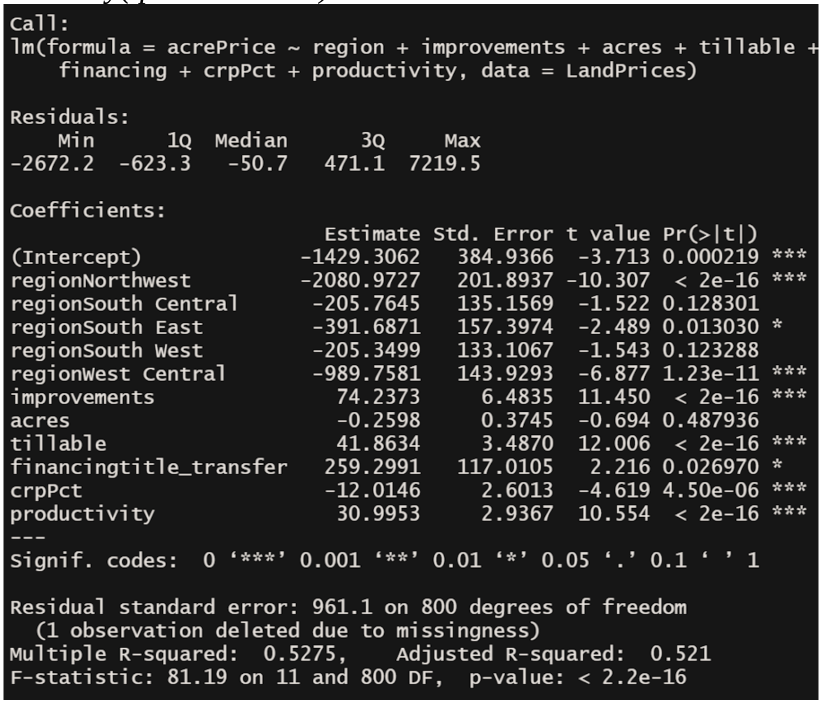
Step A) Initial Model (containing all of the original predictor variables)

#initial model

lp.initialmodel <- lm(acrePrice~region+improvements+acres+tillable+financing+crpPct +productivity,data=LandPrices)

summary(lp.initialmodel)

