STAT 330 Midterm Analysis #2 - Farm Appraisal

Jillian Maw

3/23/2022

Section 1: Introduction and Problem Background

In this report, I explore the various factors that can determine the value of a farm's price value, as used by farm appraisers in Minnesota. It is important to ensure a fair sale price is reached between buyer and seller to prevent tax fraud or monopoly. I looked at the data of past farm sales to determine what factors are used to appraise the value of a farm. The factors I explore to determine the value of a farm's price value include what percentage of the property value was due to buildings, what percentage of the farm was rated arable, what type of financing was used to sell the farm, what percentage of the land was enrolled in conservation reserve (or protected land), how productive the land was (on a numeric scale), which region of Minnesota the farm in question was located, and what the sale price of the farm ultimately was (per acre of land). Determining which of the former listed factors are used from the data set will allow me to create a multiple linear regression model any farm appraiser in Minnesota can use to determine the values of any new farms in the future, excluding changes in price from inflation.

Below are tables and graphics highlighting main points from the exploration of the data set. The covariance matrix shows which quantitative variables have positive or negative relationships with each other. Specifically looking at response variable, `acrePrice`, or the price of the farm's sale price per acre, it can be seen that the explanatory variable `crpPct`, or the percentage of the land that was enrolled in conservation reserve, is the only explanatory variable to have a negative relationship with the farm's sale price per acre. The correlation matrix shows that the strength of most of the quantitative variable's relationships is weak, meaning they have low correlations with each other. The only exception is the productivity of the land and the farm's sale price per acre, which is of medium strength at 0.5644 out of 1.

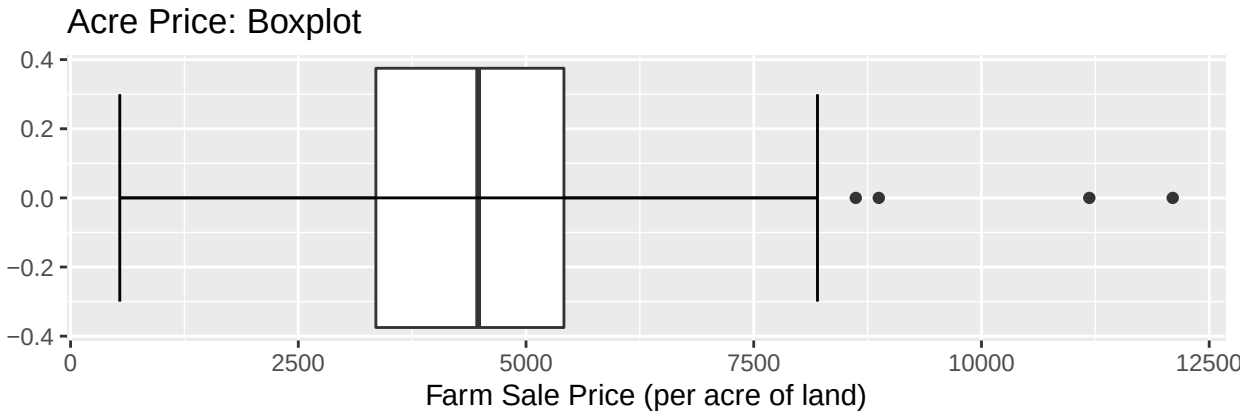
Table 1: Covariance Matrix

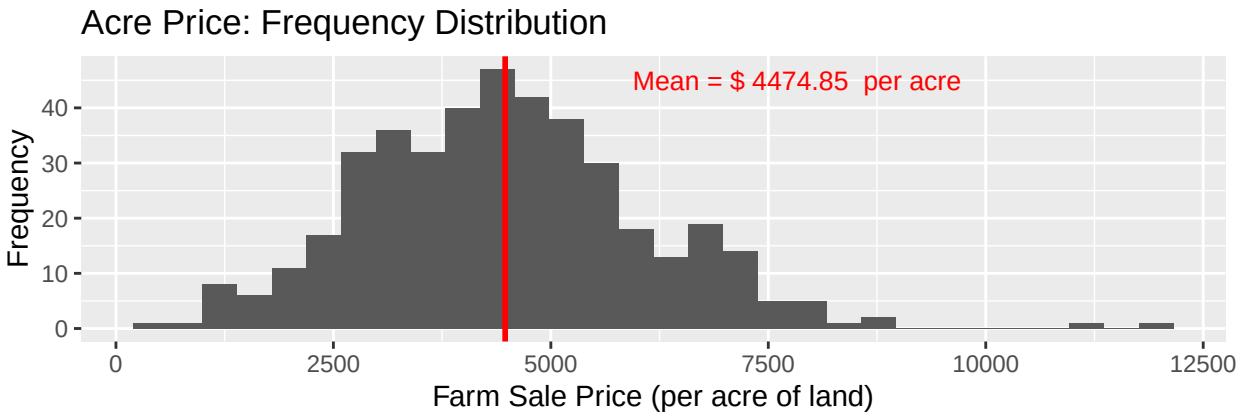
	improvements	tillable	crpPct	productivity	acrePrice
improvements	71.3939	-28.5062	1.9322	-5.7223	2974.789
tillable	-28.5062	157.1129	-30.4970	43.2293	5262.695
crpPct	1.9322	-30.4970	338.8730	-95.7043	-8316.942
productivity	-5.7223	43.2293	-95.7043	236.6903	13993.376
acrePrice	2974.7886	5262.6947	-8316.9419	13993.3760	2597306.749

