Nate Jenkins
November 2017

Farm Appraisal Analysis

Introduction and Background:

Farm appraisers are in charge of valuing farms prior to sale. In order to make the best appraisals, it is important to know which factors most affect the prices of farms. This analysis uses data collected on farms and their sale prices in order to help determine which factors are the best indicators of farm price.

The collected data includes specific information about each farm including per acre price, percentage of land that is tillable, type of financing, percentage of land that is enrolled in conservation reserve, productivity score, percentage of property that is due to buildings, and farm location. The goal of this study is to use statistical modeling to quantify and explore the relationships between these variables and uncover which data about a farm are the best indicators of price.

Multiple linear regression is a good candidate for modeling this data because there are several numerical and categorical features that can be used to predict farm price. Before modeling these relationships, however, it is important to explore the data for obvious trends and analyze the viability of multiple linear regression.

First, in order to preform multiple linear regression, it is important that the potential predictor variables are linearly related to the dependent variable, price per acre. Figure 1 below is a scatter plot matrix of the numerical data in the data set. Although productivity score is clearly linearly related to price, the others lack a clear linear relationship.

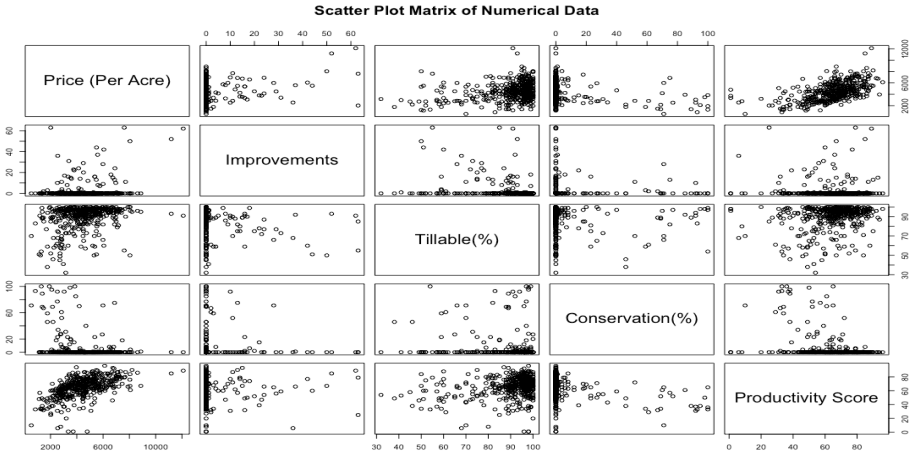


Figure 1

In order to account for this apparent lack of linearity, a number of transformations were attempted on the data in order to increase the linearity. The transformation that appears to have the best results is log transforming the price variable. Figure 2 below shows that this transformation has the effect increasing the linearity between the predictor variables and price

per acre. Additionally, the correlation matrix in figure 3 shows that there is correlation between price and the predictor variables. Some of the variables such as productivity score and conservation (%) are potentially collinear. It will be important to be aware of this relationship when modeling the data. Overall, however, the numerical data appears suitable for preforming multiple linear regression.

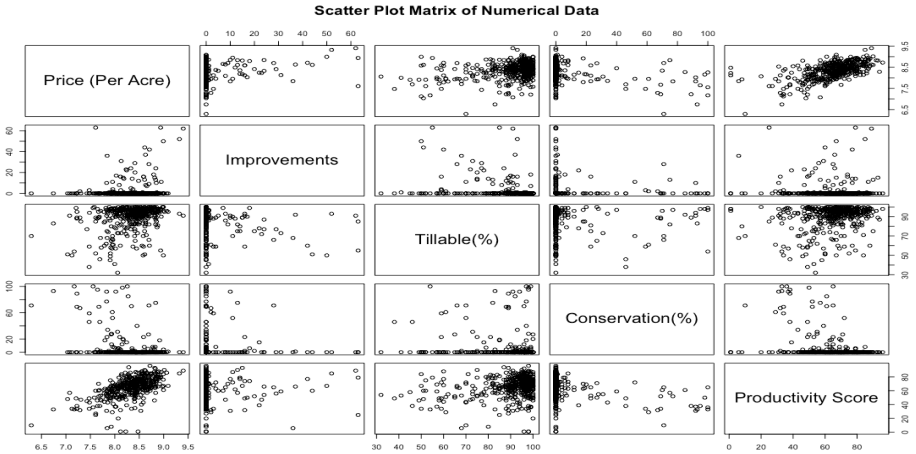


Figure 2

Correlation Matrix

	Price (Per Acre)	*Improvements*	*Tillable (%)*	*Conservation (%)*	*Productivity Score*
Price (Per Acre)	*1.00*	*0.14*	*0.30*	*-0.37*	*0.60*
Improvements	*0.14*	*1.00*	*-0.27*	*0.01*	*-0.04*
Tillable (%)	*0.30*	*-0.27*	*1.00*	*-0.13*	*0.22*
Conservation (%)	*-0.37*	*0.01*	*-0.13*	*1.00*	*-0.34*
Productivity Score	*0.60*	*-0.04*	*0.22*	*-0.34*	*1.00*

