Lecture 2: Three Justifications for OLS

Dennis Sun

June 24, 2015

Today, we will study three derivations of the OLS estimator

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{1}$$

To save space, we will use the shorthand $X^+ \stackrel{def}{=} (X^T X)^{-1} X^T$. X^+ is called the pseudoinverse of X because it satisfies $X^+ X = I$. (It's not a proper inverse, though, because $xx^+ \neq I$.)

1 As the Best Linear Unbiased Estimator

Recall that our model for the data is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\text{E}[\boldsymbol{\epsilon}|X] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\epsilon}|X] = \sigma^2 I$.

An estimator is any rule for estimating $\boldsymbol{\beta}$ based on the data $\boldsymbol{y}$. $\hat{\boldsymbol{\beta}}^{\text{OLS}} = X^+ \boldsymbol{y}$ is a natural estimator, of course, but there are others. For instance, we could ignore the data and always estimate $\boldsymbol{\beta}$ as $\mathbf{0}$. This is not a very good estimator, but an estimator nonetheless.

What makes $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ a better estimator than $\mathbf{0}$? A criterion that is often used to compare estimators is mean-squared error (MSE):

$$\text{MSE}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \text{E}(\hat{\beta}_j - \beta_j)^2. \tag{3}$$

The MSE measures how far the estimator is from the truth, in expectation. If one estimator has a smaller MSE than another, for all values of $\boldsymbol{\beta}$, then the first estimator is clearly better.

An important fact about the MSE is that it can be decomposed into the sum of a bias and a variance term:

$$\begin{aligned}
\text{MSE}_{\boldsymbol{\beta}}[\hat{\boldsymbol{\beta}}] &= \sum_{j=1}^{p} \text{E}(\text{E}[\hat{\beta}_j] - \beta_j + \hat{\beta}_j - \text{E}[\hat{\beta}_j])^2 \\
&= \sum_{j=1}^{p} \underbrace{(\text{E}[\hat{\beta}_j] - \beta_j)^2}_{\text{bias}^2} + \underbrace{\text{E}(\hat{\beta}_j - \text{E}[\hat{\beta}_j])^2}_{\text{Var}[\beta_j]} + 2 \underbrace{\text{E}[(\text{E}[\hat{\beta}_j] - \beta_j)(\hat{\beta}_j - \text{E}[\hat{\beta}_j])]}_{0}.
\end{aligned}$$

We can decrease MSE by either decreasing the variance or the bias. In particular, it is sometimes desirable to trade a little bias for a massive reduction in variance.

If we only consider estimators that are unbiased (i.e., $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$), then the problem of minimizing the MSE becomes one of minimizing the variance. Therefore, the "best" estimator is the one which minimizes the variance.

Unfortunately, it's not quite true that $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ minimizes the variance among all unbiased estimators. (It would be true if we assumed that the errors $\boldsymbol{\epsilon}$ are also normally distributed.) However, it *is* true that $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ minimizes the variance among all *linear* unbiased estimators. (A *linear estimator* is an estimator of the form $\hat{\boldsymbol{\beta}} = A\boldsymbol{y}$ for some matrix A.) For this reason, we say that $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is the best linear unbiased estimator (BLUE).

Theorem 1. *$\hat{\boldsymbol{\beta}}^{OLS}$ is the best linear unbiased estimator. That is, if $A\boldsymbol{y}$ is any other linear unbiased estimator, then*

$$\mathrm{MSE}_{\boldsymbol{\beta}}[A\boldsymbol{y}] = \sum_{j=1}^{p} \mathrm{Var}_{\boldsymbol{\beta}}[(A\boldsymbol{y})_j] > \sum_{j=1}^{p} \mathrm{Var}_{\boldsymbol{\beta}}[\hat{\beta}_j^{OLS}] = \mathrm{MSE}_{\boldsymbol{\beta}}[\hat{\boldsymbol{\beta}}^{OLS}]. \tag{4}$$

Proof. First, let's see what unbiasedness tells us about A. Since $\mathrm{E}[A\boldsymbol{y}] = AX\boldsymbol{\beta} = \boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, this means that $AX = I$.

