ECON 2041
Introductory Econometrics

Drawing the line

Plan for the semester

Plan for today

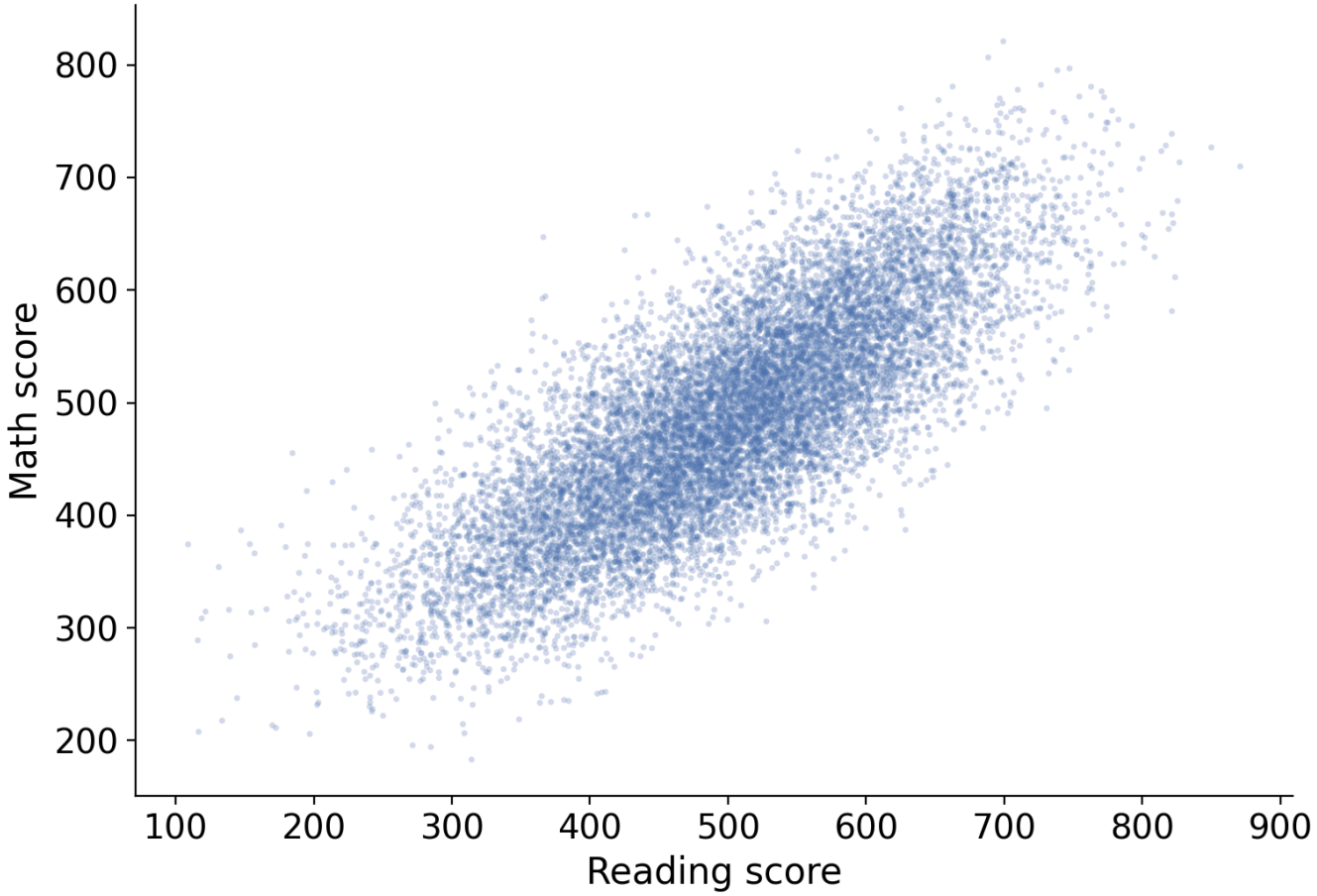
- A correlation gallery
- Where a covariance comes from
- You draw the line
- Conditional means become a line
- Residuals, and what least squares means

Can your eye see a correlation?

Three correlations

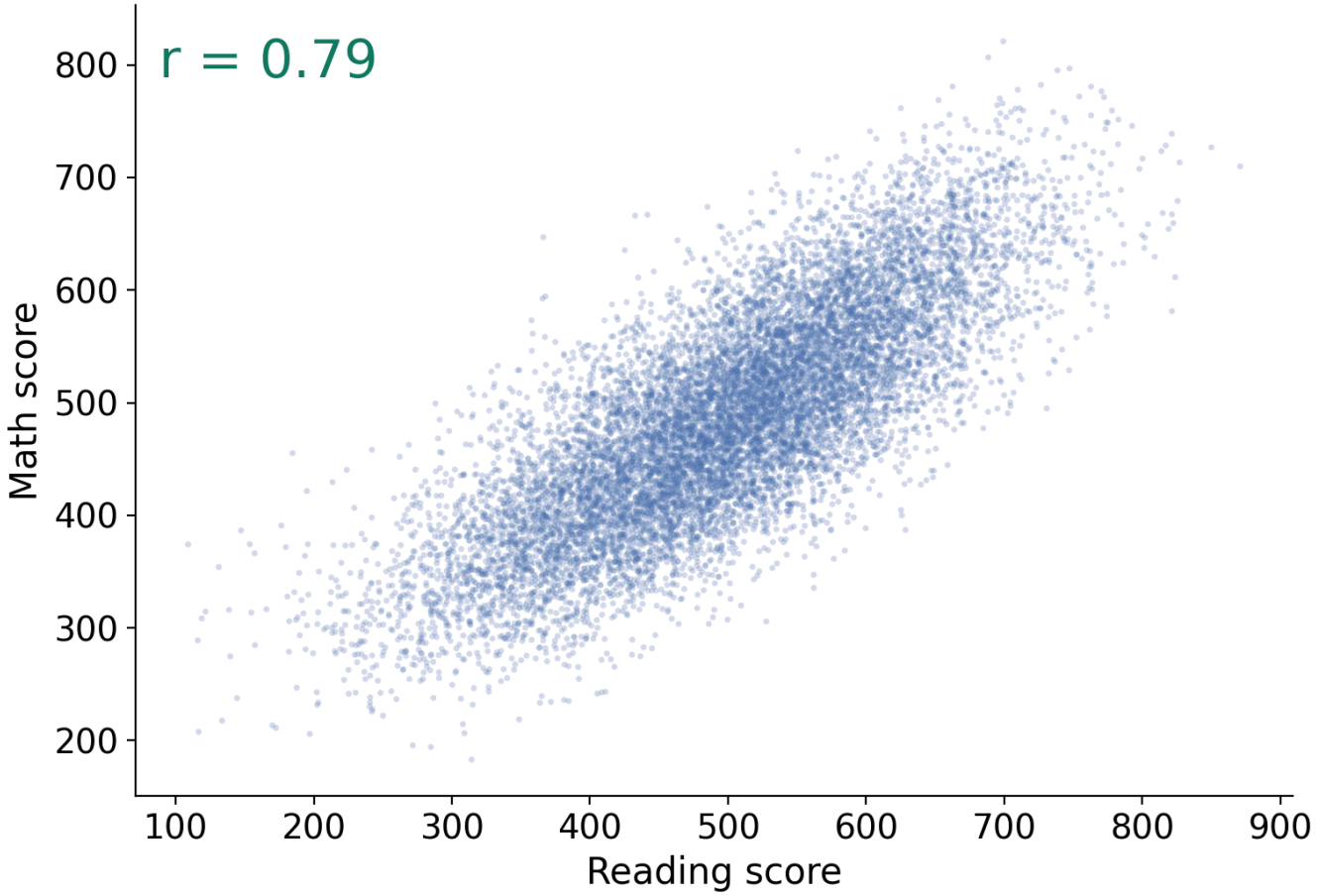
For each one: guess the correlation

Gallery 1: math & reading



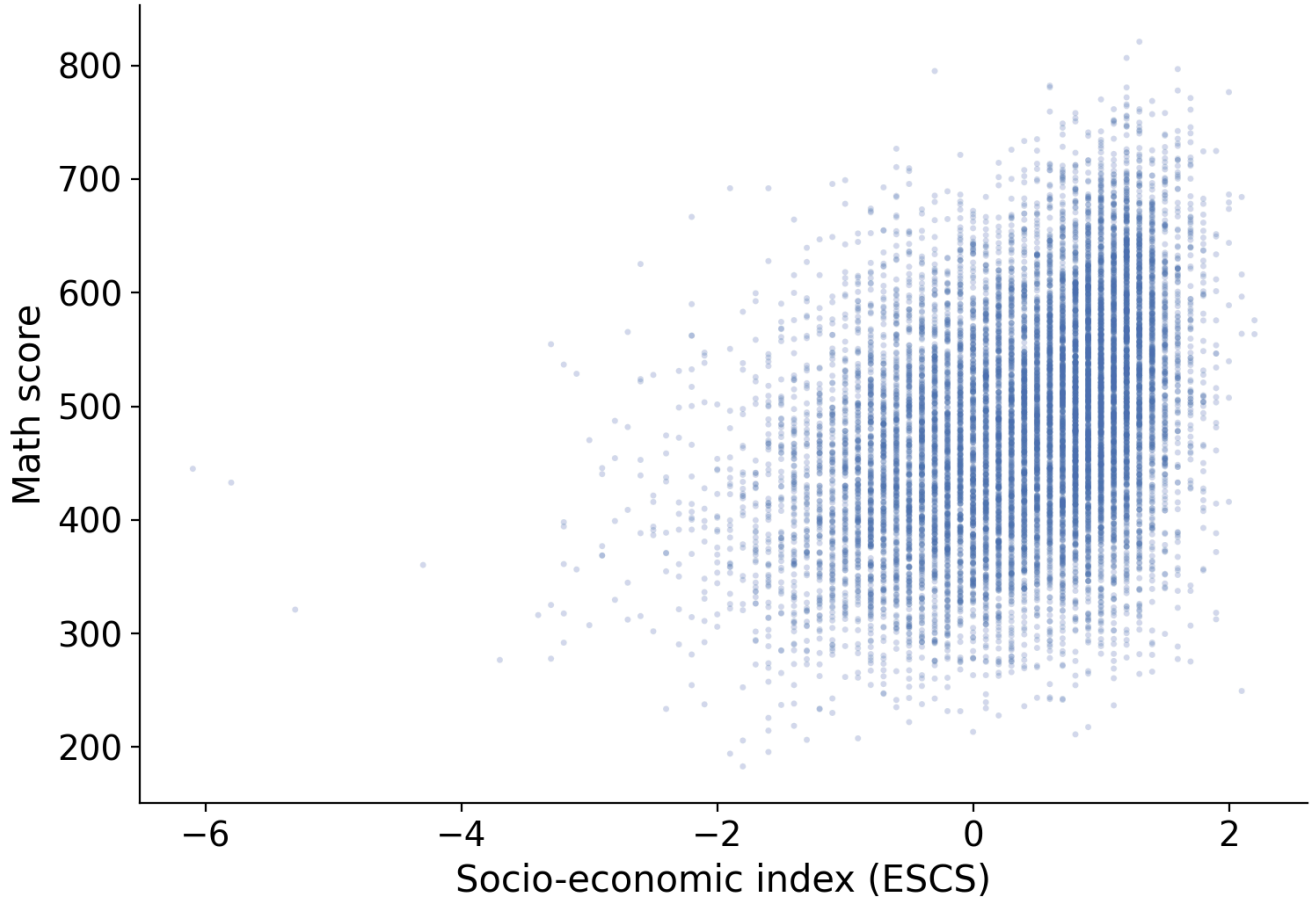
Data: PISA 2022, Australia (OECD)

Gallery 1: math & reading



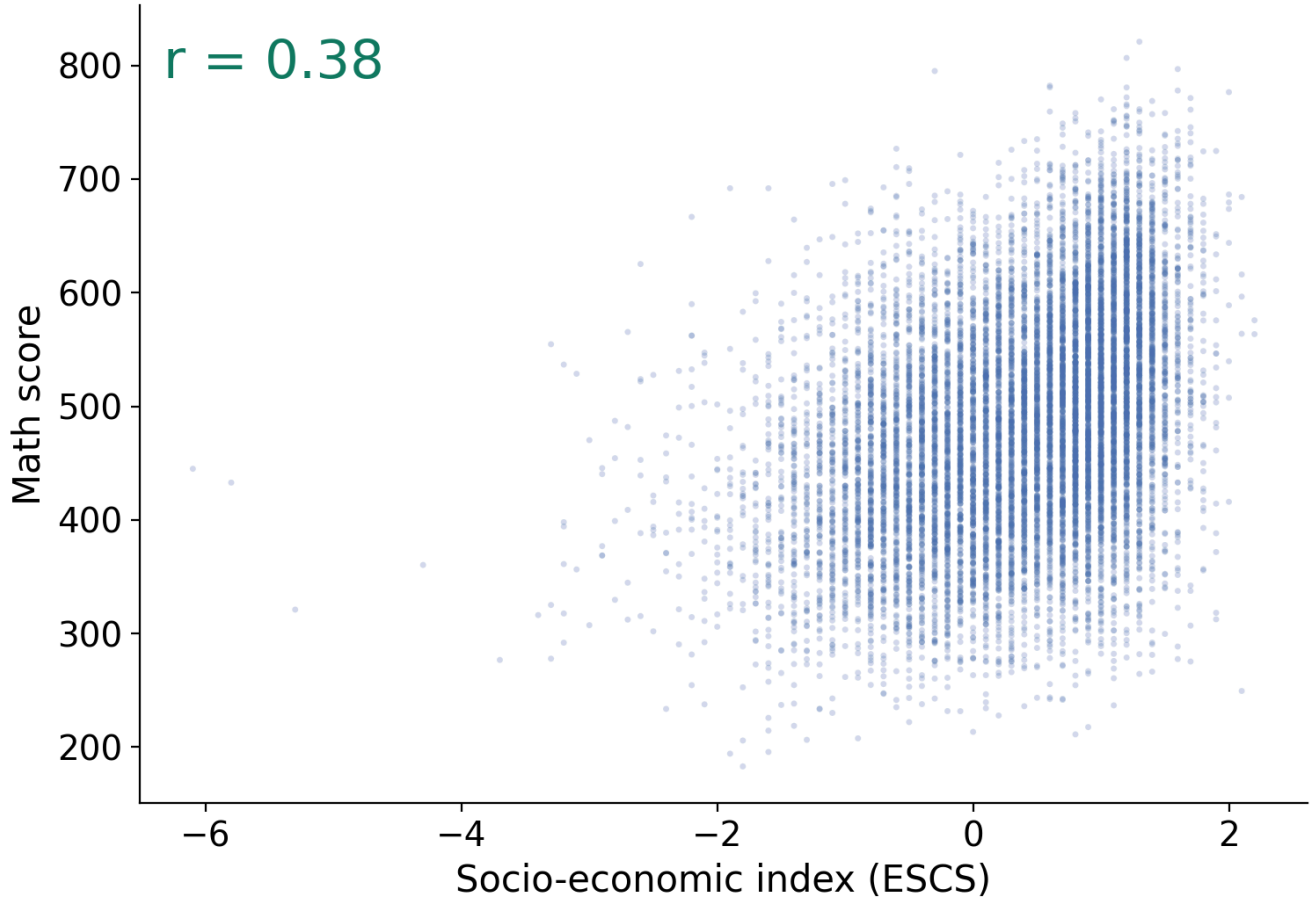
Data: PISA 2022, Australia (OECD)

Gallery 2: math & socio-economic status



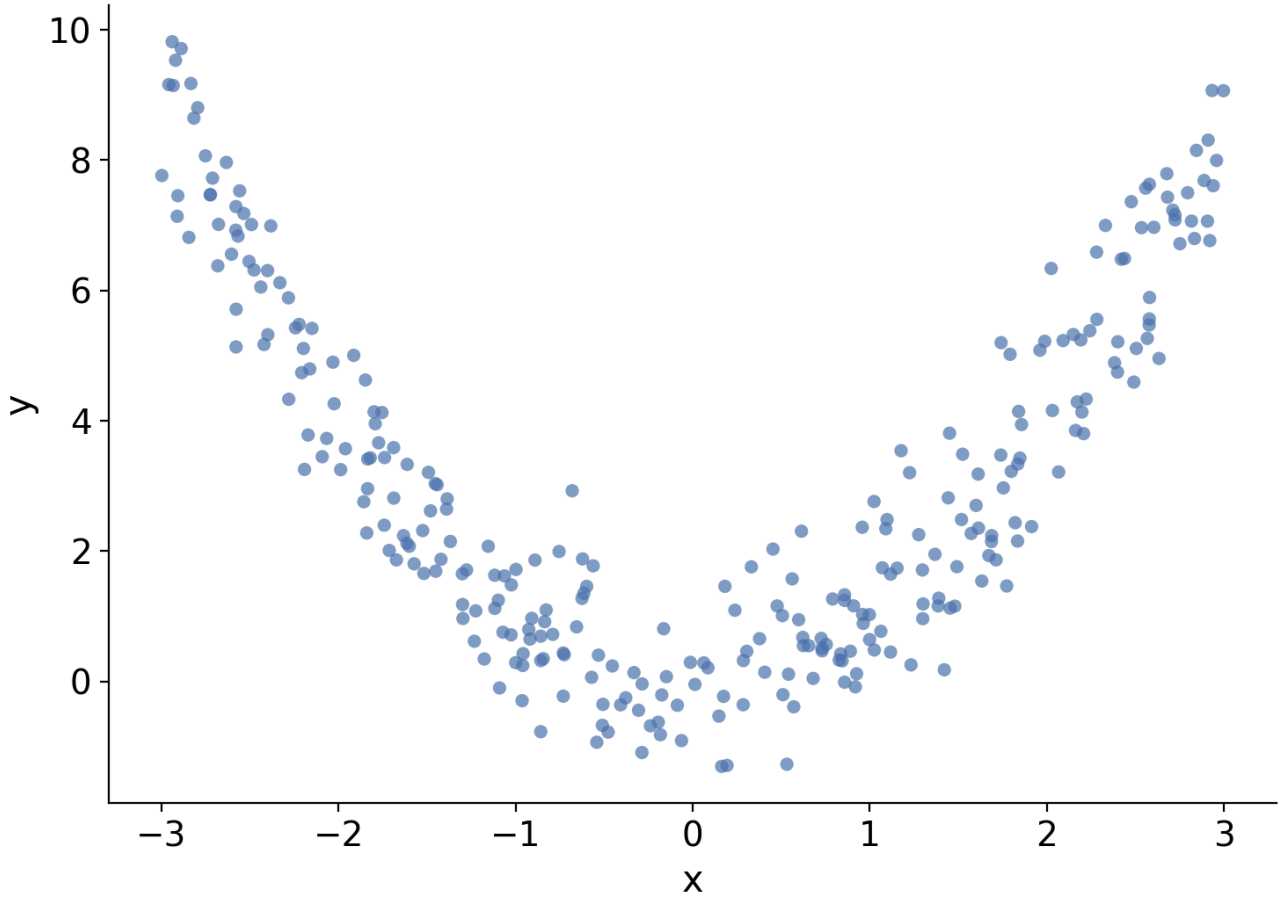
Data: PISA 2022, Australia (OECD)

