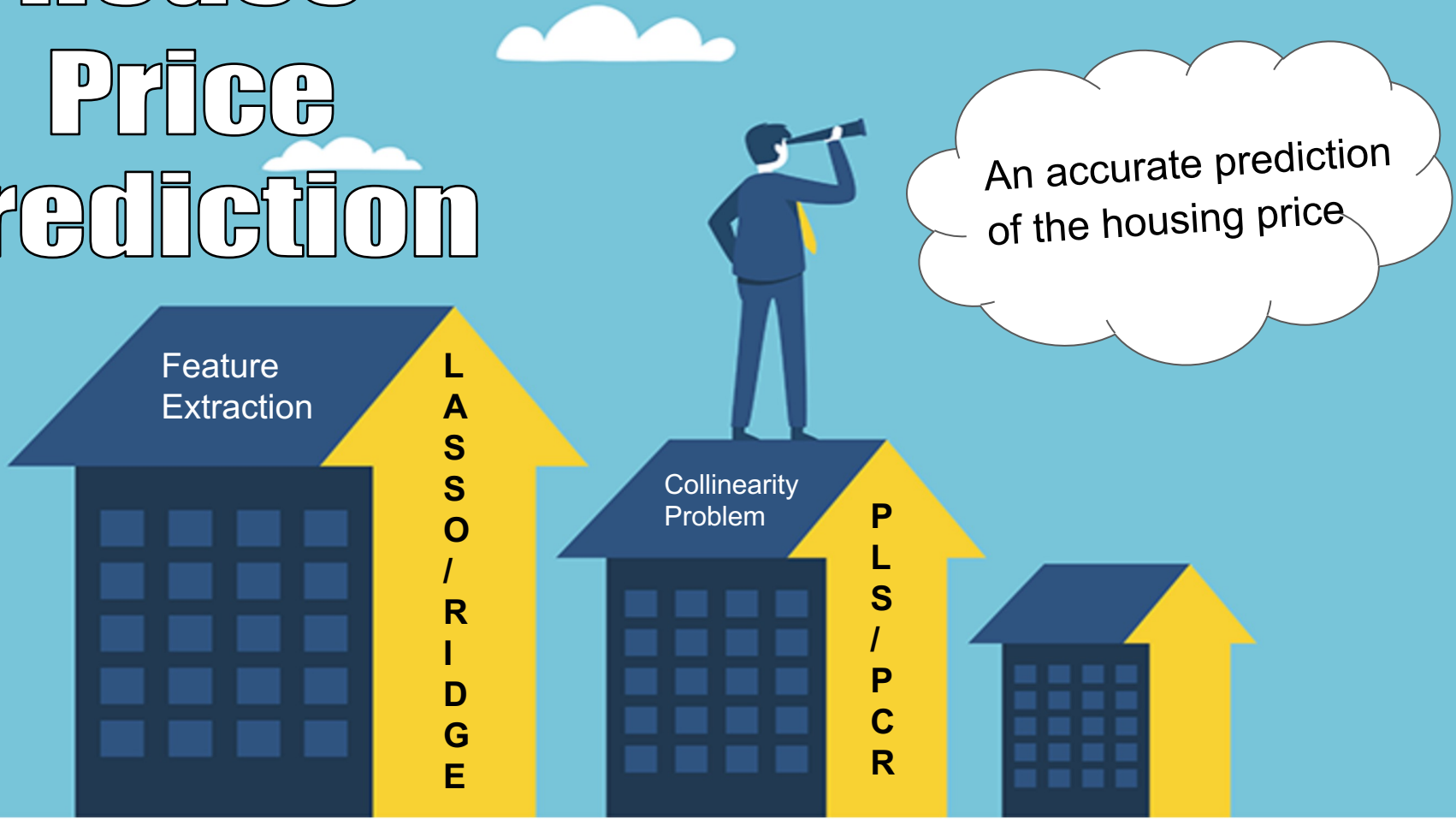
House Price Prediction



Model Selection

Feature Extraction

```
print(dim(train))
```

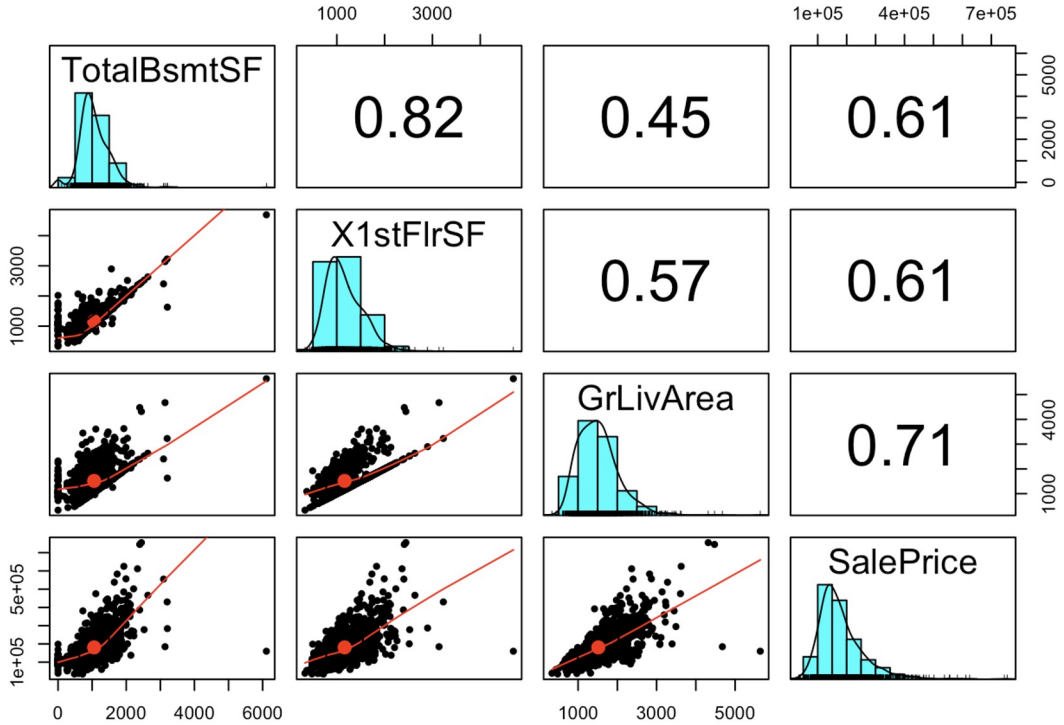
```
## [1] 1460   81
```

	SalePrice
LotFrontage	0.33477085
LotArea	0.26384335
MasVnrArea	0.47261450
BsmtFinSF1	0.38641981
BsmtFinSF2	-0.01137812
BsmtUnfSF	0.21447911
TotalBsmtSF	0.61358055
X1stFlrSF	0.60585218
X2ndFlrSF	0.31933380
LowQualFinSF	-0.02560613
GrLivArea	0.70862448
BsmtFullBath	0.22712223
BsmtHalfBath	-0.01684415
FullBath	0.56066376
HalfBath	0.28410768
BedroomAbvGr	0.16821315
KitchenAbvGr	-0.13590737
TotRmsAbvGrd	0.53372316
Fireplaces	0.46692884
GarageCars	0.64040920
GarageArea	0.62343144
WoodDeckSF	0.32441344
OpenPorchSF	0.31585623
EnclosedPorch	-0.12857796
X3SsnPorch	0.04458367
ScreenPorch	0.11144657
PoolArea	0.09240355
MiscVal	-0.02118958
SalePrice	1.00000000

➡ **Lasso / Ridge**

Model Selection

Collinearity Problem



PLS / PCR

Model Assumption

- Lasso and Ridge are special cases of the General Linear Model
- Assumptions need to diagnose

The normal linear regression model assumes:

$$Y_i = \beta_o + \sum_{k=1}^{p} \beta_k X_k + e_i$$

$$e_i \overset{iid}{\sim} N(0, \sigma^2)$$

(1) Mean Function: $E(e_i \mid X) = 0$.

