Make the first regression model explicit

Define the regression equation, its coefficients and the error term.

> Population: $Y_i = \beta_0 + \beta_1 X_i + u_i$

> Conditional-mean assumption: $E(u_i \mid X_i) = 0$

> Fit: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$; residual: $e_i = Y_i - \widehat{Y}_i$

A population conditional mean is different from a fitted sample line

The error is an unobserved population remainder; the residual follows an estimated line.

> Population: $Y = \beta_0 + \beta_1 X + u$, with $E[u \mid X] = 0$.

> Population conditional mean: $E[Y \mid X = x] = \beta_0 + \beta_1 x$.

> Sample: $\widehat{Y_i} = b_0 + b_1 X_i$; residual: $\widehat{u_i} = Y_i - \widehat{Y_i}$.

Inspect variables before running the model

Check the data, then use a saved do-file to reproduce the estimates.

> Check observation counts, units, ranges and missing values.

> Outcome first; regressors afterward.

A slope is a change with units

A fitted change equals the slope times the change in the regressor.

> $\Delta X = \text{new ratio} - \text{old ratio}$

> $22 \to 20$ gives $\Delta X = -2$

> Fitted change: $b_1 \Delta X \approx +4.56$ score points

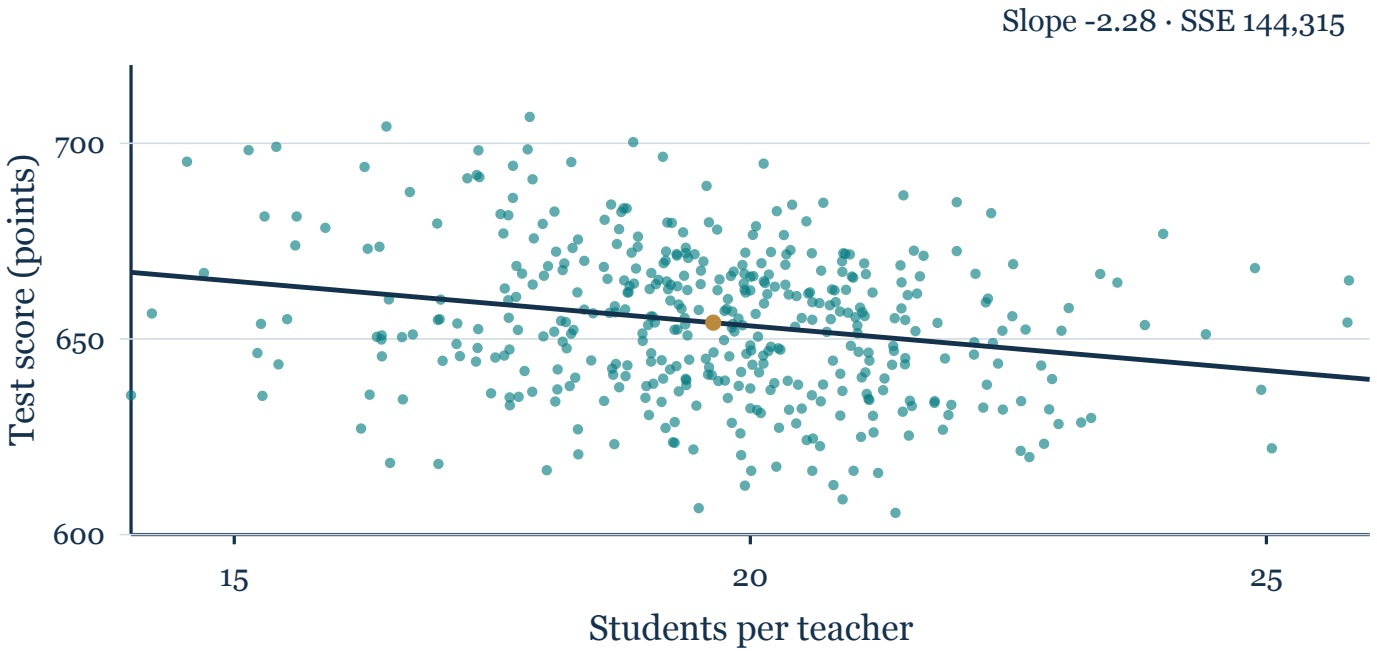
Practice

Use the same slope to calculate the fitted score change when the ratio rises from 20 to 23.

