

AMES HOUSING data analysis

Project 2 – STAT 897



November 1, 2017

Daljeet singh

Penn State

Table of Contents

[Introduction 1](#_Toc494373628)

[Analysis 1](#_Toc494373629)

[Least squares regression model using all ten predictors 1](#_Toc494373630)

[Best subset selection using BIC to select the number of predictors 2](#_Toc494373631)

[Best subset selection using 10-fold cross-validation to select the number of predictors 3](#_Toc494373632)

[Ridge regression model using 10-fold cross-validation 3](#_Toc494373633)

[Lasso model using 10-fold cross-validation 3](#_Toc494373634)

[Results 4](#_Toc494373635)

[Conclusion 4](#_Toc494373636)

[Appendix i](#_Toc494373637)

# Introduction

# Data Cleansing / Preparation

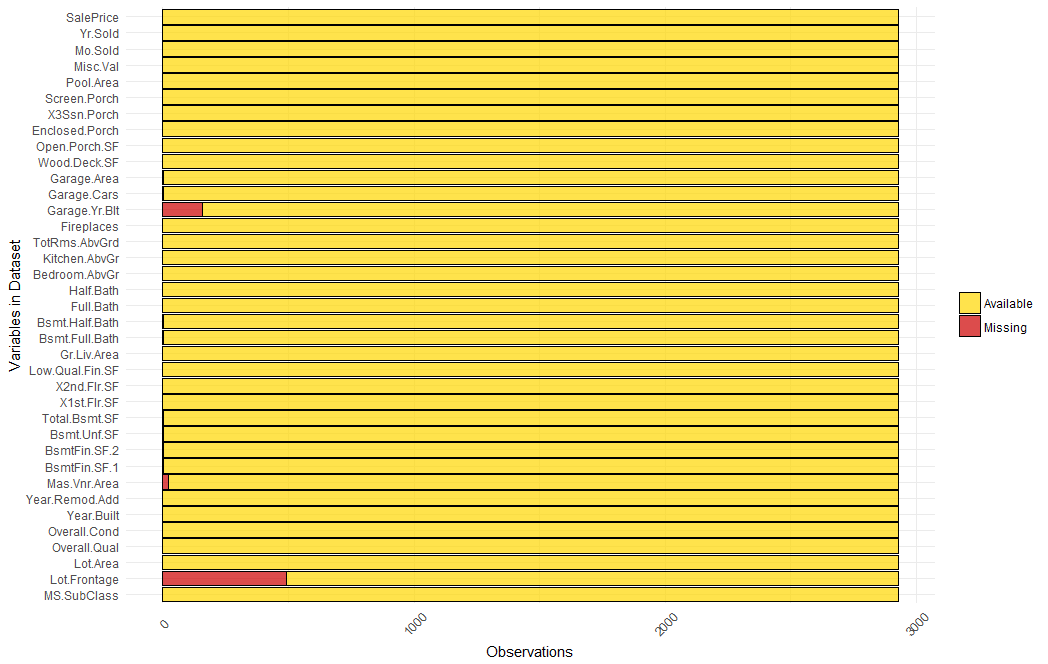
We will be analyzing this dataset with R and the first step was to analyze the data – look for missing values and try to understand the relationships. We see that we have 2930 rows and 82 columns.

As a first step we drop the columns: Order and PID as they serve no purpose in the modeling exercise leaving us with 80 columns.

