Module 3 Wolfe

Tyler Wolfe

9/13/2020

## Problem 1:

The data file jevons in the alr4 package is a study of the weight of 274 gold coins that were collected in 1868 in Manchester, England. For each coin the weight after cleaning was recorded to the nearest 0.001 gram, and the date of the issue. The data includes a summary of this information the variable Age, which is the age of the coin in decades, n, the number of coins with that age, Weight, the average weight of the n coins of that age, SD, the standard deviation of the n coins of that age. The standard for new coins was 7.9876g, and legally the coins had to exceed 7.9379g.

Use weighted least squares to analyze the data appropriately to define the relationship between weight and age. Interpret the regression coefficients, and compare the result to using original least squares (no weights) and discuss the differences. How do your predictions at ages 1-5 relate to the legal minimum for a coin?

## Hint: Plot the relationship of SD to the mean to look to see if the variance is constant. Remember the Variance of any mean is SD^2/n, which you will need to create your weight.

First step is to open the data and explore the variables

library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

help(jevons)  
head(jevons)

## Age n Weight SD Min Max  
## 1 1 123 7.9725 0.01409 7.900 7.999  
## 2 2 78 7.9503 0.02272 7.892 7.993  
## 3 3 32 7.9276 0.03426 7.848 7.984  
## 4 4 17 7.8962 0.04057 7.827 7.965  
## 5 5 24 7.8730 0.05353 7.757 7.961

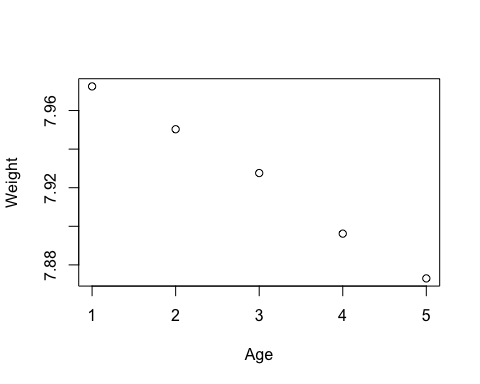
We see that the head only produces 5 observations instead of the typical 6 observations. Taking a look at the dimensions, we can see the size of the data set.

dim(jevons)

## [1] 5 6

We see that there are only 5 observations and 6 variables.

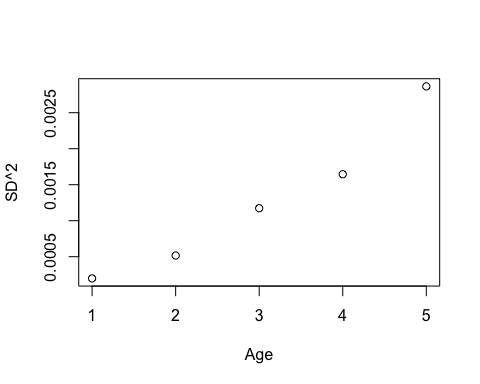
plot(Weight~Age,data=jevons)

 In the above plot, we see that there is a negative correlation characterized by the downward trend between Weight and Age, where when Age increases, Weight decreases. Understanding that the Weight variable, in relation to Age, in this dataset is an average of all of the coins of that age. For instance, the average Weight shown by Age=1 is representative of 123 observed different weighted coins that are 1 decade old. The Weight data point for Age=2 is a representatin of the average weight for 78 observed coins that are 2 decades old and etc.

When we revisit our head from above we see that the standard deviation is also shown, as well as the minimum weight and maximum weights observed within the individual vectors, as sorted by age. For example, we can see that the minimum weight of all 123 observations in Age=1 range from 7.900 grams - 7.999 grams, but are represented in the plot above as the mean of 7.9725 iwith a Standard Deviation of .01409.

