PHS End-of-Summer Camp 2019: Relationships between random variables and regression

Christopher Boyer

August 30, 2019

Harvard University

Review

In your own words, explain to me...

What is a random variable?

In your own words, explain to me...

What is an expected value?

In your own words, explain to me...

What is variance?

Relationships between random variables

Relationships between random variables

One of the primary aims of statistics in the population health sciences is to describe the relationship between two or more random variables, e.g.

- what is the relationship between income and health?
- are people who smoke more likely to develop lung cancer?
- is increased air pollution associated with excess mortality in children?

Covariance

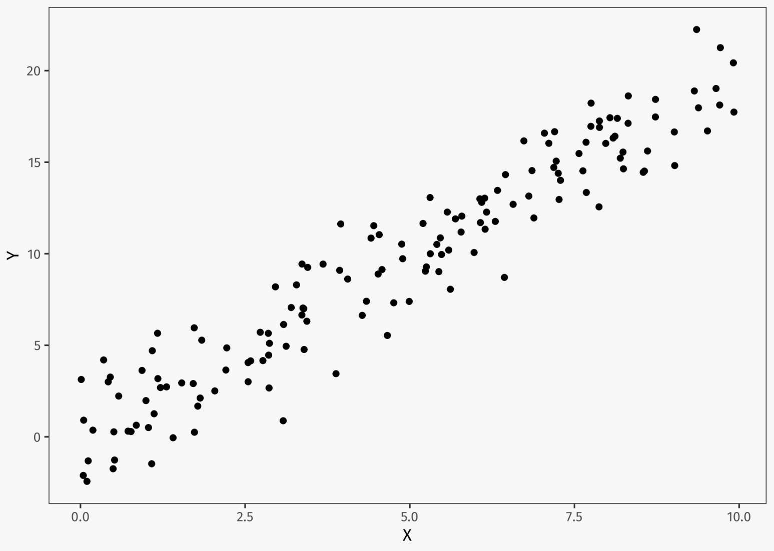
One way we can assess the relationship between two random variables is their covariance:

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

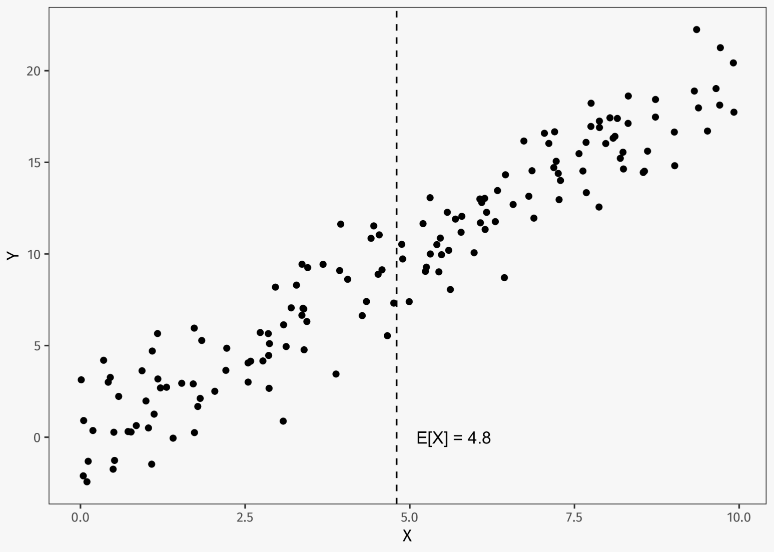
This measures the tendency of two random variables to "move together". If they tend to move in similar directions, the covariance is positive; if they tend to move in opposite directions, it's negative.

In one, sense it is the natural generalization of variance to the bivariate case.

