PHS Summer Camp 2020

Relationships and Regression

Matthew Lee

Harvard University

https://phs-summr2020.netlify.app/regressionslides/slides.html#1

submit panel questions

Please use the link below to submit anonymous questions for the student panel this afternoon, no later than 1pm EST today if possible!

Submit a question here!

review

So far this week, we've discussed the idea of random variables and their properties including:

- Expected values $\rightarrow \mathbb{E}(X)$
- Variance $\rightarrow \mathbf{Var}(X)$
- Probability distributions, like the *Binomial distribution* for a discrete variable or *Normal distribution* for one that's continuous.

But why do we actually care about these things? Why do we even need to worry about crazy expressions like the one below?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

i.e. the PDF of a normal distribution, **please** don't actually memorize this -- it's on Wikipedia

a data generating world

Ultimately, we're interested in these concepts because we can think of these distributions of random variables as an approximation of the world we live in -- and of the processes we wish to understand. In Public Health we might think of:

- The number of events that occur in a given period of time (such as the number of hospitalizations per week) as a Poisson process.[1]
- Whether or not someone experiences pre-term birth as a Binomial/Bernoulli process.[2]

a data generating world

Ultimately, we're interested in these concepts because we can think of these distributions of random variables as an approximation of the world we live in -- and of the processes we wish to understand. In Public Health we might think of:

- The number of events that occur in a given period of time (such as the number of hospitalizations per week) as a Poisson process.[1]
- Whether or not someone experiences pre-term birth as a Binomial/Bernoulli process.[2]
- From your favorite field of interest, what's another example of a random variable and what sort of distribution could it be represented by?

a data generating world

More often than not, we as public health researchers want to describe the relationship between *two or more* random variables. For example:

- 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?
- Does exposure to a sugary beverages tax decrease risk of obesity?

the plan for today

1. Introduce the FÜN Study
2. Relationships between variables
3. Intro to linear regression
4. Wrapping up + conclusions

the plan for today

We will use `R` to explore different ways to assess relationships between variables. Interactive exercises can be found on the website below, but feel free to work on your own computer if you'd like.

https://phs-summr2020.netlify.app/

This material is meant to introduce or refresh your memory of certain concepts, but it is totally ok if you don't understand everything: we will be returning to much of this over the course of the Fall semester.

the plan for today

We will use `R` to explore different ways to assess relationships between variables. Interactive exercises can be found on the website below, but feel free to work on your own computer if you'd like.

https://phs-summr2020.netlify.app/

This material is meant to introduce or refresh your memory of certain concepts, but it is totally ok if you don't understand everything: we will be returning to much of this over the course of the Fall semester.

Questions: If you have a question, feel free to type it into the Zoom chatbox and we'll return to it at various points in the presentation. You can always email me (mlee8@g.harvard.edu) or any of the other TF's if you think of something later on.

the plan for today

1. Introduce the FÜN Study
2. Relationships between variables
3. Intro to linear regression
4. Wrapping up + conclusions

the FÜN study

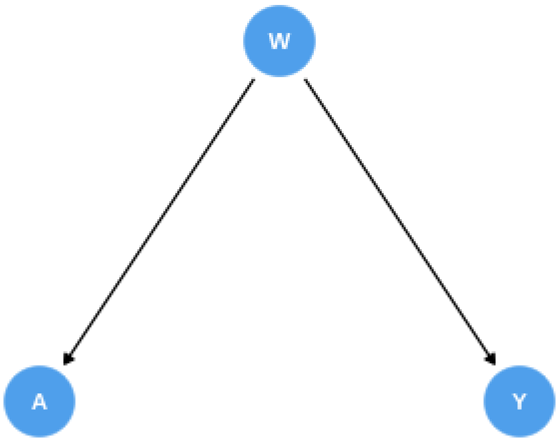
To build our intuition of ideas, let's look at a silly made-up dataset from the Follow-up of Über-cool StudeNts (FUN). As part of the study, 10,000 doctoral students pursuing health-related degrees were asked to provide information on:

- `W`: Whether the student is currently in their "dissertating" phase
- `A.con`: # hours student slept last night
- `A.bin`: Whether student slept at least 8 hours (yes/no)
- `Y.con`: # times student used a food delivery service (FDS) last week
- `Y.bin`: Whether FDS comprised ≥50% of the week's meals (yes/no)

A note: our outcome `Y.con` is continuous, rather than discrete, to take into account fractions of meals a student ate (e.g. snacks, second breakfasts)

the FÜN study

Suppose that we actually know the **true** relationships between these variables and how they were generated in the population. Specifically:



That is, student sleep hours and FDS use is affected by whether a student is writing their dissertation, but student sleep itself **does not** cause FDS use. We'll come back to this when we talk about **confounding** and regression.

* Full details on how this data were simulated can be viewed here.

the plan for today

1. Introduce the FÜN Study
2. Relationships between variables
3. Intro to linear regression
4. Wrapping up + conclusions

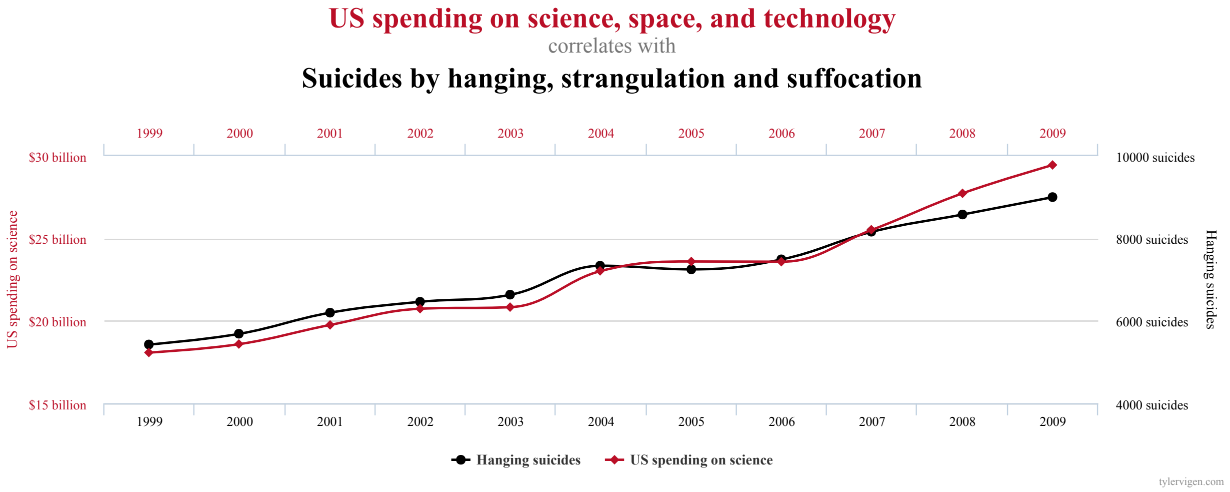
relationships between r.v.'s

By now you've probably heard the phrase:

> *correlation does not imply causation" (or something similar).*

But what do we mean by correlation in the first place? And why doesn't it imply causation?

