Analysis on Rice Harvest Prices in Andhra Pradesh

CSP 571 – Final Project

Santosh Reddy Edulapalle
Venkata Siva Rupesh Akurati
Jobin Joyson

Outline

- Introduction
  - What are we trying to answer?
- Data Processing and Transformation
  - Data cleaning and transformation.
- Data Visualization
  - What can the data tell us so far?
- Model Building
  - Linear Regression, AIC Model, Lasso Regression
- Results and Conclusion
  - What did we learn from this?

Introduction

- Affordable crop prices are important, especially with countries with many suffering from starvation
- Rice is a widely available crop and India is the second largest rice producing country in the world
  - Andhra Pradesh is among the top 5 states in India for rice production
- Many factors play into rice production and the overall price of rice. Our goal is:
  - Understand what factors in rice production have a significant influence on rice harvest price
  - Construct an accurate model that can predict the harvest price of rice in Andhra Pradesh

Data Processing and Transformation

- Data Processing
  - 13 districts considered within the state of Andhra Pradesh
  - Deleted rows < 1990 and > 2015
  - Missing Values were handled with district mean values according to the year
  - Redundant columns spread throughout months summarized into three agricultural seasons:
    - Kharif, Rabi, Zaid
- Data Transformation
  - Combined dataset was scaled due to spread between ranges

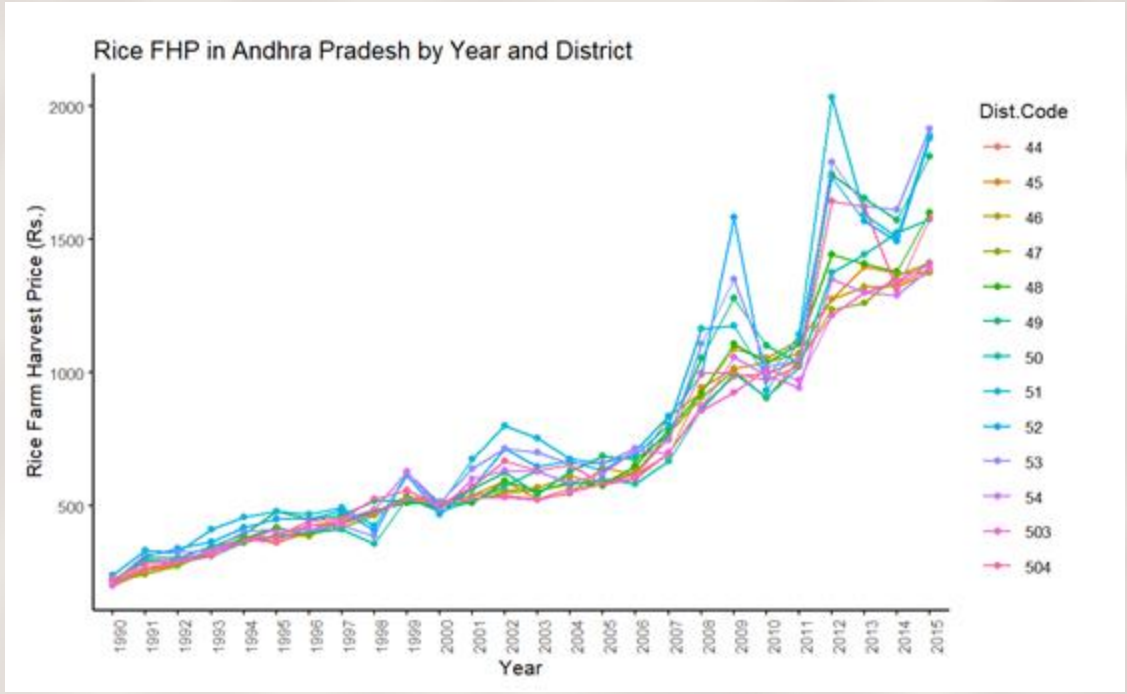
$$z = \frac{x - \bar{x}}{s},\ to\ scale$$

$$z = (x \cdot s) + \bar{x},\ re - scaling$$

    - Where x = vector, $\bar{x}$ = mean of x, and s = standard deviation of x
  - Categorical data factorized

