Akanksha Bansal **Assignment 2**
Data Mining **Feb 20, 2014**

Problem 1 Sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\Rightarrow \epsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$n \geq \hat{p} * (1-\hat{p})(\frac{z_{\frac{\alpha}{2}}}{\epsilon})^2$

$n \geq p * (1-p)(\frac{z_{\frac{\alpha}{2}}}{\epsilon})^2$

Part a. Replacing $p = 0.5$ in the above equation $\alpha = 0.1$ thus $z_{\frac{\alpha}{2}} = 1.645$ and $\epsilon = 0.02$

$n = 0.5 * (0.5)(\frac{1.645}{0.02})^2 = 1691.3$

$n = 0.95 * (0.05)(\frac{1.645}{0.02})^2 = 321.34$

Part b. For $\alpha = 0.1$ and given p, ϵ we have $n = 1000$. $\Rightarrow z_{\frac{\alpha}{2}} = 1.645$ For $\alpha = 0.05$ thus $z_{\frac{\alpha}{2}} = 1.96$ therefore value of n will change by $(\frac{z_{0.05/2}}{z_{0.1/2}})^2$.

$n = 1000 * (\frac{1.96}{1.645})^2 = 1419.64$

For $\epsilon = 0.02$ and given p, α we have $n = 1000$. For $\epsilon = 0.01$ therefore value of n will change by $(\frac{0.02}{0.01})^2$.

$n = 1000 * (\frac{0.02}{0.01})^2 = 4000$

Part c. Replace ϵ with $\gamma\hat{p}$

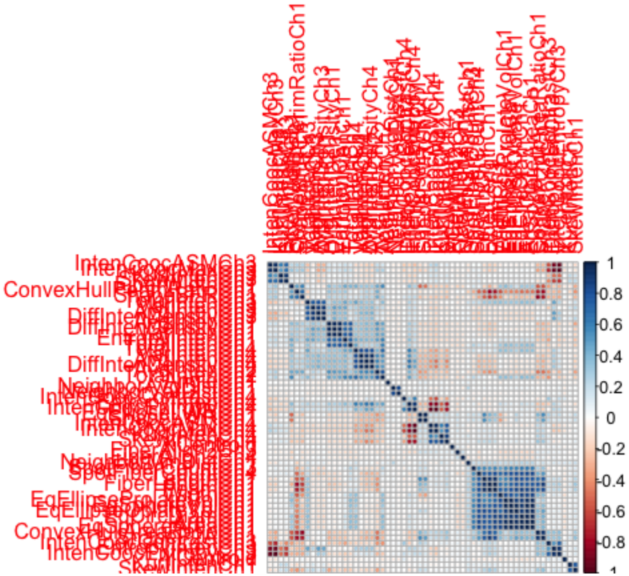
$\gamma\hat{p} = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\gamma^2 p^2 = \frac{p(1-p)}{n} {z_{\frac{\alpha}{2}}}^2$

