OLS Proof in Matrix Form

The following provides a concise proof for estimating OLS $\hat{\beta}$ coefficients first articulated by G. Udney Yule (1897).

Proof. We want to prove that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ minimizes the sum of squared errors for a set, $\boldsymbol{X}$, with dimensions $n \times (k+1)$.

1. First, we begin with the sum of the squared errors:

$$SSR(\boldsymbol{b}) \quad = \quad \sum_{t=1}^{n} \boldsymbol{u}_t^2 \tag{1}$$

   Note that equation (1) multiplies two column vectors together. As a rule, when we have two column vectors multiplied times one another, $\boldsymbol{a} \cdot \boldsymbol{b}$, the result is that we transpose the first vector such that $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}'\boldsymbol{b}$. Thus:

$$\sum_{t=1}^{n} \boldsymbol{u}_t^2 \quad = \quad \boldsymbol{u}'\boldsymbol{u}. \tag{2}$$

   Observe that we wind up with a row vector of dimension $1 \times n$ multiplied by a column vector with $n \times 1$ dimensions. Put differently, we were always going to end up with a scalar, or a single number, as the result of equation (1). For example, take a hypothetical $\boldsymbol{u_1} = \{1, 0, 1\}$, which is a row vector of dimension 1×3. If we transpose it, we get a column vector, $\boldsymbol{u_1'}$, which has dimensions 3×1. The resulting product between $\boldsymbol{u_1'}\boldsymbol{u_1}$ will have dimensions 1×1, or a scalar. This is true for any $\boldsymbol{u}$ of length n. In the above example, $\boldsymbol{u_1'}\boldsymbol{u_1} = 2$.

2. Our objective is to differentiate equation (2) with respect to the partial slope coefficients, $\boldsymbol{b}$ so that we may minimize the sum of squared errors:

$$\frac{\partial SSR}{\partial \boldsymbol{b}} = \frac{\boldsymbol{u}'\boldsymbol{u}}{\partial \boldsymbol{b}} = 0. \tag{3}$$

3. We can start by expanding the two expressions in $\boldsymbol{u}'\boldsymbol{u}$ and then, using substitution, gradually build out the expression we ultimately differentiate with respect to $\boldsymbol{b}$. Remember that in matrix algebra, $(AB)' = B'A'$. That fact comes into play going from the second to third expression in equation (4):

$$\begin{aligned}
\boldsymbol{u}'\boldsymbol{u} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}), \\
&= [\boldsymbol{y}' - (\boldsymbol{X}\boldsymbol{b})'](\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}), \\
&= (\boldsymbol{y}' - \boldsymbol{b}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}), \\
&= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{b} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}, \\
&= \boldsymbol{y}'\boldsymbol{y} - (\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y})' - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}, \qquad (4)
\end{aligned}$$

4. The last line in equation (4) holds due to our rules related to transposing matrix products. Suppose we have three matrixes, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, such that we multiply across them and take the transpose, $(\boldsymbol{ABC})'$. To accomplish this, we start by treating $\boldsymbol{AB}$ as its own matrix. So, $(\boldsymbol{ABC})' = \boldsymbol{C}'(\boldsymbol{AB})'$. And now we just need to calculate $(\boldsymbol{AB})'$ in the normal fashion, giving us: $(\boldsymbol{ABC})' = \boldsymbol{C}'\boldsymbol{B}'\boldsymbol{A}'$.

5. So take a look at the expression, $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{b}$ in the fourth line of equation (4). We want to prove that $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{b} = (\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y})'$, where the right-hand side of the previous equation derives from the final line of equation (4). So let's set it up:

$$\begin{aligned}
\boldsymbol{y}'\boldsymbol{X}\boldsymbol{b} &= (\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y})', \\
&= \boldsymbol{y}'(\boldsymbol{b}'\boldsymbol{X}')', \\
&= \boldsymbol{y}'\boldsymbol{X}\boldsymbol{b},
\end{aligned}$$

   which is true. Equation (4) holds.

6. Next, note that because the expression, $\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y}$ is a $[1 \times 1]$ scalar, we can express it as its transpose, $(\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y})'$. Thus, we can rewrite equation (4) as:

$$\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}. \qquad (5)$$

7. Now we can differentiate equation (5) with respect to the unknown vector of slope coefficients, $\boldsymbol{b}$:

$$\frac{\partial(\boldsymbol{u}'\boldsymbol{u})}{\partial\boldsymbol{b}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}, \qquad (6)$$

where the expression, $2\boldsymbol{X'Xb}$, is found due to the rules of matrix differentiation. First, let's recognize that $\boldsymbol{X}$ has dimensions, $n \times (k+1)$, and $\boldsymbol{X'}$ has dimensions $(k+1) \times n$. Let $\boldsymbol{Z} = \boldsymbol{X'X}$, which has dimensions $(k+1) \times (k+1)$. Therefore, $\boldsymbol{Z}$ is a square matrix. By the rules of matrix differentiation, if we differentiate some expression that takes the form, $f = \boldsymbol{x'Zx}$, then $\frac{\partial f}{\partial \boldsymbol{x}} = 2\boldsymbol{Zx}$. Using this generalization, we find that $\frac{\partial \boldsymbol{b'X'Xb}}{\partial \boldsymbol{b}} = 2\boldsymbol{Zb} = 2\boldsymbol{X'Xb}$.

8. Finally, we can solve for the optimal values of the slope coefficients, which I'll now denote as $\hat{\boldsymbol{\beta}}$, to indicate these are the values that satisfy the first order condition:

$$\begin{aligned} -2\boldsymbol{X'y} + 2\boldsymbol{X'X}\hat{\boldsymbol{\beta}} &= \boldsymbol{0}, \\ \boldsymbol{X'X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X'y}, \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}. \end{aligned} \tag{7}$$

Thus, so long as $\boldsymbol{X'X}$ is invertible—which is to say, so long as there is no perfect collinearity in $\boldsymbol{X}$—then $(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ produces the values in $\hat{\boldsymbol{\beta}}$ that minimize the sum of squared errors.

□