Gallery 2: math & socio-economic status

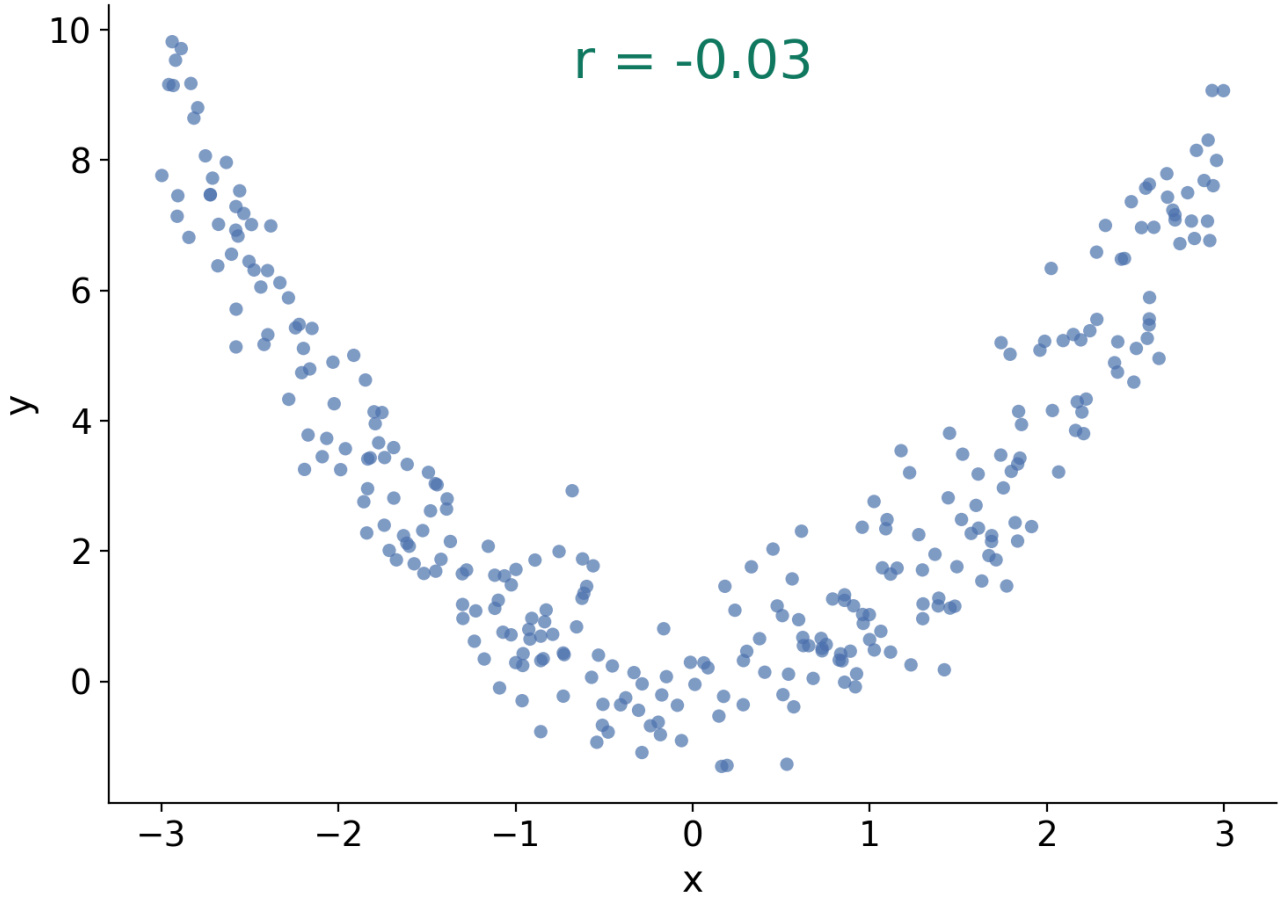


Data: PISA 2022, Australia (OECD)

Gallery 3: simulated data



Gallery 3: simulated data



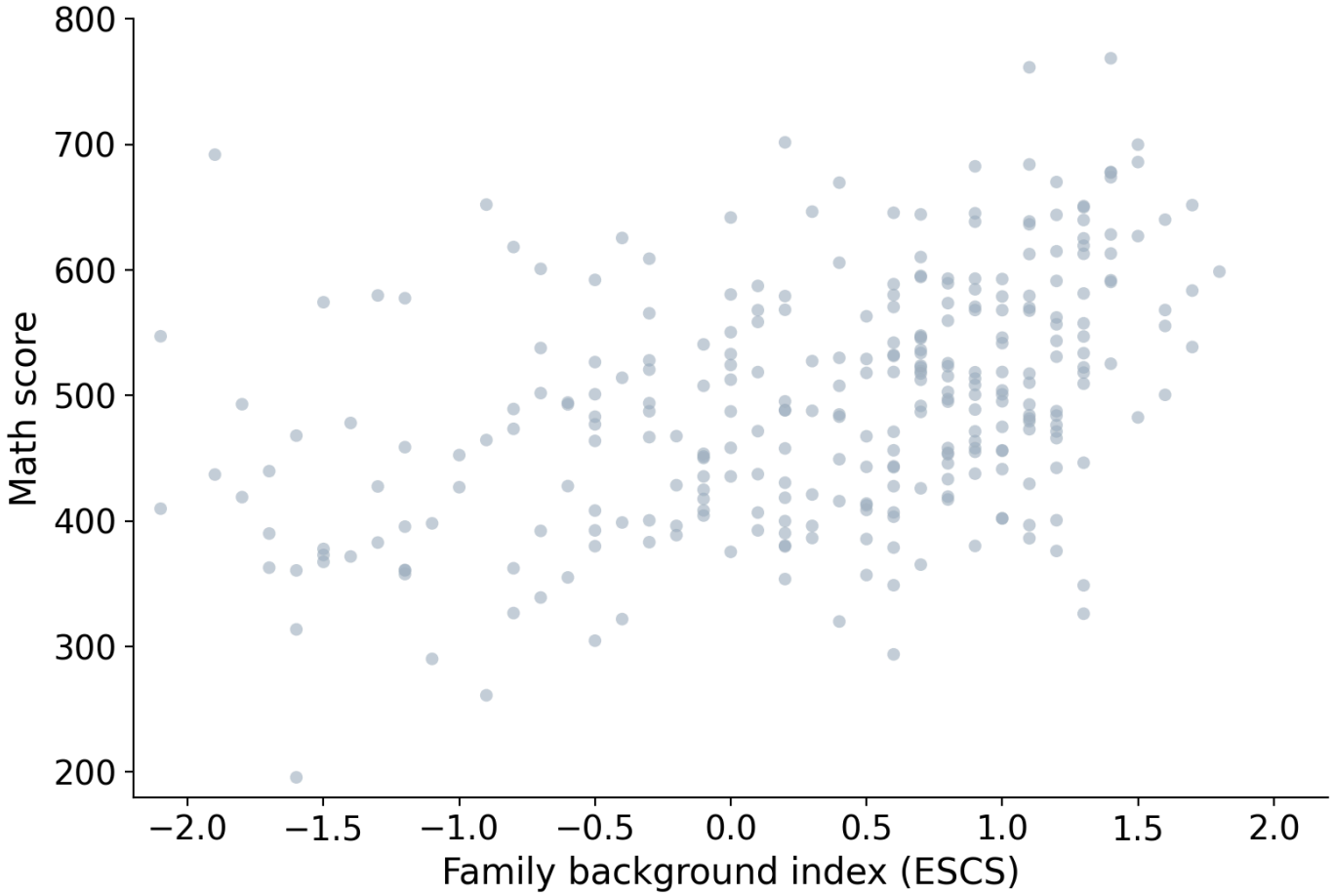
Understanding covariance

Covariance adds products of deviations

$$\operatorname{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

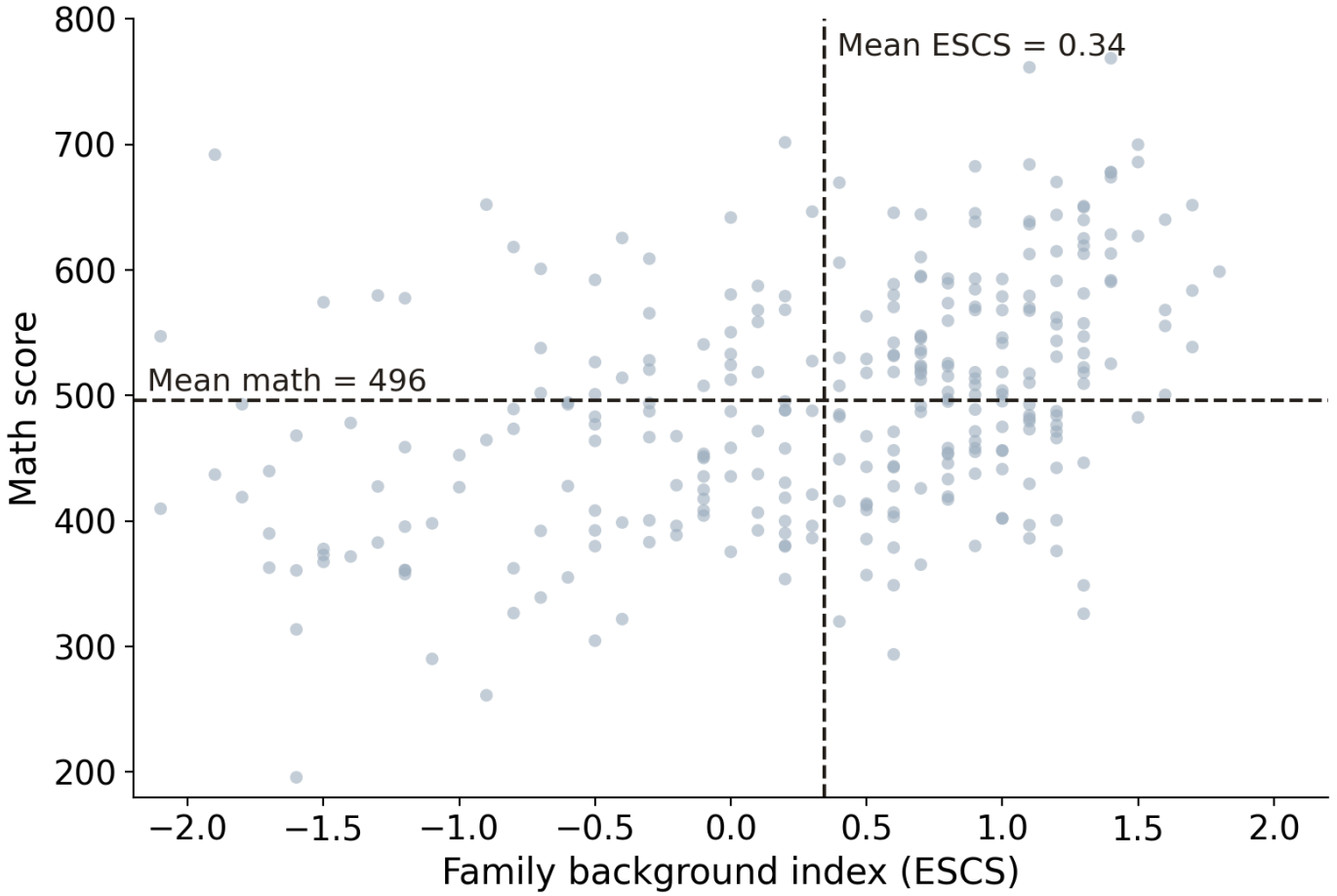
1. Find the two sample means, $\bar{x}$ and $\bar{y}$
2. For every student i, measure their deviation from the mean: $x_i - \bar{x}$ and $y_i - \bar{y}$
3. Multiply, then add the signed products
4. Divide by $n - 1$

Math vs. socio-economic status (N=300)



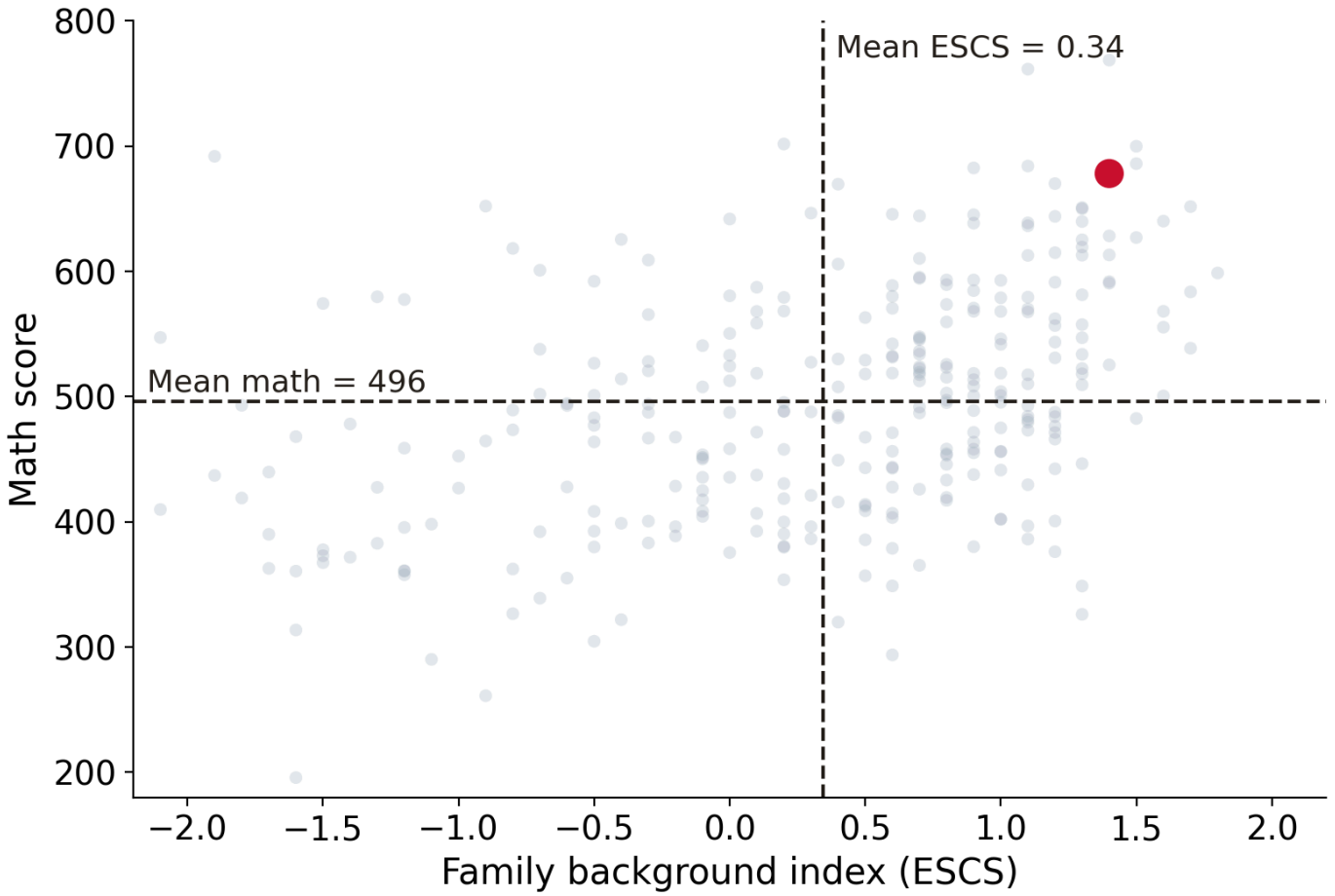
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 1: find the two means