When two variables are correlated, we are trying to get at this idea that two variables are related. Let's look at how to *quantify* this relationship.



the simple 2x2

When we have two variables that are both Bernoulli distributed (i.e. they take on values of `0` or `1` only), the easiest thing we can do is draw up a 2x2 contingency table. Going back to our FUN study example, we can count how many students (recall your set notation!):

1. Got ≥8 hours of sleep and used FDS ≥50% of the week
   - (`A.bin` = 1 ∩ `Y.bin` = 1)
2. Got ≥8 hours of sleep and did not use FDS 50% of the week
   - (`A.bin` = 1 ∩ `Y.bin` = 0)
3. Got <8 hours of sleep and did not use FDS 50% of the week
   - (`A.bin` = 0 ∩ `Y.bin` = 1)
4. Got <8 hours of sleep and used FDS ≥50% of the week
   - (`A.bin` = 0 ∩ `Y.bin` = 0)

the simple 2x2

Thankfully, we can do this easily in `R`, rather than going through every row of the data and tallying things up

```
xtabs(~Y.bin + A.bin, data = big.FUN)
```

```
##      A.bin
## Y.bin    0    1
##     0 7022 1797
##     1 1151   30
```

We can use this information to calculate the prevalence ratio, comparing the prevalence of ≥50% FDS use between those who got 8 hours of sleep to those who did not:

$$PR = \frac{P(\texttt{Y.bin} = 1 \mid \texttt{A.bin} = 1)}{P(\texttt{Y.bin} = 1 \mid \texttt{A.bin} = 0)}$$

$$PR = \frac{30}{1827} \bigg/ \frac{1151}{8173} = 0.117$$

the simple 2x2

$$PR = \frac{30}{1827} \Big/ \frac{1151}{8173} = 0.117$$

What does this mean?

> This suggests that the proportion of students who used FDS for ≥50% of their weekly meals among those who got at least 8 hours of sleep was 88.3% lower than the proportion of students who used FDS for ≥50% of their weekly meals among those who got less than 8 hours of sleep.

In other words, those who get at least 8 hours of sleep appear to be much less likely to use food delivery services for more than half of their weekly meals.

the simple 2x2

Other statistics you might be familiar with that are often used to assess relationships between two Bernoulli random variables are:

- Odds ratios
- Risk ratios
- Hazard ratios
- Risk differences

Each has its own interpretation, you will learn more about each one in PHS 2000A and EPI 201/202!

But what about continuous variables?

i.e. you still haven't told me what correlation is yet

covariance

When we have two continuous random variables X and Y, one statistic we can use to assess their relationship is their covariance:

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

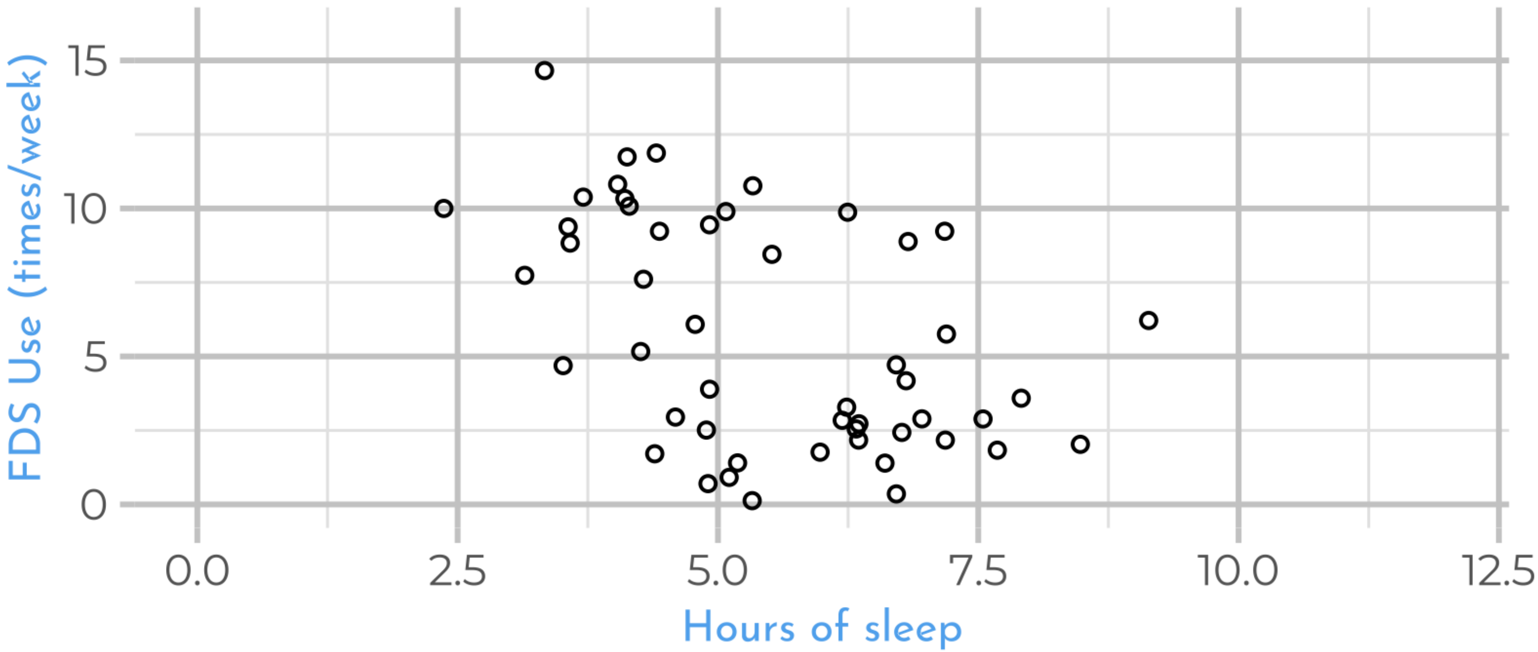
This measures the tendency of two random variables to“move together”. If they tend to move in similar directions, the covariance is positive. If they tend to move in opposite directions, it’s negative.

In other words, the covariance answers the multi-part question: How variable is X? How variable is Y? Does variation in X increase as variation in Y increases? Is X more variable when Y is more variable?