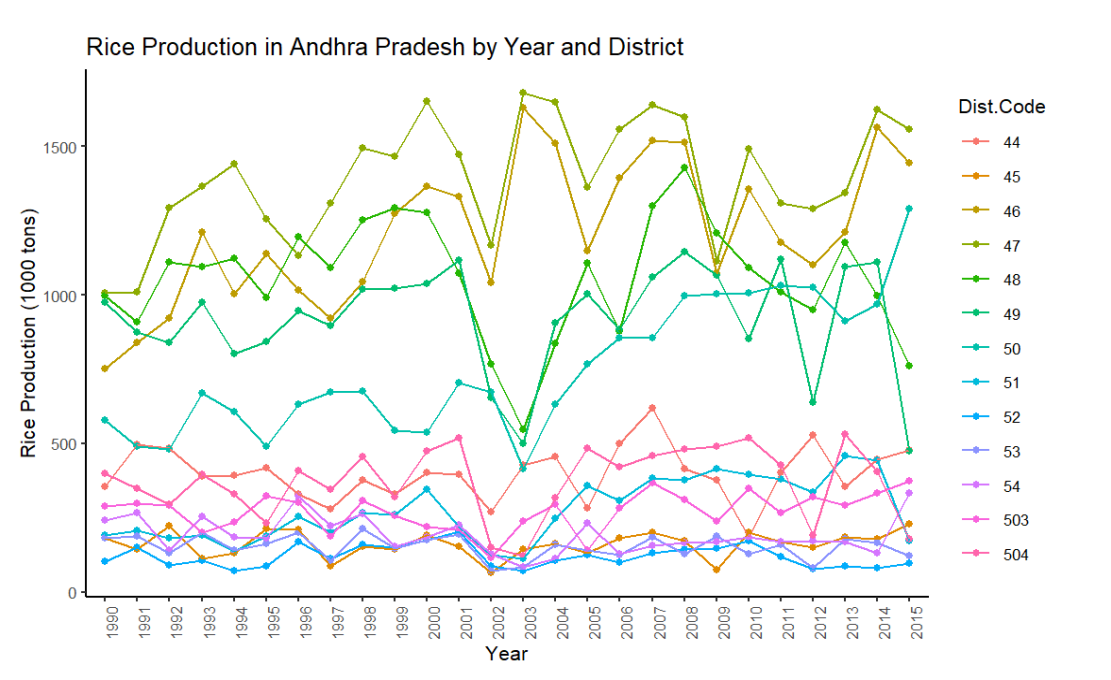
Data Visualization

Rice FHP in Andhra Pradesh 1990-2015

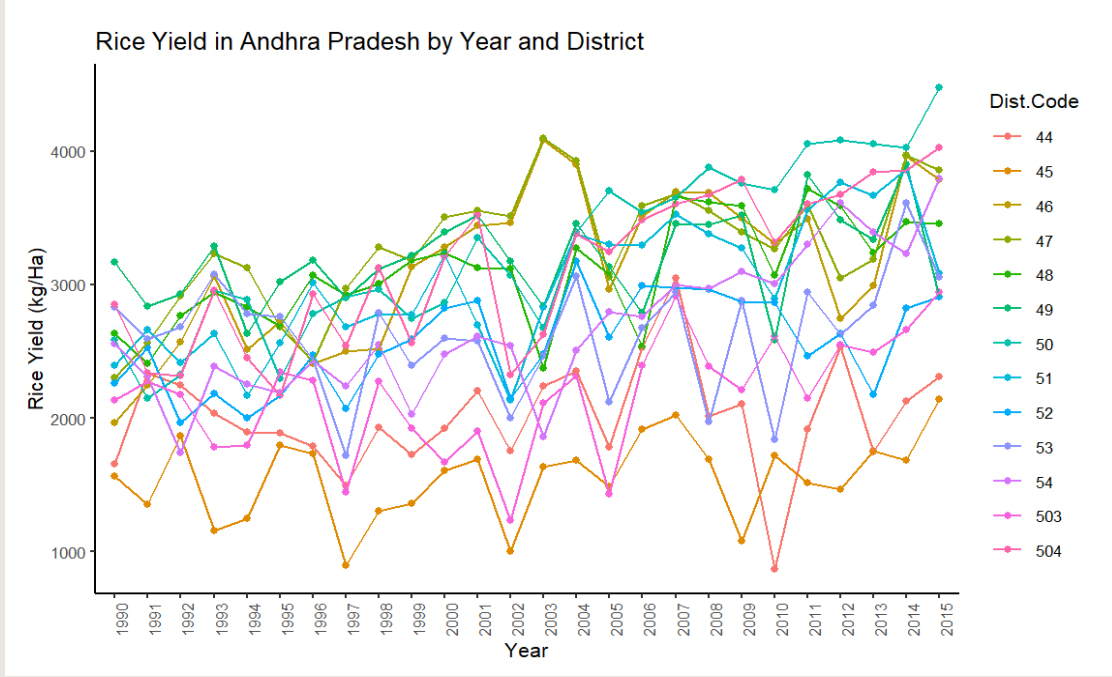


- Harvest price of rice steadily increased across the years in all districts
- Quicker changes occur after 2008 and early 2010s

