Daniel Witt

**Farm Appraisal Analysis**

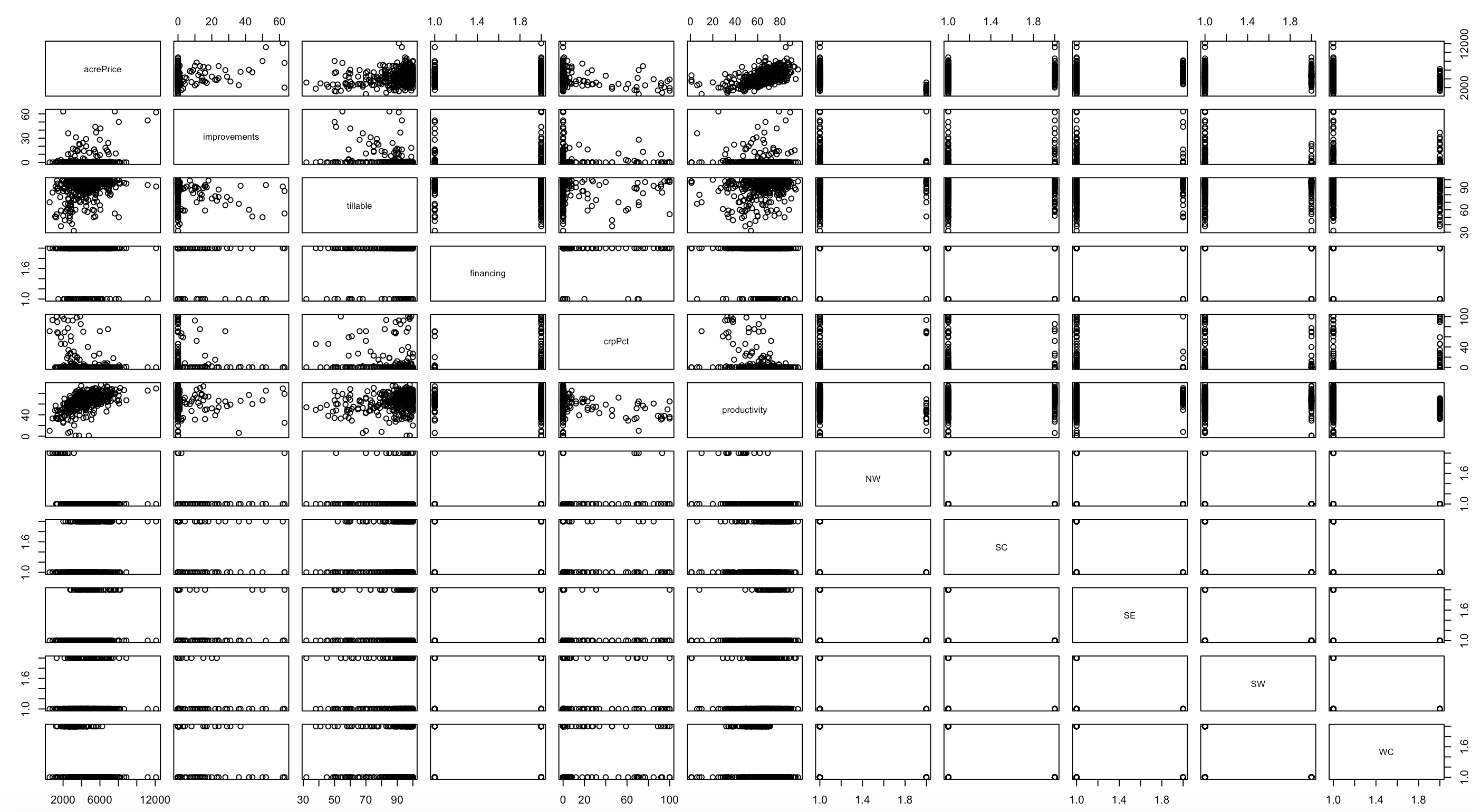
**Section 1: Introduction and Problem Background**