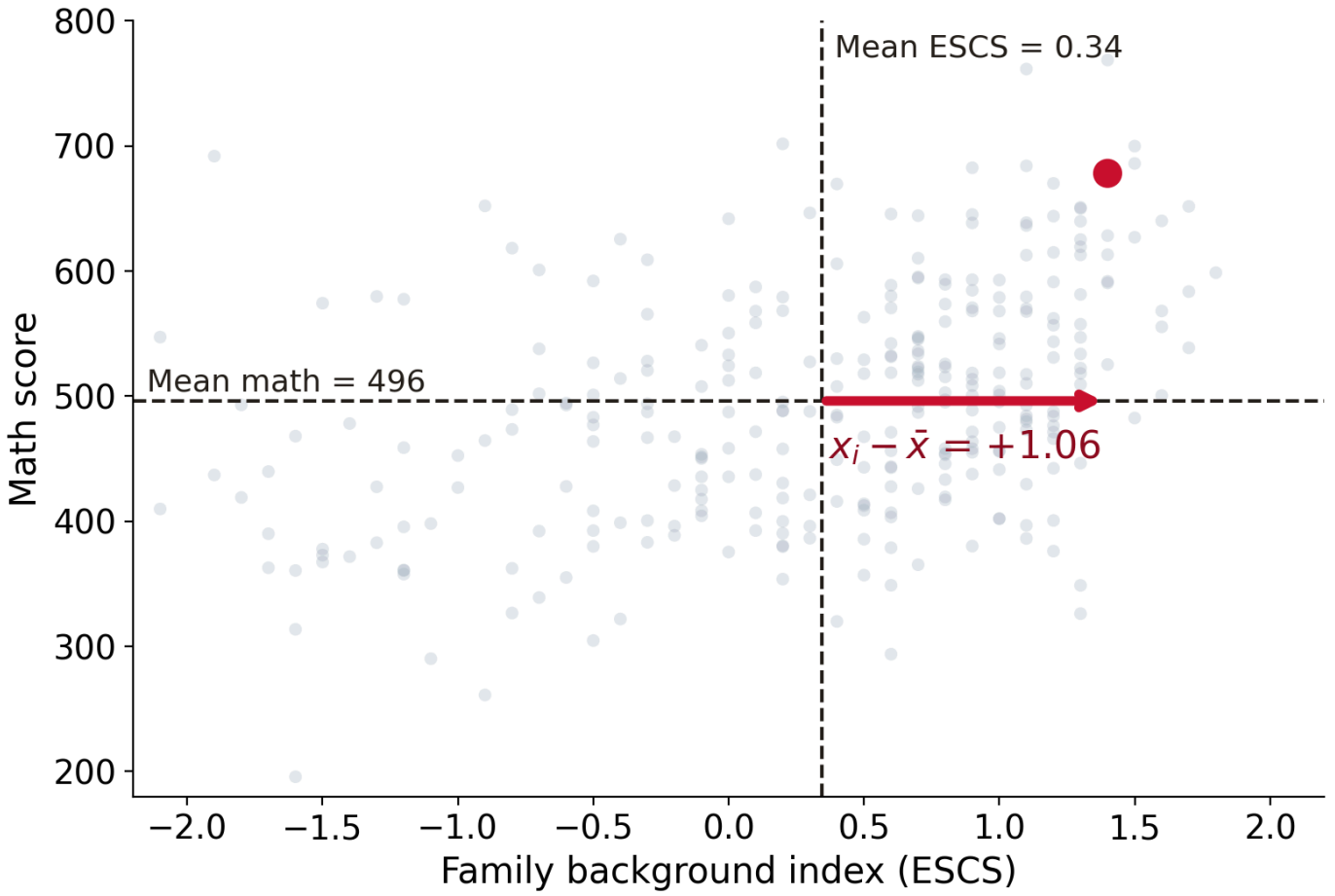
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 2: measure deviations for each student



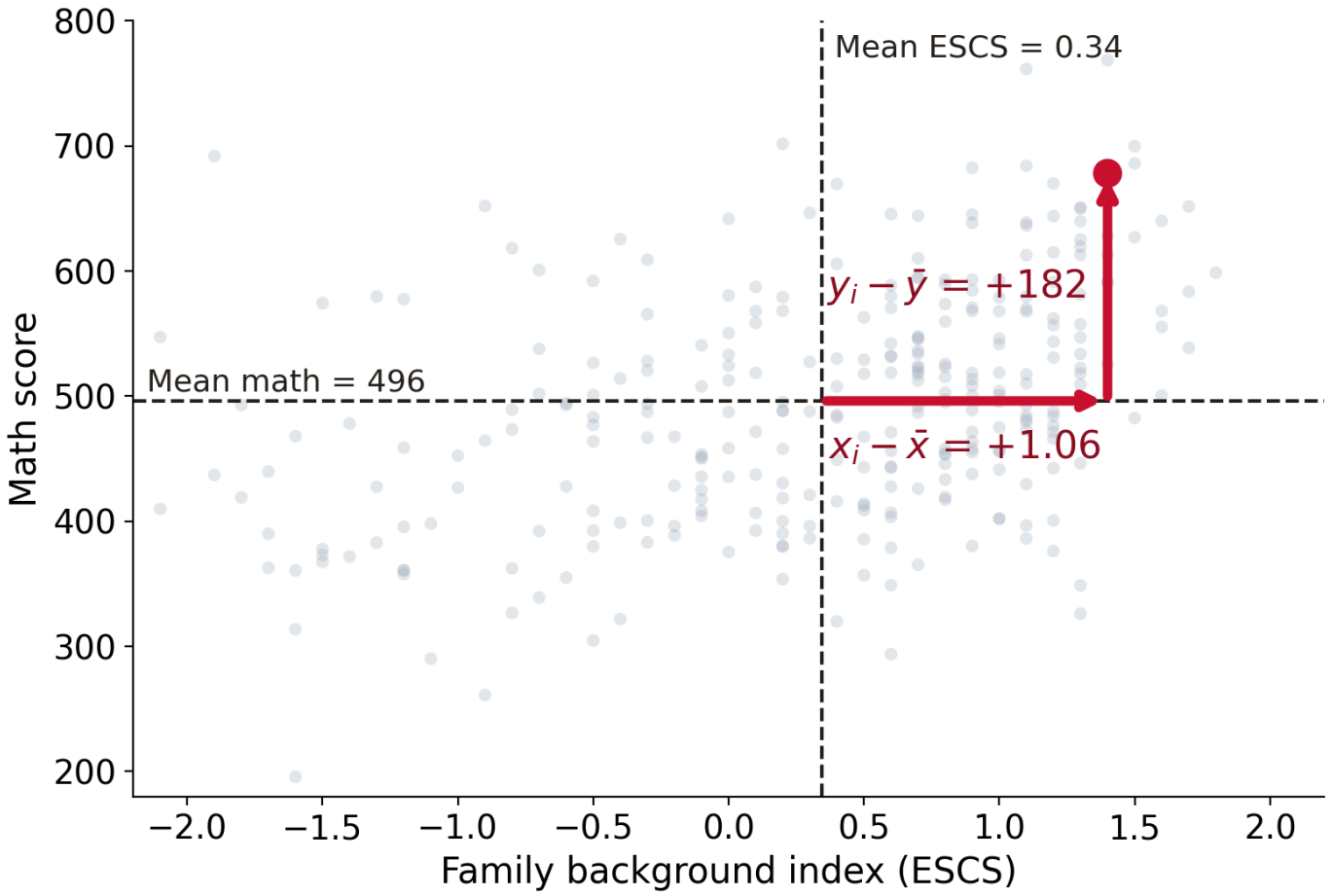
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 2: find the X-deviation



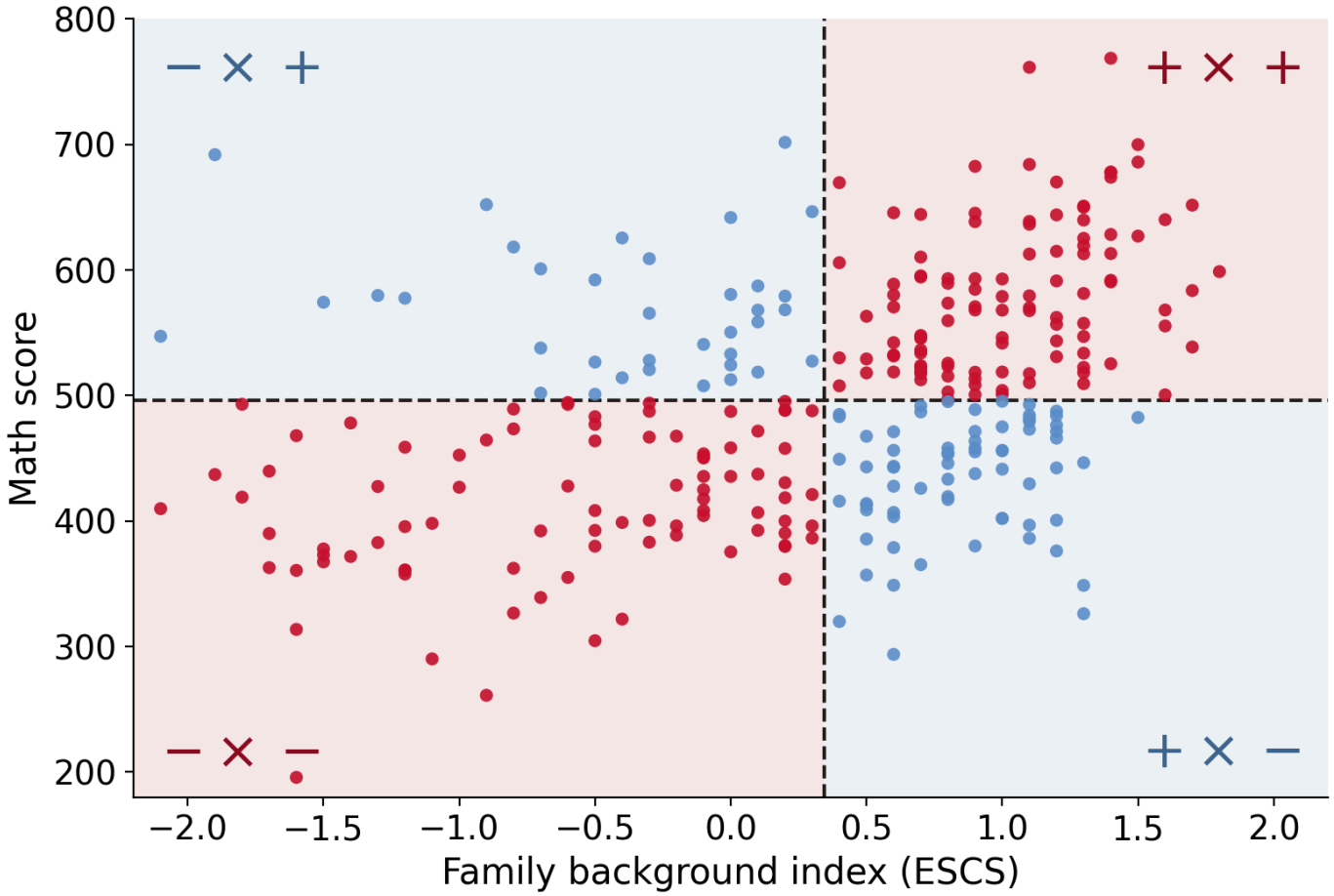
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 2: find the Y-deviation



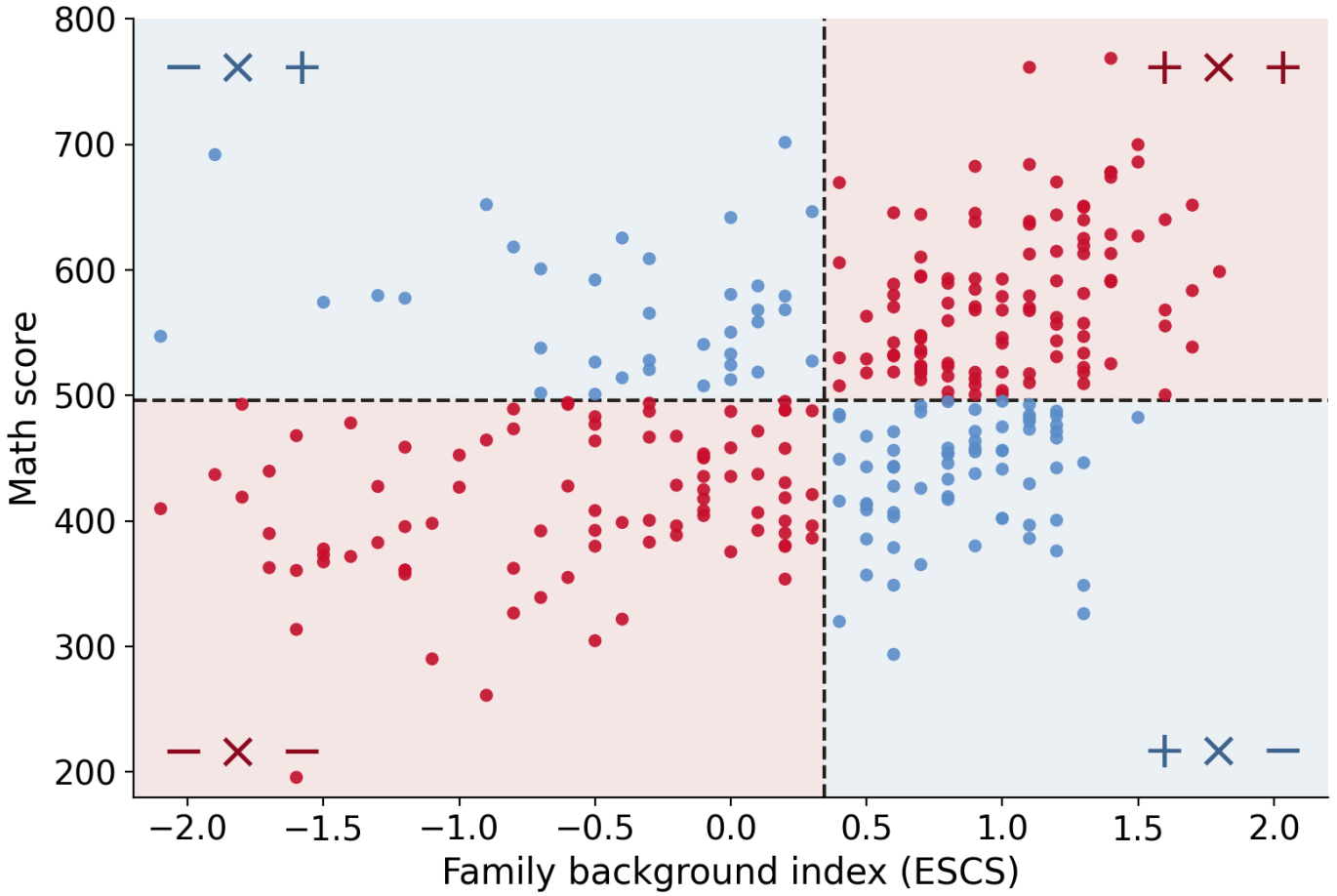
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 3: multiply & sum



Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 3: multiply & sum



+14,520

positive products

-4,046

negative products

Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 4: divide by (N−1)

$$\operatorname{cov}(X, Y) = \frac{14{,}520 - 4{,}046}{300 - 1} = 35 \text{ index points} \times \text{score points}$$

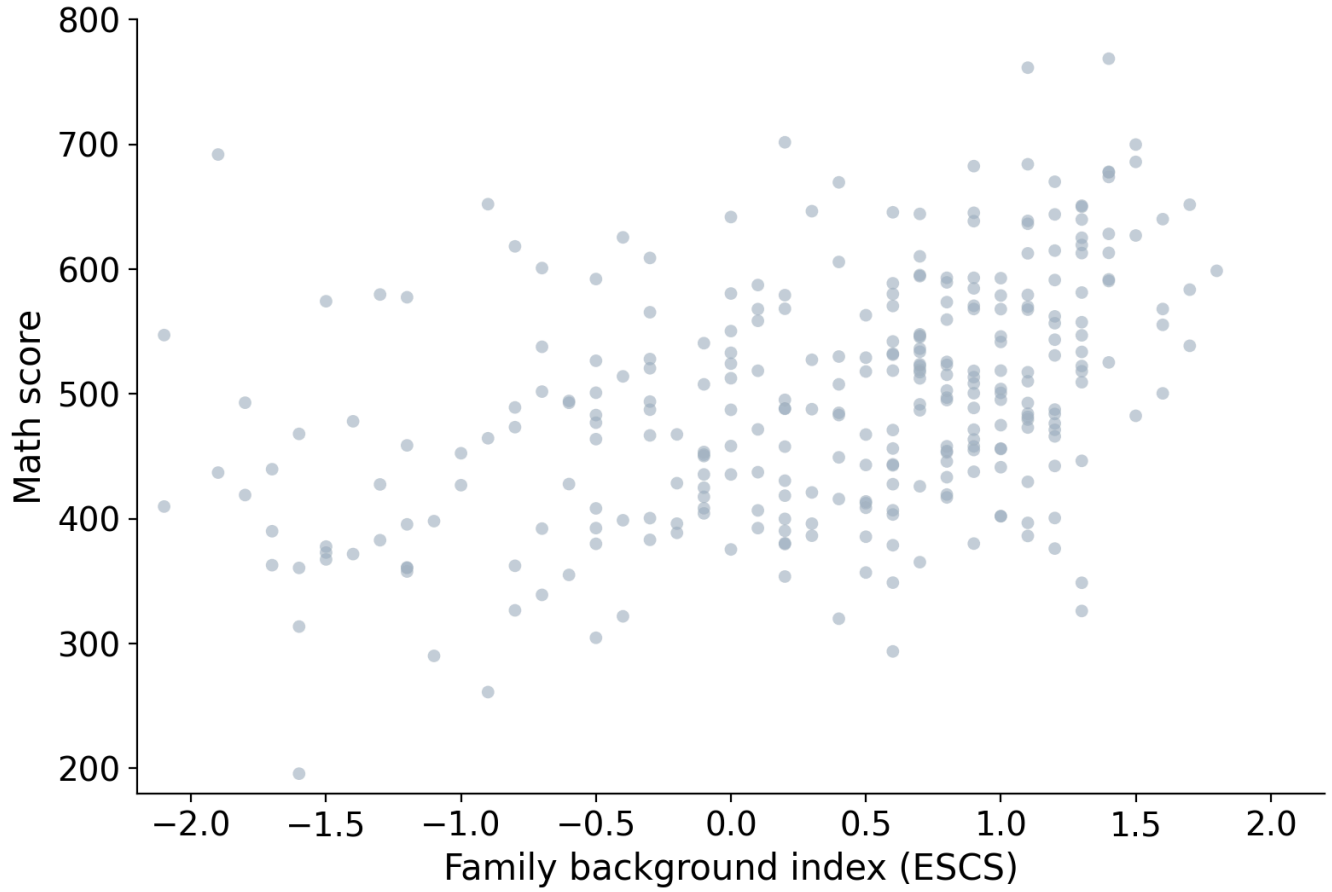
Fixing the nonsense units

$$\operatorname{cov}(X, Y) = \frac{14{,}520 - 4{,}046}{300 - 1} = 35 \text{ index points} \times \text{score points}$$

$$r = \frac{\operatorname{cov}(X, Y)}{s_X \, s_Y} = \frac{35}{0.88 \times 95.3} = 0.42$$

