Tyler Cann

Stat 330 – Heaton

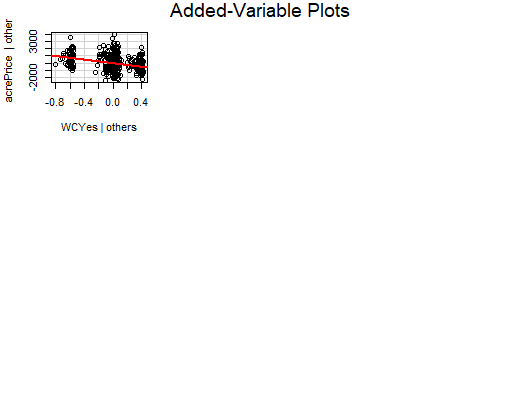
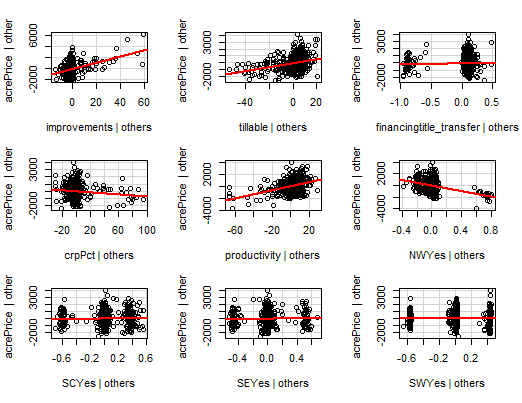
**Farm Appraisal**

**Section 1: Introduction and Problem Background**

A Farm appraiser wishes to be able to predict the value of farms prior to sale. To make sure a fair price is reached between the buyer and seller, the appraiser will have to determine which factors affect the price of a farm and use those factors to appraise the value of the farm. As a statistician, I’ll be using a fitted multiple linear regression model to help this appraiser determine which farm factors affect price the most, and determine what a fair price for the farm will be.

We are going to use the Farms.txt data we collected from previously observed farms to fulfill our goals stated above. We want to look at the relationship between the price per acre of these farms and the other farm attribute variables (improvements, tillable, financing, crppct, productivity, and location [NW,SC,SE,SW,WC,C]).

First we want to analyze the data using some simple graphics and summaries.

We want to check how linear the data is, so we’ll run some avPlots of the data:

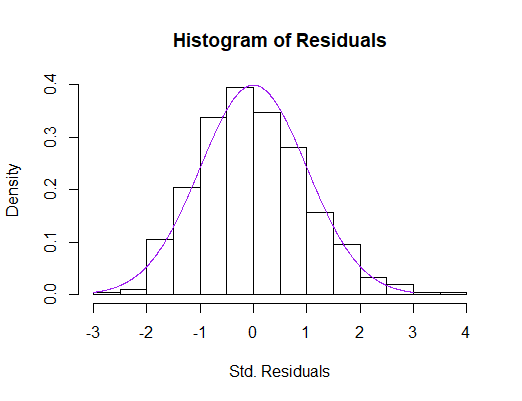
The above plots are a summary of the data we were given. We can see that there’s a linear relationship between the price per acre and the various farm attributes, with varying levels of correlation. All the plots have straight lines, leading us to understand that there are linear relationships between them. We may have to transform the data to observe a stronger relationship though.

The correlation coefficient, a numerical representation of the relationship, is only 0.21 (a 0 means there’s no relationship, 0.5 means there’s a moderate relationship, and 1 means there’s direct causation) between price per acre and improvements. The correlation between price per acre and other variables were: .26 with arable land, -.28 with protected land, and .56 with productive land. Given that these values (minus that of productive land) are all close to 0, we know it’s a weak relationship. Productive land’s is close to .5, meaning it has a moderate relationship. The other variables were categorical, so a correlation coefficient couldn’t be collected. Judging by the plots though, it appears there’s a moderate negative relationship with the farm being in the west central region and the north west region, with the rest appearing like there’s not much of a relationship.

