We start from the linear regression model.

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$$

The first part,

$$\beta_0 + \beta_1 \cdot X_i$$

is called the **Population Regression Function (PRF)**. It shows the effect of X on Y assuming no other factors affect Y. The final part,

$$u_i$$

is the **error term**. It includes all the factors *other than* X that affects Y.

Suppose we arbitrarily pick estimates of β_0 and β_1. Let's denote these estimates as $\hat{\beta}_0$ and $\hat{\beta}_1$. How do we know if they are "good" estimates? One criteria is to measure how well they predict Y. Define the **predicted value** of Y, which we will call $\hat{Y}$ ("yhat"), as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i$$

Define the difference between the actual and predicted values of Y as $\hat{u}$ ("uhat"), called **residuals**.

$$\hat{u}_i = Y_i - \hat{Y}_i$$

"Good" estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ should preduce "good" predictions and therefore "small" residuals. But how do we measure if the residuals are small? One way is simply to sum them and pick the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize this sum. The problem with this method is that good and bad predictions may produce the same sum of residuals. Suppose we have a data set of two observations, and that the first set of estimates of β_0 and β_1 produces residuals of -1 and 1. Also suppose another set of estimates of β_0 and β_1 produces residuals of -8 and 8. They both sum to zero, yet the first set of predictions are clearly better than the second. The problem is that the negative and positive residuals cancel each other out. One solution is to take the absolute value of the residuals, and pick estimates of β_0 and β_1 that minimuze their sum. This works and is the best method in some situations but is messy. Instead, the solution used most often is to square the residuals, and pick esimates of β_0 and β_1 that minimize the sum of squared residuals, or RSS. That is find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

With only two observations this sum is

$$RSS = \left(Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1\right)^2 + \left(Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2\right)^2$$

To find the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize this sum, we take derivatives of this with respect to $\hat{\beta}_0$ and with respect to $\hat{\beta}_1$ and set both equal to zero. Let's start with $\hat{\beta}_0$.

$$\frac{d}{d\hat{\beta}_0} RSS = 2 \cdot \left(Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1\right) \cdot (-1) + 2 \cdot \left(Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2\right) \cdot (-1) = 0$$

Recall that

$$Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1 = \hat{u}_1$$

and

$$Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2 = \hat{u}_2$$

Then

$$\begin{aligned} 0 &= -2 \cdot \hat{u}_1 - 2 \cdot \hat{u}_2 \\ \frac{0}{-2} &= \frac{-2}{-2} \cdot \hat{u}_1 \frac{-2}{-2} \cdot \hat{u}_2 \\ 0 &= \hat{u}_1 + \hat{u}_2 \end{aligned}$$

Thus, ***the residuals must sum to zero***.

Now take the derivative with respect to $\hat{\beta}_1$ and set equal to zero.

$$
\begin{aligned}
0 &= \frac{d}{d\hat{\beta}_1}\left[\left(Y_1 - \hat{\beta_0} - \hat{\beta}_1 X_1\right)^2 + \left(Y_2 - \hat{\beta_0} - \hat{\beta}_1 X_2\right)^2\right] \\
&= 2 \cdot \left(Y_1 - \hat{\beta_0} - \hat{\beta}_1 X_1\right)(-X_1) + 2 \cdot \left(Y_2 - \hat{\beta_0} - \hat{\beta}_1 X_2\right)(-X_2) \\
&= -2\hat{u}_1 X_1 - 2\hat{u}_2 X_2 \\
&= \hat{u}_1 X_1 + \hat{u} X_2
\end{aligned}
$$

Thus***, the product of the residuals and X must sum to zero.***

Recall

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

Sum both sides and divide by n. This is the same as averaging both sides. Note that the average of a constant is itself. Also, within one particular sample, $\hat{\beta}_0$ and $\hat{\beta}_1$ are constants.

$$
\begin{aligned}
\frac{\sum Y_i}{n} &= \frac{\sum \hat{\beta}_0}{n} + \frac{\sum \hat{\beta}_1 X_i}{n} + \frac{\sum \hat{u}_i}{n} \\
\overline{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \overline{X} + 0
\end{aligned}
$$

Subtract this equation from the equation above.

$$
\begin{aligned}
Y_i - \overline{Y} &= \hat{\beta}_0 - \hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_1 \overline{X} + \hat{u}_i \\
Y_c &= \hat{\beta}_1 X_i - \hat{\beta}_1 \overline{X} + \hat{u}_i \\
Y_c &= \hat{\beta}_1 X_c + \hat{u}_i
\end{aligned}
$$

Multiply both sides by X_c and sum.

$$
\begin{aligned}
X_c Y_c &= \hat{\beta}_1 X_c^2 + X_c \hat{u} \\
\sum X_c Y_c &= \hat{\beta}_1 \sum X_c^2 + \sum X_c \hat{u}
\end{aligned}
$$

Recall that

$$\sum X_c \hat{u} = 0$$

Then

$$
\begin{aligned}
\sum X_c Y_c &= \hat{\beta}_1 \sum X_c^2 \\
\hat{\beta}_1 &= \frac{\sum X_c Y_c}{\sum X_c^2}
\end{aligned}
$$

Recall

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$$

Thus

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Because these estimators minimize the sum of square residuals they are called the **ordinary least squares** estimators, or **OLS**.

Remember the two key points from the derivation. If we calculate $\hat{\beta}_0$and $\hat{\beta}_1$ correctly, then

1. $\sum_{i=1}^{n} \hat{u}_i = 0$
2. $\sum_{i=1}^{n} \hat{u}_i X_{ci} = 0$