Explanation: The change in the ratio is +3. The fitted score change is 3×(−2.279808), about −6.84 score points. Moving right along the downward-sloping line lowers its fitted score.

Choose a line by its errors

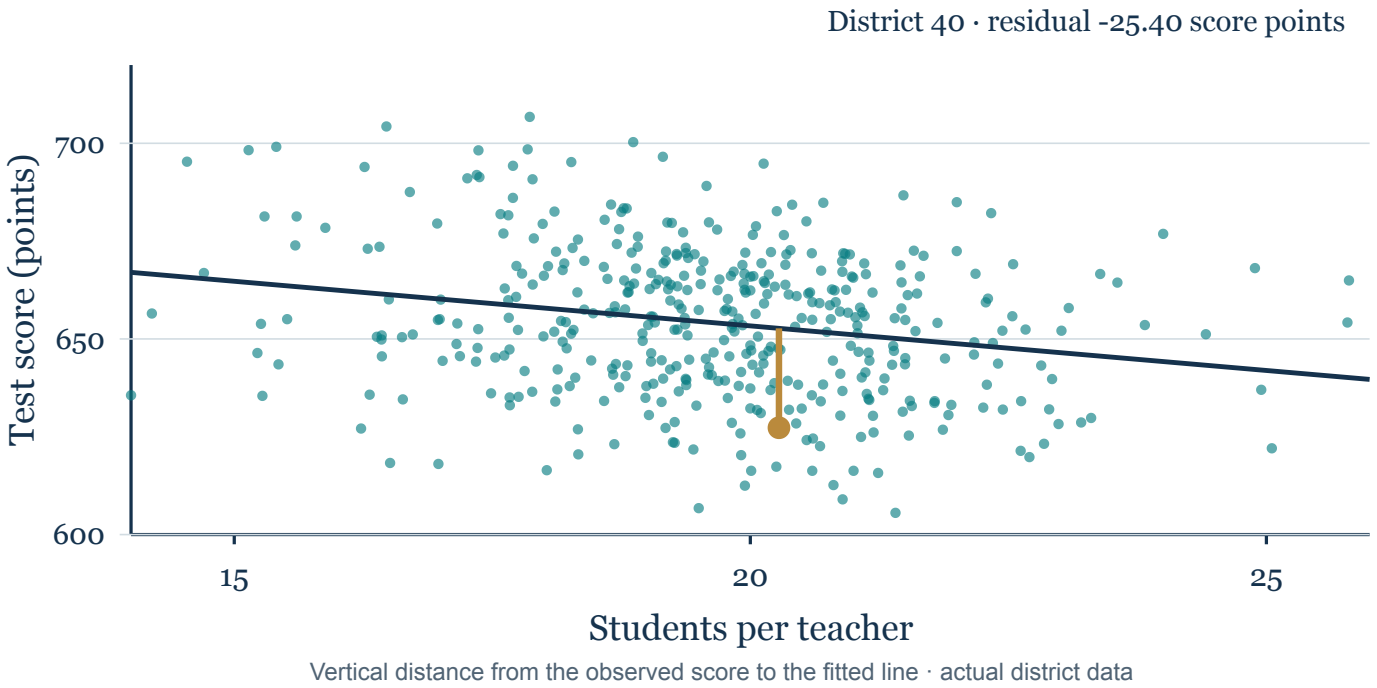
OLS chooses the line with the smallest sum of squared residuals.



Rotate around the mean point. The loss uses all 420 observed districts.

Follow one residual

For one district, compare the observed score with the score on the line.



Calculate the OLS slope from the data

Compare each variable with its average, then combine the deviations.

$$b_1 = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Without regressor variation, the slope cannot be separately estimated.

The normal equations explain the slope formula

Differentiate the squared-error criterion, then solve the two conditions.