Next, we want to check whether the data is independent.

When thinking about the overreaching problem, we need to think about whether each response variable (Farm Price Per Acre) is affected by the other. The price per acre of one farm won’t have an effect on another’s price per acre. Because of this, independence is met.

Third, we want to check normality, so we’ll run a histogram of the residuals of the plot:



The histogram above shows that the data is normal. It follows the bell curve shape well and doesn’t have any issues with tails being particularly too strong on either side of the curve.

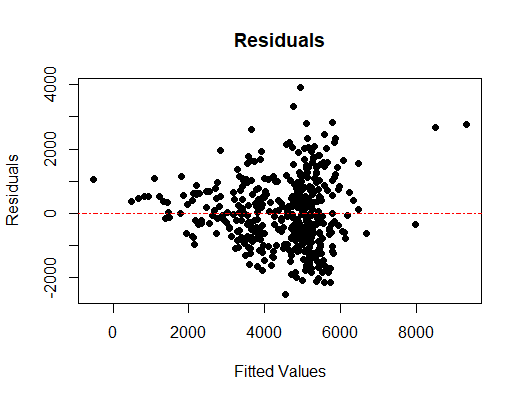
If we perform the following Kolmogorov-Smirnov hypothesis test,

H0: data come from a normal

Ha: data do not come from a normal

Our resulting p-value is .4317. Thus, we would fail to reject the null hypothesis and conclude we have insufficient evidence to say the data do not come from a normal distribution. .4317 is above .05, hence our results.

Next, we want to check if the data has equal variance, so we’ll plot the residuals vs fitted values:



This plot shows that the data isn’t scattered equally. Most of it falls above the line we drew through the middle, along with the data being extremely packed around the middle. Because of this, there isn’t equal variance. If there *was* equal variance, then there would be an even distribution of circles on both sides of the line.

We will do a Breusch-Pagan test to verify. When testing,

H0: Data have equal variance

Ha: Data do not have equal variance

We find our p-value to be .00000387, leading us to reject our null hypothesis and conclude we have sufficient evidence to say our data does not have equal variance. .00000387 is below .05, hence our results.

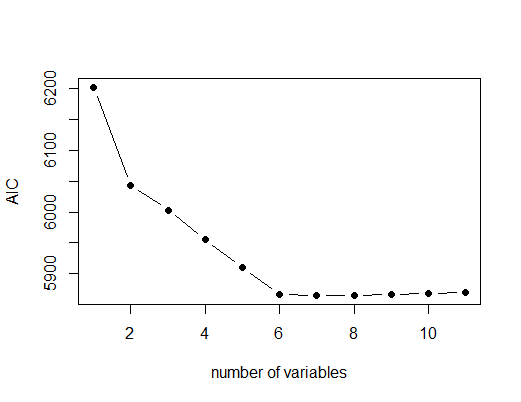
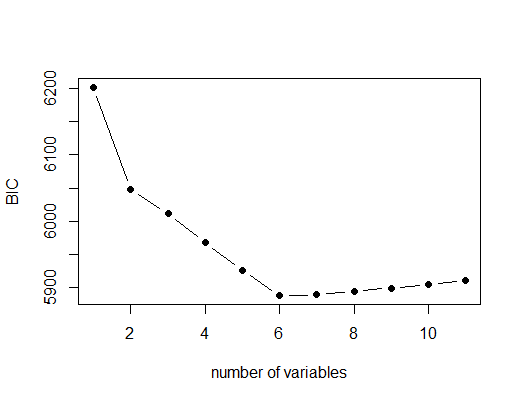
With the data in its current state, a multiple linear regression (MLR) model would NOT be appropriate. The data is linear, independent, and normal. However, the data isn’t equally variant. In order to run linear regression, we’re going to have to transform this data. A suitable transformation would be a logarithmic transformation. That means that we’re going to use a natural log function like you’d find in a calculator and run the data through it to see how the data looks after.

