Laura Bialozynski and Priscilla Long
STAT 334, Winter 2021

Final Project Report

Part A: Initial Model

We know that we have 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). **Figure 0A** shows the output that we get from running this initial model.

Part B: Checking Assumptions, Multicollinearity, and Influential Observations

We checked linearity and equal variance assumptions by looking at the residual vs. predicted plots. **Figure 0B** shows all the residual vs. predicted plots. We can see that the linearity regression assumption is not violated since none of the residual plots show curvature. 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, the equal variance regression assumption is violated. We are not told in the study design whether or not our data were randomly sampled. Since 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 the independence assumption as we work towards our final model. We will consider this condition okay because there is nothing in our analysis that we could do to fix this. **Figure 0D** shows the QQ plot of residuals above has a distinct curve away from the diagonal towards the end. **Figure 0E** shows that the p-value for the Shapiro-Wilk test is <2.2e-16, which is very small. Thus, the normality assumption is violated. However, since we have such a large sample size there is a small effect on the regression assumption. Thus, we don't have to worry about the normality violation. All the VIF's are below 4 which suggests no multicollinearity, as is shown in **Figure 0F**. The cut-off when looking at the studentized residuals for outliers is less than -3 or greater than 3. From **Figure 0G** we see that there are 9 data points with studentized residual points greater than 3 while there are none less than -3. For leverage, we computed our cutoff value for high leverage. In this case, we have k=13(number of predictors) and n=813(number of observations). 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 **Figure 0H** 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. The cutoff for influence in this data set for Cook's distance is 0.918782 which we got from running code. 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). From **Figure 0I** there are no influential values.

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

We began by transforming our y variable since we have problems with equal variance, normality, and high leverage points. For equal variance, the acres vs residual plot shows the fan is larger on the left and narrower on the right. Because of both of these issues that need to be fixed, we will transform y up the ladder of powers. We ran code to get a suggested starting transformation power. From this, R suggested 0.3629529 as the power, but we know that we want a transformation power closest to the nearest 0.5. This value from R is closest to 0.5, so we use the square root transformation. In R, we tried powers 0 and 1 to ensure that 0.5 was in fact the best transformation to use, however, ½ remained the best transformation. The y transformation fixes our residual plots, as can be seen in **Figure 1A**. We can still see a fan shape in acres and there is still potential for a fan shape in tillable, but the single point at 60 with a high residual value prevents us from concluding this. The normality violation is still not much of a concern because

of the large sample size (n = 816). To fix the high leverage points we know that we should transform the x variables. We can see from the residual plots from **Figure 2A** that transforming acres to a power of 0, as a log transformation, immensely fixes our equal variance problems. This is the only transformation to a predictor variable that is needed in order to address our regression assumption violations.

D. Variable reduction

We began by running all possible interactions, the output of which can be seen in **Figure 3**. In this figure, we can see all the statistically significant interactions and from there we can get our next model. **Figure 4** is our transformed model with our statistically significant interactions. However, in our most updated model, model 4, some of our predictor variables have high p-values and don't have interactions. Since the p-values are high and not statistically significant we are able to proceed with variable reduction. **Figure 5** shows our summary output with only statistically significant variables and interactions. **Figure 5A** shows us our severe problem with model 5 multicollinearity as most of them are in their hundreds. Since our multicollinearity is so high, we use a stepwise function as another way to find significant interactions seen in **Figure 3A**. Even though this process reduces our VIF's, they are still considerably high to be seen as a problem. These issues with multicollinearity lead up to believe that something else needs to change with our predictor variables to bring the VIFs down. To reduce our multicollinearity further, in **Figure 6** we have the summary output of our centered quantitative variables. We centered all five of our quantitative predictor variables, and it is important to note that, for acres, we first took the log of the variable and then centered it. If we centered before taking the log, this would not work, because we cannot take the log of a negative value, which we would get from centering first. Because we decided in part C above that the acres predictor variable needs to be log-transformed, this is how we will address this issue. Since we are centering our variables, we want to check all possible interactions again. **Figure 7** includes all significant variables and interactions after the insignificant terms from **Figure 6** are removed. Our variables cProd and financing are not significant and do not have any significant interactions, so we can remove these two variables for our next model. **Figure 8** is the summary output of our final model. This model includes only statistically significant variables and interactions.