covariance

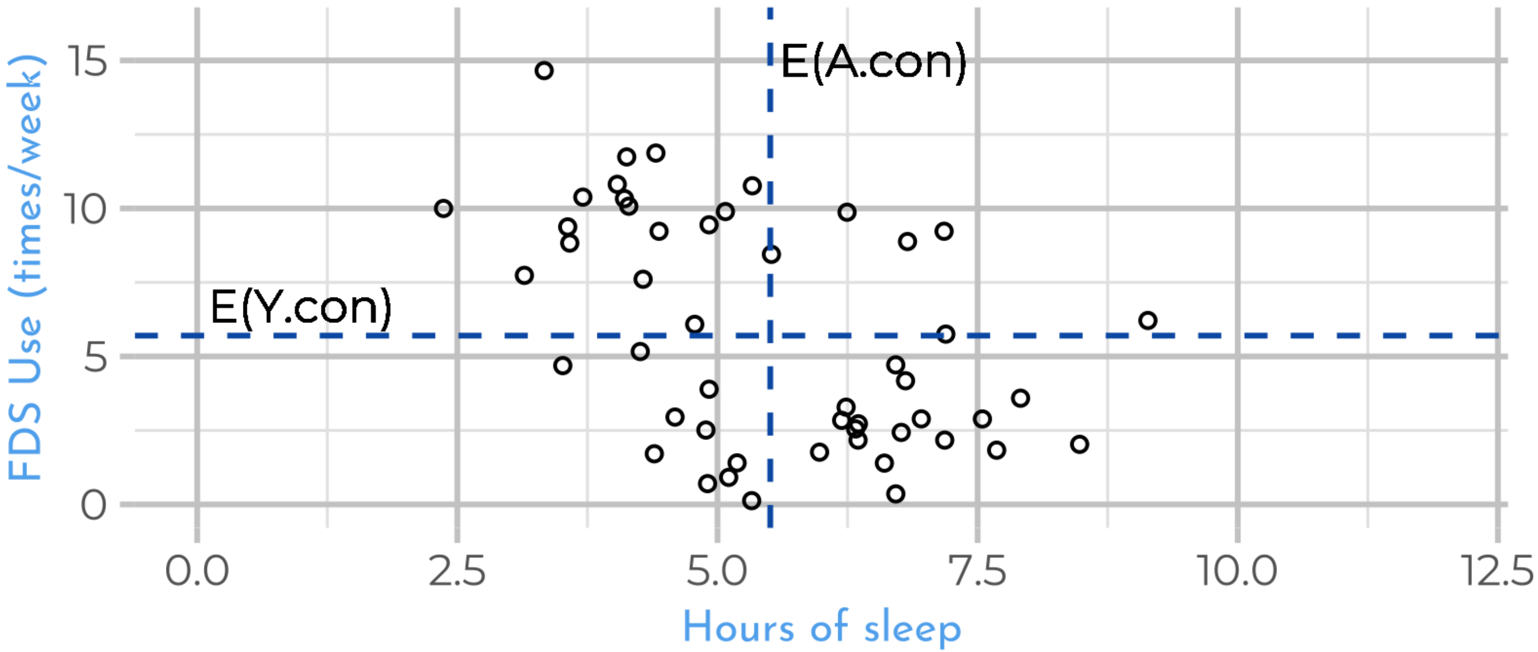
Another way to understand what the covariance represents is with a plot. Returning to our FUN study example, let's examine the relationship between the hours slept last night (`A.con`) and the number of times food delivery services were used that week (`Y.con`), both as continuous variables.

covariance



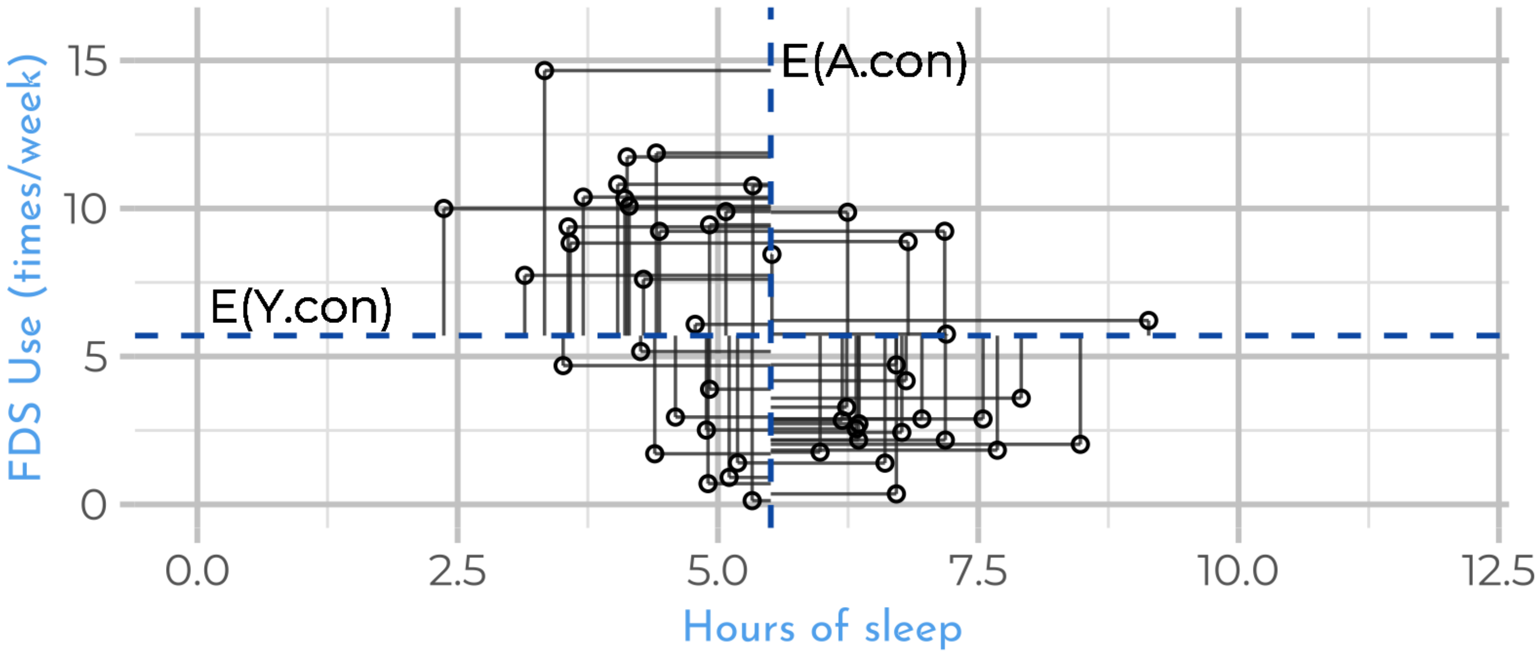
We'll start by simply looking at a scatter plot with sleep hours on the x-axis and FDS use on the y-axis. Here, we've taken a small random sample of 50 students so we can see what's going on more clearly.

covariance



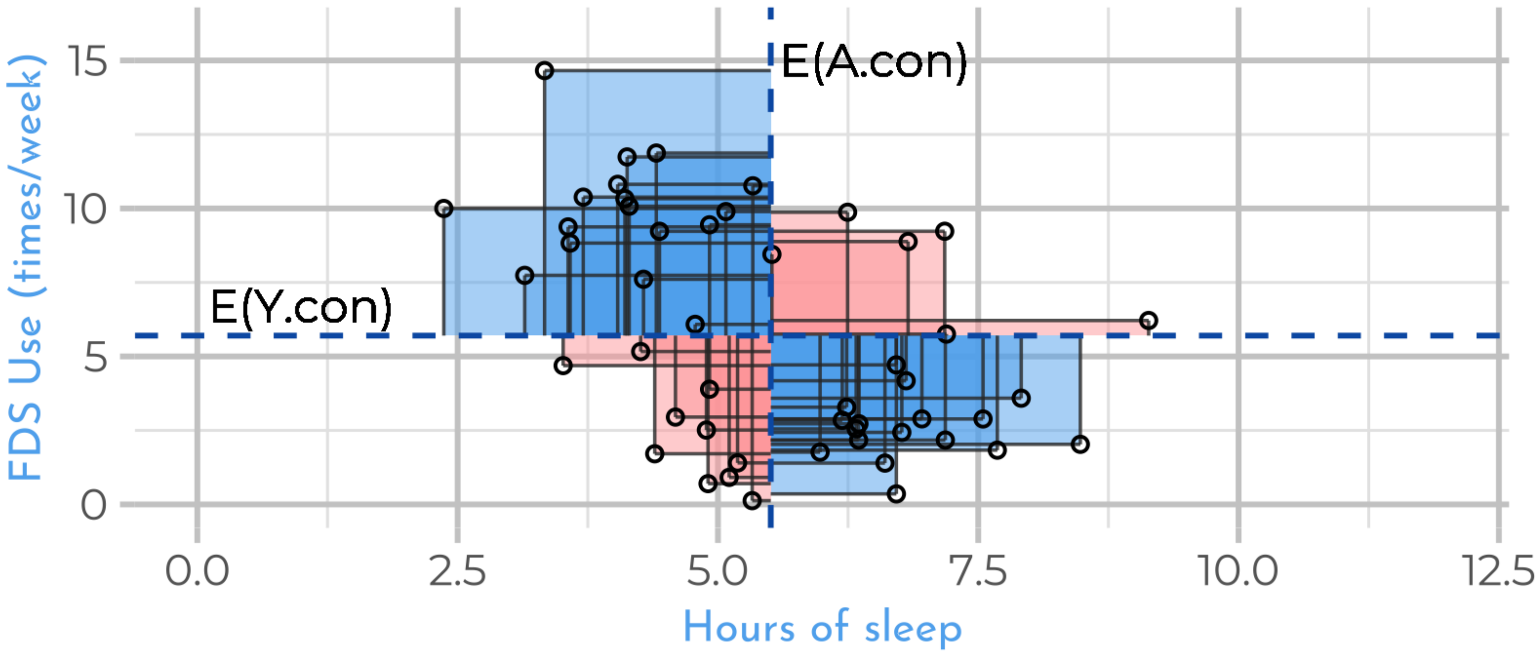
Now we've added dashed lines representing the mean hours of sleep and the mean FDS use across these 50 students.

