Farm Appraisal Analysis

Jonathan Wilson

November 22, 2019

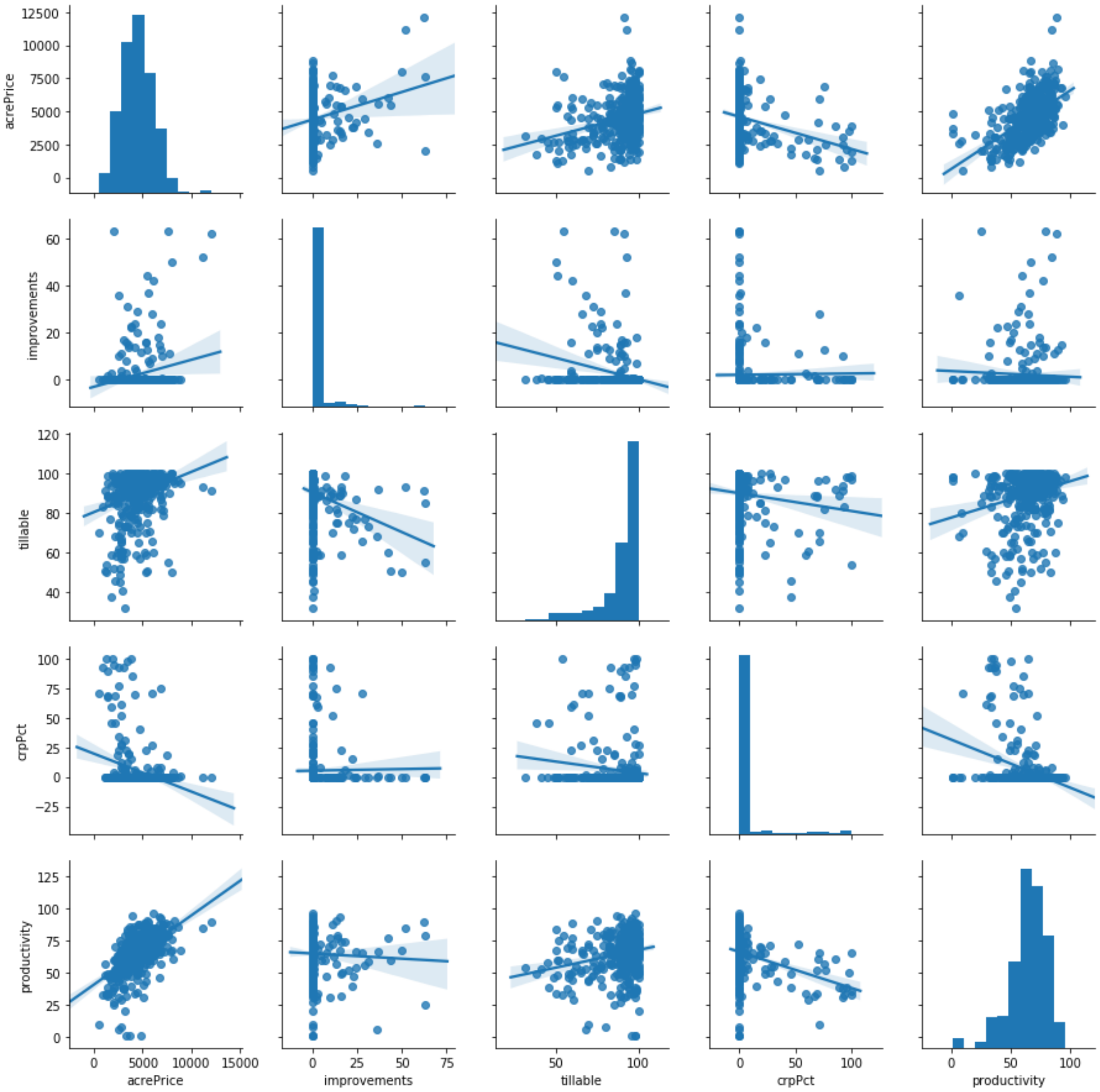
Introduction and Problem Background

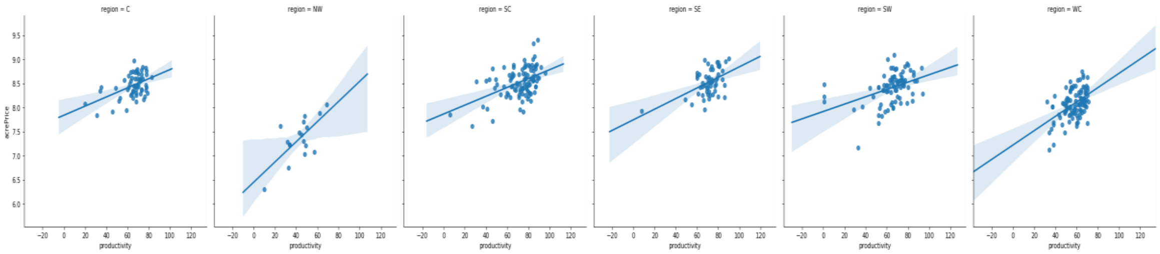
Farm appraisers are interested in understanding what gives a farm its value. Knowing what factors increase or decrease the value of a farm will be valuable to an appraiser in determining a fair price for both the seller and buyer. Given the right factors we would be able to make predictions for the price of a farm. Being able to do this would not only save time and money but also potential mistakes that a novice or even an experienced appraiser might make when determining a price. There is an appraiser who believes that the effect of on price in the NW is different than in other areas.

The data we have on hand is from Farm.txt which includes variables such as productivity of the land and the region this farm is in. The response or predictive value would be the acrePrice. With these data we would like to create a model that would allow us to make these predictions we previously discussed. Before we do that, we should explore the data and see which model and techniques are necessary to get the right predictions.

The scatter plot matrix below shows individual plots for each variable. The data for acrePrice had several outliers so I went ahead and transformed the data so that we make sure we follow the assumptions we are making for multiple linear regression. A noticeable correlation is between productivity and acrePrice where there is a somewhat strong positive correlation between them. There may also be a few variables that have collinearity which I can address shortly. The reason we concern ourselves with collinearity is that it can inflate our standard errors which could throw off our estimates as they would be highly sensitive to changes in our observations. Calculating variance inflation factors helps to see which variables are collinear or if there is even any collinearity at all.