Now, let's write the variance of $A\boldsymbol{y}$ in terms of the variance of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$:

$$\begin{aligned}\mathrm{Var}[A\boldsymbol{y}] &= \mathrm{Var}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}} + \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] \\ &= \mathrm{Var}[\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] + \mathrm{Var}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] + 2\,\mathrm{Cov}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}, \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}].\end{aligned}$$

The final term is zero:

$$\begin{aligned}\mathrm{Cov}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}, \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] &= \mathrm{Cov}[(A - X^+)\boldsymbol{y}, X^+\boldsymbol{y}] \\ &= (A - X^+)\,\mathrm{Var}(\boldsymbol{y})(X^+)^T \\ &= \sigma^2(A - X^+)(X^+)^T \\ &= \sigma^2(A - (X^TX)^{-1}X^T)(X(X^TX)^{-1}) \\ &= \sigma^2 \underbrace{(AX - I)}_{0}(X^TX)^{-1} \\ &= 0.\end{aligned}$$

Therefore, we have established that:

$$\mathrm{Var}[A\boldsymbol{y}] = \mathrm{Var}[\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] + \mathrm{Var}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}]. \tag{5}$$

The MSE is the sum of the diagonal entries of the variance matrix (i.e., the trace). By taking the trace of both sides of (5), we see that

$$\mathrm{MSE}[A\boldsymbol{y}] = \mathrm{tr}(\mathrm{Var}[A\boldsymbol{y}]) = \mathrm{tr}(\mathrm{Var}[\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}]) + \mathrm{tr}(\mathrm{Var}[A\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}]) = \mathrm{MSE}[\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}] + \begin{matrix}\text{something} \\ \text{positive}\end{matrix},$$

so $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ must have the smallest MSE among all estimators of the form $A\boldsymbol{y}$. □

2 As the Maximum Likelihood Estimator

If we assume that the errors $\boldsymbol{\epsilon}$ are normal, then under model (2), $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$. One general strategy for estimating parameters is maximum likelihood estimation. It yields estimators that are asymptotically efficient (meaning that they achieve the smallest possible variance as $n \to \infty$).

Theorem 2. *The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ when $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$ is $\hat{\boldsymbol{\beta}}^{OLS}$.*

Proof. $\boldsymbol{y}$ has a multivariate normal distribution with likelihood

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{n/2}(\det(\sigma^2 I))^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\sigma^2 I)^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\}.$$

We typically take logs before attempting to maximize the likelihood to make things easier.

$$\log L(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{y} - X\boldsymbol{\beta}||^2. \tag{6}$$

Notice that the first term is just a constant (it doesn't depend on $\boldsymbol{\beta}$). So maximizing (6) is equivalent to maximizing $-\frac{1}{2\sigma^2}||\boldsymbol{y} - X\boldsymbol{\beta}||^2$. To find the value of $\boldsymbol{\beta}$ that maximizes this, we can alternatively minimize $||\boldsymbol{y} - X\boldsymbol{\beta}||^2$. But this is simply $\hat{\boldsymbol{\beta}}^{\text{OLS}}$.

□

Remark 1. *In this example, there was a closed-form for the MLE because it reduced to a least-squares problem. However, most MLEs cannot be written down in closed form and have to be computed iteratively using optimization algorithms. All optimization algorithms work basically as follows: you start at a initial guess of the parameter and then iteratively move in a direction of increasing likelihood until you cannot increase the likelihood anymore. Two of the most commonly used methods (which we will use later in the course) are gradient descent and Newton's method.*

3 As a Sample Estimate of the MMSE Predictor

Suppose $\boldsymbol{x}_0$ and y_0 are random variables. We might want to predict y_0 using some function $\hat{f}(\boldsymbol{x}_0)$. A reasonable requirement is that this function minimize (mean square) prediction error, i.e.,

$$\underset{f}{\text{argmin}} \ \text{E}(y_0 - f(\boldsymbol{x}_0))^2.$$

It turns out that the minimum MSE (MMSE) predictor is the conditional expectation of y_0 given $\boldsymbol{x}_0$.