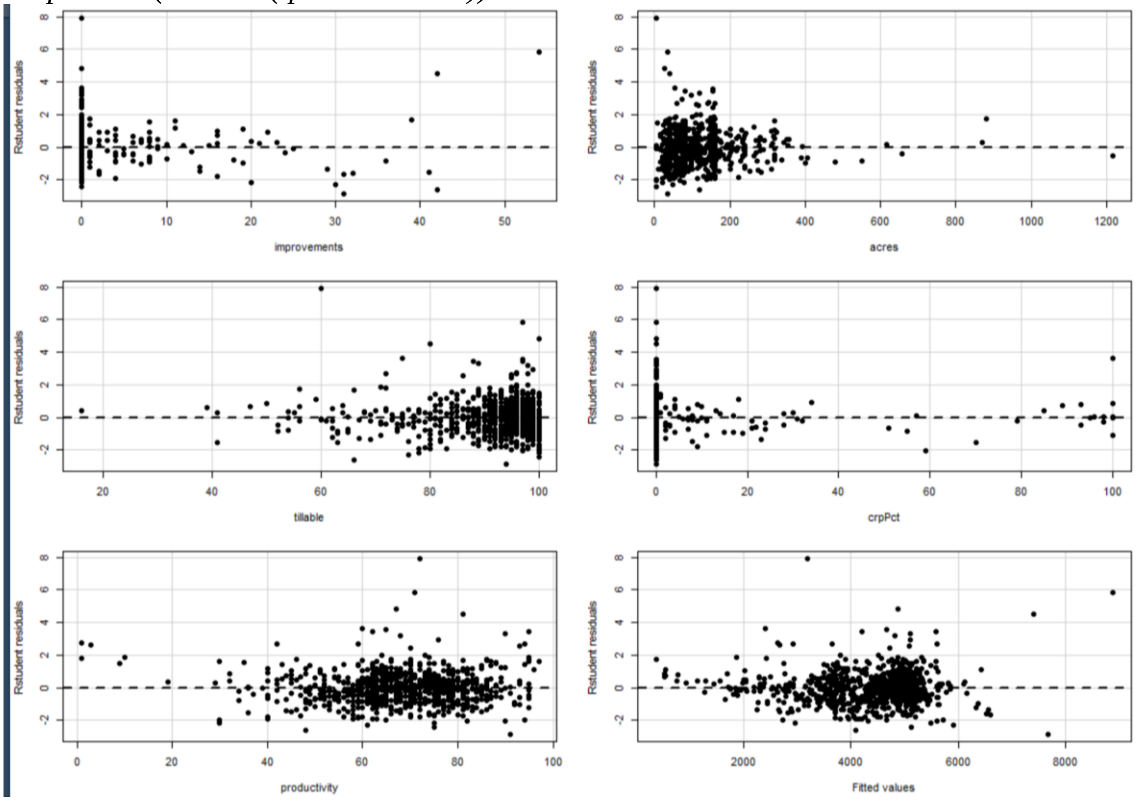


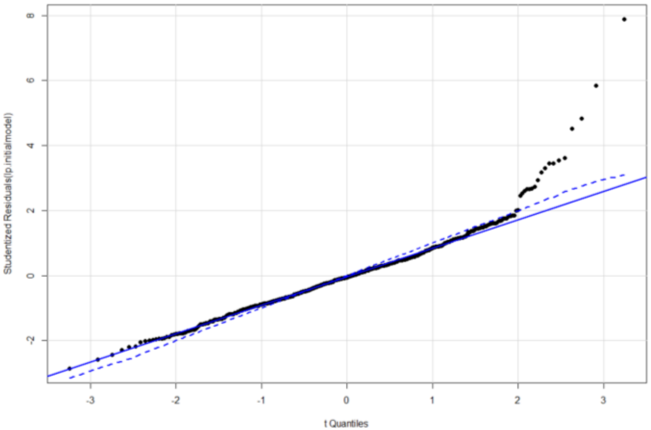
```
Call:
lm(formula = acrePrice ~ region + improvements + acres + tillable +
    financing + crpPct + productivity, data = LandPrices)

Residuals:
    Min      1Q  Median      3Q     Max
-2672.2  -623.3   -50.7   471.1  7219.5

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1429.3062   384.9366  -3.713 0.000219 ***
regionNorthwest         -2080.9727   201.8937 -10.307  < 2e-16 ***
regionSouth Central      -205.7645   135.1569  -1.522 0.128301
regionSouth East         -391.6871   157.3974  -2.489 0.013030 *
regionSouth West         -205.3499   133.1067  -1.543 0.123288
regionWest Central       -989.7581   143.9293  -6.877 1.23e-11 ***
improvements               74.2373     6.4835  11.450  < 2e-16 ***
acres                      -0.2598     0.3745  -0.694 0.487936
tillable                   41.8634     3.4870  12.006  < 2e-16 ***
financingtitle_transfer   259.2991   117.0105   2.216 0.026970 *
crpPct                    -12.0146     2.6013  -4.619 4.50e-06 ***
productivity               30.9953     2.9367  10.554  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 961.1 on 800 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5275,    Adjusted R-squared:  0.521
F-statistic: 81.19 on 11 and 800 DF,  p-value: < 2.2e-16
```

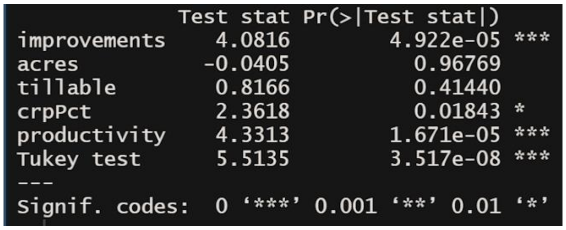
We know that this is the correct initial model because it has all of our original predictor variables (except for CRP, which was excluded because crpPct is part of the model, and we are told to only use one of these two). The output above is what we get from running the initial model.

Step B) Check assumptions, multicollinearity, and influential observations for Initial Model

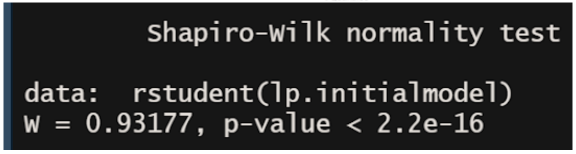
```
#check assumptions
residualPlots(lp.initialmodel,type="rstudent",pch=16,quadratic=FALSE,id=FALSE)
qqPlot(lp.initialmodel,pch=16,envelope=FALSE,id=FALSE)
# Shapiro-Wilk test for non-normality
shapiro.test(rstudent(lp.initialmodel))
```







```
            Test stat Pr(>|Test stat|)
improvements    4.0816       4.922e-05 ***
acres          -0.0405         0.96769
tillable        0.8166         0.41440
crpPct          2.3618         0.01843 *
productivity    4.3313       1.671e-05 ***
Tukey test      5.5135       3.517e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
```



```
        Shapiro-Wilk normality test

data:  rstudent(lp.initialmodel)
W = 0.93177, p-value < 2.2e-16
```

Linearity - check residuals vs. predicted and residuals vs. each x plot - OK if no curves in any plots

> This regression assumption looks to be okay. None of the plots above have a distinct curve pattern in them.

