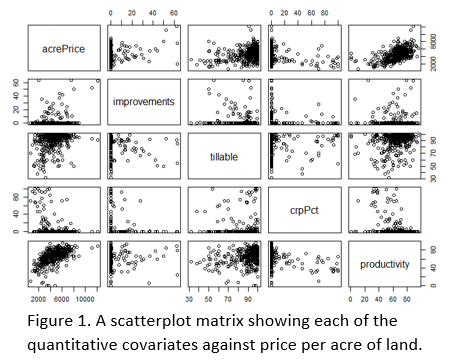
Farm Appraisals Analysis

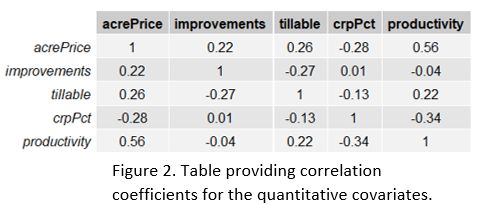
Carly Lundgreen

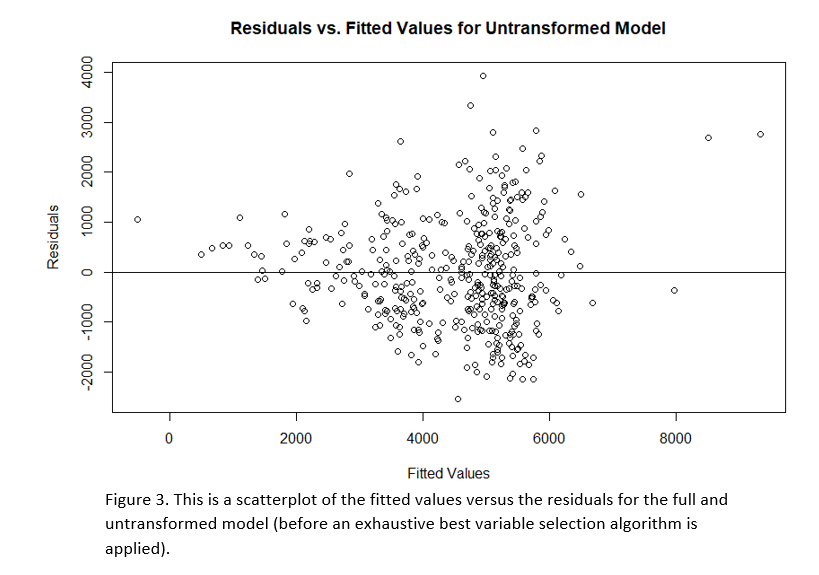
STAT 330 – Brigham Young University

11/20/18

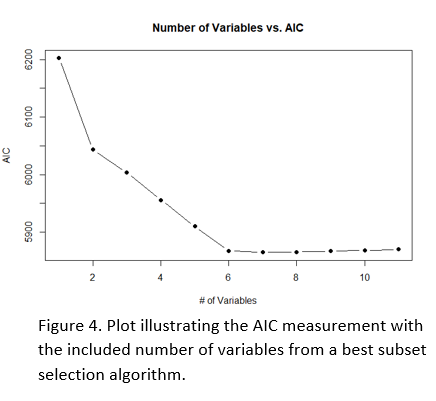
**Introduction**

In order to accurately appraise a farm, several factors must be considered. The goal of this analysis is to determine the relationship between several factors and the sale price (per acre) for a farm. These factors include the percentage of property value stemming from buildings, the percentage of arable land on the farm, the type of financing, percentage of land that is protected, productivity of the farm, and general location of the farm. In order to appraise a farm, or predict an accurate price per acre for the property, a statistical model may be used. This data, collected from multiple farms, will allow the appraiser to make an accurate appraisal that is fair for both the buyer and the seller.

The data set provides values for each of the previously mentioned variables (acre price, location, type of financing, etc.) for 420 farms. Some of the explanatory variables (covariates) are categorical in nature, such as the region where the farm is located, and some are quantitative, such as the percentage of tillable (arable) land. Figure 1, a scatterplot matrix, illustrates the relationship between acre price and each of the quantitative covariates provided in the data. Figure 2, provided below, provides the correlation coefficients between each of these quantitative variables. We can see that there are not very strong linear relationships between each of these variables.

An untransformed multiple linear regression (MLR) model does not appear to be suitable for these data. The condition of homoskedasticity (equal variance of responses about the regression line) seems to be violated, based on Figure 3. This plot shows a clear “funnel” pattern: smaller predicted acre prices appear to be less spread out around the line than larger prices. A logarithmic transformation on acre price is proposed before fitting a statistical model.