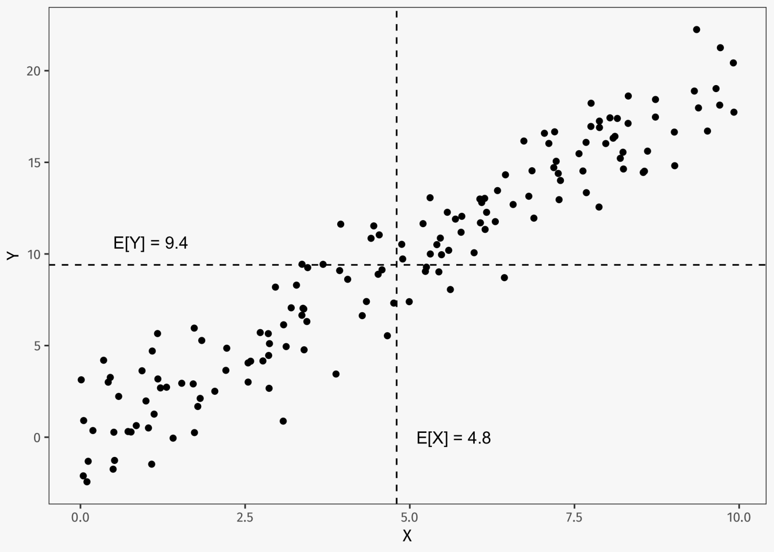
Covariance: intuition



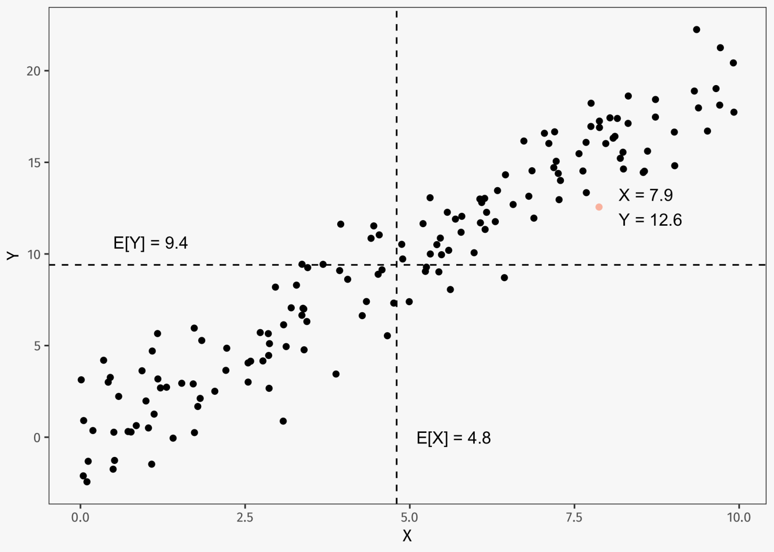
Covariance: intuition



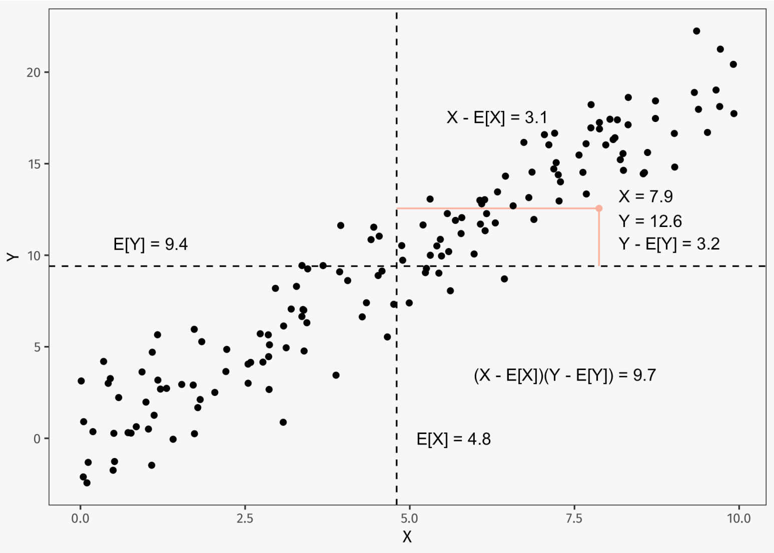
Covariance: intuition



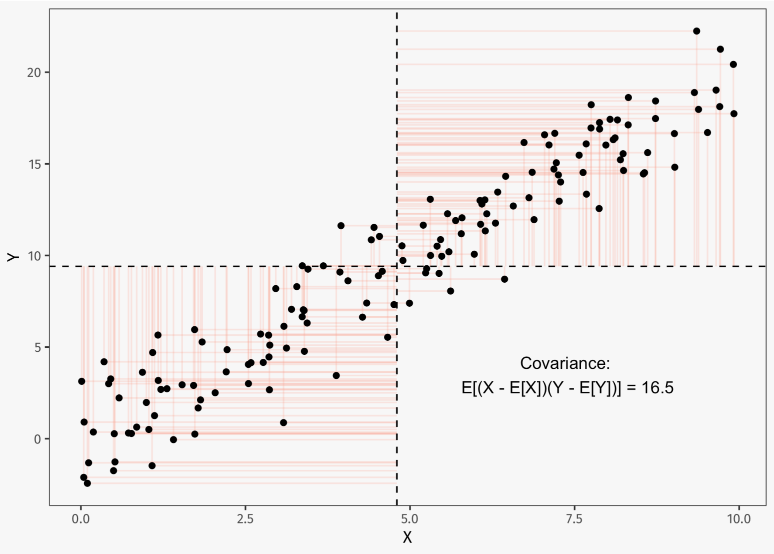
Covariance: intuition



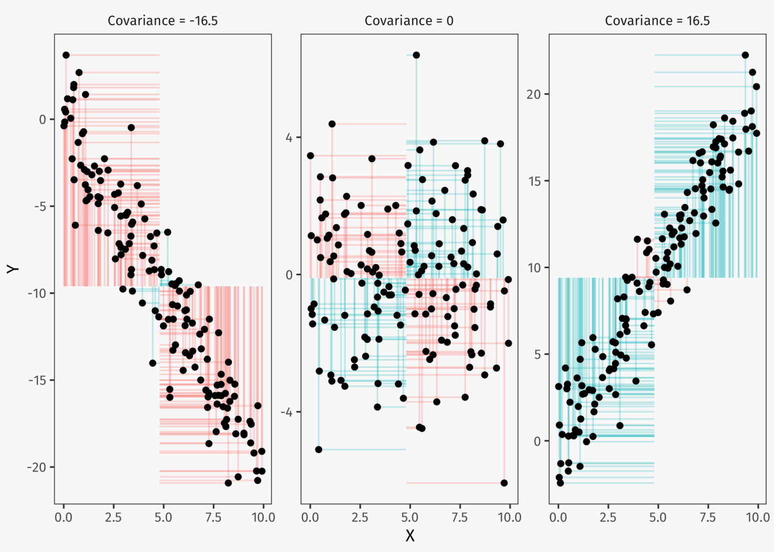
Covariance: intuition



Covariance: intuition



Covariance: intuition



Properties of covariance

Some important properties of the covariance:

- As with expectation and variance, $\mathrm{Cov}[\cdot,\cdot]$ is an operator not a function so $\mathrm{Cov}[X,Y]$ is a constant.
- The covariance is symmetric, i.e. $\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$.
- The covariance of a random variable with itself is just the variance, i.e. $\mathrm{Cov}[X,X] = \mathrm{Var}[X]$.

Sample covariance

Applying the plug-in principle, we can calculate the sample covariance by exchanging expectations for sample means.

$$\widehat{\text{Cov}}[X, Y] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})$$

This estimates the "true" population covariance under the normal regularity conditions.

Sample covariance: example

We observe the following data of course satisfaction ratings and whether or not the instructor brought candy to lecture:

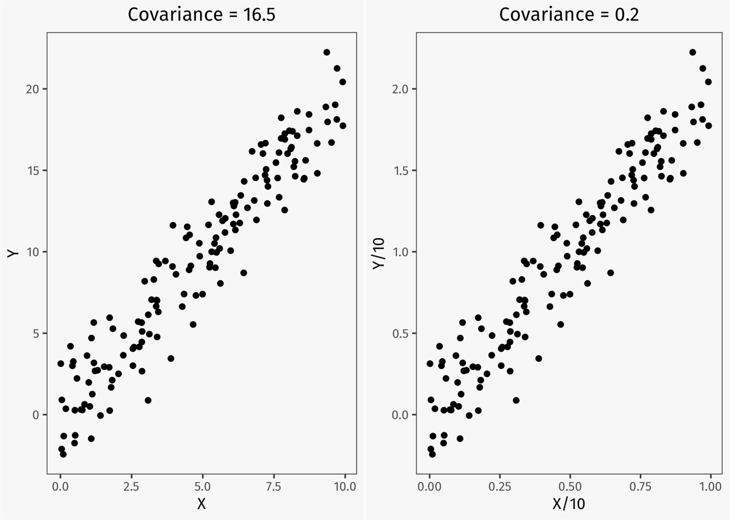
	Satisfaction (Y)			
	1	2	3	4
candy ($X = 1$)	2	5	2	19
no candy ($X = 0$)	32	14	4	22

$$\widehat{\text{Cov}}[X, Y] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z}) = 0.231$$

The fact that this is positive tells us that larger values of Y (higher satisfaction) tend to occur more often with large values of X (lectures with candy).

