OLS Regression and Its Properties

2021-09-02

We have previously defined the population regression model (using scalar algebra) as:

$$y_i = \beta_0 + \beta_1(x_i) + \epsilon_i$$

where the outcome (y) is assumed to be statistically and linearly related to the predictor (x) and ϵ. We also assume that the error term, ϵ, is a random variable.

Recall that the least squares estimators can be analytically computed as:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1(\bar{x})$$

Representing the Population Regression Model Using Matrix Algebra

Using the subject-specific subscripts ($1, 2, 3, \dots, n$), we can write out each subject's equation:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1(x_1) + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1(x_2) + \epsilon_2 \\
y_3 &= \beta_0 + \beta_1(x_3) + \epsilon_3 \\
\vdots \quad &\quad \vdots \quad\quad \vdots \quad\quad\quad \vdots \\
y_n &= \beta_0 + \beta_1(x_n) + \epsilon_n
\end{aligned}$$

These can be arranged into a set of vectors and matrices, namely,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where,

- $\mathbf{y}$ is an $n \times 1$ vector of observations on the outcome variable.

- $\mathbf{X}$ is an $n \times k$ matrix (called the *design matrix*) consisting of a column of ones and the observations for *k* independent predictors. In the simple regression example, $k = 2$, and the design matrix has two columns—a column of ones and a column of observations for the predictor *X*.
- β is a $k \times 1$ vector of unknown population parameters that we want to estimate. In the simpke regression model, $\mathbf{b}$ is a 2×1 vector consisting of β_0 and β_1.
- ϵ is a $n \times 1$ vector of residuals.

Estimating the Regression Coefficients

In a regression analysis, one goal is often to estimate the values of the parameters in the β vector using sample data (i.e., the *y* and *x* values. We denote the estimates of the regression parameters using the roman letters; the vector of the sample estimates for the β-values are denoted as $\mathbf{b}$. Similarly the sample residuals are denoted as $\mathbf{e}$ rather than ϵ. (It is common to refer to the population errors as "errors" and the sample estimates as "residuals.") Thus, the sample equivalent of the model is:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

In ordinary least squares (OLS) estimation, the estimated coefficients minimize the sum of the squared *sample residuals* (i.e., the SSE). Using scalar algebra, the SSE can be expressed as: SSE $= \sum e_i^2$. The SSE can be expressed in matrix notation as:

$$\begin{aligned} \text{SSE} &= \mathbf{e}^\top \mathbf{e} \\ &= \begin{bmatrix} e_1 & e_2 & e_3 & \dots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \end{aligned}$$

Re-arranging the sample regression equation, we can express the residual vector as $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$. The SSE can then be expressed as:

$$\text{SSE} = (\mathbf{y} - \mathbf{Xb})^\top (\mathbf{y} - \mathbf{Xb})$$

This can be re-written as:

$$\begin{aligned} \text{SSE} &= \mathbf{y}^\top \mathbf{y} - \mathbf{b}^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{Xb} + \mathbf{b}^\top \mathbf{X}^\top \mathbf{Xb} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{b}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{b}^\top \mathbf{X}^\top \mathbf{Xb} \end{aligned}$$

To find the values for the elements in $\mathbf{b}$ that minimize the equation, we use calculus to differentiate this expression with respect to $\mathbf{b}$

> Although calculus, especially calculus on matrices, is beyond the scope of this course, Fox (2009) gives the interested reader some mathematical background on optimization (i.e., minimizing). For now you just need to understand we can optimize a function by computing its derivative, setting the derivative equal to 0, and solve for any remaining unknowns.

This gives the expression:

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{y}$$

This expression is referred to as the *Normal Equations*. Note that the $(\mathbf{X}^\top\mathbf{X})$ matrix has two important properties:

- It is square; and
- It is symmetric.

To solve for the elements in $\mathbf{b}$, we pre-multiply both sides of the equation by $(\mathbf{X}^\top\mathbf{X})^{-1}$.

$$(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$$

$$\mathbf{I}\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

As long as $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists, the vector of regression coefficients is given as:

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

This implies that the vector of regression coefficients can be obtained directly through manipulation of the design matrix and the vector of outcomes. In other words, the OLS coefficients is a direct function of the data. Note that as of yet, we have made no assumptions about the residuals. The coefficients can be estimated making no assumptions about the distributions of the residuals.

Extending the Model

Using matrix algebra to compute the OLS regression coefficients gives us the same values as using the analytic formulas. So why use matrix algebra? The simple reason is that we can use the same matrix algebra computation of $\mathbf{b}$ regardless of how many predictors we include in the model (it is extensible). The analytic formulas change and become quite difficult to manipulate. For example, consider an example where we want to estimate the coefficients for a model that includes two main effects (x_1 and x_2) and an interaction between these effects. The population model written in scalar algebra is:

$$y_i = \beta_0 + \beta_1(x_{1i}) + \beta_2(x_{2i}) + \beta_3(x_{1i})(x_{2i}) + \epsilon_i$$

If we express this using matrix notation, we get:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}(x_{21}) \\ 1 & x_{12} & x_{22} & x_{12}(x_{22}) \\ 1 & x_{13} & x_{23} & x_{13}(x_{23}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n}(x_{2n}) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Adding predictors expands the size of the design matrix and the length of the β matrix, but the compact notation ($\mathbf{y} = \mathbf{X}\beta + \epsilon$) is exactly the same, so estimating the values in the **b** vector for multiple regression models is identical to doing so for the simple regression model!

