# MSDS 6372 Project 1

2/16/2020

**Jonathon Roach**
**Sanjay Pillay**
**Fred Poon**

Project Professor: Dr Jacob Turner

Citations

Spline Model

Polynomial Interpretation

Dataset

## INTRODUCTION

Based on  the data set published by The National Agricultural Statistics Service (NASS) we will gain insight using a few regression models on how agricultural land values are impacted in different regions across the United States and correlate any factors driving the land values. In addition will conduct a two way analysis to investigate changes in values in certain regions and how prices vary across types of land (Overall farmland, pasture, and cropland).

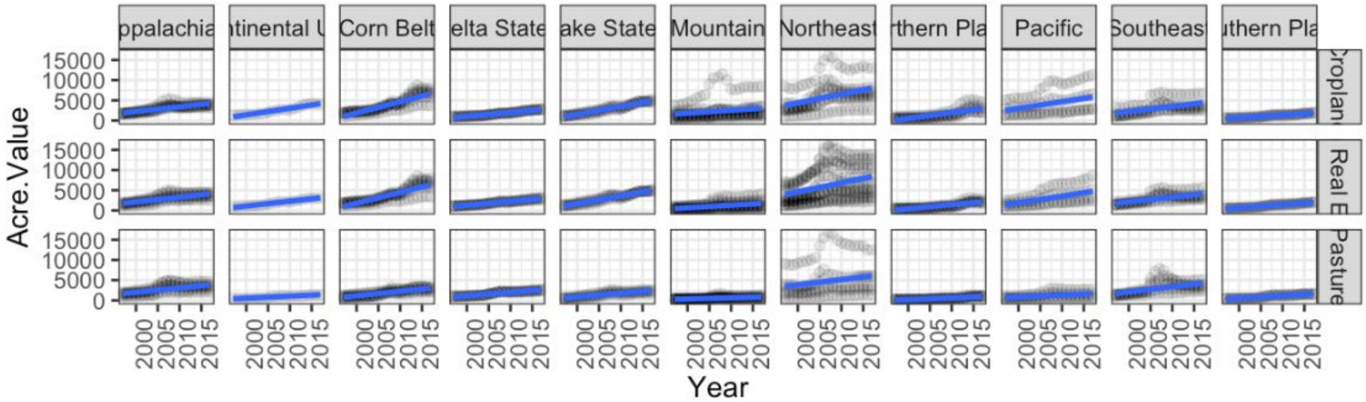
## DATA DESCRIPTION

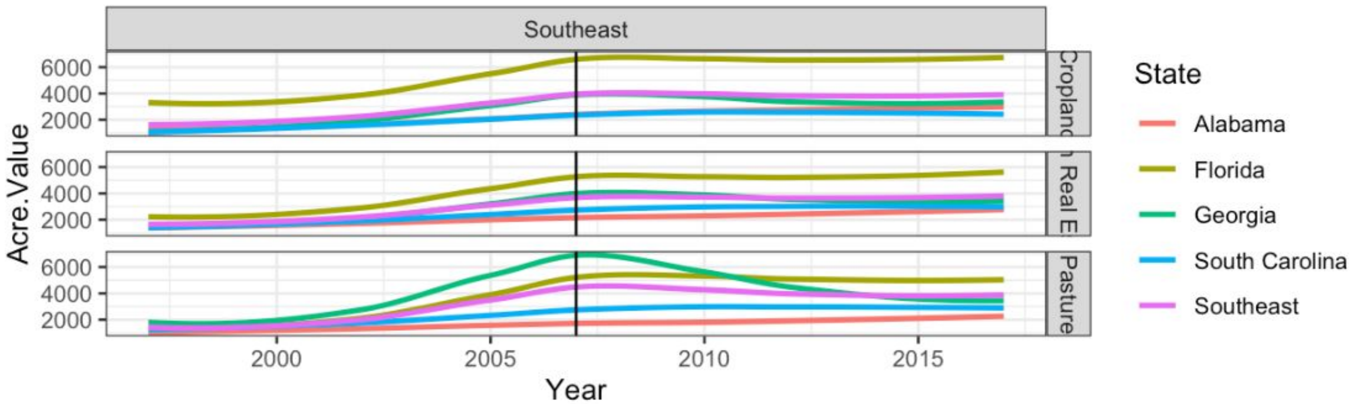
We are using a cleaned up version of this dataset hosted on kaggle which was extracted from NASS. The dataset has about 3500 observations.

| **Variable** | **Description** |
|---|---|
| State | 49 states in the US where agricultural land values were recorded. |
| LandCategory | Three factors of land category (Pasture, Farm Real Estate, Cropland). |
| Region | Ten economic categorical regions that US farmland is identified by (Mountain, Pacific, Cornbelt etc..). |
| RegionState | Variable identifying the observation as either State or Region (More info in EDA section). |
| Year | The year of the observation. |
| Acre.Value | Estimated land value in US Dollars. |

## EDA

Upon initial inspection of the data, we immediately noticed idiosyncrasies in the way the data were organized. Namely, the “State” data element did not exclusively contain US states. For most states, there were exactly 63 observations of acreage prices for crop land, farm real estate, and pastures over a 20-year period (it’s worth noting here that a small number of states in the Northeast had only 21 or 42 observations). In addition, each region had 63 observations and was also captured in the “State” data element. Because most of the variation in acreage pricing is captured by the region, our regression analysis focuses on predicting acreage by region (summary statistics), with some exceptions (see figures below).




For regions in which there is little variability in the acreage pricing over the 20-year period, there was no need to include information at the state level. For regions such as the **Northeast, Southeast, and the Pacific**, there were significant differences between acreage over time for each state. As a result, for modeling purposes, we created a new variable called "Region_New" that reported the state for 7 of the regions, but only reported the region for the remaining 4.

## OBJECTIVE 1

### Problem Statement

We will investigate the best possible model to predict land values in the region of **Northeast, Southeast, and the Pacific** using various regression models.

### Model Selection

We will also analyse the optimal number of features to fit using forward/backward selection model/ With the dataset having a limited set of features we can see both forward and backward and lasso indicate using all three fields.

| Model | ASE Three variables (LandCat, Region, Year) |
|---|---|
| Forwards | 0.8286628 0.7423832 0.6714105 |
| Backward | 0.8286628 0.7423832 0.6714105 |

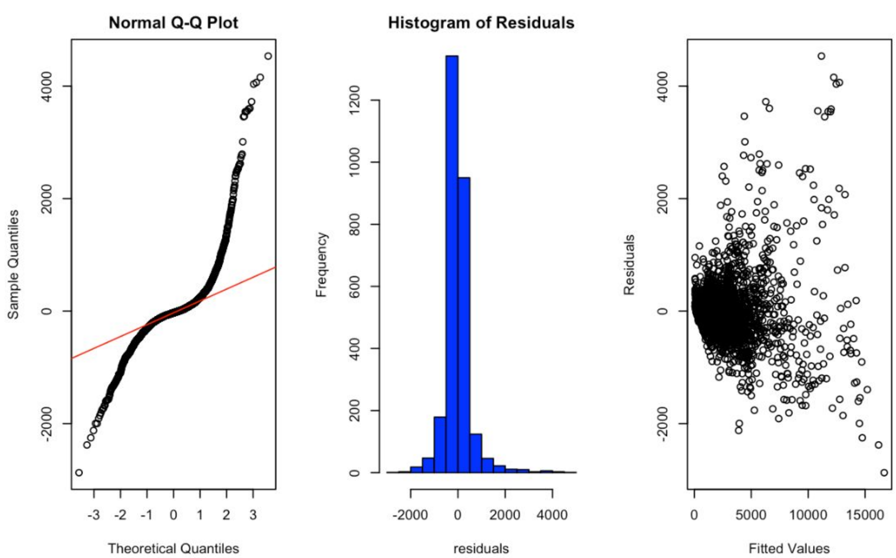
**Regression Models**

Before beginning the modeling process, training and test data sets were created with an 80/20 split for cross-validation. Given the figure  in the EDA above, there are clearly differences in variation of acreage prices over time between land categories and regions. Therefore, the following model seemed appropriate:

Model 1 (Non Transformed)

**Acre.Value ~ Year*LandCategory*Region_new**

With an adjusted R-squared value of 0.9452 and an RMSE of 688.595 on the test set, this model's performance is not bad, but it is clear that we have violated some regression assumptions in the plot below.