Figure 3

It is also important to get a sense of the categorical data used in the analysis. Figure 4 displays each region plotted against price per acre as well as financing type against price per acre. There appears to be a difference in price in some of the areas, especially in the North-West Region (NW). The difference in price for financing type is less apparent.

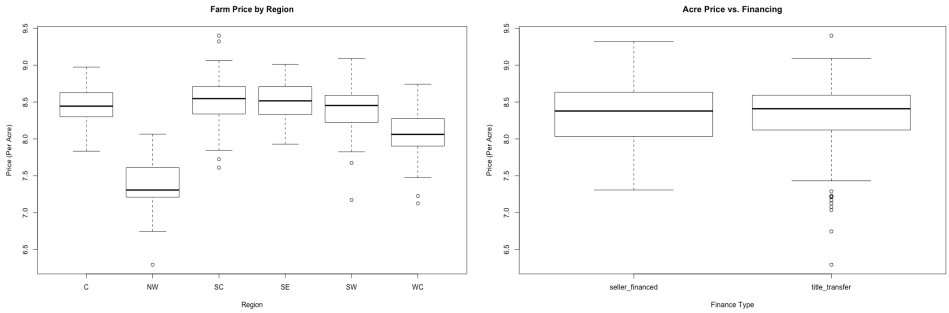


Figure 4

Nate Jenkins
November 2017

Statistical Modeling:

Because a number of variables are present in this data set, best subset model selection is used to choose the best combination of predictor variables for the model. Best subset selection tests all potential combinations of predictor variables and compares them using a specific criterion. In this analysis, the subset selection is preformed using "BIC" as the model criterion. BIC is relatively exclusive criterion that compares the potential models to each other and incorporates weighted penalties for increasing number of predictor variables. Because BIC is relatively exclusive, it is a good fit for this analysis. Farm appraisers need to be able to quickly focus on the important aspects of a farm for appraisal so having fewer variables offers more information on the important variables. The graph in figure 5 shows that the BIC is minimized when 7 predictor variables are present.

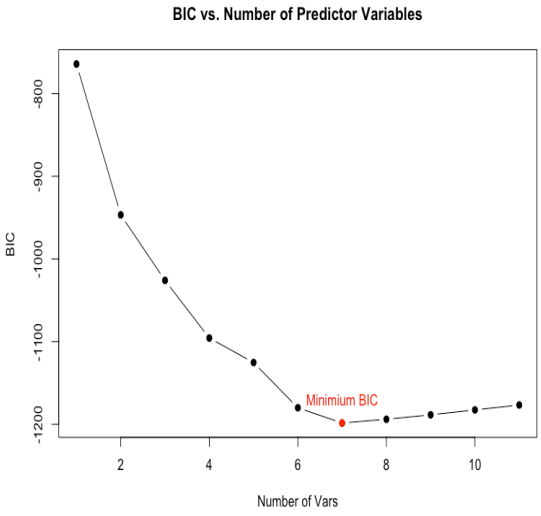


Figure 5

The variables selected by BIC model selection are improvements, tillable, conservation (%), NW, WC, and productivity score. In addition to these variables, I also observed a potential interaction between farm productivity and farm region. As show in figure 6, the smoothed scatter lines show that the effect of productivity may be different depending on region. Although the graph shows all the regions, the interaction between NW and productivity appears strongest and thus I decided test adding it to the model.