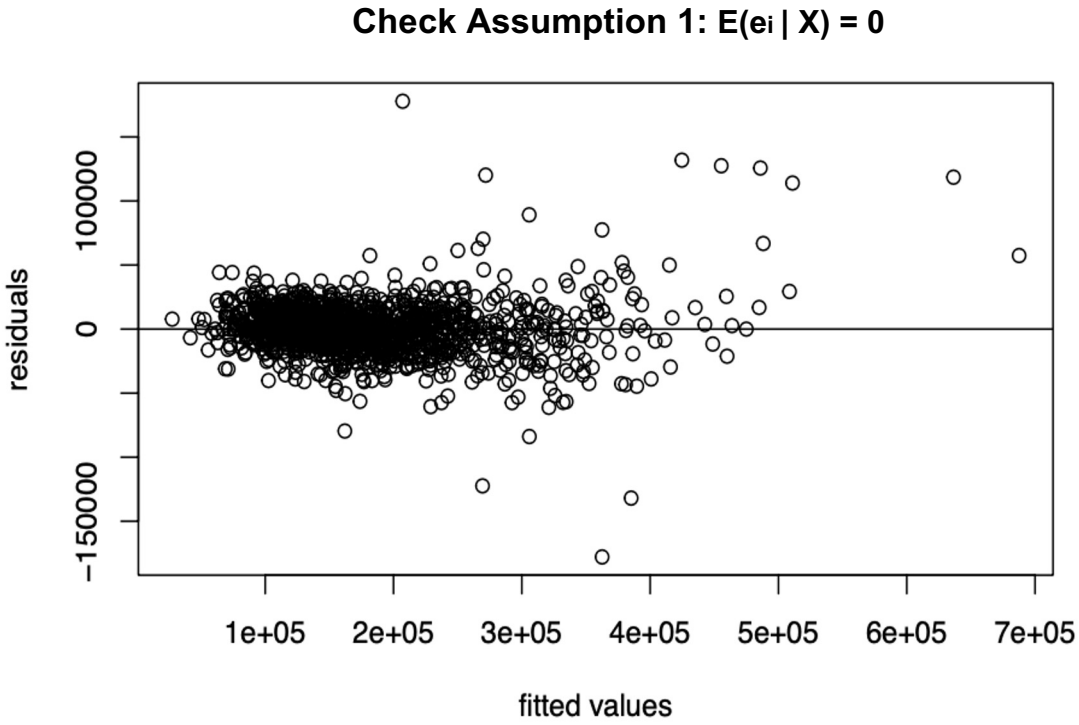
(2) Variance Function: $Var(e_i \mid X) = \sigma^2$.

(3) Normality of the errors.

(4) Independence of the errors.