covariance



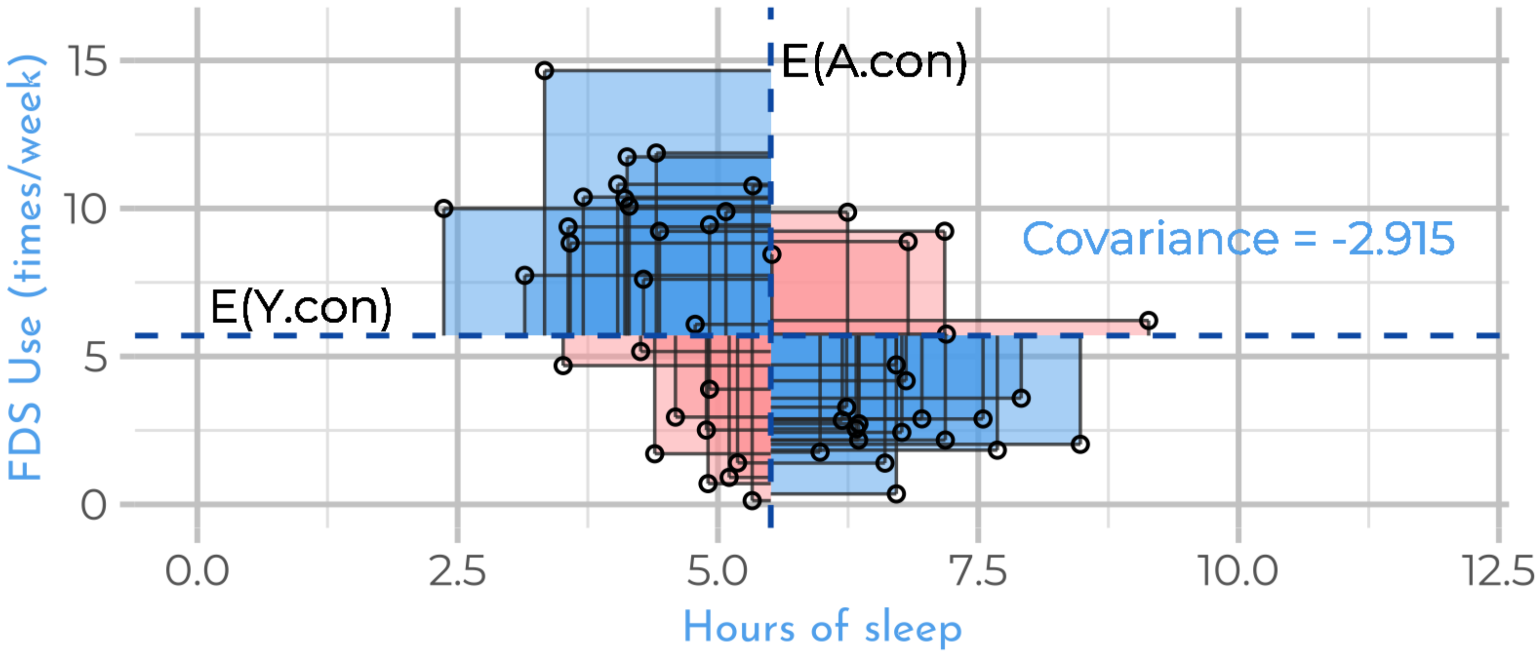
If we draw vertical and horizontal lines between each point and these dashed mean lines, we get a series of rectangles where each rectangle's height is (Y.con - E(Y.con)) and each rectangle's width is (A.con - E(A.con)).

covariance



Multiplying these together, we'll get the area of each rectangle, that is: (Y.con - E(Y.con)) (A.con - E(A.con)). Some rectangles will have negative areas (blue) and others will have positive areas (pink).

covariance



Once we add all these areas up and divide by the number of rectangles (i.e. obtain the mean of the areas), we get the quantity E[(Y.con - E(Y.con)) (A.con - E(A.con))], which is (suprise), the expression we saw for the population covariance!

covariance

Question: Imagine you have 1 million observations of X and 1 million observations of Y, but all values of X are the same and all values of Y are the same. What's the covariance between X and Y?

covariance

Question: Imagine you have 1 million observations of X and 1 million observations of Y, but all values of X are the same and all values of Y are the same. What's the covariance between X and Y?

i.e. We can't study the relationship between two variables when either variable doesn't vary (or in practice varies very little). If we want to design a study to look at this relationship, we need to keep this in mind!

covariance --> correlation

Why would we use the covariance to quantify relationships?

covariance --> correlation

Why would we use the covariance to quantify relationships?

- $\mathbf{Cov}(X, Y)$ is a constant
- $\mathbf{Cov}(X, Y)$ is symmetric, so $\mathbf{Cov}(X, Y) = \mathbf{Cov}(Y, X)$
- $\mathbf{Cov}(X, X) = \mathbf{Var}(X)$

covariance --> correlation

Why would we use the covariance to quantify relationships?

- $\mathbf{Cov}(X, Y)$ is a constant
- $\mathbf{Cov}(X, Y)$ is symmetric, so $\mathbf{Cov}(X, Y) = \mathbf{Cov}(Y, X)$
- $\mathbf{Cov}(X, X) = \mathbf{Var}(X)$

Why wouldn't we use the covariance to quantify relationship?

covariance --> correlation

Why would we use the covariance to quantify relationships?

- $\mathbf{Cov}(X, Y)$ is a constant
- $\mathbf{Cov}(X, Y)$ is symmetric, so $\mathbf{Cov}(X, Y) = \mathbf{Cov}(Y, X)$
- $\mathbf{Cov}(X, X) = \mathbf{Var}(X)$

Why wouldn't we use the covariance to quantify relationship?

