Ridge Regression, a.k.a $\hat{\beta}^{Ridge} = \arg\min \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$

Andersen Ang

First created: 2014. Last update : 2017-Feb

1 Summary

Consider the regression problem

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \;\; \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

Now consider $n = p$, our goal is to estimate β.

The Ordinary Least Square Estimator $\hat{\beta}^{OLS} = \arg\min \|y - X\beta\|_2^2$ has the analytical closed form $\hat{\beta}^{OLS} = \left(X^T X\right)^{-1} X^T y$ and distribution $\hat{\beta}^{OLS} \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(X^T X\right)^{-1}\right)$ for β_0 is the true value.

The OLS estimator had the problem of *overfitting*. So a penalized OLS, called the *Ridge Regression* is invented. It has the following form

$$\hat{\beta} = \arg\min \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

2 The Ridge Regression derivation

Denote the true β be β_0 and our estimate be $\hat{\beta}$. Our $\hat{\beta}$ can be found by solving the optimization problem

$$\hat{\beta} = \arg\min \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Thus cost function J is a sum of the error and the size of the β

$$J = \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Since for a vector v, $||v||_2^2 = v^T v$, so

$$\|y - X\beta\|_2^2 = \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right) = y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}$$

$$\|\beta\|_2^2 = \beta^T \beta$$

Thus

$$J = y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta} + \lambda\hat{\beta}^T\hat{\beta}$$

Now, as we want to minimize J by setting $\dfrac{\partial J}{\partial \hat{\beta}}$

$$\frac{\partial J}{\partial \hat{\beta}} = \frac{\partial}{\partial \hat{\beta}}\left[y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta} + \lambda\hat{\beta}^T\hat{\beta}\right]$$

Formula for Matrix calculus

$$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0} \quad \frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \quad \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A} \quad \frac{\partial \mathbf{x}^\mathbf{T}\mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}^\mathbf{T} \quad \frac{\partial \mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{cases} \left(\mathbf{A} + \mathbf{A}^\mathbf{T}\right)\mathbf{x} & \text{General } \mathbf{A} \\ 2\mathbf{A}\mathbf{x} & \text{Symmetric } \mathbf{A} \end{cases}$$

Since X^TX is symmetric matrix, and therefore

$$\frac{\partial J}{\partial \hat{\beta}} = \left[-y^TX - y^TX + 2X^TX\hat{\beta} + 2\lambda\hat{\beta}\right]$$

$$\frac{\partial ||y - X\hat{\beta}||_2^2}{\partial \hat{\beta}} = -2X^Ty + 2\left(X^TX + \lambda I\right)\hat{\beta}$$

Now set this derivative equal to zero, we get

$$-2X^Ty + 2\left(X^TX + \lambda I\right)\hat{\beta} = 0$$

$$X^Ty = \left(X^TX + \lambda I\right)\hat{\beta}$$

$$\hat{\beta} = \left(X^TX + \lambda I\right)^{-1}X^Ty$$

Therefore, now we have the analytical solution of $\hat{\beta}$ that minimize J.

3 The derivation of $\hat{\beta}^{Ridge} \sim \mathcal{N}\left(\left[\left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1} - I\right]\beta_0, \sigma^2\left(X^TX + \lambda I\right)^{-1}X^T$

It can be shown that $\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\beta_0 + \left(X^TX + \lambda I\right)^{-1}X^T\varepsilon$

First, the original expression of $\hat{\beta}$ is

$$\hat{\beta} = \left(X^TX + \lambda I\right)^{-1}X^Ty$$

Take out the X^TX

$$\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\left(X^TX\right)^{-1}X^Ty$$

Since $y = X\beta + \varepsilon$, where β is the "true β", denoted as β_0, then

$$\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\left(X^TX\right)^{-1}X^T\left[X\beta_0 + \varepsilon\right]$$