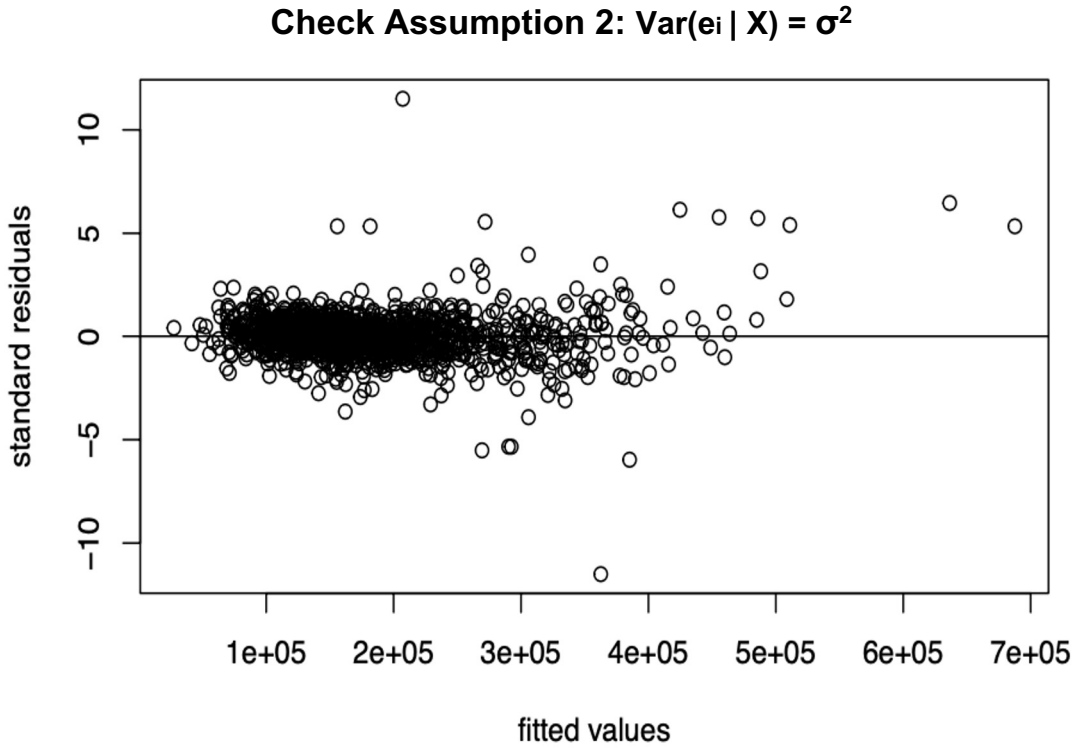
(5) Little/No Multicollinearity in data.

Model Assumption



- Most dots are around 0.
- The fitted mean function is appropriate.

