Mathematical Sciences 885
Project I
January 16, 2007

1 Description of Data

The data set is taken from Ryan, Joiner and Ryan (1985), in the literature referred to as the Minitab blackcherry data set. The purpose for collecting these data was to provide a way of predicting the volume of timber in unfelled trees, from their height and diameter measurements, using a regression model. The initial model

$$volume \quad = \quad \beta_0 + \beta_1 diameter + \beta_2 height + \epsilon$$

The original data is given at the end of this report. Our first issues is to determine visually if the proposed model is appropriate. We include in Figure 1 a pairwise scatterplot of the variables. Taking into consideration specfically the visual relationship between the dependent variable ***volume*** and the preditor variables ***diameter*** and ***height***.

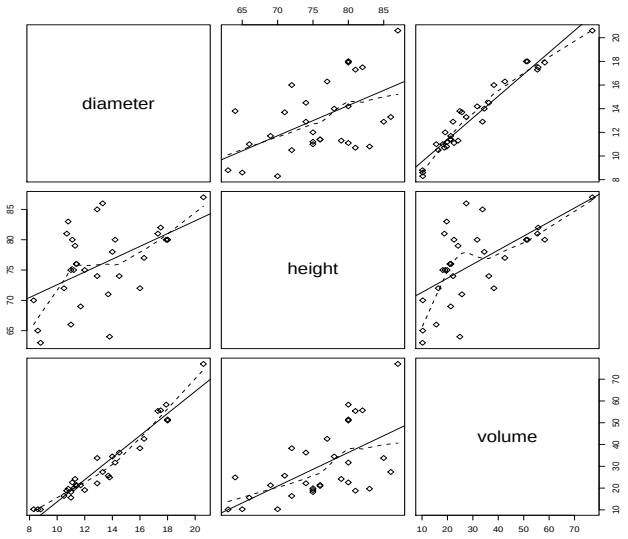

Figure 1: Scatterplot matrix of dependent and predictor variables.

2 Analysis

We first describe the initial multiple linear model. The summary below is taken directly for R summary window.

```
Call: lm(formula = volume ~ diameter + height, data =
blackcherry.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
```

```
diameter      4.7082     0.2643  17.816  < 2e-16 ***
height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,  Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value:     0
```

This suggest a significant regression of the two variables diameter and height in predicting tree volume. Further it suggests that 95% of the variation in the volume can be accounted for by the two variables diameter and height.

As suggested, the next stage would be an examination of the residuals generated by the fitted model. A plot of the residuals against each explanatory variable in the model. The presence of a curvilinear relationship, for example, may suggest that a higher-order term, perhaps a quadratic in the explanatory variable, should be added to the model. Figure 2 shows the standardized residuals plotted against values of explanatory variables. There appears to be a slight curvature in the plot of the residuals vs. the diameter. This is considered later.

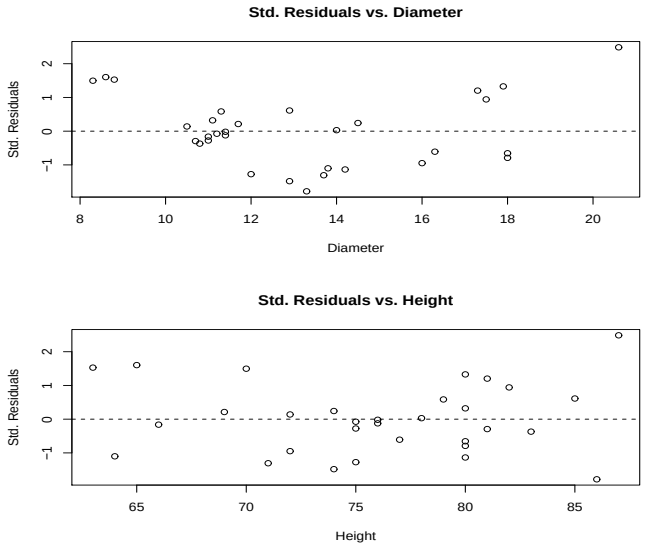

Figure 2: Standardized residuals plotted against values of explanatory variables.

A plot of the residuals against predicted values of the response variable. If the variance of the response appears to increase with predicted value, a transformation of the response may be in order. Figure 3 shows the standardized residuals plotted against the fitted values of the response variables.

A normal probability plot of the residuals. After all the systematic variation has been removed from the data, the residuals should look like a sample from the normal distribution. A plot of the ordered residuals against the expected order statistics from a normal distribution provides a graphical check of this assumption. We show a normal probability plot in figure 4.