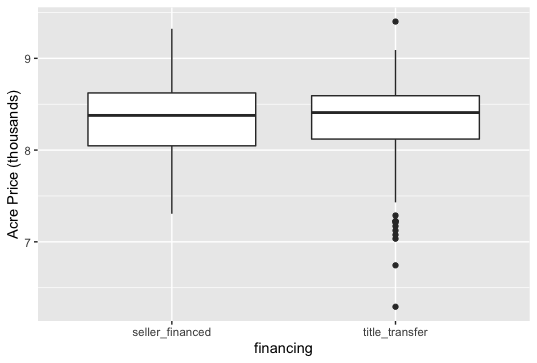
A farm appraiser is someone who determines the value of a farm based on different factors that the farm possesses. In order to correctly value a farm, it is important to determine which factors are the most valuable when determining the price of a farm, as well as correctly predicting an appropriate selling price for the farm. The goal of this analysis is to make an inference as to which factors are most important when valuing a farm, and to be able to use those factors to predict the price that the farm should be sold at.

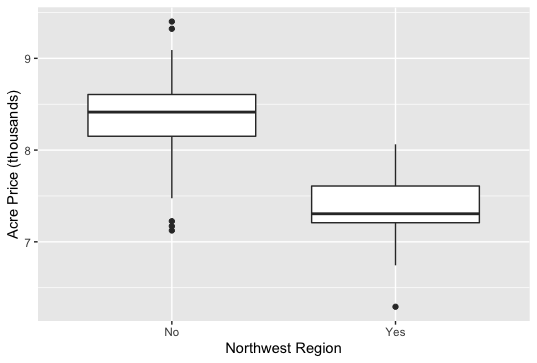
In order to do this, a dataset was collected that contains a list of 420 different farms and the following variables associated with them:

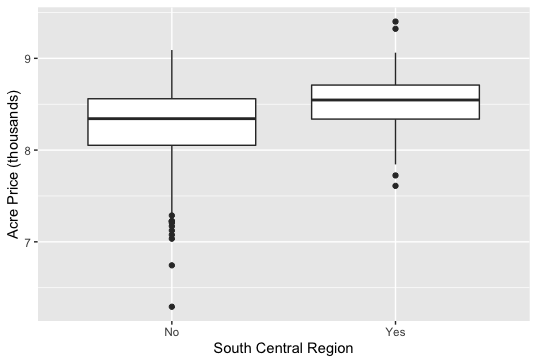
* Acre Price – Sale price per acre of land
* Improvements – Percentage of property value due to buildings
* Tillable – Percentage of farm rated arable
* Financing – Type of financing
* CrpPct – Percentage of land that is enrolled in conservation reserve
* Productivity – Numeric score between 1 and 100 where 100 indicates productive land
* NW – Farm is located in Northwest region
* SC – Farm is located in South Central region
* SE – Farm is located in Southeast region
* SW – Farm is located in Southwest region
* WC – Farm is located in West Central region

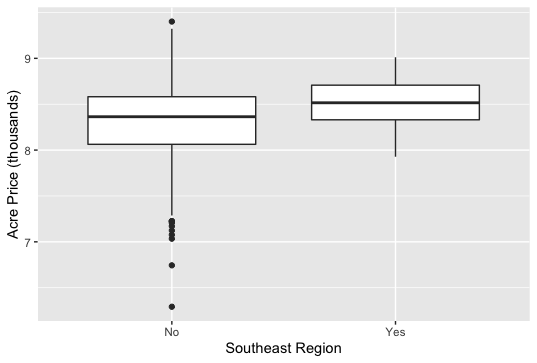
Using this data, a Multiple Linear Regression model can be developed in order to determine which of these factors is most important in valuing a farm, and then use the model to predict the appropriate acre price for the farm. First we will graph the data using scatter plots and box plots:

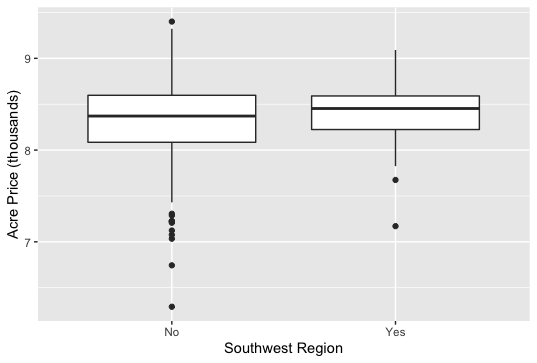


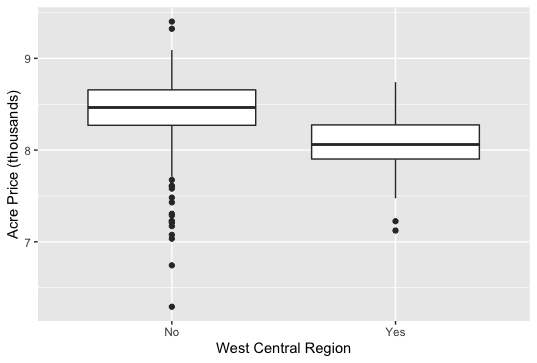






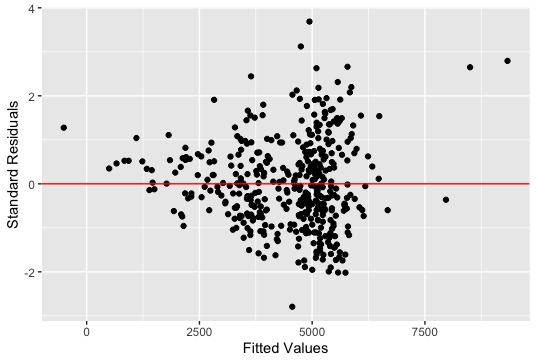






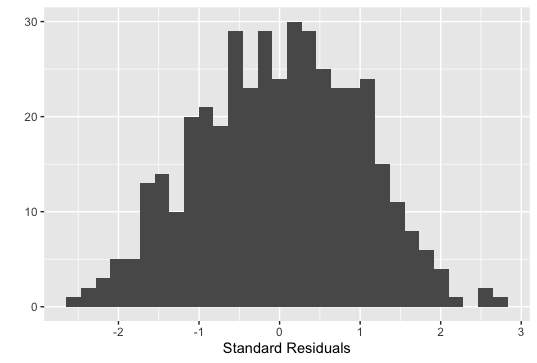
When looking at the scatterplot matrix, we want to focus on the top row, which graphs the acre price on the y-axis and each variable along the x-axis. We can see that the points for all of the quantitative variables, like the tillable variable have some interesting patterns. There seems to be a linear pattern for the most part with a positive relationship for the quantitative variables, meaning as acre price increases, the variable increases as well. However, there seem to be many outliers in the data. The boxplots of each of the categorical variables show that there are many outliers in the data as well, for example, in the southwest region there are many outliers on the lower end of acre price for those who are not in the southwest region.

We want to determine if we can use this data to perform our Multiple Linear Regression (MLR) model, so we need to check the assumptions of our model, which are Linearity, Independence, Normality, and Equal Variance. We can quickly check equal variance by looking at the graph of a fitted MLR model with its standard residuals to see if the variance about the regression line is equal:

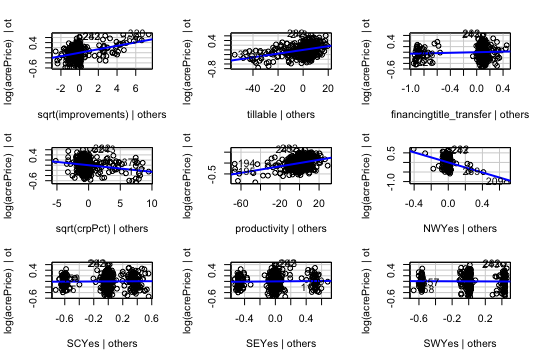


We can see that the variance about the line increases as the fitted values increases, so our equal variance assumption does not seem to be met. In order to confirm this, we performed a BP test, where the null hypothesis states the variance is equal, and the alternative states that the variance is not equal, with 0.05 significance. We receive a p-value of essentially zero, which confirms that our equal variance assumption is not met. This means that we cannot use the untransformed data for our MLR model. We will need to transform the data in order to proceed with the model to obtain our analysis goals.