**Statistical Modeling**

Using a best subset selection algorithm, we will select the model that minimizes the AIC measurement on this data. Best subset is suitable due to the relatively small (i.e. less than about 30) number of variables. In this case, we only have 11 variables, so R has the capability to process each possible combination of variables to minimize AIC and provide the model with the optimal variables. AIC is an appropriate metric to utilize because of the appraiser’s interest in predicting a selling price per acre, or a prediction of selling price, for a certain farm. Using Figure 4, we can see that including 6 variables minimizes AIC and thus is an appropriate number of variables to include in the model. Because the appraiser believes that there is an interaction between productivity and price in the NW region, we will include an interaction term in our model accounting for that relationship.

The following equation illustrates a proposed model to use for analysis of this data set:

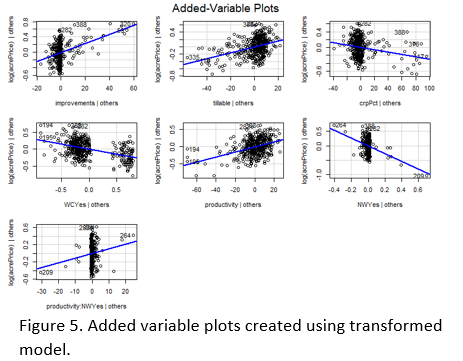
log(acrePrice) = +(improvements) + … (Productivity) + (NW) + (NW)\*(Productivity) +

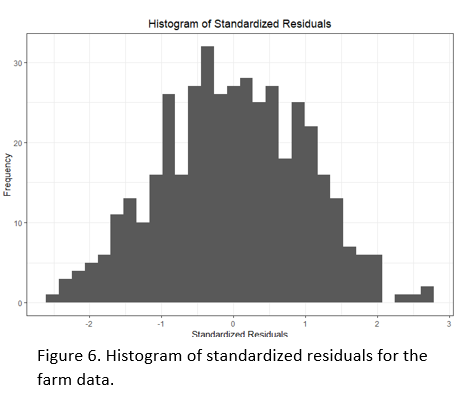
Where log(acrePrice) represents the logarithm of the price per acre for a farm. represents the logarithm of the selling price per acre we would expect for a farm located in neither the Northwest or West Central region with 0% of its value coming from buildings (“improvements”), 0% arable land (“tillable”), 0% protected land (“crpPct”), and zero productivityHolding all other variables constant, the average expected increase in the log of acre price for a 1% increase in property value due to buildings is represented by . We would expect a farm located in the Northwest region to have a higher selling price per acre than a farm located outside this region, holding all other variables constant. Finally, as a farm’s productivity score increases by 1, the logarithm of the average selling price per acre goes up by if the farm is not in the Northwest region and + if the farm is in the Northwest region.

Finally, represents the residual of each observed log(acrePrice), or the distance between the logarithm of the observed prices and the logarithm of the expected prices (the regression line). Using this model, we will be able to assess the relationship between price per acre for a farm and each of the variables included in the model, while also accounting for the appraiser’s belief that productivity has a different effect on farm prices in the NW region than it does in other regions.

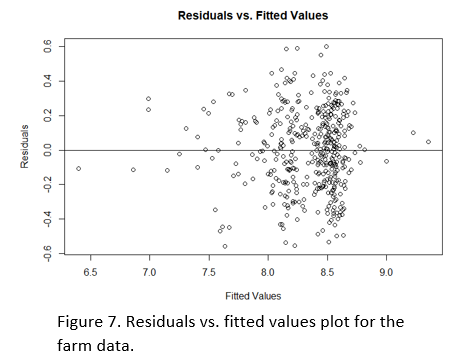
By fitting this model to the data, we assume that the transformed data follows a linear pattern. We also assume that each recorded—and transformed—farm price is independent of the next and that the residuals for the transformed prices are distributed normally with an average value of 0 and a variance represented by . This distribution can be rewritten as: ~ *N*(0,). That is, we assume that the transformed acre prices are distributed normally for each covariate value. Finally, we assume that these transformed prices vary equally about the regression line, or the predicted average logarithm of the prices per acre for individual farms.

**Model Verification**

After transforming the data, there should be a linear pattern visible in a scatterplot matrix. Looking at the added variable plots for our transformed model, seen in Figure 5, the patterns are much more linear—particularly between the “tillable” variable (percentage of arable land) and the log of price per acre.

We also assume independence between the log price per acre for each farm. This seems to be reasonable—there is no reason why the percentage of property value due to buildings for one farm should affect the property value for another farm, for example.

