Gauss-Markov Theorem for OLS is the best linear unbiased estimator

Andersen Ang

First created: 2014. Last update : 2017-Feb

1 The regression problem

$$y = X\beta + \varepsilon$$

Gauss-Markov assumptions on ε

- $\mathbb{E}\varepsilon = 0$
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$

It means, the noise is zero-mean
The variance $Var\left(\varepsilon_i\right) = Cov\left(\varepsilon_i, \varepsilon_i\right) = \sigma^2 < \infty$, all noise have same variance.
The covariance $Cov\left(\varepsilon_i, \varepsilon_i\right) = 0 \; \forall i \neq j$, all noise are uncorrealted

Then under these assumptions, OLS estimator is the best linear unbiased estimator. That means, for all unbiased estimators for this regression, OLS estimator has the smallest variance.

Note. The assumptions only requires the noise to be zero-mean and $Cov\left(\varepsilon_i, \varepsilon_j\right) = \sigma^2 \delta_{ij}$. It does not require the noise to be Gaussian noise.

2 The proof

Recall that $\hat{\beta}^{OLS} = \left(X^T X\right)^{-1} X^T y$, $\mathbb{E}\left[\hat{\beta}^{OLS}\right] = \beta_0$, $Var\left[\hat{\beta}^{OLS}\right] = \sigma^2 \left(X^T X\right)^{-1}$
Now let $\hat{\beta}^{Other} = \left(\left(X^T X\right)^{-1} X^T + D\right) y$ where D is a non-zero matrix.
Then

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

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

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

$$\mathbb{E}\left[\hat{\beta}^{Other}\right] = \mathbb{E}\left[\hat{\beta}^{OLS}\right] + \mathbb{E}\left[D\left(X\beta_0 + \varepsilon\right)\right]$$

$$\mathbb{E}\left[\hat{\beta}^{Other}\right] = \beta_0 + \mathbb{E}\left[DX\right]\beta_0 + \underbrace{\mathbb{E}\left[D\varepsilon\right]}_{D\mathbb{E}[\varepsilon]=0}$$

$$\mathbb{E}\left[\hat{\beta}^{Other}\right] = \left[I + DX\right]\beta_0$$

So $\hat{\beta}^{Other}$ is unbiased iff $DX = 0$.

$$Var\left[\hat{\beta}^{Other}\right] = Var\left[\left(\left(X^TX\right)^{-1}X^T + D\right)y\right]$$

$$Var\left[\hat{\beta}^{Other}\right] = \mathbb{E}\left[\left\{\left(\left(X^TX\right)^{-1}X^T + D\right)y\right\}\left\{\left(\left(X^TX\right)^{-1}X^T + D\right)y\right\}^T\right]$$

$$Var\left[\hat{\beta}^{Other}\right] = \mathbb{E}\left[\left(\left(X^TX\right)^{-1}X^T + D\right)yy^T\left(\left(X^TX\right)^{-1}X^T + D\right)^T\right]$$

$$Var\left[\hat{\beta}^{Other}\right] = \left(\left(X^TX\right)^{-1}X^T + D\right)\mathbb{E}\left[yy^T\right]\left(\left(X^TX\right)^{-1}X^T + D\right)^T$$

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

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

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

$$Var\left[\hat{\beta}^{Other}\right] = \sigma^2\left[\left(X^TX\right)^{-1}X^T\left(\left(X^TX\right)^{-1}X^T\right)^T + \left(X^TX\right)^{-1}X^TD^T + D\left(\left(X^TX\right)^{-1}X^T\right)^T + DD^T\right]$$

Since $DX = 0$ so $X^TD^T = \left(DX\right)^T = 0^T = 0$

$$Var\left[\hat{\beta}^{Other}\right] = \sigma^2\left[\left(\left(X^TX\right)^{-1}\right)^T + \left(X^TX\right)^{-1}\underbrace{X^TD^T}_{0} + \underbrace{DX}_{0}\left(\left(X^TX\right)^{-1}\right)^T + DD^T\right]$$

$$\left(\left(X^TX\right)^{-1}\right)^T = \left(\left(X^TX\right)^T\right)^{-1} = \left(X^T\left(X^T\right)^T\right)^{-1} = \left(X^TX\right)^{-1}$$

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

$$Var\left[\hat{\beta}^{Other}\right] = \underbrace{\sigma^2\left(X^TX\right)^{-1}}_{Var\left[\hat{\beta}^{OLS}\right]} + \sigma^2DD^T$$

$$Var\left[\hat{\beta}^{Other}\right] = Var\left[\hat{\beta}^{OLS}\right] + \sigma^2DD^T$$

Since $DD^T > 0$ (DD^T is positive semi-definite matrix)

$$Var\left[\hat{\beta}^{Other}\right] \geq Var\left[\hat{\beta}^{OLS}\right]$$

Therefore the OLS estimator has the lowest variance, and hence it is the best estimator.