Rice Production and Yield 1990-2015

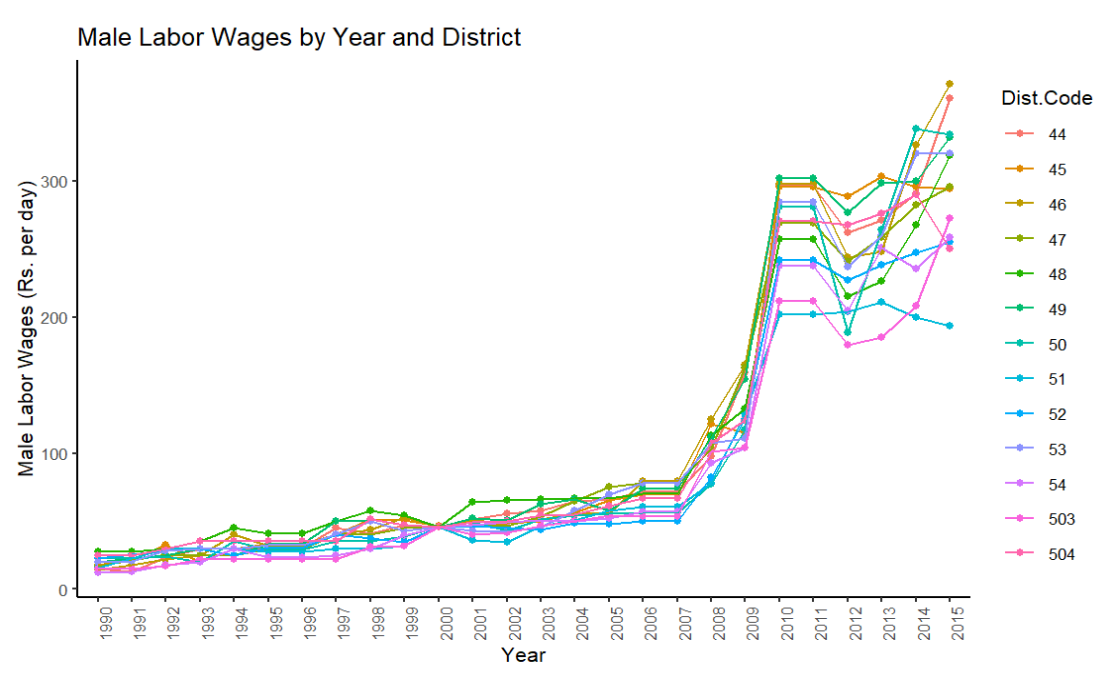


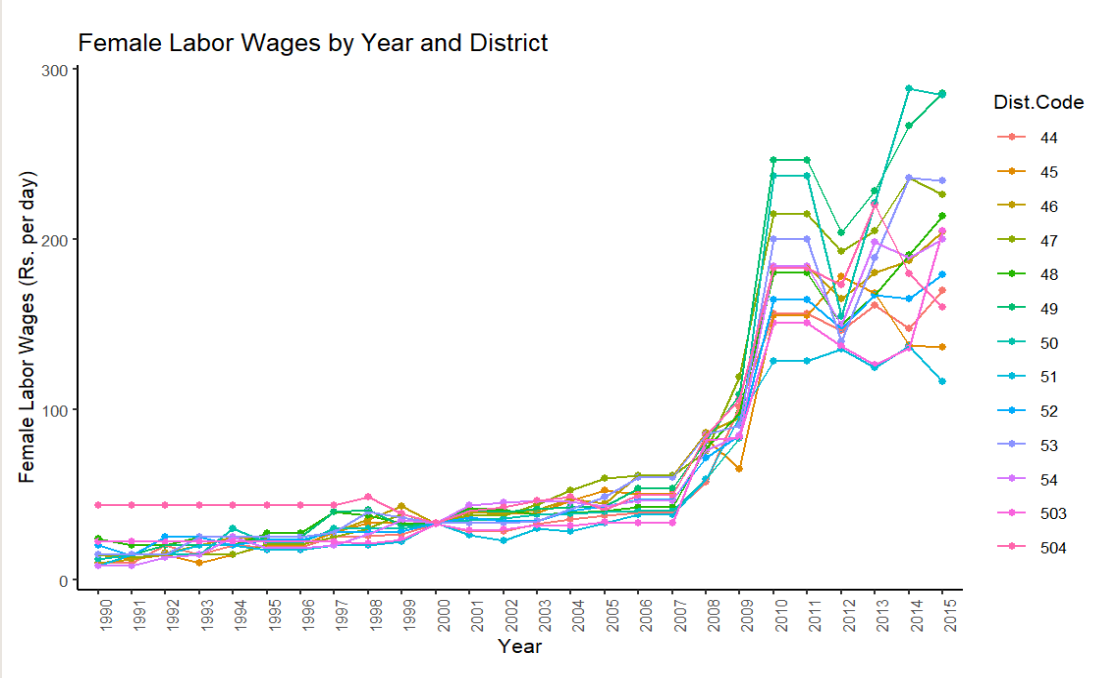
Production of rice has had many ups and downs but overall stayed within a range per district