The QQ-plot and the scatter plot indicate that log transformations may be needed. The following are the results of a log-linear model:

Model 2 (Log Linear)

**log(Acre.Value) ~ Year*LandCategory*Region_new**

This model yields and adjusted R-squared of 0.9633 and an RMSE of 759.2183 on the test set, and our residual plots, while still not perfect, look better than in the previous model.

The QQ-plot and histogram indicate that the transformation succeeded in making the data almost normal, but the scatter plot of fitted values vs. residuals shows that this model still violates the assumption of uniform variance in the residuals. The next logical step is to also take the log of the response variable.

### Model 3 (Log-Log)

**log(Acre.Value) ~ log(Year)*LandCategory*Region_new**

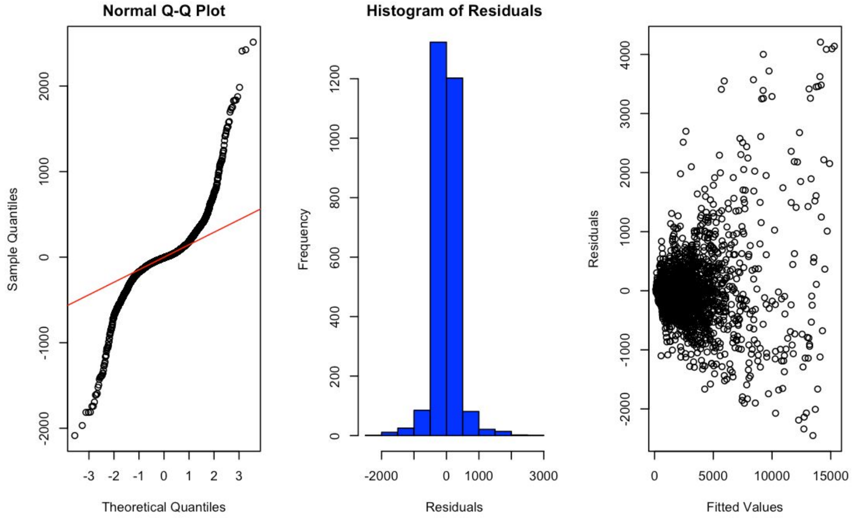
This model yields an adjusted R-squared of 0.9633 and an RMSE of 759.2183 on the test set. In both performance and adherence to model assumptions, this model is nearly identical to the previous one. See residual diagnostics

Due to the fact that log transformations couldn't help us satisfy model assumptions, it is reasonable to hypothesize that acreage prices over time are non-linear. We will therefore implement a model with a polynomial term, as well as a model fit with splines. The polynomial model is defined as follows:

### Model 4 (Linear Model With Splines)

**Acre.Value ~ (Year + X)*LandCategory*Region_new**

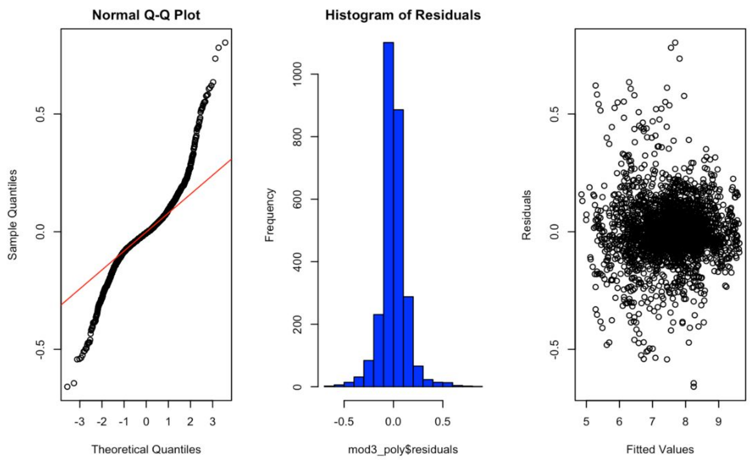
This model is fit using splines. The "X" variable is a product of two components: a binary (1/0) indicator that shows whether the year is past a change point (2007) and the number of years past the change point. The accuracy of this model is extremely high with an R-squared of 0.978 and an RMSE of 370.86. Despite this advantage in accuracy, the model does not adhere to the regression assumptions of normally distributed residuals with uniform variance.



### Model 5

**log(Acre.Value) ~ poly(Year,2)*LandCategory*Region_new**

This model is arguably the best we've seen so far in terms of adherence to regression assumptions. With an adjusted R-squared of 0.9747 and an RMSE of 535.9538 on the test set, the accuracy is comparatively high as well.




## Assumptions

After applying required transformations, Model 5 provided all the assumption validation as follows:

Normality: Residual plot/qq plot show normal distribution

Constant variance: Residual scatter plot of fitted values indicate constant variance

Independence: We assume observations were independent of each other

## Parameter Estimates/Coefficients

The interpretation of the coefficients is not straightforward due to the fact that it is polynomial rather than linear, and our response variable is on a log scale rather than a non-transformed scale. At a high level, one can interpret the coefficient of, say, the "Year" variable as the growth rate of the rate of change (in other words, acceleration) of the log of acreage price. With all other variables held constant, or every unit change in years, the rate at which the polynomial function determining log acreage price decreases by 3.71.

For example in the three states of California, Colorado and Connecticut, we see for the same year in the state of California the direction and steepness of the median acre value function increases by $\beta^{0.811}$ or by 2.25 with respect to other states and has a statistically significant p-value. ([Source of Interpretation])

Region_newCalifornia 0.81198 0.06274 12.942 < 2e-16 ***

Region_newColorado -0.78137 0.06273 -12.456 < 2e-16 ***

Region_newConnecticut 1.50866 0.08870 17.009 < 2e-16 ***

### Conclusion

For purposes of prediction accuracy, the spline model looks to be the best. With that said, as stated above, **model 5** (the polynomial model) meets all of the basic regression assumptions without sacrificing much accuracy. This is ultimately the model we have chosen as optimal. Due to the fact that the model is built on the interaction of year, land category, and region, the model has 408 terms (see appendix for table of terms with parameter estimates and confidence intervals). Furthermore, the interpretation of the coefficients is not straightforward due to the fact that it is polynomial rather than linear, and our response variable is on a log scale rather than a non-transformed scale. At a high level, one can interpret the coefficient of, say, the "Year" variable as the growth rate of the rate of change (in other words, acceleration) of the log of acreage price. With all other variables held constant, or every unit change in years, the rate at which the polynomial function determining log acreage price decreases by 3.71.

