100)\%$ change in tree diameter is associated with a $(2.19997\cdot\frac{\Delta D}{D}\cdot 100)\%$ change in lumber volume obtained. Again, equation 12 allows for our entertaining any prescribed change in diameter. This approximation is most accurate when $\frac{\Delta D}{D}$ is small, and should only be used for such cases. For large changes in tree diameter, the exact methodology in equation 11 should be used.

The 95% confidence interval for the percentage change in mean volume given a prescribed percentage change in diameter amounts to finding the 95% confidence interval for the parameter β.

$$(\hat{\beta}\ -\ t(.975\ ;\ df=29)\ s\{\hat{\beta}\}\ \leq \beta\ \leq\ \hat{\beta}\ +\ t(.975\ ;\ df=29)\ s\{\hat{\beta}\})$$

$$(2.0162\ \leq \beta\ \leq\ 2.3837)$$

With this interval, we are 95% confident that the true value of the β (the multiplier from equation 12) lies within this interval; we use a method that produces intervals containing β in 95% of random samples. **Caution:** From equation 12, the parameter β is the ratio of the the percentage change in V associated with the percentage change in D. This should be interpreted as the *elasticity* of lumber volume with respect to tree diameter. In other words, the interpretation of the parameter is different for model 2 than for 1 and it should not be interchanged.

5.7 Model 2: Confidence Interval for Mean Responses

Applying the process in section 5.3 for model 2, an estimate of mean lumber volume at each tree size group is as follows:

$$\text{Small tree} = 15.0638 \text{ ft}^3$$

$$\text{Medium tree} = 31.5799 \text{ ft}^3$$

$$\text{Large tree} = 54.8941 \text{ ft}^3$$

The 95% confidence interval for the average volume of lumber at each tree size:

$$(14.1435 \text{ ft}^3 \leq E\{Y_{10}\} \leq 16.0440 \text{ ft}^3)$$

$$(30.1958 \text{ ft}^3 \leq E\{Y_{14}\} \leq 33.0275 \text{ ft}^3)$$

$$(50.9609 \text{ ft}^3 \leq E\{Y_{18}\} \leq 59.1308 \text{ ft}^3)$$

Each of these intervals are interpreted as saying that we are 95% confident that the true value of the average volume of lumber at each tree size lies within their respective interval; again, the interpretation of confidence intervals and risks of errors in inferences on the mean response is in terms of taking repeated samples. Note that, compared to the 95% confidence intervals in section 5.3, the confidence intervals computed for model 2 shift and change in width significantly. The confidence intervals for mean volumes at the small ($D = 10$) and the medium ($D = 14$) tree size groups are both more efficient (i.e. yielding tighter limits) for model 2 than for model 1. This is true for both the original and robust confidence intervals constructed for model 1. The confidence interval for the average volume of lumber at the large ($D = 18$) tree size group is narrower for model 1 than for 2. However, this is only true for the original confidence interval constructed using the non-corrected standard error. The robust confidence interval for mean lumber volume at the large tree size group for model 1 is wider than the corresponding confidence interval for model 2. In this sense, model 2 yields more efficient confidence intervals for the average lumber volume than does model 1.

5.8 Model 2: Prediction Intervals for the Total Lumber Volume

We estimate the total lumber volume of a new pallet of sixteen trees at each tree size group, multiplying the estimated mean volumes at each tree size group in section 5.7 by 16.

$$\text{Small tree} = 241.0204 \text{ ft}^3$$

$$\text{Medium tree} = 505.2789 \text{ ft}^3$$

$$\text{Large tree} = 878.3057 \text{ ft}^3$$

However, the 95% prediction intervals for the sum of the lumber volume of a new pallet of sixteen trees at each tree size group in the original units (ft^3) cannot be obtained since model 2 regresses the log-transformed volume on the log-transformed diameter. The limits for the 95% prediction interval will be a statement about the average *natural log* of volume of sixteen trees for a given tree size group. Multiplying the limits by sixteen as we did for model 1 will actually provide the limits for the sum of the *natural log* of lumber volume of sixteen trees at each tree size group. These limits are not equivalent to the limits for the sum of the volumes of sixteen trees obtained in section 5.5. Therefore, the best we could do to satisfy your request is to provide the 95% prediction intervals for the sum of the *natural log* of volume of sixteen trees at each tree size group, which, as we will show, is essentially the 95% prediction intervals for the product of the volume of sixteen trees at each tree size group in their original units. The prediction of the mean log of volume of sixteen trees can be constructed as follows:

$$\begin{array}{lcccl}
(\text{lower limit} & \leq & \text{Mean Natural Log of Lumber Volume} & \leq & \text{upper limit}) \\
(\text{lower limit} & \leq & \dfrac{\sum_{i=1}^{16}\ln(V_i)}{16} & \leq & \text{upper limit}) \\
(\text{lower limit} \cdot 16 & \leq & \dfrac{\sum_{i=1}^{16}\ln(V_i)}{\cancel{16}} \cdot \cancel{16} & \leq & \text{upper limit} \cdot 16) \\
(\text{lower limit} \cdot 16 & \leq & \displaystyle\sum_{i=1}^{16}\ln(V_i) & \leq & \text{upper limit} \cdot 16)
\end{array}$$

By the product rule of logarithm, the sum of the logarithms of the V_i is the logarithm of the product of the V_i:

$$\begin{array}{lcccl}
(\text{lower limit} \cdot 16 & \leq & \ln(\displaystyle\prod_{i=1}^{16} V_i) & \leq & \text{upper limit} \cdot 16)
\end{array}$$