We also assume that the transformed prices are distributed normally about the regression line. In other words, we assume that the residuals are distributed normally. This is affirmed by Figure 6, which is a histogram of the standardized residuals for the transformed prices. No significant skewness is seen in the histogram, so we can assume This is affirmed by a Kolmogrov-Smirnov test of significance. A p-value of 0.8983 shows that there is *not* significant evidence that the transformed data are not normal, thus we fail to reject the null hypothesis that the data are distributed normally at a significance level of 0.05.

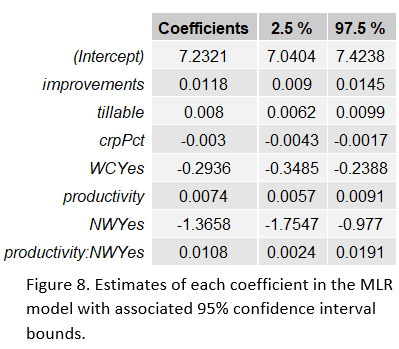
The final assumption we make when utilizing this MLR model is that the log prices vary normally about the regression line. We can affirm this with Figure 7, a plot of the fitted values against the residuals. There seems to be some clustering around log prices between 8 and 8.5, but we don’t see any extremely concerning “megaphone” patterns that would lead us to say that there is relatively equal variance about the regression line. There are a few potential outliers near log price values of 6.5 and about 9.5, but because Figure 6 looked sufficiently normal, we can proceed with the analysis. A BP test for homoscedasticity yielded a p-value of 0.1849, so we fail to reject the hypothesis that the transformed data is homoscedastic.

The R2 value for this model is calculated to be 0.6787, indicating that 67.87% of the variation in the log of selling price per acre for farms is explained by property value from buildings, percentage of arable land, percent of protected land, numeric productivity score, whether or not the farm was in the NW or WC region, and an interaction between productivity and NW region status. The average predictive bias across 250 cross-validation studies is calculated to be -105.43, which means our model underpredicts the average selling price per acre by about $100. Because the range of prices in the original data set is $11,559.27, this underprediction is not very concerning due to its small size with respect to the actual range of the data. The average root predictive mean square error is found to be 1016.32, indicating that predictions using this model will be off by about $1,000. Considering the range of selling prices, this is about 1/11 of the entire range, so this error is relatively small as well. Similarly, the average width of 95% prediction intervals is found to be 4214.14, which is about 1/3 of the range of the original prices, so this range is decently small in comparison to the range. The average coverage of these 95% prediction intervals was found to be 0.954, indicating that 95.4% of all prediction intervals will contain the actual selling price per acre for a farm. This model will predict prices accurately.

**Results**

The following equation is calculated after fitting the linear model to the log-transformed data:

log(acrePrice) = 7.232 + 0.0117(improvements) + … + 0.0074(productivity) – 1.366(NW) + 0.0108(NW)\*(Productivity)

This indicates that we expect the log-transformed selling price per acre to increase, on average, by 0.0117 for a 1% increase in property value due to buildings, holding all other variables constant. We also expect the log selling price per acre to increase by 0.0074, on average, when a farm’s productivity score increases by 1 if that farm is not in the Northwest region—holding all else constant. If it is in the northwest region, we expect the log price per acre to increase by 0.018 (.0074 + .0108) on average, again holding all other variables constant. 95% confidence intervals for each coefficient are displayed, along with the model’s estimated average value for each coefficient from the model, in Figure 8. We are 95% confident that the true average increase in log selling price per acre for a 1% increase in property value due to buildings, holding all else constant, is contained within the interval (.0062,.0099). Similarly, we are 95% confident that the true increase in log selling price for a farm in the Northwest region that sees a 1-unit increase in productivity, holding all else constant, is contained in the interval (.0024, .0191). Finally, we are 95% confident that the average increase in log selling price for a farm in the West Central Region, as opposed to a farm in another region, is contained in the interval (-.3485, -.2388). If a farm moved from another region to the West Central region, it is likely that their average log selling price per acre will decrease.

A scatterplot of log selling price with respect to productivity is shown in Figure 9. Based on this figure, we can see that farms in the NW region have lower log prices per acre than farms that are not in the NW region. Because of this, it seems that the appraiser’s intuition that the effect of productivity is different for farms in the Northwest than it is for farms in other areas is valid. To check, we look at the p-value for a t-test on the estimate for . This p-value is .0115, which is less than the significance level of .05, so we there is significant evidence of an interaction between productivity and the region in which the farm is located (NW or otherwise).