You draw the line

Fit the line



https://emiliatjernstrom.com/econ2041/portal.html

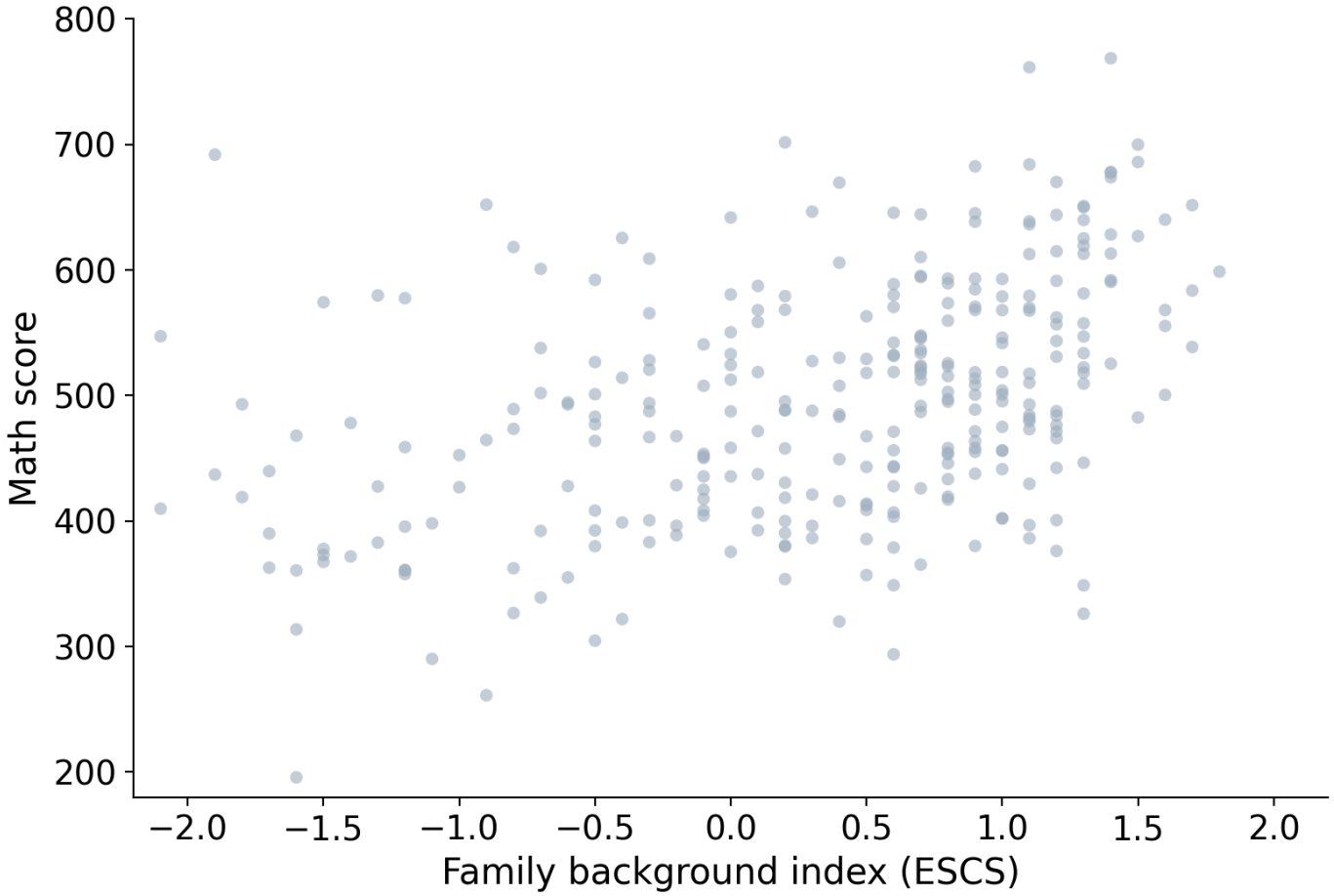
Two questions turn a line into regression

You drew a line that "looks right"

1. Why is a line a sensible summary of the relationship?
2. When many lines look reasonable, which one should we choose?

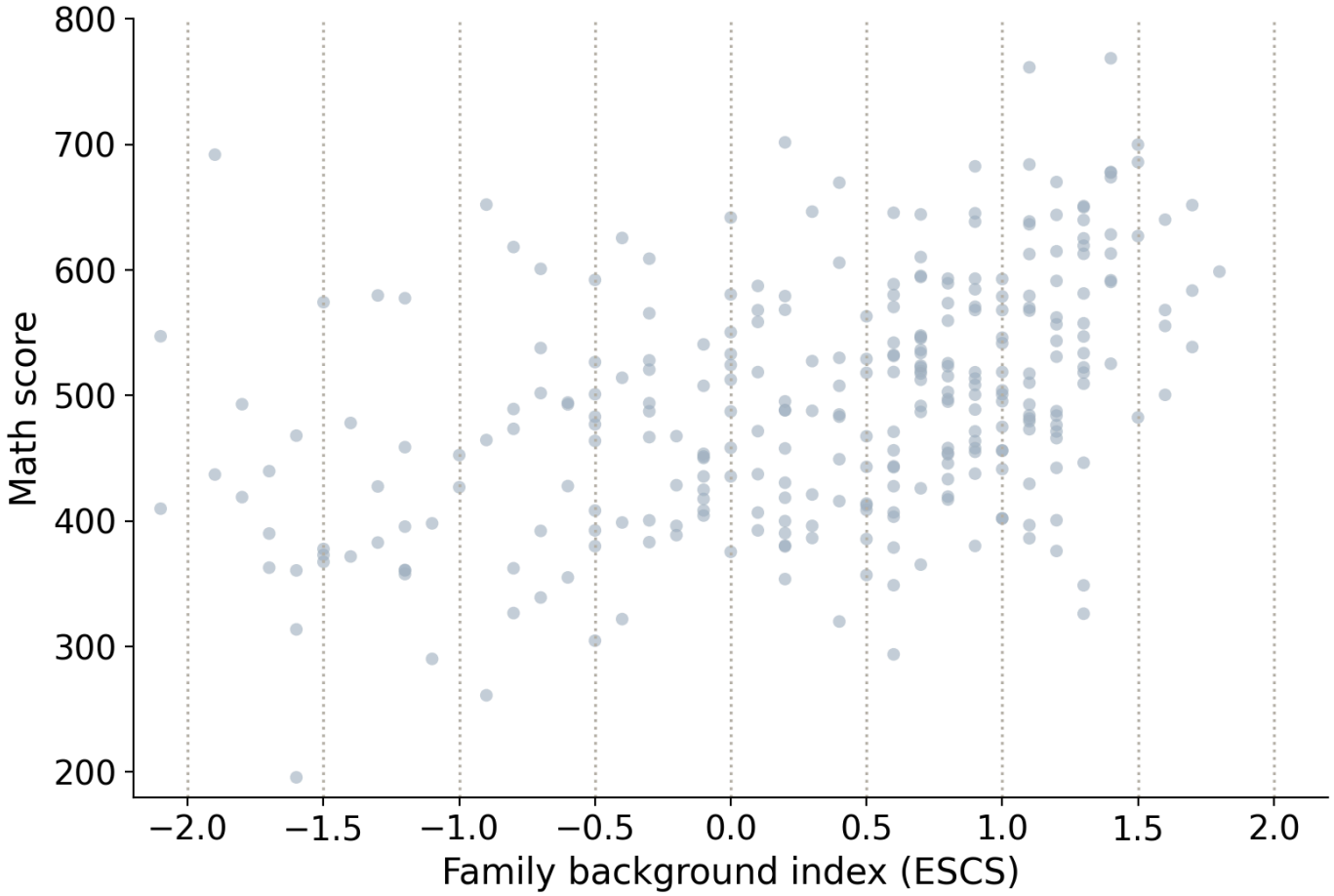
From averages to a relationship

Average math score for each level of ESCS



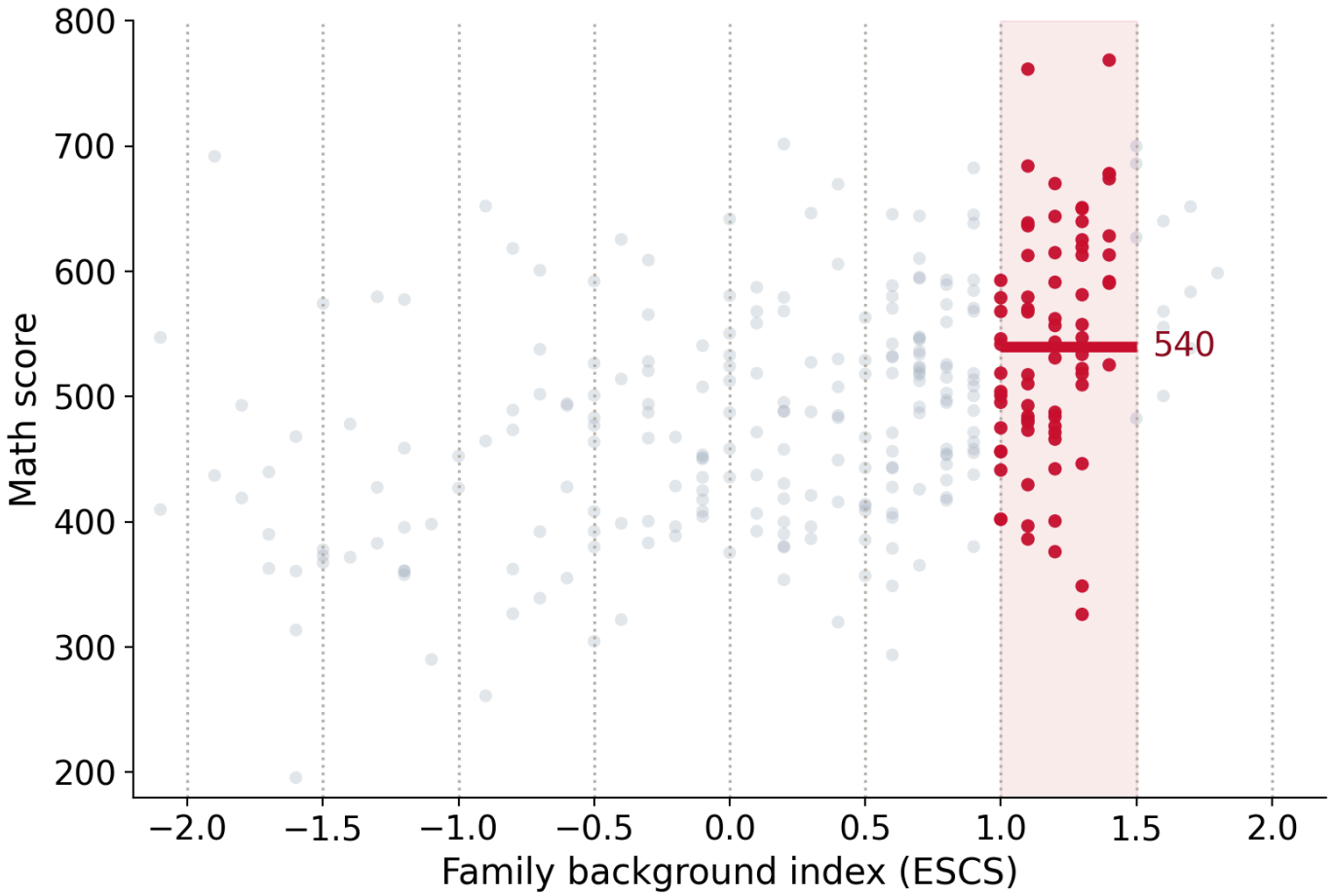
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Average math score for each level of ESCS



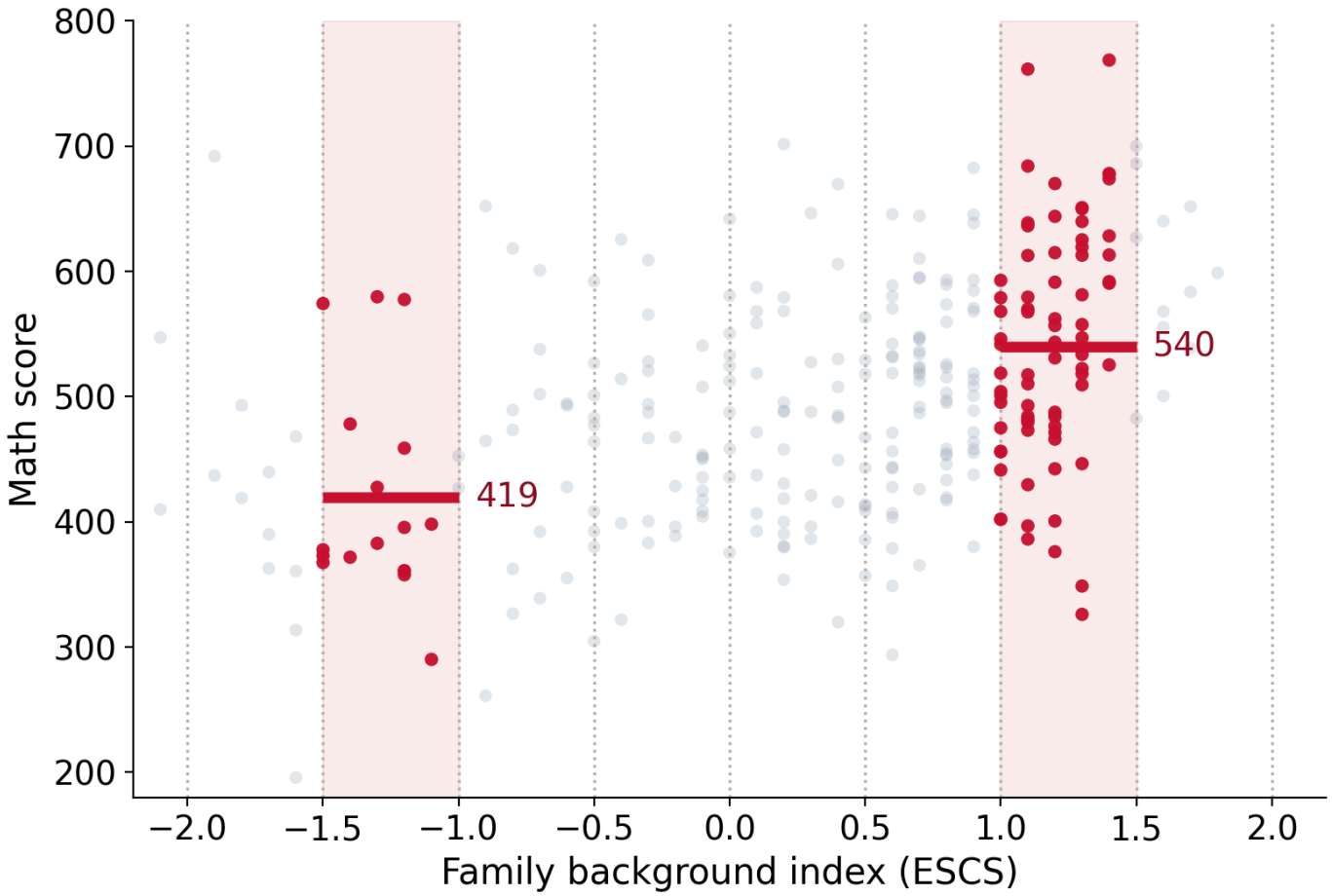
Data: PISA 2022, Australia (OECD), 300 students drawn at random