The ordinary residuals given by residuals, however, have a distribution that is scale dependent since the variance of each ϵ_i is a function of both σ^2 and the diagonal values of the so-called “hat” matrix, $\mathbf{H}$, given by:

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

Consequently it is more useful to work with a standardized version of the residuals that does not depend on either of these quantities. The standardized residuals are calculated as

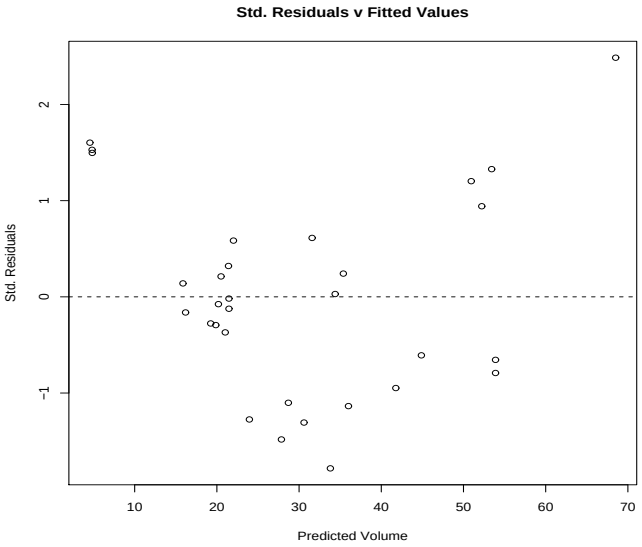

Figure 3: Standardized residuals plotted against values of fitted response.

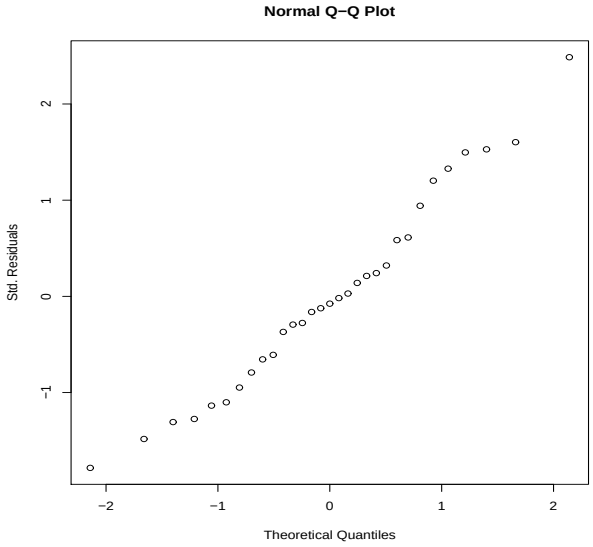

Figure 4: Normal probability plot of standardized residuals.

$$\mathbf{r_i} = \frac{y_i - \widehat{y_i}}{s\sqrt{1-h_{ii}}}$$

Examination of the enhanced normal probability plot (see Figure 5), indicates that the residuals show little departure from normality. Few lie outside the constructed confidence region.

The "hat" matrix is also helpful in identifying "strange" or "peculiar" data points, that is, those having an unusually large potential effect on the regression. Such points are indicated by relatively high values in the appropriate position in the diagonal of **H**. (The maximum value of any diagonal element is one.) Technically these points are referred to as having high leverage. As seen previously, the required diagonal values can be found as one of the elements in the list returned by the *lm.influence* function. The values of the diagonal of the hat matrix is given from **R** below:

```
   > diag(blackcherry.hat)
[1] 0.11582883 0.14720958 0.17686186 0.05919131 0.12066468 0.15575111
```

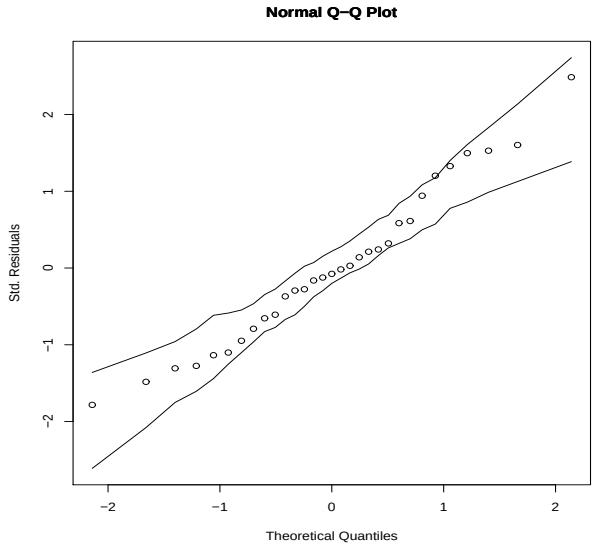

Figure 5: Enhanced normal probability plot of standardized residuals.

```
 [7] 0.11480262 0.05148096 0.09200658 0.04797237 0.07382512 0.04809206
[13] 0.04809206 0.07275901 0.03764563 0.03566543 0.13130916 0.14346152
[19] 0.06665975 0.21123665 0.03580935 0.04541796 0.04994875 0.11142518
[25] 0.06930648 0.08841762 0.09603041 0.10641665 0.10982638 0.10982638
[31] 0.22705852
>
```