An appropriate transformation of the data to meet our MLR model assumptions would be to take the natural log of acre price (our response variable), as well as take the square root of our improvements and crpPct variables. When we do this, it will transform the data in a way that will meet our model assumptions. We conduct a BP test again to see if it has fixed our equal variance assumption and we receive a p-value of 0.07357. This means that we are in favor of our null hypothesis, which means the variance is equal for our data. We also check normality using a histogram of the standard residuals:



We can see that the data appears to be normal. We also check normality using a KS test, where the null hypothesis states that the data is normal, and the alternative stating that it is not normal. We receive a p-value of 0.6752, which means the data is normal, so this assumption is met. Looking at the Added-Variable plots of our data can assess our assumption of linearity:



The data seems to be relatively linear in shape, so we can assume that our linearity assumption is met for our model. We don’t see any strange patterns in our data, so we can also assume our independence assumption is met. With all of our assumptions met, we can proceed to developing our MLR model for the farm appraisal.

**Section 2: Statistical Modeling**

It is important to check our model for collinearity, which occurs when two or more variables are highly correlated with one another. When this happens, our standard errors become inflated, significant variables are harder to detect, and our predictions of farm price will be thrown off. We can check for collinearity by looking at the variance inflation factors for each variable:

|  |  |  |
| --- | --- | --- |
| Sqrt(improvements) | Tillable | Financing |
| 1.14 | 1.16 | 1.05 |
| Sqrt(crpPct) | Productivity | NW |
| 1.17 | 1.48 | 12.11 |
| SC | SE | SW |
| 2.09 | 1.70 | 1.9 |
| WC | Productivity:NW |  |
| 2.12 | 11.32 |  |

Any VIF that is above 10 indicates collinearity issues, so we can see that the NW variable has collinearity issues, along with our interaction variable between NW and Productivity, which we are including because the farm appraiser believes there is a significant difference in productivity in the NW region. In order to fix this, we can use variable selection to determine the most important variables to include in our model.

In order to determine which variables are most important to include in our MLR model for the natural log transform of acre price, we will use the AIC variable selection technique. The reason we will use this specific technique and not another variable selection technique, like BIC, is because it maximizes our adjusted R-squared value. Also, since we are mainly concerned with predicting the acre price of the farm, we want to use AIC because this is the technique that is generally better for prediction. We will use the AIC technique and best subset selection, meaning that we will run through all possible combinations of variables and select the best model that maximizes our R-squared value. When we do this, these are the variables that we will include in our model:

* Sqrt(Improvements)
* Tillable
* Financing – Title Transfer
* Sqrt(CrpPct)
* Productivity
* NW – Yes
* WC – Yes
* Productivity-NW interaction

The Multiple Linear Regression model for our data will look like this:

*Yi = β0 + β1(Sqrt(improvements)+ β2(tillable) + β3I(financing = title transfer) + β4(Sqrt(crpPct)) + β5(productivity) + β6I(NW = Yes) + β7I(WC = Yes) + β8I(NW = Yes)I(productivity) + εi*

*εi ~ Ν(0, σ2)*

Where:

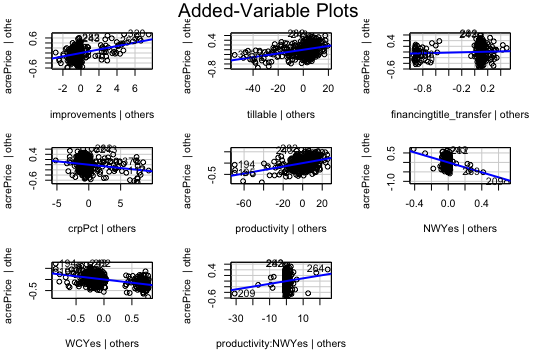
* Yi = The natural log of the predicted acre price for a farm
* i =1, …, n
* β0 = Intercept coefficient = The natural log of the acre price of the farm if the square root of improvements, tillable, the square root of crpPct, and productivity were all equal to 0, and the farm was not financed by a title transfer and was not in the NW or WC regions, on average
* β1 = Slope coefficient = Holding all other variables constant, if the square root of improvements increases by 1, the natural log of acre price will increase by β1, on average
* β2 = Slope coefficient = Holding all other measurements constant, if the tillable variable were to increase by 1, the natural log of acre price would increase by *β2,* on average
* β3 = Slope coefficient = Holding all else constant, if the farm is financed by a title transfer, compared to not a title transfer, the natural log of acre price will increase by β3, on average
* β4 = Slope coefficient = Holding all other variables constant, if the square root of crpPct increases by 1, the natural log of acre price will increase by β4, on average
* β5 = Slope coefficient = Holding all other variables constant, if the productivity of the farm increases by 1, the natural log of acre price will increase by β5, on average
* β6 = Slope coefficient = Holding all other variables constant, if the farm is located in the NW region, compared to all other regions, the natural log of acre price will increase by β6, on average
* β7 = Slope coefficient = Holding all other variables constant, if the farm is located in the WC region, compared with all other regions, the natural log of acre price will increase by β7, on average
* β8 = Slope coefficient = Holding all else constant, If the farm is located in the NW region, the interaction between NW region and productivity will increase the natural log of acre price by β8 on average
* εi = Residuals or distance to the mean
* σ = for any measurement variable, 99.7% of the acre price measurements will be within 3 σ of Yi

When using this model, we will use the assumptions of linearity, independence, normality, and equal variance, which we will justify later in the analysis.

**Section 3: Model Justification and Verification**

The assumptions we used for our model were linearity, independence, normality, and equal variance. We can justify these assumptions as follows:

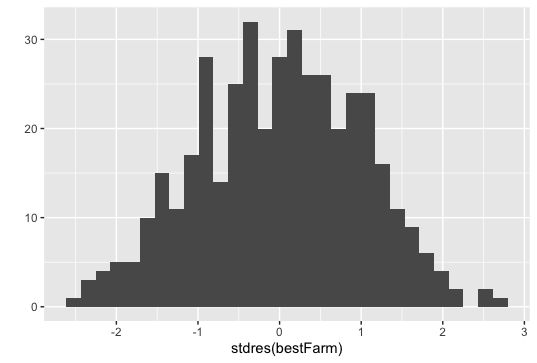
Linearity – We can look at the Added-variable plots of the data to verify this assumption



We can see that the data follows a relatively linear pattern, so we can assume that this assumption is met

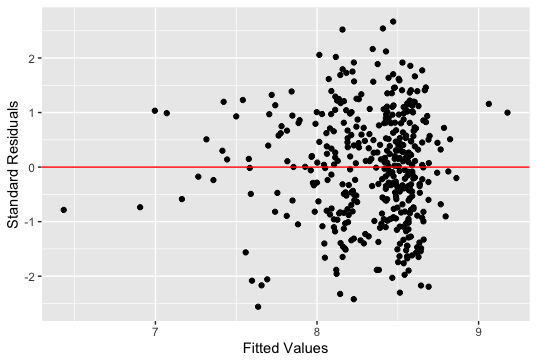
Independence – We can assume this assumption is met because the price of one farm is not determined by the price of another farm. There does not seem to be any unusual patterns in the data that would cause us to believe that the data was not independent

Normality – We can graph the standard residuals of the fitted model in a histogram to see if the data is normal or not:



The data looks to be normal, so we can assume that normality is met. We can verify this as well by conducting a KS test like we did earlier in the analysis for the original data. When we do this on our new model, we receive a p-value of 0.7103, which means that our data is normal, so this assumption is met.