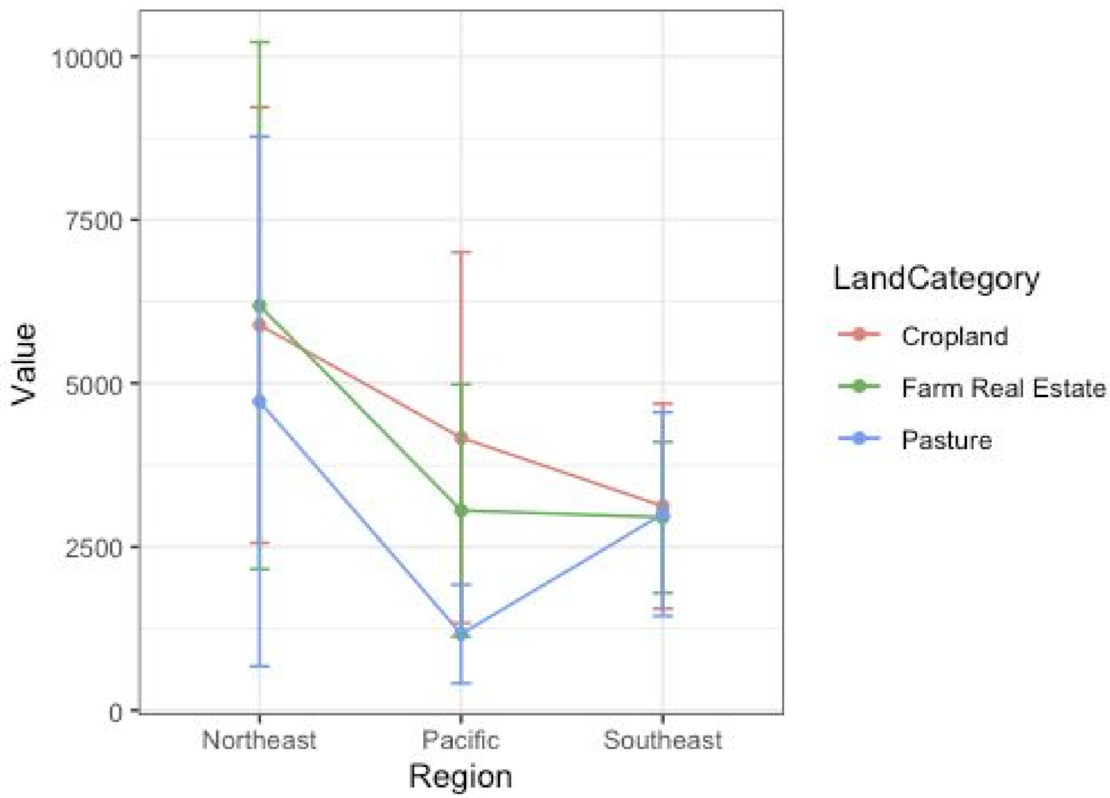
| Model number | Formula | Statistics |
|---|---|---|
| Non Transformed (Model 1) | Acre.Value ~ Year*LandCategory*Region_new | $R^2$ = 0.9452<br>RMSE = 688.595 |
| Log Linear (Model 2) | log(Acre.Value) ~ Year*LandCategory*Region_new | $R^2$ = 0.9633<br>RMSE = 759.2183 |
| Log - Log (Model 3) | log(Acre.Value) ~ log(Year)*LandCategory*Region_new | $R^2$ = 0.9633<br>RMSE = 759.2183 |
| Linear Model With Splines (Model 4) | Acre.Value ~ (Year + X)*LandCategory*Region_new | $R^2$ = 0.978<br>RMSE=370.86 |
| **Log- Polynomial (Model 5)** | **log(Acre.Value) ~ poly(Year,2)*LandCategory*Region_new** | $R^2$ = **0.9452**<br>**RMSE=454.1732** |

## OBJECTIVE 2

In our second objective, we ran a 2-way ANOVA on the data. For our categorical values, we chose to use region and land category. With our region category and the above analysis in objective 1, we know that the sample sizes are very different. This is significant as we deep dive into particular regions in our analysis.

As we've seen in our analysis in Objective 1, the sample sizes in the different regions made it unclear on how to handle regions where it was more dense like the Northeast. In our ANOVA analysis, we'll continue using the same data, **but filtered on 3 regions: Northeast, Pacific, and Southeast regions.** Breaking out the Regions will help us provide an understanding in the significance of land category in its corresponding region.

## Analysis



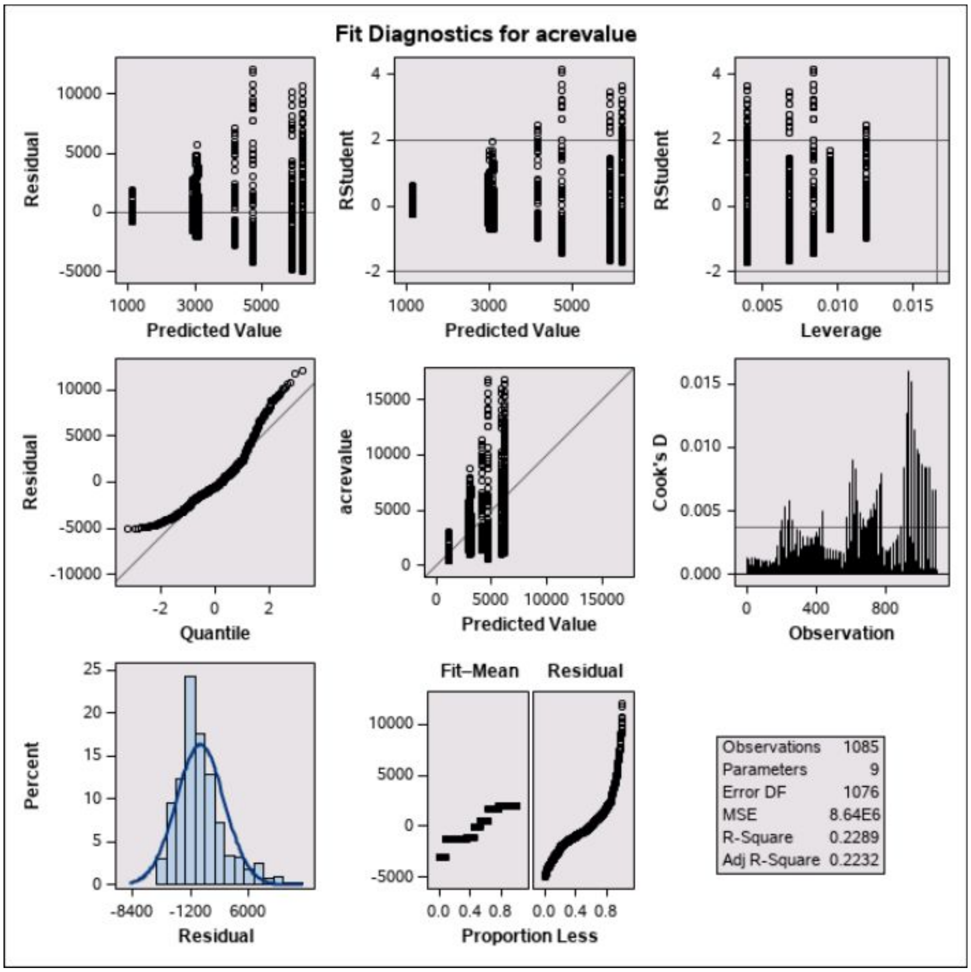

## Ftest

In our Type III sum of squares table, our interaction term is significant. This defines an **non additive model** as we reject the null hypothesis (p-value <0.001). Because this is a subset of data from the larger dataset, we know there is interaction but to what extent can not be determined.

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| landcategory | 2 | 347056043 | 173528021 | 20.07 | <.0001 |
| region | 2 | 1848227438 | 924113719 | 106.90 | <.0001 |
| landcategory*region | 4 | 241908553 | 60477138 | 7.00 | <.0001 |

### Assumption

The assumptions for 2-way ANOVA are independence, constant variance and normality. Under normality, the qq-plot is not linear, showing a possible violation. This is possibly due to the unequal standard deviations in our regions and (less of a possibility), the land categories. Also, the histogram looks normally distributed, but the tall center does show as a concern. We also see potential outliers which contribute to the right skew in our histogram.