Independence - non-series: check study design - OK if data were randomly sampled

> We are not told in the study design whether or not our data were randomly sampled. We are however told that it comes from 816 farm sales in Minnesota in 2010. We do not know for sure whether this data was randomly sampled, however, if it is not, there is no way for us to fix this as we work towards our final model.

Equal Variance - check residuals vs. predicted and residuals vs. each x plot - OK if no fan shape in any plots

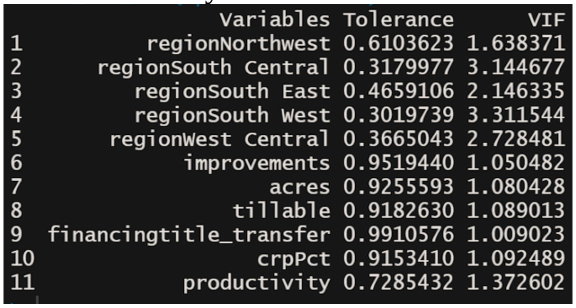
> There is a very distinct fan shape in the residuals vs. acres plot, as lower acreage has a greater variance, and the variance decreases drastically as acreage increases. Residuals vs. tillable might have a fan shape, but there is a single observation that deters us from drawing this conclusion. Because of the first fan shape mentioned, this regression assumption is violated.

Normality - QQ plot of residuals, Shapiro-Wilk test - OK if points fall near the diagonal

> The QQ plot of residuals above has a distinct curve away from the diagonal towards the end. We can also see from the output above that the p-value for the Shapiro-Wilk test is <2.2e-16. Because this is very small, we conclude that the regression errors are not normal. So this regression assumption is violated. There is evidence of a violation.

The normality violation is not as much of a concern because of the large sample size (n = 816).

Multicollinearity

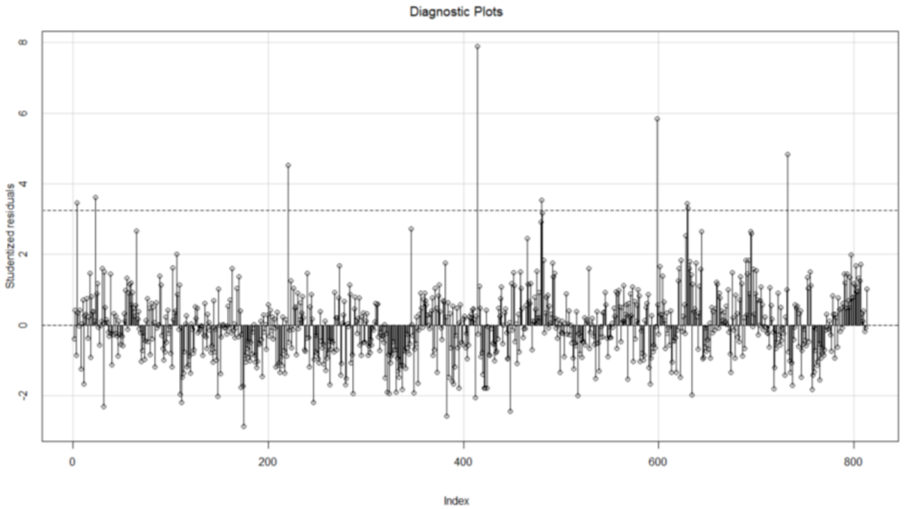


```
                  Variables Tolerance      VIF
1           regionNorthwest 0.6103623 1.638371
2       regionSouth Central 0.3179977 3.144677
3          regionSouth East 0.4659106 2.146335
4          regionSouth West 0.3019739 3.311544
5        regionWest Central 0.3665043 2.728481
6              improvements 0.9519440 1.050482
7                     acres 0.9255593 1.080428
8                  tillable 0.9182630 1.089013
9  financingtitle_transfer 0.9910576 1.009023
10                   crpPct 0.9153410 1.092489
11             productivity 0.7285432 1.372602
```

We can see from the output above that there is no indication of multicollinearity issues in our initial model. All of the VIFs above are below 4, which means there is not even an indication of moderate multicollinearity, much less of severe multicollinearity.

