Least-Squares Assumptions and Estimation in Stata

16 Sep 2026 | Chapter 4

ECO 351 – Introduction to Econometrics

Shi Qiu – Fall 2026

Recall: Population and fitted objects

From Lecture 7

Lecture 7 distinguished the unknown population model from the line and residuals calculated using the California sample.

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i$$

Which objects are computed from the sample, and which population objects remain unknown?

Why a fitted line is not automatically causal

Answer

The estimates $\widehat{\beta}_0, \widehat{\beta}_1$, fitted values, and residuals are computed from the sample. The population parameters β_0, β_1 and population errors u_i remain unknown. The new question is when the fitted sample line can inform that population relationship.

When can a fitted sample line teach us about the unknown population line?

The fitted line and the population line answer different questions

What OLS guarantees

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

$$\sum_i \widehat{u}_i = 0$$

What we want to learn

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

$$E(u_i \mid X_i) = 0$$

Today's question

What assumptions allow the left-hand sample calculation to inform the right-hand population relationship?

The error contains the competing explanations

$$Score_i = \beta_0 + \beta_1 STR_i + u_i$$

1. Teacher quality, family resources, student preparation, and language background can affect scores.
2. If one of these factors also moves systematically with STR, the simple line mixes its contribution with class size.
3. The problem is not merely that other factors exist; it is that their average may differ across STR values.

Ask the substantive question

Among districts with a given STR, are the omitted score determinants centered at the same benchmark?

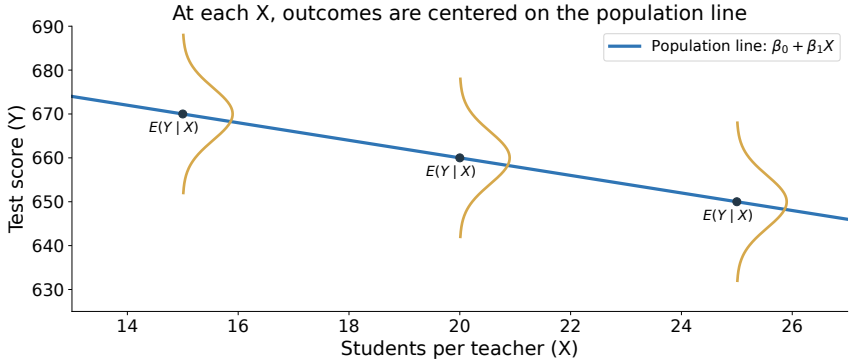
At every STR, omitted determinants must average to zero

$$E(u_i \mid STR_i = x) = 0 \quad \text{for every relevant } x$$

$$E(Score_i \mid STR_i = x) = \beta_0 + \beta_1 x$$

1. Errors need not equal zero for individual districts.
2. Positive and negative omitted contributions must average to zero within each STR group.
3. Then the population regression line is the conditional mean of scores given STR.

Conditional distributions are centered on the population line



At STR 15, 20, and 25, individual scores can vary. Zero conditional mean requires each conditional distribution to be centered on the line.

Evaluate a teacher-assignment rule

A school system gives extra teachers to districts whose students need the most support. Student need is omitted from the simple regression. Explain the two links and whether zero conditional mean is plausible.

Solution: Evaluate a teacher-assignment rule

1. Greater student need tends to lower test scores.
2. Greater student need leads to more teachers and therefore a lower STR.
3. Low-STR districts consequently have systematically less favorable omitted score determinants.

$$E(u_i \mid STR_i) \neq 0$$

The simple STR slope combines the class-size relationship with targeted teacher assignment. Its bias need not have the same sign as in a system where disadvantaged districts instead have larger classes.

A residual identity cannot verify a population assumption

$$\underbrace{\sum_i \widehat{u}_i = 0}_{\text{true after OLS with an intercept}}$$

$$\underbrace{E(u_i \mid X_i) = 0}_{\text{assumption about unobserved population errors}}$$

1. Residuals use the fitted sample line; errors use the unknown population line.
2. OLS centers residuals overall and makes their sample covariance with X equal zero.
3. Those algebraic identities hold even when an omitted determinant is systematically related to X.