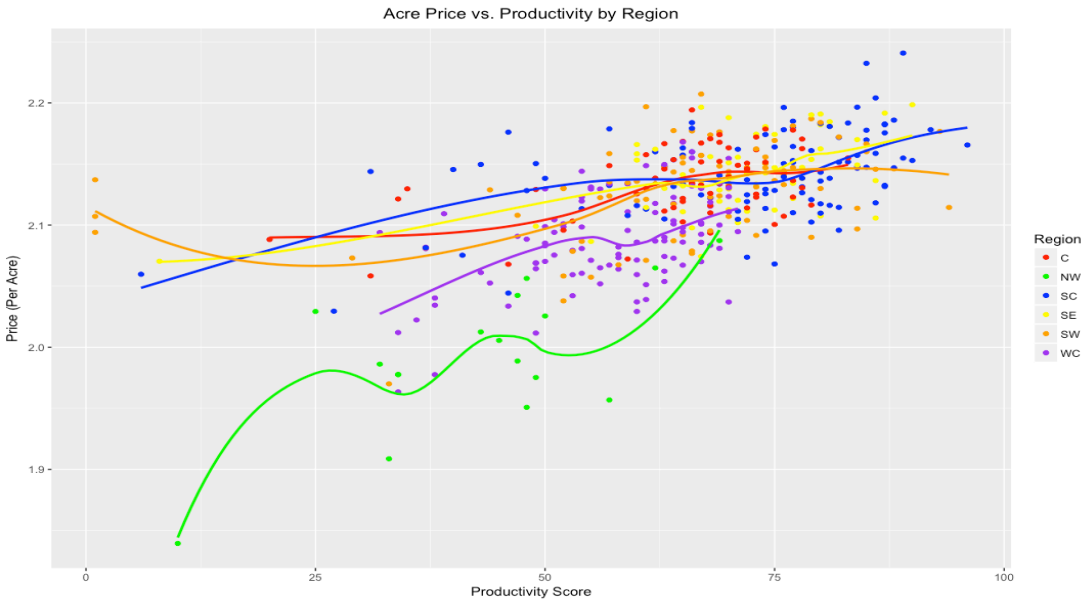


Figure 6

Nate Jenkins
November 2017

The final multiple linear regression model produced by the analysis is the following:

$$\text{Log}(P) = \beta_0 + \beta_1(I) + \beta_2(T) + \beta_3(C) + \beta_4(P) + \beta_5 I(NW = Yes) + \beta_6 I(WC = Yes) + B_7 I(NW = Yes)*(P) + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
7.23	0.01	0.008	-.003	.007	-1.37	-0.29	0.01

- P: The response variable, or price per acre for a farm
- I/T/C/P: Numerical predictor variables corresponding to improvements, percent tillable, percent conservation land, and productivity score
- NW/WC: Categorical predictor variables corresponding to the region that a farm is in. The region is all areas except for NW and WC if no region is specified in the data set
- β_0: The y-intercept, or the expected average price of a farm not in the NW or WC regions with 0% conservation land, 0% tillable land, 0 improvements, and a productivity score of 0
- β_1- β_4: Coefficients for numerical predictors. Holding all else equal, these coefficients represent the average increase in log per acre price for every 1 unit increase in the numerical variables. The only exception is Productivity which increases by (β_6+ β_7) for every 1 unit increase in Productivity if it is in the NW region.
- β_5- β_6: Coefficients for categorical variables, or the expected average increase in log acre price if one of farms changes its region value from any other region to NW or WC holding all else equal. NW, however, also is impacted by the interaction effect. Therefore, log of price increases by $\beta_5 + \beta_7$*P if a farm is in the NW as opposed to elsewhere.
- β_7: The interaction term for NW and Productivity. This is the additional effect of productivity on per acre price if a farm is in the NW region.
- ε: The random error

There are several assumptions made for this model. First, it is assumed that log(P) is linearly related to the predictor variables. Secondly, an assumption is made that the data observations are independent of one another. Also, the model assumes normality of data distributed about the regression line. Lastly, we assume that the data is homoskedastic, or that it is equally distributed along the entirety of the regression line. These assumptions will be discussed in the next section of this report. Once validated, the model will be used to see which of the predictor variables are strong indicators of farm price.

Model Justification and Verification:

In order to ensure the suitability of multiple linear regression, the assumptions discussed above must be investigated in detail. First, the data must be linear. This can be observed in Figure 7 which displays added variable plots for the predictor variables in the model. These added variable plots show the effect of regressing the predictors using the other variables and the y variable using the others as well. Based on this set of added variable plots, the predictors are linearly related to log of price per acre and thus fit the assumption for linearity.

