Multicollinearity: Ridge: ad hoc and Bayes approach

Manas Patnayakuni: BS2017

October 25,2022

1 Recapitulation

We studied that ridge regression is a unique way to deal with multicollinearlity in the observed data. Our introduction to this concept was rather ad hoc so we shall now discuss it in a more detailed manner.

$$\hat{\vec{\beta}}(\lambda) = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

Previously, we have intuitively concluded that adding λI (where $\lambda \geq 0$) to $X^T X$ would help in getting rid of the singular nature of $X^T X$. This should make $X^T X + \lambda I$ non-singular and positive definite for a large enough value of λ. The new estimator that we obtain is called the ridge regression estimator and it is a better estimator than the least square estimator because of its lower MSE.

2 Bayesian Interpretation

To understand ridge regression in a more systematic and rigorous approach, we take help of the Bayesian formulation. This approach is similar to how we perceive the least squares method as a maximum likelihood technique under a particular model. Instead of looking at the least square method as a way of minimizing the norm of $\vec{\epsilon}$, we can express it as a statistical assumption.

Similarly, we can look at ridge regression as a Bayesian problem. From our assumptions, we have the following Gauss Markov model written as a distribution of the observed data $\vec{y}$:

$$\vec{y} \sim N_n(X\vec{\beta}, \sigma^2 I)$$

To setup our Bayesian model, we need to consider $\vec{\beta}$ to be a random vector following a Gaussian distribution. With this, our new model can be written as:

$$\vec{y}|\vec{\beta} \sim N_n(X\vec{\beta}, \sigma^2 I)$$

In this model, $\vec{\beta}$ is a random vector so we need to specify a distribution called the prior which is:

$$\vec{\beta} \sim N_p(\vec{0}, \tau^2 I)$$

2.1 Bayesian Procedure

Let us carry out the Bayesian procedure given the above assumptions. The posterior distribution $f(\vec{\beta}|\vec{y})$ is:

$$\begin{aligned}
f(\vec{\beta}|\vec{y}) &= \frac{f(\vec{y}|\vec{\beta})f(\vec{\beta})}{\int_\beta f(\vec{y}|\vec{\beta})f(\vec{\beta})\,\mathrm{d}\beta} \\
&\propto f(\vec{y}|\vec{\beta})f(\vec{\beta}) \\
&\propto exp[-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)] * exp[-\frac{1}{2\tau^2}\beta^T\beta] \\
&= exp[-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta) - \frac{1}{2\tau^2}\beta^T\beta]
\end{aligned}$$

Now we have to find the β for which the above expression is maximum,i.e, the posterior mode. This mode is supposed to be the ridge regression estimator.

$$\begin{aligned}
\hat{\vec{\beta}}_{ridge} &= \arg\min_\beta[exp[-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta) - \frac{1}{2\tau^2}\beta^T\beta]] \\
&= \arg\min_\beta[\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta) + \frac{1}{2\tau^2}\beta^T\beta] \\
&= \arg\min_\beta[(y-X\beta)^T(y-X\beta) + \frac{\sigma^2}{\tau^2}\beta^T\beta] \\
&= \arg\min_\beta[(y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta) + \frac{\sigma^2}{\tau^2}\beta^T\beta] \\
&= \arg\min_\beta[(y^Ty - \beta^TX^Ty - \beta^TX^Ty + \beta^TX^TX\beta) + \frac{\sigma^2}{\tau^2}\beta^T\beta] \\
&\because (y^TX\beta)^T = y^TX\beta \\
&= \arg\min_\beta[(y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta) + \frac{\sigma^2}{\tau^2}\beta^T\beta] \\
&= \arg\min_\beta[SS]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial SS}{\partial \beta} &= \frac{\partial((y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta) + \frac{\sigma^2}{\tau^2}\beta^T\beta)}{\partial \beta} \\
&= -2X^Ty + 2X^TX\beta + 2\frac{\sigma^2}{\tau^2}\beta = 0
\end{aligned}$$

Solving the above equation, we get the ridge regression estimator:

$$\hat{\vec{\beta}}_{ridge} = (X^TX + \frac{\sigma^2}{\tau^2}I)^{-1}X^T\vec{y}$$

For the previous formulation, we have used λ as the tuning parameter in the ridge regression estimator. However, in the Bayesian formulation, τ is the tuning parameter and there is a corresponding value of λ for every τ. From the Bayesian procedure, we get the following relation between the two tuning parameters:

$$\lambda = \frac{\sigma^2}{\tau^2}$$