Rice Yield has also had a similar change. Many ups and downs but overall stayed within a range per district

Labor Wages per District 1990-2015





Wages of workers have increased steadily until 2007. Wages increased rapidly from 2008-2015. Possible reasoning is inflation. We can see that throughout time, the increase of wages follows a similar trend with rice harvest prices. Building accurate models will help us confirm this relationship and possibly find other major influencers of rice harvest price within the production process of rice in Andhra Pradesh.

Model Building

Linear Regression

AIC Model

Lasso Regression

Linear Regression

- Most general model that contained all predictors
- Adjusted r-squared value of 0.9293
- It's alright, but we are sure not ALL 77 predictors are necessary

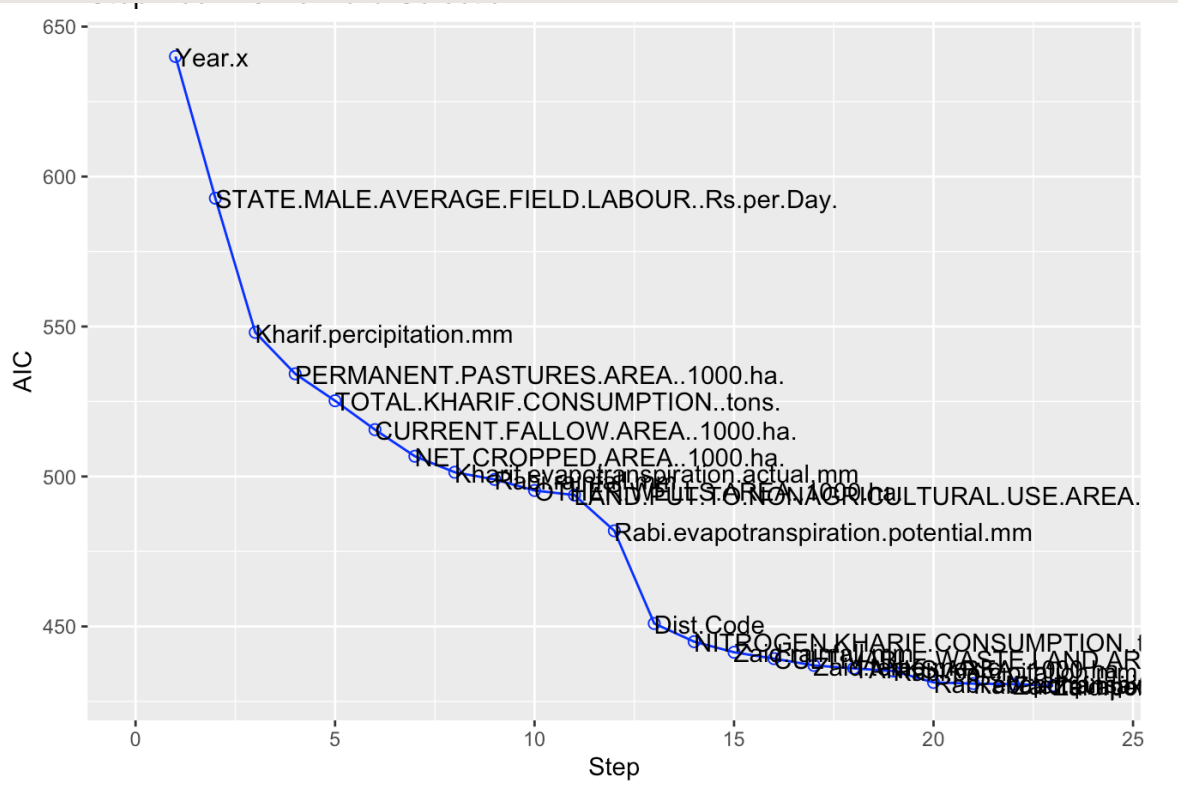
```
##
## Residual standard error: 0.4909 on 193 degrees of freedom
## Multiple R-squared:  0.9498, Adjusted R-squared:  0.9293
## F-statistic: 46.23 on 79 and 193 DF,  p-value: < 2.2e-16
```