> $S(b_0, b_1) = \sum_i (Y_i - b_0 - b_1 X_i)^2.$

> $\partial S / \partial b_0 = -2 \sum_i \widehat{u}_i = 0, \partial S / \partial b_1 = -2 \sum_i X_i \widehat{u}_i = 0.$

> $b_0 = \overline{Y} - b_1 \overline{X}, b_1 = \dfrac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2}.$

A small calculation you can finish

Use three constructed districts to see the formula work.

Practice

Find the slope, intercept, fitted scores and residuals for X=(18,20,22) and Y=(680,670,650).

Explanation: b_1=−7.5 and b_0=816⅔. Fitted scores are 681⅔, 666⅔ and 651⅔. Residuals are −1⅔, +3⅓ and −1⅔, which sum to zero. Small differences from rounding are acceptable.

Zero conditional mean is stronger than zero covariance

An average restriction need not hold at every regressor value.

> $\mathrm{Cov}(X, u) = E[(X - EX)(u - Eu)]$.

> $E[u \mid X] = 0 \Rightarrow E[Xu] = E[X E(u \mid X)] = 0$.

> Counterexample: $P(X = -1) = P(X = 0) = P(X = 1) = 1/3$, $u = X^2 - 2/3$.

With an intercept, fitted and residual variation add up

Define the sums of squares before reading fit statistics.

> $SST = \sum_i (Y_i - \overline{Y})^2$, $SSR = \sum_i \widehat{u}_i^2$, $ESS = \sum_i (\widehat{Y}_i - \overline{Y})^2$.

> With an intercept: $\sum_i (\widehat{Y}_i - \overline{Y})\widehat{u}_i = 0$, so $SST = ESS + SSR$.

> $R^2 = 1 - SSR/SST$, for $SST > 0$.

Rescaling a variable changes coefficients and preserves fitted values

Units belong to the coefficient, not to the fit.

> $X^\star = cX \Rightarrow b_1^\star = b_1/c,\, c \neq 0.$

> $Y^\star = aY \Rightarrow b_0^\star = ab_0,\ b_1^\star = ab_1,\, a > 0.$

> Changing units preserves the translated fit; removing an intercept changes the model.

Read a genuine Stata result

Read the fitted coefficients and their units in the actual Stata display.

```
. quietly regress score str

. matrix list e(b), format(%10.6f)

e(b)[1,2]
           str       _cons
y1   -2.279808  698.932953
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

> $\widehat{\beta}_1 = -2.279808$ points per student per teacher

> Fitted contrast: $(-2)(-2.279808) = 4.559616$ points

Calculate the policy-sized fitted difference

lincom translates a slope into the particular comparison we want.

```
. display %10.6f -2*_b[str]
  4.559617
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

$$\Delta X = 20 - 22 = -2$$

$$\Delta \widehat{Y} = \widehat{\beta}_1 \Delta X = 4.559616 \text{ points}$$

Logs provide a second measurement scale for a simple regression

Define the logarithm before using percentage language.

> $\log Y_1 - \log Y_0 = \log(Y_1/Y_0)$; exact proportional change: $\exp(d) - 1$.

> Log-level: $\log Y = b_0 + b_1 X$; one-unit fitted change: $100[\exp(b_1) - 1]\%$.

> Log-log: $\log Y = b_0 + b_1 \log X$; b_1 is a local elasticity.

Check the normal equations and fit decomposition

Try the comparison before opening the explanation.

Practice

Estimate score on str with an intercept using schools.csv. Verify that the residual sum and staffing-residual cross-product are approximately zero, reconstruct SST and SSR, and recover R squared. Explain why these numerical identities do not establish exogeneity.

Check the normal equations and fit decomposition: explanation

Follow the calculation and interpret its units.

> OLS with an intercept makes the two sample cross-products zero up to numerical precision. Here SST is about 152109.594, SSR about 144315.484, and reconstructed R squared is about 0.0512401, matching Stata. The overleaf_technical.do reference demonstrates the calculation. These are consequences of minimisation, not evidence that the unknown population error has zero conditional mean.