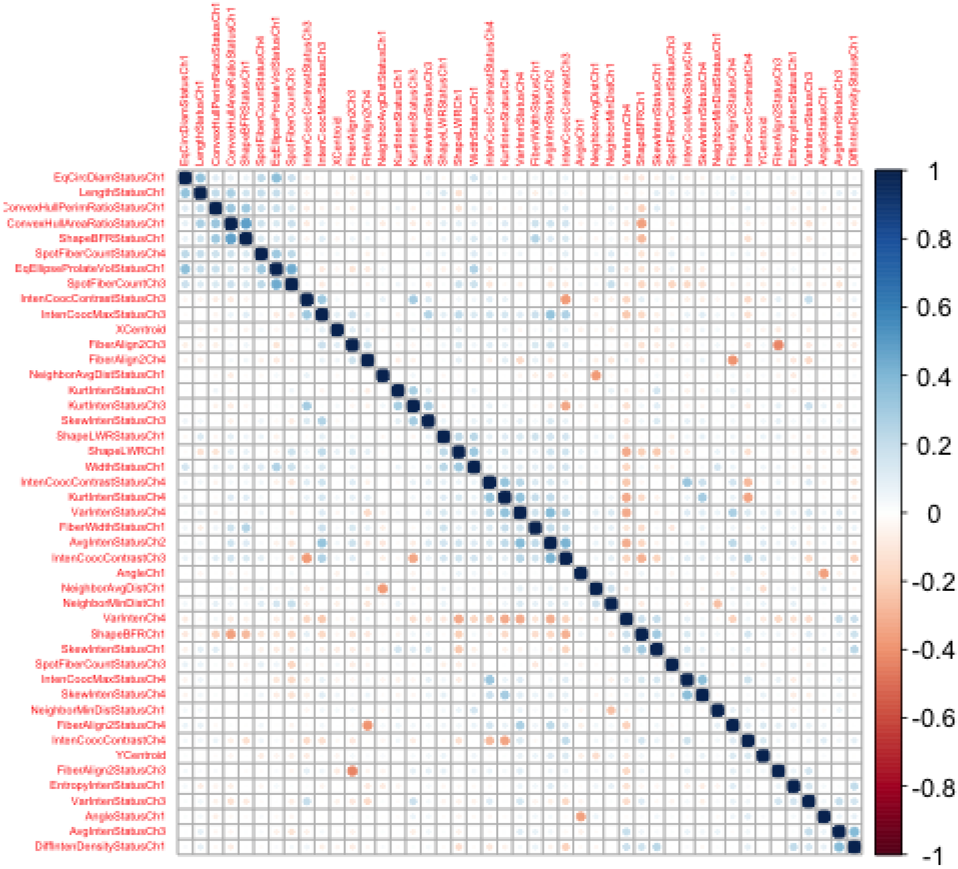
$n \geq \frac{(1-p)}{p}(\frac{z_{\alpha/2}}{\gamma})^2$

Problem 2 a. When the cutoff is set to 0.5, 69/113 features are eliminated form the model.
When the cutoff is set to 0.75, 43/113 features are eliminated form the model.
When the cutoff is set to 0, 113/113 features are eliminated form the model.

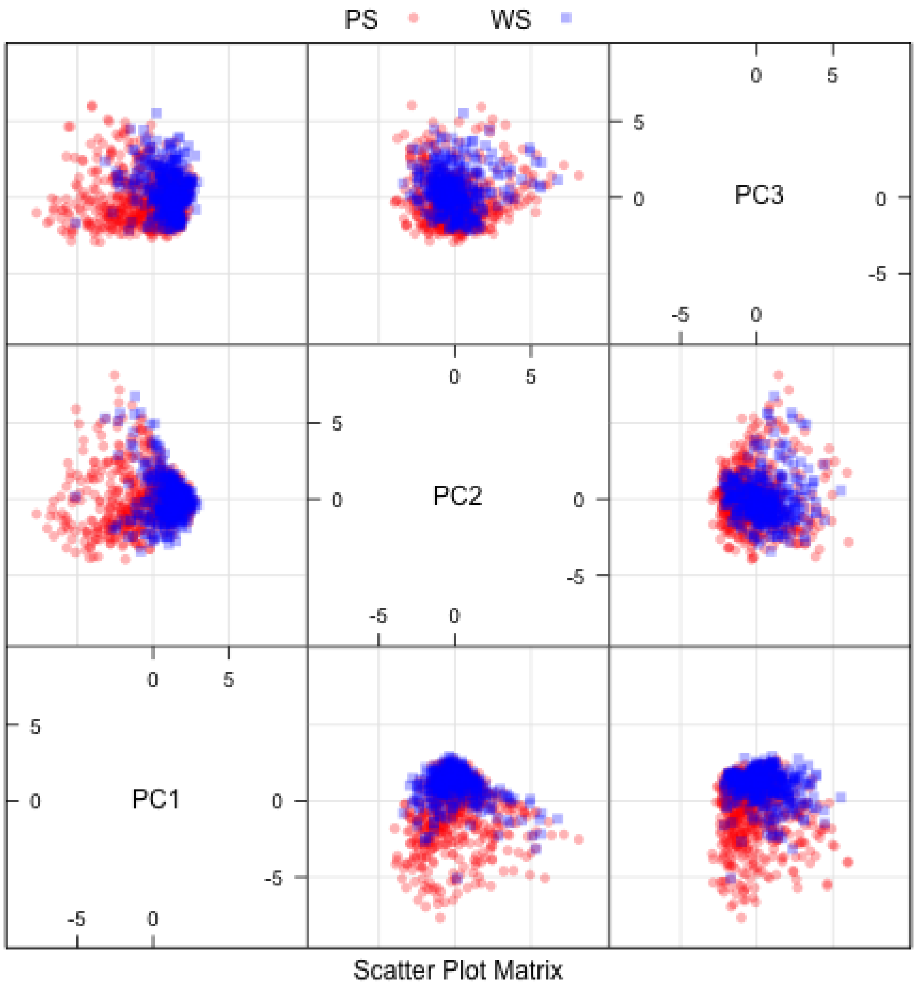
Now the top three Principal Components have a reduction in overlap. The values are still centered around 0, thus small part of the variance is now being explained captured by them. The Well segmented(WS) Cells are completely contained within the poorly segmented(PS) cells. Even in this image there are no significant outliers being shown. The graph with PC2 and PC3 doesn't show any kind of separation between WS cells and PS cells. Thus we can conclude that PC1 might be the only feature which can distinguish to some extent the WS vs PS cells.