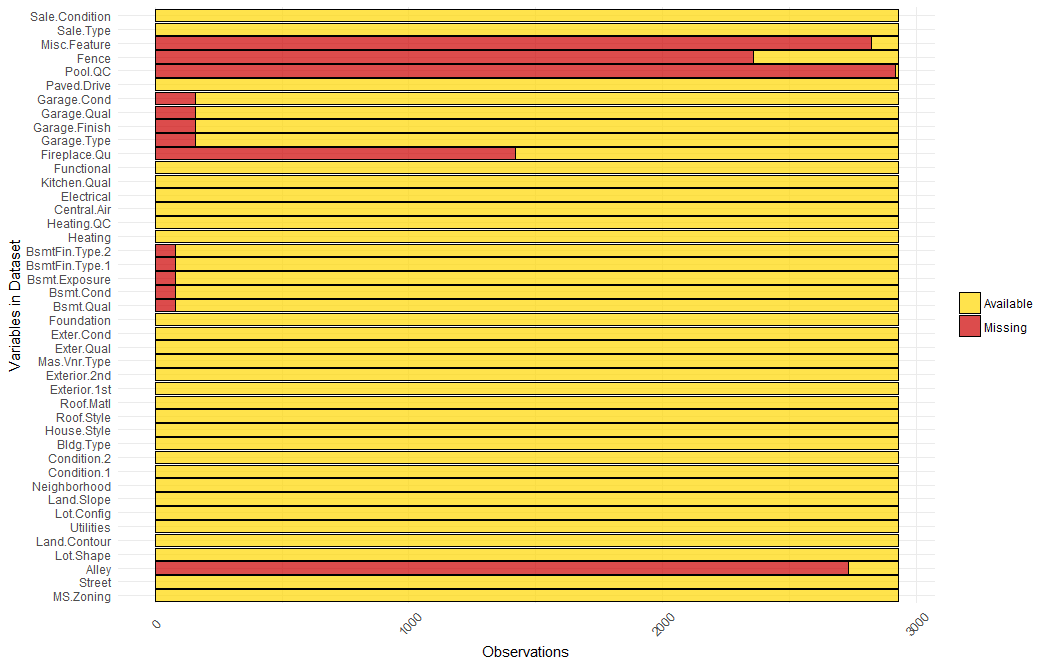
## Handling Missing Values

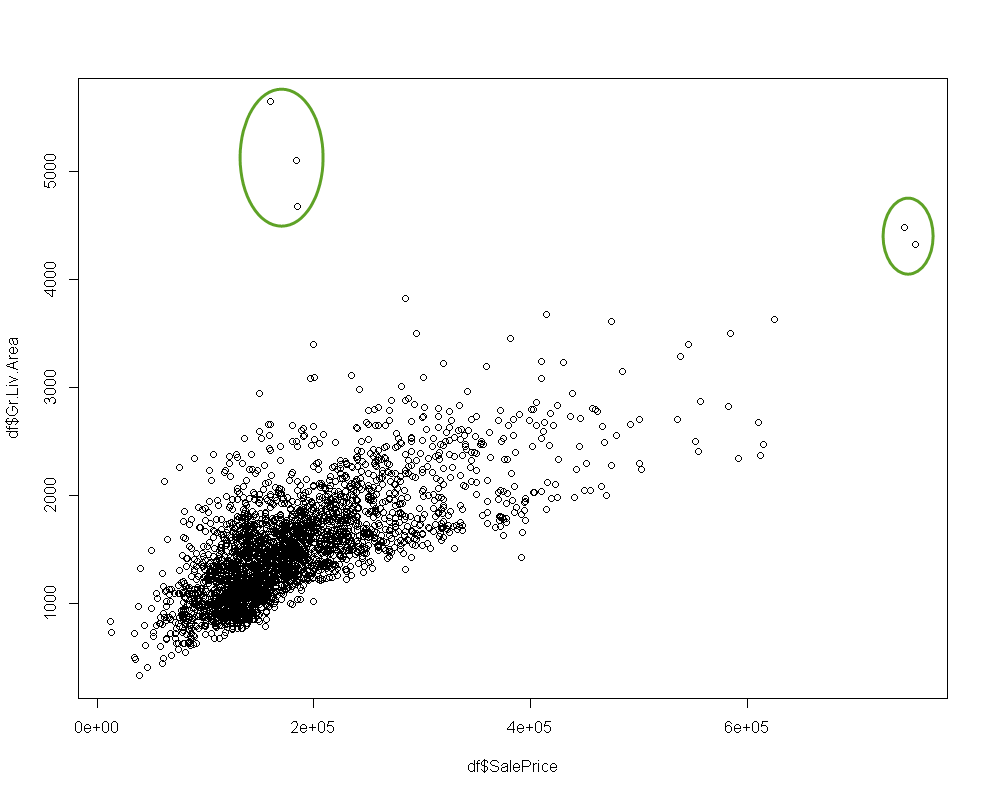
We start to see the breakdown of the missing values in our data set. We find that there are 13960 cells that contain NA (missing values). Let’s see their breakdown within numeric and categorical variables.