As we consider the overall implications on the data because of the inequality of the data points and the observations they represent, it may be worth looking into the variances.

plot(SD^2~Age, data=jevons)

 By plotting the Standard Deviation squared in comparison with Age, we see that as the age increases, so does the the variance. Because the variance is not constant (because the data points do not form a horizontal line across the plot), we will claculate the variance of each data point in Plot(Weight~Age, data=jevons). To do so we will use .

model1<-lm(Weight~Age,weights=n/(SD^2),data=jevons)  
summary(model1)

##   
## Call:  
## lm(formula = Weight ~ Age, data = jevons, weights = n/(SD^2))  
##   
## Weighted Residuals:  
## 1 2 3 4 5   
## -0.2091 0.5017 0.3875 -0.5383 -0.4339   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.9965218 0.0013220 6049 9.96e-12 \*\*\*  
## Age -0.0237562 0.0008797 -27 0.000111 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5549 on 3 degrees of freedom  
## Multiple R-squared: 0.9959, Adjusted R-squared: 0.9945   
## F-statistic: 729.2 on 1 and 3 DF, p-value: 0.0001114

In the summary of our model, we see that as Age increases by one unit, we should expect weight to decrease by .02 grams. The p-value of .0001 is smaller than .05 which confirms our inference from above that Age and Weight have a relationship. Before we plot the least squares model, lets look at a simple linear regression.

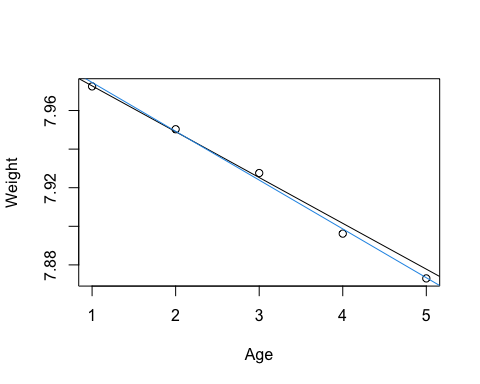
model2<-lm(Weight~Age,data=jevons)  
summary(model2)

##   
## Call:  
## lm(formula = Weight ~ Age, data = jevons)  
##   
## Residuals:  
## 1 2 3 4 5   
## -0.00204 0.00107 0.00368 -0.00241 -0.00030   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.9998500 0.0030123 2655.74 1.18e-10 \*\*\*  
## Age -0.0253100 0.0009082 -27.87 0.000101 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.002872 on 3 degrees of freedom  
## Multiple R-squared: 0.9962, Adjusted R-squared: 0.9949   
## F-statistic: 776.6 on 1 and 3 DF, p-value: 0.0001014

Comparing the two summaries, we see that there are some discrepencies. These differences in coefficients are because model1 is including a weighted least squared model accounting for non-constant variance, where model 2 is only comparing the relationship between Weights and Age assuming constant variance.

Lets plot the models to see the differences visually.

plot(Weight~Age,data=jevons)  
abline(model1)  
abline(model2,col=4)

 As seen in the above plot, the two lines (for model1 (black) and model2 (blue)) vary only by small amounts. Because the standard deviation of the Age=1 is smaller, and slowly grows as we approach Age=5, we can be more certain that we can predict in Age=1. The variance becomes less predictable as we move to the right on the x-axis towards Age=4 and Age=5.

So lets consider when currency should be taken out of circulation to remain 95% confident that the currency still means the legal weight of 7.9379 grams.

predAge1=predict(model1, newdataWLS = data.frame(weight=7.9379), interval = "confidence")  
predAge1

## fit lwr upr  
## 1 7.972766 7.970603 7.974928  
## 2 7.949009 7.946302 7.951717  
## 3 7.925253 7.920188 7.930319  
## 4 7.901497 7.893773 7.909222  
## 5 7.877741 7.867284 7.888198

Concerning predictability of whether the coins maintain the legal minimum for circulation, we can be 95% confident that after one or two decades of circulation currency still meets the legal specifications. However, after two decades of circulation we see that the upper bounds of the predicted weight for Age=3, 4, and 5 illustrate that the coins are not in compliance with minimum legal specifications. By this rational, we should remove the coins after 20 years of circulation.

## Problem 2:

## Using the stopping data in the alr4 package, fit a model that estimates the stopping distance (Distance) given the speed is known (Speed). Create a scatter plot, and then investigate a possible transformation that allows OLS to fit this model. Interpret the model and present any insights you may have.