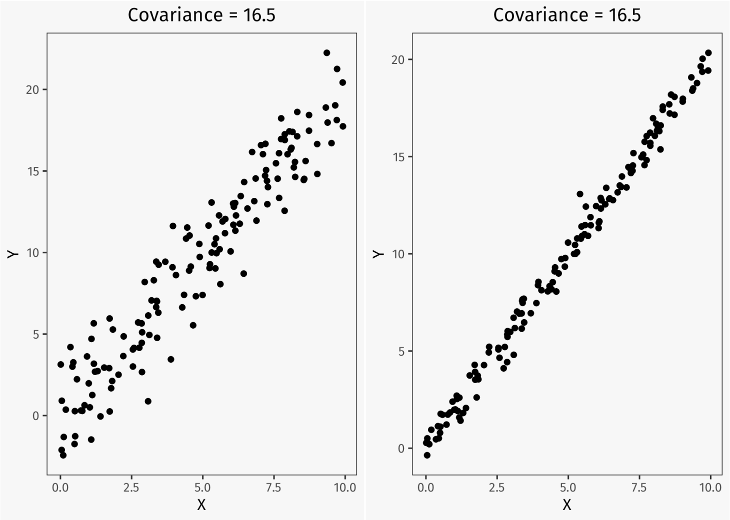
Limitations of covariance

The covariance is sensitive to the scale of the random variables.



Limitations of covariance

The covariance can't tell you about the strength of the relationship between random variables.



Correlation

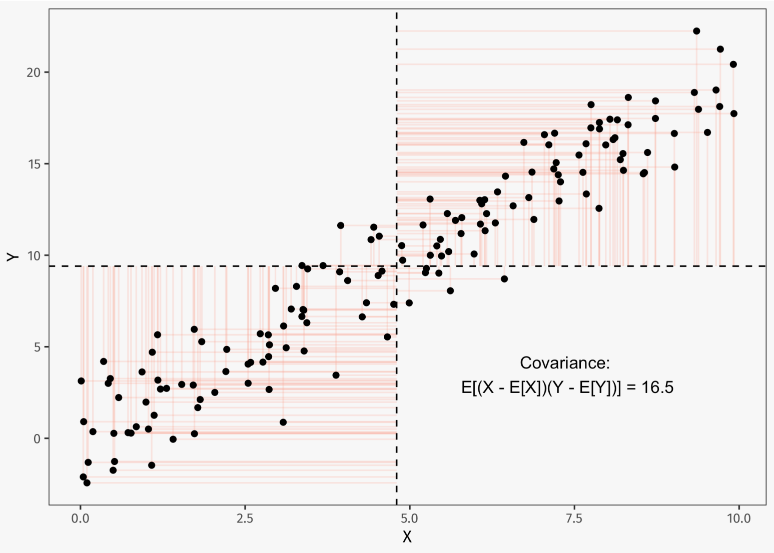
One way we could overcome these limitations is to develop a standard scale for the covariance.

Indeed, if we standardize the covariance by dividing by the product of the standard deviation ($\sigma[X] = \sqrt{\mathrm{Var}[X]}$), we get the correlation, which we often refer to with ρ.

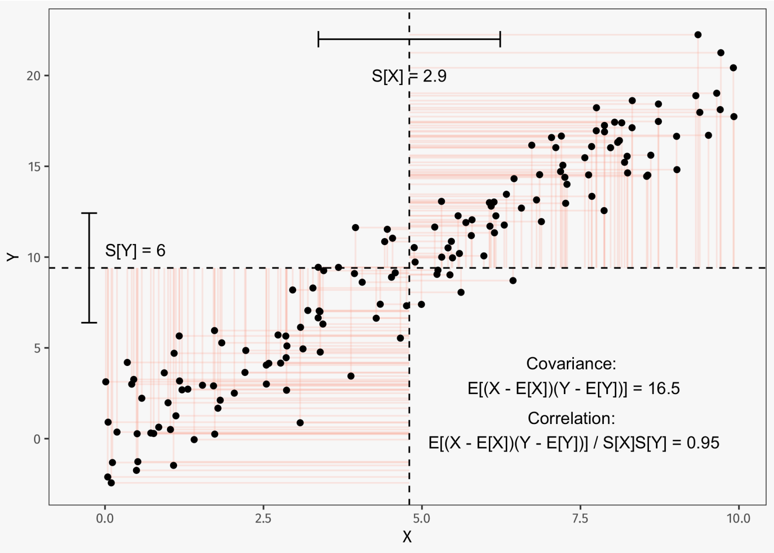
$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma[X]\sigma[Y]}$$

The correlation is another useful summary of the relationship between random variables.

