REGRESSION REVIEW

Fundamentals of

PROGRAM EVALUATION

JESSE LECY

THE ROAD MAP

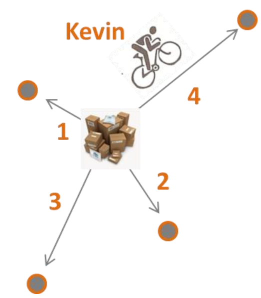
	Of the Mean:	Of the Slope:
Variance:	$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ (for x)	$\sigma_\varepsilon^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$ (using the residual)
↓		
Standard Deviation:	$\sigma_x = \sqrt{\sigma_x^2}$	$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$
↓		
Standard Error:	$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2}}$
↓		
Confidence Interval	$\mu = \bar{x} \pm t \cdot SE_{\bar{x}}$ (of the mean)	$\beta_1 = b_1 \pm t \cdot SE_{b_1}$ (of the slope)

All of the statistical concepts that you have learned in the previous course using variance, standard errors, and confidence intervals of a estimates of the mean from a single variable apply to regression, but they have to be adapted.

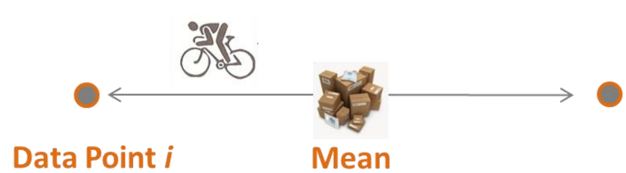
Make note that statistical concepts always need to be followed by the phrase “of the” because they are general concepts and the specific calculations are determined by the variables you are working with. The standard error around an estimated mean is different than the standard error around an estimated slope.

USEFUL METAPHORS

Variance

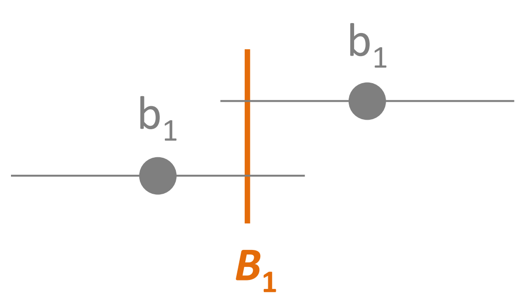


Standard Deviation



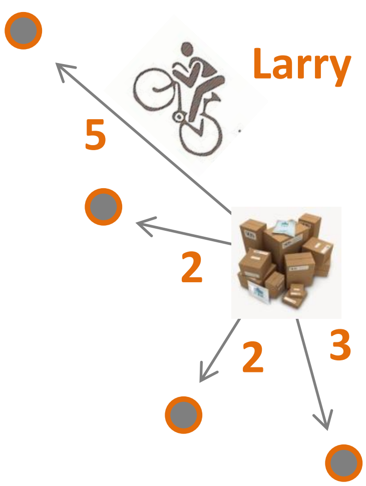
Standard Error

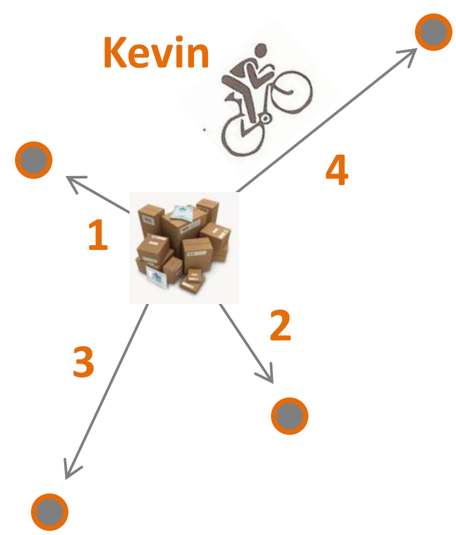
Confidence Interval



VARIANCE

Which cyclist rode the furthest?

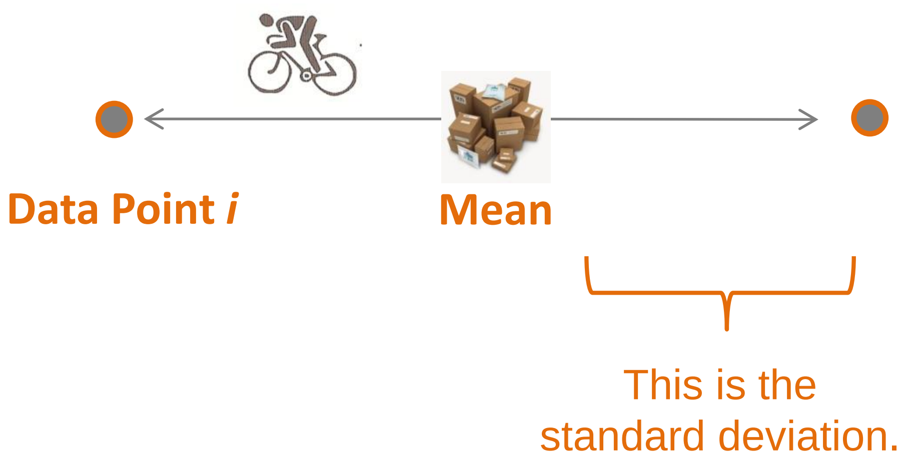




$$\frac{3+2+5+2}{4}=3$$

$$\frac{3+2+4+1}{4}=2.5$$

This is a visual metaphor for variance.



- Variance is a measure of average dispersion.
- You first need a reference point in order to calculate average distance. You can't ask "how far have I traveled if I am in Chicago?" without knowing where you started from. Use the mean as the starting point for each distance. Dispersion is the distance from the mean.
- The problem is that when you use the mean then the sum of all distances from the mean will always be zero. As a result, you must square them first.
- Divide by N so you have an average squared distance from the mean. For an estimate is actually N-1 for reasons we will not discuss here. This is variance.