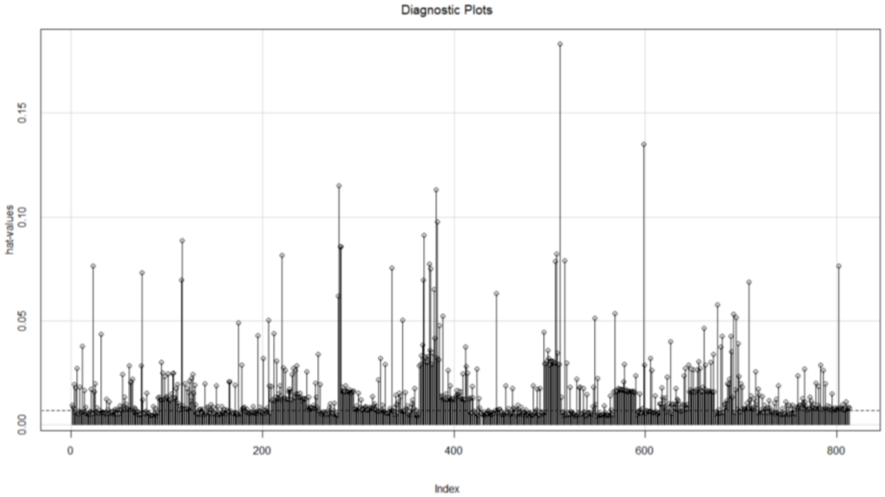
Influential Observations

```
#influential observations
#outliers
influenceIndexPlot(lp.initialmodel,vars="Studentized",id=FALSE)
abline(h=3,lty=2)
```



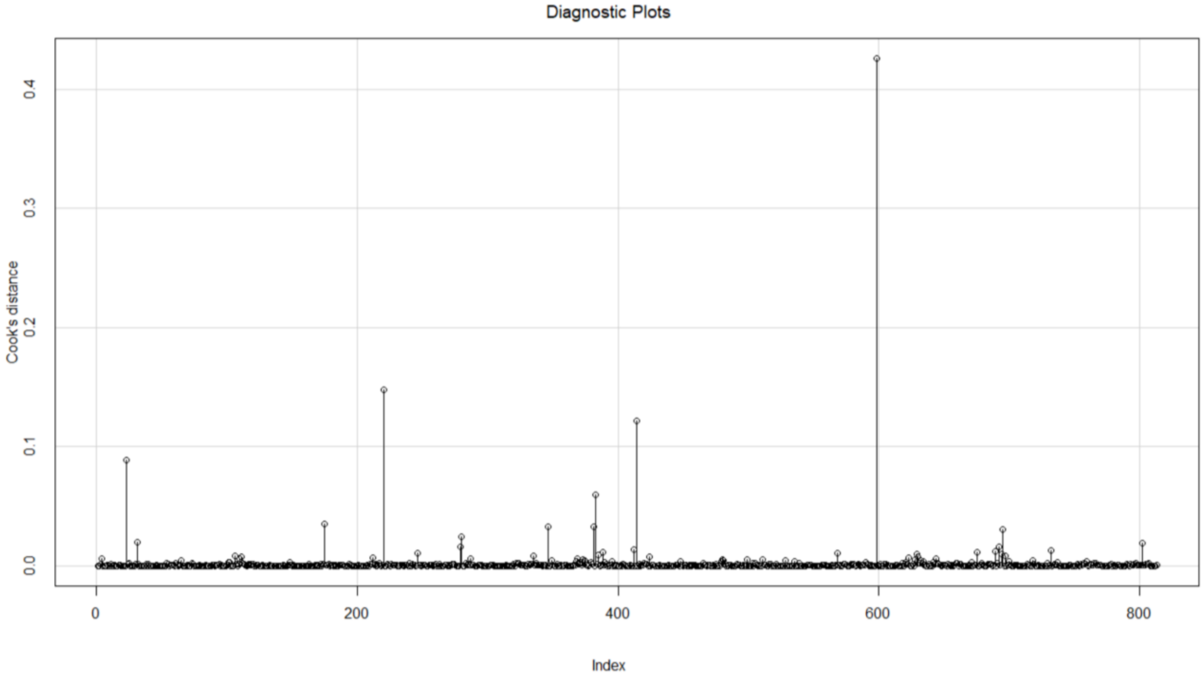
We can see from the plot of the studentized residuals above that there are a few data points that are outliers, as they are greater than 3 (the dashed line). There are none less than -3. We can see that there are 9 data points with a studentized residual greater than 3.

```
#leverage
influenceIndexPlot(lp.initialmodel,vars="hat",id=FALSE)
abline(h=0.0295203,lty=2)
```



For leverage, we need to compute our cutoff value for high leverage. Let k be the number of predictors in the regression equation, and n be the number of observations. In this case, we have k=13 and n=813. High leverage is considered anything greater than 3(k+1)/n = 3(13+1)/813 = 3(14)/813 = 42/813 = 0.05166052. We can see from the plot output above that there are a very large number of data points that have hat-values greater than our computed high-leverage cutoff value. There are at least 25 data points with high leverage.