AIC Forward Selection

- AIC estimates relative information lost by given model
  - Less information lost = more accurate model
- Prevent overfitting

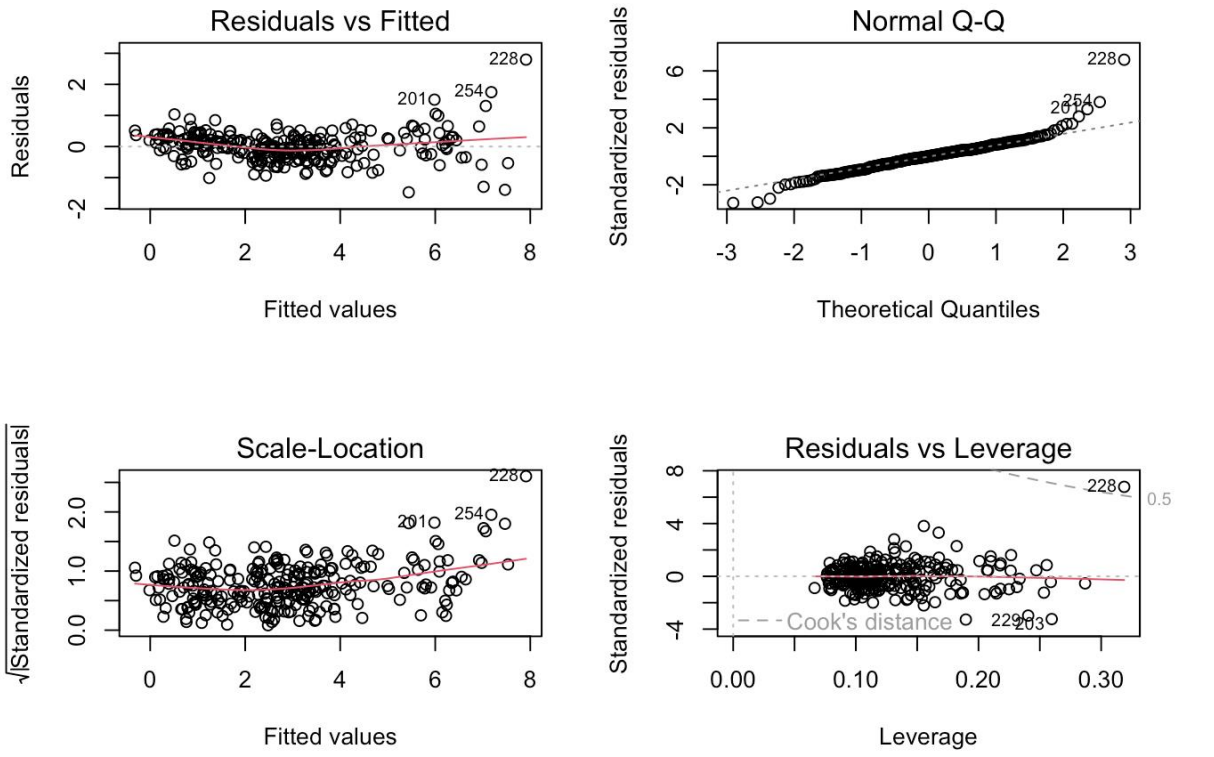
- Forward selection process throughout entire model
- Predictors within AIC criteria selected for the model