Table 2: Correlation Matrix

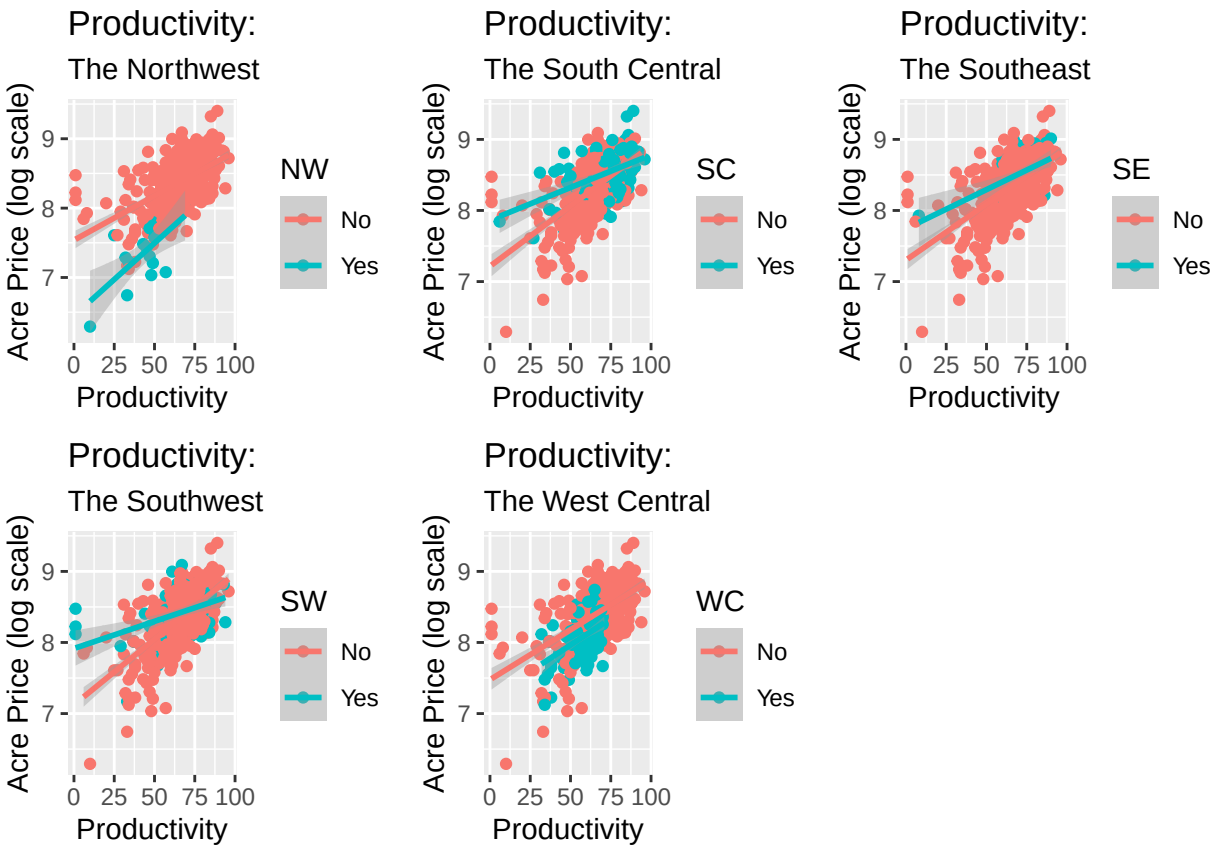
	improvements	tillable	crpPct	productivity	acrePrice
improvements	1.0000	-0.2692	0.0124	-0.0440	0.2185
tillable	-0.2692	1.0000	-0.1322	0.2242	0.2605
crpPct	0.0124	-0.1322	1.0000	-0.3379	-0.2803
productivity	-0.0440	0.2242	-0.3379	1.0000	0.5644
acrePrice	0.2185	0.2605	-0.2803	0.5644	1.0000

The boxplot and frequency distribution below show that the response variable `acrePrice`, the price of the farm's sale price per acre, is skewed right by outliers, and may possibly need a transformation to allow us to apply the linearity, equal variance, and normality assumptions needed to use multiple linear regression models.





Finally, to round out the last of the data exploration, I took a look at the relationship between the explanatory variable `productivity`, of the land, the explanatory variable of the region the farm was located in (i.e. Northwest, South Central, Southeast, Southwest, and West Central; I could not specify just the Central region in the graphs), and the response variable `acrePrice`, the price of the farm's sale price per acre. I transformed the acre price to the log scale to more strongly show the possible interaction between productivity and the Northwest region that the appraiser believed to exist. The graphs below suggest there is a distinct difference in productivity in the Northwest region compared to the other regions; specifically, I think there might be a reduction in the farm sale price due to lower productivity in the Northwest region of Minnesota.



Due to the right skew of the acre price response variable in the data, an untransformed multiple linear regression model would be inappropriate because it would not fulfill all of the necessary requirements for the statistical modeling assumptions necessary to create a true multiple linear regression model. Those assumptions are linearity, independence, normality, and equal variance. Skewed data violates the normality assumption. A log transformation of the acre Price response variable fixes the violation of the normality assumption, allowing for a multiple linear regression model to be used.

Section 2: Statistical Modeling

Due to the number of explanatory variables in the data being less than 40, I can use the "best subset selection" variable selection procedure because it is the best method for minimizing the Aikake Information Criteria, the Bayesian Information Criteria, or the Predictive Error, and maximizes the Adjusted R^2 for my multiple linear regression model. These criteria are the best indicators for determining the fit of my model. In addition, the "best subset selection" procedure tests all possible variables with each other, allowing me to know with certainty that I have the best model. I decided to use the Aikake Information Criteria, or AIC, model comparison criterion, because I am looking to make predictions on the true sale price of a farm in Minnesota, and the AIC model comparison criterion is optimized for making predictions, compared to the Bayesian Information Criteria, or BIC, which is optimized for making inferences. After using the "best subset selection" variable selection method and the AIC model comparison criteria, I decided the most important variables to include were `improvements` (the percentage of the property value due to buildings), `tillable` (the percentage of the farm that was rated arable), `crpPct` (the percentage of the land that was enrolled in conservation reserve), `productivity` (how productive the land was), `NW` (whether or not the farm was located in the Northwest region of Minnesota or not), and `WC` (whether or not the farm was located in the West Central region of Minnesota or not), and the log of `acrePrice` (the log of the farm's sale price per acre of land).