Using the model, we predict an average selling price of $4,317.44 per acre for a farm in the NW with the given metrics. A 95% prediction interval for this estimate is found to be ($2,286.13, $8,153.62). We are 95% confident that the true average selling price for a NW farm with 0% property value stemming from buildings, 94% tillable land, 0% protected land, and a productivity score of 96 is contained in this interval.

**Conclusions**

We found that there is a significant relationship between the logarithm of selling price per acre for a farm and the percent of the land’s property value that comes from buildings, the percent of land that is arable, the percent of land that is protected, the farm’s numeric productivity score (on a scale of 1:100), and whether the farm is located in the Northwest or West Central region. We also noted that productivity has a different effect on log price per acre in the Northwest region than it does in other regions. This was accounted for in our model. Each of these variables will affect the price of the farm. The prices of the farms from the original data set were transformed using a logarithm function to ensure that the model provided accurate predictions and that our inference was sound.

I would recommend that the appraiser explore further potential variables that may affect the price of the farm, such as the time of year at which the farm is appraised or perhaps the farm’s output from the previous year. In order to better understand and predict acre price, the appraiser could also gather data from more farms than the original data set provided.

Appendix: R Code

##################  
library(ggplot2)  
library(GGally)  
library(gridExtra)

library(grid)  
library(car)

library(MASS)

library(bestglm)

library(lmtest)

##################

farms <- read.table(file="C:/Users/Carly/Documents/Semester 6/Stat 330/farms3.txt",sep="",header=TRUE)  
farms <- farms[,c(2:ncol(farms),1)]  
  
#explore data with scatterplot and get correlations  
str(farms)

pairs(farms[,c("acrePrice","improvements","tillable","crpPct","productivity")])

corrs <- cor(farms[,c("acrePrice","improvements","tillable","crpPct","productivity")])  
corrs <- round(corrs,2)  
table <- tableGrob(corrs)  
grid.newpage()  
grid.draw(table)

#fit an untransformed model and assess it  
full.lm <- lm(acrePrice~., data=farms)  
avPlots(full.lm)

resids.full <- full.lm$residuals  
fittedvals.full <- full.lm$fitted.values  
plot(fittedvals.full,resids.full,main="Residuals vs. Fitted Values for Untransformed Model",xlab="Fitted Values",ylab="Residuals")  
abline(h=0)

#find best fitting model   
vs.res <- bestglm(farms,IC="AIC",method="exhaustive")

## binary categorical variables converted to 0-1 so 'leaps' could be used.

plot(vs.res$Subsets$AIC,type="b",pch=19,xlab="# of Variables", ylab="AIC",main="Number of Variables vs. AIC")

vs.res$BestModel

##   
## Call:  
## lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),   
## drop = FALSE], y = y))  
##   
## Coefficients:  
## (Intercept) improvements tillable crpPct productivity   
## -112.169 57.996 29.762 -6.474 34.350   
## NWYes WCYes   
## -2403.401 -1196.939

#fit new model with best variables   
new.lm <- lm(log(acrePrice)~improvements + tillable + crpPct + WC + productivity\*NW, data=farms)  
  
###########look at assumptions###############  
#linearity  
avPlots(new.lm)

#Histogram of residuals  
stdresids <- stdres(new.lm)  
ggplot(farms, aes(x=stdresids))+geom\_histogram()+theme\_bw()+  
 labs(x="Standardized Residuals",y="Frequency",title="Histogram of Standardized Residuals")+  
 theme(plot.title = element\_text(hjust = 0.5))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

#KS test for normality   
ks.test(stdresids,"pnorm")

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: stdresids  
## D = 0.027944, p-value = 0.8983  
## alternative hypothesis: two-sided

#Get residuals vs. fitted values plot   
resids <- new.lm$residuals  
fitted.vals <- new.lm$fitted.values  
  
plot(fitted.vals,resids,main="Residuals vs. Fitted Values",xlab="Fitted Values",ylab="Residuals")  
abline(h=0)

#BP test for equal variance  
bptest(new.lm)

##   
## studentized Breusch-Pagan test  
##   
## data: new.lm  
## BP = 10.065, df = 7, p-value = 0.1849

################################################  
  
##Assess fit and predictive accuracy of model   
summary(new.lm)$r.squared

## [1] 0.6787371

n.cv <- 250   
bias <- rep(NA,n.cv)  
rpmse <- rep(NA,n.cv)  
cvg <- rep(NA,n.cv)  
width <- rep(NA,n.cv)  
n.test <- 10  
  
for(cv in 1:n.cv){  
 ##split into test and training sets   
 test.obs <- sample(1:nrow(farms), n.test)  
 test.set <- farms[test.obs,]  
 train.set <- farms[-test.obs,]  
   
