Method

For the project, we will analyze the given data with five different approaches: OLS(Ordinary Least Squares), RR(ridge regression), LR(lasso regression), PCR(principle components regression) and PLSR(partial least squares regression).

1) Ordinary Least Square Method(OLS)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Ordinary least squares(OLS) is estimataing parameters in a linear regression model by minimizing the sum of squares of the distance between observed responses and predicted responses. This approach is perfomed under certain assumptions: errors are finite, homoscedastic, uncorrelated, and normally distributed. Under these assumption OLS provides the minimum-variance and unbiased mean estimators.

Shrinkage Methods

Shrinkage Methods starts with fitting a modeling with all p predictors. However, it forces the estimated coefficients to be close to zero or to be zero. Also, it requires the data to be standardized before applying the method. There are two methods of shrinkage methods: Ridge Regression Method, and Lasso Regression Method. Depending on the method, some of the cofficients forced to be estimated to be exactly zero, which is the case of Lasso Regression. Thus, shrinkage Method does variable selection.

2) Ridge Regression(RR)

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$minimize_{\beta} \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \} \quad subject\ to \quad \lambda \sum_{j=1}^{p} \beta_j^2 \leq s$$

$$where\ every\ value\ of\ \lambda\ there\ is\ a\ corresponding\ s$$

Equations above are two different representations of Ridge Regression. The first equation explains the similarity between OLS regression and Ridge Regression. Ridge regression calculates the coefficients by minimizing $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$. If λ, the tunning parameter, is equal to zero, the estimated coefficients will be same as those of OLS regression.

The advantage of Ridge regression over OLS regression is bias-variance trade-off. As λ increases, the estimated coefficients forced to be close to zero, leading bias estimators and decreasing the variance. As λ decreases, the estimated coefficients do not have to be close to zero, leading less - bias estimators and increasing the variance.

3) Lasso Regression(LR)

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$minimize_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} \quad subject\ to \quad \lambda\sum_{j=1}^{p}|\beta_j| \leq s$$

$$where\ every\ value\ of\ \lambda\ there\ is\ a\ corresponding\ s$$

Equations above are two different representations of Lasso Regression. The first equation explains the similarity between OLS regression, Ridge regression and Lasso regression. Lasso regression calculates the coefficients by minimizing $RSS + \lambda\sum_{j=1}^{p}|\beta_j|$. If λ, the tunning parameter, is equal to zero, the estimated coefficients will be same as those of OLS regression.

Lasso regression has the same advantage over OLS regression as Ridge regression does: bias-variance trade-off. However, as λ increases, the Lasso regression forces esimates to be exactly equal to zero. Thus, Lasso shrinkage method performs the variable selection and a model generated by Lasso regression is easier to interpret than a model generated by Ridge regression. Lasso regression involves only a subset of the variables referred as sparse-model.

Dimension Reduction Methods

$$Z_m = \sum_{j=1}^{p}\phi_{jm}X_j$$

$$where \quad \sum_{j=1}^{p}\phi_{jm}^2 \ = \ 1$$

$$y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i \quad i \ = \ 1, ..., n$$

$$\sum_{m=1}^{M}\theta_m z_{im} \ = \ \sum_{m=1}^{M}\theta_m\sum_{j=1}^{p}\phi_{jm}x_{ij} \ = \ \sum_{j=1}^{p}\sum_{m=1}^{M}\theta_m\phi_{jm}x_{ij} \ = \ \sum_{j=1}^{p}\beta_j x_{ij}$$

$$where \quad \beta_j \ = \ \sum_{m=1}^{M}\theta_m\phi_{jm}$$

Dimenstion reduction method is transforming the predictors,
$Z_m = \sum_{j=1}^{p}\phi_{jm}X_j$ and fit a least square models using transformed variables,
$y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i \quad i \ = \ 1, ..., n$. This approach is used when there is a large number of predictor variables. By computing M different linear combinations, transforming the predictors, we can project p predictors into a M-dimensional subspace where M is less than or equal to p. Then M projections can be used as predictors to fit a linear model by ordinary least squares.

If M is equal to p, there is no dimention reduction occurs, yielding the same result equivalent to performing ordinary least squares on the p predictors.

There are two major dimension reduction methods: principal components regression and partial least squares regression. Those methods work in two steps. First step is transforming the predictors,
$Z_m = \sum_{j=1}^{p}\phi_{jm}X_j$. Second steop is selecting ϕ_{jm}.

4) Principle Components Regression(PCR)

Principle Components Regression is based on the principal component analysis. Principal components analysis extracts the most variation of datasets to reduce the dimension of the data.

Find the first principal component Z_1

$$Z_1 \;=\; \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$$

Z_1 is the first factor maximizing the dispersion of obervations.

Then find the second principal component Z_2

$$Z_2 \;=\; \phi_{12}X_1 + \phi_{22}X_2 + ... + \phi_{p2}X_p$$

Z_2 is the second factor maximizing the dispersion of observations as well as orthogonal to the first factor.

By doing this process iteratively up to desired M, it gives us the Principal Components Regression on a given M space.

Principle Components Regression shows a linear regression between the data on the factors that are correlated. It enables us to solve the colinearity between preditor variables and dimensionality of the given data. That is, the number of samples, n, does not necessarily bigger than the number of predictor variables.

5) Partial Least Squares Regression(PLSR)

Like Principle Componenet Regression, Partial Least Square Regression calculuates $Z_1, .., Z_m$ and fit a linear model on those transformed predictors.However,Partial Least Square Regression takes into account the structure of both the explanatory and dependent variable unlike Principle Component Regression. That being said, it uses response variable Y to find directions that help explaining both the reponse and the predictors.

Find the first direction Z_1

$$Z_1 \;=\; \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$$

$$by\ setting\ \phi_{j1} \;=\; \beta_j \quad where\ \beta j\ is\ from\ simple\ linear\ regression\ Y\ onto\ X_j\ and\ j = 1, ..., p$$

The first direction Z_1 consequently places the weights on each predictor variables proportional to its correlation with the resoponse variable.

To find the second direction Z_2, we need to regressing each variable on Z_1 and take the residuals. Then compute the Z_2 with the residuals that are orthogonal to the original data by applying the same procedure to compute the first directon Z_1

By doing this process iteratively up to desired M, it gives us the Partial Least Squares Regression on a given M space.