Numeric columns – We have 37 numeric columns

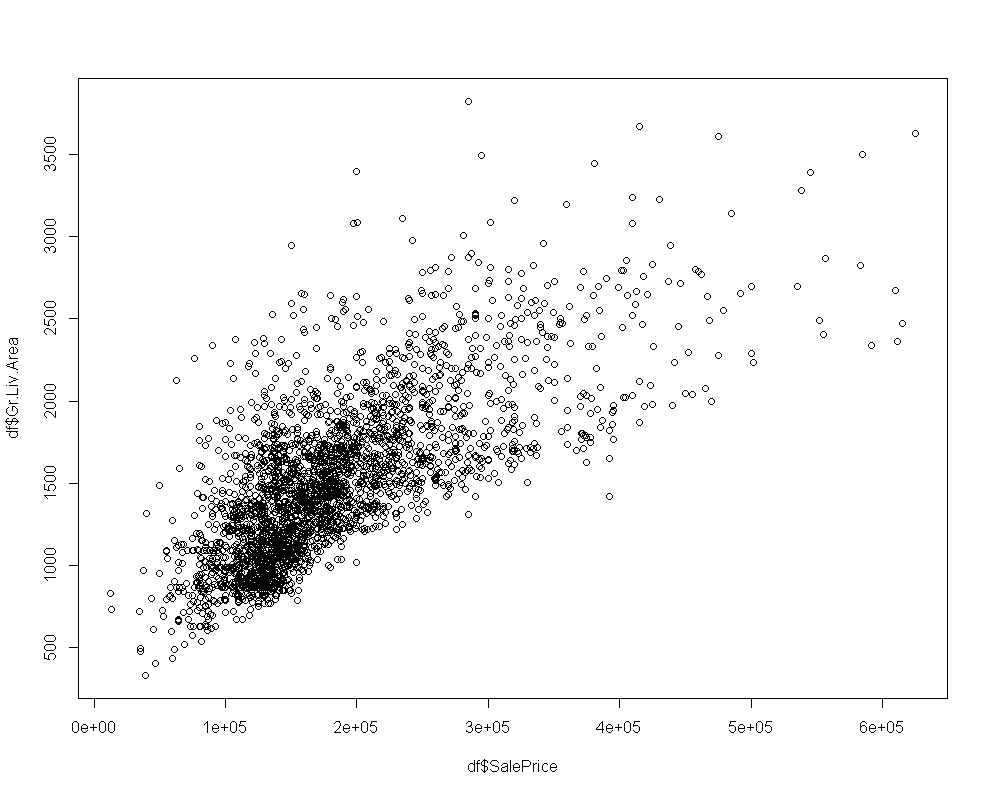


Categorical columns – We have 43 categorical columns



The data documentation mentions that there are 5 outlier rows. Let’s view them on a plot:

We will remove these rows and keep the rows that have Gr.Liv.Area<=4000. After removing we get the plot:



Let’s see the list of columns in the order of the number of missing values:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pool.QC** | **Misc.Feature** | **Alley** | **Fence** | **Fireplace.Qu** | **Lot.Frontage** |
| 2914 | 2820 | 2737 | 2354 | 1422 | 490 |
| **Garage.Yr.Blt** | **Garage.Qual** | **Garage.Cond** | **Garage.Type** | **Garage.Finish** | **Bsmt.Qual** |
| 159 | 158 | 158 | 157 | 157 | 79 |
| **Bsmt.Cond** | **Bsmt.Exposure** | **BsmtFin.Type.1** | **BsmtFin.Type.2** | **Mas.Vnr.Area** | **Bsmt.Full.Bath** |
| 79 | 79 | 79 | 79 | 23 | 2 |
| **Bsmt.Half.Bath** | **BsmtFin.SF.1** | **BsmtFin.SF.2** | **Bsmt.Unf.SF** | **Total.Bsmt.SF** | **Garage.Cars** |
| 2 | 1 | 1 | 1 | 1 | 1 |
| **Garage.Area** |  |  |  |  |  |
| 1 |  |  |  |  |  |