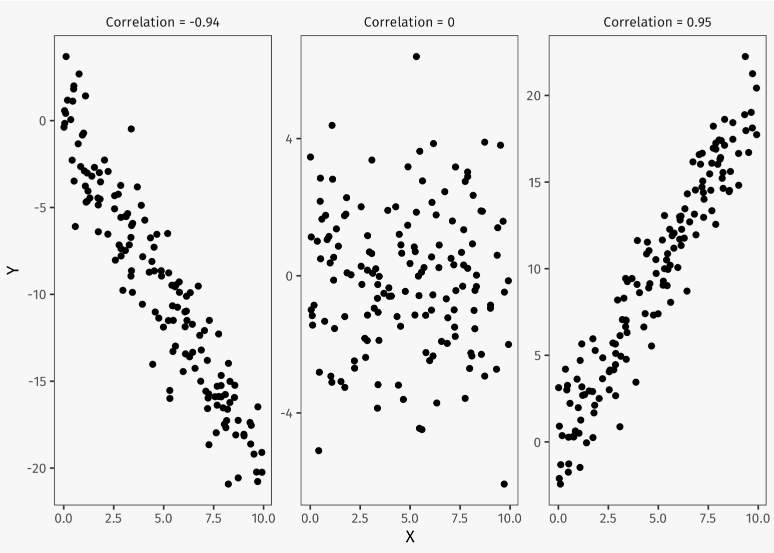
Correlation: intuition



Correlation: intuition

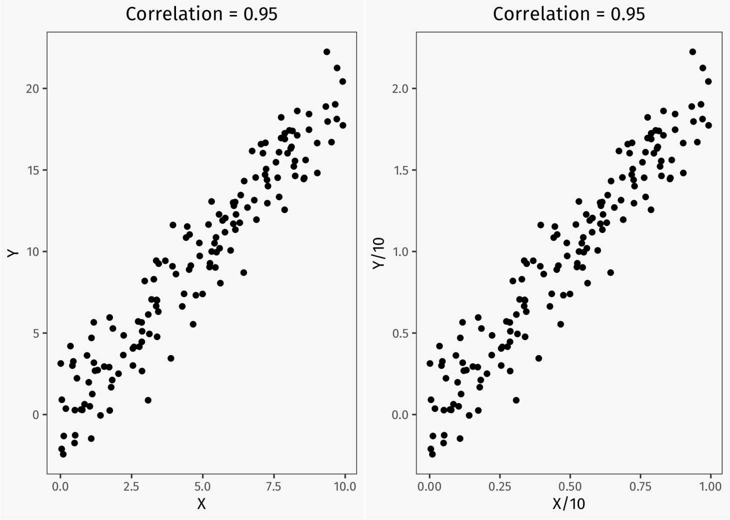


Correlation: intuition



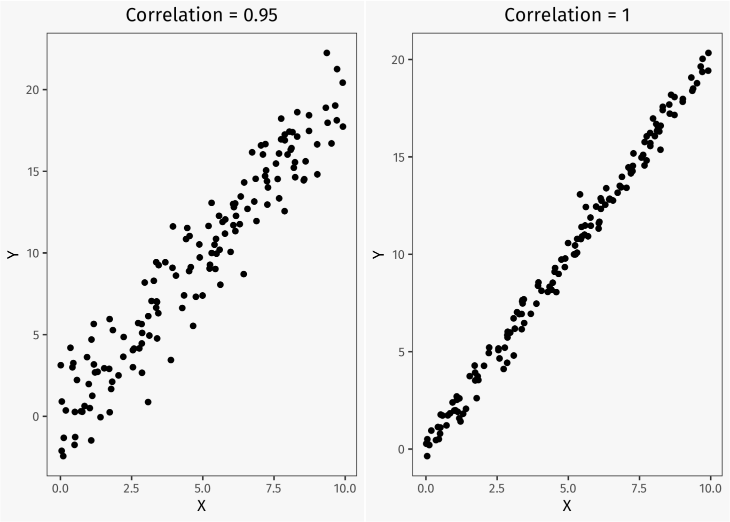
Improvements over covariance

The correlation is NOT sensitive to the scale of the random variables.



Improvements over covariance

The correlation DOES tell you something about the strength of the relationship between random variables.



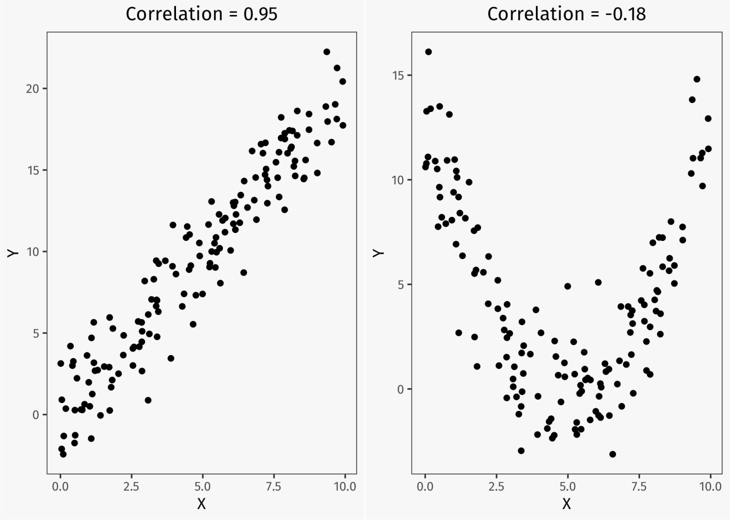
Properties of correlation

Some important properties of the correlation:

- $\rho[\cdot, \cdot]$ is an operator not a function so $\rho[X, Y]$ is a constant.
- The correlation is constrained to be between -1 and 1, i.e. $-1 \leq \rho \leq 1$.
- Like the covariance the correlation is symmetric, i.e. $\rho[X, Y] = \rho[Y, X]$.
- The correlation of a random variable with itself is always one, i.e. $\rho[X, X] = 1$.

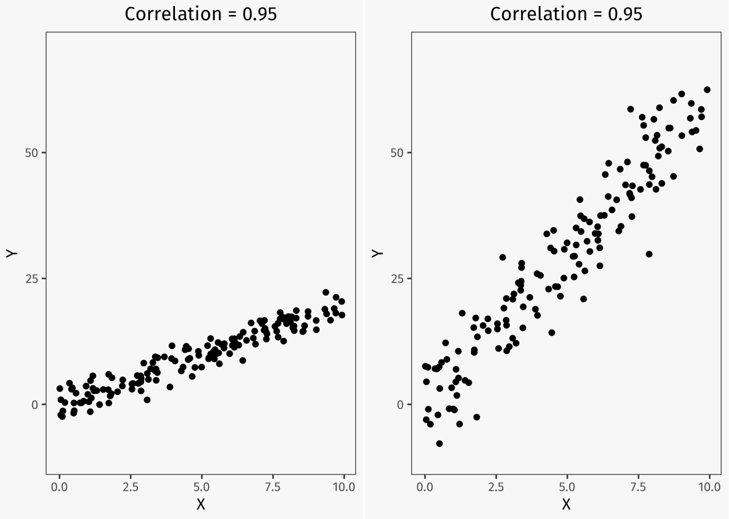
Limitations of correlation

The correlation only tells you about the linear dependence between random variables.



Limitations of correlation

The correlation doesn't tell you how much one random variable changes with the other (i.e. slope).



Aside: Independence

Recall, two random variables, X and Y, are said to be independent if knowing the outcome for one provides no information about the probability of any outcome for the other, i.e. if their distributions do not depend on other.

$$f(x, y) = f(x)f(y)$$

We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent

Independence, correlation, and covariance

Independence, correlation, and covariance are tightly bound concepts.

If X and Y are independent then their correlation and covariance are necessarily zero, i.e. $X \perp\!\!\!\perp Y$ implies:

$$\mathrm{Cov}[X, Y] = 0$$

$$\rho[X, Y] = 0$$

HOWEVER, the converse is not true; a zero correlation or covariance does NOT imply that X and Y are independent (for one just look at previous slide).

Try it yourself!

Open the file `regression-1.R` and complete the exercises

Conditional Expectation

Conditional Expectation

Thus far, we've talked about covariance and correlation and found them both in some sense wanting. Another way we can describe the relationship between random variables is the conditional expectation.