None of these values appear to be significantly influential in the current model. In general, some form of index plot of these values is preferable to presenting them in a table. One such plot is to show the deviation of each component value of *h* from the average of the values. This type of plot is shown in Figure 6.

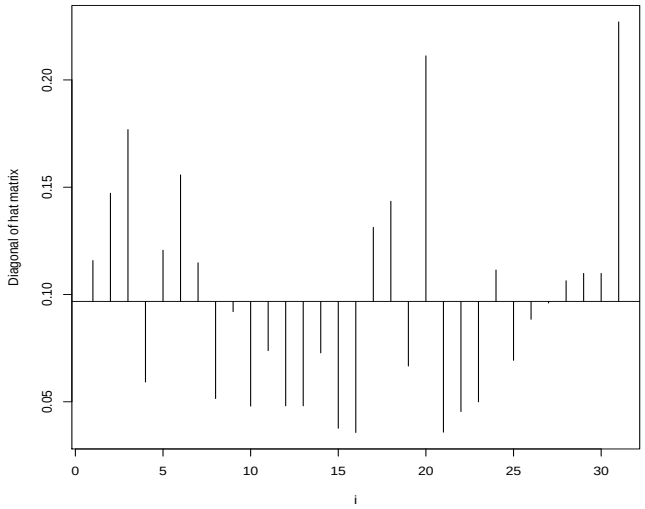

Figure 6: Index plot of leverage measures.

Here there seem to be no obvious problem points which might be unduly affecting the estimation process. The leverage values are all relatively low.

Returning now to the evidence from the residual plots, a new model involving a quadratic term in *Diameter* might now be considered. Such a model is fitted very simply using the *lm* function, although the extra term.

*Diameter*Diameter* needs to be enclosed in the identity function, **I**(), when specifying the model, to protect the special character,*.

The summary on the quadratic model is given below:

```
Call:
lm(formula = volume ~ diameter + I(diameter * diameter) + height)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -9.92041   10.07911  -0.984 0.333729
diameter              -2.88508    1.30985  -2.203 0.036343 *
I(diameter * diameter)  0.26862    0.04590   5.852 3.13e-06 ***
height                 0.37639    0.08823   4.266 0.000218 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771, Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value:     0
```

Similarly, we give residual plots for the quadratic models and see no significant departures for this assumed quadratic model.

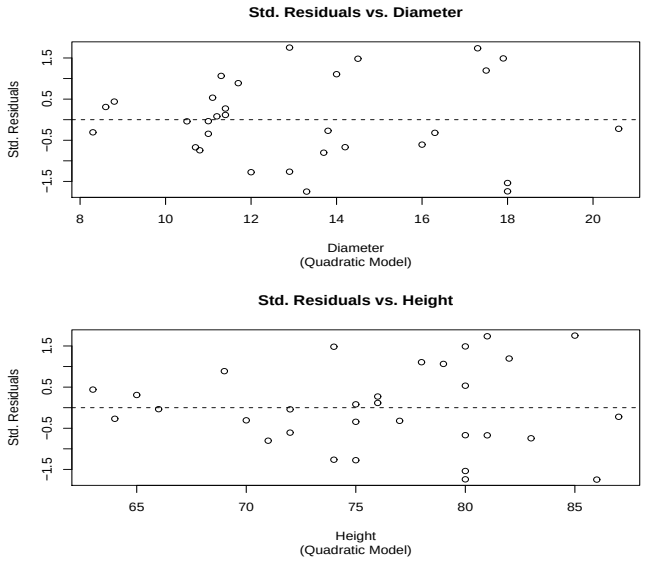

Figure 7: Standardized residuals plotted against values of explanatory variables model including quadratic terms in diameter.