From the data description we can see that some of NA's mean that the house doesn't have the feature. They are not really missing values so we relabel those variables as none. We choose the following variables where the NA is replace with none: "Pool.QC", "Misc.Feature", "Alley", "Bsmt.Qual", "Bsmt.Cond",

"Bsmt.Exposure", "BsmtFin.Type.1", "BsmtFin.Type.2", "Fireplace.Qu",

"Garage.Type", "Garage.Finish", "Garage.Qual", "Garage.Cond", "Fence"

After this we recompute the number of missing values and find them to be substantially less.

We now only have 682 cells with NA.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lot.Frontage** | **Garage.Yr.Blt** | **Mas.Vnr.Area** | **Bsmt.Full.Bath** | **Bsmt.Half.Bath** |
| 490 | 159 | 23 | 2 | 2 |
| **Total.Bsmt.SF** | **Garage.Cars** | **Garage.Area** | **BsmtFin.SF.2** | **Bsmt.Unf.SF** |
| 1 | 1 | 1 | 1 | 1 |

### Zoning Classification

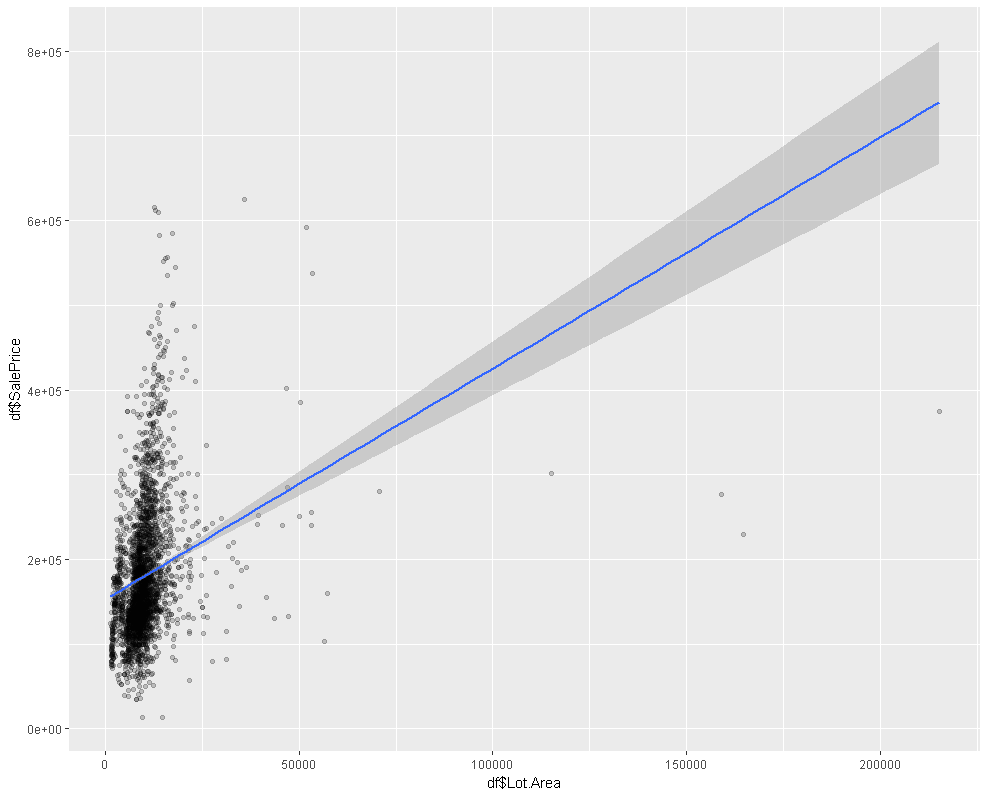
There are 8 different zoning classifications and they range from agricultural to residential. We simplified their names.