After looking at the variance inflation factors and correlations we can see that there does not seem to be any collinearity.





After making the proper transformation I went ahead and checked for any kind of interactions. The plots above show the different regions side by side with acrePrice and productivity as the axes. Upon plotting several different plots, I found that regions had distinct clusters across each group and the NW region in particular. For the most part the clusters appear to be linear. Most of the regions have roughly the same slope or they have roughly the same increase in acrePrice for each unit increase of productivity. However, the NW region seems to increase more sharply (having a different slope than the rest) which would suggest an interaction between productivity and region.

We looked at the data, explored trends within the data, made a transformation on our response to conform to the assumptions we are making when using a multiple linear regression model, and checked for collinearity as well as any kind of interactions. Now we are ready to start building a statistical model.

Statistical Modeling

In the case of farm appraisals there are several factors that may or may not have an impact on the price of land. Our main goal is to use the variables that are important in our model. We can use a variable selection technique that will take out certain variables in such a way we do not reduce our model too much while also optimizing our model. For this model I used best subset selection procedure which considers all possible combinations of variables. Since we had a reasonable number of variables (not computationally intensive) this method seemed ideal. I also based my variable selection procedure on the AIC. I chose this procedure as it seems to optimize for making predictions. Prediction is what we would like to achieve in this analysis. See the code in the appendix for the process I went through to build this model. Below is the model I found using the variable selection technique.

