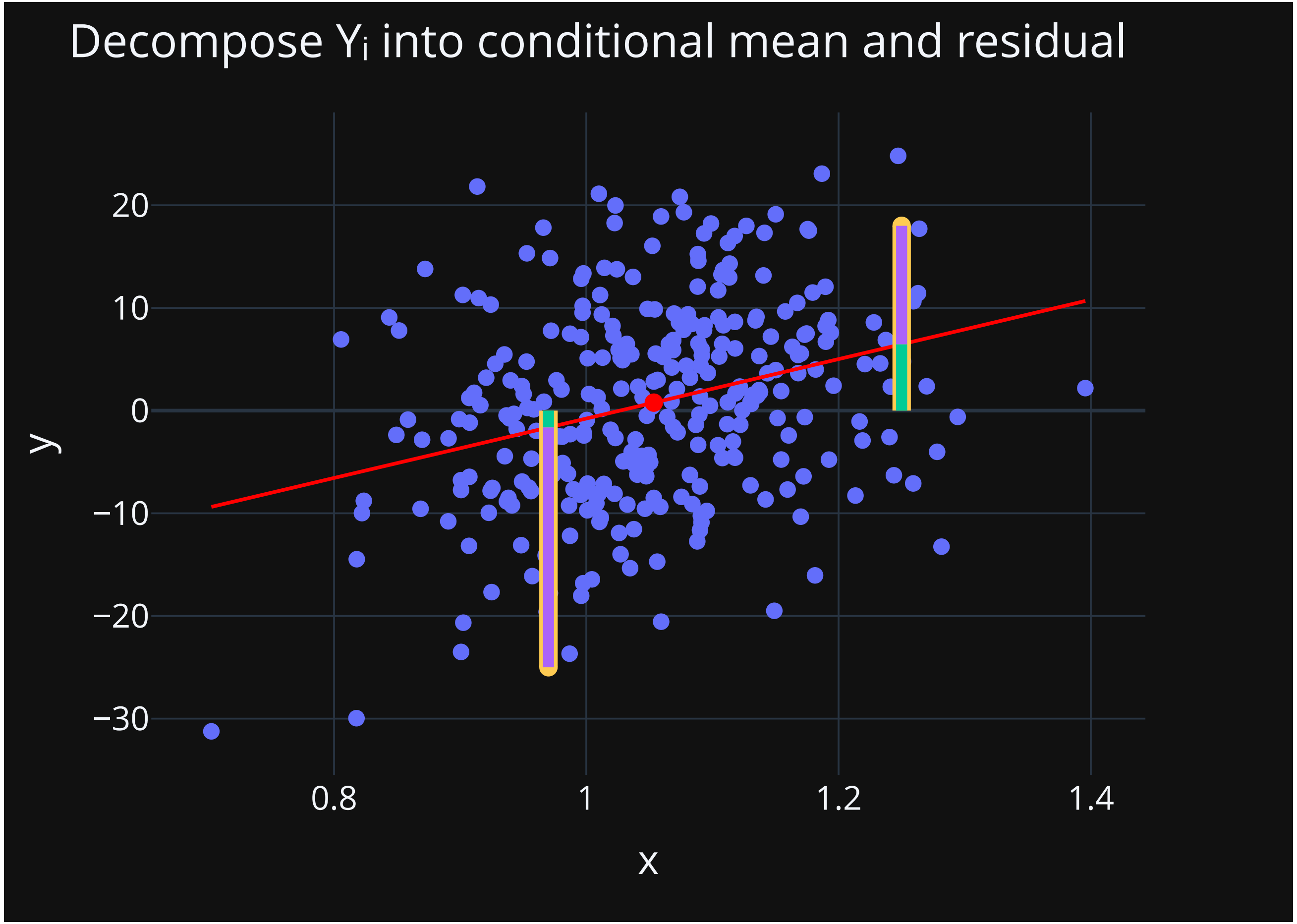
Applied Data Analytics

Statistics — Measures for bivariate data

Ordinary Least Squares (OLS): Derivation

Hans-Martin von Gaudecker and Aapo Stenhammar



Decompose Yᵢ into conditional mean and residual
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x

OLS: Setup

Rewrite

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

as

$$U_i = Y_i - \beta_0 - \beta_1 X_i$$

and pick β_0, β_1 as to minimize the sum of squares of the U_i

Minimisation problem

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1) &= \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} U_i^2 \\
&= \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2
\end{aligned}
$$

Steps for $\hat{\beta}_0$

Differentiate objective function with respect to b_0:

$$\frac{\partial \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2}{\partial b_0} = \sum_{i=1}^{n} 2 \cdot (Y_i - b_0 - b_1 X_i) \cdot (-1)$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the values solving this:

$$\sum_{i=1}^{n} -2 \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \stackrel{!}{=} 0$$

Divide by -2

Steps for $\hat{\beta}_0$

$$\sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = 0$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_i = 0$$

$$\sum_{i=1}^{n} Y_i - \hat{\beta}_0 \sum_{i=1}^{n} 1 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} Y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

Steps for $\hat{\beta}_0$

Divide by n and rearrange:

$$\sum_{i=1}^{n} Y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

Steps for $\hat{\beta}_1$

Differentiate objective function with respect to b_1:

$$\frac{\partial \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2}{\partial b_1} = \sum_{i=1}^{n} 2 \cdot \left(Y_i - b_0 - b_1 X_i\right) \cdot \left(-X_i\right)$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the values solving this:

$$\sum_{i=1}^{n} -2 \cdot \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i \stackrel{!}{=} 0$$