A C FV I RH RL RM

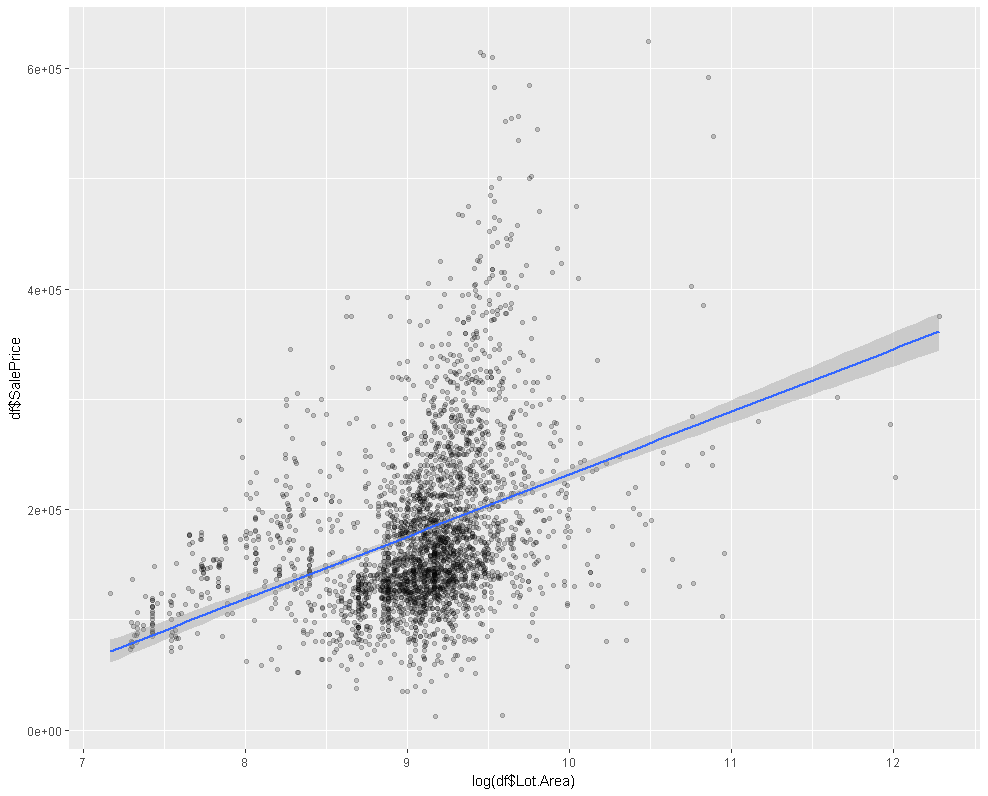
2 25 139 2 27 2268 462

### Lot Area

Here we plot Sale Price against Lot Area and get the following plot:



Clearly we see that we need some transformation in order to perform linear regression. We try to use logarithm of the lot area and plot against the SalePrice to get:



This looks much better and therefore we add this column to the data frame and drop the original lot area.

## MS.Subclass

We dropped the MS.SubClass variable as it is a combination of bldg.type, house.style, and year.built.

## Bath Rooms

Let’s address these columns individually.

Garage.Yr.Blt - It is logical to assume that the garages were built same time with the houses – we find that this is true for 76% of the rows. We set the value to be same as year built.

Mas.Vnr.Area – Here we find that for all the rows where this column in NA we have an empty value for the Mas.Vnr.Type. Mas.Vnr.Type (Masonry veneer type) & Mas.Vnr.Area (Masonry veneer area in square feet) are related to each other - those missing value will be imputed as "None" & "0".

Basement square footage - We observe that Total.Bsmt.SF is a sum of BsmtFin.SF.1 , BsmtFin.SF.2 and Bsmt.Unf.SF

Therefore the columns Total.Bsmt.SF and Bsmt.Unf.SF are overlapping. We will drop Bsmt.Unf.SF

We also observer that Gr.Liv.Area is X1st.Flr.SF+X2nd.Flr.SF+Low.Qual.Fin.SF

length(which(dat$GarageYrBlt == dat$YearBuilt))

### Cleaning Data

#### Neighborhood

 The GrnHill and Landmark neighborhoods are very underrepresented and are missing frontage data for every listed property in the dataset. We dropped these as both neighborhoods only account for three observations in the original dataset.