 ##fit a LM using training data only  
 train.lm <- lm(log(acrePrice)~improvements + tillable + crpPct + WC + productivity\*NW, data=train.set)  
   
 ##Get predictions and prediction intervals   
 pred <- predict.lm(train.lm,newdata=test.set,interval="prediction")  
 ##untransform predictions  
 realpreds <- exp(pred)  
   
 ##Calculate results   
 bias[cv] <- mean(realpreds[,'fit'] - test.set[,'acrePrice'])  
 rpmse[cv] <- sqrt(mean((realpreds[,'fit'] - test.set[,'acrePrice'])^2))  
 cvg[cv] <- mean(realpreds[,'lwr'] < test.set[,'acrePrice'] & realpreds[,'upr'] > test.set[,'acrePrice'])  
 width[cv] <- mean(realpreds[,'upr'] - realpreds[,'lwr'])  
   
}  
mean(bias)

## [1] -78.23246

mean(cvg)

## [1] 0.9596

mean(rpmse)

## [1] 997.5089

mean(width)

## [1] 4129.802

max(farms$acrePrice) - min(farms$acrePrice)

## [1] 11559.27

############Results##################  
summary(new.lm)

##   
## Call:  
## lm(formula = log(acrePrice) ~ improvements + tillable + crpPct +   
## WC + productivity \* NW, data = farms)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.55963 -0.15819 0.00263 0.16786 0.59967   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2320609 0.0975173 74.162 < 2e-16 \*\*\*  
## improvements 0.0117561 0.0013942 8.432 5.77e-16 \*\*\*  
## tillable 0.0080477 0.0009594 8.388 7.95e-16 \*\*\*  
## crpPct -0.0029712 0.0006621 -4.487 9.37e-06 \*\*\*  
## WCYes -0.2936437 0.0278932 -10.527 < 2e-16 \*\*\*  
## productivity 0.0074247 0.0008705 8.529 2.84e-16 \*\*\*  
## NWYes -1.3658287 0.1978117 -6.905 1.91e-11 \*\*\*  
## productivity:NWYes 0.0107805 0.0042492 2.537 0.0115 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2304 on 412 degrees of freedom  
## Multiple R-squared: 0.6787, Adjusted R-squared: 0.6733   
## F-statistic: 124.3 on 7 and 412 DF, p-value: < 2.2e-16

#Get confidence intervals for estimates   
Coefficients <- new.lm$coefficients  
coefs <- data.frame(Coefficients)  
coefs2 <- round(coefs, 4)  
int <- confint(new.lm)  
int <- round(int,4)  
  
estimates <- tableGrob(coefs2)  
intervals <- tableGrob(int,rows=NULL)  
grid.newpage()  
grid.arrange(gtable\_combine(estimates,intervals),ncol=1)

#plot scatterplot to explore interaction   
ggplot(farms,aes(y=log(acrePrice),x=productivity,color=NW))+geom\_point()+  
 labs(x="Productivity",y="Log Price per Acre",title="Price vs. Productivity")+  
 theme(plot.title = element\_text(hjust = 0.5))

summary(new.lm)

##   
## Call:  
## lm(formula = log(acrePrice) ~ improvements + tillable + crpPct +   
## WC + productivity \* NW, data = farms)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.55963 -0.15819 0.00263 0.16786 0.59967   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2320609 0.0975173 74.162 < 2e-16 \*\*\*  
## improvements 0.0117561 0.0013942 8.432 5.77e-16 \*\*\*  
## tillable 0.0080477 0.0009594 8.388 7.95e-16 \*\*\*  
## crpPct -0.0029712 0.0006621 -4.487 9.37e-06 \*\*\*  
## WCYes -0.2936437 0.0278932 -10.527 < 2e-16 \*\*\*  
## productivity 0.0074247 0.0008705 8.529 2.84e-16 \*\*\*  
## NWYes -1.3658287 0.1978117 -6.905 1.91e-11 \*\*\*  
## productivity:NWYes 0.0107805 0.0042492 2.537 0.0115 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2304 on 412 degrees of freedom  
## Multiple R-squared: 0.6787, Adjusted R-squared: 0.6733   
## F-statistic: 124.3 on 7 and 412 DF, p-value: < 2.2e-16

#Use model to predict Price  
dframe <- data.frame(improvements=0, tillable=94, crpPct=0,productivity=96, WC = "No",NW = "Yes")  
prediction <- predict.lm(new.lm,newdata=dframe,interval="prediction",level=.95)  
#un-transform prediction  
real.pred <- exp(prediction)  
real.pred

## fit lwr upr  
## 1 4317.436 2286.131 8153.624