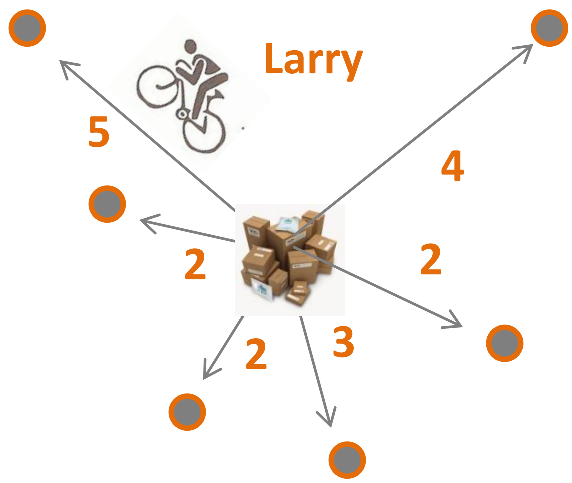
$$\sigma^2 = \frac{1}{N}\sum\left(x_i - \bar{x}\right)^2$$

- We want to reconcile the units so that the number is meaningful. We squared everything, so we take the square root. This is the standard deviation.

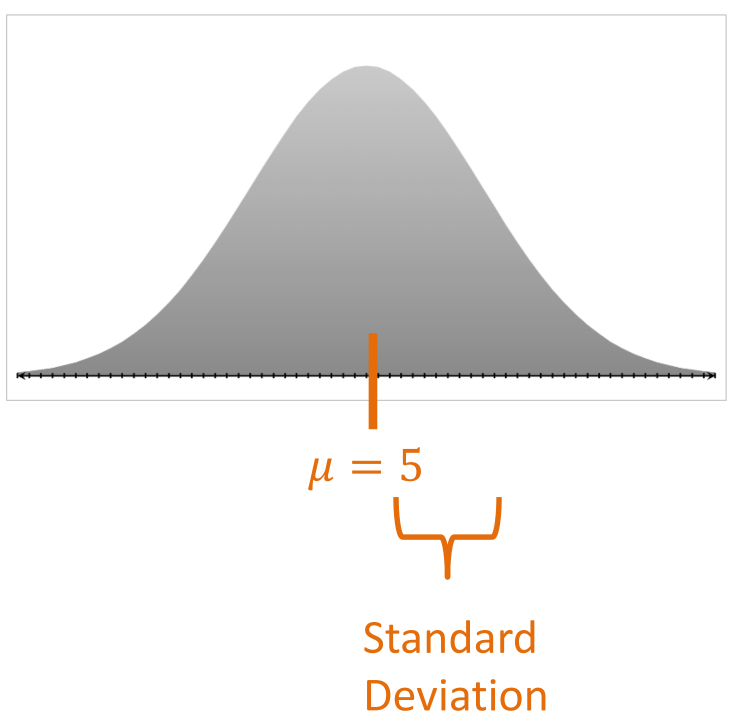
$$\sigma = \sqrt{\frac{1}{N}\sum\left(x_i - \bar{x}\right)^2}$$

- **Intuitively, it is the "average" amount each point must travel to reach the mean.**

VARIANCE



All Trips by Larry as Histogram



STANDARD DEVIATION VERSUS STANDARD ERROR

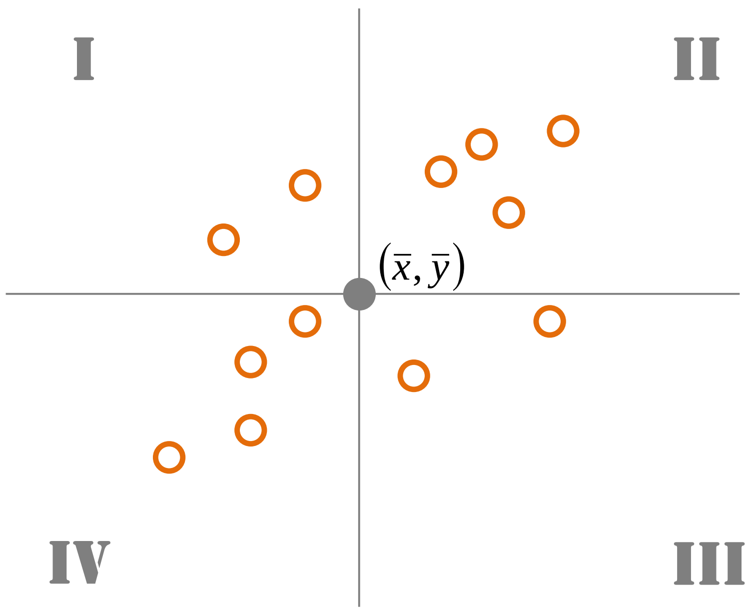
The standard deviation is, **how far the data is from the mean**, on average.

~

The standard error is, **how far our best guest is from 'the truth'**, on average.

The 'truth' means different things depending upon what kind of standard error you are calculating.

COVARIANCE



Covariance tells us, on average when X is above-average, do we expect Y to also be above average?

It helps us measure the strength of a relationship.

WHAT IS COVARIANCE?

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Classroom size and test scores

X	Y	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)
3	5	3 - 6 (-)	5 - 3 (+)	(-)(+) = (-)
5	2	5 - 6 (-)	2 - 3 (-)	(-)(-) = (+)
10	2	10 - 6 (+)	2 - 3 (-)	(+)(-) = (-)

First we convert all of our measures into distances from the mean. This allows us to determine whether an individual case is above-average (positive number) or below-average (negative number) on a measure.