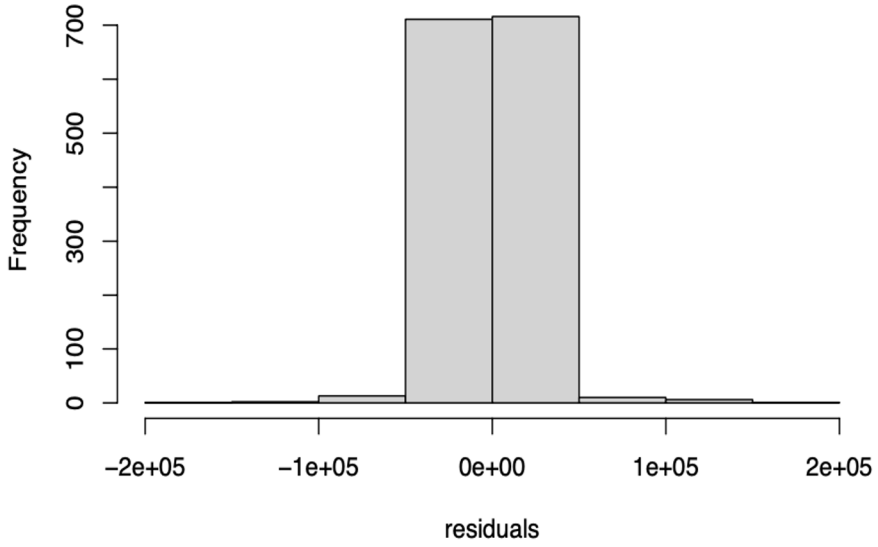
Model Assumption

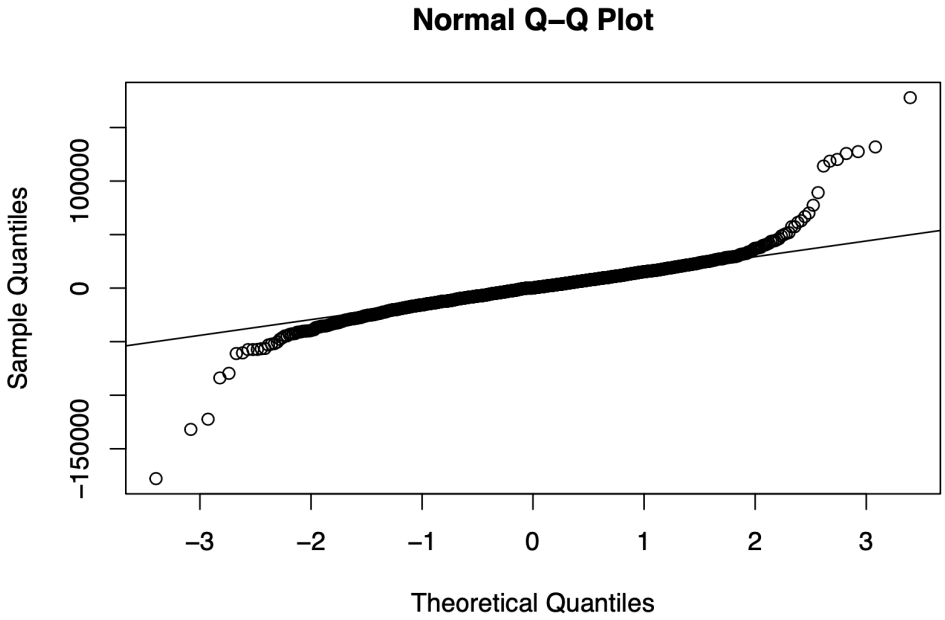


- Most dots are around 0 (a constant).
- The fitted variance function is appropriate.

Model Assumption

Check Assumption 3: Normality



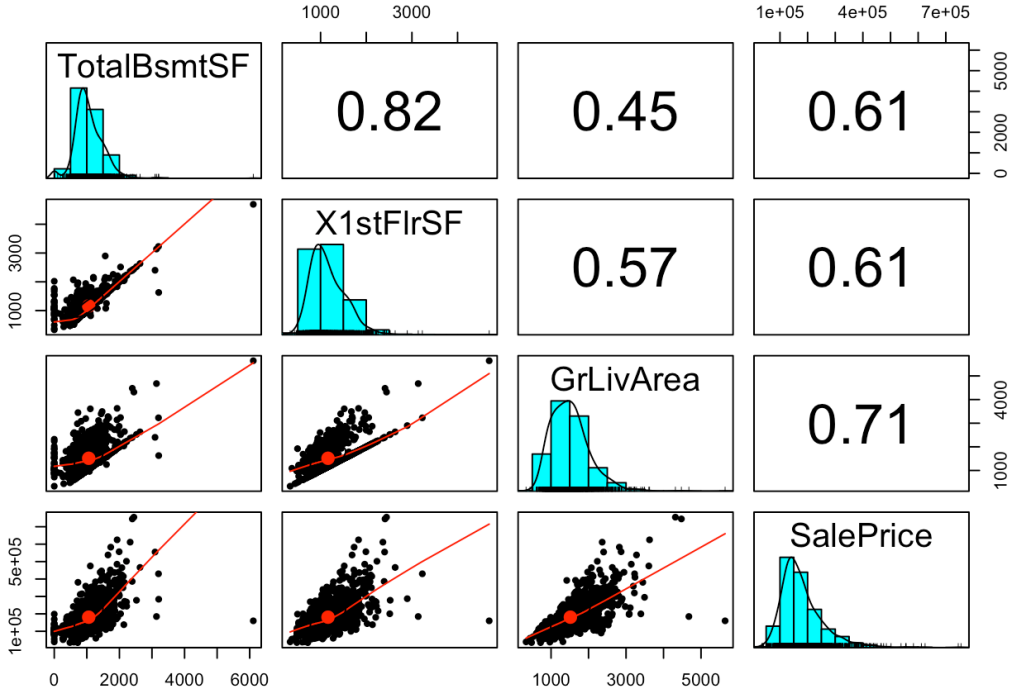


- Points on the lower-end have lower measurement than the Normal model predicts and the points on the upper-end have higher measurement than the Normal model predicts.
- Most points are approximately on the line. It might be due to the outliers.
- Thus, the residuals are normally distributed.