When we exponentiate the lower and upper limits using base e, we are left with the product of the volume of sixteen trees at each tree size group (small, medium, and large) in their *original units*:

$$\begin{array}{lcccl}
(e^{\text{lower limit}\cdot 16} & \leq & \cancel{e}^{\cancel{\ln}(\prod_{i=1}^{16} V_i)} & \leq & e^{\text{upper limit}\cdot 16}) \\
(e^{\text{lower limit}\cdot 16} & \leq & \displaystyle\prod_{i=1}^{16} V_i & \leq & e^{\text{upper limit}\cdot 16})
\end{array}$$

The 95% prediction intervals for the sum of the natural logarithm of volume of sixteen trees at each tree size group:

(42.0176 $\leq$ Sum of the Natural Log of Lumber Volume (16 small trees) $\leq$ 44.7758)

(54.0577 $\leq$ Sum of the Natural Log of Lumber Volume (16 medium trees) $\leq$ 56.4230)

(62.5701 $\leq$ Sum of the Natural Log of Lumber Volume (16 large trees) $\leq$ 65.6029)

When we exponentiate using base e, the intervals above are equivalent to the 95% prediction intervals for the product of the volume of sixteen trees at each tree size group in their *original units*:[13]

$$
\begin{aligned}
&(1.770213 \times 10^{18} \;\leq\; \text{Product of Lumber Volume (16 small trees)} \;\leq\; 27.91638 \times 10^{18})\\
&(299.8863 \times 10^{21} \;\leq\; \text{Product of Lumber Volume (16 medium trees)} \;\leq\; 3.193017 \times 10^{24})\\
&(1.492265 \times 10^{27} \;\leq\; \text{Product of Lumber Volume (16 large trees)} \;\leq\; 30.97259 \times 10^{27})
\end{aligned}
$$

One possible use of these prediction intervals is to find the geometric mean at each tree size group. From the limits computed above, one can the find geometric mean of the individual volumes whose products equal the lower and the upper limit of the prediction intervals. We need to set up the following equation and solve for V:

$$V^{16} = \text{lower or upper limit}$$

To demonstrate, let's examine the limits for the small tree size group. We need to find an estimate for the V that equals the lower and upper limits of the prediction interval when we raise it to the power of 16:

$$
\begin{aligned}
V_{lower}^{16} &= 1.770213 \times 10^{18}\\
16\ln(V_{lower}) &= \ln(1.770213 \times 10^{18})\\
\ln(V_{lower}) &= \frac{\ln(1.770213 \times 10^{18})}{16}\\
V_{lower} &= \exp\left(\frac{\ln(1.770213 \times 10^{18})}{16}\right) \approx 13.8198\ ft^3
\end{aligned}
$$

And

$$
\begin{aligned}
V_{upper}^{16} &= 27.91638 \times 10^{18}\\
16\ln(V_{upper}) &= \ln(27.91638 \times 10^{18})\\
\ln(V_{upper}) &= \frac{\ln(27.91638 \times 10^{18})}{16}\\
V_{upper} &= \exp\left(\frac{\ln(27.91638 \times 10^{18})}{16}\right) \approx 16.4197\ ft^3
\end{aligned}
$$

Note that these are estimates, since the individual volumes will not be the same at each tree size group. Nonetheless, these geometric means may prove to be useful in the decision-making process to provide a sense of the range of possible values these random variables may take. We wish to emphasize once more that prediction intervals come with a great deal of uncertainty. While we have confidence in the estimation procedure and the model assumptions for model 2, we recommend using these predictions and calculations always with an understanding of the nature of their uncertainty in mind.

[13] As expected, these limits are astronomical numbers, and so they are represented using engineering notation.

6 Conclusion

The evidence in this report is in favor of model 2. In section 3.2, we find that model 1 suffers from lack-of-linearity and non-constant error variance. Section 3.3 Table 2 confirms our graphical analysis that model 1 indeed has non-constant error variance. For model 2, we note that all model assumptions are satisfied. In section 4, our regression results show that both tree diameter and the log-transformed tree diameter are significant predictors for lumber volume and the log-transformed lumber volume. In Table 5, the higher R^2 value for model 2 provides further indication that model 2 has the slight edge. Furthermore, the sensitivity analysis in section 4.2 reveals that model 1 is disadvantaged by an influential observation whose presence has a large effect on the parameter estimates. In section 3.4, we entertain possible measures to improve model 1 either with a transformation or with a heteroskedasticity-consistent variance-covariance estimator. Nonetheless, the strength of model 1 lies in its ease of interpretation. In section 5, we find that the interpretation of the coefficient on the predictor variables is more straightforward for model 1 than for model 2. In general, though, model 2 yields more efficient interval estimates than does model 1, as can be seen in sections 5.3 and 5.7. For the prediction intervals in sections 5.5 and 5.8, model 1 offers simplicity of interpretation while model 2 provides more reliable predictions as the model assumptions are well satisfied. All of these findings culminate to suggest that Model 1 is not without its own strength but that the model 2 should be used for predictive modeling.

There are many potential sources of error in regression analysis, but evidence shows that model 2 bypasses these sources. Since we know that the data are collected using proper random sampling techniques, we can safely assume that results obtained from our sample approximates what would have been obtained if the entire population of trees had been measured. Another source of error to consider in regression analysis is the possibility of omitted variable bias. This results when an omitted variable has an impact on the response variable and correlates with the predictor variables. However, it is difficult to imagine that there is omitted variable bias in model 2, especially considering the fact that the natural log of tree diameter leads to a 95.2% reduction in the variation of the natural log of lumber volume.

To conclude, we have evidence that model 2 is the predictive model that The Lumber Company Inc. should use when seeking to predict the volume of lumber obtained from a tree after processing. The strenghs of this model greatly outweigh any additional complications related to interpretations resulting from the log-log specification. We hope that you choose the us for any future statistical analysis requests.