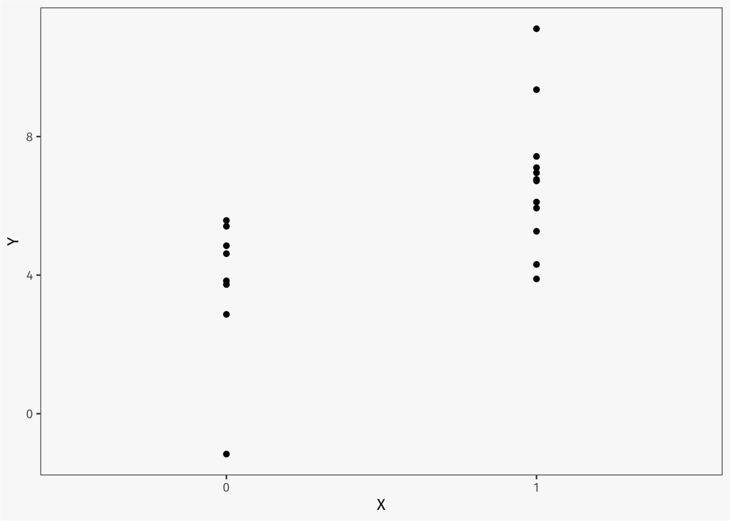
$$\mathrm{E}[Y \mid X = x] = \sum_y y f(y \mid x)$$

$$\mathrm{E}[Y \mid X = x] = \int_{\infty}^{\infty} y f(y \mid x) dy$$

These expressions may look intimidating, but the conditional expectation is just the expectation, or population average, of Y at a pariticular value of X.

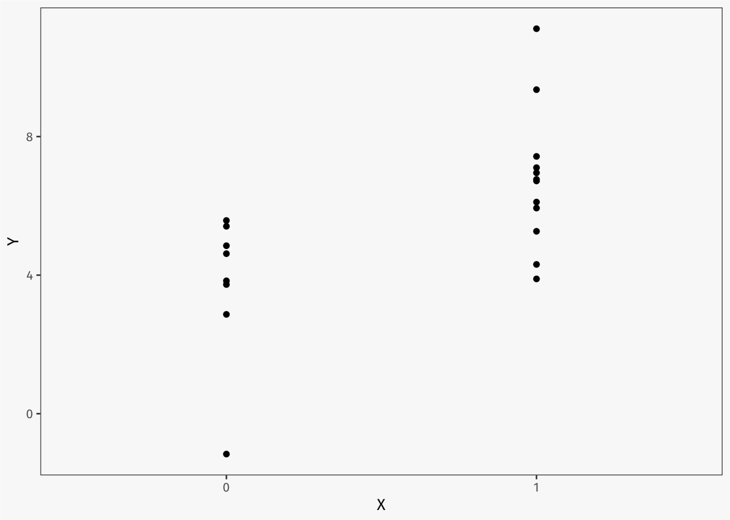
Conditional Expectation

What is your best estimate as to the value of $E[Y \mid X = 1]$?



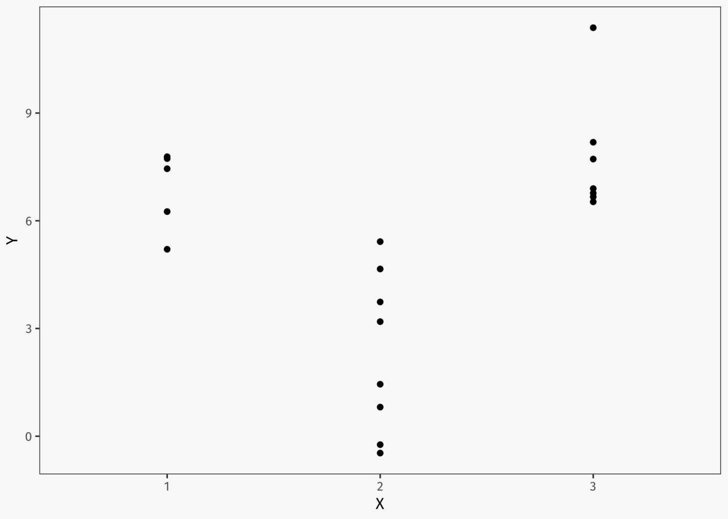
Conditional Expectation

What is your best estimate as to the value of $\mathrm{E}[Y \mid X = 0]$?



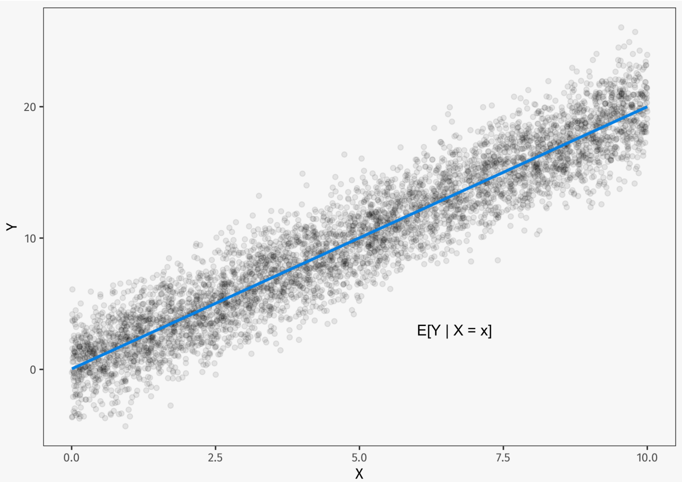
Conditional Expectation

What is your best estimate as to the value of $E[Y \mid X = 2]$?



Conditional Expectation Function

Taking this one step further we can begin to conceive of a conditional expectation function that maps the population average or expectation of Y to each value of X.

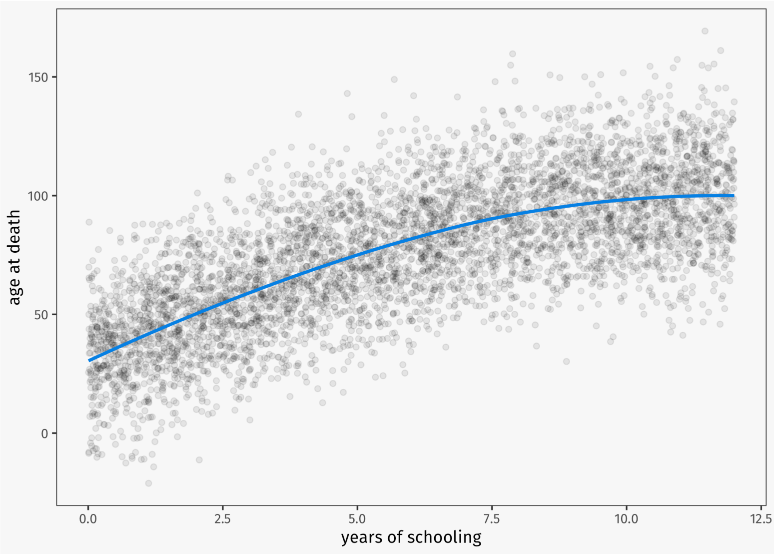


Why CEF is important

In some sense the conditional expectation function is exactly what we've been looking for to describe relationships between random variables in the population health sciences.