WHAT IS COVARIANCE?

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Classroom size and test scores

X	Y	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)
3	5	3 - 6 (-)	5 - 3 (+)	(-)(+) = (-)
5	2	5 - 6 (-)	2 - 3 (-)	(-)(-) = (+)
10	2	10 - 6 (+)	2 - 3 (-)	(+)(-) = (-)

We then multiply the two measures to determine whether they tend to be positively or negatively related.

If someone receives an above-average level of the treatment, and their outcome is above-average. That is a positive relationship.

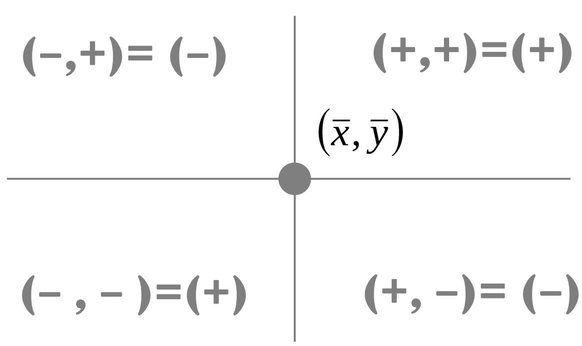
Or conversely, if they receive a below-average level of the treatment and their performance is below-average, that is also a positive relationship since (-)(-) = (+).

WHAT IS COVARIANCE?

$$\mathrm{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

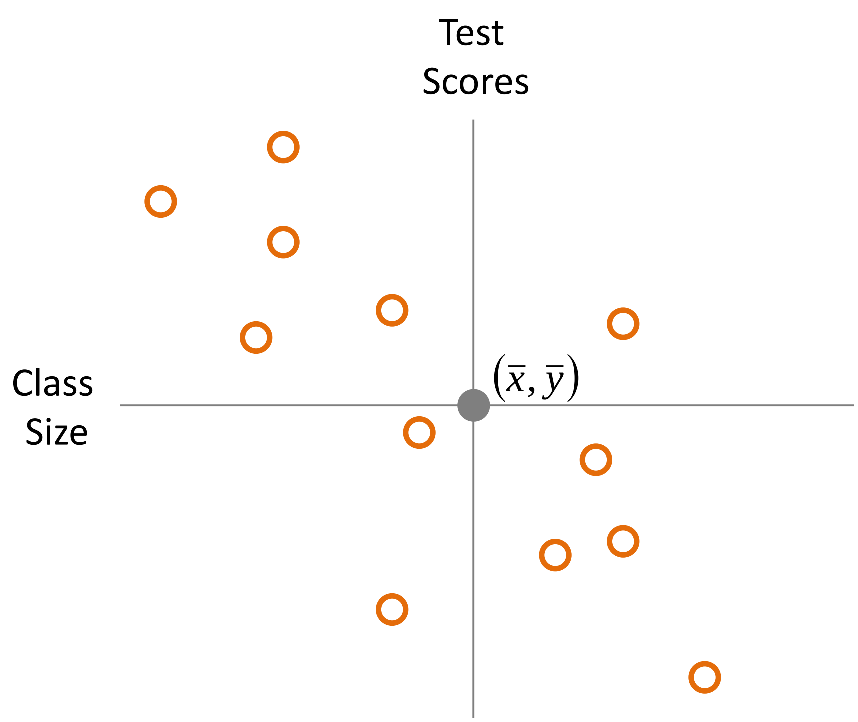
Classroom size and test scores

X	Y	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)
3	5	3 - 6 (-)	5 - 3 (+)	(-)(+) = (-)
5	2	5 - 6 (-)	2 - 3 (-)	(-)(-) = (+)
10	2	10 - 6 (+)	2 - 3 (-)	(+)(-) = (-)