To begin analyzing the data, I will examine the dimensions, head and plot the data in order to understand the data, prior to determining the appropriate transformation.

dim(stopping)

## [1] 62 2

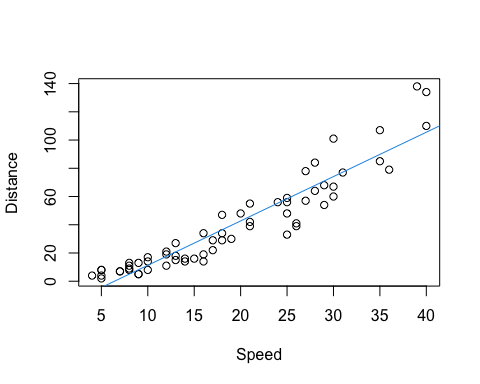
help(stopping)

By opening the help file associated with the data, we see that the speed is measured in mph, while distance is measured in feet.

head(stopping)

## Speed Distance  
## 1 4 4  
## 2 5 2  
## 3 5 4  
## 4 5 8  
## 5 5 8  
## 6 7 7

plot(stopping)  
modeld<-lm(Distance~Speed, data=stopping)  
abline(modeld, col=4)

 After plotting the data, we observe that the distance interprets as higher than on order of magnitude. Under the “Log Rule,” which states that if the values of a variable range by more than one order of magnitude and the variable is strictly positive, then use of a log transformation is appropriate. Since all observations of distance are positive, this applies.

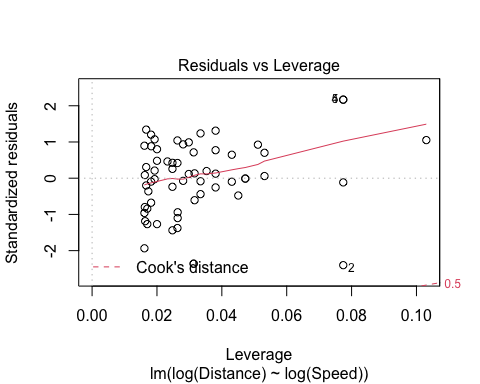
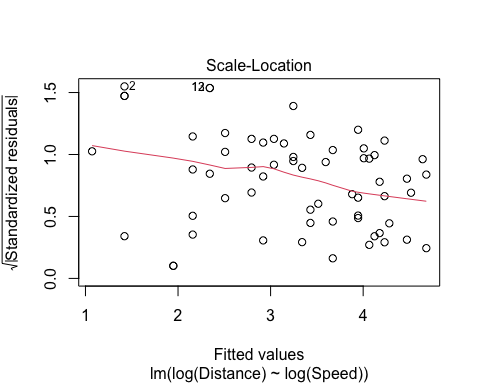
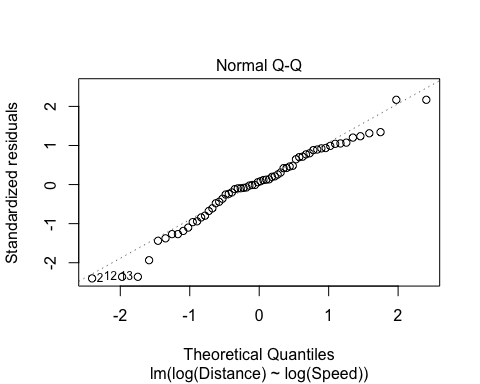
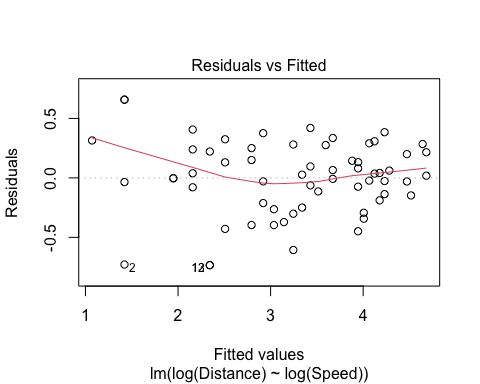
summary(modeld)

##   
## Call:  
## lm(formula = Distance ~ Speed, data = stopping)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.410 -7.343 -1.334 5.927 35.608   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -20.1309 3.2308 -6.231 5.04e-08 \*\*\*  
## Speed 3.1416 0.1514 20.751 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.77 on 60 degrees of freedom  
## Multiple R-squared: 0.8777, Adjusted R-squared: 0.8757   
## F-statistic: 430.6 on 1 and 60 DF, p-value: < 2.2e-16

modeldLog<-lm(log(Distance)~log(Speed), data=stopping)  
summary(modeldLog)