Original Correlation matrix generated in the book. (cutoff = 0.75)



Original PCA components generated from the code of the book.



Plot of the First There components of Filtered data.

Problem 3 PCA and FLDA are both used to reduce the dimension space in which the given data set can be represented. They both allow the data projection into a smaller dimension space while still being able to represent majority of the information (in some cases all) present in the original data. For both these models, eigen values represent the variation explained by their components. The feature space determined by PCA tries to maximize the data variance(variance of the overall data set). FLDA on the other hand tried to determine those features where the separation of two or more classes is maximized. Thus Inter class variance is maximized in FLDA.

Problem 4

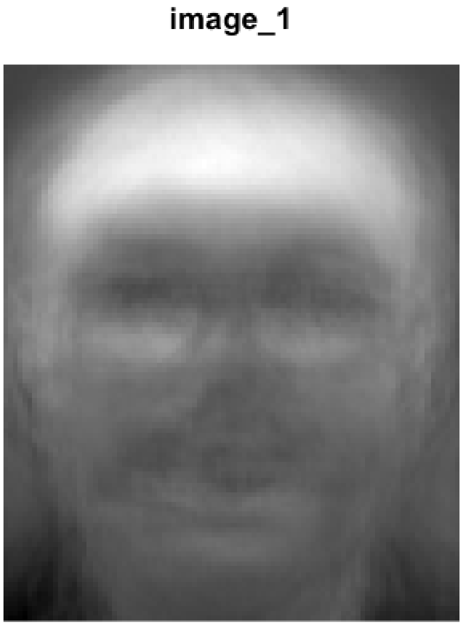


image_1

[k=1]

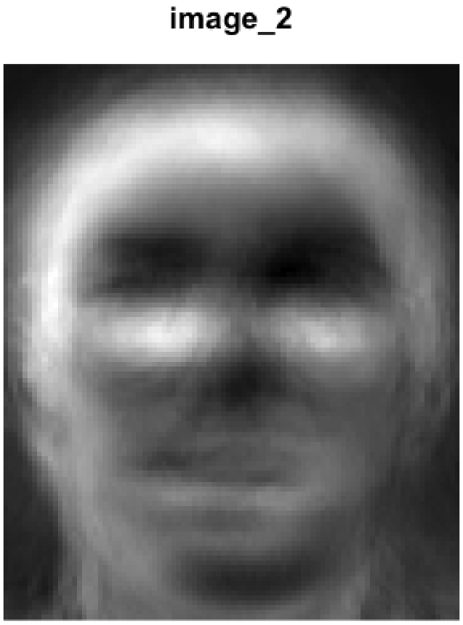


image_2

[k=10]

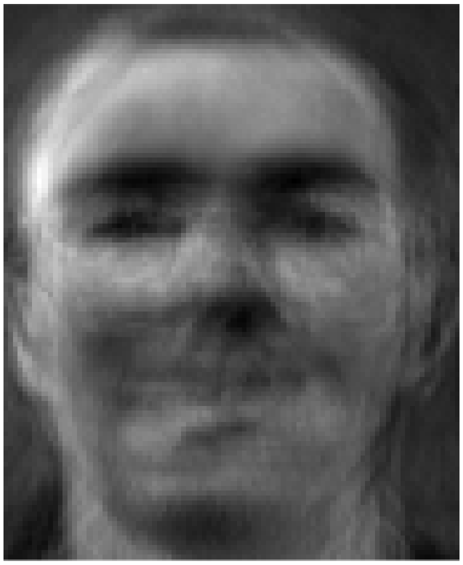


image_3

[k=50]