- $\mathbf{Cov}(X, Y)$ is sensitive to the scale of the random variables (e.g. think transformations of age or time).
- Therefore, it doesn't really provide us with useful information on the strength of relationships -- is the covariance large because the relationship is strong or because of the scale of your variables?
- $\mathbf{Cov}(X, Y)$ isn't all that easily interpreted!

correlation

Let's fix this scaling issue of the covariance by dividing it by the standard deviations of our random variables. This is called the correlation!

$$\rho_{X,Y} = \frac{\mathrm{Cov}(\mathrm{X},\mathrm{Y})}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

correlation

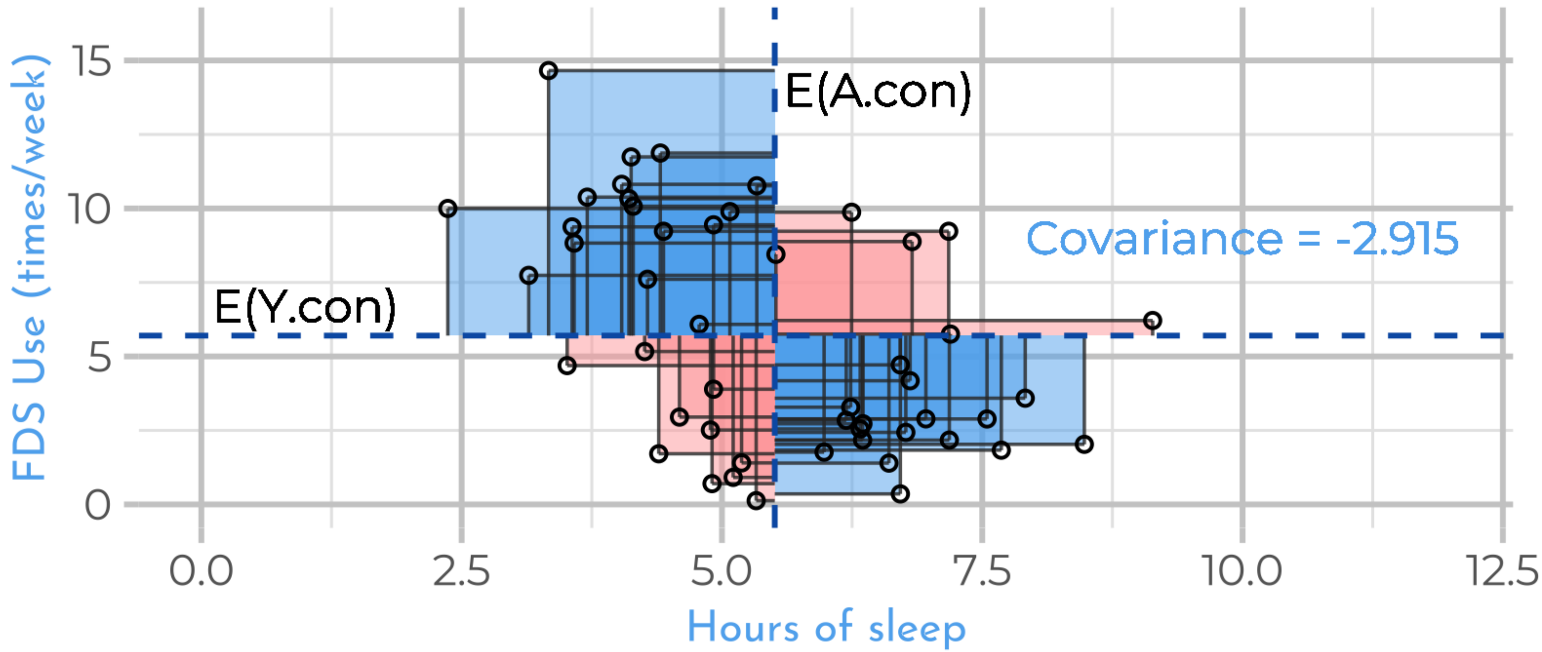
Let's fix this scaling issue of the covariance by dividing it by the standard deviations of our random variables. This is called the correlation!

$$\rho_{X,Y} = \frac{\mathrm{Cov(X,Y)}}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Unlike the covariance, the correlation ρ is:

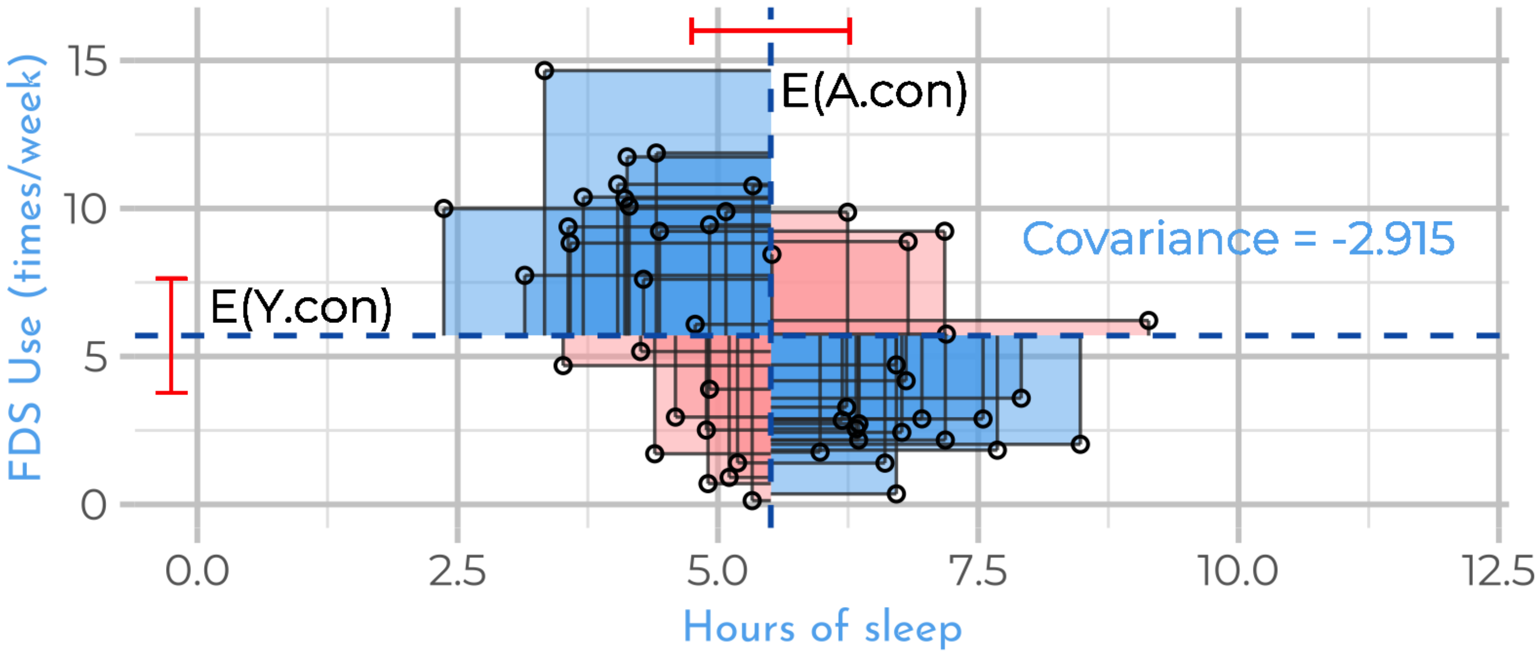
- Not sensitive to scale, and is bounded between -1 and 1
- Does tell us about the strength of the relationship
- More intuitive, the correlation between a r.v. with itself is $\rho_{X,X} = 1$

