Methods

Goal: ways in which simple lienar model can be improved by replacing plain least square fitting with alternative fitting procedures to get better prediction accuracy and model interpretability.

Ordinary Least Squares (OLS)

Ordinarly least squares regression are typicall known as the benchmark, in which it is a method that is used to estimate the unknown pareameters in a linear regression model.

In the standard multiple linear model,

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

it is important to know that Y represents the respons variable, X_i represents the i_th predictor variable, and β_i quantifies the association between the i_th predictor variable and the response. In other words, we can say that β_i is the average effect on Y for a unit increase in X_i, holding all other predictors fixed.

For our case, our linear model is in the form,

$$Balance = \beta_0 + \beta_1 xIncome + \beta_2 xLimit + \beta_3 xRating + ... + \beta_p xEthnicity + \epsilon$$

Estimating the Regression Coefficients/Parameters

From the multi-linear model above, we have to estimate the values of the regression coefficients or parameters $\beta_0, \beta_1, \beta_2, ..., \beta_p$, because they are unknown.

Given the coefficient estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$

We can get the predicition for Y:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

and ith residual (estimate of the error term) be:

$$\epsilon_i = y_i - \hat{y}_i$$

The values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ that minimizes the RSS, where

$$RSS = \sum i = 1^p (\epsilon_i)^2$$

$$RSS = \sum i = 1^p (y_i - \hat{y}_i)^2$$

$$RSS = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i 1 - \hat{\beta}_2 x_i 2 - ... - \hat{\beta}_n x_i p)^2$$

is used to estimate the multiple least square regression coefficients.

Subset Selection

Subset selection controls variance by using a subset of the original variables. From the previous homework, we have learnt the basics of subset selection.

There are 3 classical methods: * Forward stepwise selection - Starting with no variables in the model with only the intercept , we then proceed by adding a variable with the lowest RSS (greatest additional improvement). This is continued until some stopping rule is satisfied. * Backward stepwise selection - Starting with all variables in the model, we proceed to the next step by removing the variable with the largest p-value. We refit the model, and remove the variable with the largest p-value again. We stop when all p-values are below a certain threshold. * Mixed selection - Combining Forward selection and Backward selection, Mixed selection start with no variables in the model, and add variables that provide the best fit one-by-one. As new predictors are added, the p-value for one of the variables in the model can get larger. If it rises above a certain treshold, then we should remove that variable. Continue to do this until all variables in the model have a low p-value, and all variables outside the model have a large p-value if added to the model.

Cross-validation

In order to implement the methods, we can use validation and cross-validation, to determine which of the models are the best. Meaning that, it can be used to estimate the accuracy of the different models in order to choose the best one. Cross validation involves breaking data sets into training data and testing data to assess how the results of a statistical analysis can become an indepenedent data set. It uses the training data to estimate the test MSE.

Shrinkage Methods

The Shrinkage Method approach invloves fitting a model involving all p predictors instead of fitting a linear model that contains a subset of the predictors, which is discussed in the section above. The core concept behind the shrinkage method is constraining or regularizing the coefficient estimates of predictor variables towards zero. Having very minimal coefficient estimates results in a reduction of the variance.

There are two Shrinkage methods techniques:
1. Ridge Regression
2. The Lasso

Ridge regression (RR)

Ridge regression is very much alike to the least squares fitting procedure that minimizes

$$RSS = \sum i = 1^p(y_i - \beta_0, - \sum j = 1^n(\beta_i x_i j))^2$$

The difference between ridge regression and RSS is that the coefficients estimates $\hat{\beta}^R$ are the values that minimizes

$$RSS + \lambda \sum j = 1^n(\beta_j)^2$$

where $\lambda \geq 0$ is called a *tuning parameter*.

Just like the least square fitting, by making the RSS small, we can get the estimates that fit the data well. However, the second part of the equation $\lambda \sum j = 1^n(\beta_j)^2$, which is called the *shrinkage penalty*. The shrinkage penalty get smaller as $\beta_1, ..., \beta_n$ gets to zero, giving an effect of shrinking the estimates of β_j towards zero. Unlike leaast square fitting, ridge regression will generate a different set of coefficient estimats for $\hat{\beta_\lambda}^R$ for each value of λ.

When $\lambda = 0$, then the ridge regression will be equivalent to the ordinary least square fitting. When λ grows larger, however, the impact of the shrinkage penalty will grow, and the ridge regression coefficient estimates will approach zero. Additionally, there will also be an decrease in the flexibility of ridge regression fit, which leads to decreased variance and increased bias. Th theory behind this is rooted from the *bias-variance trade-off*.

Lasso regression (LR)

Lasso coefficeient overcomes the one limitation of ridge regression: includes all the predictors in the final model. The lasso coefficients minimizes

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

The lasso also shrinks the coefficient estimates towards zero. However, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when λ is large. Hence, the lasso performs variable selection, which generates a much easier to interpret model. Lasso also yields sparse models, meaning it involves only a subset of the variables.

When we perform the lasso, we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to how large $\sum_{j=1}^{p}|\beta_j|$ can be. It is clear that the lasso produces simpler and more interpretable model. However, it assumes that a number of the coefficients are truly equal to zero, resulting in prediction error. This shows that the ridge regression will perform better when the response is a function of many predictors with coefficients of roughly equal size while lasso regression yields a reduction in variance and can generate more accurate predictions.

Dimension Reduction Methods

From the subset selection method and shrinkage methods above, we have seen ways in which variance can be controlled by either using a subset of the original variables, or by shrinking the coefficients towards zero. Now, we are going to explore another approach that involves transforing the predictors, and fitting them to a least square model, called *dimension reduction.*

There are two Dimension Reduction methods:
1. Principal Componenents Regression
2. Partial Least Squares Regression

Principal Components regression (PCR)

Principal components regression is one of popular methods of dimension reduction. For a linear transformation, it involves transforming data from a high-dimensional space to a space of fewer dimensions. In other words, it derives a low-dimensional set from a large set of variables. The first step of principal components regression is to construct the first M principal components, $Z_1, ..., Z_M$, and use the components as predictors for the linear regression model that is fit with the ordinary least squares. A smaller number of principal components, where $M < P$ is enough to expalin the variability in the data and the relationship with the response. Using $M < P$ predictors is not a feature of subset selection because each of the M principal components used is a linear combination of all the p predictors.

Assuming that the directions of $X_1, ..., X_p$ shows that the most variations are the directions associated with Y, often times, although the assumption is not guaranteed, it turns out to be a reasonable approximation which gives good results. If the assumption is true, then we can say that fitting a leaast squares model to $Z_1, ..., Z_M$ produces better results than fitting a least square model to $Z_1, ..., Z_P$. Additionally, estimating a smaller number of coefficients can mitigate overfitting.

Before performing PCR, it is important to *standardize* each predictors. This ensures that all the variables are on the same scale, and avoids the effect the roles high-variance variables play in PCR.

Partial Least Squares regression (PLSR)