Equal variance – We can look at a scatterplot of the standard residuals versus the fitted values of the model to determine equal variance:



The variance seems to be relatively equal, but to make sure this assumption is met, we can conduct a BP test to determine if the model has equal variance. When we conduct the BP test on the model, we receive a p-value of 0.1508, which confirms that our data is equal in variance. All of our assumptions are met for our model.

We can look at the adjusted R-squared value to make a decision of how well the model fits our data. Our R-squared value from the model is 0.667, which means that the variables improvements, tillable, financing, crpPct, productivity, NW, and WC can explain 66.7% of our acre price. Although we wish that this were higher, we know this is the highest percentage we can obtain from the variables we have been given because we used best variable selection, which maximizes the R-squared value. There are other variables that were not recorded that could help us obtain a higher R-squared value, but for the data given us, this is the best fitting model.

We can assess our model’s ability to predict the price of a farm by performing cross validation, where we remove a portion of the data, then fit the model, and then predict the acre price of a farm on the data we left out of our model. We performed 100 cross validations and received these results:

Bias = 0.0019 – on average we are predicting the natural log of acre price 0.0019 too high

RPMSE = 0.234 – on average our predictions of the natural log of acre price are off by 0.234

Coverage = 0.948 – 94.8% of our data is included in our prediction interval

Prediction Width = 0.925 – the width of our prediction interval for the natural log of acre price is 0.925

When considering the range of our log(acrePrice), which is 6.292402 to 9.400938, and the standard deviation of the log(acrePrice), which is 0.403071, we can see these predictions are fairly accurate. Our model does a good job of predicting the acre price of a farm since the bias and error of our model is small, meaning it is accurately predicting the farm acre price.

**Section 4: Results**

Below is a table of each of the selected variables included in our model and their effects on log(acrePrice), with a 95% confidence interval included with each variable estimate to account for uncertainty:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient Estimate | Coefficient Lower | Coefficient Upper |
| Intercept | 7.1815752 | 6.977940610 | 7.385209751 |
| Sqrt(Improvements) | 0.0691207 | 0.052399876 | 0.085841446 |
| Tillable | 0.0078603 | 0.005944167 | 0.009776511 |
| Financingtitle\_transfer | 0.0547498 | -0.016989364 | 0.126488943 |
| Sqrt(crpPct) | -0.0257294 | -0.036784275 | -0.014674460 |
| Productivity | 0.0076747 | 0.005950929 | 0.009398419 |
| NWYes | -1.3118417 | -1.702290381 | -0.921393000 |
| WCYes | -0.2928252 | -0.348219933 | -0.237430537 |
| Productivity:NWYes | -0.2928252 | 0.001394901 | 0.018193496 |

The quantitative variables in our model can be interpreted in this way:

Holding all other variables constant, we are 95% confident that as the percentage of tillable land increases by 1, the log(acrePrice) will increase by between 0.005944167 and 0.009776511, on average.

The categorical variables in our model can be interpreted in this way:

Holding all other variables constant, we are 95% confident that if the financing for the farm is through a title transfer, as opposed to seller financed, the log(acrePrice) will increase by between -0.016989364 and 0.126488943, on average.

