Multicollinearity: Ridge Regression

Ishan Paul,bs2033

October 2022

In this section, we will learn about "Ridge Regression." It is a technique, used to get better estimates, in terms of MSE, of the coefficients and their variances. Ridge Regression is more popular than the subset selection method, introduced in the previous section, since here, we do not discard any data and hence, make use of more information.

Before jumping into definitions, it would be beneficial to review the problem at hand. Consider the following model:

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

$$\vec{\epsilon} \sim (\vec{0}, \sigma^2 I)$$

$\vec{Y}$ and $\vec{\epsilon}$ are n×1 vectors, $\vec{\beta}$, a p×1 vector, and X, an n×p matrix. We assume X is of full-column rank, but is very close to being not full column rank, due to multicollinearity. The least square estimator of the coefficients is the following:

$$\hat{\vec{\beta}} = (X'X)^{-1}X'\vec{y}$$

The variance-covariance matrix of $\hat{\vec{\boldsymbol{\beta}}}$ is given by $\sigma^2(X'X)^{-1}$. Multicollinearity makes X'X "almost singular." So, the entries of $(X'X)^{-1}$ in this case are much larger. As a result, the variances of the estimators are also very large, making the least square estimators unreliable. A naive solution to the problem is to get rid of the singularity, by adding a positive, non-singular part λI to X'X.

$$\hat{\vec{\beta}}(\lambda) = (X'X + \lambda I)^{-1}X'\vec{y} \qquad \lambda \geq 0$$

$\hat{\vec{\boldsymbol{\beta}}}$ is called the ridge estimate and the non-negative parameter, λ, the ridge parameter. For λ large enough, $X'X + \lambda I$ will be non-singular and postive definite.

Although this bounds the variance within acceptable range, several things remain unclear. Not only does this make the estimator biased, we also have no clue as to which values of λ would work well in this case. Therefore, it would be fruitful to plot measures of the error against different values of λ.

In the following plots, we consider β_i an element of $\vec{\beta}$ and consider the corresponding least square estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$. Here $1 \leq i \leq p$. The plot of $\text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}})$ against λ is as follows:

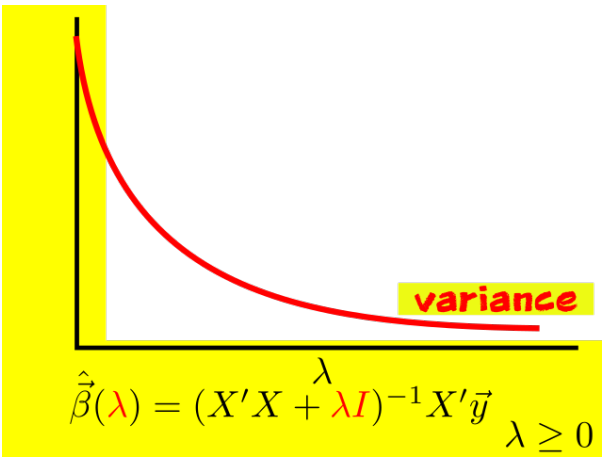

Figure 1: $\text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}})$ vs λ

At $\lambda = 0$, the highest variance is observed. This is nothing but the unaltered, large variance that we wanted to avoid. As λ increases, it dominates X'X and the variance decreases, assymptotically approaching 0. However, the information provided by X gets less and less weightage as the estimator assymptotically approahes the constant estimator. So, both very small and very large values of λ are undesirable, and we expect the sweet spot to be somewhere in between.

Now we plot the $\boldsymbol{bias^2}$ of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ against λ. The following graph is observed.