Although the results in the previous summaries indicate that the regression coefficients of both Height and Diameter are significantly different from zero, it is often useful to explore a number of models in an attempt to find the simplest that adequately describe the data. Essentially this involves adding or deleting terms from an existing model and assessing the effect of the change. Two **R** functions, **add1** and **drop1**, can be used to look

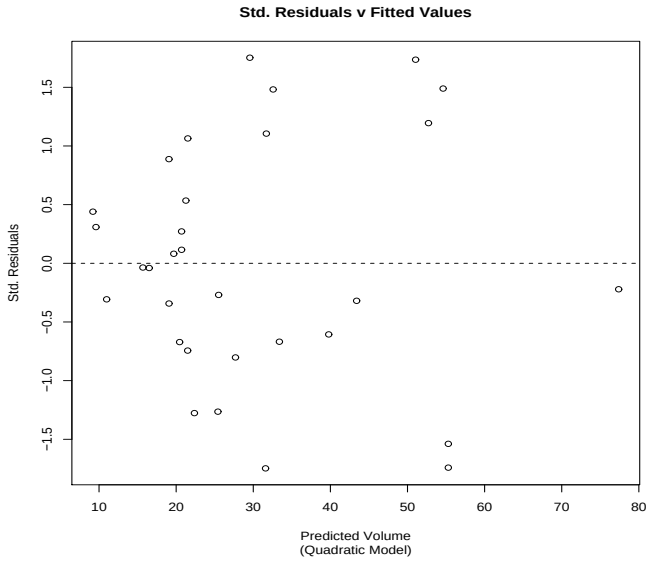

Figure 8: Standardized residuals plotted against fitted values of explanatory variables model including quadratic terms in diameter.

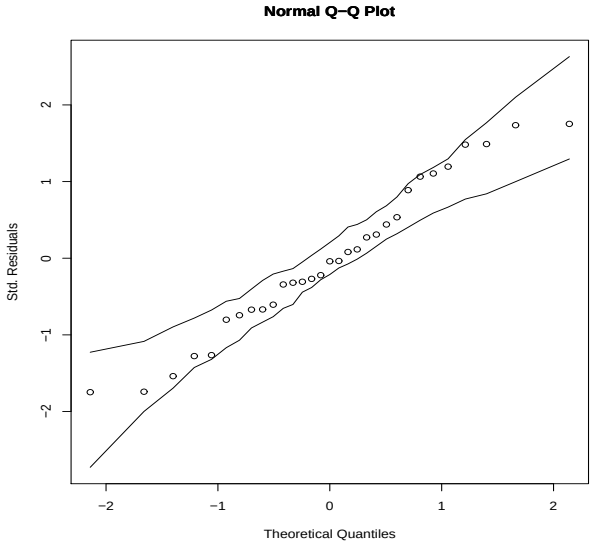

Figure 9: Normal plot of standardized residuals with simulated confidence interval for model including quadratic effect of diameter.

at the effects of adding or dropping single terms from a model. For example, *blackcherry.fit* involves a linear regression model with two explanatory variables *Diameter* and *Height*. The two models that can be formed from deleting either variable can each be examined by using:

```
> blackcherry.drop1 <-drop1(blackcherry.fit)
```

The information contained in *blackcherry.dropl* is detailed in the summary below. Sums of squares due to the deleted terms and residual sums of squares for the reduced model are given. Here their values indicate the great importance of Diameter in the model. Also given are the values of *AIC* (Akaike's Criterion). In models involving large numbers of explanatory variables, this statistic can be helpful in identifying important subsets.

```
> blackcherry.drop1
Single term deletions
```

```
Model:
volume ~ diameter + height
         Df Sum of Sq    RSS    AIC
<none>                 421.9   86.9
diameter  1    4783.0 5204.9  162.8
height    1     102.4  524.3   91.7
```

The opposite approach starts from a model and adds on terms. If, for example, the current model for the data is one involving only an intercept:

```
> blackcherry0.fit

Call: lm(formula = volume ~ 1, data = blackcherry.dat)

Coefficients:
(Intercept)
      30.17

> summary(blackcherry0.fit)

Call: lm(formula = volume ~ 1, data = blackcherry.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-19.971 -10.771  -5.971   7.129  46.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.171      2.952   10.22 2.75e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.44 on 30 degrees of freedom
```

This function is slightly different from the S-Plus function **add1**. The **add1** function in **R** assumed that the variables being added or removed are categorical in nature. We show only **drop1** here. To see what the **add1** function does in R, type ?add1 at the input cursor >.

The predicted values:

```
> predict(blackcherry1.fit,blackcherry1.dat,se.fit=T)
$fit
       1        2        3        4        5        6
18.20203 20.11100 29.82670 42.04388 78.06641 97.81696

$se.fit
        1         2         3         4         5         6
1.0041753 0.7772944 1.7471998 0.7039942 2.1687669 3.5557485

$df
[1] 27

$residual.scale
[1] 2.624753
```

Exercise

1. Several authors, including Sprent (1982) have commented on that the shape of a tree trunk is rather like that of a cone. Consequently, it might be sensible to consider models of the form

$$V = khd^2$$

Aitkinson (1987) suggests two such models

(a) (Volume)$^{1/3}$ on height and diameter.