### Contrast

In our **contrast analysis,** comparing land category paired with region, The Pacific region and Pasture land category separated itself from the pack showing to be significantly different from the rest (p-value of < 0.0001 at 95% confidence interval). The below charts are results from running our analysis with tukey adjustment.

| landcategory | region | acrevalue LSMEAN | LSMEAN Number |
|---|---|---|---|
| Cropland | Northeast | 5891.70068 | 1 |
| Cropland | Pacific | 4169.76190 | 2 |
| Cropland | Southeast | 3122.28571 | 3 |
| Farm Real Estate | Northeast | 6189.40476 | 4 |
| Farm Real Estate | Pacific | 3055.00000 | 5 |
| Farm Real Estate | Southeast | 2952.19048 | 6 |
| Pasture | Northeast | 4724.20168 | 7 |
| Pasture | Pacific | 1168.96429 | 8 |
| Pasture | Southeast | 3001.52381 | 9 |

Least Squares Means for Effect landcategory*region
t for H0: LSMean(i)=LSMean(j) / Pr > |t|

Dependent Variable: acrevalue

| i/j | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 4.281874<br>0.0007 | 7.371653<br><.0001 | -0.97562<br>0.9880 | 7.053906<br><.0001 | 7.824414<br><.0001 | 3.220117<br>0.0356 | 11.74383<br><.0001 | 7.693098<br><.0001 |
| 2 | -4.28187<br>0.0007 | | 2.433724<br>0.2666 | -5.45216<br><.0001 | 2.457144<br>0.2544 | 2.828926<br>0.1082 | -1.32325<br>0.9244 | 6.61432<br><.0001 | 2.714305<br>0.1439 |
| 3 | -7.37165<br><.0001 | -2.43372<br>0.2666 | | -8.9808<br><.0001 | 0.156333<br>1.0000 | 0.419175<br>1.0000 | -4.06919<br>0.0017 | 4.538381<br>0.0002 | 0.2976<br>1.0000 |
| 4 | 0.975621<br>0.9880 | 5.452161<br><.0001 | 8.980805<br><.0001 | | 8.461535<br><.0001 | 9.478859<br><.0001 | 4.480314<br>0.0003 | 13.55301<br><.0001 | 9.334407<br><.0001 |
| 5 | -7.05391<br><.0001 | -2.45714<br>0.2544 | -0.15633<br>1.0000 | -8.46154<br><.0001 | | 0.238869<br>1.0000 | -3.9838<br>0.0023 | 4.157176<br>0.0012 | 0.124248<br>1.0000 |
| 6 | -7.82441<br><.0001 | -2.82893<br>0.1082 | -0.41918<br>1.0000 | -9.47886<br><.0001 | -0.23887<br>1.0000 | | -4.50126<br>0.0003 | 4.143178<br>0.0012 | -0.12157<br>1.0000 |
| 7 | -3.22012<br>0.0356 | 1.323254<br>0.9244 | 4.069187<br>0.0017 | -4.48031<br>0.0003 | 3.983802<br>0.0023 | 4.501263<br>0.0003 | | 8.485111<br><.0001 | 4.375947<br>0.0005 |
| 8 | -11.7438<br><.0001 | -6.61432<br><.0001 | -4.53838<br>0.0002 | -13.553<br><.0001 | -4.15718<br>0.0012 | -4.14318<br>0.0012 | -8.48511<br><.0001 | | -4.2578<br>0.0008 |
| 9 | -7.6931<br><.0001 | -2.7143<br>0.1439 | -0.2976<br>1.0000 | -9.33441<br><.0001 | -0.12425<br>1.0000 | 0.121575<br>1.0000 | -4.37595<br>0.0005 | 4.2578<br>0.0008 | |

## Conclusion

Our 2-way ANOVA analysis helped us understand the acre value and the correlation between land category and the three selected regions. We conclude land values are not significantly different in Southeast irrespective of the land type (p-value close to 1). While in pacific the value has a statistically significant difference (p-value .001 corp/pasture and .0012 pasture/realest)

# APPENDIX

## Data Set

https://www.kaggle.com/jmullan/agricultural-land-values-19972017

https://www.nass.usda.gov/Publications/Todays_Reports/reports/land0818.pdf

## Economic Categories

**Economic Regions**


**Economic Regions:**

**Northeast:**.................... Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.

**Lake States:**................. Michigan, Minnesota, Wisconsin.

**Corn Belt:**.................... Illinois, Indiana, Iowa, Missouri, Ohio.

**Northern Plains:** ......... Kansas, Nebraska, North Dakota, South Dakota.

**Appalachian:**............... Kentucky, North Carolina, Tennessee, Virginia, West Virginia.

**Southeast:**.................... Alabama, Florida, Georgia, South Carolina.

**Delta States:** ................ Arkansas, Louisiana, Mississippi.

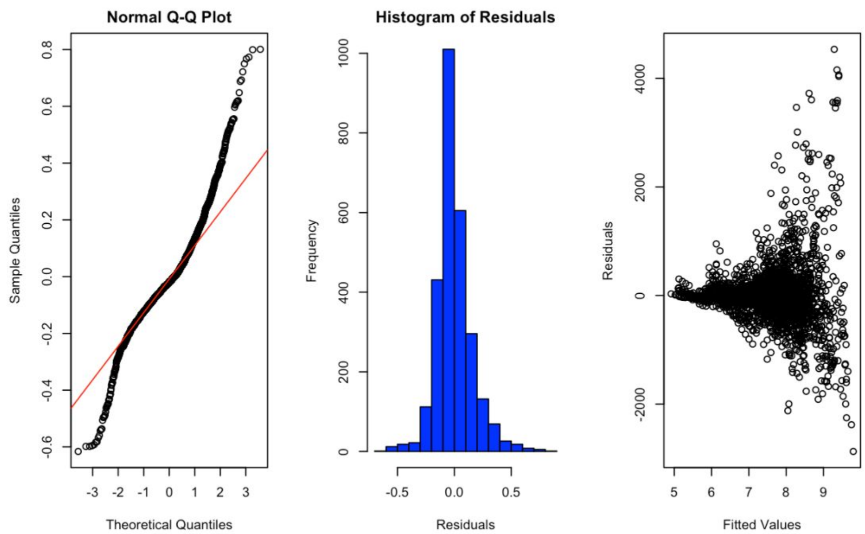
**Southern Plains:** ......... Oklahoma, Texas.

**Mountain:**.................... Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming.

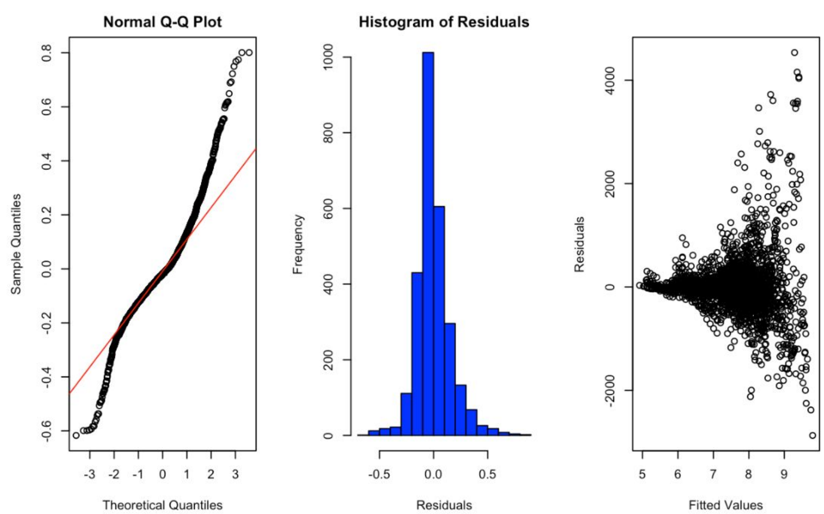
**Pacific:**......................... California, Oregon, Washington.