##   
## Call:  
## lm(formula = log(Distance) ~ log(Speed), data = stopping)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73374 -0.17858 0.02258 0.23500 0.65795   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.10221 0.19003 -5.80 2.64e-07 \*\*\*  
## log(Speed) 1.56806 0.06683 23.46 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3158 on 60 degrees of freedom  
## Multiple R-squared: 0.9017, Adjusted R-squared: 0.9001   
## F-statistic: 550.6 on 1 and 60 DF, p-value: < 2.2e-16

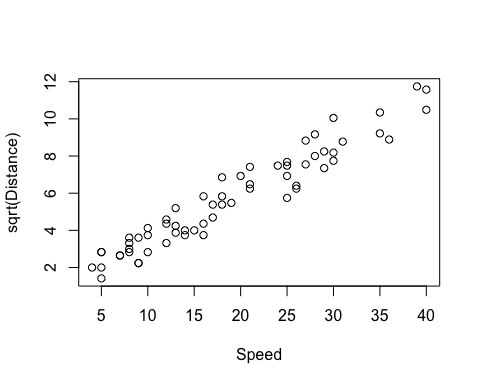
plot(modeldLog)

 After transforming the model to log(Distance) as response and log(Speed) as predictor, we can see by our residuals vs. fitted plot that there is an obvious curve among our residuals. As a square root transformation is a common practice for stabilizing variance among residuals, I will attempt to execute a square root transformation model, first only transforming the response, then moving on to transform the predictor if necessary to further stabilize variance in the residuals.

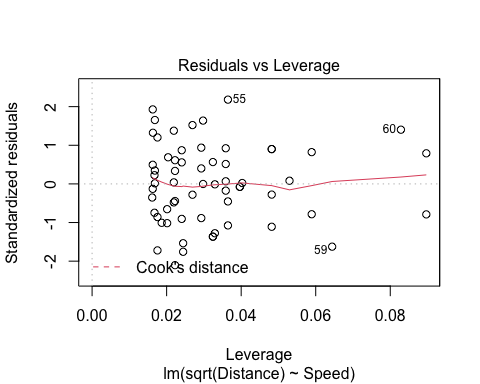
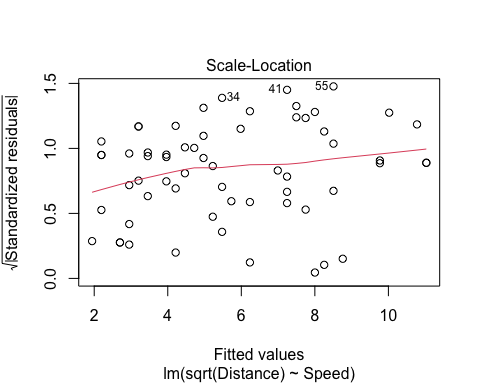
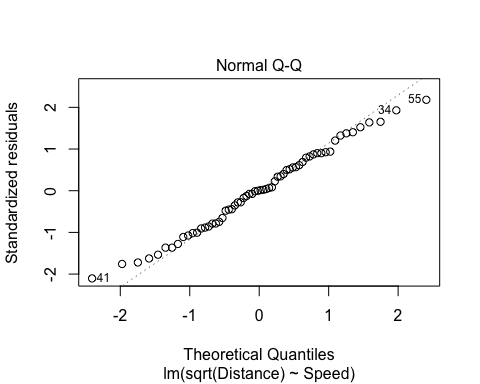
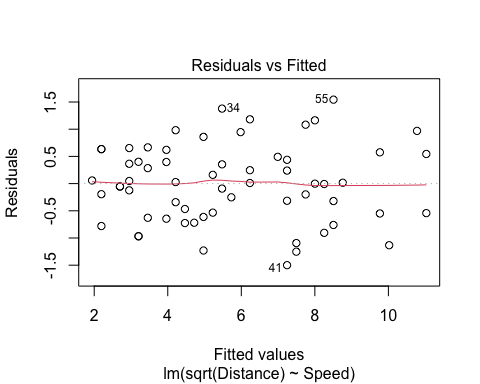
modeld3<-lm(sqrt(Distance)~Speed, data=stopping)  
summary(modeld3)