Theorem 3. *The MMSE predictor is the conditional expectation $f(\boldsymbol{x}_0) = \text{E}[y_0|\boldsymbol{x}_0]$.*

Proof. Add and subtract $\mathrm{E}[\boldsymbol{y}|\boldsymbol{x}]$, then expand:

$$\begin{aligned}\mathrm{E}(y_0 - f(\boldsymbol{x}_0))^2 &= \mathrm{E}(y_0 - \mathrm{E}[y_0|\boldsymbol{x}_0] + \mathrm{E}[y_0|\boldsymbol{x}_0] - f(\boldsymbol{x}_0))^2 \\ &= \mathrm{E}(y_0 - \mathrm{E}[y_0|\boldsymbol{x}_0])^2 + (\mathrm{E}[y_0|\boldsymbol{x}_0] - f(\boldsymbol{x}_0))^2 + 2\,\mathrm{E}[(y_0 - \mathrm{E}[y_0|\boldsymbol{x}_0])(\mathrm{E}(y_0|\boldsymbol{x}_0) - f(\boldsymbol{x}_0))]\end{aligned}$$

We can see that the last term is zero by conditioning further on $\boldsymbol{x}_0$. To avoid notational overload, we will denote $\mathrm{E}[y_0|\boldsymbol{x}_0]$ by $\mu(\boldsymbol{x}_0)$:

$$\begin{aligned}\mathrm{E}\left[(y_0 - \mu(\boldsymbol{x}_0))(\mu(\boldsymbol{x}_0) - f(\boldsymbol{x}_0))\right] &= \mathrm{E}\left[\mathrm{E}\left[(y_0 - \mu(\boldsymbol{x}_0))\ (\mu(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)) \mid \boldsymbol{x}_0\right]\right] \\ &= \mathrm{E}\left[(\mu(\boldsymbol{x}_0) - f(\boldsymbol{x}_0))\,\mathrm{E}\left[y_0 - \mu(\boldsymbol{x}_0) \mid \boldsymbol{x}_0\right]\right] \\ &= \mathrm{E}\left[(\mu(\boldsymbol{x}_0) - f(\boldsymbol{x}_0))\,\underbrace{(\mathrm{E}[y_0|\boldsymbol{x}_0] - \mu(\boldsymbol{x}_0))}_{0}\right] \\ &= 0.\end{aligned}$$

Thus, we have shown that

$$\mathrm{E}(y_0 - f(\boldsymbol{x}_0))^2 = \mathrm{E}(y_0 - \mathrm{E}[y_0|\boldsymbol{x}_0])^2 + (\mathrm{E}[y_0|\boldsymbol{x}_0] - f(\boldsymbol{x}_0))^2.$$

How should we choose f to minimize this? The first term on the right-hand side does not depend on f, so we have no control over it. But we can force the second term to zero by choosing $f(\boldsymbol{x}_0) = \mathrm{E}[y_0|\boldsymbol{x}_0]$. Hence, this must be the choice that minimizes the MSE. □

What does this conditional expectation look like? It depends on the distribution. But if $(\boldsymbol{x}_0, y_0)$ are jointly normal, then $\mathrm{E}[y_0|\boldsymbol{x}_0]$ is linear. In particular, suppose

$$\begin{pmatrix}\boldsymbol{x}_0 \\ y_0\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}\right). \tag{7}$$

Then, the conditional expectation and variance are given by

$$\mathrm{E}[y_0|\boldsymbol{x}_0] = \Sigma_{yx}\Sigma_{xx}^{-1}\boldsymbol{x}_0 \tag{8}$$

$$\mathrm{Var}[y_0|\boldsymbol{x}_0] = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{9}$$

Unfortunately, we usually cannot compute Σ_{xx}, Σ_{xy}, or Σ_{yy} in practice, as these are unknown population quantities. But they can be estimated from data $(\boldsymbol{x}_i, y_i)$:

$$\hat{\Sigma}_{xx} = \frac{1}{n}X^T X \tag{10}$$

$$\hat{\Sigma}_{xy} = \frac{1}{n}X^T \boldsymbol{y} \tag{11}$$

$$\hat{\Sigma}_{yy} = \frac{1}{n}\boldsymbol{y}^T \boldsymbol{y}. \tag{12}$$

Substituting these into (8), we estimate the MMSE predictor function as

$$\hat{f}(\boldsymbol{x}_0) = \frac{1}{n}\boldsymbol{y}^T X \left(\frac{1}{n}X^T X\right)^{-1} \boldsymbol{x}_0.$$

If you don't recognize this yet, try canceling the $\frac{1}{n}$'s and transposing it, which should be the same (since this is a 1×1 matrix). You should obtain:

$$\hat{f}(\boldsymbol{x}_0) = \boldsymbol{x}_0^T (X^T X)^{-1} X^T \boldsymbol{y} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}.$$

This is of course how you would predict y_0 once you had the OLS coefficients.