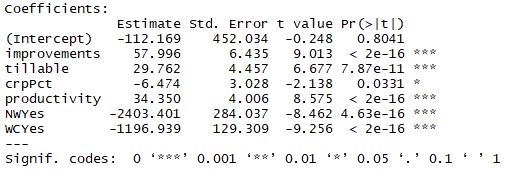
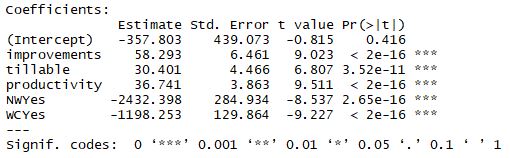
**Section 2: Statistical Modeling**

All of the variables had a variance inflation factor below 10, so it was hard to tell initially which variables could be sacrificed to improve our model.

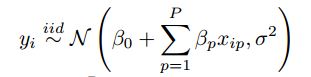
The variable selection procedure that I chose was exhaustive selection. This was because I received instructions to use the exhaustive method in previous training unless the computer program wouldn’t run. However, if one had to choose between forward and backward variable selection in this situation, I would’ve chosen backward selection because our dataset doesn’t have very many variables, so forward selection wouldn’t be necessary as it is best utilized with large datasets.

The model comparison criterion I chose was Bayesian Information Criterion. This was because the BIC only keeps the most important ones and I only want to know about which variables are the most important in finding out what affects farm prices the most. The BIC gave me 5 variables to use, while the AIC gave me 6, so I went with the BIC. The ones that only slightly affect the price per acre of a farm aren’t important to me, so I only want to see the most important variables.





As a result of selecting BIC/exhaustive procedure, we got the following mathematical MLR model:



Interpretation of variables and parameters:

**Log(yi)**: Farm’s Price per Acre

**xip**: Farm Attribute Score on most important variables (improvements, tillable, productivity, NWYes, WCYes)

**β0**: (This is the intercept) When each of the x’s (improvements, tillable, productivity, NWYes, WCYes) are at their average, we would expect a farm’s price per acre (y) to be β0.

**βp**: (This is one of the slopes) Holding all other x’s constant, as xip(improvements, tillable, productivity, NWYes, WCYes) goes up by one, we expect that a farm’s price per acre (y) goes up by an average of βp.

**σ**: For any xip (Farm Attribute Score), 99.7% of the Farm Price Per Acres will be within 3σ of the model.

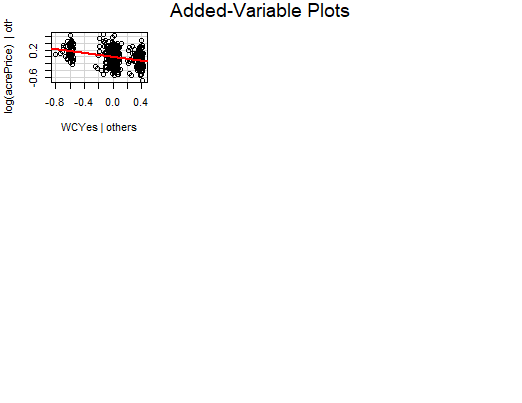
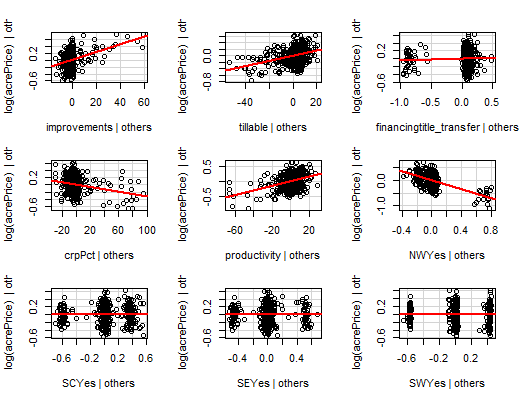
Once the model is fitted to the data, we will have estimates for β0,βp, and σ. We will then be able to use mean Farm Attribute Scores to predict a farm’s price per acre.