In order to better interpret these variables, we can untransform the model by taking the exponential of each estimate, so that we understand the effect of each variable on acre price:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient Estimate | Coefficient Lower | Coefficient Upper |
| Intercept | 1314.9779552 | 1072.7069684 | 1611.9658710 |
| Sqrt(Improvements) | 1.0715655 | 1.0537970 | 1.0896335 |
| Tillable | 1.0078913 | 1.0059619 | 1.0098245 |
| Financingtitle\_transfer | 1.0562763 | 0.9831541 | 1.1348369 |
| Sqrt(crpPct) | 0.9745988 | 0.9638840 | 0.9854327 |
| Productivity | 1.0077042 | 1.0059687 | 1.0094427 |
| NWYes | 0.2693236 | 0.1822656 | 0.3979643 |
| WCYes | 0.7461525 | 0.7059436 | 0.7886517 |
| Productivity:NWYes | 1.0098423 | 1.0013959 | 1.0183600 |

The quantitative variables in our model can be interpreted in this way:

Holding all other variables constant, we are 95% confident that as the percentage of tillable land increases by 1, the acre price will increase by between 1.0059619 and 1.0098245, on average.

The categorical variables in our model can be interpreted in this way:

Holding all other variables constant, we are 95% confident that if the financing for the farm is through a title transfer, as opposed to seller financed, the acre price will increase by between 0.9831541 and 1.1348369, on average.

Our model confirms the appraiser’s point of view about the effect of productivity in the NW region being different than in other areas. When we observed the effects of each variable in our model, we receive a p-value of 0.0224 for our interaction between NW and productivity. This is a significant result, which means that the interaction has a significant effect on acre price.

We can use our model to predict the price of a farm that consists of these qualities:

|  |  |
| --- | --- |
| Financing | Title transfer |
| Region | NW |
| Improvements | 0 |
| Tillable | 94 |
| crpPct | 0 |
| Productivity | 96 |

With these qualities and using our model, we predict that the price of the farm will be 4189.627. We must also account for uncertainty, so we provide a measure of uncertainty in our prediction with a 95% prediction interval. We are 95% confident that a title transfer farm in the NW region with 0 improvements, 0 crpPct, 94 tillable, and 96 productivity will be priced between 2206.569 and 7954.871, on average.

**Section 5: Conclusions**

After conducting our analysis, we have developed a Multiple Linear Regression model that helps us to understand which variables significantly affect the acre price of a farm, as well as predict the acre price of a farm. We learned that the most significant factors that contribute to acre price are improvements, tillable, crpPct, financing with a title transfer, productivity, NW region, WC region, and the interaction between productivity and NW region. These variables explain 66.7% of the acre price, which does not explain as much of our acre price as we would like, but it is our best model, as determined by variable selection using best subset selection and AIC technique. We also determined that there was a significant effect of the interaction between productivity and the NW region, as the appraiser suspected.

In order to better understand and predict acre price, the appraiser should determine more variables to collect data on in order to explain more of the acre price. Our model only explained 66.7% of the acre price, and this was our best model, so we will need new variables to help us better explain acre price.

R-code:

#read in data

farm <- read.table(file = 'https://mheaton.byu.edu/docs/files/Stat330/Exams/Midterm2/Data/Farms3.txt', header = TRUE)

head(farm)

tail(farm)

#plot data

library(ggplot2)

library(GGally)

pairs(farm)

ggplot(data=farm, mapping=aes(y=acrePrice, x=financing)) + geom\_boxplot() + ylab("Acre Price (thousands)")

ggplot(data=farm, mapping=aes(y=acrePrice, x=NW)) + geom\_boxplot() + ylab("Acre Price (thousands)") + xlab("Northwest Region")

ggplot(data=farm, mapping=aes(y=acrePrice, x=SC)) + geom\_boxplot() + ylab("Acre Price (thousands)") + xlab("South Central Region")

ggplot(data=farm, mapping=aes(y=acrePrice, x=SE)) + geom\_boxplot() + ylab("Acre Price (thousands)") + xlab("Southeast Region")

ggplot(data=farm, mapping=aes(y=acrePrice, x=SW)) + geom\_boxplot() + ylab("Acre Price (thousands)") + xlab("Southwest Region")

ggplot(data=farm, mapping=aes(y=acrePrice, x=WC)) + geom\_boxplot() + ylab("Acre Price (thousands)") + xlab("West Central Region")