Model Assumption

Check Assumption 4: Independence of Errors. --> No way to check, we assume it is correct here.

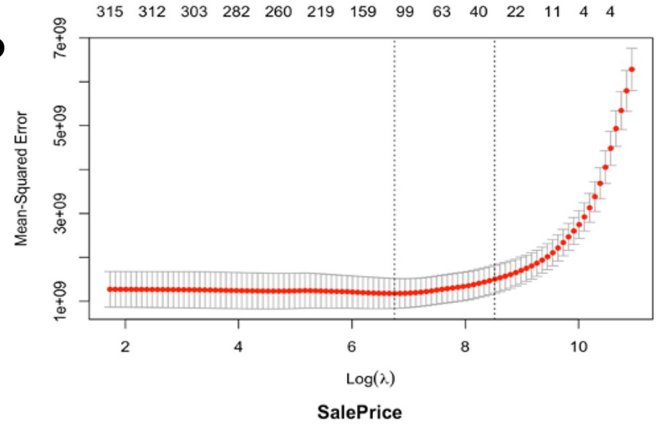
Check Assumption 5: Little / No Multicollinearity



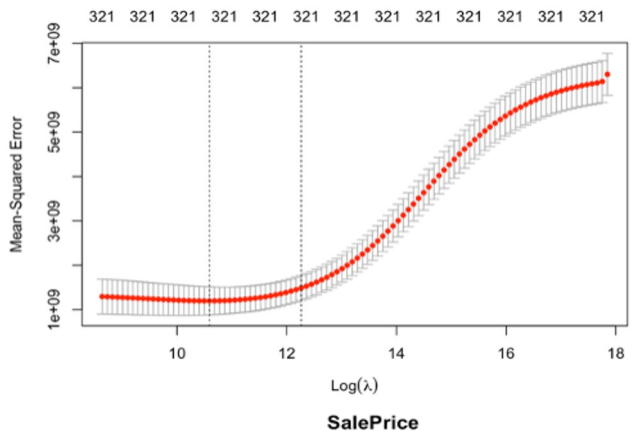
- There are some collinearity problems

Model Analysis

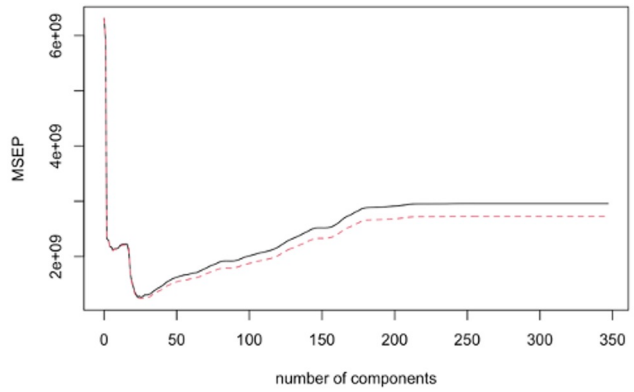
LASSO



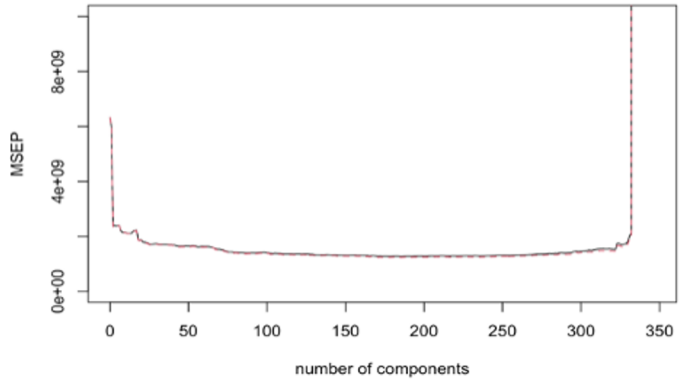
Ridge



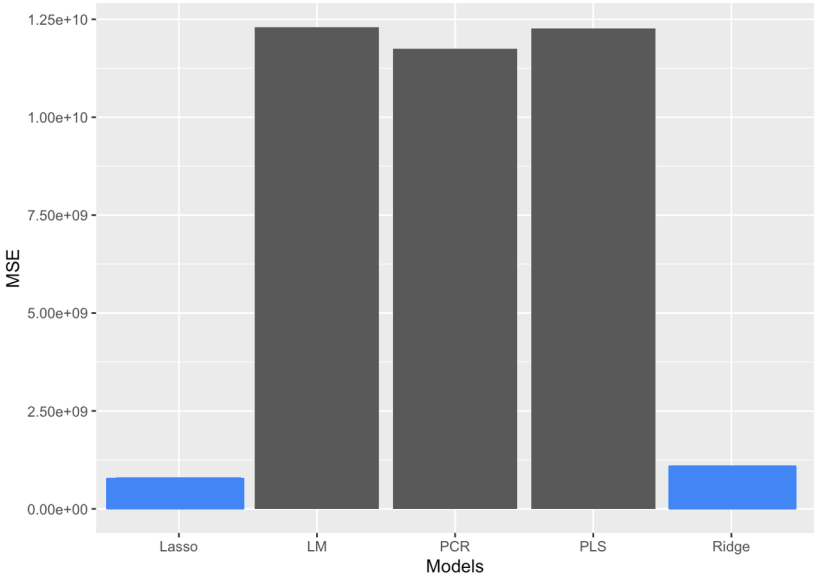