Divide by -2, multiply out

Steps for $\hat{\beta}_1$

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 X_i - \hat{\beta}_1 X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i\right) = 0$$

Focus on first two terms

$$
\begin{aligned}
&\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} \bar{X} Y_i + \sum_{i=1}^{n} \bar{X} Y_i \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} \bar{X} Y_i + \bar{X} \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} Y_i
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} \bar{X} Y_i + \bar{X} \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} Y_i \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} \bar{X} Y_i + n \bar{X} \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \bar{Y} - \sum_{i=1}^{n} \bar{X} Y_i + \sum_{i=1}^{n} \bar{X} \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X} \bar{Y} \\
&= \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \cdot \left(Y_i - \bar{Y}\right)
\end{aligned}
$$

Back to the equation for $\hat{\beta}_1$

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i \right) = 0$$

$$\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \cdot \left(Y_i - \bar{Y}\right) - \hat{\beta}_1 \cdot \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = 0$$

$$s_{XY} - \hat{\beta}_1 \cdot s_X^2 = 0$$

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

Implications: $\bar{U} = 0$, $s_{X,U} = 0$

Remember the FOC's:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = \frac{1}{n}\sum_{i=1}^{n} U_i = \bar{U} = 0$$

$$\frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i = \frac{1}{n-1}\sum_{i=1}^{n} U_i X_i = s_{X,U} = 0$$

Second equation follows from same argument as above for $\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}$ and observing that $\bar{U} = 0$.

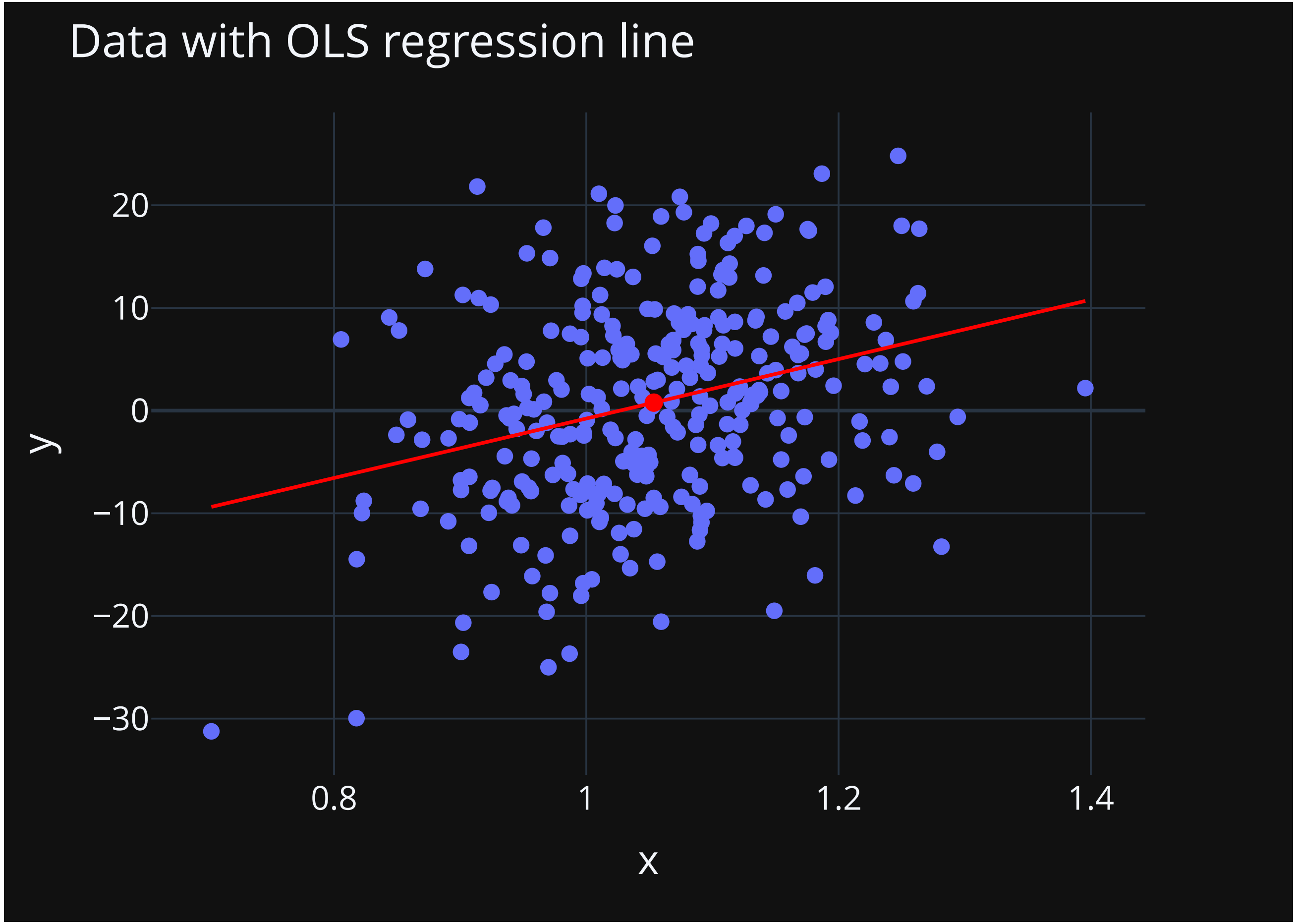
OLS estimator

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

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

These are the intercept and slope of the regression line:

- $\hat{\beta}_0$ is the mean value of Y when X is zero
- $\hat{\beta}_1$ is the mean change in Y when X changes by one unit


Data with OLS regression line
20
10
0
−10
−20
−30
y
0.8
1
1.2
1.4
x