NEGATIVE COVARIANCE EXAMPLE

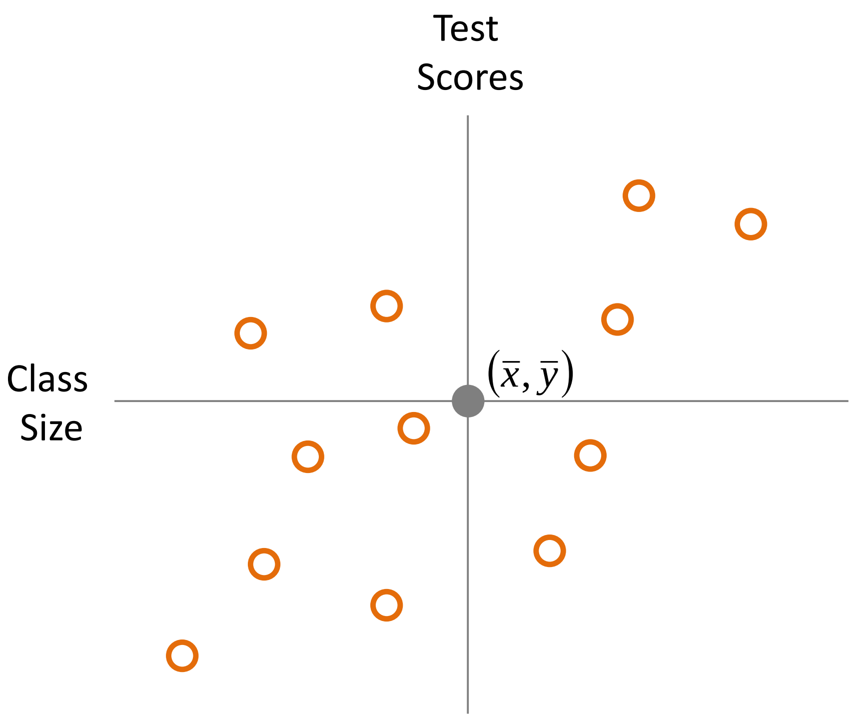
Class size and student performance



High negative correlation results in a large negative slope in a regression

POSITIVE COVARIANCE EXAMPLE

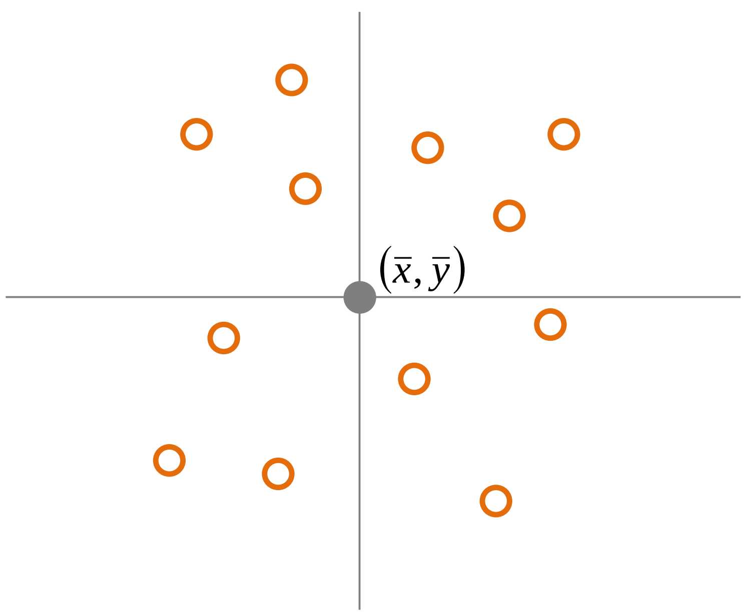
Class size and student performance



High positive correlation results in a large positive slope in a regression

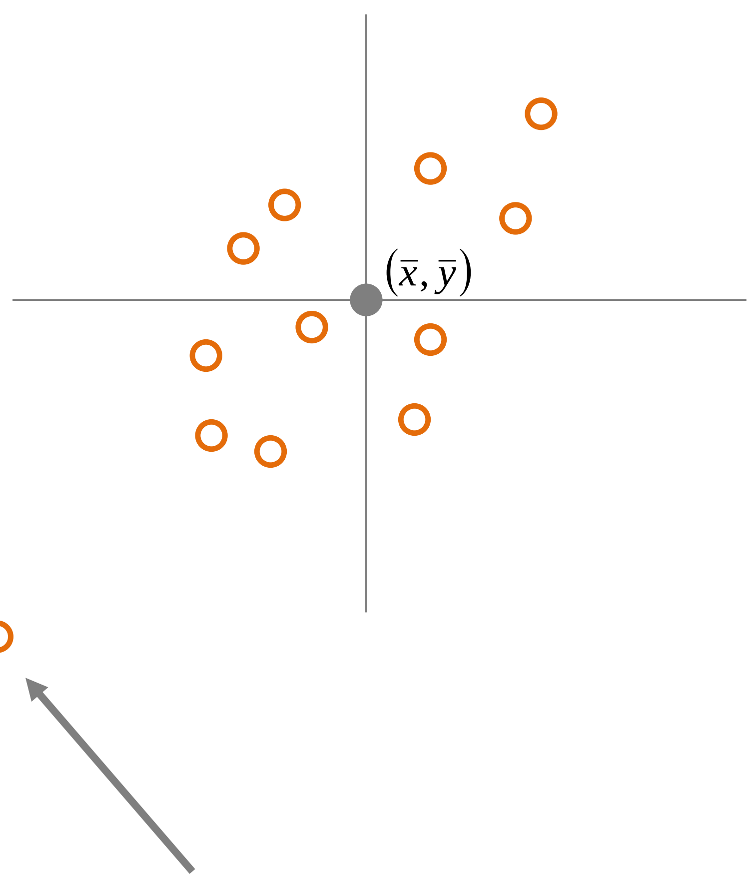
SMALL COVARIANCE

Low correlation, small b_1 in a regression



OUTLIERS

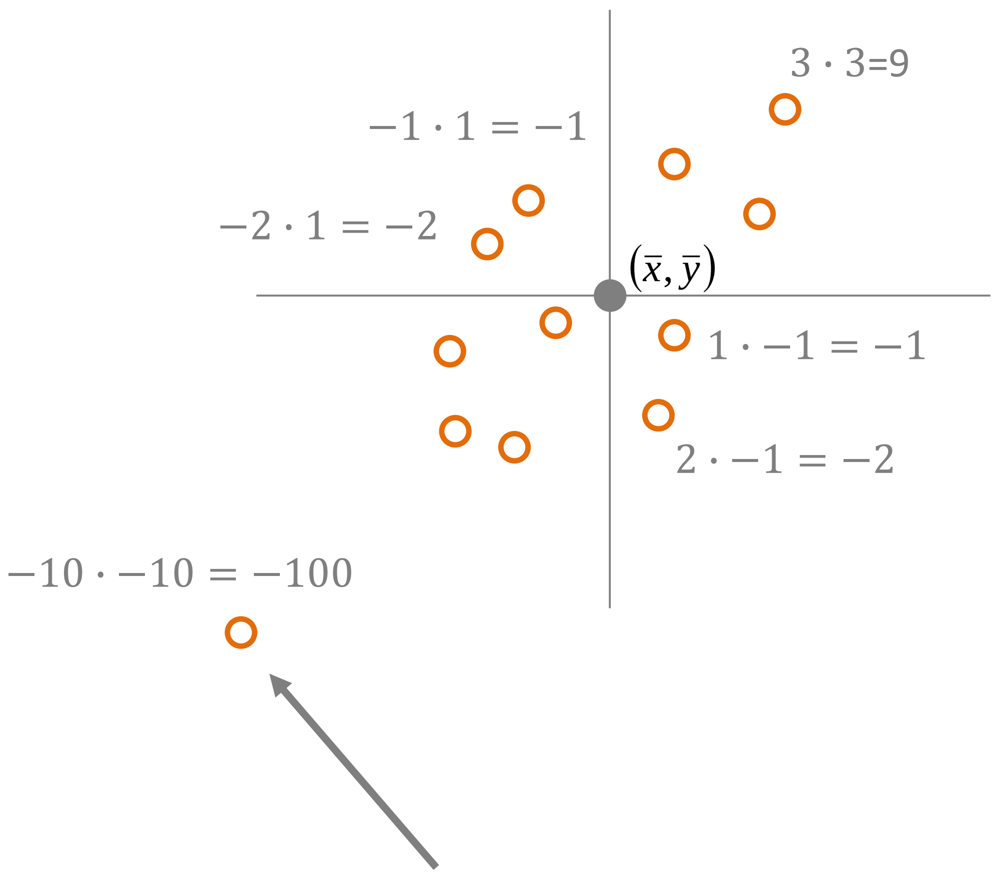
$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$