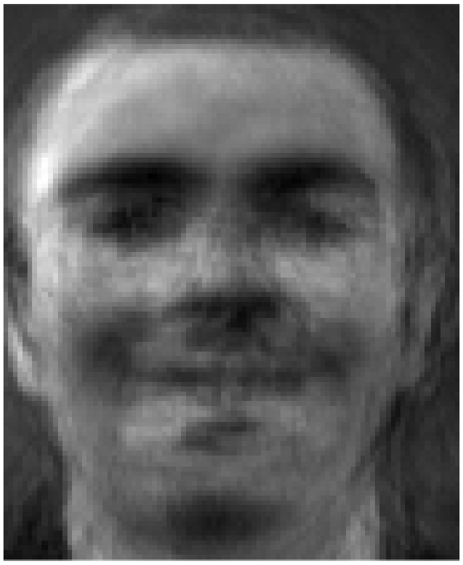
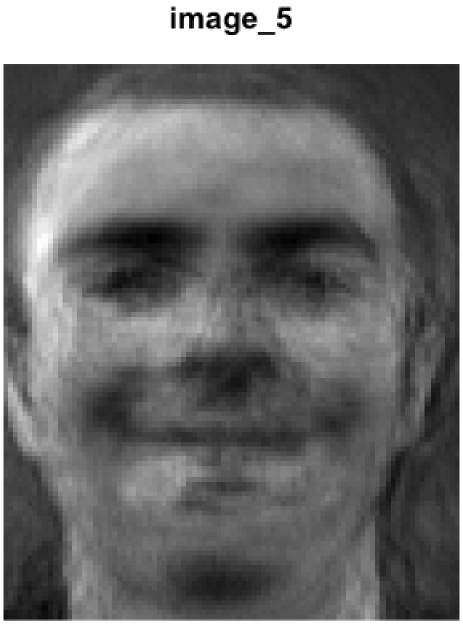
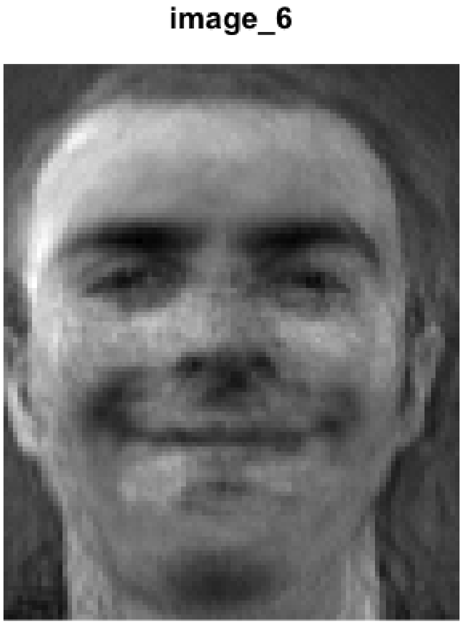


image_4

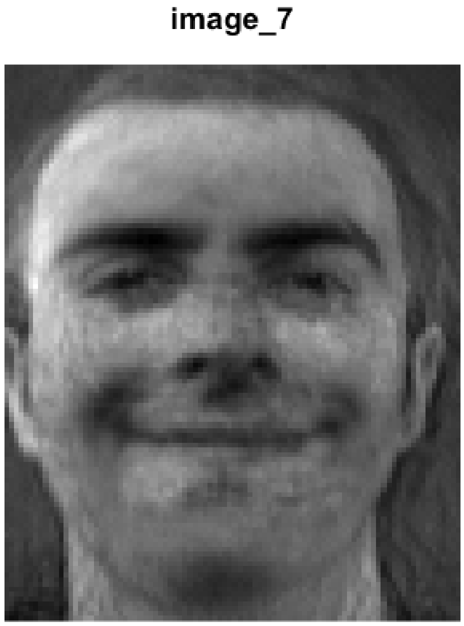
[k=100]