## Residual Plots

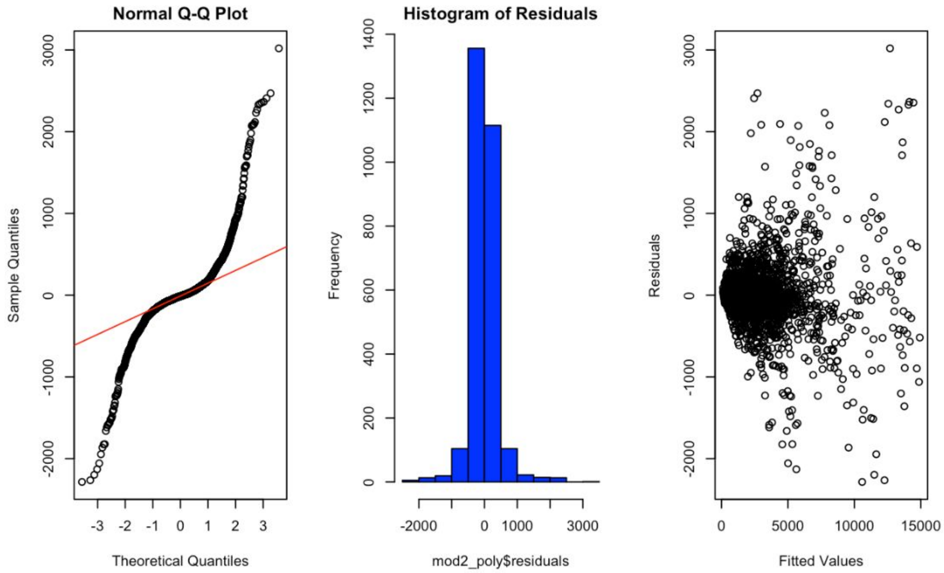
### Model 2




### Model 3



## Model 4



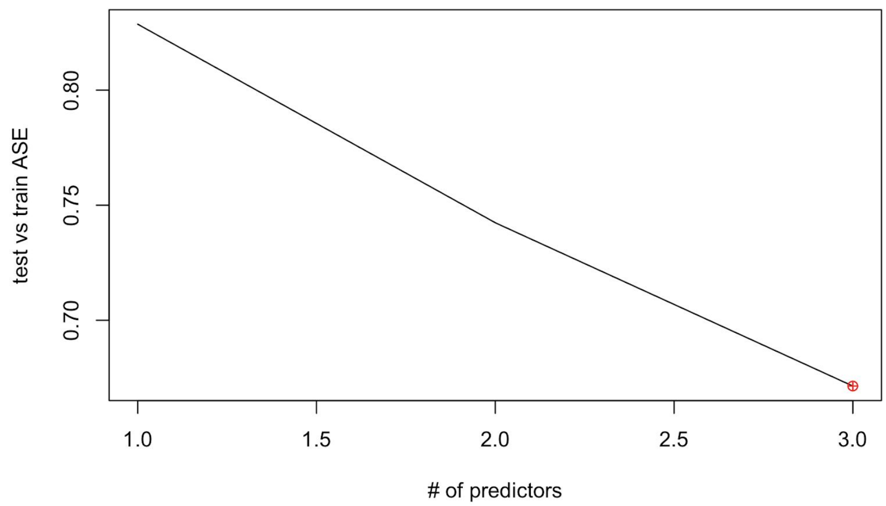

## Summary Statistics

| | Region | LandCategory | N | Mean | Min | Max | IQR | SD | SE |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Appalachian | Cropland | 21 | 3043.33333333333 | 1730 | 3890 | 1350 | 764.64588754098 | 166.859412356821 |
| 2 | Continental U.S. | Cropland | 21 | 2551.42857142857 | 1270 | 4130 | 1760 | 1035.1921836755 | 225.897454310458 |
| 3 | Corn Belt | Cropland | 21 | 3918.09523809524 | 1760 | 7000 | 3420 | 1906.18629479812 | 415.963951647609 |
| 4 | Delta States | Cropland | 21 | 1717.71428571429 | 956 | 2760 | 1000 | 609.514654693154 | 133.007002015059 |
| 5 | Lake States | Cropland | 21 | 2822.38095238095 | 1130 | 4830 | 2070 | 1273.35346531081 | 277.868506724815 |
| 6 | Mountain | Cropland | 21 | 1399.90476190476 | 904 | 1780 | 550 | 318.922859757952 | 69.5946736044096 |
| 7 | Northeast | Cropland | 21 | 4368.57142857143 | 2590 | 5590 | 2120 | 1171.36367416053 | 255.612509674391 |
| 8 | Northern Plains | Cropland | 21 | 1484.95238095238 | 633 | 3130 | 1490 | 928.215625605951 | 202.553255503812 |
| 9 | Pacific | Cropland | 21 | 4657.61904761905 | 3030 | 6570 | 2160 | 1188.02737662861 | 259.248827670495 |
| 10 | Southeast | Cropland | 21 | 3148.09523809524 | 1610 | 4380 | 1570 | 952.988032703554 | 207.959037916719 |
| 11 | Southern Plains | Cropland | 21 | 1196.42857142857 | 641 | 1930 | 672 | 416.94670779712 | 90.9852309163304 |
| 12 | Appalachian | Farm Real Estate | 21 | 2980 | 1630 | 3800 | 1360 | 782.067771999333 | 170.661179227257 |
| 13 | Continental U.S. | Farm Real Estate | 21 | 1928.09523809524 | 926 | 3080 | 1310 | 756.388650414713 | 165.057535491948 |
| 14 | Corn Belt | Farm Real Estate | 21 | 3632.38095238095 | 1610 | 6370 | 3160 | 1761.64384812 | 384.422203883962 |
| 15 | Delta States | Farm Real Estate | 21 | 1957.14285714286 | 1070 | 2910 | 1050 | 606.590000388589 | 132.368790123048 |

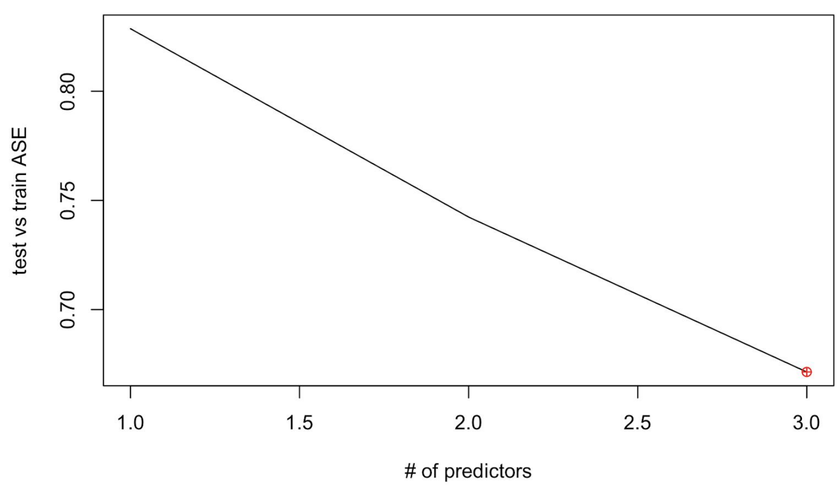
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **16** | Lake States | Farm Real Estate | 21 | 2959.04761904762 | 1200 | 4890 | 2010 | 1242.14292560037 | 271.057808596076 |
| **17** | Mountain | Farm Real Estate | 21 | 777.761904761905 | 399 | 1130 | 510 | 271.364128941521 | 59.2165077033465 |
| **18** | Northeast | Farm Real Estate | 21 | 4025.71428571429 | 2240 | 5050 | 1920 | 1068.66538929906 | 233.201906621067 |
| **19** | Northern Plains | Farm Real Estate | 21 | 1135.33333333333 | 481 | 2340 | 1044 | 678.644040225311 | 148.092270679197 |
| **20** | Pacific | Farm Real Estate | 21 | 3377.61904761905 | 1730 | 5370 | 2030 | 1170.57210269981 | 255.439774620261 |
| **21** | Southeast | Farm Real Estate | 21 | 3021.90476190476 | 1630 | 3940 | 1530 | 861.86204840229 | 188.073717876818 |
| **22** | Southern Plains | Farm Real Estate | 21 | 1234.14285714286 | 557 | 2050 | 865 | 500.794197821249 | 109.282253290979 |
| **23** | Appalachian | Pasture | 21 | 2715.71428571429 | 1510 | 3620 | 1410 | 755.324906438093 | 164.825407525619 |
| **24** | Continental U.S. | Pasture | 21 | 903.428571428571 | 466 | 1350 | 533 | 319.801433928707 | 69.7863942063675 |
| **25** | Corn Belt | Pasture | 21 | 1648.80952380952 | 756 | 2440 | 1080 | 608.908254094787 | 132.874674555846 |
| **26** | Delta States | Pasture | 21 | 1737.04761904762 | 955 | 2480 | 990 | 553.34017350907 | 120.748725245967 |
| **27** | Lake States | Pasture | 21 | 1391 | 486 | 2080 | 951 | 563.114286801534 | 122.881611627576 |
| **28** | Mountain | Pasture | 21 | 443.666666666667 | 219 | 625 | 321 | 161.516356240888 | 35.2457584974915 |
| **29** | Northeast | Pasture | 21 | 2829.04761904762 | 1890 | 3480 | 1300 | 629.689643887406 | 137.409545592563 |
| **30** | Northern Plains | Pasture | 21 | 502.142857142857 | 206 | 1040 | 399 | 293.634855852347 | 64.0763787438489 |
| **31** | Pacific | Pasture | 21 | 1315.85714285714 | 729 | 1900 | 739 | 408.22497298846 | 89.0819923471866 |
| **32** | Southeast | Pasture | 21 | 3184.28571428571 | 1340 | 5040 | 2040 | 1240.30065479533 | 270.655792147756 |
| **33** | Southern Plains | Pasture | 21 | 1037.2380952381 | 484 | 1620 | 819 | 431.786741894874 | 94.2235918481724 |

