The Linear Regression Model and Measures of Fit

14 Sep 2026 | Chapter 4

ECO 351 – Introduction to Econometrics

Shi Qiu – Fall 2026

Recall: The fitted school line

From Lecture 6

Lecture 6 moved quickly from the superintendent's class-size question to the following fitted result.

$$\widehat{Score} = 698.93 - 2.280 STR, \qquad \widehat{Score}(18) - \widehat{Score}(20) = 4.56$$

What do the hats tell us? Which quantities were calculated from the sample, and which population quantities remain unknown?

Before accepting the fitted answer

Answer

The equation with hats is the fitted sample line. The numbers 698.93 and -2.280 are coefficient estimates; fitted values and residuals are calculated from that line. The population parameters β_0, β_1 and population errors u_i remain unknown. Lecture 7 keeps these objects separate before evaluating the fitted line.

We need names that keep the unknown population relationship separate from the line fitted to one sample.

Keep the population and sample columns separate

Population

Parameters: β_0, β_1

Regression line: $\beta_0 + \beta_1 X$

Error: u_i

Observed sample

Estimates: $\widehat{\beta}_0, \widehat{\beta}_1$

Fitted line: $\widehat{\beta}_0 + \widehat{\beta}_1 X$

Residual: $\widehat{u}_i$

Central distinction

Hats mark quantities calculated from this sample. They do not turn a sample relationship into a population fact.

The population model describes each district

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

1. Y_i: the dependent variable; here, district i's test score.
2. X_i: the regressor; here, district i's student–teacher ratio.
3. $\beta_0 + \beta_1 X_i$: the population regression line evaluated at X_i.
4. u_i: all other factors that place Y_i above or below that line.

What population means

The relationship concerns the population of districts we want to understand, not merely the 420 rows in the dataset.

The fitted line and the fitted model are sample statements

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

$$Y_i = \widehat{Y}_i + \widehat{u}_i$$

1. $\widehat{Y}_i$: fitted or predicted value on the sample regression line.
2. $\widehat{u}_i = Y_i - \widehat{Y}_i$: residual, the observed vertical distance from that line.
3. $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i$: the fitted-model decomposition for observation i.

Parameter, estimator, and estimate are not synonyms

Parameter

β_1 is the fixed but unknown population slope.

Estimator

$\widehat{\beta}_1$ is the rule that turns sample observations into a slope; it would vary across repeated samples.

Estimate

-2.280 is the numerical value obtained from this particular California sample.

The same pattern

Population error u_i is unobserved; sample residual $\widehat{u}_i$ is calculated after fitting the line.

OLS turns candidate coefficients into fitted coefficients

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2}, \qquad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

1. b_0, b_1: candidate numbers considered by the fitting rule.
2. $\widehat{\beta}_0, \widehat{\beta}_1$: the candidates that minimize squared prediction mistakes.
3. Squaring prevents positive and negative mistakes from canceling.

Separate question

OLS tells us how to fit the sample line. It does not, by itself, make the fitted slope causal.

Name and calculate the sample objects

The fitted California line is $\widehat{Score} = 698.93 - 2.280 STR$. A constructed district has $STR = 21$ and an observed score of 640. (a) Find its fitted value and residual. (b) Identify the estimates, fitted value, residual, and the population quantities that remain unknown.

Solution: Name and calculate the sample objects

$$\widehat{Score} = 698.93 - 2.280(21) = 651.05$$

$$\widehat{u} = 640 - 651.05 = -11.05$$

1. Estimates: $\widehat{\beta}_0 = 698.93$ and $\widehat{\beta}_1 = -2.280$.
2. Fitted value: 651.05; residual: -11.05 score points.
3. Still unknown: population parameters β_0, β_1 and population error u_i.

The negative residual places this district below the fitted sample line.

Stata: create the fitted values and residuals

Read one displayed row horizontally: observed score and STR come from the data; fitted score and residual were created after regress.

```
do "setup.do" schools
regress score str
predict score_hat, xb
predict u_hat, residuals
list score str score_hat u_hat in 1/3, noobs abbrev(16)
summarize u_hat
```

Read the result

Stata estimates $\widehat{Score} = 698.93 - 2.280 STR$. For every listed district, verify $Score = \widehat{Score} + \widehat{u}$. The residual mean is numerically zero; the regression table reports $R^2 = 0.0512$ and Root MSE about 18.58.