One slice, one estimated conditional mean



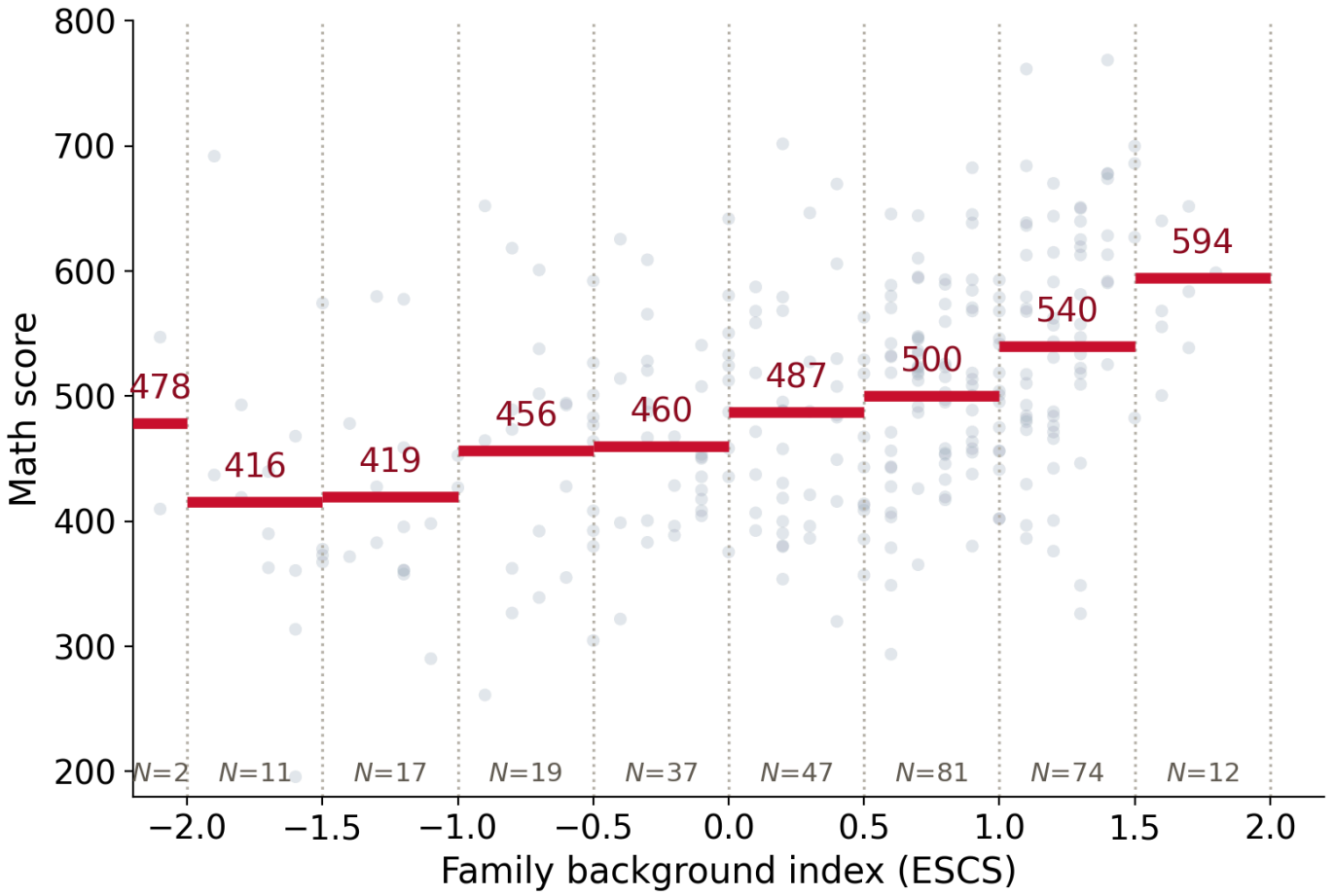
Data: PISA 2022, Australia (OECD), 300 students drawn at random

One slice, one estimated conditional mean



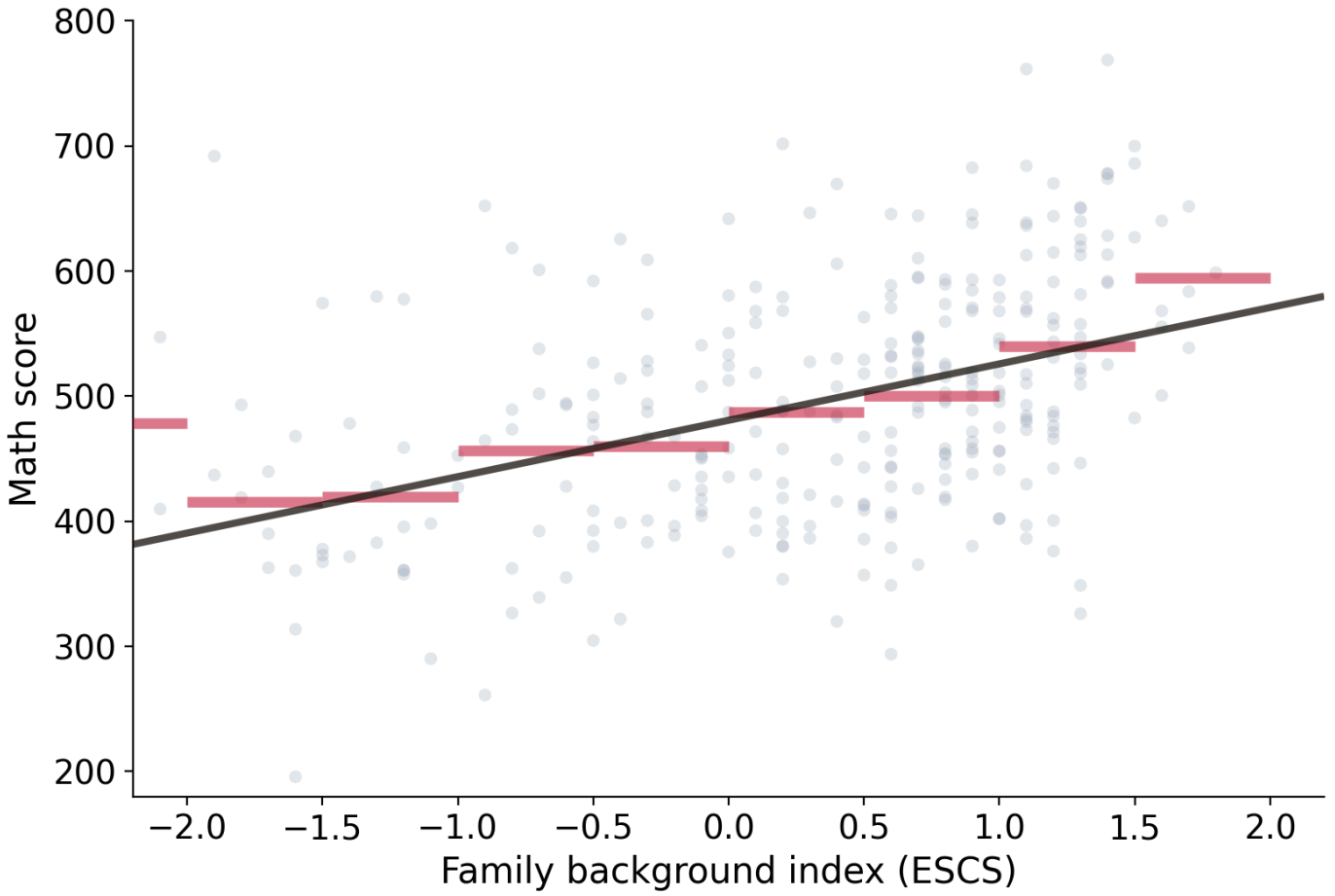
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Slice means estimate the conditional mean



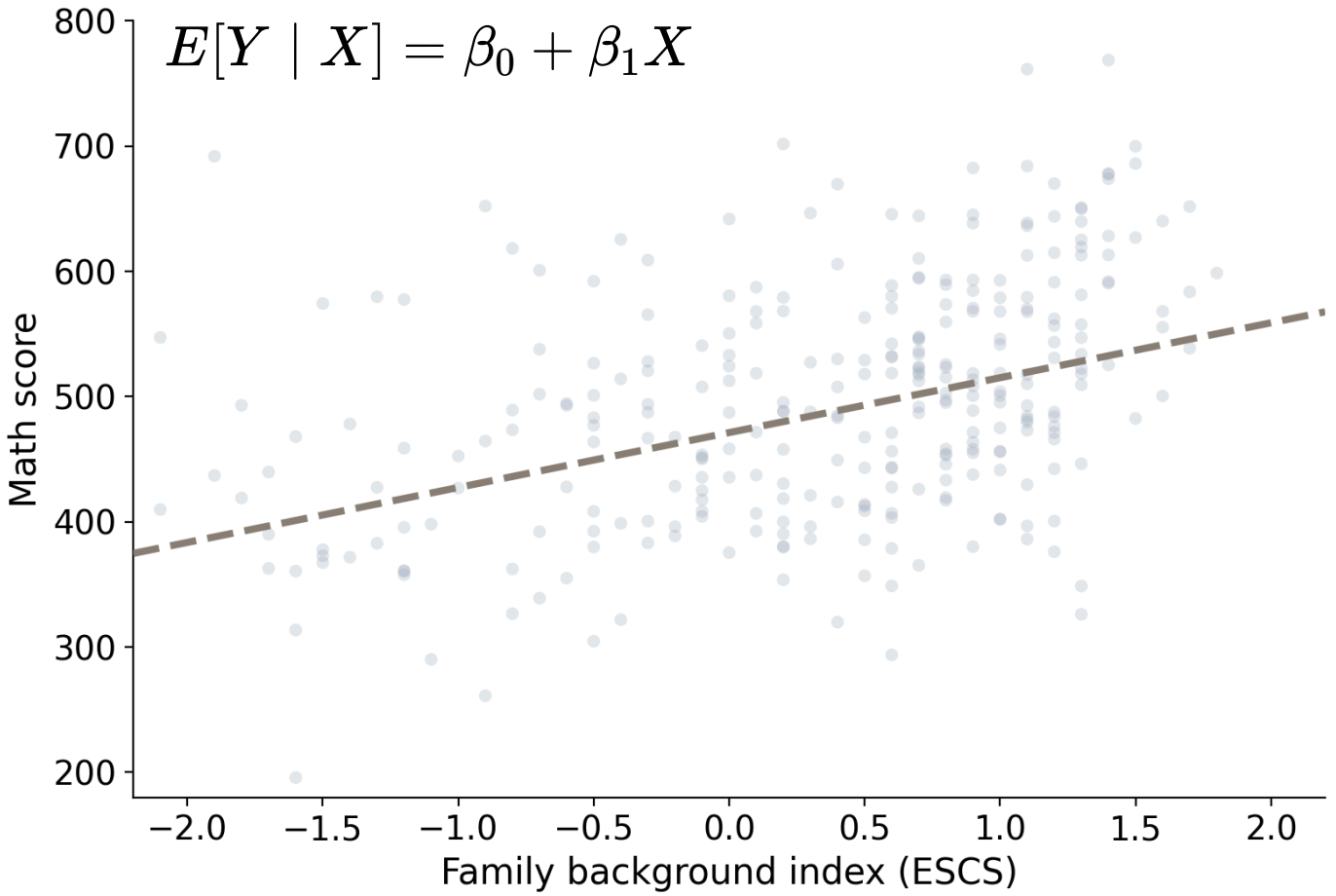
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Summarize conditional means with a line



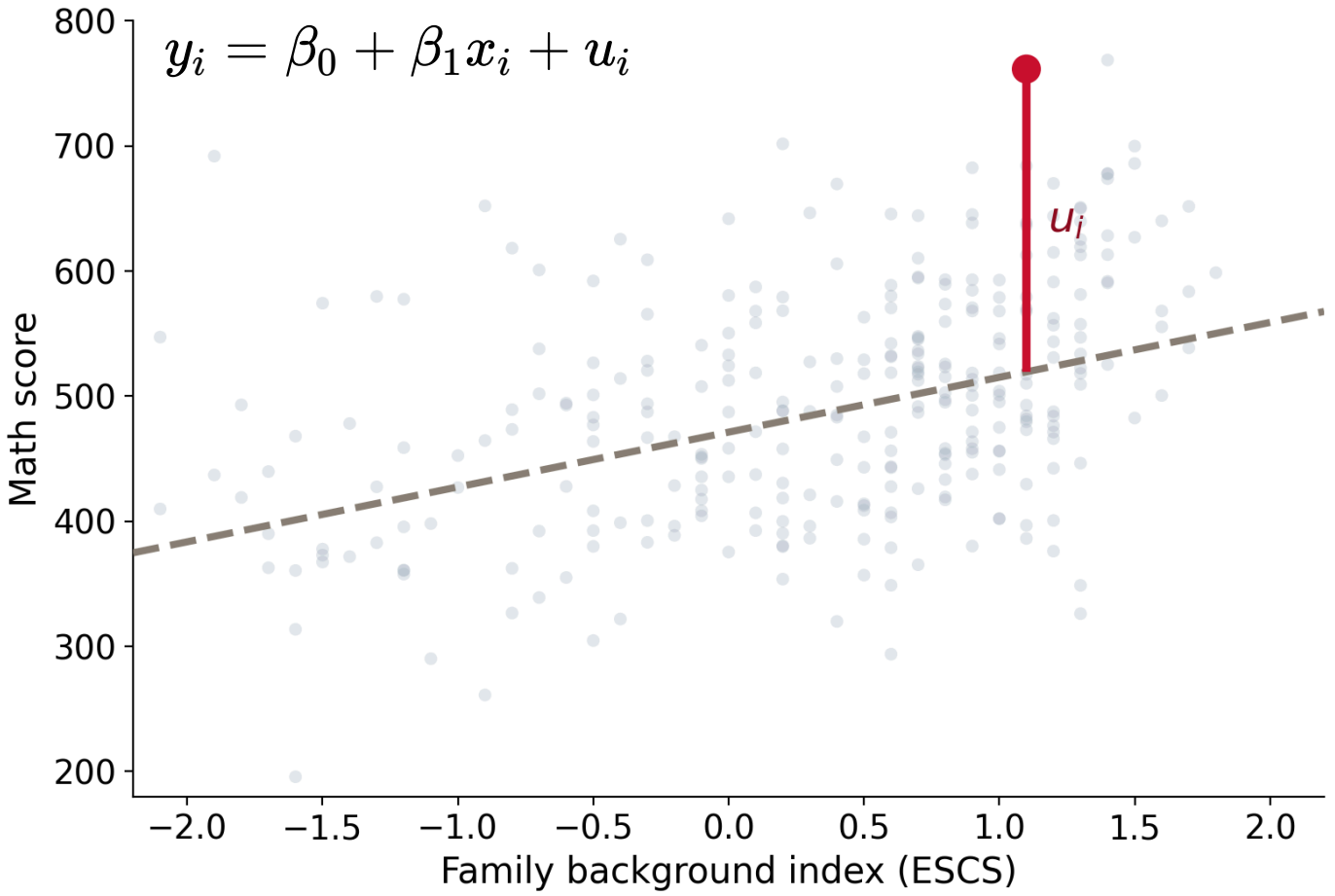
Data: PISA 2022, Australia (OECD), 300 students drawn at random

The population line is the conditional mean



Data: PISA 2022, Australia (OECD); dashed line fitted on all 12,136 students

An error is the gap from the population line



Data: PISA 2022, Australia (OECD); dashed line fitted on all 12,136 students

The sample line estimates the population line

Population target

$$E[Y \mid X] = \beta_0 + \beta_1 X$$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- β_0, β_1: fixed & unknown
- u_i: unobserved population error

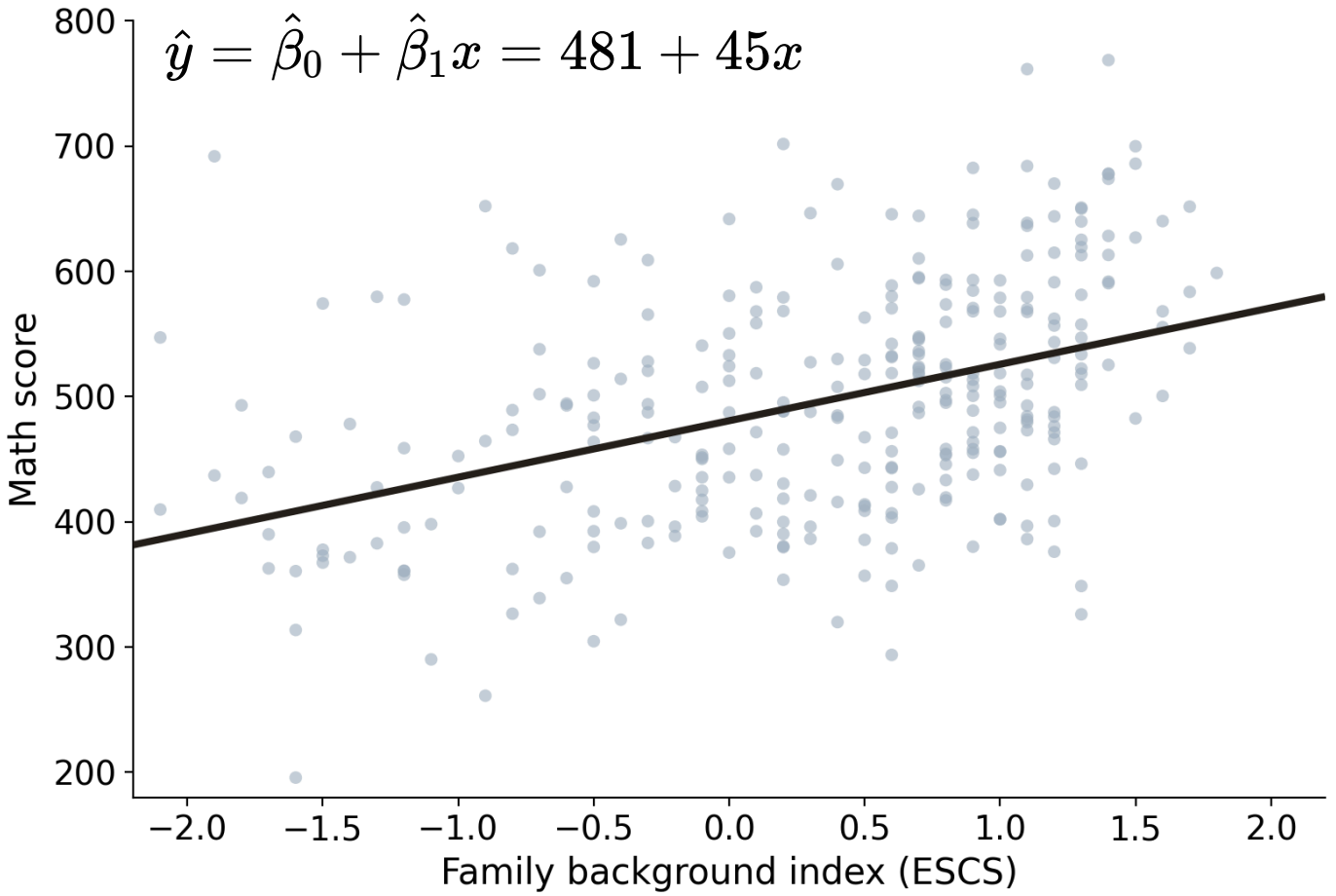
Sample estimate

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\beta}_0, \hat{\beta}_1$: computed from the sample
- $\hat{y}_i$: predicted score from the fitted line