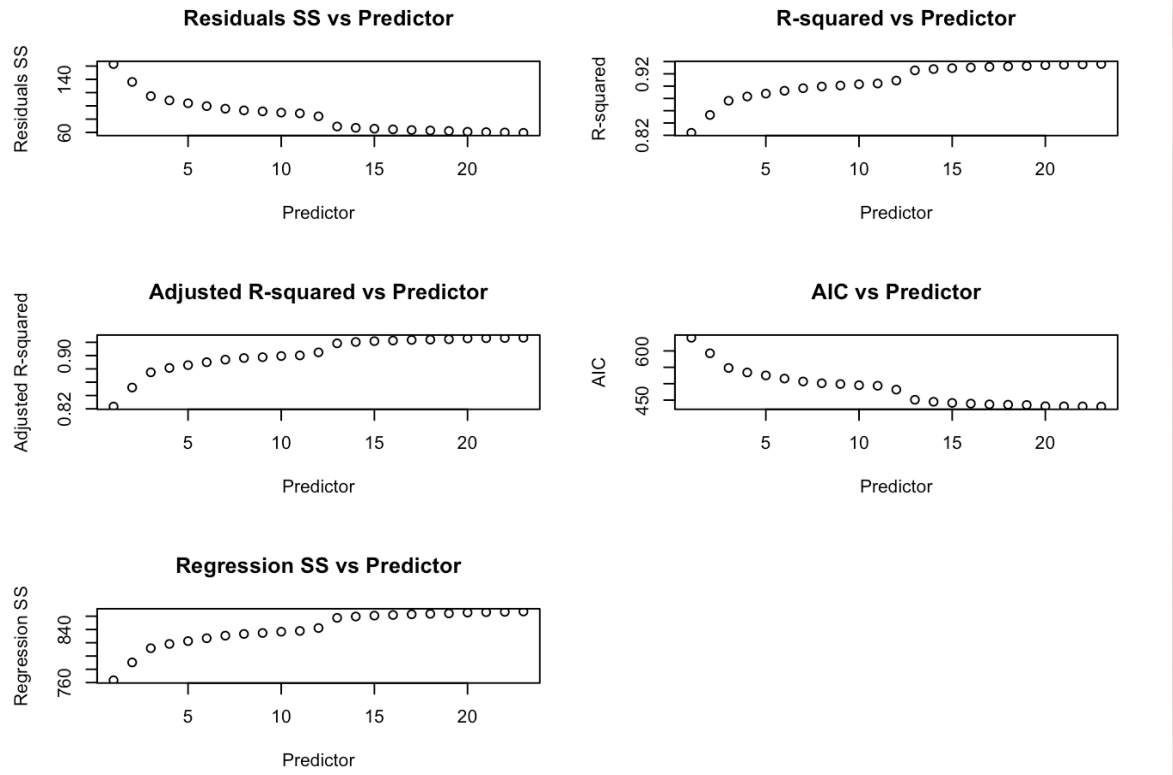
AIC cont.




```
##
##                                        Selection Summary
## ------------------------------------------------------------------------------------------------------
## Variable                                    AIC      Sum Sq       RSS      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------------------------------
## Year.x                                  640.037     763.425   163.053   0.82401        0.82336
## STATE.MALE.AVERAGE.FIELD.LABOUR..Rs.per.Day.  592.788     790.340   136.139   0.85306        0.85197
## Kharif.percipitation.mm                 548.044     811.763   114.715   0.87618        0.87480
## PERMANENT.PASTURES.AREA..1000.ha.       534.235     818.218   108.261   0.88315        0.88140
## TOTAL.KHARIF.CONSUMPTION..tons.         525.246     822.489   103.990   0.88776        0.88566
## CURRENT.FALLOW.AREA..1000.ha.           515.619     826.825    99.654   0.89244        0.89001
## NET.CROPPED.AREA..1000.ha.              506.745     830.716    95.762   0.89664        0.89391
## Kharif.evapotranspiration.actual.mm     501.405     833.257    93.222   0.89938        0.89633
## Rabi.rainfall.mm                        499.017     834.743    91.735   0.90098        0.89760
## OTHER.WELLS.AREA..1000.ha.              495.324     836.636    89.842   0.90303        0.89933
## LAND.PUT.TO.NONAGRICULTURAL.USE.AREA..1000.ha.  493.804     837.787    88.691   0.90427        0.90024
## Rabi.evapotranspiration.potential.mm    481.879     842.198    84.281   0.90903        0.90483
## Dist.Code                               450.985     857.549    68.929   0.92560        0.91840
## NITROGEN.KHARIF.CONSUMPTION..tons.      444.920     859.556    66.923   0.92777        0.92046
## Zaid.rainfall.mm                        441.380     860.900    65.578   0.92922        0.92174
## CULTIVABLE.WASTE.LAND.AREA..1000.ha.    439.266     861.881    64.597   0.93028        0.92259
## Zaid.temp.max.c                         436.999     862.883    63.596   0.93136        0.92348
## TANKS.AREA..1000.ha.                    436.029     863.571    62.907   0.93210        0.92400
## Rabi.percipitation.mm                   435.113     864.239    62.239   0.93282        0.92449
## Rabi.evapotranspiration.actual.mm       431.318     865.547    60.932   0.93423        0.92577
## Rabi.temp.max.c                         430.942     866.075    60.404   0.93480        0.92611
## Zaid.evapotranspiration.potential.mm    430.759     866.556    59.923   0.93532        0.92639
## Zaid.percipitation.mm                   430.267     867.100    59.378   0.93591        0.92675
## ------------------------------------------------------------------------------------------------------
```