What impact will this outlier have on the covariance measure?

OUTLIERS

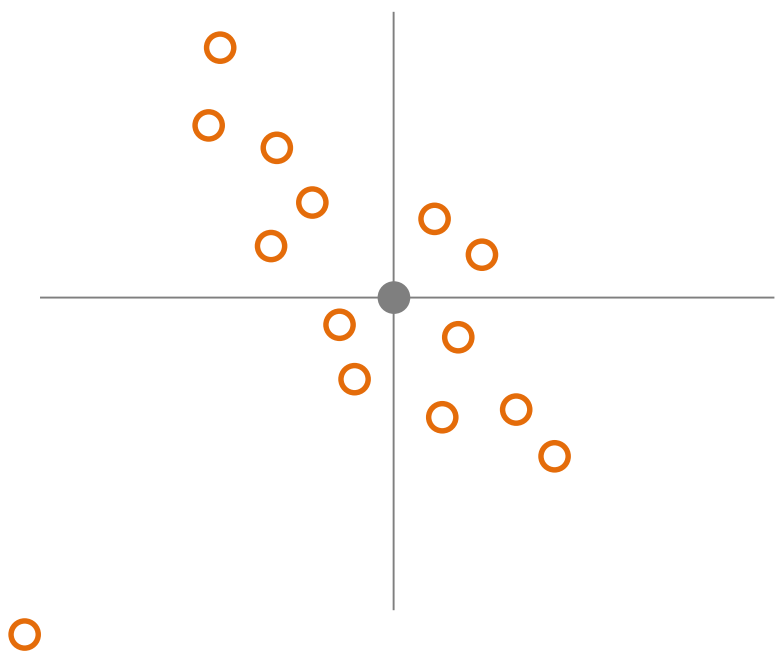
$$\operatorname{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



What impact will this outlier have on the covariance measure?

OUTLIERS

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

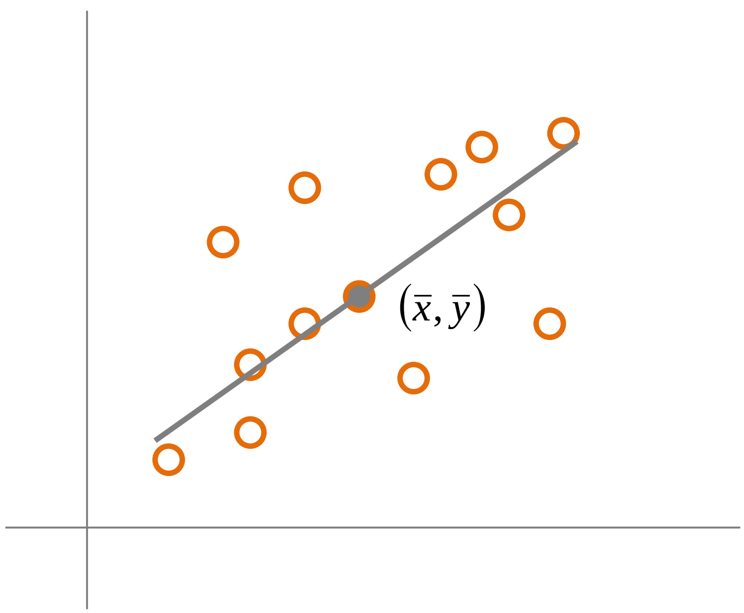


What about now?

THE REGRESSION SLOPE

THE INTUITIVE REGRESSION FORMULA

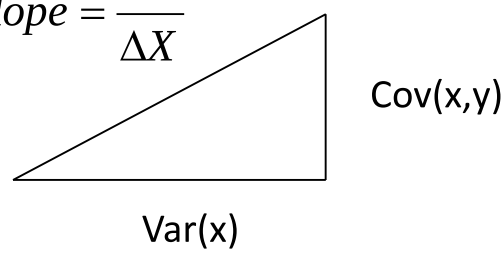
$$slope = \frac{\Delta Y}{\Delta X} = \boxed{\frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}} = \frac{\sum \cancel{(x_i - \bar{x})}(y_i - \bar{y})}{\sum \cancel{(x_i - \bar{x})}(x_i - \bar{x})}$$



Note the regression always passes through the mean of X and mean of Y.

THE INTUITIVE REGRESSION FORMULA

$$slope = \frac{\Delta Y}{\Delta X}$$



We want to interpret the slope causally, meaning the outcome is going to change ***b*₁** amount because of a one-unit change in X.