(b) log(Volume) on log(height) and log(diameter)

Investigate both models.

Solution to Question 1

Question 1 arises from the discussion following the paper by Atkinson(1982).You should locate this and read it. Sprent in the same volume noted the following:

> Is there not then a danger that this very computational simplicity may make some users not very discriminating in how they apply the approaches? Not every user of robust regression or of diagnostic plots has the insight of an Andrews or an Atkinson. For example, do the users of Andrews' method always heed his warning about appropriate starting values? Least squares estimators can be disastrous for this. Is there not room for more thought about specific models when diagnostic plots suggest several that may be appropriate? Perhaps I can illustrate my point by reference to the tree data discussed in Section 6. A forester or biologist will certainly be attracted by the model of a tree as something very like a cone. This will be a very inexact model; it may be distorted by the degree of branching of each tree, or by just how much of the total superstructure is recorded as volume. Nevertheless a model that says volume, **v**, height, **h** and base diameter, **d** are related approximately by a formula
>
> $$v = kd^2h$$
>
> where k is a constant seems a good starting point. Taking logarithms immediately gives a linear relationship and would seem to provide a good basis for multiple regression exercises; but there is still plenty of room for diagnostics, for goodness knows what sort of error structure we create by taking logarithms, or even what it was before we took logarithms. Obviously some of Dr Atkinson's plots may help to sort this out as well as highlighting other difficulties. We would of course also be worried about our model if the estimated coefficients of log **d** and log**h** differed markedly from 2 and 1 respectively. Dr Atkinson does not tell us what his coefficients were when he fitted this model. He does tell us though that his methods indicate that when the explanatory variables are logged the transformation parameter for the response is $\widehat{\lambda}$ = - 0 0672 implying that something like logging the response variable is the right thing to do. I would have been happy starting off with the logged cone type of model and refining it if need be in the light of diagnostic checks even if this meant foregoing the excursion leading to regressions such as that of $v^{1/3}$ on **h** and **d**. Despite its dimensional correctness this appeals to me neither physically nor biologically. Perhaps when we get several models that are almost equally good statistically, a choice might be aided by cross-validation techniques when there are no biological, physical or other grounds to aid a choiceproviding we have sufficient data.

Below we give the summary and plots similar to those give previously for the model:

$$vol^{1/3} = \beta_0 + \beta_1 diameter + \beta_2 height + \epsilon$$

```
Call: lm(formula = (volume)^(1/3) ~ diameter + height, data =
blackcherry.dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.159602 -0.050200 -0.006827  0.069649  0.133981

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.085388   0.184315  -0.463    0.647
diameter     0.151516   0.005639  26.871  < 2e-16 ***
height       0.014472   0.002777   5.211 1.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08283 on 28 degrees of freedom
Multiple R-Squared: 0.9777, Adjusted R-squared: 0.9761
F-statistic: 612.5 on 2 and 28 DF,  p-value:     0
```

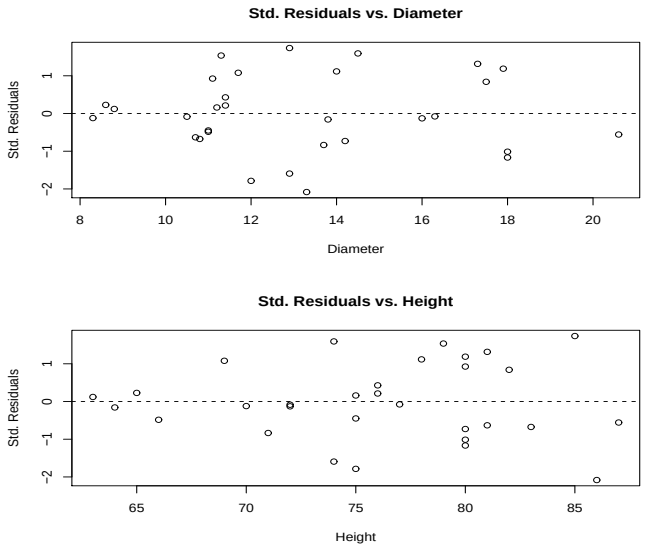

Figure 10: Standardized residuals plotted against values of explanatory variables model including model including cube root effect on volume.