#fit model and check assumptions

library(MASS)

library(car)

farmModel <- lm(log(acrePrice)~sqrt(improvements) + tillable+ financing + sqrt(crpPct) + productivity+NW+SC+SE+SW+WC+productivity:NW, data=farm)

summary(farmModel)

coef(farmModel)

qplot(farmModel$fitted.values, stdres(farmModel), geom = "point") +geom\_hline(yintercept=0,color="red") + ylab("Standard Residuals") + xlab("Fitted Values")

avPlots(farmModel)

qplot(stdres(farmModel), geom = "histogram") + xlab("Standard Residuals")

ks.test(stdres(farmModel), "pnorm")

library(lmtest)

bptest(farmModel)

#create interaction variable

prodNW <- farm$productivity:farm$NW

#determine colinearity

library(car)

vif(farmModel)

#choose and fit model

farm$acrePrice <- log(farm$acrePrice)

farm$improvements <- sqrt(farm$improvements)

farm$crpPct <- sqrt(farm$crpPct)

library(bestglm)

farm <- farm[,c(2:ncol(farm),1)]

farm

vs.resBIC <- bestglm(farm, IC="BIC", method = "exhaustive")

summary(vs.resBIC$BestModel)

vs.resAIC <- bestglm(farm, IC="AIC", method="exhaustive")

summary(vs.resAIC$BestModel)

#add interaction to best model

bestFarm <- lm(acrePrice~improvements+tillable+financing+crpPct+productivity+NW+WC+NW:productivity, data = farm)

coef(bestFarm)

summary(bestFarm)

#cross validation

n.cv <- 100

bias <- rep(NA, n.cv)

rpmse <- rep(NA, n.cv)

coverage <- rep(NA, n.cv)

pred.width <- rep(NA, n.cv)

n.test <- 10

for(i in 1:n.cv){

# Choose which obs. to put in test set

test.obs <- sample(1:nrow(farm), n.test)

# Split data into test and training sets

test.set <- farm[test.obs,]

train.set <- farm[-test.obs,]

# Using training data to fit a model

train.lm <- lm(acrePrice~improvements+tillable+financing+crpPct+productivity+NW+WC+NW:productivity,data=train.set)

# Predict test set

test.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

# calculate coverage, bias, rpmse and prediction width

coverage[i] <- mean((test.preds[,2] < test.set$acrePrice) & (test.preds[,3]>test.set$acrePrice))

bias[i] <- mean(test.preds[,1]-test.set$acrePrice)

rpmse[i] <- sqrt(mean((test.preds[,1]- test.set$acrePrice)^2))

pred.width[i] <- mean(test.preds[,3] - test.preds[,2])

}

mean(bias)

mean(rpmse)

mean(coverage)

mean(pred.width)

min(farm$acrePrice)

exp(max(farm$acrePrice))

sd(farm$acrePrice)

#prediction

prediction <- data.frame(improvements=c(0), tillable=c(94), crpPct=c(0), productivity=c(96), financing=c("title\_transfer"), NW=c("Yes"), WC=c("No"))

predict.lm(bestFarm, newdata = prediction)

estimatedFarm <- predict.lm(bestFarm, newdata = prediction)

estimatedFarm <- exp(estimatedFarm)

estimatedFarm

pred.Int <- predict(bestFarm, prediction, interval="predict")

exp(pred.Int)

#justify model assumptions

qplot(bestFarm$fitted.values, stdres(bestFarm), geom = "point") +geom\_hline(yintercept=0,color="red") + ylab("Standard Residuals") + xlab("Fitted Values")

avPlots(bestFarm)

qplot(stdres(bestFarm), geom = "histogram")

ks.test(stdres(bestFarm), "pnorm")

library(lmtest)

bptest(bestFarm)

#hypothesis testing

summary(bestFarm)

exp(bestFarm$coefficients)

#confidence interval

exp(confint(bestFarm))