Random sampling determines what population the sample represents

1. Treat (X_i, Y_i) as independently and identically distributed draws from the population of interest.
2. Identically distributed means each observation follows the same population sampling rule.
3. Independence rules out one sampled district mechanically determining another sampled district's observation.

Scope matters

A sample of California districts does not automatically represent schools in every country, grade level, or year.

Variation and controlled tails keep the slope informative

1. STR must vary: $\text{Var}(STR) > 0$. Otherwise no slope can be learned.
2. Extremely large values of X, Y, or the error must be sufficiently unlikely; the relevant fourth moments are finite.
3. Normal errors are not required for the large-sample OLS results used in this course.

Why this matters

A few extreme observations can dominate covariance, the OLS slope, and its sampling behavior.

Each assumption has one main job

Assumption	Main role
$E(u \mid X) = 0$	Centers OLS on the intended population line
i.i.d. sampling	Supports repeated-sample generalization
Finite fourth moments	Keeps large-sample approximations well behaved
$\mathrm{Var}(X) > 0$	Makes a slope estimable

Do not substitute jobs

More observations can reduce sampling noise; they cannot repair systematic sorting between X and omitted determinants.

Will a larger sample repair targeted assignment?

Increase the sample from 420 to 4,200 districts, but keep the same rule that assigns extra teachers to districts with greater unobserved student need. What is likely to improve, and what can remain wrong?

Solution: Will a larger sample repair targeted assignment?

Likely improvement

Sampling uncertainty can shrink because the estimator averages information from more districts.

Problem that remains

The association between STR and omitted student need is systematic, so zero conditional mean can still fail.

A large sample can estimate the wrong population comparison very precisely. Sample size addresses sampling noise, not the target created by the assignment process.

Stata: inspect, estimate, and retain the residuals

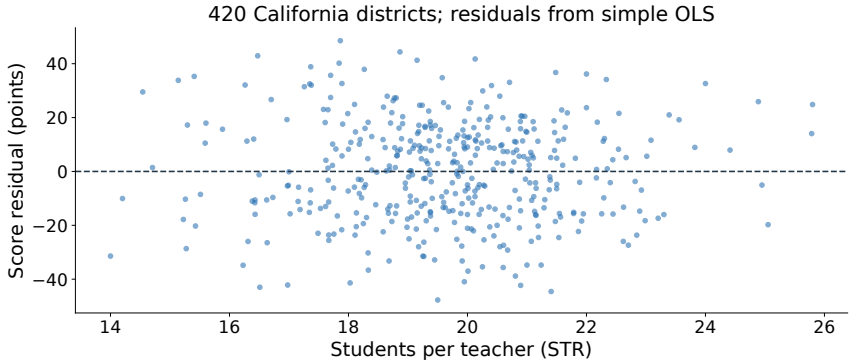
First verify the observation unit, ranges, and complete-case count. Then fit the same sample line and inspect what remains observable after fitting.

```
do "setup.do" schools
summarize score str el
count if missing(score,str,el)
regress score str
predict u_hat, residuals
summarize u_hat
twoway scatter u_hat str, yline(0)
```

Read the result

All 420 districts are complete on score, STR, and EL. OLS again gives $\widehat{Score} = 698.93 - 2.280 STR$. The residual mean is numerically zero by construction.

The residual plot can diagnose only visible sample patterns



The horizontal line is zero. The residuals are dispersed around it, with several large positive and negative deviations.

Does a clean residual plot establish zero conditional mean?

A residual plot has no obvious curve, and its residual mean is essentially zero. A student concludes that the STR slope identifies the effect of reducing class size. Is that conclusion justified?

Solution: Does a clean residual plot establish zero conditional mean?

No. A zero residual mean is imposed by OLS with an intercept, and a shapeless residual plot cannot reveal unmeasured student need, family resources, or teacher allocation. The plot is reassuring about visible functional-form problems, but a population and policy interpretation still requires a credible assignment argument.

What the least-squares assumptions add

Conclusion

OLS guarantees sample fitting identities. Learning about the population line requires zero conditional mean, an appropriate sampling framework, sufficient variation in X, and controlled tails. The central empirical question is whether omitted score determinants move systematically with STR; neither a large sample nor a tidy residual plot can answer that automatically.

Which assumption concerns the target of the slope, and which assumptions mainly support repeated-sample approximation?