##   
## Call:  
## lm(formula = sqrt(Distance) ~ Speed, data = stopping)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.49948 -0.54761 0.00469 0.53153 1.54350   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.932396 0.197909 4.711 1.5e-05 \*\*\*  
## Speed 0.252466 0.009274 27.223 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7209 on 60 degrees of freedom  
## Multiple R-squared: 0.9251, Adjusted R-squared: 0.9239   
## F-statistic: 741.1 on 1 and 60 DF, p-value: < 2.2e-16

plot(sqrt(Distance)~Speed, data = stopping)



plot(modeld3)



in comparison, the new residuals vs. fitted based on the square root transformation seems to have stabilized the variance in the residuals.

We see in the summary of modeld3, that we can expect that for every one mph of speed, we can expect the necessary stopping distance to increase by .25 feet.

Unfortunately, using a square root transformation also is known to create issues for interpretations. For instance, in the above discussion concerning the relationship between speed and distance, we cannot infer that 1 mph increase will also increase stopping distance by .25ft^2 (.0625 ft) because there is no way to transform the result to an accurate number.

Hopefully soon we will learn another alternative method of stabilizing variance within residuals and conversion to linear data that does not create issues with inference.

## Problem 3:

## Please look at the help file for the data in this problem. We have been asked to analyze the MinnLand data in the alr4 package. For now we will analyze it without using the region or financing variables included in the data (DO NOT USE THIS IN YOUR MODEL YET). Using all other variables, fit a model to explain acrePrice using all other variables. Decide on appropriate transformations and diagnostic methods for the data. Fit your final model and provide any insights it provides on the data. You should include any interpretation you can provide.

First I will load the data and explore dimensions and identify variables in the data by examining the head.

library(alr4)  
dim(MinnLand)

## [1] 18700 9

head(MinnLand)

## acrePrice region improvements year acres tillable financing crpPct  
## 1 766 Northwest 0 2002 82 94 title\_transfer 0  
## 2 733 Northwest 0 2003 30 63 title\_transfer 0  
## 3 850 Northwest 4 2002 150 47 title\_transfer 0  
## 4 975 Northwest 0 2003 160 86 title\_transfer 0  
## 5 886 Northwest 62 2002 90 NA title\_transfer 0  
## 6 992 Northwest 30 2003 120 83 title\_transfer 0  
## productivity  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

help(MinnLand)

In the above printouts, we see that the data has 18700 observations over 9 vectors. In the head we observe that the help file was correct in saying that not all of the vectors contain exclusively values, as some counties in the state of Minnesota have not factored in productivity or had land surveyed to account for if the land is tillable by an assessor. Lets execute a take a look at the summary of the data.

summary(MinnLand)

## acrePrice region improvements year   
## Min. : 108 Northwest :3799 Min. : 0.000 Min. :2002   
## 1st Qu.: 1425 West Central :3297 1st Qu.: 0.000 1st Qu.:2004   
## Median : 2442 Central :4198 Median : 0.000 Median :2006   
## Mean : 2787 South West :2583 Mean : 4.493 Mean :2006   
## 3rd Qu.: 3702 South Central:2832 3rd Qu.: 0.000 3rd Qu.:2008   
## Max. :15000 South East :1991 Max. :100.000 Max. :2011   
## NA's :50   
## acres tillable financing crpPct   
## Min. : 1.0 Min. : 0.00 title\_transfer :16601 Min. : 0.000   
## 1st Qu.: 47.0 1st Qu.: 72.00 seller\_financed: 2099 1st Qu.: 0.000   
## Median : 80.0 Median : 92.00 Median : 0.000   
## Mean : 112.7 Mean : 80.67 Mean : 4.163   
## 3rd Qu.: 153.0 3rd Qu.: 97.00 3rd Qu.: 0.000   
## Max. :6970.0 Max. :100.00 Max. :100.000   
## NA's :1212   
## productivity   
## Min. : 1.00   
## 1st Qu.:59.00   
## Median :68.00   
## Mean :66.63   
## 3rd Qu.:76.00   
## Max. :99.00   
## NA's :9717