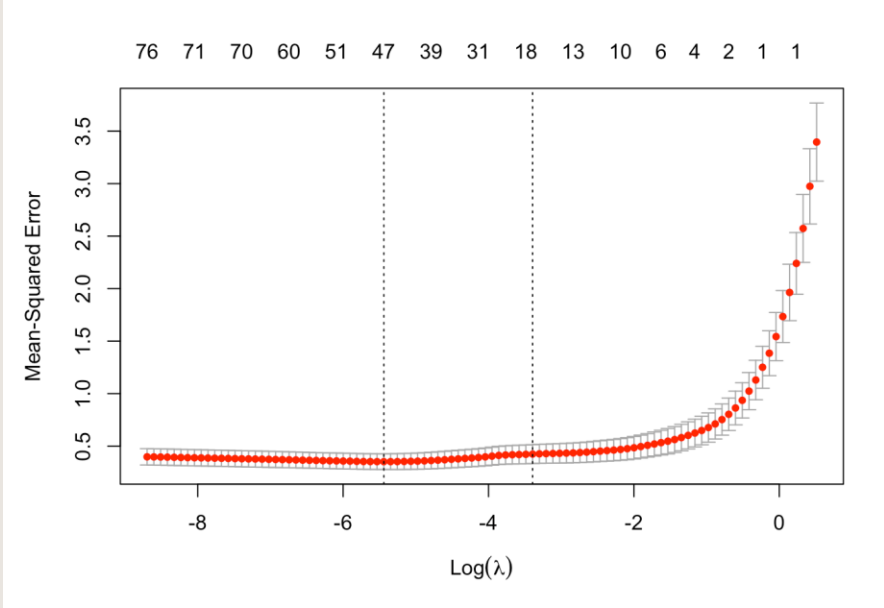
AIC model resulted in a R-squared value of 0.9268, very close to Linear regression. 34 predictors with the most influencing variables being Year, followed by Male Wages and Kharif Precipitation


Residuals vs Fitted
Residuals
Fitted values
228
201
254
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
0.5


Residuals SS vs Predictor
Residuals SS
R-squared vs Predictor
R-squared
Adjusted R-squared vs Predictor
Adjusted R-squared
AIC vs Predictor
AIC
Regression SS vs Predictor
Regression SS
Predictor

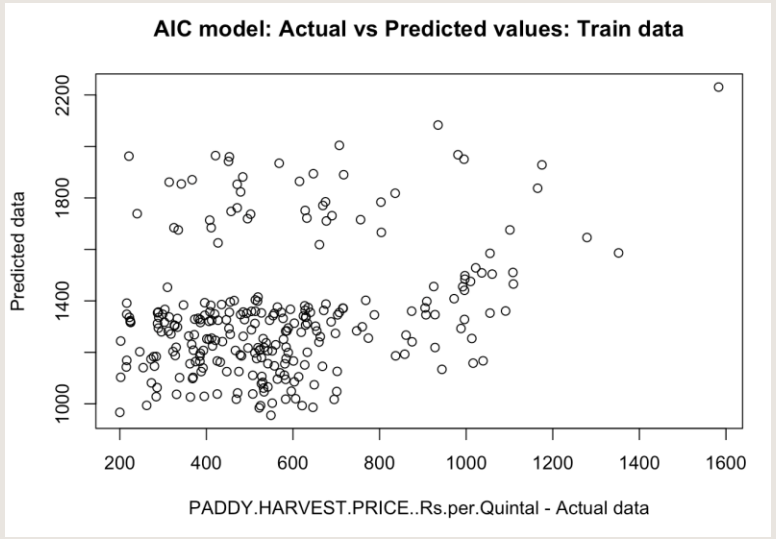
Lasso Regression

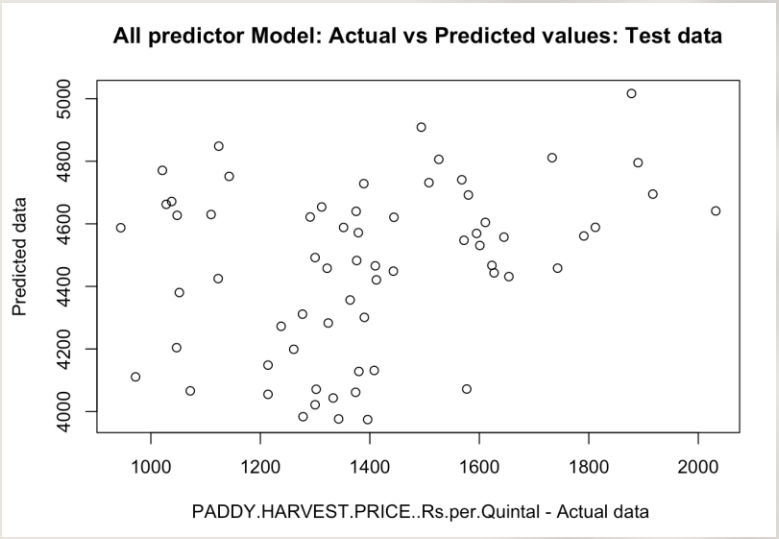
- Variable selection and regularization
- Shrinkage of coefficients
  - Bring coefficients of insignificant predictors to 0