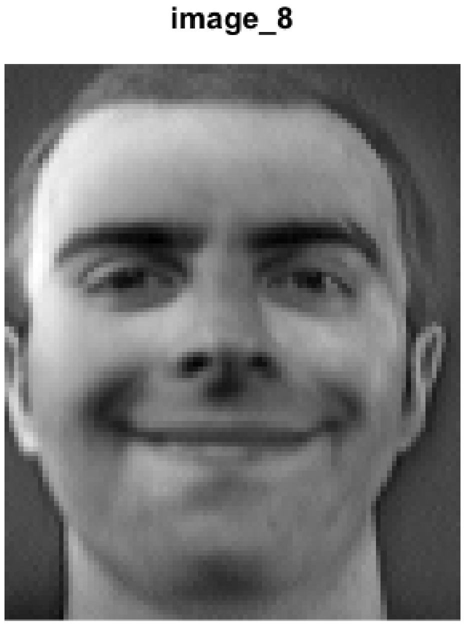
[k=150]



[k=200]



[k=250]



[k=400]

Problem 5 When data points are statistically independent and the residuals have a theoretical mean of zero and a constant variance of σ^2,
$E[MSE] = \sigma^2 + ModelBias^2 + ModelVariance$

where
E is the expected value of the Mean Square Error.
σ^2 is the "irreducible noise" that cannot be eliminated by any model.
Squared bias of the model reflects how close the functional form of the model can get to the true relationship between the predictors and the outcome.
Variance of the model

The bias-variance tradeoff is encountered in all the supersede learning models. The above equation represents the fact that it is not possible in model to minimize the Error and at the same time reduce the bias and variance observed in a model. This terms is used to represent the fact that at the same time a chosen model can either capture the regularities in its training data, or generalize well on unseen data. Models can have high bias, meaning they impose restrictions on the kind of regularities that can be learned, or they can have high variance, meaning they can learn many kinds of complex regularities including noise in the training data. To achieve good performance on data outside the training set, a tradeoff must be made.

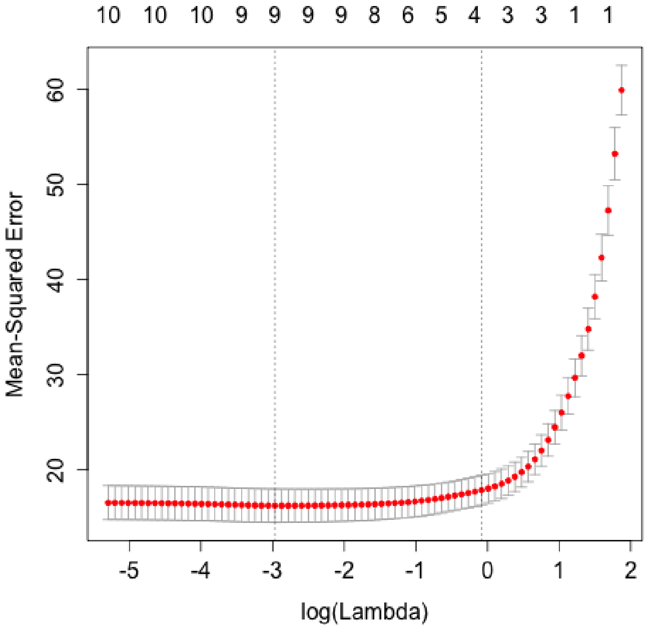
More complex models can have very high variance, which leads to over-fitting. On the other hand, simple models tend to under-fit if they are not flexible enough to model the true relationship (thus high bias). However, for simple models the variance in the model is small. Also, highly correlated predictors can lead to collinearity issues and this can greatly increase the model variance.

As the data increases by huge amount we can afford a complex model with more features because then the amount of data can result in reduction in variance of the model. Anyway the bias has increased because we have chosen a complex model. Basically we will not overfit the training data if the model is not changed. But model is made more complex with increase in training data, this might not hold true.

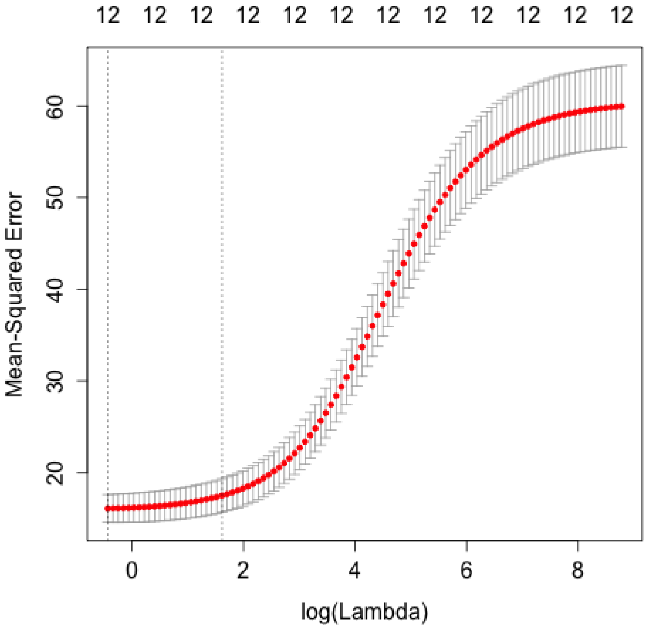
Problem 6 Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of
$E(w) = \sum_{i=1}^{N} \left\{ w^T \phi(x_n) - t_n \right\}^2 + \frac{\lambda}{2}|x|^2$,