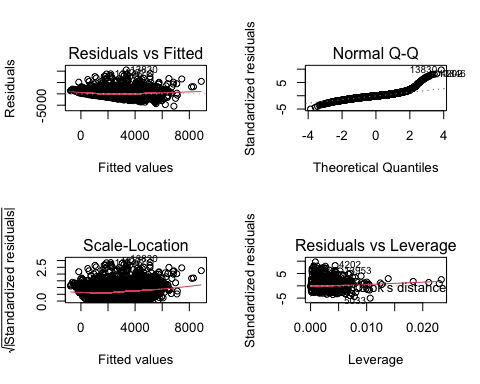
In the summary, we see that several values have a magnitude of 1 or higher so we could consider using log transformation. But first, lets execute a non-transformed linear regression.

modela<-lm(acrePrice~improvements+year+acres+tillable+crpPct+productivity, data=MinnLand)  
summary(modela)

##   
## Call:  
## lm(formula = acrePrice ~ improvements + year + acres + tillable +   
## crpPct + productivity, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5414.6 -573.8 -85.6 429.8 10544.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.809e+05 9.651e+03 -70.554 < 2e-16 \*\*\*  
## improvements 5.437e+01 1.403e+00 38.742 < 2e-16 \*\*\*  
## year 3.393e+02 4.812e+00 70.513 < 2e-16 \*\*\*  
## acres -1.979e+00 1.323e-01 -14.960 < 2e-16 \*\*\*  
## tillable 6.687e+00 8.263e-01 8.092 6.65e-16 \*\*\*  
## crpPct -9.822e+00 8.759e-01 -11.214 < 2e-16 \*\*\*  
## productivity 4.064e+01 9.383e-01 43.314 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1087 on 8763 degrees of freedom  
## (9930 observations deleted due to missingness)  
## Multiple R-squared: 0.5104, Adjusted R-squared: 0.51   
## F-statistic: 1522 on 6 and 8763 DF, p-value: < 2.2e-16

After executing the model, we should look further into our residuals.

par(mfrow=c(2,2))  
plot(modela)

 In our residuals vs. fitted plot we can see that there is a curve in our residuals, indicating that we should try to transform the data in order to execute a more accurate regression model.

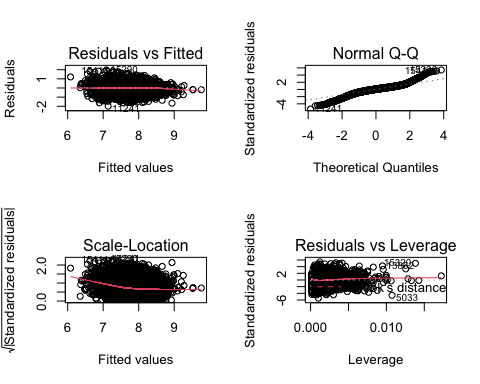
So lets try a log transformation.

modelaLog<-lm(log(acrePrice)~improvements+year+log(acres)+log(tillable)+crpPct+productivity, data=MinnLand)  
summary(modelaLog)

##   
## Call:  
## lm(formula = log(acrePrice) ~ improvements + year + log(acres) +   
## log(tillable) + crpPct + productivity, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.98676 -0.16654 0.02461 0.19444 1.97712   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.376e+02 3.239e+00 -73.37 <2e-16 \*\*\*  
## improvements 1.469e-02 4.653e-04 31.58 <2e-16 \*\*\*  
## year 1.216e-01 1.616e-03 75.25 <2e-16 \*\*\*  
## log(acres) -1.082e-01 5.556e-03 -19.48 <2e-16 \*\*\*  
## log(tillable) 2.155e-01 1.564e-02 13.78 <2e-16 \*\*\*  
## crpPct -4.732e-03 2.941e-04 -16.09 <2e-16 \*\*\*  
## productivity 1.630e-02 3.118e-04 52.27 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.365 on 8763 degrees of freedom  
## (9930 observations deleted due to missingness)  
## Multiple R-squared: 0.5605, Adjusted R-squared: 0.5602   
## F-statistic: 1863 on 6 and 8763 DF, p-value: < 2.2e-16

By comparing the residuals of modela and modelaLog, we have successfully reduced the residuals, but lets explore a little more.