correlation



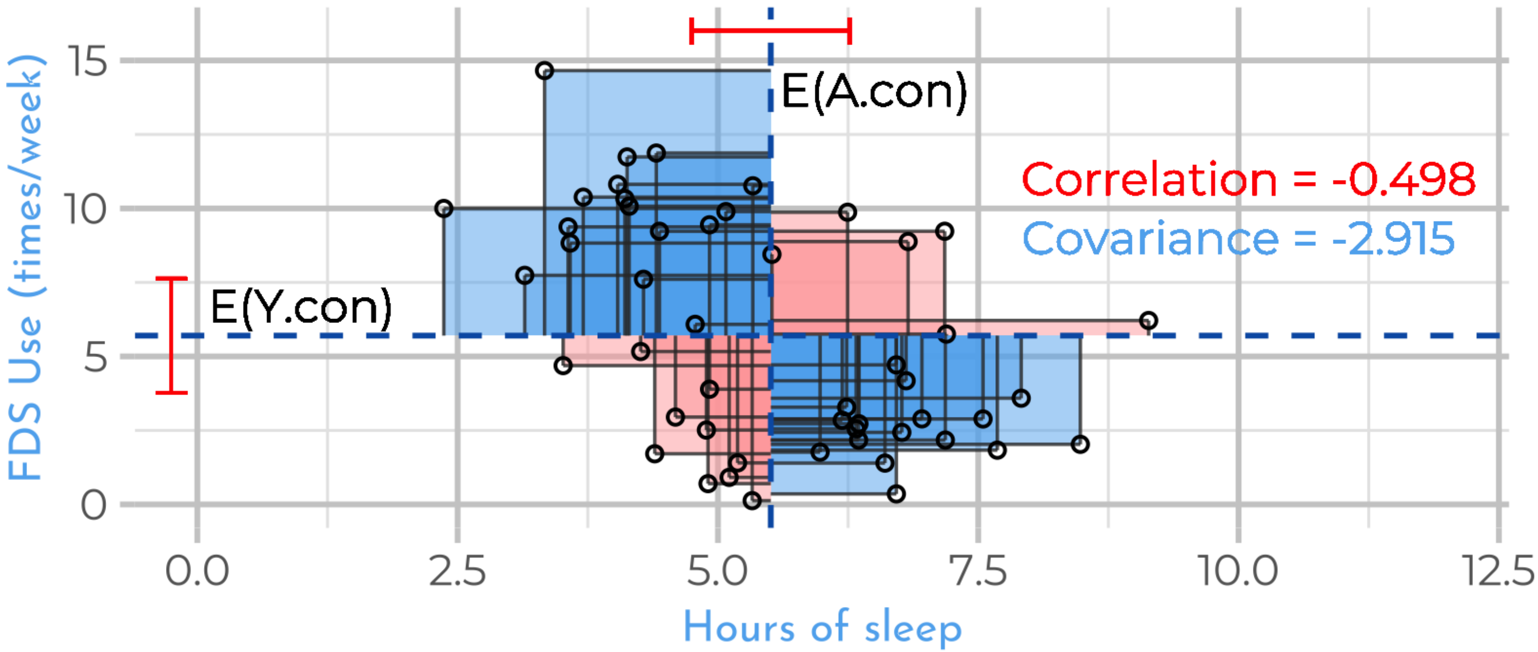
Let's use some geometry again to help illustrate what the correlation measures. Recall our plot of the covariance as a series of rectangles

correlation



The standard deviation of our `A.con` or `Y.con` variable is the average deviation between each point and their group means, i.e.
$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n}}$, and is on the original scale of the variable

correlation



The correlation is then the average area of each rectangle divided by the product of the lengths of the two red bars.

Question: Why does this solve the scaling issue?

correlation

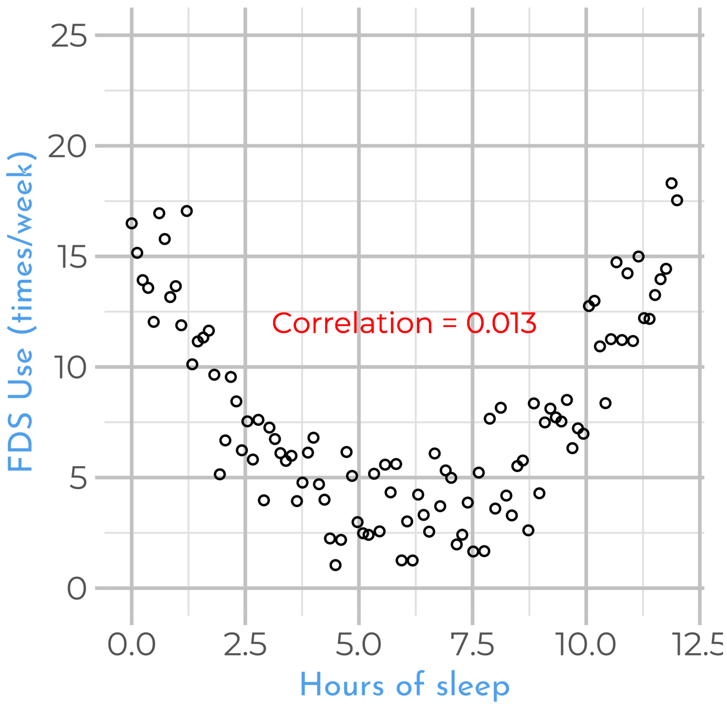
Great!! Now we can say that (1) the hours of sleep a student gets is inversely correlated with food delivery service use, and (2) that this relationship is moderately strong**

But what if in actuality, students were more likely to use FDS when they slept both very little and a lot? This plot might look like:

correlation

Great!! Now we can say that (1) the hours of sleep a student gets is inversely correlated with food delivery service use, and (2) that this relationship is moderately strong**

But what if in actuality, students were more likely to use FDS when they slept both very little and a lot? This plot might look like:



correlation

Clearly, these two variables are related. This demonstrates another limitation of the correlation, which is that ρ is only useful in cases where the relationships between random variables is linear.

correlation

Clearly, these two variables are related. This demonstrates another limitation of the correlation, which is that ρ is only useful in cases where the relationships between random variables is linear.

Second, correlations are often not meaningful for public health practice. They don't tell us anything about how much of a change in one variable is related to a change in another variable.

- For example, I could tell you that sleep hours is related to FDS use with a correlation of -0.498, but I wouldn't be able to tell you the actual decrease in FDS use for every additional hour of sleep a student got.

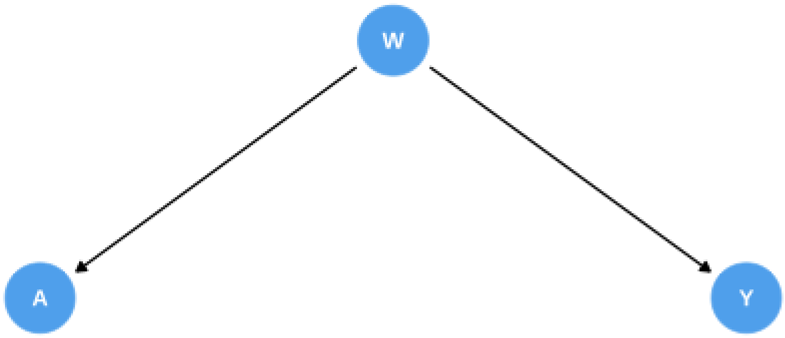
correlation, maybe causation?