Below, I include the mathematical formula (with Greek letters) for my multiple linear regression model:

$$\log(y_i) = \beta_0 + \beta_1(\text{Improvements}_i) + \beta_2(\text{Tillable}_i) + \beta_3(\text{crpPct}_i) +$$

$$\beta_4(\text{Productivity}_i) + \beta_5 I(\text{Region} = \text{Northwest}_i) + \beta_6 I(\text{Region} = \text{West Central}_i) +$$

$$\beta_7 I(\text{Region} = \text{Northwest}_i)(\text{Productivity}_i) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

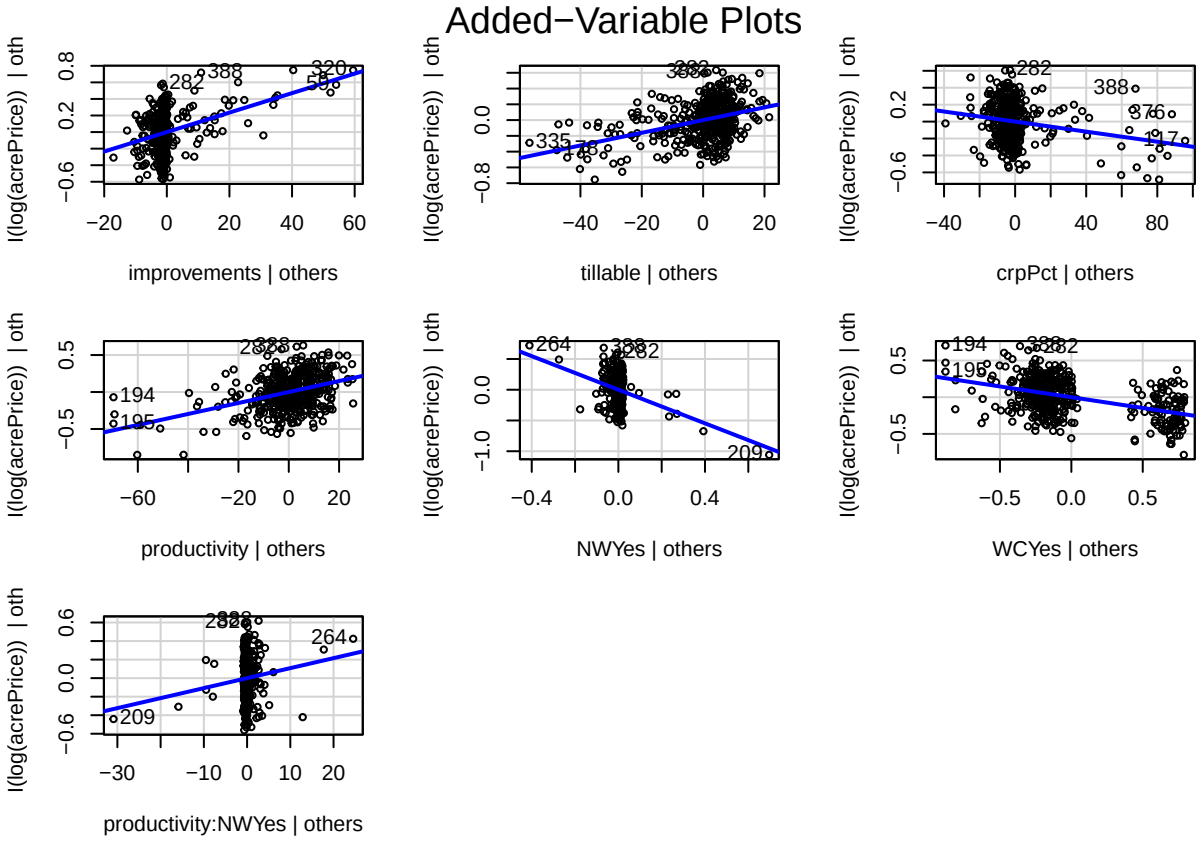
The Greek letters and unique statistical notation are commonly used in statistical models and are explained following this sentence. The i represents the i^{th} farm in the data set, or the individual Minnesota farm being represented by the multiple linear regression model. y represents the sale price of the farm, per acre of land. In the model, the log of y better fits the needed assumptions for multiple linear regression models. The coefficient β_0 represents the intercept, or when all the other explanatory variables in the model are zero and the region where the farm is located is *not* the Northwest or the West Central regions of Minnesota, then y is β_0, on average; this is a next-to-useless interpretation, however, and it is better to think of β_0 as the base, average sale price of a Minnesota farm. The coefficient β_1 represents the change to a farm's sale price based on the percentage of the property value due to buildings, increasing by 1, assuming all other variables do not change value as well. The coefficient β_2 represents the change to a farm's sale price based on the percentage of the farm that was rated arable, increasing by 1, assuming all other variables do not change value as well. The coefficient β_3 represents the change to a farm's sale price based on the percentage of the land that was enrolled in conservation reserve, or is protected land, increasing by 1, assuming all other variables do not change value as well. The coefficient β_4 represents the change to a farm's sale price based on how productive the land was rated increasing by 1, assuming all other variables do not change value as well. I is an indicator for a variable that is encoded as a 0 or 1 in the data set and, in this model, is used for indicating which region the farm is located in. The coefficient β_5 represents the change to a farm's sale price based on if the location of the farm was in the Northwest region of Minnesota, assuming all other variables do not change value as well. The coefficient β_6 represents the change to a farm's sale price based on if the location of the farm was in the West Central region of Minnesota, assuming all other variables do not change value as well. The last coefficient, β_7, represents the change to a farm's sale price from the interaction of productivity and the farm's location in the Northwest region of Minnesota, as the farm appraiser believes to exist, assuming all other variables do not change value as well. ϵ represents the residual errors, or the difference from the true average of the farm's sale price per acre of land. The N is short for Normal distribution, meaning the multiple linear regression model's residuals (or, the difference between the true sale price of a farm and the predicted sale price of a farm) follow a Normal distribution's shape and behaviors, standardized at a mean of \$0 and a standard deviation of $\$\sqrt{\sigma^2}$. The symbol $\overset{iid}{\sim}$ means "independent and identically distributed", which means I assume the model meets two of the assumptions needed for multiple linear regression modeling. σ^2 represents the variance of the data around the regression line fitted to the data by this model. Another way to think of the variance is that it is the square of the standard deviation. The standard deviation shows that for any farm's appraisals, 99.7% of the response variables will be within three standard deviations of the regression line made by the model.