I also compared the reduced and full models to verify that adding an interaction term was important. After running an F-test for an interaction and got a p-value below a .05 and thus concluded that adding an interaction term was statistically significant.

Simple Multiple Linear Regression Model: $log(y_i) = \beta_0+\beta_1(improvements)+\beta_2(tillable)+ \beta_3(crpPct)+ \beta_4(productivity)+ \beta_5 I(region = NC)+ \beta_6 I(region = SC)+ \beta_7 I(region = SE)+ \beta_8 I(region = SW)+\beta_9 I(region = NW)+\beta_{10}(productivity)x(I(region = NW))$

with interpretations:

ϵ_i *distributed as (iid)* $N(0, \sigma^2)$ **-** What this means is that our model is dependent on following the assumptions:

linear, independent, normal, and equal variance which we will test shortly. In or order to use the simple multiple linear regression model we must follow these assumptions.

ϵ_i **-** The error associated with an observation's distance between a dot and the line. $log(y_i)$ **-** the response variable for the ith data point interpreted as the log of the acrePrice.

β_0 **-** The intercept. On average this is what we would expect the acrePrice to be if the farm was located in the Central region (as the Central Region was our base factor) and all other variables were zero.

(improvements) - This is a data point for improvements used to predict the acrePrice. Other variables like this can be interpreted this way as well.

β_1 **-** The slope for the 1st explanatory variable improvements. Holding all else constant if improvements increases by 1 unit then acrePrice would increase by β_1 on average. Other slope coefficients can be interpreted in the same manner.

Interactions - As productivity goes up by 1, then acrePrice goes up by β_4 for the central region, $\beta_4 + \beta_5$ for the NC region, $\beta_4 + \beta_6$ for the SC region, $\beta_4 + \beta_7$ for the SE region, $\beta_4 + \beta_8$ for the SW region, and $\beta_4 + \beta_{10}$ for the NW region. σ^2 **-** The variance about the line or how far our distances are from the average. Or how much acrePrice varies from the average given the explanatory variables.

After fitting our model to the data we should be able to predict the an acrePrice for a farm given our variables and region group.

Model Verification and Justification

Now we will test the fit of our model to determine how well is fits our data as well as check the assumptions we made previously about using a regression model.