## Feature selection

### Backward




### Forward



Parameter Estimates for regression

```
Call:
lm(formula = log(Acre.Value) ~ poly(Year, 2) + LandCategory +
    Region_new, data = train)
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.80943    0.04502 173.483  < 2e-16 ***
poly(Year, 2)1             18.31853    0.31601  57.968  < 2e-16 ***
poly(Year, 2)2             -2.92956    0.31571  -9.279  < 2e-16 ***
LandCategoryFarm Real Estate -0.15715    0.01497 -10.494  < 2e-16 ***
LandCategoryPasture        -0.61874    0.01494 -41.409  < 2e-16 ***
Region_newAppalachian       0.35453    0.04763   7.443 1.31e-13 ***
Region_newArizona           0.23960    0.06339   3.780 0.000160 ***
Region_newCalifornia        0.81198    0.06274  12.942  < 2e-16 ***
Region_newColorado         -0.78137    0.06273 -12.456  < 2e-16 ***
Region_newConnecticut       1.50866    0.08870  17.009  < 2e-16 ***
Region_newContinental U.S.  -0.22904    0.06127  -3.738 0.000189 ***
Region_newCorn Belt         0.31366    0.06155   5.096 3.71e-07 ***
Region_newDelaware          0.94059    0.06993  13.449  < 2e-16 ***
Region_newDelta States     -0.09619    0.04937  -1.948 0.051496 .
Region_newFlorida           0.78721    0.06182  12.733  < 2e-16 ***
Region_newGeorgia           0.40112    0.06101   6.574 5.84e-11 ***
Region_newIdaho            -0.19264    0.06274  -3.071 0.002157 **
Region_newIllinois          0.48174    0.06409   7.516 7.61e-14 ***
Region_newIndiana           0.45662    0.06127   7.452 1.23e-13 ***
Region_newIowa              0.34478    0.06183   5.577 2.70e-08 ***
Region_newLake States       0.10304    0.04930   2.090 0.036707 *
Region_newMaine            -0.12209    0.10148  -1.203 0.229062
Region_newMaryland          1.01904    0.06448  15.805  < 2e-16 ***
Region_newMassachusetts     1.45281    0.09072  16.015  < 2e-16 ***
```

```
Region_newMissouri         -0.07659    0.06305  -1.215 0.224590
Region_newMontana          -1.24960    0.06183 -20.210  < 2e-16 ***
Region_newMountain         -0.96745    0.06305 -15.344  < 2e-16 ***
Region_newNevada           -0.78297    0.06659 -11.758  < 2e-16 ***
Region_newNew Hampshire     0.53108    0.09071   5.855 5.36e-09 ***
Region_newNew Jersey        1.80753    0.06373  28.360  < 2e-16 ***
Region_newNew Mexico       -1.35600    0.06305 -21.507  < 2e-16 ***
Region_newNew York         -0.23674    0.06183  -3.829 0.000132 ***
Region_newNortheast         0.62602    0.06273   9.980  < 2e-16 ***
Region_newNorthern Plains  -0.90860    0.04851 -18.730  < 2e-16 ***
Region_newOhio              0.47134    0.06274   7.513 7.80e-14 ***
Region_newOklahoma         -0.77241    0.06075 -12.714  < 2e-16 ***
Region_newOregon           -0.47360    0.06486  -7.302 3.71e-13 ***
Region_newOther States      1.07799    0.06934  15.547  < 2e-16 ***
Region_newPacific           0.27716    0.06273   4.418 1.03e-05 ***
Region_newPennsylvania      0.59506    0.06568   9.060  < 2e-16 ***
Region_newRhode Island      1.67685    0.08358  20.062  < 2e-16 ***
Region_newSouth Carolina    0.15832    0.06242   2.536 0.011259 *
Region_newSoutheast         0.48756    0.06486   7.517 7.56e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3152 on 2717 degrees of freedom
Multiple R-squared:  0.8728,	Adjusted R-squared:  0.8705
F-statistic: 388.4 on 48 and 2717 DF,  p-value: < 2.2e-16
```

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -1.110257e+02 | -104.13728145 |
| Region_newAppalachian | 2.780906e-01 | 0.49552500 |
| Region_newArizona | 1.380097e-01 | 0.36218823 |
| Region_newCalifornia | 6.979659e-01 | 0.91540035 |
| Region_newColorado | -8.962330e-01 | -0.67879861 |
| Region_newConnecticut | 1.391389e+00 | 1.70034966 |
| Region_newContinental U.S. | -3.299382e-01 | -0.11250373 |
| Region_newCorn Belt | 1.993144e-01 | 0.41674880 |
| Region_newDelaware | 8.477533e-01 | 1.09132005 |
| Region_newDelta States | -1.972345e-01 | -0.02533750 |
| Region_newFlorida | 6.654120e-01 | 0.88284641 |
| Region_newGeorgia | 3.299866e-01 | 0.54742100 |
| Region_newIdaho | -2.992441e-01 | -0.08180971 |
| Region_newIllinois | 3.672185e-01 | 0.58465293 |
| Region_newIndiana | 3.488025e-01 | 0.56623693 |
| Region_newIowa | 2.340215e-01 | 0.45145597 |
| Region_newKentucky | 8.024012e-02 | 0.29767455 |
| Region_newLake States | 2.438522e-02 | 0.19628223 |
| Region_newMaine | -2.942880e-01 | 0.01467234 |
| Region_newMaryland | 8.972300e-01 | 1.12139190 |
| Region_newMassachusetts | 1.325895e+00 | 1.63485555 |
| Region_newMissouri | -1.831312e-01 | 0.03430321 |
| Region_newMontana | -1.363730e+00 | -1.14629606 |
| Region_newMountain | -1.049666e+00 | -0.83223143 |
| Region_newNevada | -8.748404e-01 | -0.64226650 |
| Region_newNew Hampshire | 3.913884e-01 | 0.70034869 |
| Region_newNew Jersey | 1.691798e+00 | 1.90923290 |
| Region_newNew Mexico | -1.441141e+00 | -1.22370647 |
| Region_newNew York | -3.527347e-01 | -0.13530025 |
| Region_newNorth Carolina | 4.875701e-01 | 0.70500449 |
| Region_newNortheast | 5.158387e-01 | 0.73327317 |
| Region_newNorthern Plains | -9.756039e-01 | -0.80717988 |
| Region_newOhio | 3.707145e-01 | 0.58814889 |
| Region_newOklahoma | -8.664547e-01 | -0.64902024 |
| Region_newOregon | -5.811268e-01 | -0.36369241 |
| Region_newOther States | 9.748632e-01 | 1.21842511 |
| Region_newPacific | 2.090176e-01 | 0.42645200 |
| Region_newPennsylvania | 4.920097e-01 | 0.70944416 |
| Region_newRhode Island | 1.536373e+00 | 1.84533289 |
| Region_newSouth Carolina | 3.647997e-02 | 0.25391440 |
| Region_newSoutheast | 3.231994e-01 | 0.54063388 |
| Region_newSouthern Plains | -6.909319e-01 | -0.47349749 |
| Region_newTennessee | 2.890351e-01 | 0.50646957 |
| Region_newTexas | -6.473402e-01 | -0.42990580 |
| Region_newUtah | -4.084664e-01 | -0.19103194 |
| Region_newVermont | -7.878021e-03 | 0.30108232 |
| Region_newVirginia | 4.694701e-01 | 0.68690457 |
| Region_newWashington | -4.962496e-01 | -0.27881519 |
| Region_newWest Virginia | -8.684038e-02 | 0.13059405 |
| Region_newWyoming | -1.414163e+00 | -1.19672849 |