PCR

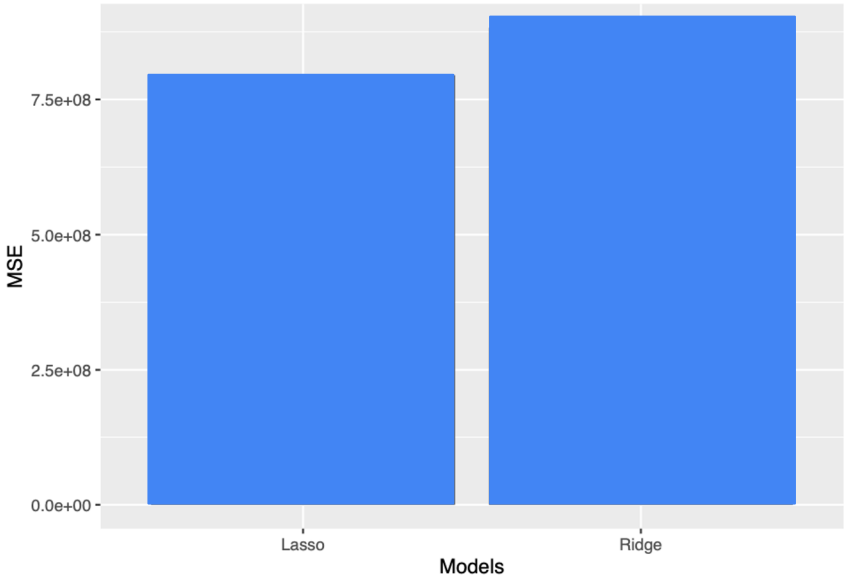


PLS



Models Comparison - MSE





(1) **Linear Regression Model** has larger MSE than **Lasso and Ridge Regression**.
(2) **PLS** and **PCR** have comparably similar MSE, while **Lasso and Ridge Regression** have comparably similar MSE.
(3) **Lasso and Ridge Regression**'s MSE much smaller than **PLS** and **PCR**'s MSE.
(4) **Lasso** has the lowest MSE value.