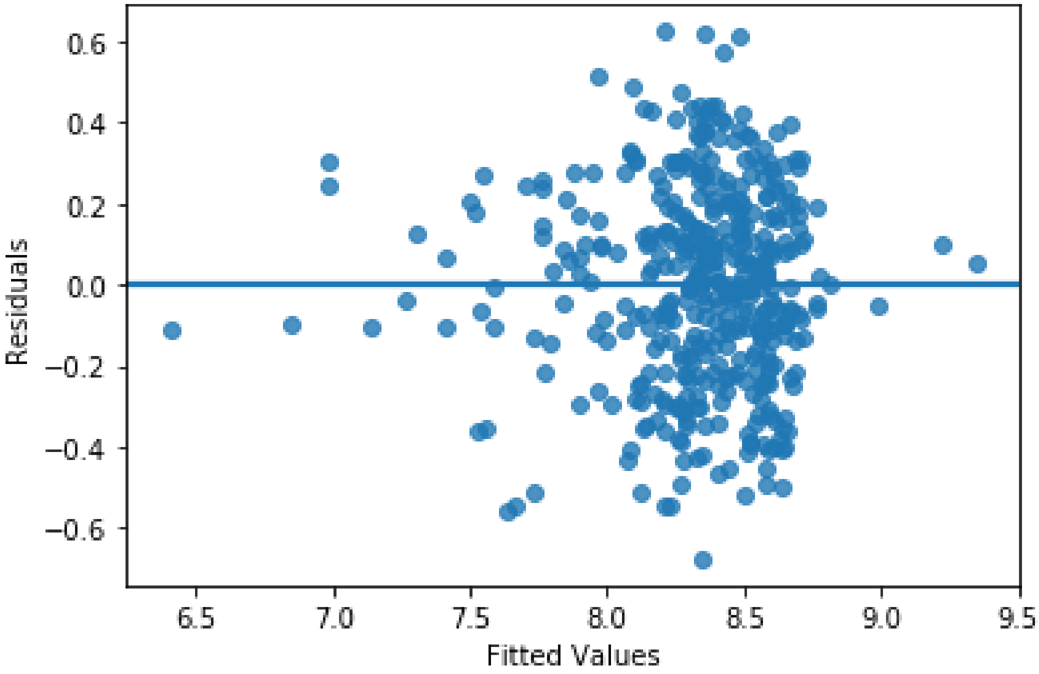
Assumptions

As noted earlier in order to use this model we must follow four assumptions. These four assumptions are listed below.

Equal Variance and independence

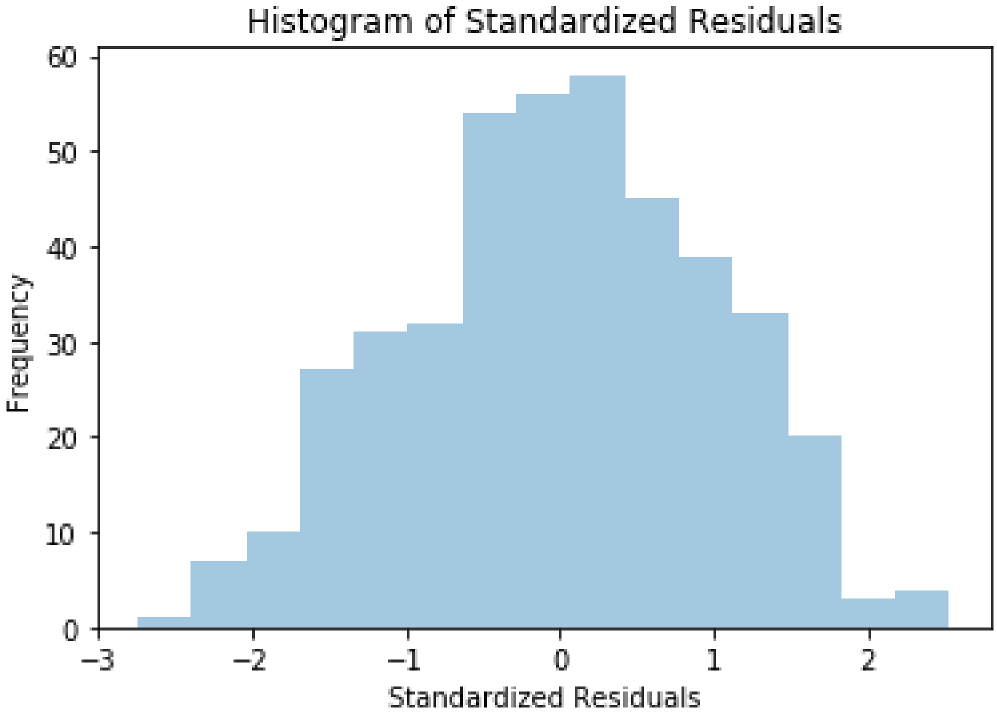
The Scatterplot of Standardized Residuals below shows that the residuals are mostly evenly spread out. Just to be sure there is equal variance I ran a Breusch-Pagan test. The test returned p-values that were all high enough to conclude that these data are have equal variance and do not violate homoscedasticity, so this assumption was met.

As far as independence knowing about one data point may in fact have an affect another. Perhaps knowledge of someone being more productive in the same region might influence another's productivity score. I am going to go ahead and assume independence however.