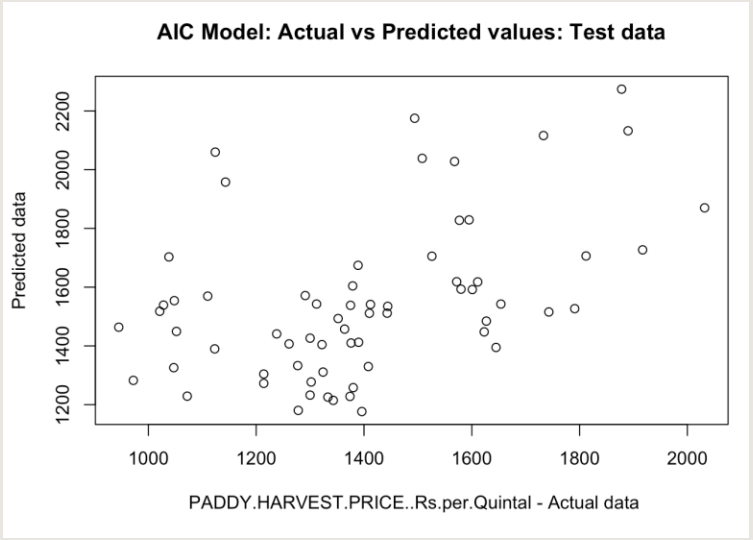


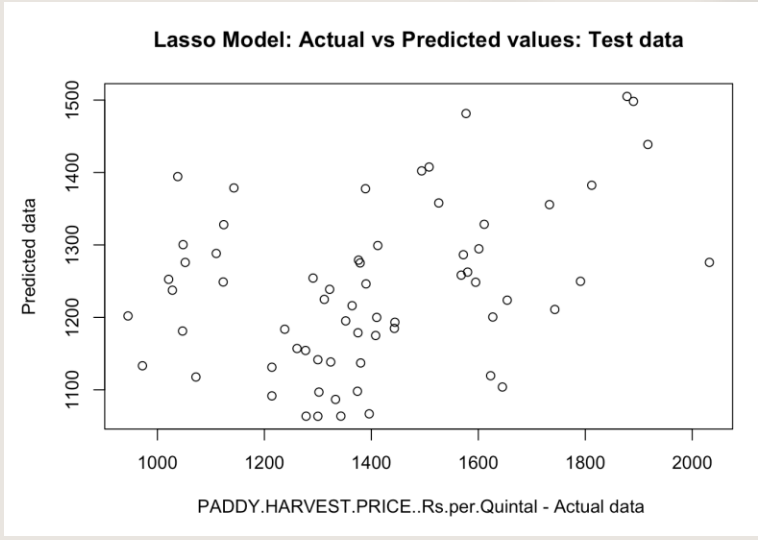
Results

The following are the test results for all three models


AIC model: Actual vs Predicted values: Train data
Predicted data
1000
1400
1800
2200
200
400
600
800
1000
1200
1400
1600
PADDY.HARVEST.PRICE..Rs.per.Quintal - Actual data


All predictor Model: Actual vs Predicted values: Test data
Predicted data
4000
4200
4400
4600
4800
5000
1000
1200
1400
1600
1800
2000
PADDY.HARVEST.PRICE..Rs.per.Quintal - Actual data


AIC Model: Actual vs Predicted values: Test data
Predicted data
1200
1400
1600
1800
2000
2200
1000
1200
1400
1600
1800
2000
PADDY.HARVEST.PRICE..Rs.per.Quintal - Actual data


Lasso Model: Actual vs Predicted values: Test data
Predicted data
1100
1200
1300
1400
1500
1000
1200
1400
1600
1800
2000
PADDY.HARVEST.PRICE..Rs.per.Quintal - Actual data

Conclusion

- RMSEs from models
  - 3076 for Linear Regression
  - 304 for AIC
  - 281 for Lasso
- We were able to find the main influencers of rice price through production was labor wages and amount of precipitation during the Kharif season
- Both AIC and Lasso are accurate models that can help us predict rice harvest price in Andhra Pradesh