1 Method

1.1 Least Squares Method

OLS estimators are considered as our base case in this project. OLS estimators are obtained by minimizing the sum of squares $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_{ij}))^2$ where $\beta_0, ..., \beta_p$ are our coefficient estimates. This is the most common regression method, and it works the best when we have homoscedastic and uncorrelated errors within the data.

1.2 Shrinkage Method

Similar to the OLS method, shrinkage method also fits a model containing all independent variables but with a focus on constrains the coefficient estimates toward 0. By shrinking the coefficient estimates, the variance in estimators is significantly reduced compared to that of OLS.

1.3 Ridge Regression

Instead of minimizing RSS, Ridge regression minimizes $RSS + \lambda \sum bj^2$. This regression abandons the requirement of an unbiased estimator in order to obtain a more precise prediction intervals. The addition term is a shrinkage penalty and since it's a sum of coefficient estimators, it will be small when βs are all close to 0. Here, λ is called a tuning parameter and it controls how much RSS and shrinkage penalty affects the regression model. When $\lambda = 0$, ridge regression is the same as the least squares models. Because each λ corresponds to a different set of coefficient estimates, we usually try a wide range of λs and pick the best tuning parameter with the smallest MSE.

One of the main feature of ridge regression is the bias-variance trade-off. When $\lambda = 0$, which is the least square model, the variance is high the estimators are unbiased. As lambda increases, variance decreases significantly but bias only sightly increase. Since $MSE = variance + squaredbias$ along with the relative effect of the change, when λ is smaller than 10, the variance drops rapidly, with very little increase in bias, so MSE decreases.

1.4 Lasso Regression

The major difference between Lasso and Ridge is that Lasso method aims to minimize $RSS + \lambda \sum |\beta_j|$. This is to compensate for the fact that ridge regression with always generate a model with all predictors since it doesn't involve a process of variable reduction. Therefore, lasso improves on this by allowing some of the coefficient estimates to be exactly 0 if the tunning parameter is sufficiently large. Therefore, this variable selection process makes the model much easier to interpret with a reduced amount of predictors.

1.5 Comparison between Ridge and Lasso

In general, lasso is expected to outperform ridge when we have a small amount of predictors having significant coefficients, and the rest close to 0. Ridge regression will perform between when the response is affected by many regressors with equal-sized coefficients.

1.6 Dimension Reduction Methods

Dimension reduction method involves a transformation of variables before we fit in the least squares model. Instead of estimating $p + 1$ coefficients $\beta_0, ..., \beta_p$ when we have p regressors, it transform the data to $Z1, ..., Z_m, M < p$ where Z_m represent linear combinations of the original predictors, so that we only need to estimate $M + 1$ coefficients $\theta_0, \theta_1, ..., \theta_M$.

1.7 Principal Components Regression

In principal components regression, we first perform principal components analysis on the original data which constructs M principal components, $Z_1, ..., Z_m$. Then, we use these components as the predictors and fit the

least squares model to obtain coefficient estimates. The key assumption we hold in this regression model is that the directions in which $X_1, ..., X_p$ show the most variation are the directions that are associated with Y. If the assumption holds, then fitting a least squares model to $Z_1, ..., Z_m$ will lead to better results than fitting a least squares model to $X_1, ..., X_p$, since most or all of the information in the data that relates to the response is contained in $Z_1, ..., Z_m$. With fewer predictors, we can also reduce the risk of overfitting. PCR performs the best when the first few principal components are sufficient to capture most of the variation in the predictors and their relationship with the response.

1.8 Partial Least Squares

In comparison to PCR regression, which uses unsupervised method to identify the principal components, partial least squares method takes the advantage of a supervised learning process. It uses the response variable Y to identify new vectors that are not only similar to existing regressors, but also select the ones that are related to the response. Therefore, as indicated in the textbook, the PLS approach attempts to find directions that help explain both the response and the predictors.