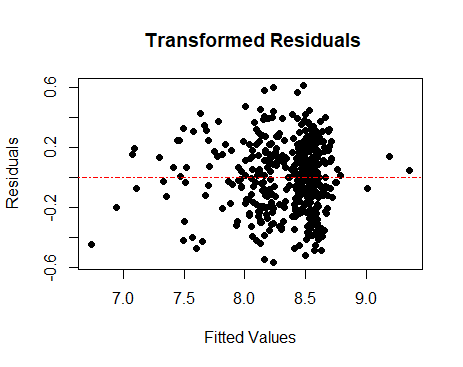
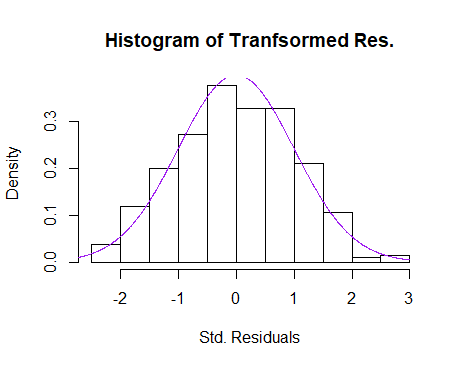
The assumption that we make when using this model is that the data follows a linear pattern, the Yi’s are independent of one another, the residuals follow a normal distribution and there is equal variance about the line. Lastly, in order for this model to work, the Yi’s must also be transformed logarithmically.

**Section 3: Model Justification and Verification**

As mentioned previously, we fitted our assumed model using a logarithmic transformation of the price per acre variable.

Transformed avPlots:





As seen in the two plots and the histogram above, our assumptions are justified.

Our linearity assumption is still met with the transformed avPlot like we mentioned previously, and The data was independent, also mentioned previously. Nothing changed in that regard.

Normality is still met. It follows a bell curve, like before, and has no outliers. Also, all of the data lies within 3 standard deviations. After a KS Test, we get a p-value of .8085. This is still above .05 like the last test, so we know that normality is met.

It looks like the data is more equally variant this time. Both sides of plot look to have about the same amount of data. There appear to be some outliers near the right side of the plot, and the data tends to bunch up around the 8.5 part as well., After another BP Test, the p-value increased dramatically to a value of .04 (compared to .00000387 earlier). While this is still below .05, this is likely due to the outliers, so we can be confident in the equal variance assumption.

We also calculated an **R2** of about 0.67. This means that about 67% of the variation in a log(farm’s price per acre) can be explained by the percentage of property value due to buildings, percentage of arable land, productive land score, location in the northwest region, and location in the west central region. Because this number is above .50, we learn that there is a moderate relationship between the farm’s price per acre and the aforementioned farm attributes.

We ran a cross validation procedure to help us know the predictability of our model. This means we took a sample of data points from our original data set and used the remaining data as our “training data set.” We fit a model to our training data and then looked to see how far off our predictions would be from our sample of data points we took. We repeated this process 100 times to get a good idea of how well our model predict on average. This gave us an average **bias** of -$127.86 price per acre. We can interpret this as meaning that our model tends to underpredict by this value on average. Our **root predictive mean square error(RPMSE),** how much our model is off on average, was $1074. So, on average, our model is off by this number. We’re dealing with prices per acre going into the thousands, so our even though the $1074 value is large, we can still predict with it and not be too far off.

The cross validation procedure gave us a **coverage** (Which looks at the percentage of intervals that contain the actual price per acre.) of .9583. That means 95.83% of our intervals were correct. It also gave us a mean **width** of 41.67. This means our prediction intervals are 41.67% wide on average.

The model fits the data well, as evidenced in the previous paragraphs of section 3, along with the graphics of the transformed data. Because the data we’re analyzing appears to have a moderate correlation, we should be able to make predictions.