#influential observations
influence.cutoff <- qf(.5,8,805)
influence.cutoff
influenceIndexPlot(lp.initialmodel,vars="Cook",id=FALSE)
abline(h=influence.cutoff,lty=2)

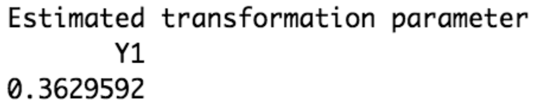


We computed from the code above that the cutoff for influence in this data set for Cook’s distance is 0.918782. We have this because our numerator degrees of freedom is 8 (k+1 = 7+1) and our denominator degrees of freedom is 805 (n-k-1 = 813-7-1 = 813-8 = 805). So any Cook’s distance higher than 0.92 for this regression would be highly influential. There are no influential observations, as can be seen from the plot above.

Step C) Changes to your data or model to fix any problems in the initial model

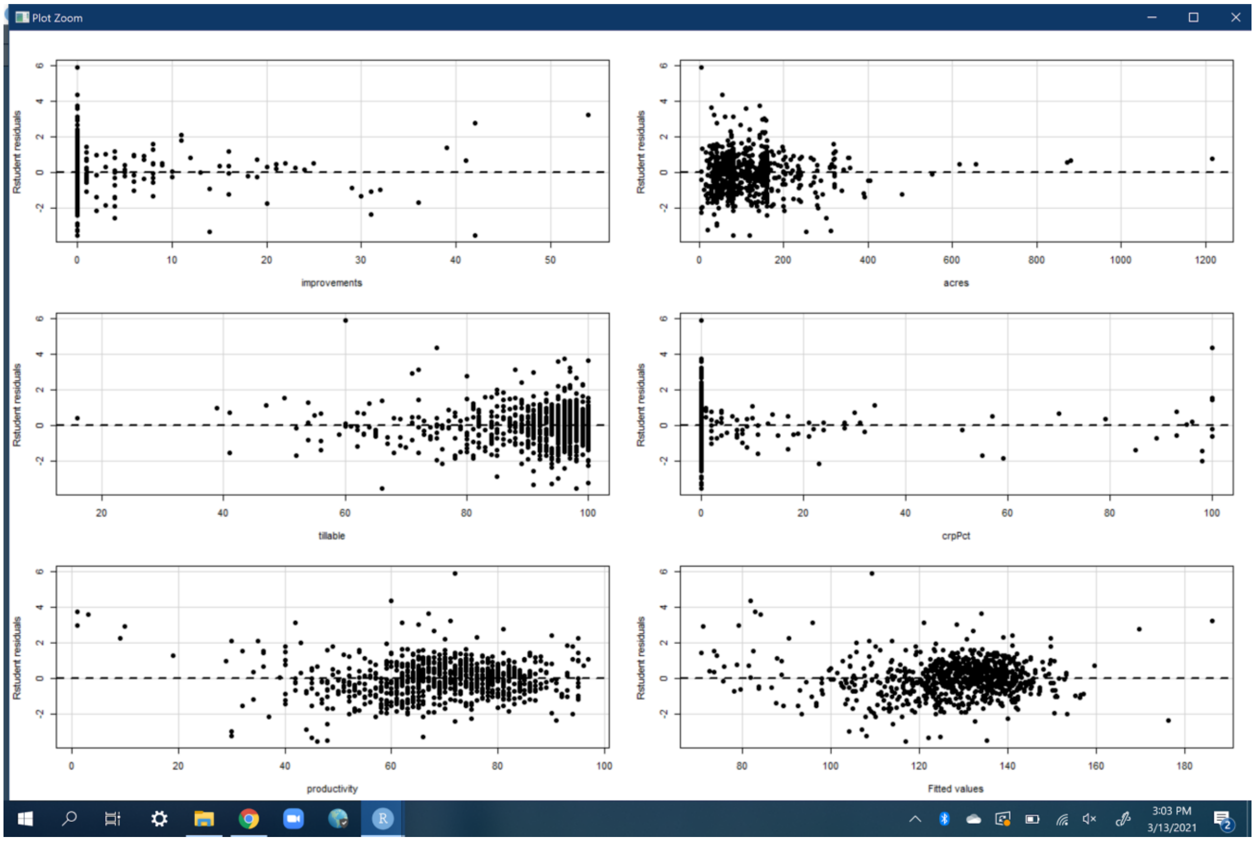
We saw from our checks above that we have problems with the equal variance assumption, Normality assumption, and we have points with very high leverage. The normality violation is not as much of a concern because of the large sample size (n = 816). Because we need to fix variance and normality, we will need to transform our y variable. To fix our high leverage issues, we will need to transform our x variable.