We further consider the model:

$$log(volume) \quad = \quad \beta_0 + \beta_1 log(diameter) + \beta_2 log(height) + \epsilon$$

```
Call: lm(formula = log(volume) ~ log(diameter) + log(height), data =
blackcherry.dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

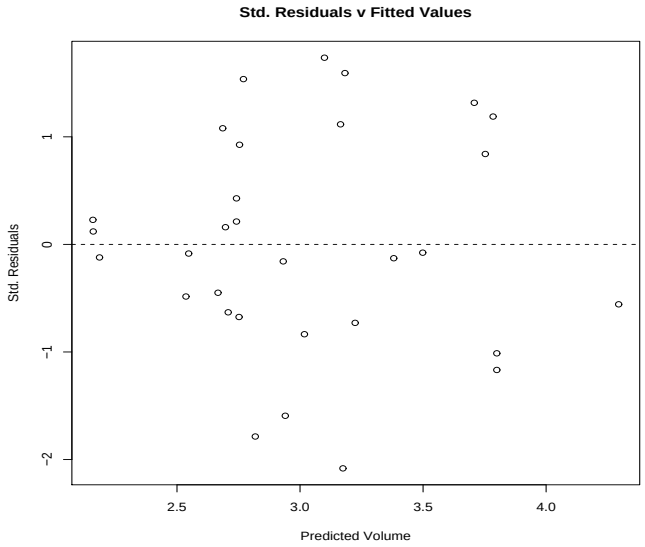

Figure 11: Standardized residuals plotted against fitted values of explanatory variables model including cube root effect on volume.

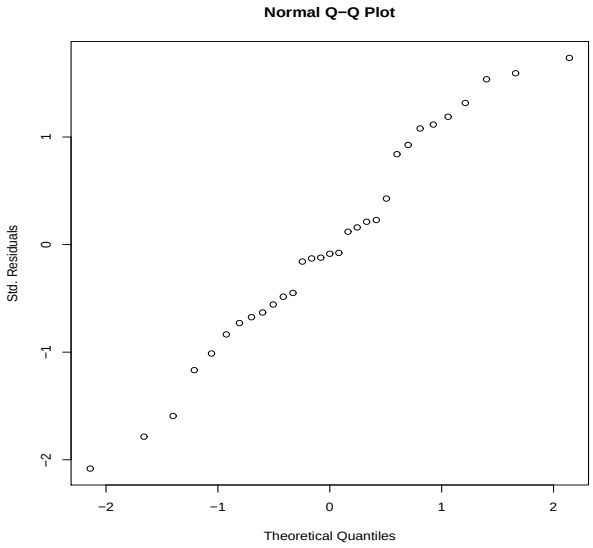

Figure 12: Normal plot of standardized residuals with simulated confidence interval for model including cube root effect on volume.