In the above model, I assume I am able to meet the assumptions of linearity, independence, equal variance, and normality in order to create a multiple linear regression model. I cannot use multiple linear regression modeling accurately without meeting the requirements for these assumptions. If the linearity assumption is broken, the data is not linear, and I cannot use multiple linear regression to model the data's patterns; I would have to use another modeling tool. If the independence assumption is not met, then the estimates of a farm's sale price would be unbiased (i.e. would still be close to the true average of a farm's sale price), but the accuracy of those estimate, the standard errors ϵ_i, would be too small to capture the true value of a farm's sale price. Likewise, the normality assumption being broken would result in the estimates of a farm's sale price still being unbiased (i.e. still being close to the true average of a farm's sale price), but the confidence and prediction intervals would be wrong, and I wouldn't be able to use a t-distribution to determine how certain (or likely) a predicted value of a farm's sale price would be. Finally, if the equal variance assumption is not met, the model's estimates of a farm's sale price are still unbiased but the standard errors would be wrong to a point that cannot be accounted nor corrected for. Below, I will prove why I believe these assumptions are justified in being used.

Section 3: Model Justification and Verification

When I fit the numeric coefficients to the model, it becomes: $\log(\hat{y}_i) = 7.2321 + 0.0118$ (Improvements$_i$)+ 0.008 (Tillable$_i$)+ -0.003 (crpPct$_i$)+ 0.0074 (Productivity$_i$)+ -1.3658 I(Region = Northwest$_i$)+ -0.2936 I(Region = West Central$_i$)+ 0.0108 I(Region = Northwest$_i$)(Productivity$_i$) $+ \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, 0.0531)$, where $\log(\hat{y}_i)$ represents my best prediction for the log of a farm's sale price per acre of land. I will be proving the assumptions needed for this multiple linear regression model (that is, the linearity, independence, normality, and equal variance assumptions mentioned above) with the graphics and summary statistics following this paragraph.

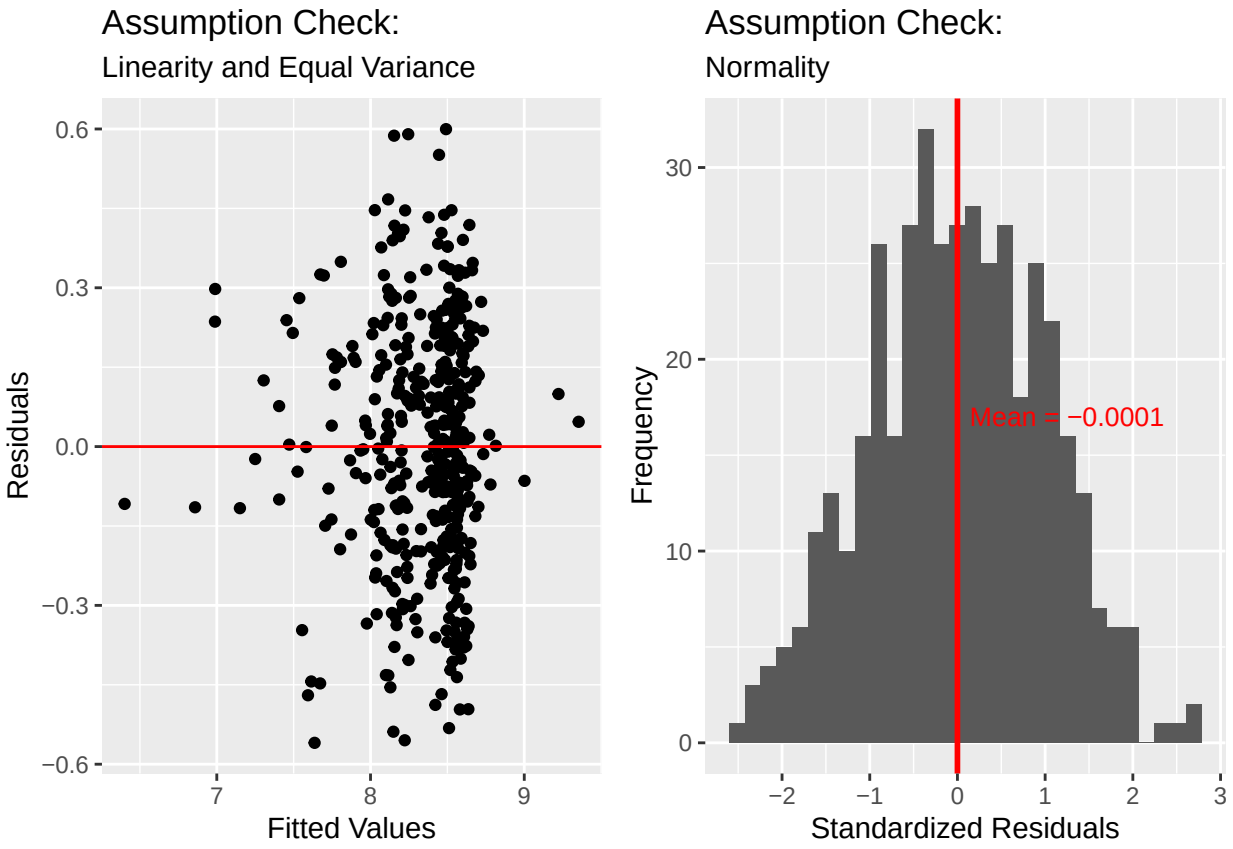
Below are the Added-Variable Plots. The graphs all seem to show clustering data at some central point, with outliers possibly affecting the slope, in that they pull the slope through a cluster of points in one direction over another. However, there are no patterns in the plots showing that other graphical line would be appropriate (e.g. exponential), so I can assume the linearity assumption is met. I will double check the linearity with the soon-to-be shown fitted values versus standardized residuals plot after the Added-Variable Plots.