E. Check of assumptions, multicollinearity, and influential observations for the final model

We have found our final model, lp.model8 to best represent our data set. Now we want to check our assumptions. Independence has not changed from our initial model, so we assume this condition is still okay. From **Figure 8A**, the residual plots don't show any sign of curvature or fan shape. Thus, we can conclude that linearity and equal variance are not violated. We can check for normality in **Figure 8B** and **Figure 8C.** The QQ plot shows a distinct curve away from the diagonal which indicates a violation. The p-value for the Shapiro-Wilk test is 7.337x10^-11 which is very low, so we can conclude that there is a normality assumption violation. However, since we have such a large sample size this violation will have a relatively small effect on the overall regression, therefore we don't have to worry about this violation, similar to in the initial model. Our VIF's are significantly better in this model. Although we have a few high VIFs shown in **Figure 8D**, these are unfixable. In **Figure 8E**, we are checking outliers with the same cutoff value of +/- 3. There are some outliers however these are also unfixable. With a cutoff of 3(k+1)/n = 3(9+1)/813 = 0.0369 for high leverage, **Figure 8F** shows a lot of the points exceeding this cutoff. In R, we can find the cutoff for influence at 0.918782. From **Figure 8G** we can see that there are still no influential observations, as all the observations stay below 0.5. We were

able to confirm that model 8 is our best model because it includes all the variables from the best subset output in **Figure 8H.** The subset with the highest adjusted R^2 and lowest Cp, AIC, and SBC are the same terms in our final model.

F. Discussion of any remaining (unfixable) problems with your final model

We were unable to fix the normality assumption violation. This violation does not affect the regression assumption because we have a very large sample size of 816. From **Figure 8D** cImp, cTill, and cPct have high VIF's that are considered moderate or severe. However, we are unable to reduce these even further because this means that they are correlated to another predictor variable. We used variable reduction to conclude that this correlation must be due to the categorical predictor variable, region. We are not able to test the correlation between quantitative variables and our categorical variable region, and we are also not able to center a categorical variable to reduce multicollinearity as we would with a quantitative predictor variable. Model 8 is also the best model with the lowest amount of multicollinearity compared to other models. We also have outliers and high leverage points. Since there are so many data points from a large number of observations, we are unable to control the points that don't fall close to the center of all the data points.

G. Interpretation of Significant Predictors (final model)

We concluded above that the significant predictors included in our final model are centered improvements, centered tillable, centered crpPct, region, and the interactions between region and each of the previously-mentioned quantitative variables. We can see from our summary output of our final model in **Figure 8** that our final regression equation can be written as an equation where the square root of the response variable is equal to the coefficients of each significant predictor variable and interaction times the corresponding variable/interaction. We can see from our output that Central is considered the baseline group for our categorical variable, region. So, for a farm in the Central region of Michigan, we have the following regression equation: sqrt(acrePrice) = 68.36 + 0.39*cImp + 0.52*cTill + -0.17*cPct. This means that every increase in one above the average in the percentage of property value due to improvements is associated with an increase in the square root of acrePrice of 0.39, which means that an increase in 1 in centered improvements in associated with an increase of $0.15 in the sale price per acre for a farm, adjusting for all other predictors. Similarly, every increase in one above the average in percentage of the farm that is tillable is associated with an increase in the square root of acrePrice of 0.52, which means that an increase in 1 in centered tillability percentage is associated with an increase in $0.27 in the sale price per acre for a farm, adjusting for all other predictors. Every increase in one above average in the percentage of the farm acres enrolled in CRP is associated with a decrease in the square root of acrePrice of 0.17, which means that an increase in 1 in centered crpPct is associated with a decrease of $0.03 in the sale price per acre for a farm, adjusting for all other predictors. Finally, this means that, when improvements, percent tillable, and crpPct are all average, we have that the square root of acrePrice is our intercept, 68.36, which means that our intercept, when all other predictor variables are average, and we are in our baseline group of Central region, an acre of land is predicted to be approximately $4673. For the other 5 regions, we would find similar interpretations, however, with the other regions, we would need to include the coefficients of the interaction terms. Because of this, each of the other 5 regions will have both a different b_0 intercept as well as different values for the effect of improvements, tillability, and crpPct on the sale price in dollars per acre.