Here λ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other). The values do not become zero therefore feature selection is not performed in Ridge Regression. It protects against the potentially high variance of gradients estimated in the short directions. The implicit assumption is that the response will tend to vary most in the directions of high variance of the inputs. This is often a reasonable assumption, since predictors are often chosen for study because they vary with the response variable, but need not hold in general.

Problem 7

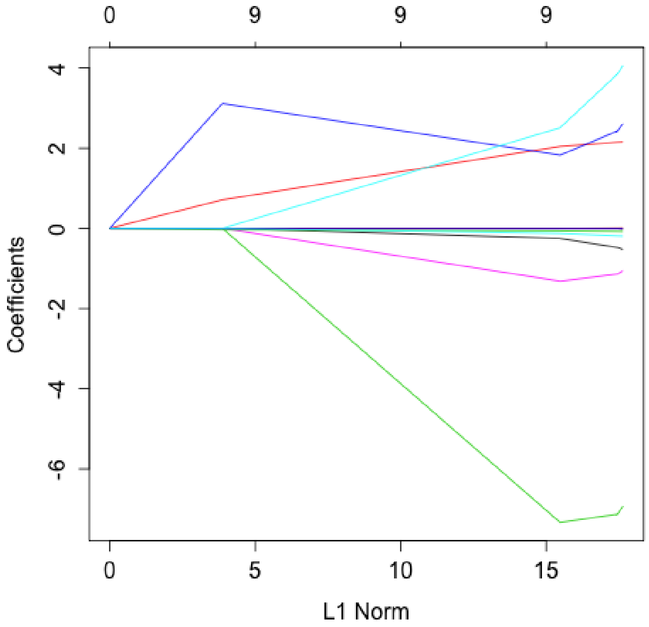


MSE when Lambad varies for Lasso Regression

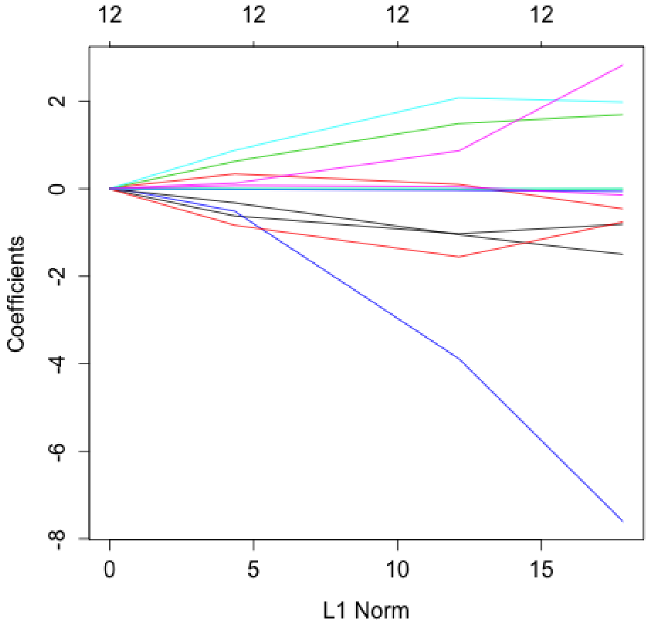


MSE when Lambda varies for Ridge Regression

Part 7.c



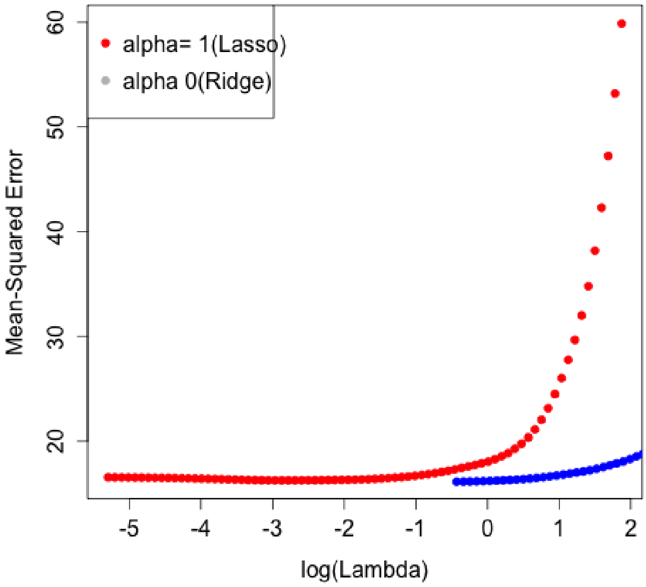
Lasso Coefficients When varied with lambda



Ridge Coefficients When varied with lambda

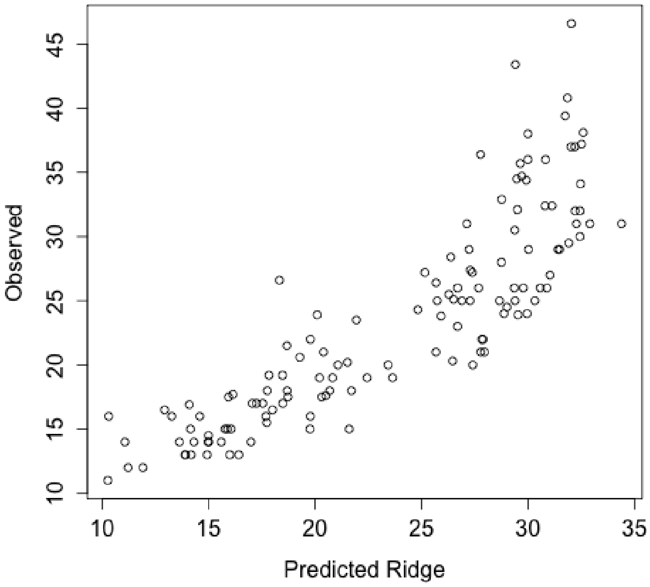