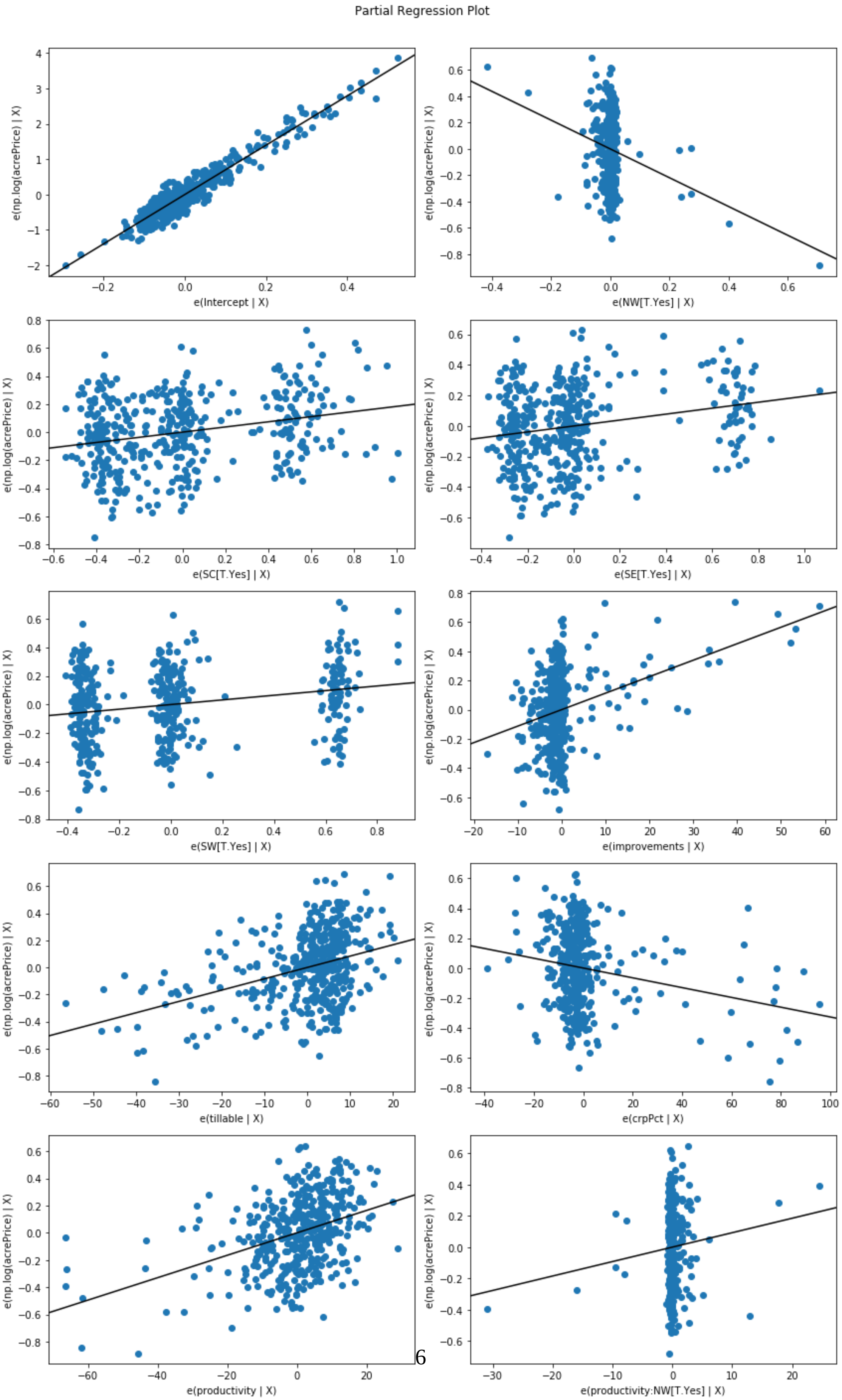
Normality

The histogram below appears normal but just to be sure I ran a Kolmogorov-Smirnov test for normality of residuals which says we cannot reject the null hypothesis that the data do come from a normal so we can conclude that the data are normal given a p-value of 0.8783.