## Least squares regression model using all ten predictors

The first step is fitting a regular least squares multiple linear regression model using all the ten variables. The coefficients are:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 151.885 3.065 49.555 < 2e-16 \*\*\*

age -6.387 71.461 -0.089 0.928837

sex -257.173 72.951 -3.525 0.000485 \*\*\*

bmi 513.830 78.892 6.513 2.84e-10 \*\*\*

map 335.714 77.309 4.342 1.89e-05 \*\*\*

tc -779.357 507.431 -1.536 0.125550

ldl 481.739 407.534 1.182 0.238047

hdl 85.036 262.514 0.324 0.746203

tch 262.487 197.443 1.329 0.184650

ltg 649.500 205.962 3.153 0.001766 \*\*

glu 117.226 76.871 1.525 0.128252

A lot of variables are not significant (using an alpha of 0.05). Interestingly we also find that age, tc, ldl, hdl, tch and glu are not significant predictors of disease progression. This is interesting because these variables intuitively appear that they should be significant for instance age.

The plots of the residuals:

|  |  |
| --- | --- |
|  |  |

The plots look reasonable. There is a slight non linearity and the variance appears to be varying but both the things appear to be within reasonable limits.

**Test MSE for the full linear regression** model is: 2511.981

## Best subset selection using BIC to select the number of predictors

Next, we will apply the best subset selection using BIC to select the number of predictors.

|  |  |
| --- | --- |
|  | The lowest and therefore the best BIC is for 5 predictors. The BIC value for a model with 5 predictors is -194.7 |

The respective coefficients are:

|  |  |
| --- | --- |
|  | Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 151.801 3.067 49.491 < 2e-16 \*\*\*  sex -242.577 72.001 -3.369 0.000845 \*\*\*  bmi 530.590 76.902 6.900 2.72e-11 \*\*\*  map 346.631 74.405 4.659 4.64e-06 \*\*\*  hdl -353.984 80.206 -4.413 1.38e-05 \*\*\*  ltg 426.250 78.251 5.447 1.01e-07 \*\*\* |

**Test MSE for the best subset** model based on BIC with 5 parameters is: 2506.565

## Best subset selection using 10-fold cross-validation to select the number of predictors

When we use the cross-validation approach with 10 folds, the training data is divided into 10 parts and the approach loops through, each time using 1-fold as validation data and the remaining as training data. With this approach, too we get the minimum cross validation error for the same 5 parameter model as above. The coefficients are listed below (they will be same as the last output).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | (Intercept) | 151.8015 | | sex | -242.577 | | bmi | 530.5897 | | map | 346.6307 | | hdl | -353.984 | | ltg | 426.25 | |

**Test MSE for the best subset** model based on 10-fold cross validation is: 2506.565

## Ridge regression model using 10-fold cross-validation

We will now move to the ridge regression approach. We used 10-fold cross validation and selected the largest value of λ such that the cross-validation error is within 1 standard error of the minimum.

|  |  |
| --- | --- |
|  | Largest value of λ such that the cross-validation error is within 1 standard error of the minimum = **60.7213** |

The coefficients with ridge are:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | age | sex | bmi | map | tc | ldl | hdl | tch | ltg | glu |
| 152.10 | 29.72 | -105.00 | 333.38 | 221.14 | 4.64 | -20.74 | -178.61 | 148.64 | 255.54 | 133.29 |

**Test MSE for the ridge regression** model is: 2852.122

## Lasso model using 10-fold cross-validation

In the last step, we will use lasso model. We used 10-fold cross validation and selected the largest value of λ such that the cross-validation error is within 1 standard error of the minimum.

|  |  |
| --- | --- |
|  | Largest value of λ such that the cross-validation error is within 1 standard error of the minimum = **7.662159** |

The coefficients with ridge are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | bmi | map | hdl | ltg | glu |
| 152.332 | 491 | 192.21 | -146.9 | 381.2 | 24 |

**Test MSE for the lasso regression** model is: 2599.714

# Results

|  |  |
| --- | --- |
| Model | Test MSE |
| Full linear regression | 2511.981 |
| Best subset model based on BIC | 2506.565 |
| **Best subset model based on 10-fold cross validation** | **2506.565** |
| Ridge Regression | 2852.122 |
| Lasso Regression | 2599.714 |

# Conclusion

Based on the Test Mean squared error, we find that the best subset models behave the best. Lasso and full linear regression are very close while ridge regression performs the worst. In conclusion, the recommended model is the best subset model for the given analysis.