Intuitively we can think of the math as:

$$slope = \frac{\Delta Y}{\Delta X} = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \frac{\sum \cancel{(x_i - \bar{x})}\boxed{(y_i - \bar{y})}}{\sum \cancel{(x_i - \bar{x})}\boxed{(x_i - \bar{x})}}$$

Change in Y

Change in X

This is not entirely mathematically correct, but the intuition is there.

WHAT IS CORRELATION?

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\downarrow$$

$$cor(x, y) = \frac{\text{cov}(x, y)}{sd(x) \cdot sd(y)}$$

$$= \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The correlation is the covariance in simple units

WHAT IS CORRELATION?

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\downarrow$$

$$cor(x,y) = \frac{\mathrm{cov}(x,y)}{sd(x)\cdot sd(y)}$$

$$= \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Recall that the standard deviation translates the variance from sums-of-squares back into the original units.

Since we have two variables in the covariance, which unit should we use to describe the measure?

WHAT IS CORRELATION?

$$\text{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Test Score · Class Size

(this is an odd unit)

$$\downarrow$$

$$cor(x,y) = \frac{\text{cov}(x,y)}{sd(x) \cdot sd(y)}$$

$$= \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\frac{\cancel{\text{Test Score}} \cdot \cancel{\text{Class Size}}}{\cancel{\text{Test Score}} \cdot \cancel{\text{Class Size}}}$$

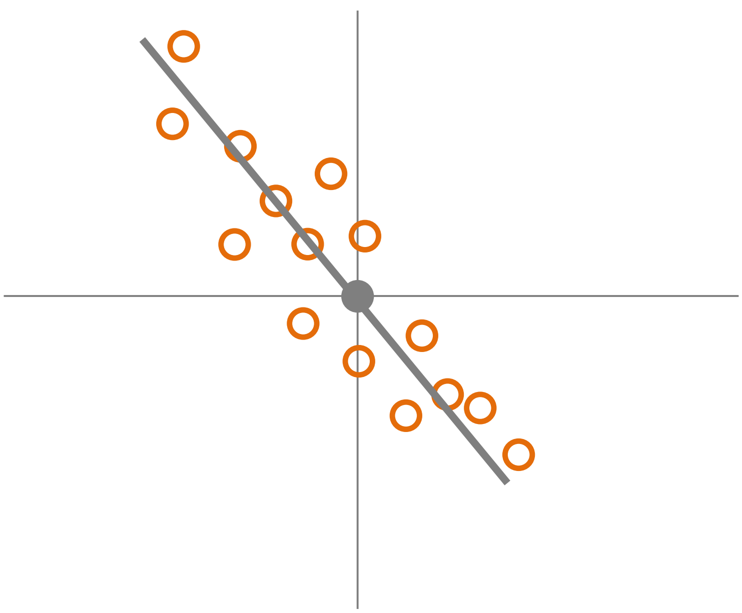
(now we have no units)

The correlation is the covariance in "standard" units:

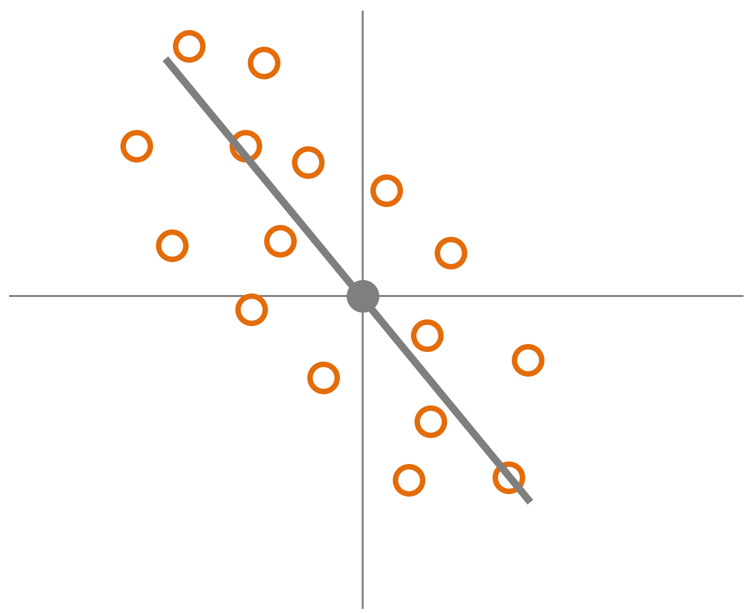
-1 < *cor(x,y)* < +1

STRENGTH OF CORRELATION

CASE A



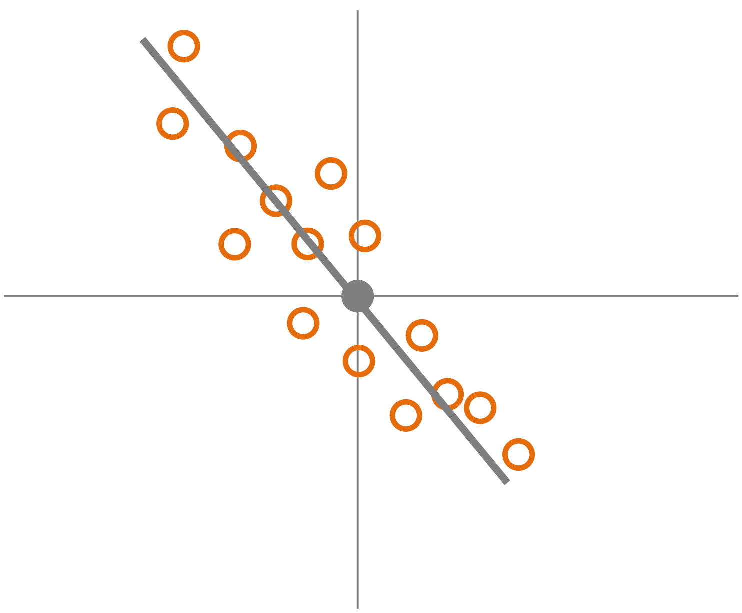
CASE B



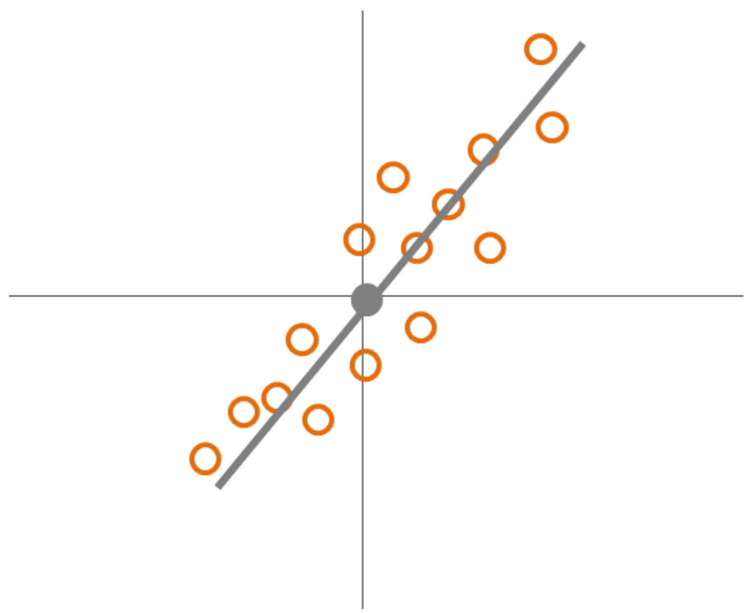
Which of these has the stronger correlation?

STRENGTH OF CORRELATION

CASE A



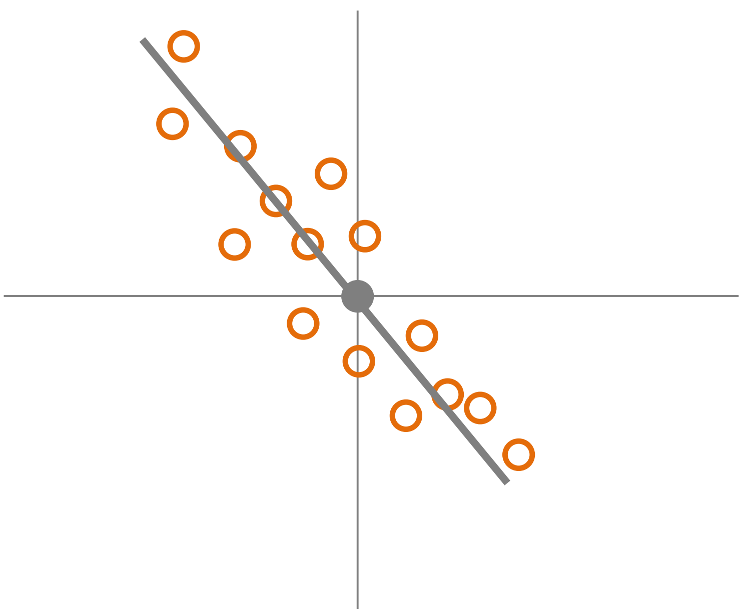
CASE B



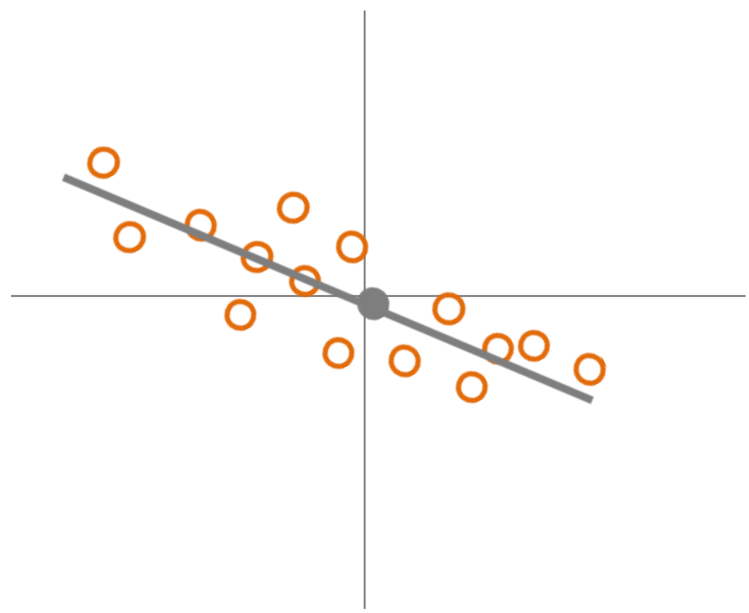
Which of these has the stronger correlation?