Our model does a moderately good job at predicting accurately. According to our model, our R2 is 0.67, so about 67% of the variation in a log(farm’s price per acre) can be explained by the percentage of property value due to buildings, percentage of arable land, productive land score, location in the northwest region, and location in the west central region. 67% will tell you a decent amount of what you want to know. Unfortunately, Our RMPSE was fairly large too. Most of our data lies between $3000 and $6000 in the price per acre range, so being off by $1074 means that you’re potentially off by quite a bit. Our high coverage makes me confident our model does a well-enough job though.

**Section 4: Results**

|  |  |  |
| --- | --- | --- |
| Variable | Coefficient | 95% Confidence Interval |
| β0 (intercept) | -357.8 | (-1220.89, 505.29) |
| β1 improvements | 58.29 | (45.59, 70.99) |
| β2 tillable | 30.4 | (21.62, 39.18) |
| β3 productivity | 36.74 | (29.15, 44.33) |
| β4 NW (yes) | -2432.4 | (-2992.5, -1872.3) |
| β5 WC (yes) | -1198.25 | (-1453.53, -942.98) |

According to the model, our average farm price per acre should be -$357.8 ignoring all other variables. This means that the land lose you -$357.8 per acre without the other variables.

Regarding the “improvements” variable (and holding all other variables constant): As the percentage of property value due to buildings goes up by one, we expect that a farm’s price per acre will go up by an average of $58.29.

Regarding the “tillable” variable (and holding all other variables constant): As the percentage of arable land goes up by one, we expect that a farm’s price per acre will go up by an average of $30.4.

Regarding the “productivity” variable (and holding all other variables constant): As productive land score goes up by one, we expect that a farm’s price per acre will go up by an average of $36.74.

Regarding the “NW” variable (and holding all other variables constant): If the farm is located in the Northwest region, we expect that a farm’s price per acre will go down by an average of $2432.4.

Regarding the “WC” variable (and holding all other variables constant): If the farm is located in the West Central region, we expect that a farm’s price per acre will go down by an average of $1198.25.

I do not agree with the appraiser’s prior intuition that the effect of productivity in the NW is different than in other areas. I disagree because a farm located in the NW will always have a significantly lower price per acre no matter how productive the land is. It could just be that farms in the NW have a lower value because of a historical bias against farms of that region or other factors we aren’t taking into account in our dataset. We don’t have enough evidence to say that the productivity in the NW is different.

According to our model, the farm price for a title transfer farm in the NW with improvements=0, tillable=94, crpPct=0, and productivity=96 will result in a farm with a **price per acre of $3594.64** with a 95% prediction interval of (1384.49, 5804.80). This means we can be 95% confident that a farm with these attributes will have a price per acre between $1384 and $5804.

**Section 5: Conclusions**

We found that the factors that affect the price per acre of a farm the most are the percentage of property value due to buildings, percentage of arable land, productive land score, location in the northwest region, and location in the west central region. The relationship between these factors was moderately strong and required a logarithmic function to fully analyze. By finding this relationship and exploring the data, we were able to create a mathematical model that would let us predict how much a farm’s price per acre would be. The model’s predictions won’t always be accurate, but will give us a ballpark idea of what to expect.

1) Farm appraisers should gather more data that puts the farm location in a “group” factor with different location responses rather than different locations with a “yes” or “no” response. This would help us find the interaction between these values rather than taking them into account individually.

2) Until then, the farm appraiser should use this data to help educate farmers on what they need to do increase the value of their land.

**R CODE**:

setwd("C:/Users/thety/Desktop/330/Exam 2")

farm <- read.table("Farms.txt", header = TRUE)

#b

farm2 <- farm[,-c(7:11)]

farm2 <- farm2[,-4]

cor(farm2)

library(car)

farmlm <- lm(acrePrice~.,data=farm)

avPlots(farmlm)

vif(farmlm)

coef(farmlm)

library(lmtest)

plot(farmlm$fitted.values,farmlm$residuals,pch=19, ylab = "Residuals", xlab = "Fitted Values",main="Residuals")

abline(0,0,lty=4,col="red")

bptest(farmlm)

library(MASS)

res <- stdres(farmlm)

hist(res, freq=FALSE, main="Histogram of Residuals", xlab="Std. Residuals")

resso <- seq(-3,3,length=1000)

lines(resso,dnorm(resso),col="purple")

ks.test(res,"pnorm")

summary(farmlm)