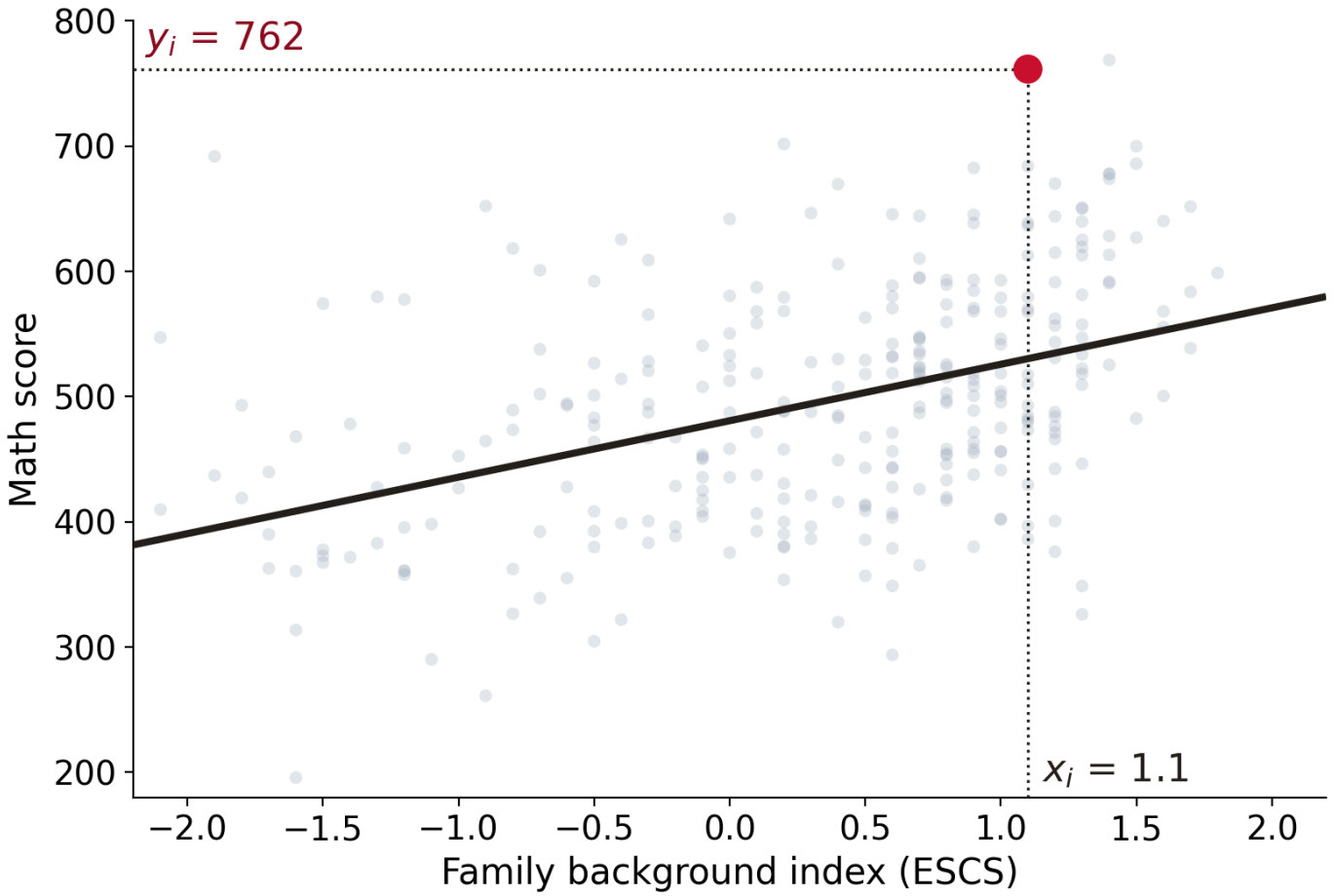
Estimate the sample line

The sample gives us a fitted line



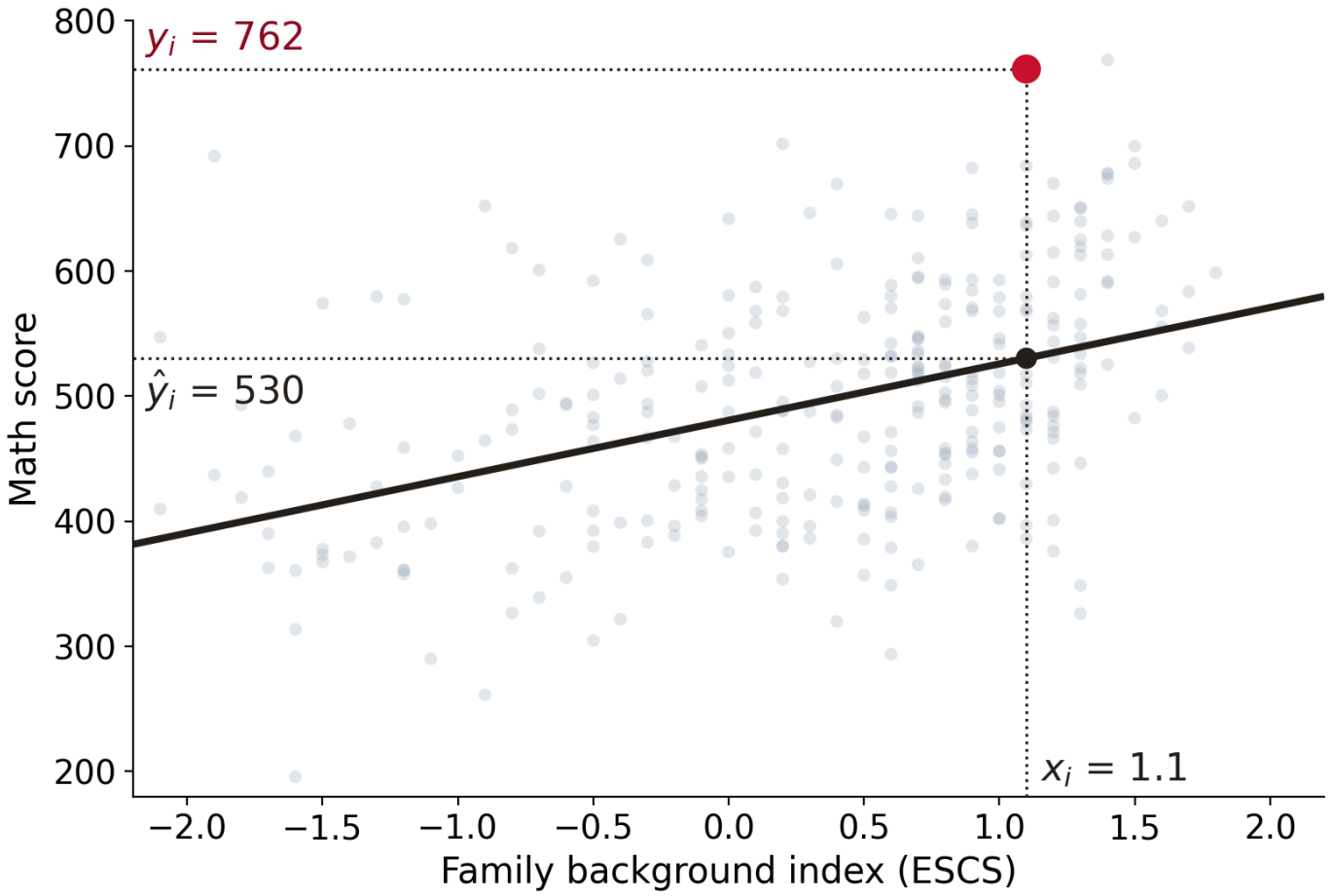
Data: PISA 2022, Australia (OECD), 300 students drawn at random

The sample gives us a fitted line



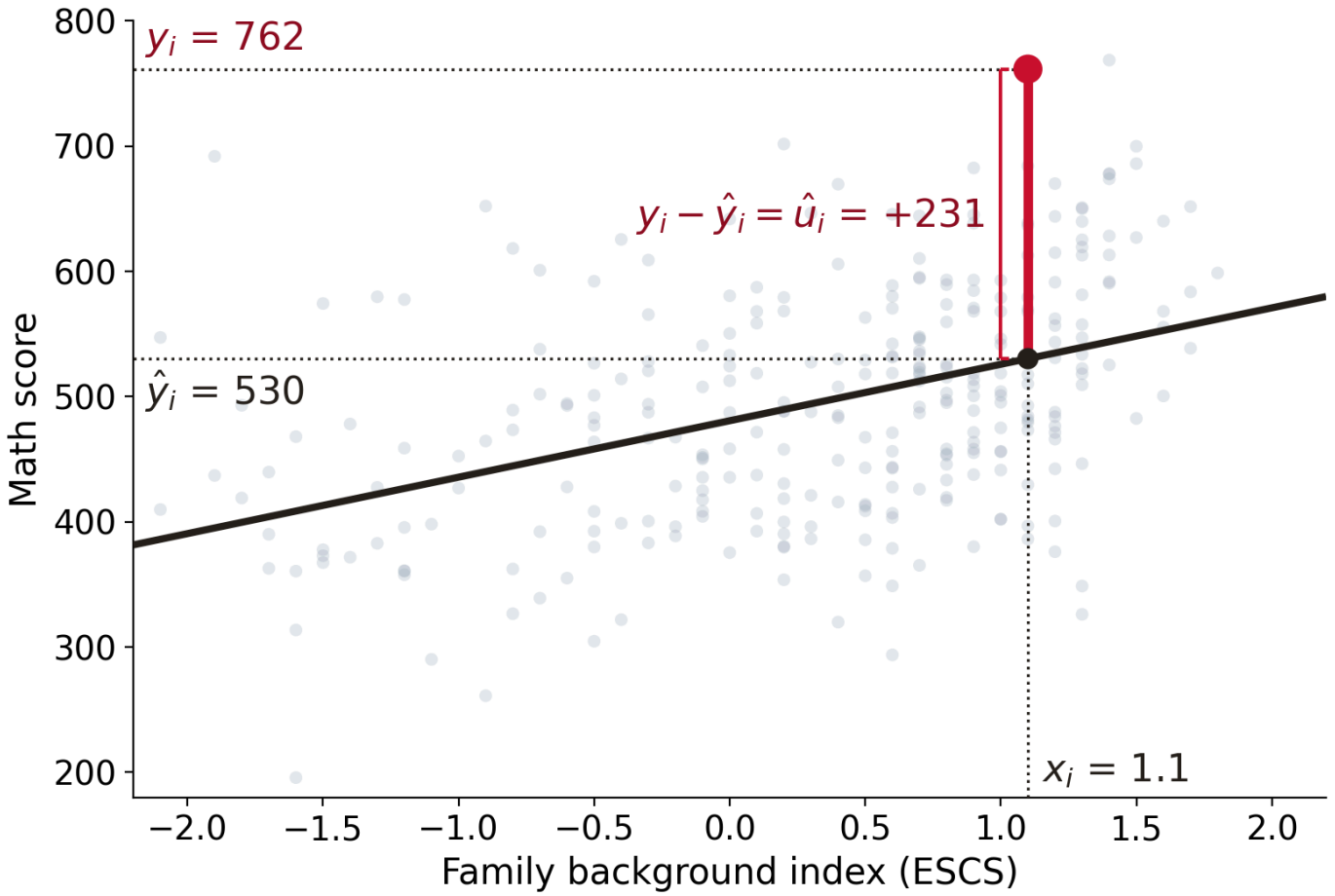
Data: PISA 2022, Australia (OECD), 300 students drawn at random

The line's prediction for student i



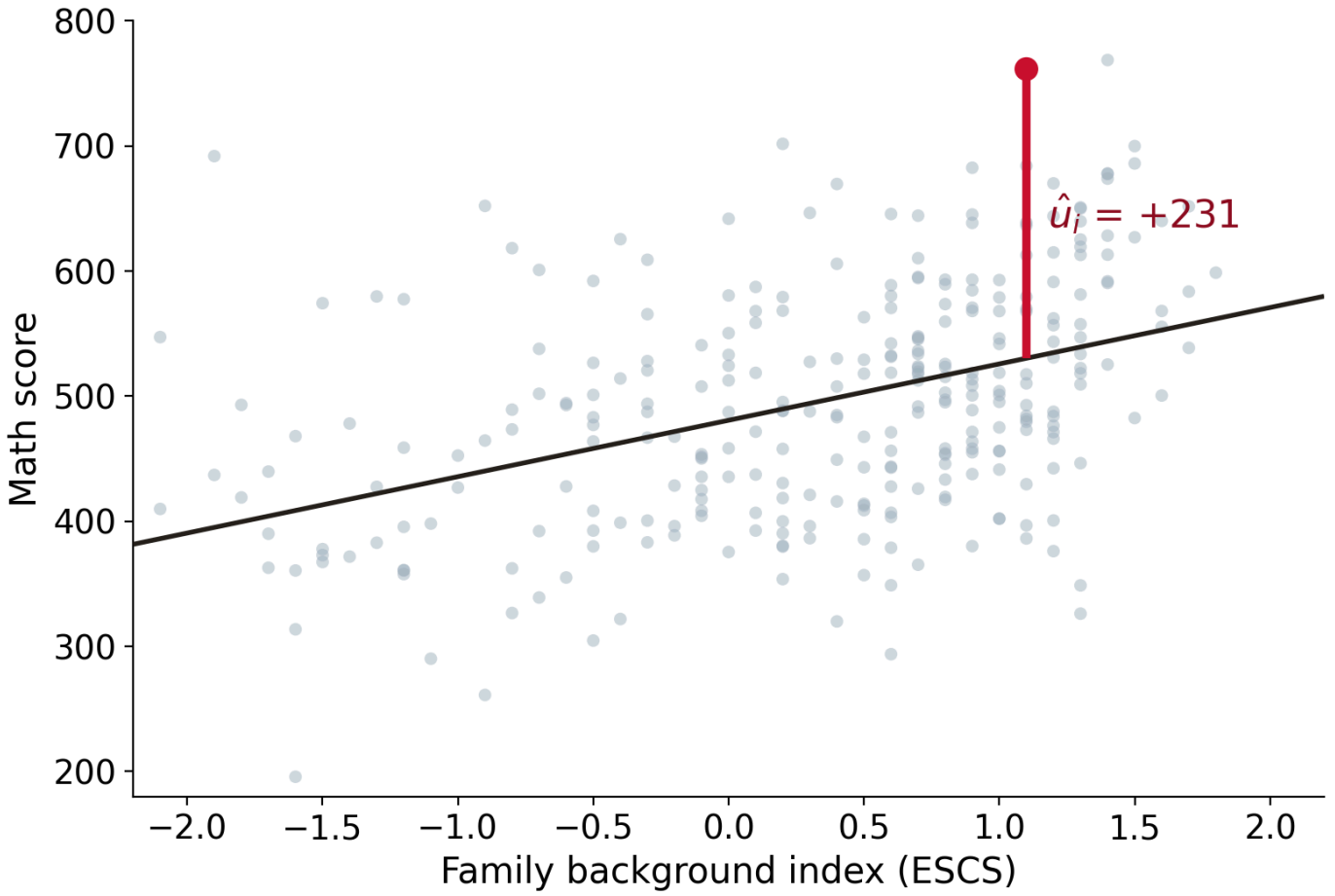
Data: PISA 2022, Australia (OECD), 300 students drawn at random