I can assume the farm sale price data is independent enough for this multiple linear regression model, because the sale price of one of the farms should not affect the sale price of another farm, the farms are spread out over several regions of Minnesota, I am testing for the interaction between productivity and the Northeast region, and the variable measurements of one one farm should not affect another farm. Even if a few farms are right next door to each other in one or two regions, they are not all (or at least, not in great numbers) next to each other, so I feel safe in assuming that the farm sale prices are independent of each other, as are the explanatory variables used for appraising farms.

The graphics below are a scatterplot of fitted values (predicted value for the farm sale price in the data set) versus residuals (the difference between the observed and the predicted farm sale price) and a histogram of

standardized residuals (residuals transformed to show their difference from the observed farm sale price's mean, if the mean farm sale price was \$0). The fitted values versus residuals scatterplot helps me reaffirm the linearity assumption if the points lacks patterns in plotting and affirm the equal variance assumption if the points have constant variance, or random, scattered points on the plot. Since the plots shows a constant variance with a lack of patterns, the data has equal variance and those assumption appear to be proven correct. The histogram shows a generally Normal distribution, with only one curve with relatively balanced sides, so I can assume the normality assumption is also met.



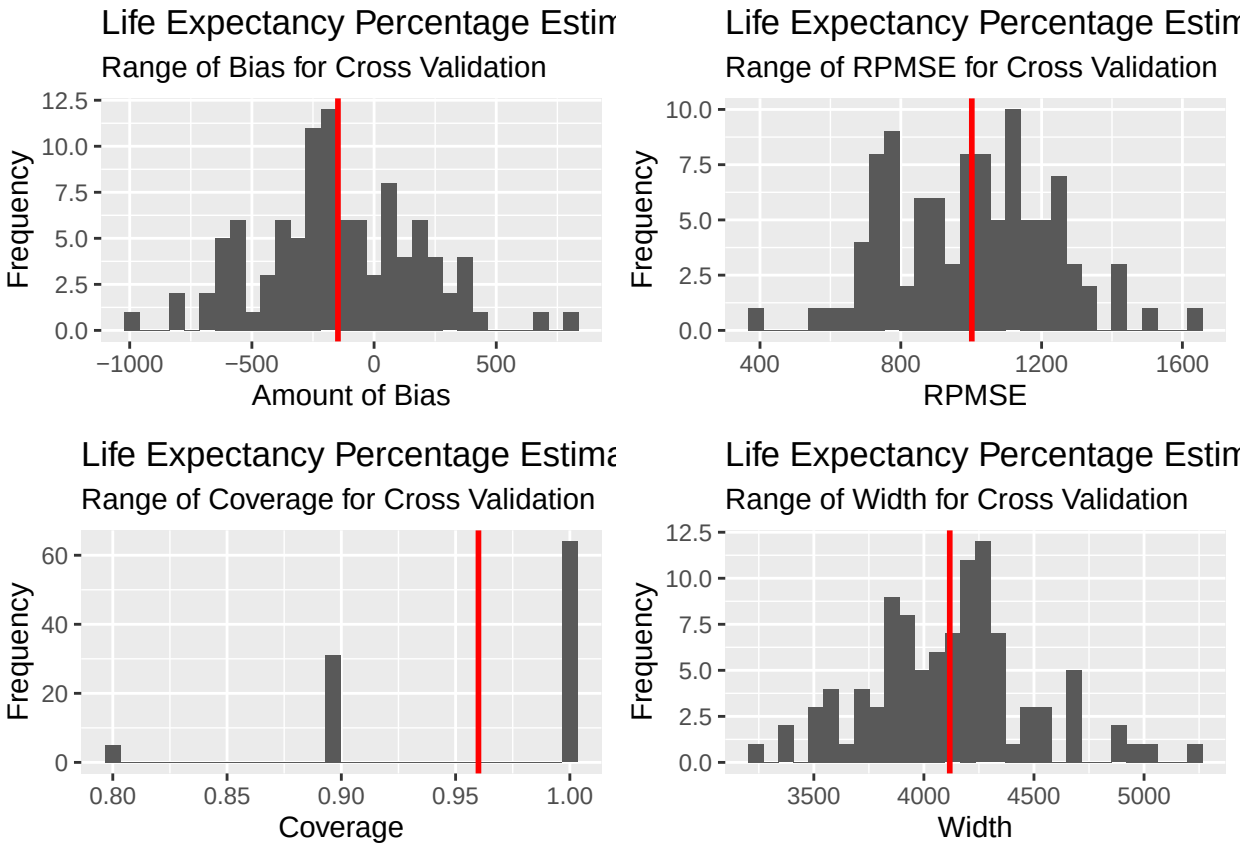
To further prove the data meets the Normality assumption for the multiple linear regression model, I conducted a One-sample Kolmogorov-Smirnov test, also called a KS-test, and a Jarque-Bera test for normality, also called a JB-test. These tests conduct hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution *is* skewed. I set the p-value to be 0.05 for both tests, to prove significance. The KS-test produced a p-value of 0.8983, so I failed to reject the null hypothesis. The JB-test for normality produced a p-value of 0.191, so I failed to reject the null hypothesis for both tests. I accept that I have the normality assumption met for our data.