OLS imposes useful identities on the fitted sample

$$\sum_i \widehat{u}_i = 0, \qquad \sum_i X_i \widehat{u}_i = 0, \qquad \overline{\widehat{Y}} = \bar{Y}$$

1. Residuals average to zero when the fitted regression includes an intercept.
2. The sample covariance between the regressor and residual is zero.
3. The average fitted value equals the average observed outcome.

Do not confuse identities with assumptions

$\sum_i X_i \widehat{u}_i = 0$ holds because OLS fitted this sample. It does not demonstrate that $E(X_i u_i) = 0$ in the population.

The fitted difference has an economic scale

$$\widehat{Score}(17.7) - \widehat{Score}(19.7) = (-2.280)(-2) = 4.56$$

1. An STR of 19.7 is the sample median; 17.7 is near the 10th percentile.
2. A score of 654.5 is the sample median; adding 4.56 reaches about 659.1, just below the 60th percentile.
3. The contemplated staffing change is large, while the fitted score movement is noticeable but not transformative.

Still a fitted association

This comparison describes movement along the sample regression line, not yet the effect of hiring teachers in one district.

A fitted line should not be trusted far beyond the sample

$$\widehat{Score}(5) = 698.93 - 2.280(5) = 687.53$$

Why the arithmetic is not enough

The smallest observed STR is about 14. The data contain no districts near STR 5, so the fitted line has no empirical support there.

1. Interpolation compares values inside the observed regressor range.
2. Extrapolation extends the fitted relationship beyond that range.

Practical rule

Always compare the proposed X value with the data range before reporting a prediction.

Fit separates variation represented by the line from what remains

$$Y_i - \bar{Y} = (\widehat{Y}_i - \bar{Y}) + \widehat{u}_i$$

$$TSS = ESS + SSR$$

1. $TSS = \sum_i (Y_i - \bar{Y})^2$: total outcome variation.
2. $ESS = \sum_i (\widehat{Y}_i - \bar{Y})^2$: variation represented by the fitted line.
3. $SSR = \sum_i \widehat{u}_i^2$: variation left in the residuals.

R-squared is a fraction of sample variation

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \qquad 0 \leq R^2 \leq 1$$

California schools

$R^2 = 0.0512$: STR represents 5.1% of the sample variation in district scores; 94.9% remains in the residual sum of squares.

Simple-regression connection

With one regressor and an intercept, $R^2 = r_{X,Y}^2 \approx (-0.226)^2$.

R-squared does not measure causality or identify omitted factors.

SER measures residual spread in outcome units

$$SER = \sqrt{\frac{SSR}{n-2}}$$

California schools

$SER \approx 18.58$ score points: scores remain widely dispersed around the STR-only fitted line.

Why n minus 2?

Estimating the intercept and slope uses two degrees of freedom.

SER is residual outcome spread; a coefficient SE is sampling uncertainty.

Compare the fit of two sample regressions

Two regressions use the same outcome and the same 102 observations, so both have $TSS = 10{,}000$. Model A has $SSR = 6{,}400$; Model B has $SSR = 3{,}600$. Each model estimates an intercept and one slope. Compute R^2 and SER for both. Which line fits this sample more closely?

Solution: Compare the fit of two sample regressions

$$R_A^2 = 1 - \frac{6400}{10000} = 0.36, \qquad SER_A = \sqrt{\frac{6400}{100}} = 8$$

$$R_B^2 = 1 - \frac{3600}{10000} = 0.64, \qquad SER_B = \sqrt{\frac{3600}{100}} = 6$$

Model B represents more of the outcome variation and leaves a smaller residual spread, so it fits this sample more closely. This comparison alone does not establish that Model B is causal or correctly specified.

What a fitted regression line tells us

Conclusion

The population model contains unknown parameters and errors. OLS uses one sample to produce coefficient estimates, a fitted regression line, fitted values, and residuals. The California line gives a 4.56-point fitted comparison, but its low R-squared and 18.58-point SER show that STR alone leaves most score variation unrepresented.

Which statements in this lecture are sample identities, and which claims would require assumptions about the population error?