This gap is called the residual



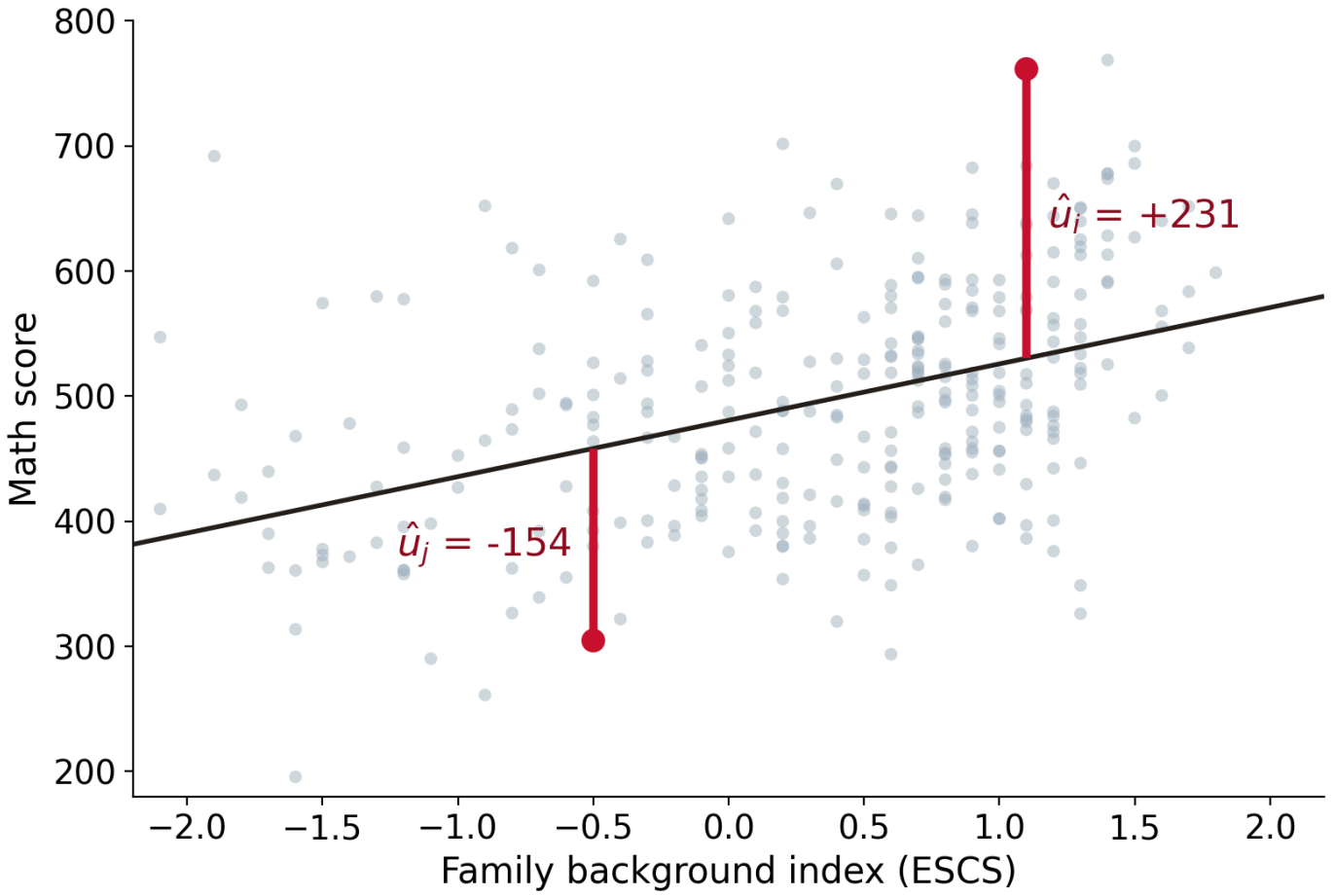
Data: PISA 2022, Australia (OECD), 300 students drawn at random

This gap is called the residual



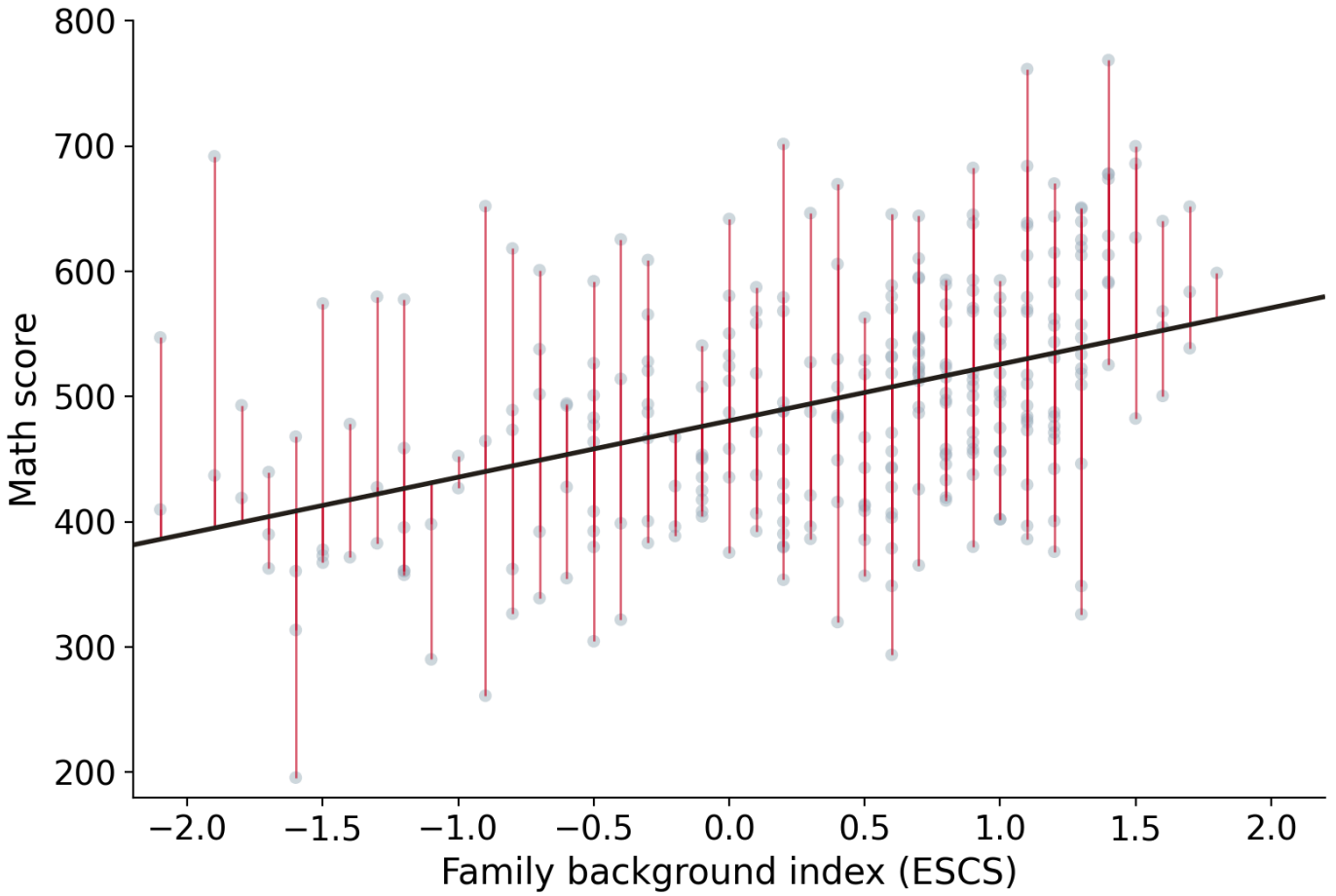
Data: PISA 2022, Australia (OECD), 300 students drawn at random

This gap is called the residual



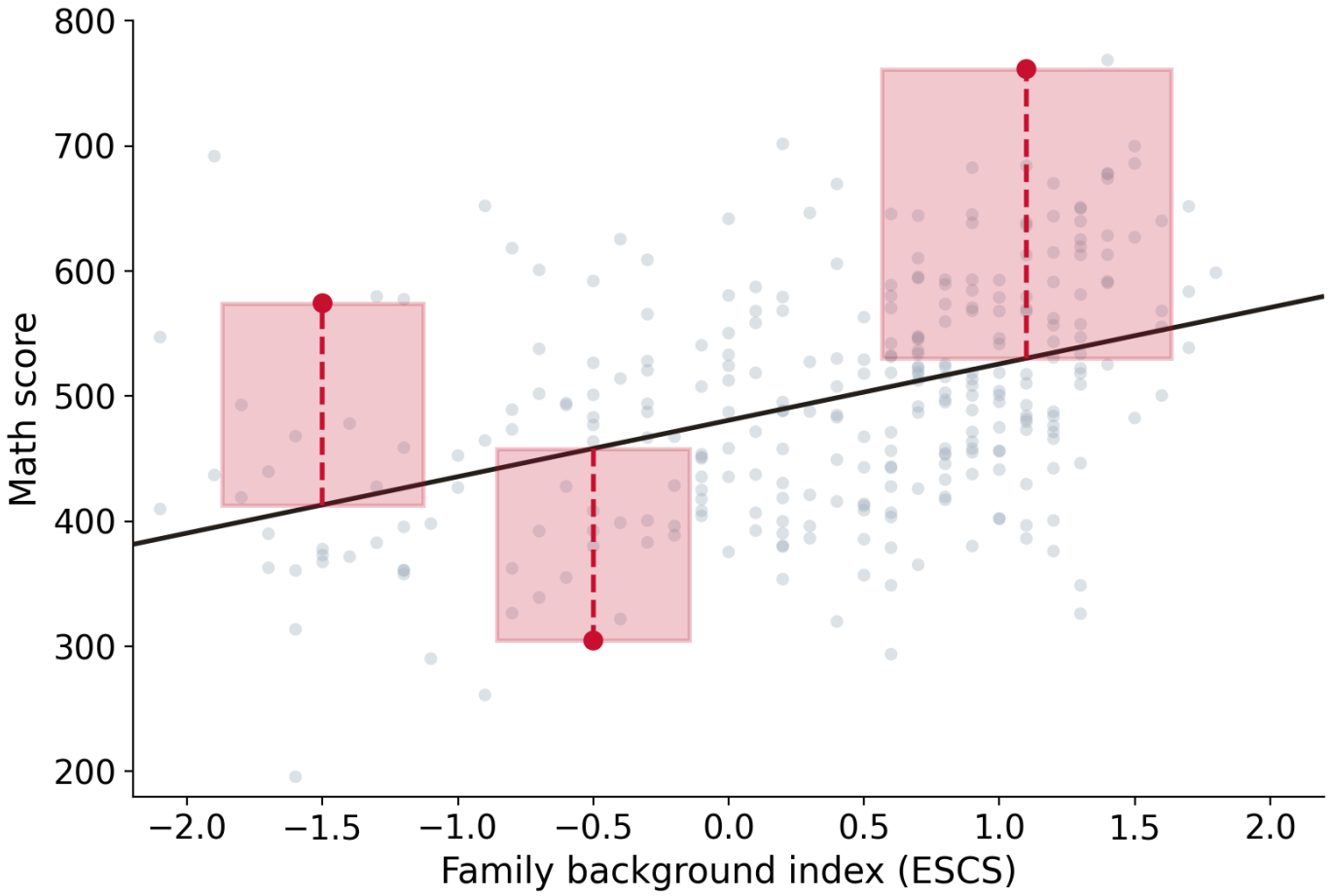
Data: PISA 2022, Australia (OECD), 300 students drawn at random

The residuals: what our summary leaves out



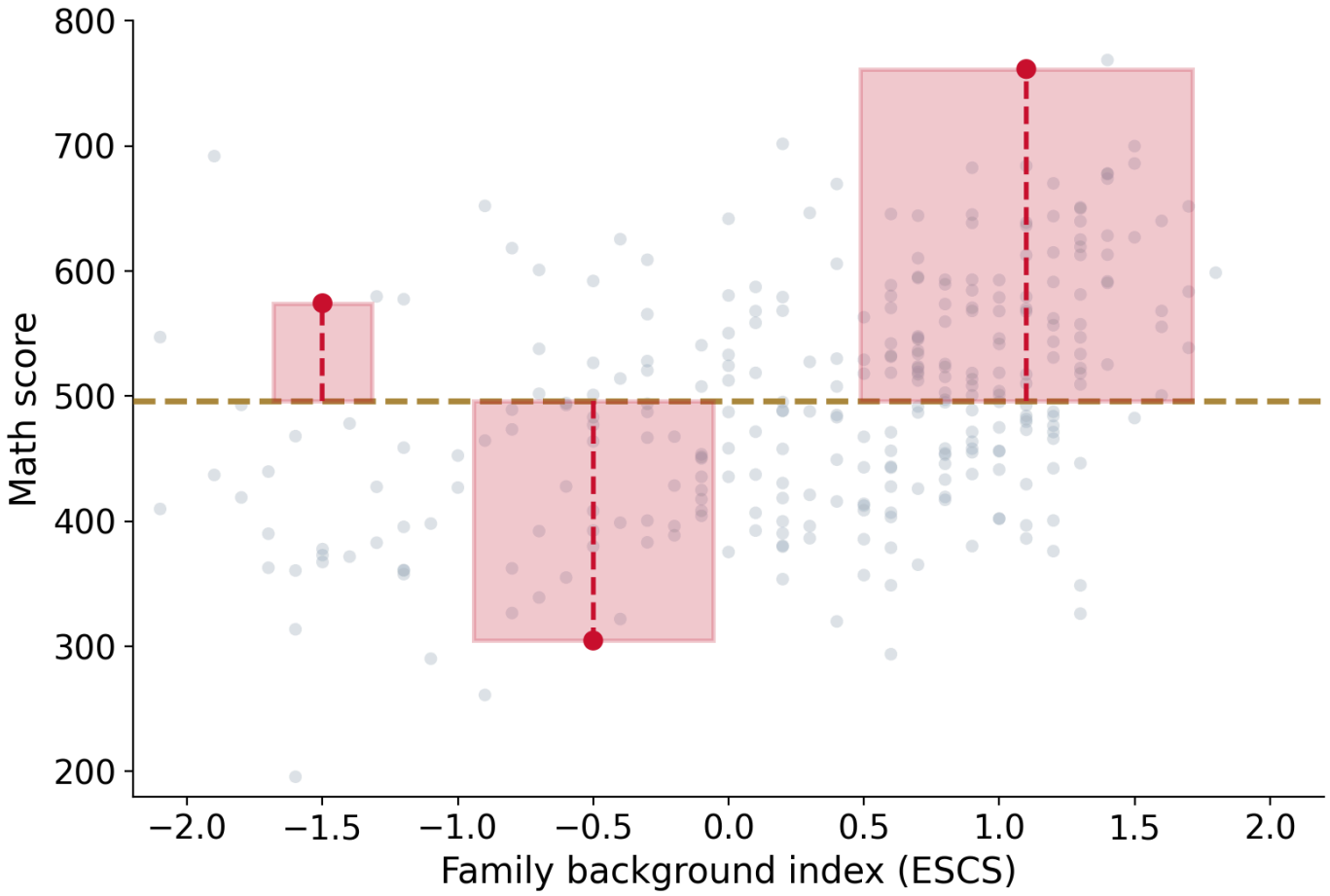
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Minimize the sum of squared residuals



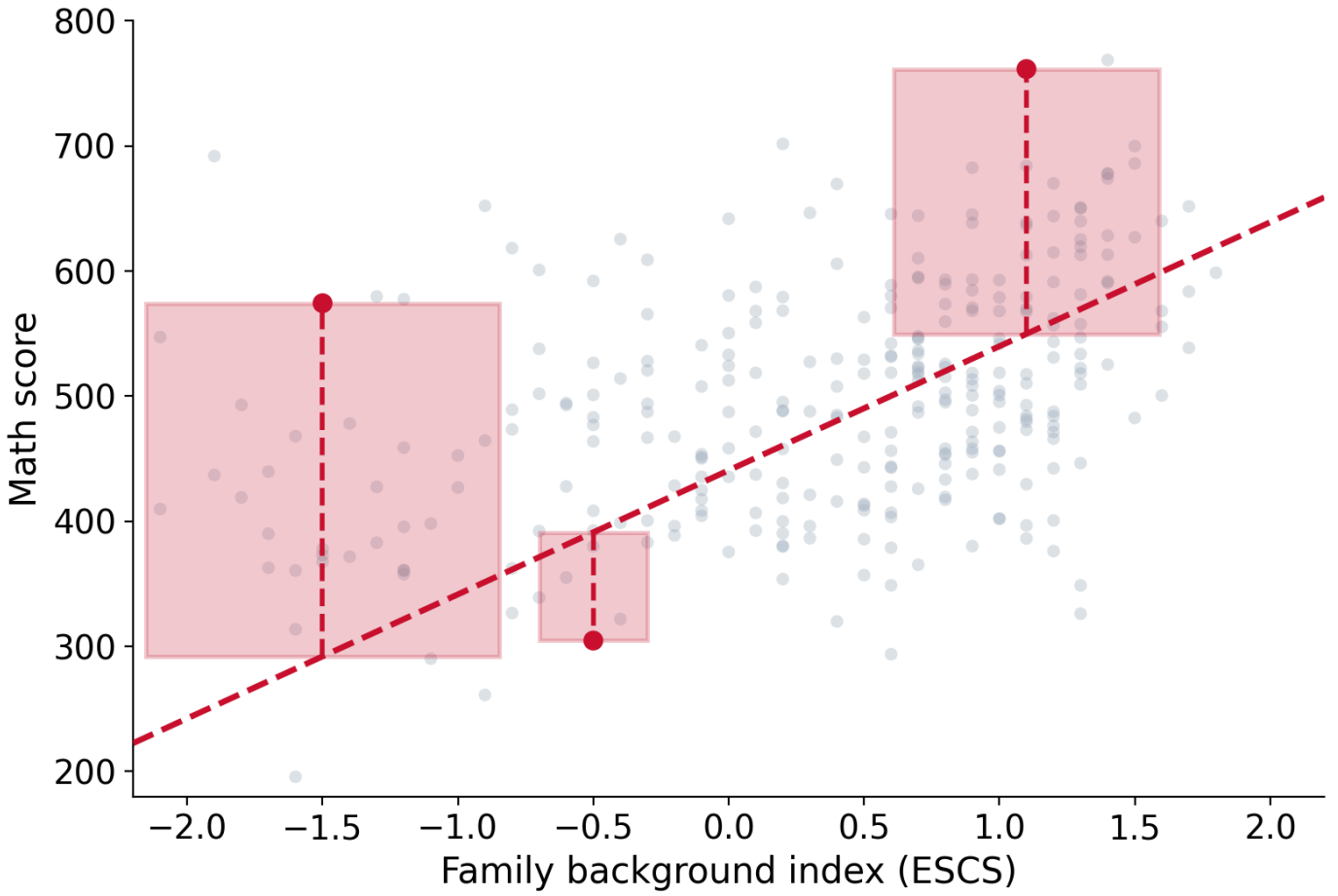
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Squared residuals, flat line