Two Way Anova Code

```
#Get DataSet
agri_data <- read.csv("Combined_Clean.csv")

#Summarize data
head(agri_data)
summary(agri_data)

#Visualize data
plot(agri_data$Year,agri_data$Acre.Value)
ggplot(data = agri_data) + geom_point(mapping=aes(x=agri_data$Year,
y=agri_data$Acre.Value, color=agri_data$LandCategory)) +
facet_wrap(~Region)


#Create model
model <- aov(Acre.Value~Region+Year, data=agri_data)
#QQ Plot
qqnorm(model$residuals)

#Log Data
agri_data$logAcre.Value <- log(agri_data$Acre.Value)
#Summarize data
head(agri_data)
summary(agri_data)

#Create model with logged data
model2 <- aov(logAcre.Value~Region+Year, data=agri_data)

#Plot / QQ Plot of model with logged values
plot(model2$fitted.values, model2$residuals,ylab="Residuals")
qqnorm(model2$residuals)

#2Way Anova on both models
anova(model)
anova(model2)

/** SAS **/
/** load the data **/
data agridata;
infile '/home/u41023123/6372_stats2/midterm
project/Combined_Clean.csv' dlm=',' firstobs=2;
informat state $30.;
informat landcategory $30.;
informat region $30.;
informat regionvsstate $30.;
input state $ landcategory $ region $ regionvsstate $ year
```

```
acrevalue;
run;

/** log the data **/
data agridata;
set agridata;
logacrevalue = log(acrevalue);
logyear = log(year);
run;

/**  plot the data **/
proc sgplot data=agridata;
vbox acrevalue / category=year group=landcategory;
run;

/** breaking out denser plots that may have skewed the data **/
data agridata;
set agridata;
if region = 'Northeast'
           | region = 'Pacific'
           | region = 'Mountain'
           | region = 'Southeast'
           | region = 'Corn Belt'
           | region = 'Appalachian'
           | region = 'Southern Plains'
           then newregion = state;
     else newregion = region;
run;

/** anova **/
proc glm data=agridata plots=(DIAGNOSTICS RESIDUALS);
class year landcategory newregion;
model logacrevalue = newregion year landcategory
newregion*year*landcategory;
/* lsmeans newregion year landcategory / pdiff tdiff adjust=bon; */
run;
```

Regression Model Code

````
```{r}


library(ggplot2)

library(gridExtra)

library(broom)
````
````

````
library(MASS)
library(leaps)
theme_set(theme_bw())

dat <- read.csv('Combined_Clean.csv')

head(dat)

```

```{r}

str(dat)

```

```{r}

unique(dat$Year)

```



```{r}



p1 <- ggplot(dat, aes(Year, Acre.Value)) +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_point(alpha=0.1) +
  geom_smooth(method = "lm") +
````

````
  facet_grid(vars(LandCategory), vars(Region))

```

```{r}

p2 <- ggplot(dat[dat$Region == 'Southeast',], aes(Year, Acre.Value, group=State, color=State)) +
  geom_smooth(method = 'loess', se=F) +
  facet_grid(vars(LandCategory), vars(Region)) +
  geom_vline(xintercept = 2007)

grid.arrange(p1, p2)

```

```{r}

ggplot(dat[dat$Region == 'Pacific',], aes(Year, Acre.Value, group=State, color=State)) +
  geom_smooth(method = 'loess', se=F) +
  facet_grid(vars(LandCategory), vars(Region)) +
  geom_vline(xintercept = 2007)


```

```{r}

ggplot(dat[dat$Region == 'Southern Plains',], aes(Year, Acre.Value, group=State, color=State)) +
````

````
  geom_point() +

  geom_smooth(method = 'lm', se=F) +

  facet_grid(vars(LandCategory), vars(Region))


```





```{r}


dat$Region_new <- ifelse(dat$Region == 'Northeast' | dat$Region ==
'Pacific' | dat$Region == 'Mountain' | dat$Region == 'Southeast' |
dat$Region == 'Corn Belt' | dat$Region == 'Appalachian' |
dat$Region == 'Southern Plains',as.character(dat$State),
as.character(dat$Region))


dat$xbar <- ifelse(dat$Year >= 2007,1,0)


dat$diff <- dat$Year - 2007


dat$X <- dat$xbar*dat$diff


dat <- dat[!is.na(dat$Acre.Value),]


tail(dat)


```





```{r}
````

```
index <- sample(seq_len(nrow(dat)), floor(.8*nrow(dat)))

train <- dat[index,]
test <- dat[-index,]

nrow(train) + nrow(test) == nrow(dat)
```

```{r}

mod <- lm(formula = Acre.Value ~ Year*LandCategory*Region , data = train)

summary(mod)

```

```{r}

test$predictions_mod <- predict(mod, newdata = test)

rmse_mod <- sqrt(sum((test$Acre.Value - test$predictions_mod)^2) / nrow(test))

rmse_mod
```

````
```

```{r}

mod2 <- lm(formula = Acre.Value ~ Year*LandCategory*Region_new ,
data = train)

summary(mod2)

```

```{r, warning=FALSE, message=FALSE}

test$predictions_mod2 <- predict(mod2, newdata = test)

rmse_mod2 <- sqrt(sum((test$Acre.Value - test$predictions_mod2)^2)
/ nrow(test))

rmse_mod2

```

Residual diagnostics

```{r}

par(mfrow = c(1,3))
qqnorm(mod2$residuals)
````

```
qqline(mod2$residuals, col='red')

hist(mod2$residuals, col = 'blue', main = 'Histogram of Residuals',
xlab='residuals')

plot(mod2$fitted.values, mod2$residuals, type = 'p', xlab = 'Fitted
Values', ylab = 'Residuals')

```

Possibly need a log transform

```{r}

mod3 <- lm(formula = log(Acre.Value) ~ Year*LandCategory*Region_new
, data = train)

summary(mod3)

```

```{r}

par(mfrow = c(1,3))

qqnorm(mod3$residuals)

qqline(mod3$residuals, col='red')

hist(mod3$residuals, col = 'blue', main = 'Histogram of Residuals',
xlab = 'Residuals')

plot(mod3$fitted.values, mod2$residuals, type = 'p', xlab = 'Fitted
Values', ylab = 'Residuals')

```
```
```

```
```{r, warning=FALSE, message=FALSE}

test$predictions_mod3 <- predict(mod3, newdata = test)

rmse_mod3 <- sqrt(sum((test$Acre.Value -
exp(test$predictions_mod3))^2) / nrow(test))

rmse_mod3

```

That helped a little bit. Let's try taking the log of year as well
.

```{r}

mod4 <- lm(formula = log(Acre.Value) ~
log(Year)*LandCategory*Region_new , data = train)

summary(mod4)

```

```{r}

par(mfrow = c(1,3))
qqnorm(mod4$residuals)
qqline(mod4$residuals, col='red')
hist(mod4$residuals, col = 'blue', main = 'Histogram of Residuals',
```
```

```
xlab = 'Residuals')

plot(mod4$fitted.values, mod2$residuals, type = 'p', xlab = 'Fitted
Values', ylab = 'Residuals')


```

```
```{r}


test$predictions_mod4 <- predict(mod4, newdata = test)


rmse_mod4 <- sqrt(sum((test$Acre.Value -
exp(test$predictions_mod3))^2) / nrow(test))


rmse_mod4



```

```
```{r}


ggplot(test,aes(Region_new, Acre.Value - predictions_mod2)) +

  theme(axis.text.x = element_text(angle = 90)) +

  geom_boxplot()


```

```
```{r}


mod2_poly <- lm(formula = Acre.Value ~
poly(Year,2)*LandCategory*Region_new , data = train)
```
```

~~~~
summary(mod2_poly)

```

```{r}

test$predictions_mod2_poly <- predict(mod2_poly, newdata = test)

rmse_mod2_poly <- sqrt(sum((test$Acre.Value -
test$predictions_mod2_poly)^2) / nrow(test))

rmse_mod2_poly