Expand

$$\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\left(X^TX\right)^{-1}X^TX\beta_0 + \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\left(X^TX\right)^{-1}X^T\varepsilon$$

$$\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\beta_0 + \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\left(X^TX\right)^{-1}X^T\varepsilon$$

For the second term, move the X^TX back

$$\hat{\beta} = \left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\beta_0 + \left(X^TX + \lambda I\right)^{-1}X^T\varepsilon$$

Now we can show that $\hat{\beta}^{Ridge} \sim \mathcal{N}\left(\left[\left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1} - I\right]\beta_0, \sigma^2\left(X^TX + \lambda I\right)^{-1}X^TX\left(X^TX + \lambda I\right)^{-1}\right)$

First, consider the bias (the mean of $\hat{\beta} - \beta_0$)

$$\mathbb{E}\left(\hat{\beta} - \beta_0\right) = \mathbb{E}\left[\left(I + \lambda\left(X^TX\right)^{-1}\right)^{-1}\beta_0 + \left(X^TX + \lambda I\right)^{-1}X^T\varepsilon - \beta_0\right]$$

Recall that for $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_{p\times p}\right)$. That means ε is having zero mean and ε_i are independent from each other ($\mathbb{E}\left[\varepsilon\varepsilon^T\right] = \sigma^2 I$)

$$\mathbb{E}\left(\hat{\beta}-\beta_0\right)=\mathbb{E}\left[\left(\left(I+\lambda\left(X^TX\right)^{-1}\right)^{-1}-I\right)\beta_0\right]+\left(X^TX+\lambda I\right)^{-1}X^T\underbrace{\mathbb{E}\varepsilon}_{0}$$

So

$$Bias\left(\hat{\beta}^{Ridge}\right)=\mathbb{E}\left(\hat{\beta}-\beta_0\right)=\mathbb{E}\left(\left(I+\lambda\left(X^TX\right)^{-1}\right)^{-1}-I\right)\beta_0$$

Therefore, $\hat{\beta}^{OLS}$ is a *biased* estimator of β

Now check for variance

$$Var\left[\hat{\beta}\right]=\mathbb{E}\left[\left(\hat{\beta}-\beta_0\right)\left(\hat{\beta}-\beta_0\right)^T\right]$$

Since only the noise part ε have nonzero variance, so

$$Var\left[\hat{\beta}\right]=\mathbb{E}\left[\left(\left(X^TX+\lambda I\right)^{-1}X^T\varepsilon\right)\left(\left(X^TX+\lambda I\right)^{-1}X^T\varepsilon\right)^T\right]$$

Since $(PQ)^T=Q^TP^T$and $(ABC)^T=C^TB^TA^T$

$$Var\left[\hat{\beta}\right]=\mathbb{E}\left[\left(X^TX+\lambda I\right)^{-1}X^T\varepsilon\varepsilon^TX\left(X^TX+\lambda I\right)^{-1}\right]$$

$$Var\left[\hat{\beta}\right]=\left(X^TX+\lambda I\right)^{-1}X^T\mathbb{E}\left[\varepsilon\varepsilon^T\right]X\left(X^TX+\lambda I\right)^{-1}$$

Recall $\mathbb{E}\left[\varepsilon\varepsilon^T\right]=\sigma^2I$

$$Var\left[\hat{\beta}\right]=\left(X^TX+\lambda I\right)^{-1}X^T\sigma^2X\left(X^TX+\lambda I\right)^{-1}$$

$$Var\left[\hat{\beta}\right]=\sigma^2\left(X^TX+\lambda I\right)^{-1}X^TX\left(X^TX+\lambda I\right)^{-1}$$

Therefore

$$\hat{\beta}^{Ridge}\sim\mathcal{N}\left(\left[\left(I+\lambda\left(X^TX\right)^{-1}\right)^{-1}-I\right]\beta_0,\sigma^2\left(X^TX+\lambda I\right)^{-1}X^TX\left(X^TX+\lambda I\right)^{-1}\right)$$