```
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(diameter)  1.98265    0.07501  26.432  < 2e-16 ***
log(height)    1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-Squared: 0.9777, Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value:     0
```

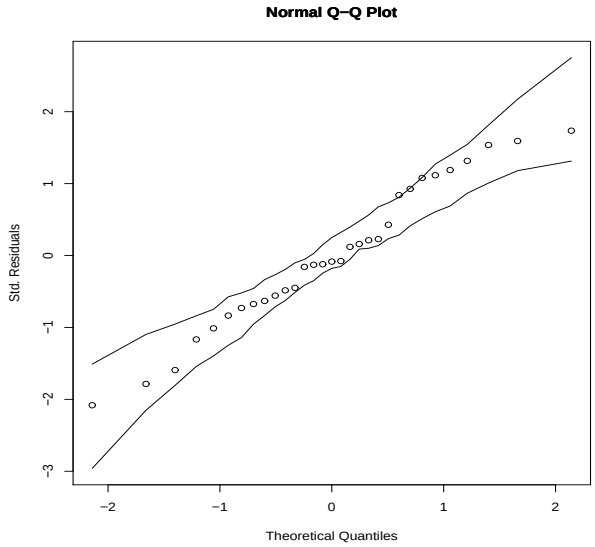

Figure 13: Normal plot of standardized residuals with simulated confidence interval for model including cube root effect on volume.

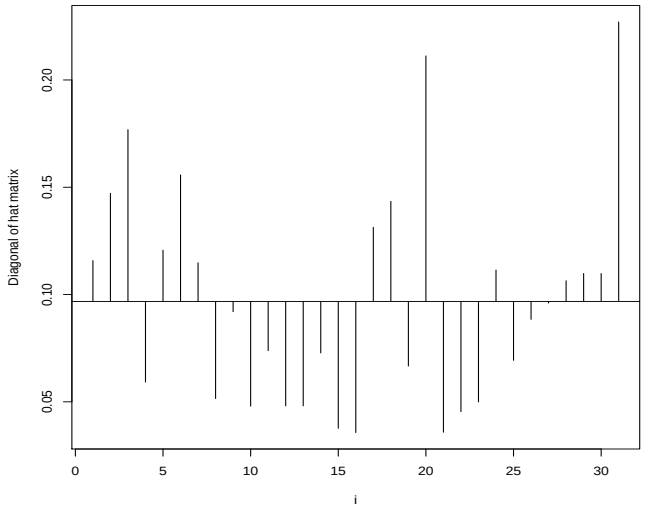

Figure 14: Normal plot of standardized residuals with simulated confidence interval for model including cube root effect on volume.

Finally, we consider all four of the models proposed.

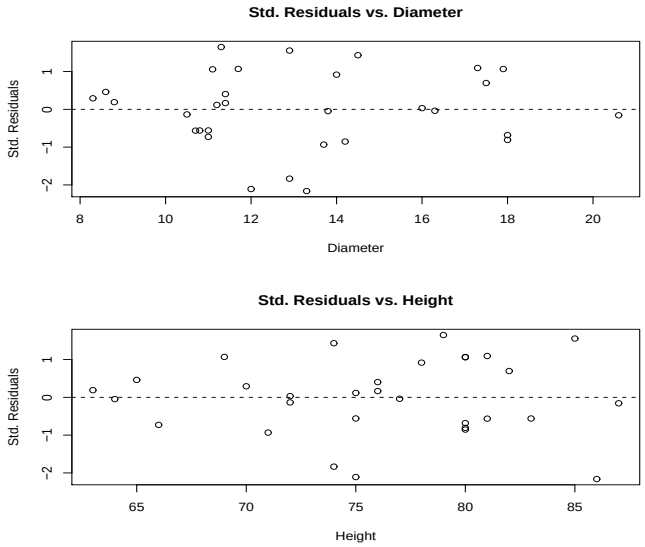

Figure 15: Standardized residuals plotted against values of explanatory variables model including log transformations on both the response, volume, and the predictors diameter and height.

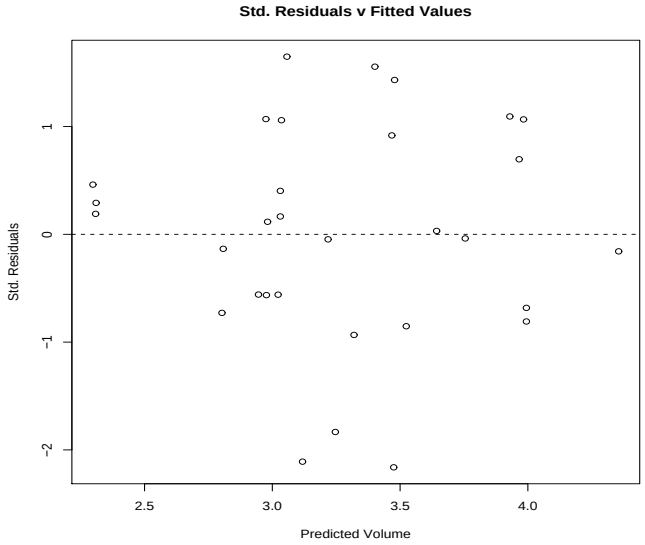

Figure 16: Standardized residuals plotted against fitted values of explanatory variables model including log transformations on both the response, volume, and the predictors diameter and height.

References

[1] Atkinson, A.C. (1982). "Regression Diagnostics, Transformations and Constructed Variables." *Journal of the Royal Statistical Society Series B*, **44**,pp. 1–36 (with discussion).

[2] Akaike, H. (1974). "A new look at statistical model identification." *IEEE Transactions on Automatic Control*, **AU-19**, 716-722.

[3] Ryan, T A., Joiner, B. L and Ryan, B. F. (1976). *Minitab Student Handhook*. North Scituate, Mass. Duxbury Press