Properties of the OLS Estimators

One property of the OLS estimators (in simple or multiple regression) is that they minimize the sum of squared residuals. There are also several other properties that the OLS estimators have. (Note: We derive these properties for the simple regression model, but they also can be extended for the multiple regression model.) Remember these estimators are based on the normal equations:

$$\mathbf{X}^\top\mathbf{X}\mathbf{b} = \mathbf{X}^\top\mathbf{y}$$

If we substitute $\mathbf{X}\mathbf{b} + \mathbf{e}$ in for $\mathbf{y}$ in this expression, we get:

$$\begin{aligned}\mathbf{X}^\top\mathbf{X}\mathbf{b} &= \mathbf{X}^\top(\mathbf{X}\mathbf{b} + \mathbf{e}) \\ &= \mathbf{X}^\top\mathbf{X}\mathbf{b} + \mathbf{X}^\top\mathbf{e}\end{aligned}$$

To make this equality work, implies that:

$$\mathbf{X}^\top\mathbf{e} = \mathbf{0}$$

Let's examine $\mathbf{X}^\top\mathbf{e}$:

$$\mathbf{X}^\top\mathbf{e} = \mathbf{0}$$

$$\begin{bmatrix} 1 & 1 & 1 & ... & 1 \\ X_1 & X_2 & X_3 & ... & X_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{0}$$

$$\begin{bmatrix} e_1 + e_2 + e_3 + ... + e_n \\ X_1e_1 + X_2e_2 + X_3e_3 + ... + X_ne_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This implies that for every column in the design matrix, $\mathbf{X}_k$, that $\mathbf{X}_k^\top\mathbf{e} = 0$. In other words, the dot product between $\mathbf{X}_k$ and **e** is zero indicating that the two vectors are independent.

P.1: The observed values of the predictor(s) are uncorrelated with the sample residuals.

Note that this does not mean that the predictor(s) are uncorrelated with the residuals in the population; that is an assumption we will have to make later on.

Properties of the OLS Regressors

If the regression model includes an intercept (the first column of the design matrix is a ones vector) then the following properties also hold.

P.2: The sum of the sample residuals is 0.

If the first column of the design matrix is a ones vector, then the first element of the $\mathbf{X}^\top\mathbf{e}$ matrix is $e_1+e_2+e_3+...+e_n = \sum e_i$, which is equal to zero since $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$.

P.3: The mean of the sample residuals is zero.

Since the mean of the residuals is computed as $\bar{\mathbf{e}} = \dfrac{\sum e_i}{n}$, and the sum (numerator) is zero, then the mean is also zero.

P.4: The regression line passes through the point $(\bar{X}, \bar{Y})$.

Remember that $\mathbf{e} = \mathbf{y} - \mathbf{bX}$. This means that:

$$\begin{aligned}\sum \mathbf{e} &= \sum \Big(\mathbf{y} - \mathbf{Xb}\Big) \\ &= \sum \mathbf{y} - \sum \mathbf{Xb} \\ &= \sum \mathbf{y} - \mathbf{b}\sum \mathbf{X}\end{aligned}$$

If we divide this expression by *n*, we get

$$\begin{aligned}\frac{\sum \mathbf{e}}{n} &= \frac{\sum \mathbf{y}}{n} - \frac{\mathbf{b}\sum \mathbf{X}}{n} \\ \bar{\mathbf{e}} &= \bar{y} - \mathbf{b}\bar{x}\end{aligned}$$

But, the mean of the residuals is zero, so:

$$\begin{aligned}0 &= \bar{y} - \mathbf{b}\bar{x} \\ \bar{y} &= \mathbf{b}\bar{x}\end{aligned}$$

That is, the predicted *y*-value when the mean of *X* is used as a predictor is the mean of *Y*. In other words, the point $(\bar{X}, \bar{Y})$ is on the regression line.

P.5: The predicted y-values are uncorrelated with the sample residuals.

Since $\hat{\mathbf{y}} = \mathbf{Xb}$ then $\hat{\mathbf{y}}^\top = (\mathbf{Xb})^\top$. If we post-multiply both sides of this expression by the residual vector **e**, we get:

$$\hat{\mathbf{y}}^\top \mathbf{e} = (\mathbf{X}\mathbf{b})^\top \mathbf{e}$$

$$= \mathbf{b}^\top \mathbf{X}^\top \mathbf{e}$$

Since $\mathbf{X}^\top \mathbf{e} = 0$, then $\hat{\mathbf{y}}^\top \mathbf{e} = 0$. This implies that $\hat{\mathbf{y}}$ and $\mathbf{e}$ are uncorrelated.

P.6: The mean of the predicted y-values is equal to the mean of the observed y-values.

We can make use of the fact that $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$. Taking the sum of both sides of the expression and dividing by n, we get:

$$\frac{\sum \mathbf{y}}{n} = \frac{\sum \hat{\mathbf{y}}}{n} + \frac{\sum \mathbf{e}}{n}$$

$$\bar{y} = \bar{\hat{y}} + 0$$

$$\bar{y} = \bar{\hat{y}}$$

IMPORTANT: These properties will always be true. They do not rely on any distributional assumptions of the residuals. Furthermore, these properties do not tell us anything about how "good" the coefficient estimates (**b**) are. Nor do these properties allow us to make inferences about the true parameters (β).

References

Fox, J. (2009). *A mathematical primer for social statistics*. Sage.