Nate Jenkins
November 2017

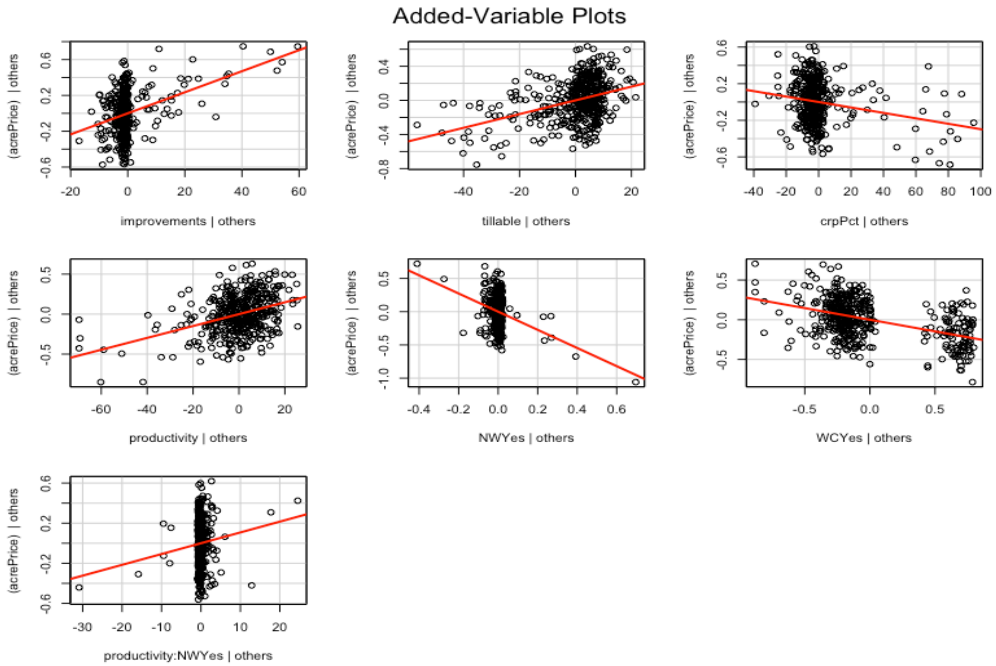


Figure 7

Variance Inflation Factors

Improvements	Tillable(%)	Conservation(%)	productivity	NW	WC	Productivity*NW
1.095381	1.141565	1.17271	1.415808	12.024539	1.139475	11.40408

Figure 8

Figure 8 displays the variance inflation factors for the coefficients in the model. These numbers represent the inflation in variance that is caused by collinearity in the data. In this data set, the only variable with a high VIF is the productivity NW interaction. These numbers, however, are all relatively low, so collinearity is not a problem in this analysis.

Secondly, the data must be independent. This means that the data observations do not affect one another. This assumption may be slightly violated if some of the farms are related in some ways not expressed in the data, such as being owned by the same person. For this analysis, however, the independence assumption appears to be relatively reasonable to accept.

Nate Jenkins
November 2017

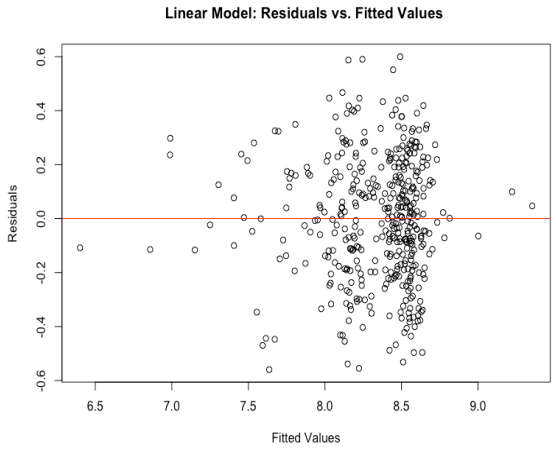


Figure 8

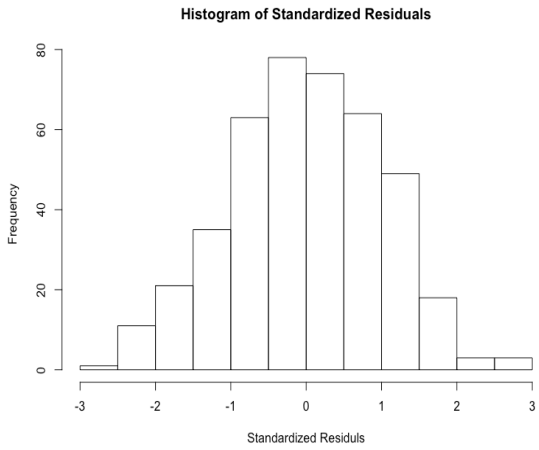
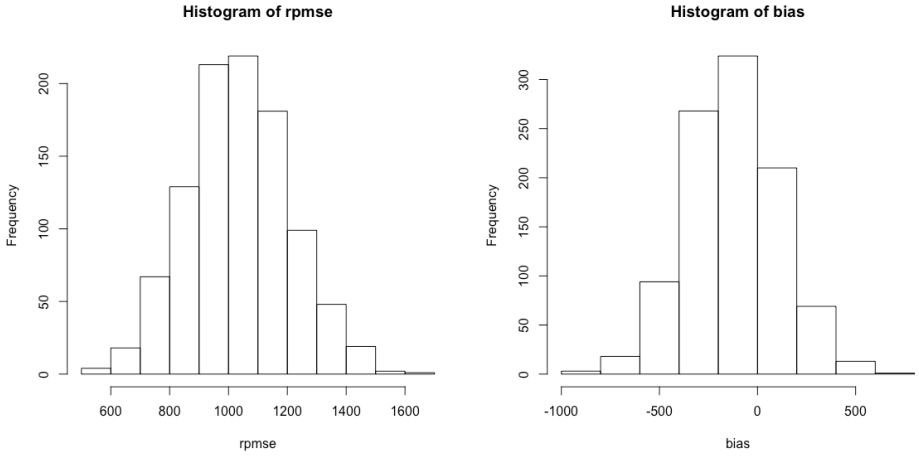