# Appendix

The code for the project is as follows:

---

title: "Stat 897 Fall 2017 Project 1"

author: "Penn State"

date: "Due October 1, 2017"

output: pdf\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

# Linear Regression, Variable Selection, Ridge Regression, Lasso

This project is to be completed individually. You may submit pdf only (Rmd is not needed and you can use another word processing tool if you like).

The diabetes data in Efron et al. (2003) will be used: ten baseline variables:

age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. The data is available in R package `lars`.

Load the data:

```{r}

#install.packages('lars')

library(lars)

library(leaps)

library(glmnet)

data(diabetes)

data.all = data.frame(cbind(diabetes$x, y = diabetes$y)) # change to normal formatting

x=model.matrix (y ~ ., data = data.all)[,-1]

y=data.all$y

```

Partition the patients into two groups: training (~75%) and test (~25%). Please use use the random number generator seed specified below before randomly splitting the data. Since you will not submit a markdown file, please use the set.seed at each noted place.

```{r}

seed = 38723

set.seed(seed) # set random number generator seed to enable reproducibility of results

test=sample (nrow(data.all), round(nrow(data.all)\*.25), replace = FALSE)

sum(test)

train=(-test)

data.train = data.all[train,]

data.test = data.all[test,]

```

# Project Requirements

Write up your results in a professional report, like you would present to a client or internal customer for your analysis. The report should be no more than 4 single-spaced pages long and submitted in PDF format.

It should include coefficient estimates for each model and test data mean prediction errors.

Include any other details from your analysis that you feel are worthy of mention.

The report should have sections (e.g., Introduction, Analysis, Results, Conclusion) and provide sufficient details that anyone with a reasonable statistics background could understand exactly what youâ™ve done and what you concluded.

Consider using tables and figures to enhance your report. You might use the package "pander" if you are using Rmarkdown for nicely formatted tables.

Do not embed R code in the body of your report (if you are using rmarkdown, use {r echo=FALSE} to supress the printing of the r code), but instead attach the code (code only, not output) in an appendix. The appendix does not count towards the page limit.

## Grading criteria (out of 25)

15 points: fulfilling the project requirements and matching results exactly (this is why you should use the specific random number generator seeds).

10 points: the quality of your report (including: clarity of writing, organization, and layout; appropriate use of tables and figures; careful proof-reading; adherence to report guidelines

## Fit the following models to the training set. For each model extract the model coefficient estimates and calculate the "mean prediction error" in the test set.

1. Least squares regression model using all ten predictors.

Let's first analyze with the least squares regression model.

```{r}

lm.fit = lm(y~., data=data.train)

summary(lm.fit)

confint(lm.fit)

plot(lm.fit)

```

```{r}

lm.predict = predict(lm.fit, data.test)

mean((lm.predict - y[test])^2)

```

2. Apply best subset selection using BIC to select the number of predictors.

Next we will perform the best subset selection using BIC

```{r}

set.seed(seed)

regfit.full=regsubsets (y~., data=data.train, nvmax =10)

reg.summary =summary (regfit.full)

reg.summary$bic

plot(reg.summary$bic, xlab =" Number of Variables ", ylab=" BIC ",type="l")

min\_bic\_pt = which.min(reg.summary$bic)

points (min\_bic\_pt, reg.summary$bic[min\_bic\_pt], col ="red",cex =2, pch =20)

min(reg.summary$bic)