Another limitation of these methods (2x2 tables, covariance, correlation), is when we are interested in not only the relationship between two random variables, but the effect one has on another.

correlation, maybe causation?

Another limitation of these methods (2x2 tables, covariance, correlation), is when we are interested in not only the relationship between two random variables, but the effect one has on another.

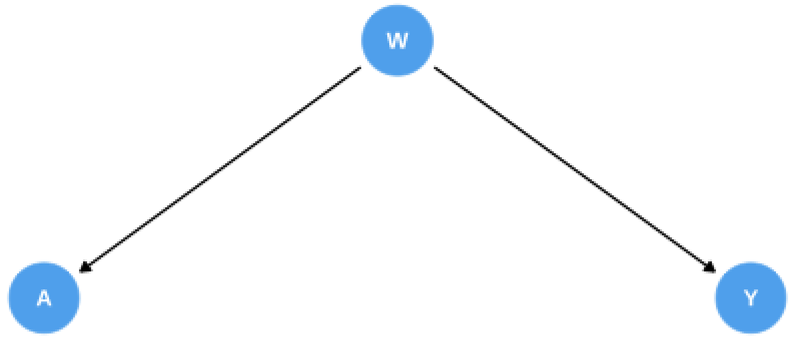
Remember our true data-generating process for the FUN study?



correlation, maybe causation?

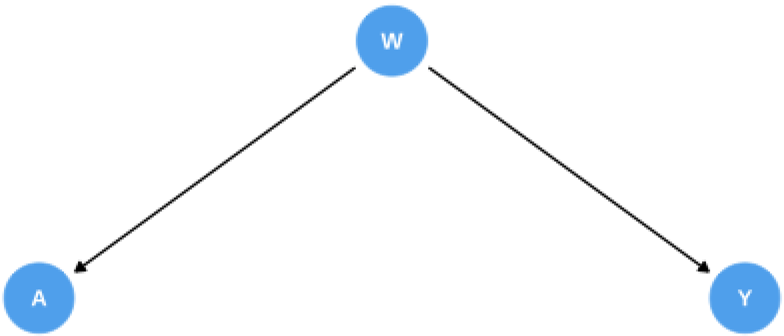
Another limitation of these methods (2x2 tables, covariance, correlation), is when we are interested in not only the relationship between two random variables, but the effect one has on another.

Remember our true data-generating process for the FUN study?



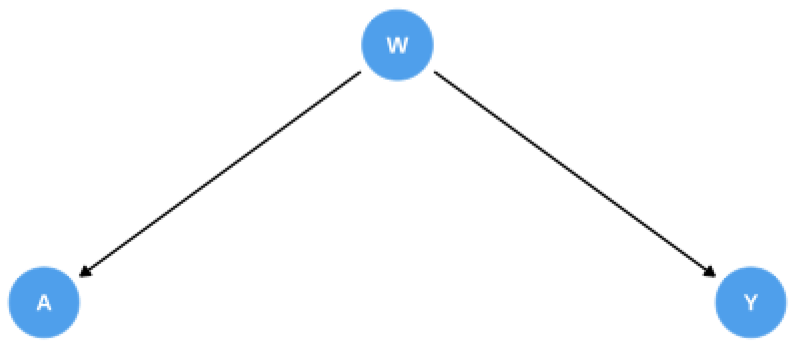
In reality, sleep doesn't have any effect on food delivery service at all, even though our calculations of the prevalence ratio, covariance, and correlation would lead us to believe otherwise.

correlation, maybe causation?



Why is this the case? Both sleep and FDS use are affected by a third variable `W`, which is an indicator of whether a student is currently writing their dissertation. Those that are dissertating are more likely to order delivery and less likely to get a full night's sleep. In other words, dissertation-writing status is a confounder of the sleep-FDS use relationship.

correlation, maybe causation?



If we don't account for this confounding in an analysis, our estimates will (usually) be spurious! Another way to think of this:

> Disseration status (`W`) is actually driving changes in food delivery service use. (`Y`) Sleep hours (`A`) might be a proxy for disseration status, so when we look at the relationship between sleep and FDS without considering dissertation writing, we see an association. However, we would be wrong to say that sleep hours itself causes FDS use.

correlation, maybe causation?

But now what? How do we move forward in the face of confounding?

- One option is to re-calculate our estimates of association (e.g. PR, covariance, correlation) within strata of our confounder. For example:

```
cor(big.FUN$A.con[big.FUN$W==1], big.FUN$Y.con[big.FUN$W==1])
```

```
## [1] 0.00291325
```

```
cor(big.FUN$A.con[big.FUN$W==0], big.FUN$Y.con[big.FUN$W==0])
```

```
## [1] 0.01592148
```

These are not exactly equal due to random noise, but both suggest little (if any) correlation between `A` and `Y` on the continuous scale. And they are both significantly different than our initial estimate of the correlation, which was -0.498!

correlation, maybe causation?

However, let's say we have not just one, but 20+ different confounders. Unless we have millions and millions of observations, there's no way we could look at the the relationships between variables in all of the (potentially infinite) number of strata. This is sometimes called the curse of dimensionality.

correlation, maybe causation?

However, let's say we have not just one, but 20+ different confounders. Unless we have millions and millions of observations, there's no way we could look at the the relationships between variables in all of the (potentially infinite) number of strata. This is sometimes called the curse of dimensionality.

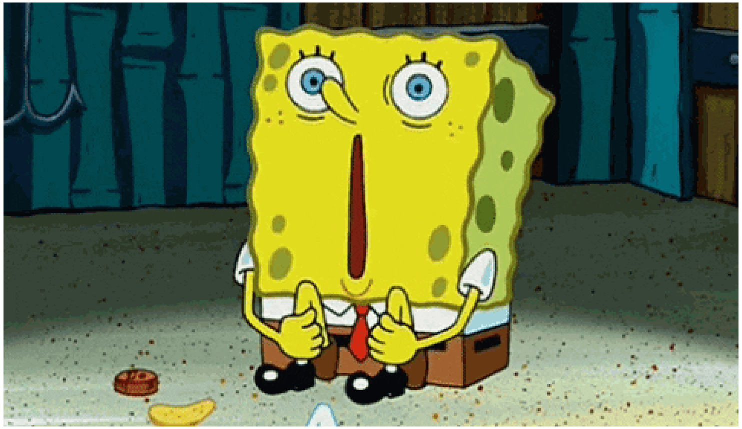
We will see how regression provides us with one way to move forward in the face of high-dimensional data.

correlation, maybe causation?

However, let's say we have not just one, but 20+ different confounders. Unless we have millions and millions of observations, there's no way we could look at the the relationships between variables in all of the (potentially infinite) number of strata. This is sometimes called the curse of dimensionality.

We will see how regression provides us with one way to move forward in the face of high-dimensional data.

But first, let's take a breather!



Questions?

Complete this R exercise on correlation and covariance **here**

the plan for today

1. Introduce the FÜN Study
2. Relationships between variables
3. **Intro to linear regression**
4. Wrapping up + conclusions

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)
4. Identify what assumptions we need to make to answer this research question (do I need to adjust for a, b, or c?)