```

Residual diagnostics

```{r}

par(mfrow = c(1,3))
qqnorm(mod2_poly$residuals)
qqline(mod2_poly$residuals, col='red')
hist(mod2_poly$residuals, col = 'blue', main = 'Histogram of
Residuals')
plot(mod2_poly$fitted.values, mod2_poly$residuals, type = 'p',
xlab='Fitted Values', ylab='Residuals')
```
~~~~

~~~~
```{r}

mod3_poly <- lm(formula = log(Acre.Value) ~
poly(Year,2)*LandCategory*Region_new , data = train)

summary(mod3_poly)

```

```{r}

test$predictions_mod3_poly <- predict(mod3_poly, newdata = test)

rmse_mod3_poly <- sqrt(sum((test$Acre.Value -
exp(test$predictions_mod3_poly))^2) / nrow(test))

rmse_mod3_poly

```

Residual diagnostics

```{r}

par(mfrow = c(1,3))
qqnorm(mod3_poly$residuals)
qqline(mod3_poly$residuals, col='red')
hist(mod3_poly$residuals, col = 'blue', main = 'Histogram of
~~~~

```
Residuals')

plot(mod3_poly$fitted.values, mod3_poly$residuals, type = 'p', xlab
= 'Fitted Values', ylab = 'Residuals')

```
```

```
```{r}


mod4_poly <- lm(formula = log(Acre.Value) ~
poly(log(Year),2)*LandCategory*Region_new , data = train)


summary(mod4_poly)


```
```

```
```{r}


test$predictions_mod4_poly <- predict(mod4_poly, newdata = test)


rmse_mod4_poly <- sqrt(sum((test$Acre.Value -
exp(test$predictions_mod4_poly))^2) / nrow(test))


rmse_mod4_poly


```
```

```
Residual diagnostics


```{r}
```

````
par(mfrow = c(1,3))

qqnorm(mod4_poly$residuals)

qqline(mod4_poly$residuals, col='red')

hist(mod4_poly$residuals, col = 'blue')

plot(mod4_poly$fitted.values, mod4_poly$residuals, type = 'p')
```



```{r}

mod2_cub <- lm(formula = Acre.Value ~
poly(Year,3)*LandCategory*Region_new , data = train)


summary(mod2_cub)


```


```{r}


test$predictions_mod2_cub <- predict(mod2_cub, newdata = test)


rmse_mod2_cub <- sqrt(sum((test$Acre.Value -
test$predictions_mod2_cub)^2) / nrow(test))


rmse_mod2_cub


```
````

Residual diagnostics
````

````
```{r}

par(mfrow = c(1,3))
qqnorm(mod2_cub$residuals)
qqline(mod2_cub$residuals, col='red')
hist(mod2_cub$residuals, col = 'blue')
plot(mod2_cub$fitted.values, mod2_cub$residuals, type = 'p')
```

```{r}

mod2_spl <- lm(formula = Acre.Value ~ (Year +
X)*LandCategory*Region_new , data = train)

spl_smry <- broom::tidy(summary(mod2_spl))

spl_confints <- broom::tidy(confint(mod2_spl))

names(spl_confints) <- c("term", "2.5_perc", "97.5_perc")

summary_df <- merge(spl_smry, spl_confints, by = 'term')

write.csv(summary_df, 'summarydf.csv')

summary(mod2_spl)

```

```{r, warning=FALSE, message=FALSE}
````
````

```
test$predictions_mod2_spl <- predict(mod2_spl, newdata = test)

rmse_mod2_spl <- sqrt(sum((test$Acre.Value -
test$predictions_mod2_spl)^2) / nrow(test))

rmse_mod2_spl

```
```

Residual diagnostics

```
```{r}

par(mfrow = c(1,3))
qqnorm(mod2_spl$residuals)
qqline(mod2_spl$residuals, col='red')
hist(mod2_spl$residuals, col = 'blue', main = 'Histogram of
Residuals', xlab = 'Residuals')
plot(mod2_spl$fitted.values, mod2$residuals, type = 'p', xlab =
'Fitted Values', ylab = 'Residuals')
```
```

```
```{r}

ggplot(test) +
```

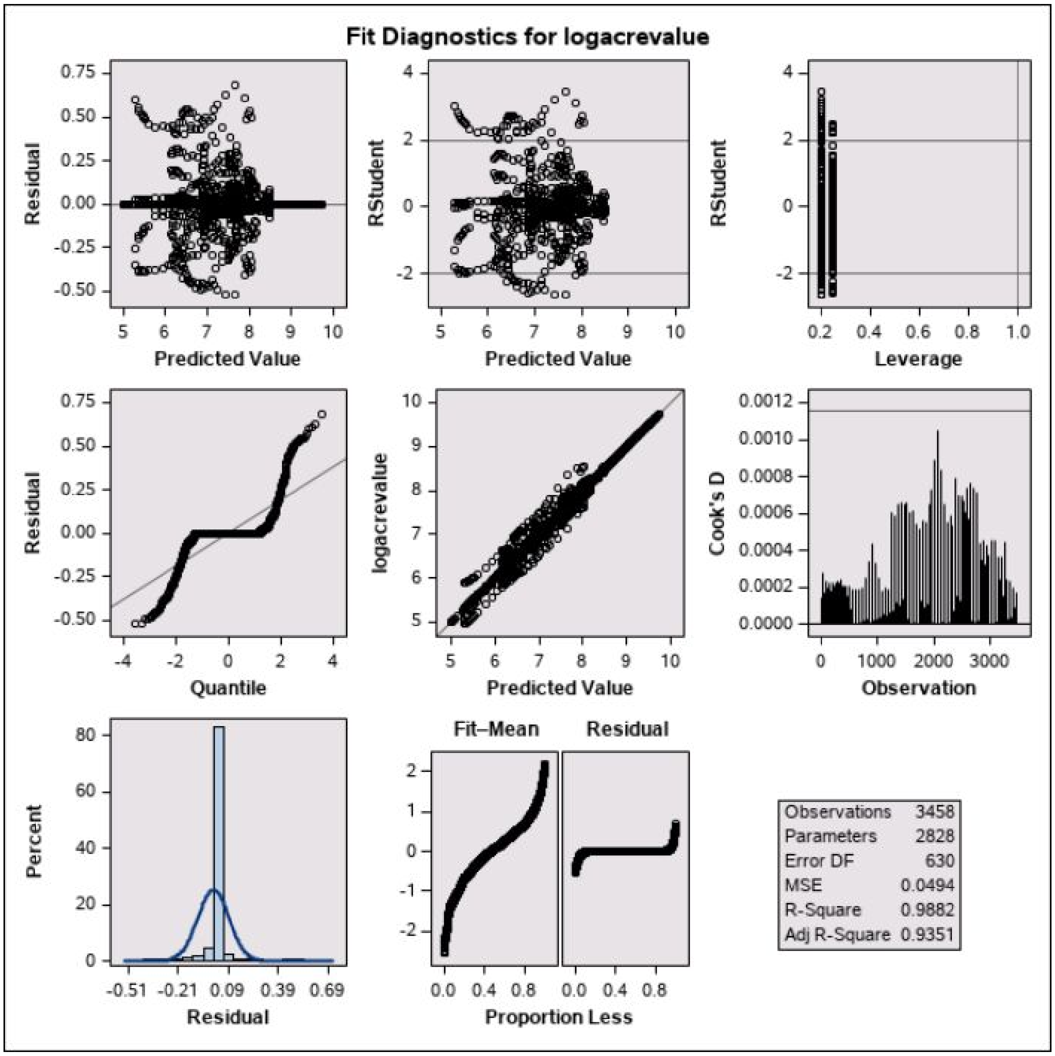
```
  geom_point(aes(Year, Acre.Value)) +
  geom_line(aes(Year, predictions_mod2_spl)) +
  facet_grid(vars(LandCategory), vars(Region))

```

```{r}

ggplot(test,aes(Region_new, Acre.Value - predictions_mod2_spl)) +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_boxplot()

```
```
```


Fit Diagnostics for logacrevalue
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
logacrevalue
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 3458
Parameters 2828
Error DF 630
MSE 0.0494
R-Square 0.9882
Adj R-Square 0.9351