Because we concluded that there for sure is a distinct fan shape in the acres vs. residuals plot, and the fan is larger on the left and narrower on the right, we would want to transform y by moving up the ladder of powers. We also see that our QQ plot of the residuals bulges up and to the left from the diagonal, which is more evidence that we should transform y by moving up the ladder of powers (or transform x down).

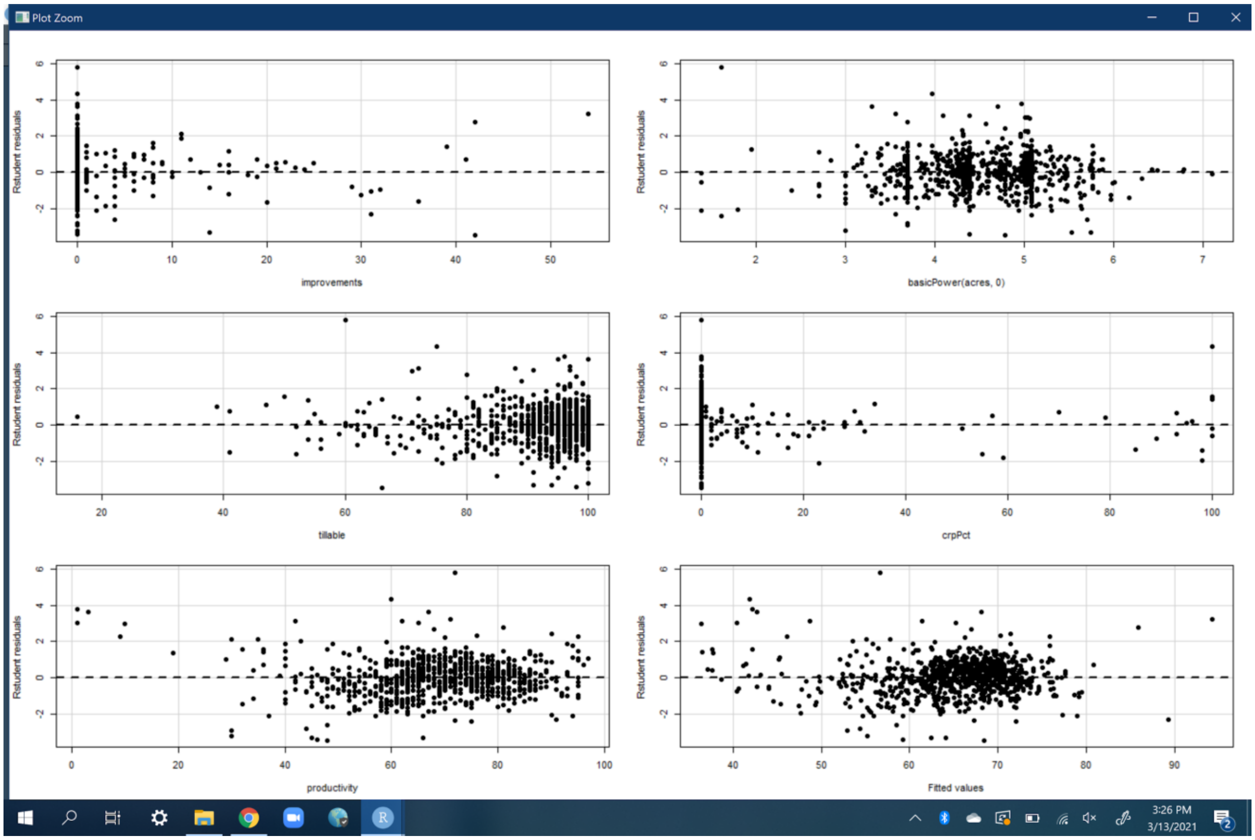


```
Estimated transformation parameter
       Y1
0.3629592
```