[4] Sprent, P. (1982). "Discussion of Dr. Atkinson's paper." *Journal of the Royal Statistical Society Series B*, **44**,pp. 22–4.

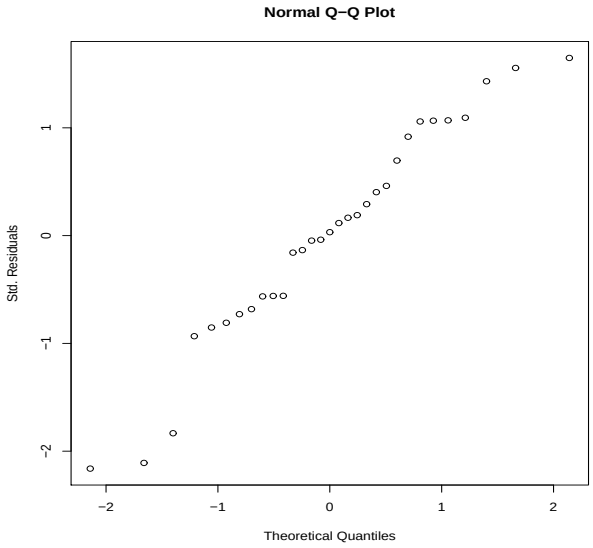

Figure 17: Normal plot of standardized residuals with simulated confidence interval for model including log transformations on both the response, volume, and the predictors diameter and height.

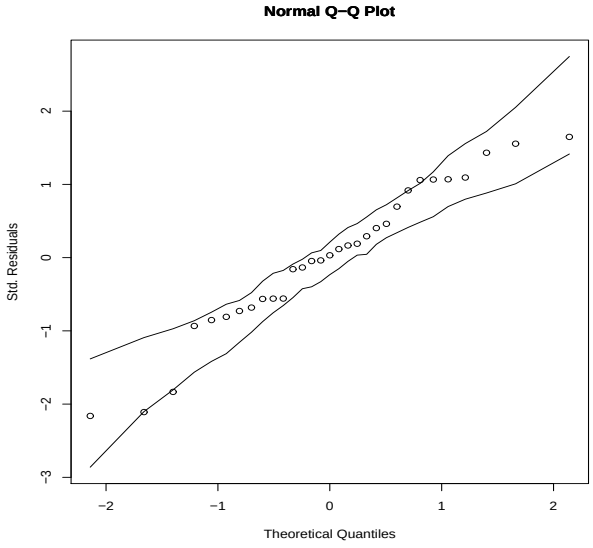

Figure 18: Normal plot of standardized residuals with simulated confidence interval for model including log transformations on both the response, volume, and the predictors diameter and height.

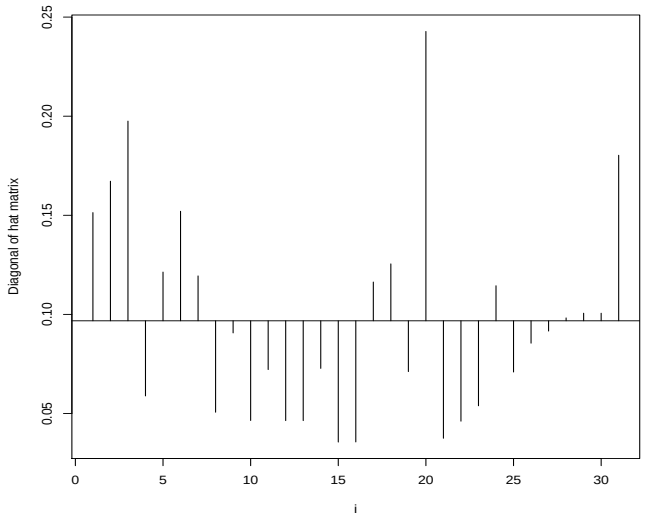

Figure 19: Normal plot of standardized residuals with simulated confidence interval for model including log transformations on both the response, volume, and the predictors diameter and height.

Model	Estimate (Standard Errors)	R^2	Significance
Model 1: v = β_0+β_1d + β_2h + ϵ	$\beta_0 = 57.9877(8.6382)$ $\beta_1 = 4.7082(0.2643)$ $\beta_2 = 0.3393(0.1302)$	0.948	< 0.001
Model 2: v = β_0+β_1d +β_2d^2 + β_3h + ϵ	$\beta_0 = -9.92041(10.07911)$ $\beta_1 = -2.88508(1.30985)$ $\beta_2 = 0.26862(0.04590)$ $\beta_3 = 0.37639(0.08823)$	0.9771	< 0.001
Model 3: v$^{1/3}$ = β_0+β_1d + β_2h + ϵ	$\beta_0 = -0.85388(0.184315)$ $\beta_1 = 0.151516(0.005639)$ $\beta_2 = 0.014472(.002777)$	0.9777	< 0.001
Model 4: log(v) = β_0+$\beta_1$log(d) + $\beta_2$log(h) + ϵ	$\beta_0 = -6.63162(0.79979)$ $\beta_1 = 1.98265(0.07501)$ $\beta_2 = 1.11712(0.20444)$	0.9777	< 0.001

Table 1: Four models considered for the Black Cherry Data

Table 2: Black Cherry Tree Data

diameter	height	volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0