Figure 9

The residuals of the regression must also be equally and normally distributed about the regression line. As seen in figure 8, the residuals appear to be homoskedastic, meaning that they maintain equal variance across the line. This is confirmed by a Breusch-Pagan Test for equal variance on the data which yields a p-value greater than .05 (0.1849). The normality assumption is also upheld as seen in figure 3. The residuals follow a normal distribution which is confirmed by a Kolmogorov-Smirnov test on the data yields a p-value greater than .05 (0.8983). Based on this analysis, all of the assumptions for multiple linear regression are upheld and multiple linear regressions is well suited for the purposes of this analysis.

The model fits the data relatively well. It has an R^2 value of 0.6787 indicating that 67.87% of the variation in the log of price is explained by our predictor variables. Additionally, an F-test for having at least one significant predictor produces a p-value less than 2.2e-16 indicating that there is at least one significant predictor in our model.



The model also has good results in terms of predictive power. Using cross validation, the model has a mean RPMSE of 1034.96 meaning that on average, our predictions for farm price are 1034.96 above or below the actual price. The mean bias is -125.488 meaning that we tend to underestimate farm price by 125.48 on average. The mean coverage for the 95% prediction interval is .957 meaning that 95.7% of the farm prices fall within their 95% predictive range. Based on these results, the model is relatively trustworthy for prediction.

Nate Jenkins
November 2017

Results:

P-Values for Coefficients

	Coefficient	t value	p-value
(Intercept)	7.2	74.16	< 2e-16
Improvements	0.01	8.432	5.77E-16
Tillable (%)	0.008	8.388	7.95E-16
Conservation (%)	-0.003	-4.487	9.37E-06
productivity	0.007	8.529	2.84E-16
NW=Yes	-1.37	-6.905	1.91E-11
WC = Yes	-0.30	-10.527	< 2e-16
Productivity*NW	0.012	2.54	0.01

Confidence Intervals

	2.50%	97.50%
(Intercept)	7.04	7.4
Improvements	0.009	0.014
Tillable (%)	0.0062	0.01
Conservation (%)	-0.0043	-0.0012
productivity	0.006	0.009
NW=Yes	-1.75	-0.977
WC = Yes	-0.35	-0.24
Productivity*NW	0.0024	0.019

Using BIC model selection and the NW productivity interaction, the final model produced has 9 variables. As shown in the table of p-values, all the coefficients individually are statistically significant. In terms of increasing per-acre price, unit increases in productivity and tillable (%) have the greatest effect. Also, the interaction between productivity and NW is statistically significant, indicating that the effect of being in the NW on productivity is not zero.

The table of confidence intervals gives a sense of the uncertainty in these estimates. For example, the coefficient for productivity is .007 meaning that the model estimates a .007 increase in log of acre price on average for each unit increase in productivity. The confidence interval indicates that we are 95% confident that the actual coefficient for productivity is between .006 and .009, or that we are 95% confident that the actual increase in log acre price for each unit increase in productivity is in the interval.

The model can be used for predicting the prices of farms. For example, a farm in the NW region with 0 improvements 94% tillable land, 0% conservation land, and a 96 productivity score is predicted to sell for 4317.44 per acre on average. The 95% prediction interval for this value is (2286.131, 8153.624) meaning that we are 95% confidant that the actual price for the farm will be between 2286.131 and 8153.624 per acre.

Conclusion:

In this analysis, we were able to use data on farm prices to find important factors in appraising farm prices. Using statistical modeling, we found that factors such as productivity and percent tillable land are extremely important for correctly estimating the prices of farms. The information in this study can help farm appraisers better understand how to go about appraising farms and even make sale predictions. Going forward, it would be helpful to collect more data on farms as there is a lot of variation in prices that is not explained by the current model. Finding additional important factors would lead to more insightful appraisal estimates and a more thorough statistical model.