STRENGTH OF CORRELATION

CASE A



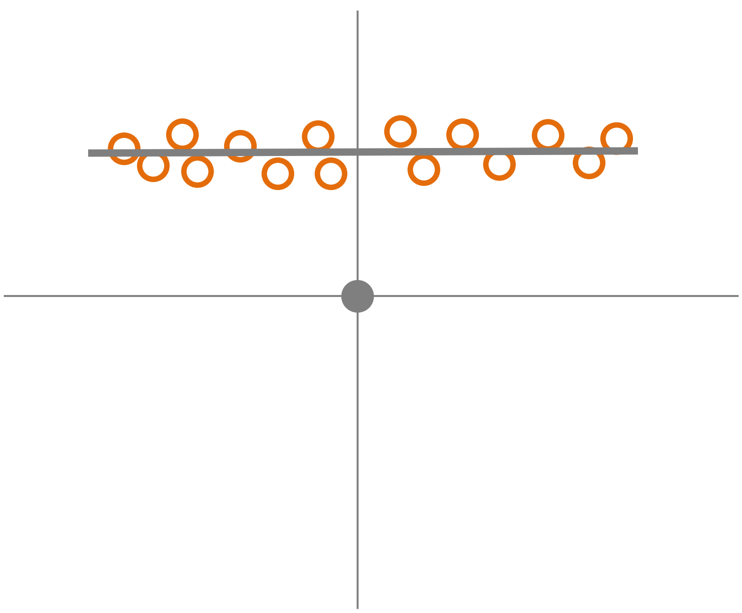
CASE B



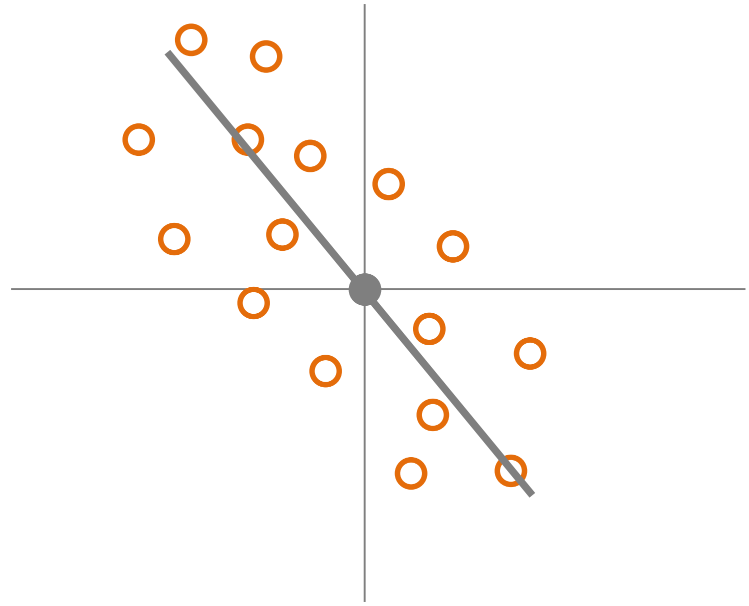
Which of these has the stronger correlation?

STRENGTH OF CORRELATION

CASE A



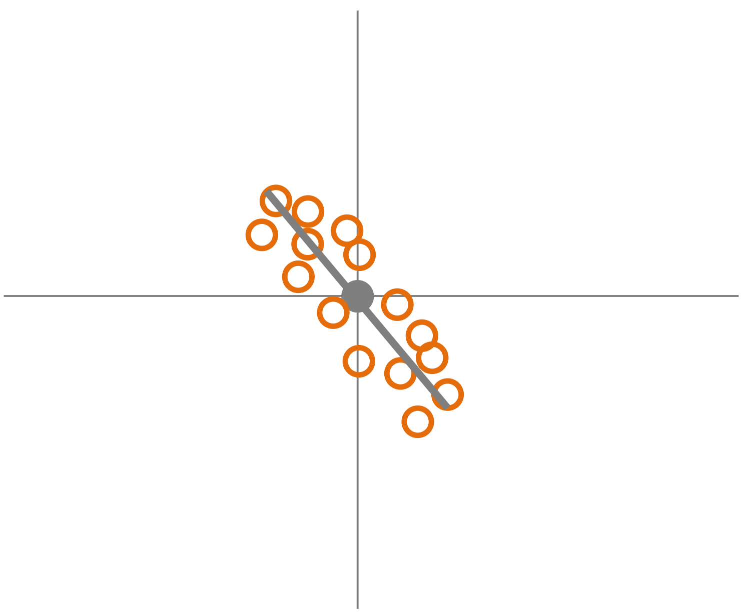
CASE B



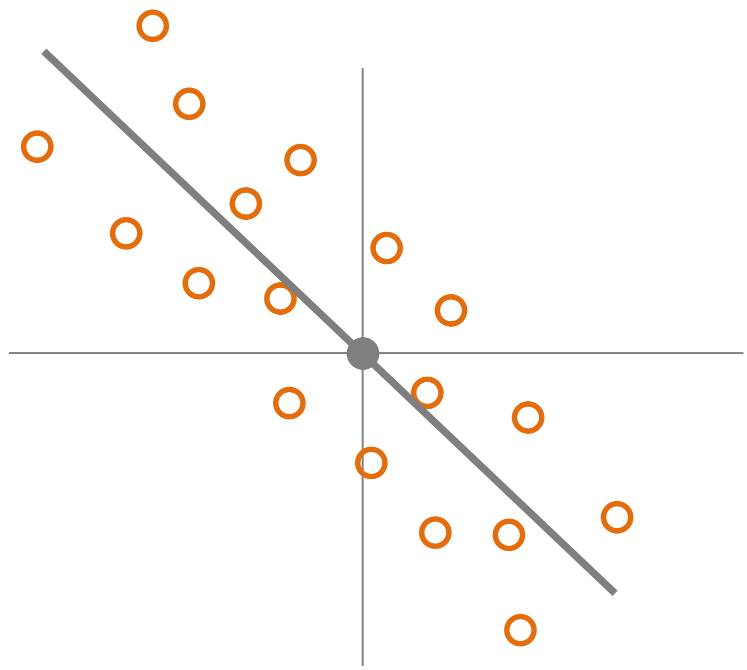
Which of these has the stronger correlation?

Strength of Correlation

Case A



Case B



Which of these has the stronger correlation?

What should be clear in my mind?

1. What is **variance** and **standard deviation**?
2. The difference between **standard deviation** and **standard error**.
3. Definitions of **covariance** and **correlation**.
4. The "intuitive" **regression** formula.