- it describes how the mean of one random variable changes with values of another
- it can be of any form (linear/nonlinear, smooth/nonsmooth)
- it is pretty straightforward to understand

Why CEF is important



Aside: CEF for more than 2 variables

You might be wondering why we've spent so much time with just relationships between two variables (i.e. X and Y). Well partially that's because the previous methods, covariance and correlation, are really only suited to bivariate relationships.

However the conditional expectation function shares no such limitations. We can extend the concept to many variable situations, e.g.

$$E[Y \mid X = x, Z = z, W = w]$$

Note that the **,** here implies "AND", e.g. when X is 1 and Z is 2 and W is 3.

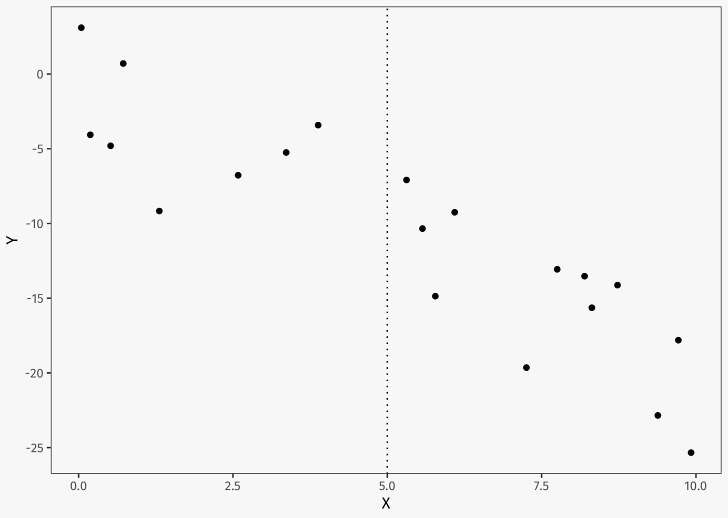
Why can't we just use the CEF

The problem is that the conditional expectation function is fundamentally a population concept.

Unless we have god-like omniscience we generally don't know what the true CEF is, but rather we have to make due with samples to learn/make inferences about what it might look like.

Why can't we just use the CEF

What is the value of the CEF at $X = 5$?



Statistical models

Statistical models

One way we can tackle the problem that we are unable to observe the true CEF, is to make some assumptions about what form it might take.

In essence this is all a model really is: a restriction on the possible values that the CEF might take, i.e.

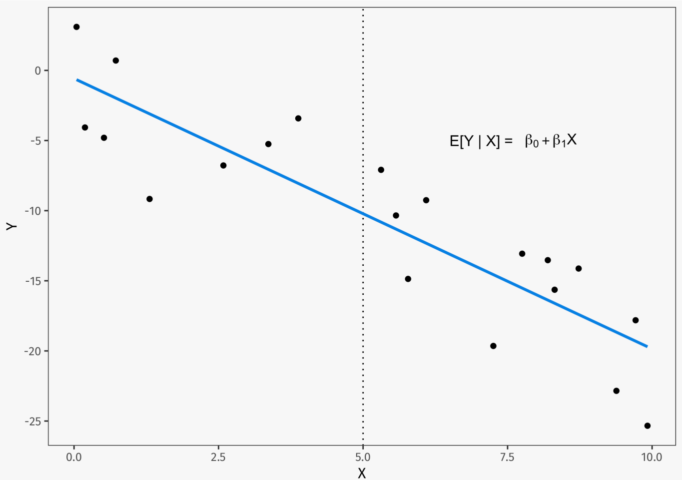
$$E[Y \mid X = x] = \mathit{function}(X)$$

“All models are wrong, but some are useful”

George Box

Statistical models

Returning to our last example, if we assume that the CEF is a linear function of X, what can we say about the likely value of $E[Y \mid X = 5]$?



A simple linear model

Consider the common model:

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

In this model all the values of the conditional mean of Y can be completely determined if we know the values of two parameters β_0 and β_1.

Why does this make sense? Think back to high school geometry.

A simple linear model

Ok but what do β_0 and β_1 represent? Well let's start by considering what happens when we set X to zero.

$$E[Y \mid X = 0] = \beta_0 + \beta_1 \cdot 0 = \boxed{\beta_0}$$

The parameter β_0 is just the value of the conditional mean of Y when X is zero or, in other words, β_0 is the intercept.

A simple linear model

Knowing this we can now also figure out what β_1 represents by using just a little math...

$$E[Y \mid X = 1] = \beta_0 + \beta_1$$

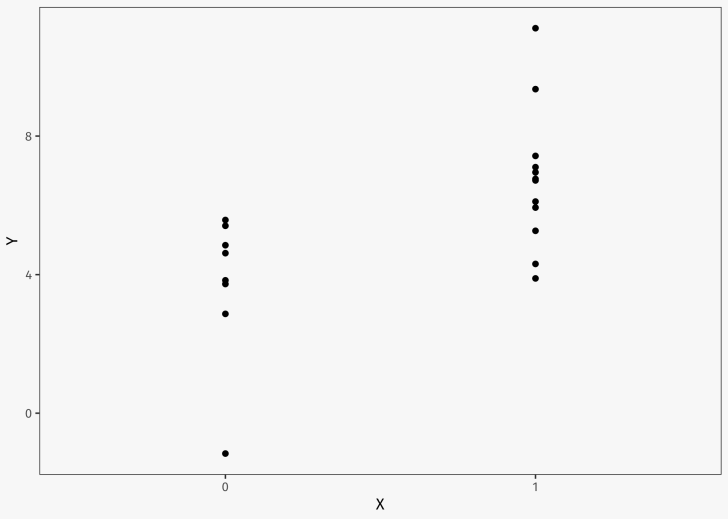
$$E[Y \mid X = 0] = \beta_0$$

$$E[Y \mid X = 1] - E[Y \mid X = 0] = (\beta_0 + \beta_1) - (\beta_0) = \boxed{\beta_1}$$

The parameter β_1 is just the change in the value of the conditional mean of Y for a unit change in X or, in other words, β_1 is the slope of the line.

A single binary predictor

Let's return to the example of a single binary predictor. What assumptions is the model $E[Y \mid X = x] = \beta_0 + \beta_1 X$ imposing?