#2

#a

#move your response variable to be the last column in the data frame before running it

farm3 <- cbind(farm[,-1],farm[,1])

names(farm3)[11]<-"acrePrice"

library(bestglm)

vs.res <- bestglm(farm3, IC="BIC", method = "exhaustive")

vs.res$BestModels

vs.res$Subsets

plot(vs.res$Subsets$BIC,type="b",pch=19,xlab="number of variables",ylab="BIC")

#they're numbering from 0 up, our plot goes from 1 up (so you just need to know it goes 1 further than needed)

best.lm <- vs.res$BestModel

summary(best.lm)

vs.res2 <- bestglm(farm3, IC="AIC", method = "exhaustive") #Not a lot of difference between 1st and 2nd best

vs.res2$BestModels #true means its included, false means it isnt

vs.res2$Subsets #minimize AIC maximize Rsquared

plot(vs.res2$Subsets$AIC,type="b",pch=19,xlab="number of variables",ylab="AIC")

#the graph gives us a 10, but the original gave us a 9

best.lm2 <- vs.res2$BestModel

summary(best.lm2)

#3

#a b and c

#### AFTER A LOG TRANSFORMATION ####

lm2 <- lm(log(acrePrice)~.,data=farm)

avPlots(lm2)

vif(lm2)

coef(lm2)

plot(lm2$fitted.values,lm2$residuals,pch=19, ylab = "Residuals", xlab = "Fitted Values",main=" Transformed Residuals")

abline(0,0,lty=4,col="red")

bptest(lm2) #still below .05 but probs due to outliers, we're okay

res <- stdres(lm2)

hist(res, freq=FALSE, main="Histogram of Tranfsormed Res.", xlab="Std. Residuals")

resso <- seq(-3,3,length=1000)

lines(resso,dnorm(resso),col="purple")

ks.test(res,"pnorm")

summary(lm2)

n.cv <- 100

bias <- rep(NA,n.cv)

rpmse <- rep(NA,n.cv)

cvg <- rep(NA,n.cv)

width <- rep(NA,n.cv)

dv <- rep(NA,n.cv)

for(i in 1:n.cv){

## Step 1 - split into test and training sets

obs.test <- sample(1:nrow(farm),round(.1\*nrow(farm)))

head(obs.test)

test.data <- farm[obs.test,] #putting a blank after a comma will give you all of teh columns

head(test.data)

training.data <- farm[-obs.test,]

training.data

## Step 2 - fit model to training data

my.model <- lm(log(acrePrice)~.,data=training.data)

## Step 3 - predict for test data

test.predict <- exp(predict.lm(my.model,newdata=test.data, interval="prediction", level = .95))

test.predict

## Step 4 - calculate bias and RMPSE

bias[i] <- mean(test.predict[,1]-test.data$acrePrice)

rpmse[i] <- sqrt(mean((test.predict[,1]-test.data$acrePrice)^2))

cvg[i] <- mean(test.predict[,2]<test.data$acrePrice & test.predict[,3]>test.data$acrePrice)

mcvg <- mean(cvg)

width[i] <- mean(test.predict[,3]-test.predict[,2])

mw <- mean(width)

}

mean(bias)

mean(rpmse)

mcvg

mw

mean(farm$acrePrice)

hist(farm$acrePrice)

#using the BIC model...

#4

#a

farm4 <- lm(acrePrice~improvements+tillable+productivity+NW+WC,data=farm3)

coef(farm4)

confint(farm4)

#c

predicto <- data.frame(improvements=0,tillable=94,productivity=96,NW="Yes",WC="No")

predict.lm(farm4,newdata=predicto,interval="prediction")