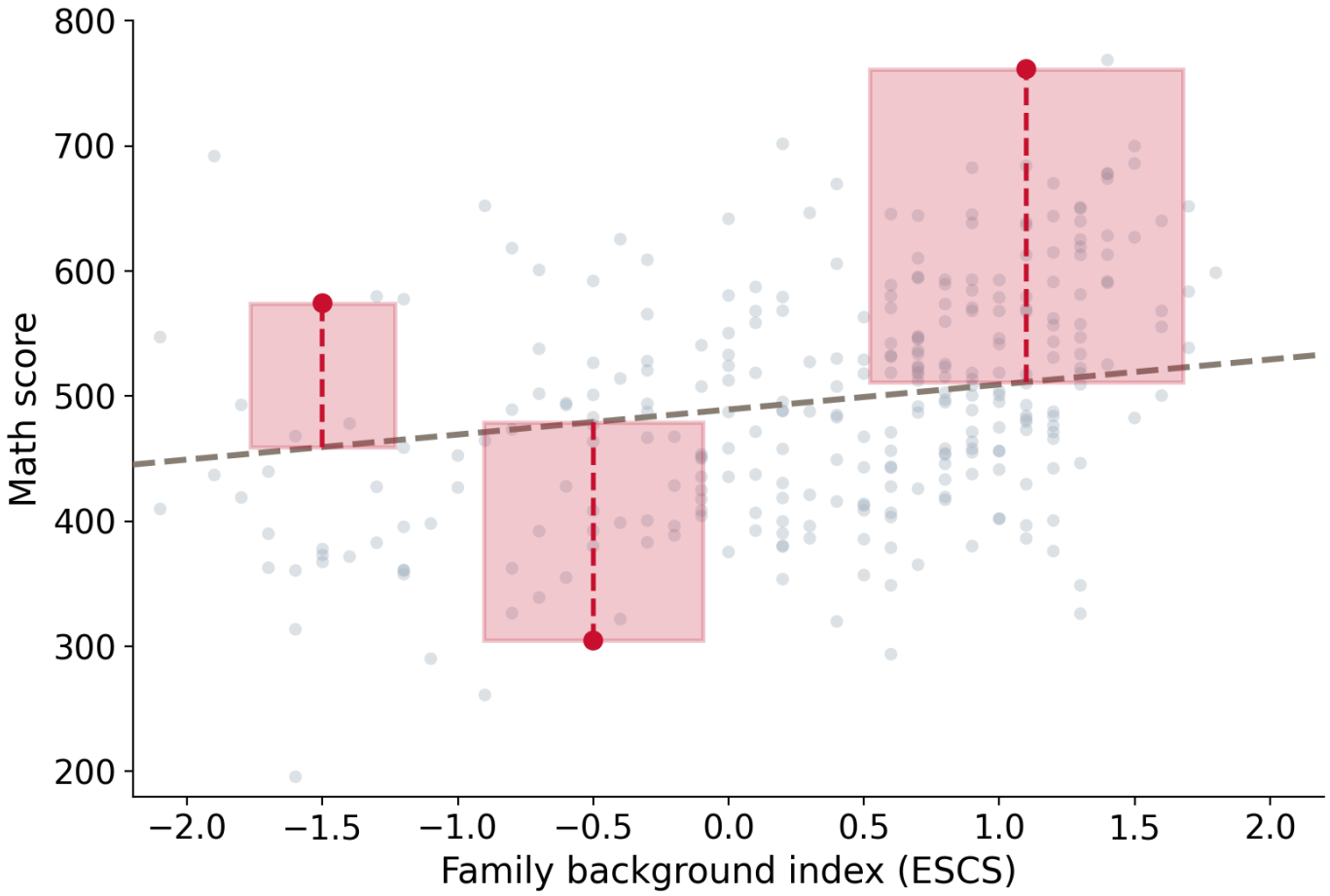
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Squared residuals, steep line



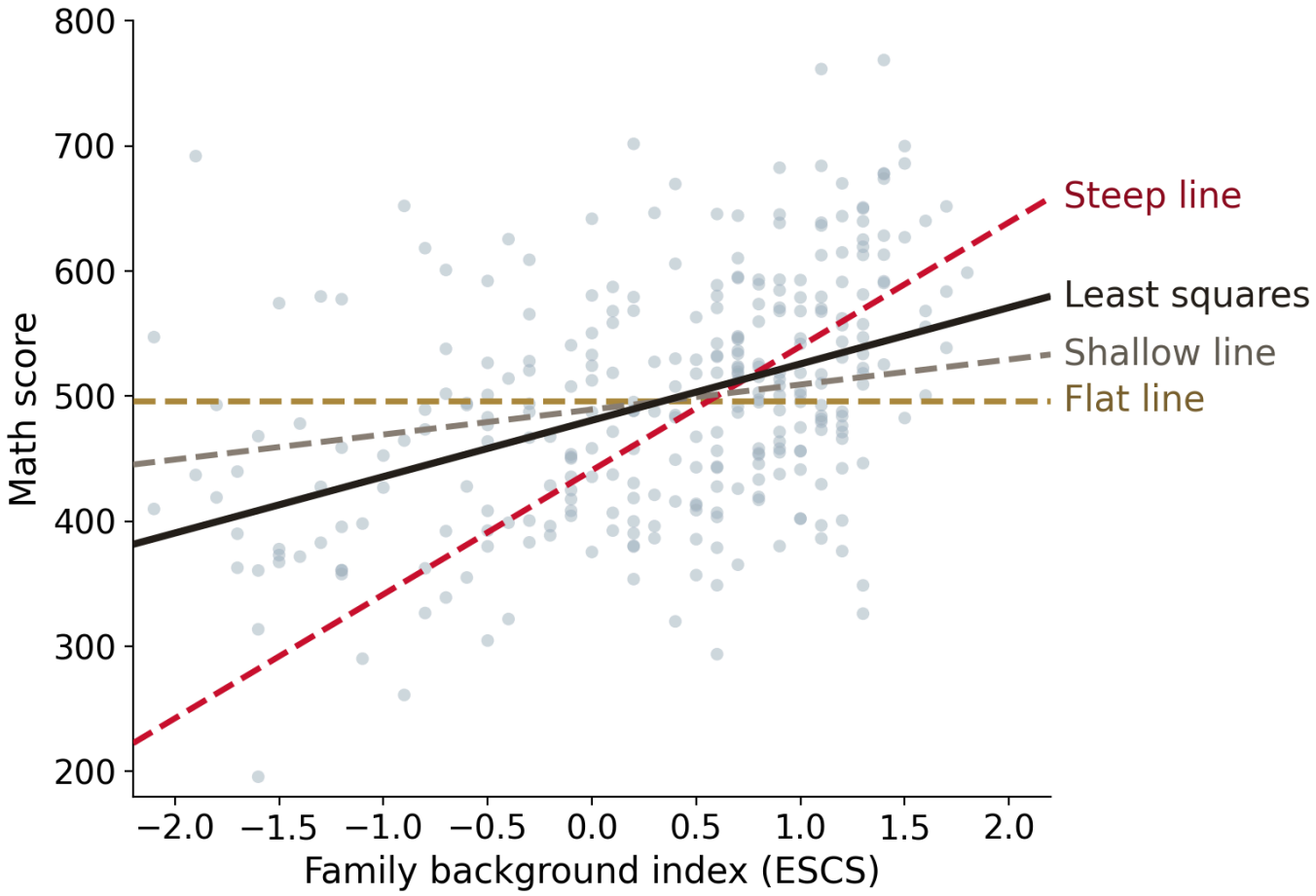
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Squared residuals, shallow line



Data: PISA 2022, Australia (OECD), 300 students drawn at random

Least squares: the smallest total area



Data: PISA 2022, Australia (OECD), 300 students drawn at random

The slope is covariance divided by variance

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)} = \frac{35.0}{0.777} = 45.1$$

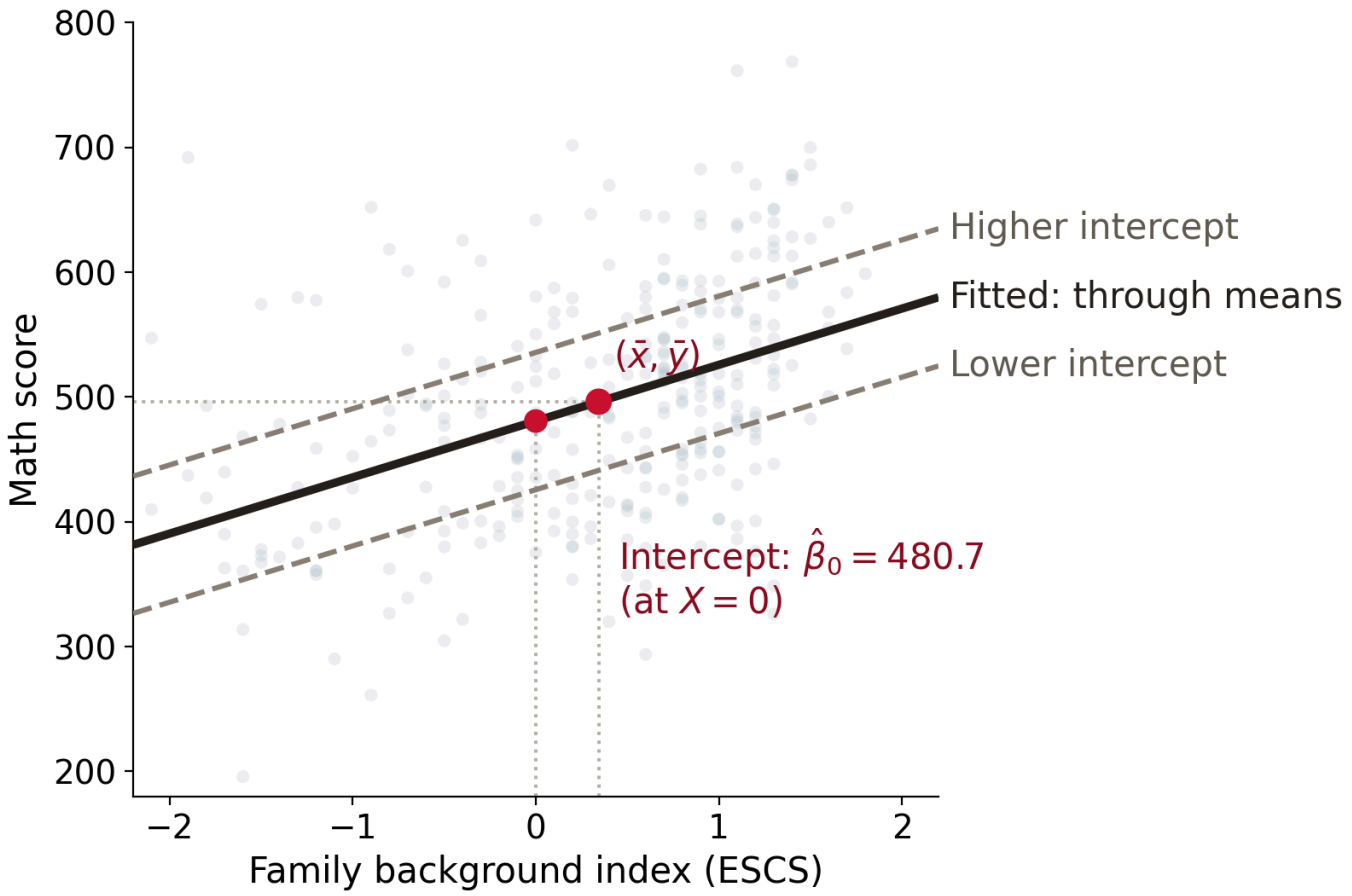
The intercept anchors the line at the means

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 496.2 - 45.07(0.343)$$

$$= 480.7$$

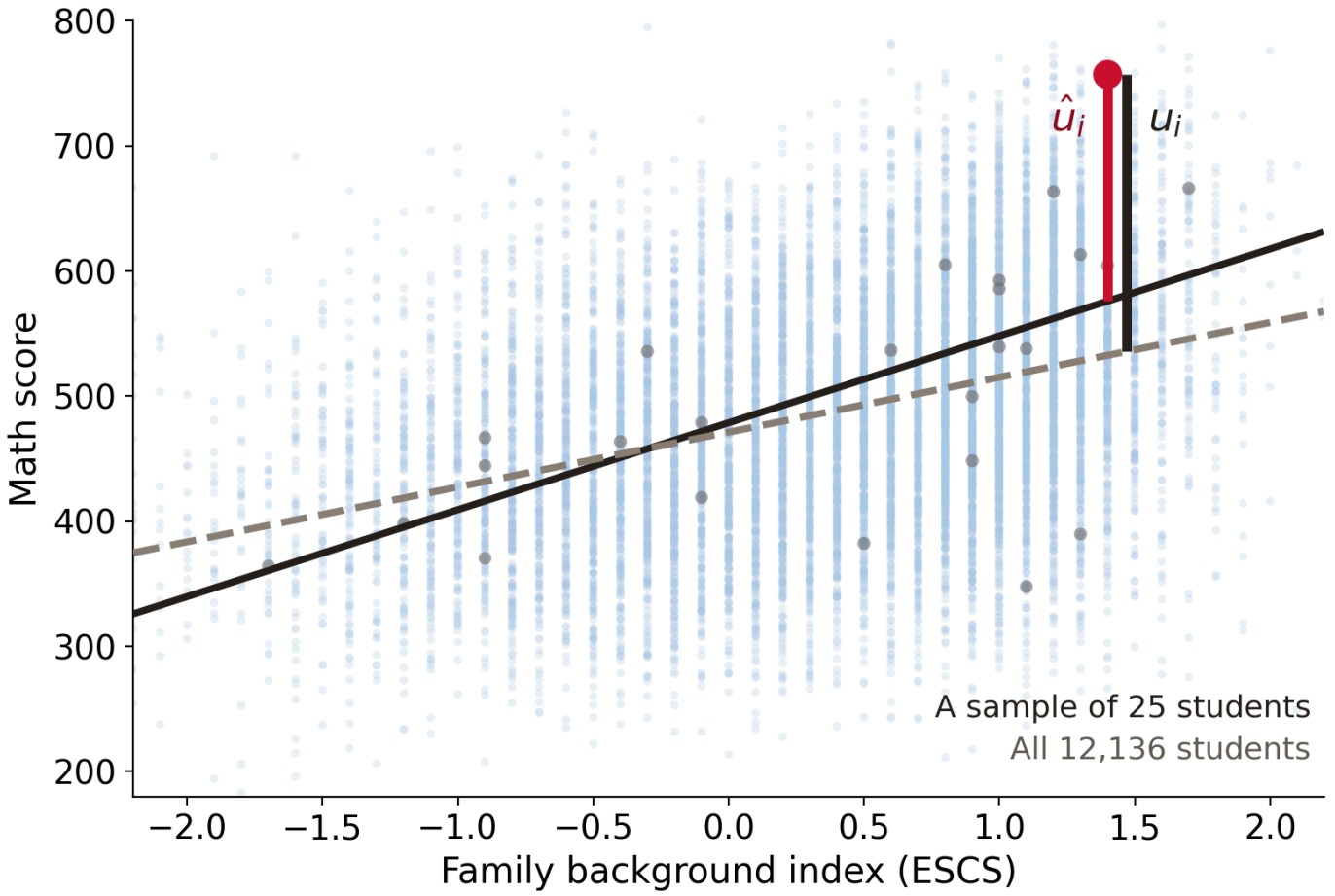


Interpret the fitted slope and intercept

$$\widehat{\text{math}} = 481 + 45 \times \text{ESCS}$$

- **Slope:** students whose ESCS index is 1 point higher score 45 points higher on average
- **Intercept:** a student at ESCS = 0, is predicted to score 481

Small samples can move the fitted line



Data: PISA 2022, Australia (OECD); dashed line fitted on all 12,136 students, solid line on a sample of 25

Some take-aways

- Covariance has nonsense units
- When estimated conditional means line up, we can summarize them by a straight line
- Least squares picks the line with the smallest sum of squared residuals
- The slope of our regression line is $\mathrm{cov}(X, Y)/\mathrm{var}(X)$
- Residuals are signed gaps to our line