Saturated models

We call models like the previous one saturated or nonparametric models because they contain a parameter for every possible value of X.

In general these occur when models have only discrete predictor variables (e.g. binary and categorical predictors) and include all possible interaction terms.

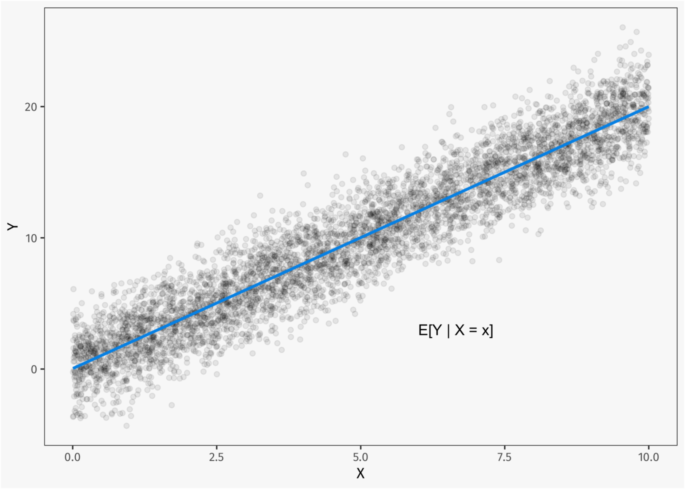
More flexible models

What if we wanted to make our model a bit more flexible? For instance what if we believed the true CEF might follow a quadratic form?

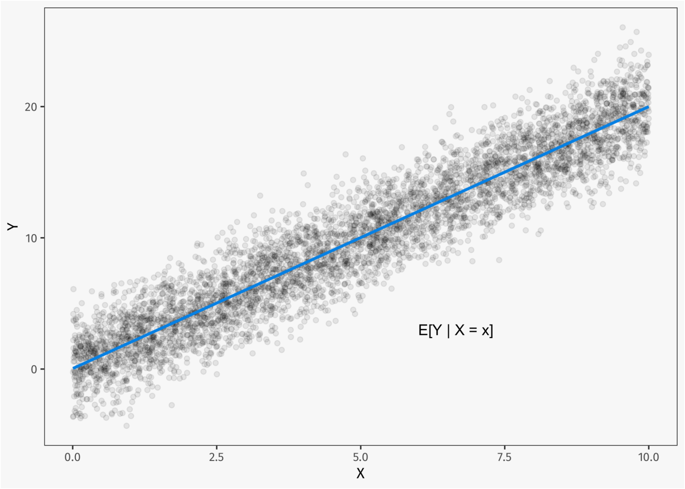
$$E[Y \mid X = x] = \beta_0 + \beta_1 X + \beta_2 X^2$$

Voila! Let's just add another parameter β_2 to capture this possible quadratic relationship.

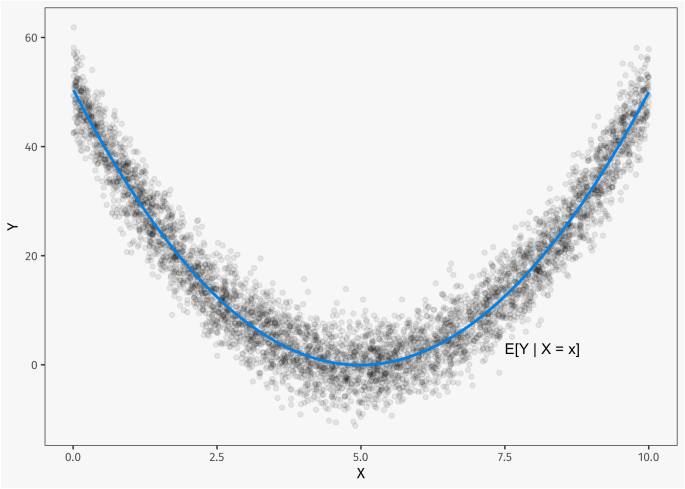
Side note: is this still a linear model?



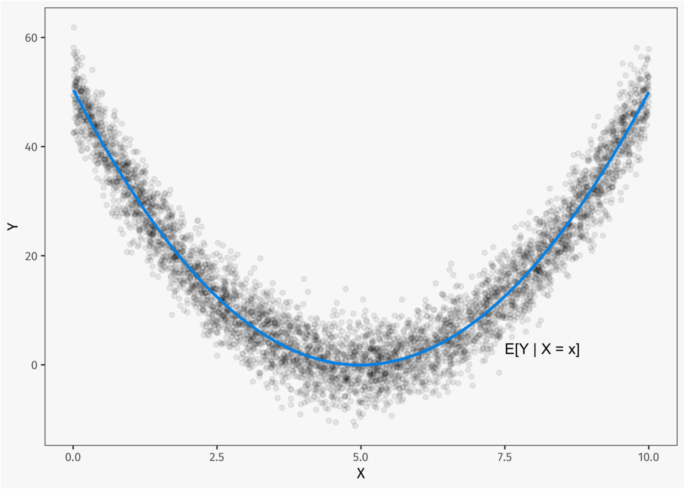
If the graph above represents the true CEF will the model $E[Y \mid X = x] = \beta_0 + \beta_1 X$ correctly estimate the CEF? A. Yes B. No



If the graph above represents the true CEF will the model $E[Y \mid X = x] = \beta_0 + \beta_1 X + \beta_2 X^2$ correctly estimate the CEF? A. Yes B. No



If the graph above represents the true CEF will the model $E[Y \mid X = x] = \beta_0 + \beta_1 X$ correctly estimate the CEF? A. Yes B. No



If the graph above represents the true CEF will the model $E[Y \mid X = x] = \beta_0 + \beta_1 X + \beta_2 X^2$ correctly estimate the CEF? A. Yes B. No

Which model imposes more restrictions on (makes more assumptions about) the CEF?

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

B. $E[Y \mid X = x] = \beta_0 + \beta_1 X + \beta_2 X^2$

Regression

Estimating parameters of statistical models

A logical question you might have had in the previous section is how do I actually get numerical values for the parameters (i.e. the βs) in my statistical model?

Regression is a tool for estimating the parameters of a statistical model. In that vein you can think of it just like any other recipe like the sample mean.

An important by product of this is that regression tells us how to get the coefficients for our models, but it tells us nothing about whether those models are right.

Reminder about estimation terminology

The estimand is the population quantity of interest whose true value you want to know.

$$\mathrm{E}[Y \mid X = x] = \beta_0 + \beta_1 X$$

An estimator is a method for estimating the estimand.