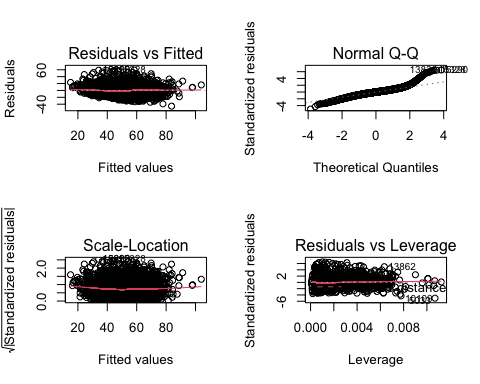
par(mfrow=c(2,2))  
plot(modelaLog)

 Still not satisfied with the shape of the residuals vs. fitted, I will try a square root transformation.

modelaSQRT<-lm(sqrt(acrePrice)~improvements+year+sqrt(acres)+sqrt(tillable)+crpPct+productivity, data=MinnLand)  
summary(modelaSQRT)

##   
## Call:  
## lm(formula = sqrt(acrePrice) ~ improvements + year + sqrt(acres) +   
## sqrt(tillable) + crpPct + productivity, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.671 -4.920 -0.143 4.434 59.566   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.252e+03 8.124e+01 -76.95 <2e-16 \*\*\*  
## improvements 4.291e-01 1.176e-02 36.49 <2e-16 \*\*\*  
## year 3.126e+00 4.052e-02 77.15 <2e-16 \*\*\*  
## sqrt(acres) -5.409e-01 2.818e-02 -19.20 <2e-16 \*\*\*  
## sqrt(tillable) 1.278e+00 1.087e-01 11.76 <2e-16 \*\*\*  
## crpPct -1.036e-01 7.373e-03 -14.05 <2e-16 \*\*\*  
## productivity 3.897e-01 7.871e-03 49.51 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.153 on 8763 degrees of freedom  
## (9930 observations deleted due to missingness)  
## Multiple R-squared: 0.5609, Adjusted R-squared: 0.5606   
## F-statistic: 1865 on 6 and 8763 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(modelaSQRT)

 Understanding that square root transformation can create issues with interpretation, as discussed in problem 2, as well as that the shape of the line on the right side of the residuals vs. fitted plot, I have decided to interpret the data using the log model.

summary(modelaLog)

##   
## Call:  
## lm(formula = log(acrePrice) ~ improvements + year + log(acres) +   
## log(tillable) + crpPct + productivity, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.98676 -0.16654 0.02461 0.19444 1.97712   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.376e+02 3.239e+00 -73.37 <2e-16 \*\*\*  
## improvements 1.469e-02 4.653e-04 31.58 <2e-16 \*\*\*  
## year 1.216e-01 1.616e-03 75.25 <2e-16 \*\*\*  
## log(acres) -1.082e-01 5.556e-03 -19.48 <2e-16 \*\*\*  
## log(tillable) 2.155e-01 1.564e-02 13.78 <2e-16 \*\*\*  
## crpPct -4.732e-03 2.941e-04 -16.09 <2e-16 \*\*\*  
## productivity 1.630e-02 3.118e-04 52.27 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.365 on 8763 degrees of freedom  
## (9930 observations deleted due to missingness)  
## Multiple R-squared: 0.5605, Adjusted R-squared: 0.5602   
## F-statistic: 1863 on 6 and 8763 DF, p-value: < 2.2e-16

given the summary of the coefficients of modelaLog, we can observe that:

A one unit increase in improvements yields a 1.4% increase in price per acre. A one unit increase in year yields a 1.21% increase in price per acre (I feel that this may be incorrect as I am unsure how to convert year into a factor). A one percent increase in acres yields a 1.08% decrease in price per acre. A one percent increase in tillable land yields a 2.155% increase in price per acre. A one unit increase in acres enrolled in the Conservation Reserve Program yields a 4.73% decrease in price per acre. A one unit increase in average agronomic productivity yields a 1.63% increase in price per acre.