**Shrinkage Methods:**

**Ridge Regression** – This method minimizes a similar cost function as OLS (RSS) but with an added penalty term. This function is:

The penalty term is controlled by the tuning parameter, A larger value shrinks the coefficients more and a smaller value leaves them closer to their OLS values. Notice that the penalty term only applies to the independent variables in the model and not the intercept term. This means that as , the resulting model converges toward the null model with only the intercept term, so it is just the mean of the data. Conversely, when , the resulting model is just OLS. An important distinction with ridge regression is that none of the predictor variables actually go to zero, it keeps all variables in the model. For this reason, it is not a variable selection method, just a shrinkage method. It is also important to standardize the predictors before applying ridge regression, unless the variables all use the same unit of measurement. It is common to use cross-validation to find the optimal

**Lasso** – Lasso is similar to ridge regression in that it minimizes RSS from OLS but with an added penalty term, but instead of using the L2 norm (as ridge does above), it uses the L1 norm in the penalty term. The function it minimizes is:

The penalty term here is also controlled by the tuning parameter, . A smaller results in most of the predictor variables being left in with resulting in the OLS model, but conversely, a large results in some predictor variables getting zeroed out with only the most important left in. And with a large enough , the resulting model is the null model with all predictor variables getting zeroed out. This means that Lasso differs in a big way from ridge in that it can be used as a variable selection method. It yields sparse models while ridge yields the full model with all predictors. It is important to scale the predictor variables once again, and also to use cross-validation to find the optimal

**Dimension Reduction Methods** – Dimension reduction methods transform the predictors in the model and then fit a least squares model using these transformed predictors. The transformed variables are a linear combination of the original predictors and can be depicted as

Where is some constant vector applied to all p variables in the model. There are variables in the new model. If , this would just be OLS. The final model fitted is:

Where the vector contains the OLS estimates on the m newly transformed variables in the model.

**Principal Component Regression (PCR)** – PCR is an unsupervised learning technique, meaning the target variable, Y, is not used to determine principal component directions. PCR determines the linear combination of the predictor variables in the directions that contain the most variance (where the observations vary the most) in the model in ascending order, meaning the first will contain the most variance, the second will contain less, and so on and so forth. Each transformed variable is also orthogonal to one another, meaning they are all independent. You decide how many principal components you want (m transformed variables) and then fit these using OLS.

**Partial Least Squares (PLS) –** PLS is a supervised learning technique, meaning that it does utilize the target variable, Y, to determine principal component directions. That is, PLS finds directions that explain both the response and the predictors. The high level explanation is that it regresses Y onto each predictor variable X and uses the correlated coefficients from that model to build out the transformed variables.

**Scenarios**

**Ridge regression** – In the event that all variables in a model are thought to be useful and mostly uncorrelated, it would be advantageous to use ridge regression since it will only shrink the coefficients for each predictor based on which are the most important. It will also tease out which variables are correlated and eliminate that problem in its shrinkage, provided it is given an optimal tuning parameter value. This will keep all variables and give you magnitudes that help suggest which variables are the most powerful in predictions. A good scenario could be for an economic model where you want to account for all carefully procured and cleaned economic variables to determine their effect on an important outcome.

**Lasso** – In the event that only a few of your variables are important, or at least a relatively small subset, and you also want to be able to interpret the results of your model, you would want to use Lasso. This is because Lasso will select the best subset of variables that it deems are most important and help you understand your problem better by having fewer predictors to account for. A practical scenario could be for cancer research where you have a ton of data and variables from patients, but you’re not confident in a lot of them and want to find out which ones are the most important. Lasso would trim the number down and give you a smaller subset to work with and infer from.

**Dimension reduction methods** – In the event where you care most about predictions and not necessarily the interpretability of the model, you should use dimension reduction methods. The model will do a good job of summarizing the important variables while cutting out the noise that results in overfitting the model. This will lead to less variance in the predictions. A practical scenario could also be the cancer prediction model, but this time used more for immediate diagnostics for patients rather than research. You want as accurate a prediction as possible to decide whether you should do more tests for a patient.