To further prove the Equal Variance assumption for our model, I conducted a Breusch-Pagan test, or BP-test. Checking the scatterplot of fitted values versus residuals above, it appears that I have a model with linearity and mostly equal variance, so I can proceed with the BP-test. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis is that the data has homoskedasticity, while the alternative hypothesis is that the data has *heteroskedasticity*; these essentially mean "same variance" and "not same variance", respectively. I set the p-value to be 0.05, to prove significance. The test produced a p-value of 0.1849, so I failed to reject the null hypothesis. I accept that I have the Equal Variance assumption met for our data.

To prove the fit of my model, I reviewed the percent of variability in farm sale price per acre of land explained by the covariate variables in the adjusted multiple linear regression model above in the beginning of Section 4. The percent of variability in farm sale price is R^2, 67%, which is reasonable, considering the transformation we had to make of our data and the clustering that could be seen in the Added-Variable Plot. An R^2 of 67% is relatively strong. The standard deviation, or how much the true farm sale price varies around the mean of farm sale prices, is 0.23, which is quite small a difference from any mean, all things considered, giving further proof that the model fits well.

With all of these results and summary statistics, I feel it is safe to say that my predictions are rather good, relative to the original spread of the response variable, farm sale price per acre of land.

I conducted a cross validation procedure to prove my model is valuable for predicting the price of a farm. I conducted this procedure 100 times on 10 randomly selected observations of the data, newly selected each time, and calculated an average bias of -147.84. This means my predictions are, on average, somewhat lower than the true average of life expectancy, but are no means a large bias. I also calculated an average Root Predictive Mean Square Error of 1001.77, which means my predictions are off, on average, $1001.77 per acre. Considering the range of farm values is between $540.45 per acre and $12,099.72, this seems a reasonable amount of error, though it may seem a large margin for error. To see how far the predictions ranged, I calculated the width to be $4117.43 per acre of land, on average. The width should cover a smaller range than the range of data response variable, farm sale price per acre of land, so the width is appropriate for this model and data set. In addition, the coverage, or the percentage of prediction intervals that contain the true average of farm sale price per acre of land, to be 96%. Below the following paragraph are graphs showing in greater detail the results of the cross validation procedures, with the red lines representing the mean values relative to the results shown.



Section 4: Results

Based on my fitted model, I think the selected variables have a significant effect on a farm's sale price per acre of land. I can prove this significance with the results of an F-test on the model. The null hypothesis of an F-test is that there is no effect from a selected variable on the response variable, a farm's sale price, the alternative hypothesis is that there is an effect on the response variable, and the significance level is set to 0.05. With the resulting F-statistic of 124.3484248 and a p-value so far below the set level of 0.05, the F-test calls for a rejection of the null hypothesis and I determine that the selected variables do, in fact, have a non-zero effect on a farm's sale price per acre of land.

From the confidence intervals that can be created from the selected variables, I can state the effect on a farm's sale price numerically and provide a range of the likely amount of that effect. I am 95% confident that the true effect of adding buildings to the farm (the `improvements` variable) on a farm's sale price will be, on average, between \$0.009 and \$0.0145 per acre of land. I am 95% confident that the true effect of having arable land (the `tillable` variable) on a farm's sale price will be, on average, between \$0.0062 and \$0.0099 per acre of land. I am 95% confident that the true effect of having protected land (the `crpPct` variable) on a farm's sale price will be, on average, between -\$0.0043 and -\$0.0017 per acre of land. I am 95% confident that the true effect productive farm land (the `productivity` variable) on a farm's sale price will be, on average, between \$0.0057 and \$0.0091 per acre of land. I am 95% confident that the true effect of having a farm in the Northwest region of Minnesota (the `NW` variable) on a farm's sale price will be, on average, between -\$1.75 and -\$0.98 per acre of land. I am 95% confident that the true effect of having a farm in the West Central region of Minnesota (the `WC` variable) on a farm's sale price will be, on average, between -\$0.35 and -\$0.24 per acre of land. Finally, I am 95% confident that the true effect of the interaction between a Northwest regional location and land productivity on a farm's sale price will be, on average, between \$0.0024 and \$0.0191 per acre of land.

The above paragraph is a dense wall of text, but I will break down two of those confidence intervals. First, the `tillable` variable: as the percent of tillable land increases by 1, assuming all other variables stay the same, I expect the farm's sale price will increase, on average, between between \$0.0062 and \$0.0099 per acre of land. Second, the `WC` variable: if the farm is located in the West Central region of Minnesota, and assuming all other variable stay the same, I expect the farm's sale price will drop, on average, between \$0.24 and \$0.35 per acre of land.

I agree with the farm appraiser that their prior intuition that the effect of productivity in the Northwest region is different than in other areas of Minnesota, based on the previously mention exploratory analysis graphics, but I will prove it by conducting a different F-test between my model, and a new model that drops the interaction between the variables in question. After performing an F-test, to determine if the interaction between the productivity of the farm and the farm's location in the Northwest region of Minnesota is significant at the p-level of 0.05, I determined that the resulting p-value is below this level at 0.01155. This proves that the interaction is significant and should be included in my multiple linear regression model for more accurate farm sale price predictions. That interaction effect is included in the multiple linear regression model as 0.0108, but could range, as stated earlier above with the confidence intervals, between

I have been asked to use my model to predict the farm price for a title transfer farm in the Northwest region of Minnesota with `improvements` = 0, `tillable` = 94, `crpPct` = 0, and `productivity` = 96. To determine the predicted farm sale price per acre with the ratings and numbers stated previously, I input the ratings used in the multiple linear regression model (as stated in Section 4) and determined I am 95% confident that if the ratings and numbers are as previously stated, the associated, predicted farm sale price per acre would be between \$2286.13 per acre and \$8153.62 per acre, on average. To be more specific, \$2286.13 per acre is our estimated answer for this farm's sale price per acre, but could easily be in the previously stated range of \$0.0024 and \$0.0191 per acre of land,

Section 5: Conclusions

In conclusion, I found that the relationship between a farm's sale price and the variables mentioned earlier, is equally positive and negative, depending og the variable, once the data had been transformed and proven with tests and cross validation. This relationship could be affected by the assumptions made, namely that the data (once transformed) is linear, independent, follows a Normal distribution, and has equal variance about the regression model, but I did what I could to adjust for the transformation to the data to be true. I proved that my multiple linear regression model is a good fit for the data and is good at making predictions of future farms. I then created a prediction based on my model from the variable numerics given above in the end of Section 4.