The coefficients generated from Ridge tend to move to zero however they become equal to zero only when λ becomes zero. However in Lasso as the value of λ reduces, more and more features start getting dropped from the regression model. Thus Lasso model can be used for feature selection.

Part d.

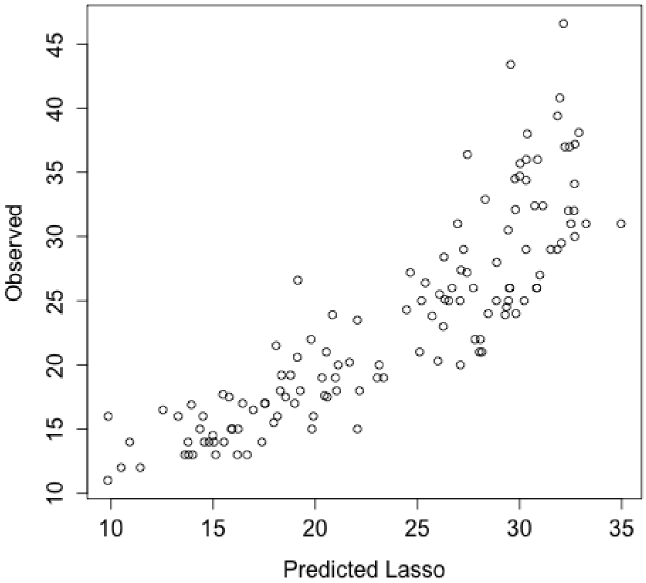


MSE Varying for Lasso and For Ridge with change in lambda

MSE of MLR = 15.73066 MSE of Lasso = 16.45156 MSE of ridge = 15.64257



Predicted values suing the Ridge Regression Plotted with the Observed Values of Test data



Predicted values suing the Lasso Regression Plotted with the Observed Values of Test data

Part e. Features being dropped by Lasso are : cylinders5 , cylinders6 , cylinders8 , displacement , acceleration , origin2

New model For MLR becomes: mpg   horsepower + weight + cylinders3 + cylinders4 + origin3

MSE for MLR with new equation : 16.28163