```

We find that BIC is minimum for model with following parameters:

```{r}

plot(regfit.full ,scale ="bic")

coef(regfit.full, min\_bic\_pt)

names(coef(regfit.full,5))

summary(lm(y~sex+bmi+map+hdl+ltg, data = data.train))

test.mat=model.matrix (y~.,data=data.test)

test.val.errors =rep(NA ,10)

for(i in 1:10){

coefi=coef(regfit.full, id=i)

pred=test.mat [,names(coefi)] %\*% coefi

test.val.errors [i]= mean(( data.test$y-pred)^2)

}

plot(test.val.errors ,type='b', xlab='# of parameters', ylab='Test MSE')

test.val.errors[5]

```

3. Apply best subset selection using 10-fold cross-validation to select the number of predictors. Please use a random number seed of 38723 immediately before entering the command.

```{r}

predict.regsubsets =function (object ,newdata ,id ,...){

form=as.formula (object$call [[2]])

mat=model.matrix (form ,newdata )

coefi =coef(object ,id=id)

xvars =names (coefi )

mat[,xvars ]%\*% coefi

}

k=10

set.seed (seed)

folds=sample (1:k,nrow(data.train), replace =TRUE)

cv.errors =matrix (NA ,k, 10, dimnames =list(NULL , paste (1:10) ))

for(j in 1:k){

best.fit =regsubsets (y~.,data=data.train [folds !=j,], nvmax =10)

for(i in 1:10) {

pred=predict (best.fit, data.train [folds ==j,], id=i)

cv.errors [j,i]=mean( (data.train$y[folds ==j]-pred)^2)

}

}

mean.cv.errors =apply(cv.errors ,2, mean)

mean.cv.errors

par(mfrow =c(1,1))

plot(mean.cv.errors ,type='b')

which.min(mean.cv.errors)

points (which.min(mean.cv.errors), min(mean.cv.errors), col ="red",cex =2, pch =20)

best.fit =regsubsets (y~., data=data.train, nvmax =10)

coefi = coef(best.fit, which.min(mean.cv.errors))

coefi

pred=test.mat [,names(coefi)] %\*% coefi

mean(( data.test$y-pred)^2)

```

4. Ridge regression model using 10-fold cross-validation to select the largest value of $\lambda$ such that the cross-validation error is within 1 standard error of the minimum (R functions glmnet and cv.glmnet in package glmnet). Please use a random number seed of 38723 immediately before entering the command.

```{r}

set.seed (seed)

cv.out =cv.glmnet (x[train,],y[train],alpha =0)

plot(cv.out)

bestlam.min =cv.out$lambda.min

bestlam.min

bestlam.1se =cv.out$lambda.1se

bestlam.1se

grid = 10^seq(10,-2,length=100)

ridge.mod =glmnet (x[train ,],y[train],alpha =0, lambda=grid, thresh =1e-12)

ridge.pred=predict (ridge.mod, s=bestlam.min, newx=x[test ,])

mean(( ridge.pred - y[test])^2)

ridge.pred=predict (ridge.mod, s=bestlam.1se, newx=x[test ,])

mean(( ridge.pred - y[test])^2)

ridge.coef=predict (ridge.mod, type ="coefficients", s=bestlam.1se )[0:11,]

ridge.coef

```

5. Lasso model using 10-fold cross-validation to select the largest value of $\lambda$ such that the cross-validation error is within 1 standard error of the minimum (R functions glmnet and cv.glmnet in package glmnet). Please use a random number seed of 38723 immediately before entering the command.

```{r}

set.seed (seed)

cv.out =cv.glmnet (x[train,],y[train],alpha =1)

plot(cv.out)

bestlam.min =cv.out$lambda.min

bestlam.min

bestlam.1se =cv.out$lambda.1se

bestlam.1se

lasso.mod =glmnet (x[train ,],y[train],alpha =1, thresh =1e-12)

lasso.pred=predict (lasso.mod, s=bestlam.min, newx=x[test ,])

mean(( lasso.pred - y[test])^2)

lasso.pred=predict (lasso.mod, s=bestlam.1se, newx=x[test ,])

mean(( lasso.pred - y[test])^2)

lasso.coef=predict (lasso.mod, type ="coefficients", s=bestlam.1se )[0:11,]

lasso.coef[lasso.coef !=0]

```