To better understand and predict a farm's sale price the appraiser could consider checking if inflation or world news could affect a farm's sale price, over the years. They could also seek to revalidate the study contained in this report, perhaps by working with another appraiser.

Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(ggplot2) #for professional looking graphics
library(GGally) #for ggpairs
library(MASS) #for standardized residuals
library(normtest) #for JB-test
library(lmtest) #for BP-test
library(car) #for variance inflation factors & added-variable plots
library(bestglm) #for variable selection procedures
library(knitr) #for pretty tables with kable
library(kableExtra) #for if kable needs to be landscape
library(gridExtra) #for making grids on same row
options(scipen = 999) #for preventing scientific notation
farm <- read.table("~/R programming/STAT_330/Farms3.txt", sep = ' ', header = TRUE,
                   stringsAsFactors = TRUE)
#reorder dataset to use bestglm
farm <- farm[,c(2:ncol(farm),1)]
farm$NW <- as.factor(farm$NW)
farm$SC <- as.factor(farm$SC)
farm$SE <- as.factor(farm$SE)
farm$SW <- as.factor(farm$SW)
farm$WC <- as.factor(farm$WC)
# (a)  In your own words, describe the background of the problem and the goals of the study.
# (b)  Summarize what data you are going to use to fulfill the goals mentioned above.  Explore the data
# (c)  Determine whether or not an *untransformed* multiple linear regression (MLR) model would be appr
#EDA
kable(cov(farm[,c(1,2,4,5,11)]), digits = 4, booktabs = TRUE, row.names = TRUE,
      caption = 'Covariance Matrix') %>%
  kable_styling(latex_options = "striped") %>%
  kable_styling(latex_options = "HOLD_position")
kable(cor(farm[,c(1,2,4,5,11)]), digits = 4, booktabs = TRUE, row.names = TRUE,
      caption = 'Correlation Matrix') %>%
  kable_styling(latex_options = "striped") %>%
  kable_styling(latex_options = "HOLD_position")
acre.box <- ggplot(data=farm, mapping=aes(x=acrePrice)) +
  geom_boxplot() +
```

```
  xlab('Farm Sale Price (per acre of land)') +
  ggtitle('Acre Price: Boxplot') +
  stat_boxplot(geom ='errorbar', width = 0.6)
acre.mean <- mean(farm$acrePrice)
acre.freq <- ggplot(data = farm, mapping=aes(x = acrePrice)) +
  geom_histogram() +
  geom_vline(xintercept = acre.mean, col = "red", lwd = 1) +
  annotate("text", x = acre.mean * 1.75, y = 45, col = "red", size = 3.5,
           label = paste("Mean = $", round(acre.mean, digit = 2), " per acre")) +
  xlab('Farm Sale Price (per acre of land)') +
  ylab('Frequency') +
  ggtitle('Acre Price: Frequency Distribution')
suppressMessages(grid.arrange(acre.box, acre.freq, nrow = 2))
productNW.scatter <- ggplot(data = farm,
                            mapping=aes(x=productivity,
                                        y=log(acrePrice),
                                        color = NW)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab('Productivity') +
  ylab('Acre Price (log scale)') +
  ggtitle('Productivity:', subtitle ='The Northwest')

productSC.scatter <- ggplot(data = farm,
                            mapping=aes(x=productivity,
                                        y=log(acrePrice),
                                        color = SC)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab('Productivity') +
  ylab('Acre Price (log scale)') +
  ggtitle('Productivity:', subtitle ='The South Central')

productSE.scatter <- ggplot(data = farm,
                            mapping=aes(x=productivity,
                                        y=log(acrePrice),
                                        color = SE)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab('Productivity') +
  ylab('Acre Price (log scale)') +
  ggtitle('Productivity:', subtitle ='The Southeast')


productSW.scatter <- ggplot(data = farm,
                            mapping=aes(x=productivity,
                                        y=log(acrePrice),
                                        color = SW)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab('Productivity') +
  ylab('Acre Price (log scale)') +
  ggtitle('Productivity:', subtitle ='The Southwest')
```

```r
productWC.scatter <-
  ggplot(data = farm, mapping=aes(x=productivity, y=log(acrePrice), color = WC)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab('Productivity') +
  ylab('Acre Price (log scale)') +
  ggtitle('Productivity:', subtitle ='The West Central')

suppressMessages(grid.arrange(productNW.scatter, productSC.scatter,
                              productSE.scatter, productSW.scatter,
                              productWC.scatter, nrow=2))
# (a) Using justifiable techniques, identify which variables (if any) are important to include in your
# (b)  Mathematically write out a justifiable (perhaps after a transformation) MLR model that you will
# (c) *Explain* any assumptions you made in your model (you will justify these later in Section 3).
vs.res <- bestglm(farm, IC="AIC", method="exhaustive", TopModels=5)
vs.res$BestModel
#the multiplication symbol means single AND together variables of interactions
farm.pred.mlr <- lm(I(log(acrePrice))~improvements+
                      tillable+crpPct+productivity*NW+WC, farm)
# (a)  Using your fitted model, justify any assumptions you made in fitting the model using appropriate
# (b)  Do you think that your model fits the data well? Explain your reasoning using necessary model fi
# (c)  Do you think your model is valuable for predicting the price of a farm?  Explain your reasoning
farm.beta.0 <- round(as.numeric(coef(farm.pred.mlr)["(Intercept)"]), digits = 4)
farm.beta.1 <- round(as.numeric(coef(farm.pred.mlr)["improvements"]), digits = 4)
farm.beta.2 <- round(as.numeric(coef(farm.pred.mlr)["tillable"]), digits = 4)
farm.beta.3 <- round(as.numeric(coef(farm.pred.mlr)["crpPct"]), digits = 4)
farm.beta.4 <- round(as.numeric(coef(farm.pred.mlr)["productivity"]), digits = 4)
farm.beta.5 <- round(as.numeric(coef(farm.pred.mlr)["NWYes"]), digits = 4)
farm.beta.6 <- round(as.numeric(coef(farm.pred.mlr)["WCYes"]), digits = 4)
farm.beta.7 <- round(as.numeric(coef(farm.pred.mlr)["productivity:NWYes"]),
                     digits = 4)
farm.var <- round(sigma(farm.pred.mlr)^2, digits = 4)
avPlots(farm.pred.mlr, ask = FALSE)
fit.vs.resids.1 <- ggplot(farm, aes(x=farm.pred.mlr$fitted.values,
                                    y=farm.pred.mlr$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Assumption Check:',
          subtitle ='Linearity and Equal Variance') +
  geom_hline(yintercept = 0, col = "red", lwd = 0.5)

