Covariance, Correlation, and Scatterplots

07 Sep 2026 | Chapter 3

ECO 351 – Introduction to Econometrics

Shi Qiu – Fall 2026

Recall: Uncertainty about a mean

From Lecture 4

A survey of 200 graduates estimated mean hourly earnings and its uncertainty as follows.

$$\bar{Y} = 22.64, \quad SE(\bar{Y}) = 1.28, \quad CI = [20.13, 25.15]$$

Does this interval contain 95% of individual wages? State the population quantity and what the 95% confidence level describes.

Answer

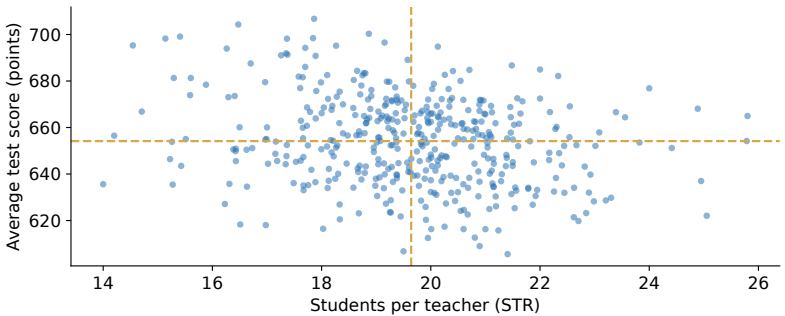
The interval estimates the population *mean* wage; it does not describe the spread of individual wages.

Lecture 4 studied one variable at a time. Today each observation carries two measurements, and their pairing matters.

For a school district, does an above-average value of one variable tend to come with an above- or below-average value of the other?

Do districts with larger classes have lower scores?

Each of 420 California districts contributes a student-teacher ratio (STR) and an average test score.



One point is one district; the dashed lines mark the two sample means. The cloud slopes downward, but it is also widely dispersed.

Measure each district relative to the two means

For district i, compare X_i with $\bar{X}$ and Y_i with $\bar{Y}$. Here X is STR and Y is test score.

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

1. Both above their means, or both below: the product is positive.
2. One above and one below: the product is negative.

Covariance records signed co-movement

A point above both means

If $(X_i - \bar{X}) > 0$ and $(Y_i - \bar{Y}) > 0$, then

$$(X_i - \bar{X})(Y_i - \bar{Y}) > 0.$$

Such a point contributes positively to covariance.

The sample summary

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

The sign summarizes whether above-average STR tends to accompany above-average or below-average scores.

A worked covariance calculation

Constructed observations: X is weekly training hours and Y is a productivity score.

X	Y	$X - \bar{X}$	$(X - \bar{X})(Y - \bar{Y})$
1	4	-1	2
2	5	0	0
3	9	1	3

$$\bar{X} = 2, \quad \bar{Y} = 6, \quad s_{XY} = \frac{2 + 0 + 3}{3 - 1} = 2.5$$

Compute covariance in a small district sample

For three illustrative districts, STR is 18, 20, 22 and score is 660, 650, 640. Find the two means, the three deviation products, and the sample covariance.

Solution: Compute covariance in a small district sample

$$\bar{X} = 20, \quad \bar{Y} = 650, \quad \text{products} : -20, 0, -20$$

$$s_{XY} = (-40)/(3-1) = -20$$

The negative sign reflects lower scores in the higher-STR districts in this constructed sample.

The size of covariance depends on the units

Suppose we report scores in tens of points instead of points. Every score deviation and every covariance contribution becomes one tenth as large.

$$Y^* = Y/10 \quad \Longrightarrow \quad s_{XY^*} = s_{XY}/10$$

The scatterplot has the same pattern after relabeling its axis. We need a standardized measure to compare linear strength across units.

Correlation divides out the two scales

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}, \qquad -1 \leq r_{XY} \leq 1$$

The denominator is positive and has the same units as covariance. The sign is unchanged; the units cancel.

The small example

$s_X = 2$, $s_Y = 10$, so $r_{XY} = -20/(2 \times 10) = -1$. The three constructed points lie exactly on a downward-sloping line.

Would changing the score units change correlation?

Given

STR remains $18, 20, 22$, with sample SD $s_{STR} = 2$.

Your task

Return to the three-district sample. Express scores as 66, 65, 64. Covariance becomes -2. Calculate the new correlation and explain why it does not change.

Solution: Would changing the score units change correlation?

$$s_{Y^*} = 1, \qquad r_{XY^*} = \frac{-2}{2 \times 1} = -1$$

Positive rescaling affects the numerator and denominator equally. A negative rescaling would reverse the sign.

Stata: return to the school districts

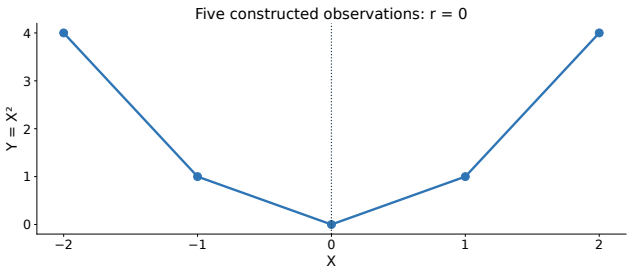
Before running the commands, predict the sign of the correlation. Should its magnitude be close to one?

```
do "setup.do" schools
twoway scatter score str
correlate score str
summarize score str
```

What the output adds

$r_{score,STR} \approx -0.226$: a negative but modest linear association, consistent with the direction and wide dispersion in the graph.

Correlation describes only the linear part



For $X = -2, -1, 0, 1, 2$, let $Y = X^2$.

Knowing X determines Y exactly, yet the left- and right-side covariance contributions cancel:

$$r_{XY} = 0.$$

Lesson

Zero correlation rules out a linear pattern, not every relationship.

What should we check about an unusual district?

In a scatterplot, one district has a very high STR and a very low score. What should we check before deciding whether to exclude it?

Solution: What should we check about an unusual district?

1. Check whether the values and units were recorded correctly.
2. Check whether the district belongs to the target population.
3. Compare the summary with and without it and explain the difference.

An inconvenient observation is not, by itself, a reason for deletion.

A downward pattern is not yet a class-size effect

Districts may differ in income, language composition, and other determinants of scores. Those characteristics may also be related to STR.

What we can report

In these data, higher STR tends to accompany lower test scores. We have not isolated what would happen if a district changed class size.

Correlation still cannot answer a prediction in points

Suppose STR falls from 20 to 18. A correlation of -0.226 does not tell us the predicted change in score points.

The next tool we will need

A regression line provides an intercept and a slope, so it can turn a specified change in STR into a predicted score change.

A disciplined way to describe two variables

1. Keep the observations paired and inspect the scatterplot.
2. Use covariance to summarize the direction of co-movement.
3. Use correlation to compare linear strength without units.
4. Check shape and unusual observations before trusting one number.

In the school data, what would be lost if we reported only $r = -0.226$?