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

Andersen Ang

First created: 2014. Last update : 2017-Feb

1 Summary

Consider the regression problem

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

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

Summary

- The Ordinary Least Square Estimator $\hat{\beta}^{OLS} = \arg\min \|y - X\beta\|_2^2$
- It has analytical solution $\hat{\beta}^{OLS} = \left(X^T X\right)^{-1} X^T y$
- It has the distribution $\hat{\beta}^{OLS} \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(X^T X\right)^{-1}\right)$ for $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$ and β_0 is the true value.

2 The Ordinary Least Square Estimator derivation

Denote the true β be β_0 and our estimate be $\hat{\beta}$.
By putting $\hat{\beta}$ into the system, $X\hat{\beta} = \hat{y}$.

The error between y and $\hat{y}$ is

$$error = y - \hat{y}$$

$$error = y - X\hat{\beta}$$

Then to find $\hat{\beta}$ that best approximate β_0, is to find the $\hat{\beta}$ that minimize this error. Typically the squared L_2 norm will be used to measure the size of the error :

$$||error||_2^2 = ||y - X\hat{\beta}||_2^2$$

That means,

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

Recalled The squared L_2 norm of a vector v equal to $v^T v$. So the squared L_2 norm of the error is thus

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

Recall that $(A + B)^T = A^T + B^T$, thus

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

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

Recall that $(AB)^T = B^T A^T$, thus

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

Now, as we want to minimize the squared L_2 norm of error, so we want to find the minimum of $||y - X\hat{\beta}||_2^2$.

Recall that to find the minimum of a function $f(x)$, we find $\dfrac{df(x)}{dx}$ and set it to zero (finding the critical point).

Thus we find $\dfrac{\partial||y - X\hat{\beta}||_2^2}{\partial\hat{\beta}}$

$$\frac{\partial||y - X\hat{\beta}||_2^2}{\partial\hat{\beta}} = \frac{\partial}{\partial\hat{\beta}}\left[y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}\right]$$

$$\frac{\partial||y - X\hat{\beta}||_2^2}{\partial\hat{\beta}} = \frac{\partial}{\partial\hat{\beta}}y^T y - \frac{\partial}{\partial\hat{\beta}}y^T X\hat{\beta} - \frac{\partial}{\partial\hat{\beta}}\hat{\beta}^T X^T y + \frac{\partial}{\partial\hat{\beta}}\hat{\beta}^T X^T X\hat{\beta}$$

Recall the following 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 $X^T X$ is symmetric matrix, and therefore

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

$$\frac{\partial||y - X\hat{\beta}||_2^2}{\partial\hat{\beta}} = -2X^T y + 2X^T X\hat{\beta}$$

Now set this derivative equal to zero, we get

$$-2X^T y + 2X^T X\hat{\beta} = 0$$

$$X^T y = X^T X\hat{\beta}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

Therefore, now we have the analytical solution of $\hat{\beta}$ that minimize the squared L_2 error.

3 The derivation of $\hat{\beta} \sim \mathcal{N}\left(\beta_0, \sigma^2\left(X^T X\right)^{-1}\right)$

It can be shown that $\hat{\beta} \sim \mathcal{N}\left(\beta_0, \sigma^2\left(X^T X\right)^{-1}\right)$, first let's consider the mean of $\hat{\beta}$.

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

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

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

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

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

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

Then to find the mean of $\hat{\beta}$

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

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

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

$$\mathbb{E}\hat{\beta} = \beta_0$$

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

Now check for variance

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

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

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

For $\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)^T$, note that $(PQ)^T = Q^T P^T$and $(ABC)^T = (A(BC))^T$

$$(ABC)^T = (A(BC))^T$$

$$(ABC)^T = (BC)^T A^T$$

$$(ABC)^T = C^T B^T A^T$$

So $\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)^T = \varepsilon^T X \left(X^T X\right)^{-1}$ and

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

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

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

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

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

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

Therefore

$$\hat{\beta} \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(X^T X\right)^{-1}\right)$$

Final words
It can be showed that OLS is the best estimator in this case by Gauss-Markov Theorem