Linearity

For linearity we will use a partial regression plot which looks at each variable individually. We could also use the scatter plots we plotted earlier in the paper as a guide as well. The plots below show that the data are linear on the individual basis that is isolating each of the variables we see that they follow a linear trend. Hence, we are fine to go ahead with using our model.



Partial Regression Plot

Model Prediction Ability

After running some numbers, we see that our R^2 is **0.6335**. What this says is that about 63% of the variation in acrePrice is explained by all variables, main effects and interaction combined. This number is not too bad. Our bias says we are under-predicting by (**-133.41**) unit price. This number is somewhat small which says our model is slightly biased. RPMSE of about **1105.5026** says that our predictions about acrePrice given all of the explanatory variables combined are off by 1105.5026 unit price. This number is a little large meaning that our predictions will be off when we make predictions about acrePrice. The coverage tells us the percentage of predication intervals that contain the true value we are looking for. For our model about **96%** of the true values we are looking for are are contained with in predication intervals. Lastly, prediction interval width is just the average width of our prediction or the difference between the upper and lower bounds of our prediction intervals. In our model the average width is roughly **4429.76** price units. This width covers roughly mean of the acrePrice or zero standard deviations. The downside to such a narrow width is that our predictions will be off.

Results

We use F-tests to determine if there are real affects on our response variable (acrePrice). Our first F-tests we determine if each covariate and interaction in our model have any real affect on acrePrice.

Below is a table of with the covariate or interaction predictors to the left, each of their 95% confidence intervals, and a p-val after conducting an F-test.

Covariate	[.025]	[.0975]	p-val
Intercept	6.770	7.169	0
NW	-1.507	-0.681	0
SC	0.119	0.249	0
SE	0.116	0.273	0
SW	0.097	0.228	0
improvements	0.008	0.014	0
tillable	0.006	0.010	0
crpPct	-0.005	-0.002	0
productivity	0.006	0.010	0
productivity:NW[T.Yes]	0	0.018	.043

Assuming an α of .05, the p-values in the table are well below α. This would indicate that these predictors have a significant affect on acrePrice. For both the region NW and the interaction of productivity and the region NW we can say that they do have an effect on acrePrice since their p-value is well below .05. We had also checked previously that the interaction was significant enough to include in our model. We can interpret the NW region categorical variable as holding all else constant we are 95% confident that acrePrice will have an increase in price between exp(-1.507) and exp(-0.681) (or .2216 and .5061 untransformed) for farms in the NW region. We can interpret the variable productivity as holding all else constant we are 95% confident that acrePrice will have an increase in price between exp(0.006) and exp(0.010) (or 1.006 and 1.01 untransformed) for farms that increase in productivity by one unit. As for interpreting the interaction term, holding all else constant we are 95% confident that farms in the NW region have between 1 and 1.0182.

The claim that the appraiser makes could hold. I ran predictions for multiple regions holding all else constant and the NW region is worth a little less than the others on average. However, since we have such a wide prediction interval, we get values for NW that overlap other region prices.

Making Predictions

We can make predictions for acrePrice by simply plugging in the parameters (i.e. productivity) into our model and run the calculation. Below is our example in question.

If we make a prediction for the data that was given in question we would expect to see a an acrePrice being exp(8.341167) or 4192.98 on average with an interval of (2122.60, 8282.799). Our prediction is around the mean for acrePrice and our prediction interval is really wide. Having such a wide interval leaves room for a lot of variance in our acrePrices.

Conclusions

Overall our model predicts well where on average we will predict the right acrePrice. The downside to our model is that we have such a wide prediction interval which means we may be getting a wide variety for acrePrice. There does seem to be a different affect with farms in the NW region and productivity also seems to have an influence on the NW region. How much of a difference is hard to tell with this model.

To better understand and predict acre price an appraiser could learn more about the NW region. It may be helpful to know more about what makes it different. The appraiser might be surprised that there is not much difference between regions and productivity.