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

An estimate is a numerical estimate of the estimand that results from the use of a particular estimator.

$$\widehat{\beta_1} = 32$$

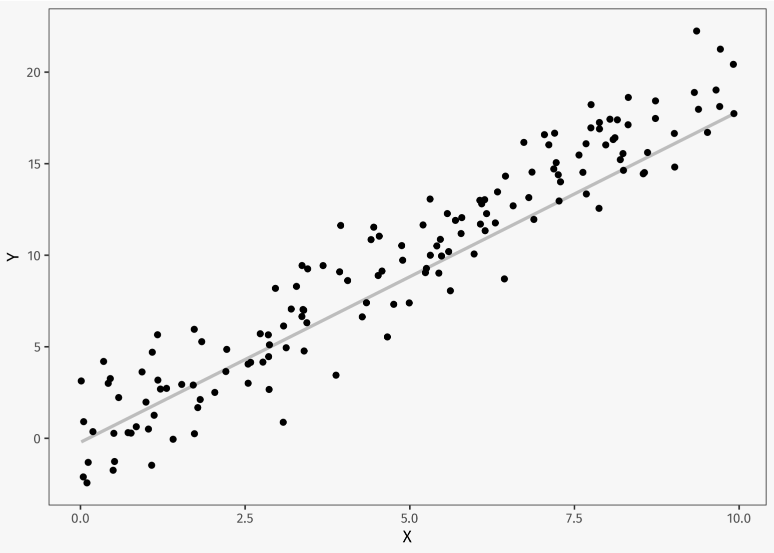
Ordinary least squares

A common method for estimating the parameters of a statistical model is to use ordinary least squares (OLS).

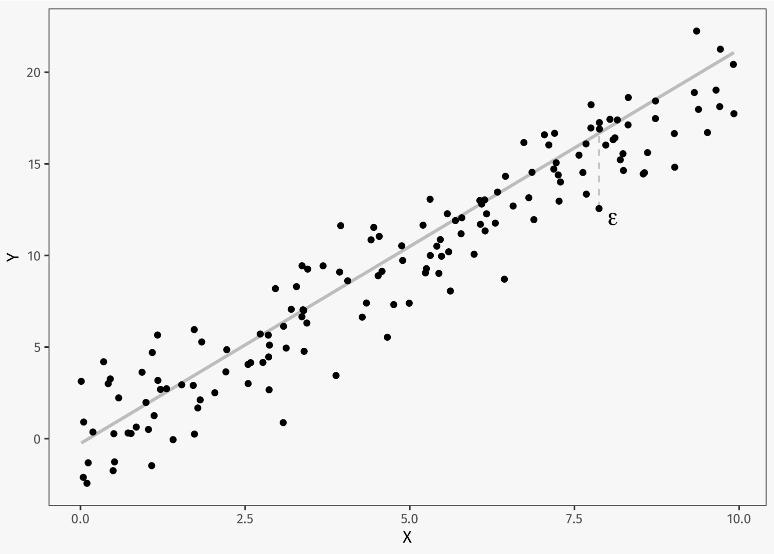
Ordinary least squares attempts to find the values of the parameters (i.e. the βs) such that the sum of squared deviations from the conditional mean are minimized.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{E[Y \mid X]})^2$$

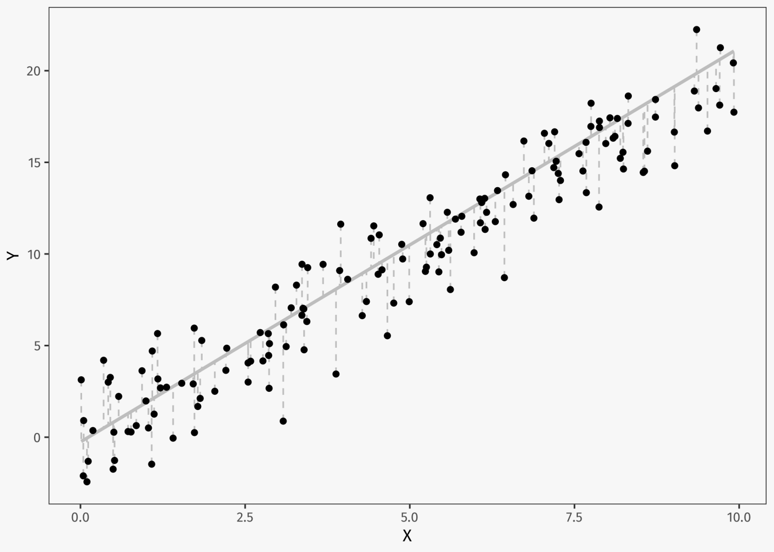
OLS graphical intuition



OLS graphical intuition



OLS graphical intuition



The OLS recipe

It turns out we can find the values of the βs that minimize the sum of squares using a bit of calculus. (Hint: it involves derivatives; for those interested in the details for how this done see me later)

Perhaps a somewhat surprising result is that the estimate for the slope e.g. β_1 is

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Which is also the sample covariance of X and Y over the variance of X!

$$\widehat{\beta_1} = \frac{\widehat{Cov[X, Y]}}{\widehat{Var[X]}}$$

OLS is BLUE

You may be wondering: why all this concern about minimizing the sum of squares?

A surprising result that we'll discuss more in the course is that minimizing the sum of squares turns out to be the best linear estimator you can come up with.

By best we mean the estimator with no bias that has the lowest variance (i.e. is the most precise). You'll sometimes hear statisticians refer to estimators that achieve this as best linear unbiased estimators (BLUE).

Try it yourself!

Open the file `regression-2.R` and complete the exercises

I run a regression of self reported happiness on an indicator of whether students attend a statistics lecture on a perfectly sunny Friday and find to my horror that students who attend are 50 points less happy than those that do not (β = -50). Does this mean that attending a statistics lecture on a perfectly sunny Friday causes students to be less happy?

A. Yes

B. No

C. I don't care just get me out of here

What if I told you that the data used in this study come from a large randomized trial in which on a given sunny Friday, students were randomly assigned to either attend a lecture or not attend a lecture. In this case would you say the results imply that attending a statistics lecture on a perfectly sunny Friday causes students to be less happy?

A. Yes

B. No

C. I still don't care... did you say it's sunny outside?

Connection to causal inference

A key insight here is that estimates obtained via regression provide a numeric estimate of how the mean of Y changes with X, but says NOTHING about the nature of that estimate. Therefore we ofter refer to these estimates as associations.

Additional inferences about whether the estimate is likely to be of a causal effect require additional assumptions about the data generation process that gave rise to the observations under study.

Or put more simply, regression is dumb!

Last word