std.resids <- stdres(farm.pred.mlr)
hline <- round(mean(std.resids), digits = 4)
farm.freq <- ggplot() +
  geom_histogram(mapping=aes(x=std.resids)) +
  xlab('Standardized Residuals') +
  ylab('Frequency') +
  ggtitle('Assumption Check:', subtitle = 'Normality') +
  geom_vline(xintercept = mean(std.resids), col = "red", lwd = 1) +
  annotate("text", x = hline + 1.25, y = 17,
           label = paste("Mean =", hline), col = "red", size = 3.5)
suppressMessages(grid.arrange(fit.vs.resids.1, farm.freq, nrow=1))
```

```
ks.res <- round(ks.test(std.resids, "pnorm")$p.value, digits = 4)
jb.res <- round(jb.norm.test(std.resids)$p.value, digits = 4)
bp.res <- round(as.numeric(bptest(farm.pred.mlr)$p.value), digits = 4)
farm.stddev <- round(sigma(farm.pred.mlr), digits = 2)
farm.r2 <- 100 * round(summary(farm.pred.mlr)$adj.r.squared, digits = 2)
#cross validation
set.seed(52) #set seed for reproducibility
n.cv <- 100 #Number of CV studies we'll run
bias <- rep(NA, n.cv) #n.cv empty biases (one for each CV)
RPMSE <- rep(NA, n.cv) #n.cv empty RPMSE (one for each CV)
coverage <- rep(NA, n.cv) #n.cv empty coverage (one for each CV)
width <- rep(NA, n.cv) #n.cv empty width (one for each CV)
n.test <- 10 #How big my test set is
for(i in 1:n.cv){
  test.obs <- sample(1:nrow(farm), n.test)
  test.set <- farm[test.obs,]
  train.set <- farm[-test.obs,]

  train.lm <- lm(I(log(acrePrice))~improvements+tillable+crpPct+productivity*NW+WC,
                 data=train.set)
  test.preds <- exp(predict.lm(train.lm, newdata=test.set, interval="prediction"))

  bias[i] <- mean(test.preds[,1]- test.set$acrePrice)
  RPMSE[i] <- sqrt(mean((test.preds[,1] - test.set$acrePrice)^2))
  coverage[i] <- mean((test.preds[,2] < test.set$acrePrice) &
                        (test.preds[,3] > test.set$acrePrice))
  width[i] <- mean(test.preds[,3] - test.preds[,2])
}

mean.bias <- round(mean(bias), digits = 2)
mean.RPMSE <- round(mean(RPMSE), digits = 2)
mean.coverage <- round(mean(coverage), digits = 2)
mean.width <- round(mean(width), digits = 2)
CV.bias <- ggplot() +
  geom_histogram(mapping=aes(x=bias)) +
  xlab('Amount of Bias') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Bias for Cross Validation') +
  geom_vline(xintercept = mean.bias, col = "red", lwd = 1)

CV.RPMSE <- ggplot() +
  geom_histogram(mapping=aes(x=RPMSE)) +
  xlab('RPMSE') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of RPMSE for Cross Validation') +
  geom_vline(xintercept = mean.RPMSE, col = "red", lwd = 1)

CV.coverage <- ggplot() +
  geom_histogram(mapping=aes(x=coverage)) +
  xlab('Coverage') +
  ylab('Frequency') +
```

```
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Coverage for Cross Validation') +
  geom_vline(xintercept = mean.coverage, col = "red", lwd = 1)

CV.width <- ggplot() +
  geom_histogram(mapping=aes(x=width)) +
  xlab('Width') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Width for Cross Validation') +
  geom_vline(xintercept = mean.width, col = "red", lwd = 1)

suppressMessages(grid.arrange(CV.bias, CV.RPMSE, CV.coverage, CV.width, nrow=2))
# (a)  Based on your fitted model, what do you think are the effect(s) (if any) of the selected variabl
# (b)  Do you agree with the appraiser's prior intuition that the effect of productivity in the NW is d
# (c)  Regardless of your answer above, use your model to predict the farm price for a title transfer f
summary(farm.pred.mlr)
f.val <- as.numeric(summary(farm.pred.mlr)$fstatistic['value'])
CI <- confint(farm.pred.mlr, level = 0.95)
farm.mlr2 <- lm(I(log(acrePrice))~improvements+tillable+crpPct+
                  productivity+NW+WC, farm)
anova(farm.pred.mlr, farm.mlr2)
farm.predict <- data.frame(improvements=0, tillable=94, crpPct=0, productivity = 96,
                           NW = 'Yes', WC = 'No')
predict.range <- exp(predict.lm(farm.pred.mlr, newdata=farm.predict,
                            interval="prediction", level=0.95))
# (a)  Briefly summarize the main findings of your analysis in 1 paragraph and without using statistica
# (b)  Identify 1-2 "next steps" that the appraiser should consider to better understand and predict ac
#End of midterm 2's code
```