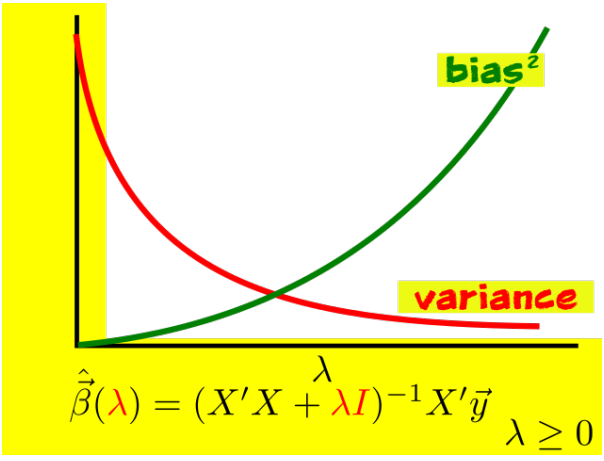

Figure 2: $bias^2$ of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ vs λ

At $\lambda = 0$, the bias is 0 as that is our original unbiased estimator. As λ increases, the λI dominates and the expected value of the estimator moves further away from the true value of the parameter. Once again, we see that λ being too large is undesirable.

Now that we have seen how variance and $bias^2$ vary individually, we should turn our attention to the Mean Square Error(MSE) of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ as it is a better measure of the error expected by the model. The MSE is given by:

$$MSE = bias^2 + variance$$

The plot of MSE of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ against λ, is as follows.

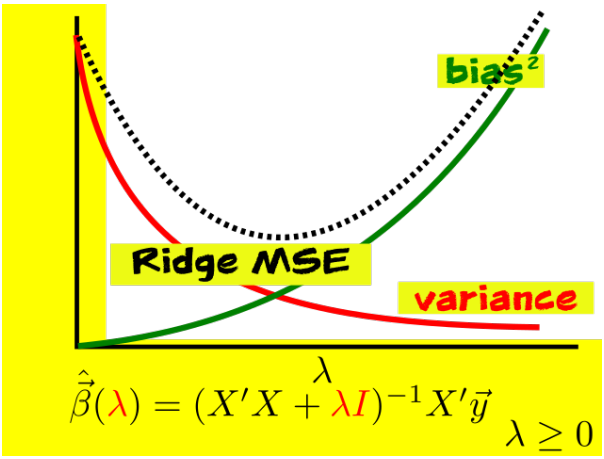

Figure 3: MSE of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ vs λ

Consistent with our previous observations, the ridge MSE decreases to a lowest value and then increases monotonically. As λ increases, the variance approaches 0, and the ridge MSE assymptotically approaches the $bias^2$ curve.

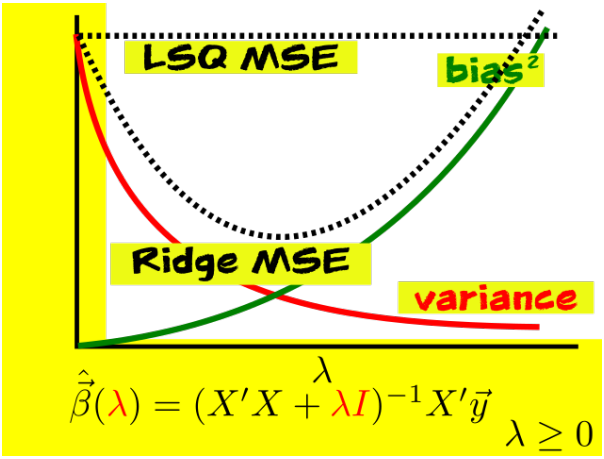

Figure 4: Ridge MSE vs least squares MSE

Compared to the ridge MSE, the least square MSE is equal to the MSE at $\lambda = 0$. We see in the above graph that it is higher than the minimum ridge MSE. The least square MSE, which is invariant of λ stays above the ridge MSE for small enough values of λ. The only problem now is to figure out, which value of λ would be optimum. There are various heuristics for that and it is not a difficult job to find the optimum λ.

Even though the ridge estimator is non-linear and biased, it is a better estimator as it has a lower MSE than the least square estimator. However, the ridge estimator is only used when the least square MSE is large due to multicollinearity. Otherwise, if the least square MSE is already small, the decrease in MSE by using the ridge estimator is not appreciable. So, we prefer the least square estimator in that case.

In the next section, we will look at the same idea with a more modern and rigorous approach.