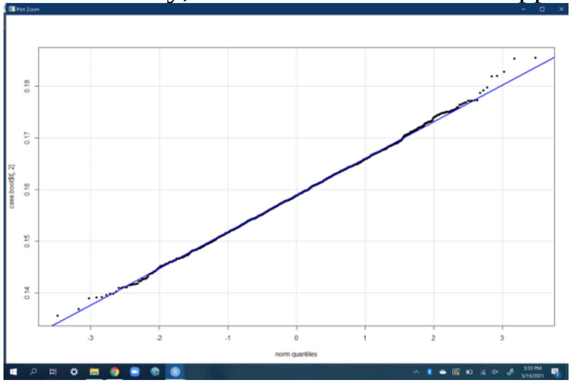
We had R suggest a starting transformation, because we know that our y-variable needs to be transformed. When we run this, we see that the suggested power is 0.3629529. We know from our notes that we should round all powers to the nearest ½. To the nearest ½, 0.3629529 becomes 0.5, or a square root transformation. (Also note here that we tried both powers of 0 and 1 but neither worked well, so we stuck with 0.5). Below are the residual plots of our newly-transformed model:



We can still see a fan shape in acres (there is also still potential for a fan shape in tillable, but the single point at 60 with a high residual value prevents us from concluding this). We can see from the residual plots output below that transforming acres to a power of 0 (log(acres)) immensely fixes our equal variance problems.



We will also look at Bootstrapping as an alternative transformation of our data. Because our initial model has issues with equal variance, we will not consider residual resampling.
We know that we can use bootstrapping because we did not find any violations with independence or linearity regression assumptions, but we did have violations with our equal variance and normality regression assumptions. Because we had issues with both equal variance and normality, we will use case bootstrapping as opposed to residual bootstrapping.



The normality violation is not as much of a concern because of the large sample size (n = 816). Due to the non-normality, only prediction intervals would be adversely affected. However, the unequal variance of a predictor variable will affect the width of our confidence intervals. Instead of a transformation, we could use bootstrapping to address these problems. The distribution of the sample slopes is close to normal. Because of the large sample size, the slope estimates still behave like a normal distribution. However, because of the unequal variance, the spread of the slope estimates is greater than we would expect.

Thus, lp.model4, where acrePrice is the square root, and acres is log is the best way for us to fix these regression assumptions.

D. Variable reduction (or "How I got from my initial model to my final model")
This step may include variable selection methods, removing variables, adding interactions, rechecking assumptions, etc., iteratively until you find a good final model.

We now need to add in interactions and eliminate insignificant terms to get to our good final model from model4.

Best model from both all subsets and best subsets:

cImp cTill cPct cPro Reg Fin

CODE:

```
library(tidyverse)
library(colorRamps)
library(car)
library(effects)
library(olsrr)

#PART A
#initial model
lp.initialmodel <-
lm(acrePrice~region+improvements+acres+tillable+financing+crpPct+productivity,data=LandPrices)
summary(lp.initialmodel)

#PART B
#check assumptions
residualPlots(lp.initialmodel,type="rstudent",pch=16,quadratic=FALSE,id=FALSE)
qqPlot(lp.initialmodel,pch=16,envelope=FALSE,id=FALSE)
# Shapiro-Wilk test for non-normality
shapiro.test(rstudent(lp.initialmodel))
#multicollinearity
ols_vif_tol(lp.initialmodel)
#influential observations
#outliers
influenceIndexPlot(lp.initialmodel,vars="Studentized",id=FALSE)
#leverage
influenceIndexPlot(lp.initialmodel,vars="hat",id=FALSE)
#influence
influence.cutoff <- qf(.5,8,805)
influence.cutoff
influenceIndexPlot(lp.initialmodel,vars="Cook",id=FALSE)

#PART C
#starting transformation (y needs to be transformed)
(power1 <- powerTransform(lp.initialmodel,family="bcPower")) #0.3629592
#suggests sqrt transformation of acrePrice
lp.model1 <-
lm(basicPower(acrePrice,0.5)~improvements+acres+tillable+crpPct+productivity+region+financing,data=LandPrices)
summary(lp.model1)
residualPlots(lp.model1,type="rstudent",quadratic=FALSE,tests=FALSE,id=FALSE,pch=16)
qqPlot(lp.model1,pch=16,envelope=FALSE,id=FALSE)
shapiro.test(rstudent(lp.model1))
# tranforming x as well - acres - logarithm
```

```
lp.model2 <-
lm(basicPower(acrePrice,0.5)~improvements+basicPower(acres,0)+tillable+crpPct+productivity
+region+financing,data=LandPrices)
summary(lp.model2)
residualPlots(lp.model2,type="rstudent",quadratic=FALSE,tests=FALSE,id=FALSE,pch=16)
qqPlot(lp.model2,pch=16,envelope=FALSE,id=FALSE)
shapiro.test(rstudent(lp.model2))

# PART D
#first try at variable reduction
#model2 from above but with all interactions possible
lp.model3 <-
lm(sqrt(acrePrice)~(improvements+log(acres)+tillable+crpPct+productivity+region+financing)^
2,data=LandPrices)
summary(lp.model3)
#model with only significant interactions based on p-values from model3 output
lp.model4 <-
lm(sqrt(acrePrice)~(improvements+log(acres)+tillable+crpPct+productivity+region+financing+i
mprovements*region+log(acres)*region+tillable*region+productivity*region+log(acres)*tillable
), data=LandPrices)
summary(lp.model4)
#model with variable reduction from model4
lp.model5 <-
lm(sqrt(acrePrice)~improvements+log2(acres)+tillable+productivity+region+improvements*regi
on+log2(acres)*region+tillable*region+productivity*region+log2(acres)*tillable,
data=LandPrices)
summary(lp.model5)
ols_vif_tol(lp.model5)
#we see very high multicollinearity, so we need to figure out how to reduce this
#checking significant interactions to compare and choose simpler model
#we use a stepwise function to try another way to find only significant interactions
step1 <- step(lp.model3,scope=~.^2,direction="both",trace=1,k=log(n))
summary(step1)
ols_vif_tol(step1)
#very high multicollinearity! something is going wrong here
#now we will try centering all of our quantitative variables to reduce multicollinearity
#center quantitative variables to reduce multicollinearity
LandPrices$cTill <- scale(LandPrices$tillable,scale=FALSE)
LandPrices$cPct <- scale(LandPrices$crpPct,scale=FALSE)
LandPrices$cProd <- scale(LandPrices$productivity,scale=FALSE)
LandPrices$cImp <- scale(LandPrices$improvements,scale=FALSE)
LandPrices$logAcre <- log(LandPrices$acres)
LandPrices$cAcre <- scale(LandPrices$logAcre,scale=FALSE)
#if we center acre first, then we will have negative values
#we cannot do a log transformation on negative values
#so we must do the log transformation first and then center the variable
```

```
#model with centered variables
lp.model6 <-
lm(sqrt(acrePrice)~cImp+cAcre+cTill+cPct+cProd+region+financing,data=LandPrices)
summary(lp.model6)
#we see from the output above that cAcre is the only term that is not significant
#so we will eliminate it from our model and continue
#we will now run a model with all possible interactions
lp.model7 <-
lm(sqrt(acrePrice)~(cImp+cTill+cPct+cProd+region+financing)^2,data=LandPrices)
summary(lp.model7)
#from this output, will will eliminate insignificant predictor variables
#we see that cProd and all of its interactions are insignificant
#we also see that financing and all of its interactions are insignificant
#so we will eliminate both of these variables from our model
lp.model8 <-
lm(sqrt(acrePrice)~(cImp+cTill+cPct+region+region*cImp+region*cTill+region*cPct),data=LandPrices)
summary(lp.model8)
#now we see that all terms and their interactions are significant
#so this is the model that we will want to stick with
#we do not want to eliminate any more variables or interactions
#define this to be our final model
lp.finalmodel <- lp.model8

# PART E
#check assumptions
residualPlots(lp.finalmodel,type="rstudent",pch=16,quadratic=FALSE,id=FALSE)
qqPlot(lp.finalmodel,pch=16,envelope=FALSE,id=FALSE)
# Shapiro-Wilk test for non-normality
shapiro.test(rstudent(lp.finalmodel))
#multicollinearity
ols_vif_tol(lp.finalmodel)
#influential observations
#outliers
influenceIndexPlot(lp.finalmodel,vars="Studentized",id=FALSE)
#leverage
influenceIndexPlot(lp.finalmodel,vars="hat",id=FALSE)
#influence
influence.cutoff <- qf(.5,8,805)
influence.cutoff
influenceIndexPlot(lp.finalmodel,vars="Cook",id=FALSE)
#verifying that our final model is the best possible
all.sub8 <- ols_step_all_possible(lp.model8)
best.sub8 <- ols_step_best_subset(lp.model8)
view(all.sub8)
view(best.sub8)
```

```
#comparing reduced and full model with partial F-test
anova(lp.finalmodel, lp.initialmodel)
```