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)
4. Identify what assumptions we need to make to answer this research question (do I need to adjust for a, b, or c?)
5. Propose a statistical model and estimate parameters

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)
4. Identify what assumptions we need to make to answer this research question (do I need to adjust for a, b, or c?)
5. Propose a statistical model and estimate parameters
6. Interpret

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)
4. Identify what assumptions we need to make to answer this research question (do I need to adjust for a, b, or c?)
5. Propose a statistical model and estimate parameters
6. Interpret

This is what your subject-matter knowledge helps with!

linear regression

Linear regression is a method that allows us to use data efficiently and flexibly to quantify relationships between random variables. For a research question of interest, there are a number of steps we can take to reach a conclusion:

1. Specify a causal model (how does the world work?)
2. Connect observed data to causal model (how do my data work?)
3. Translate our research question into a mathematical expression and statistical estimand (odds ratio? risk difference?)
4. Identify what assumptions we need to make to answer this research question (do I need to adjust for a, b, or c?)
5. Propose a statistical model and estimate parameters
6. Interpret

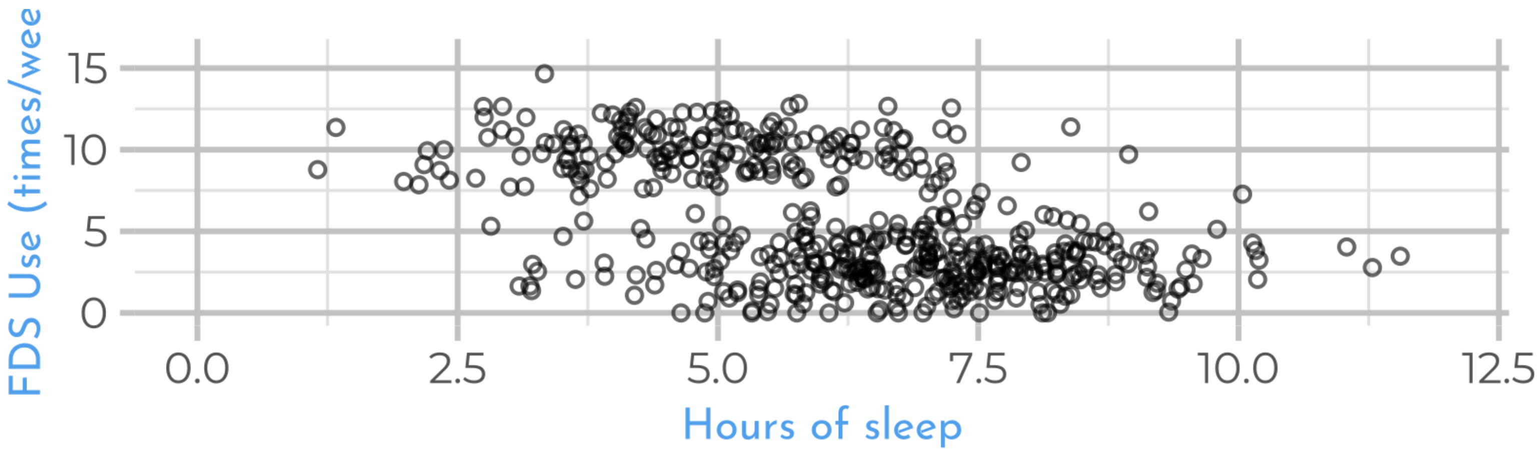
This is what linear regression can help us with! We will discuss steps 3., 5., and 6. (and come back to 4.)

linear regression

Today, we will discuss linear regression in the context of two continuous random variables, but over the course of this semester we will also learn what to do with discrete, time-to-event, and Bernoulli variables.

3. question of interest

Let's look at our FUN study example again, plotting sleep time against delivery service use. This time, we'll take a random sample of 500 students from the 100,000.



We might be interested in the question: For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average? Neither correlation or covariance directly answers this question

3. question of interest

Let's translate this quantity into math: For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average

3. question of interest

Let's translate this quantity into math: For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average

Let X represent any arbitrary number of hours of sleep per night. Let Y represent any arbitrary number of food delivery service uses per week. In other words, we want to know how the mean of Y changes given a 1 unit increase in X. This is our target estimand

3. question of interest

Let's translate this quantity into math: For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average

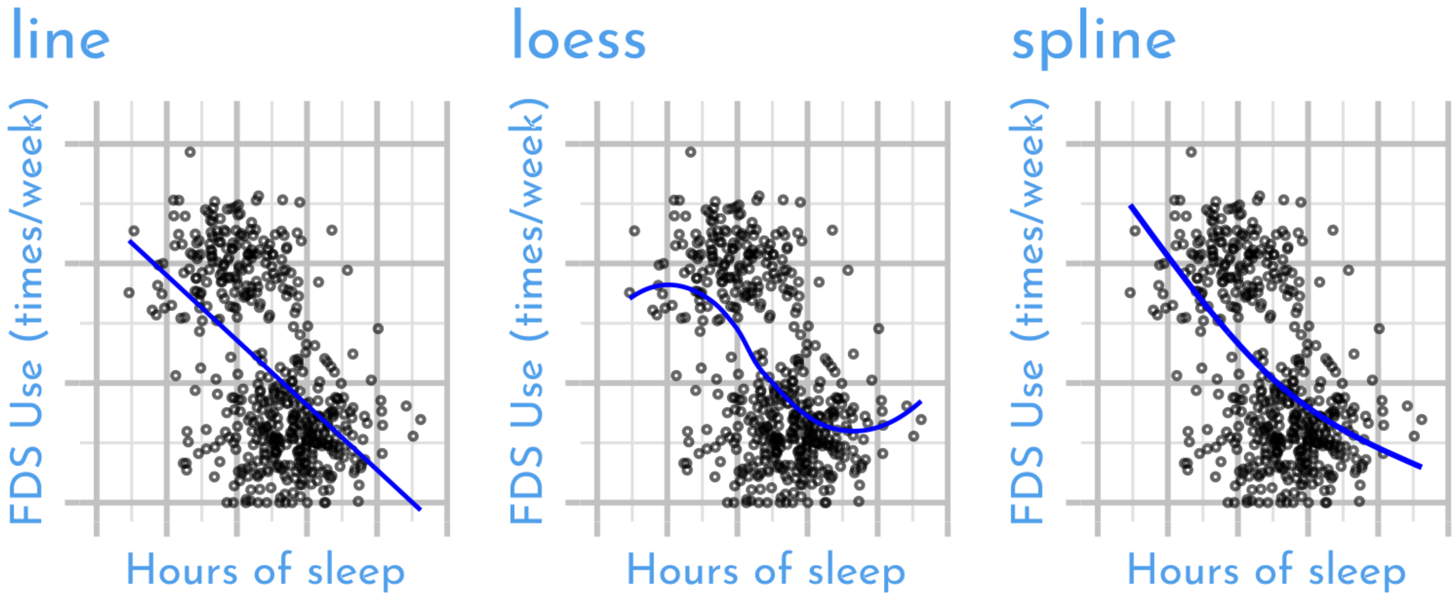
Let X represent any arbitrary number of hours of sleep per night. Let Y represent any arbitrary number of food delivery service uses per week. In other words, we want to know how the mean of Y changes given a 1 unit increase in X. This is our target estimand

$$E(Y \mid X = (x+1)) - E(Y \mid X = x)$$

5a. statistical model

In order to relate these two variables, we need to map our change in X to our change in $E(Y)$, making some assumption about their relationship. This is where we have a choice!

$$(X+1)-(X) \overset{?}{\mapsto} \Delta E(Y)$$



5a. statistical model

Let's assume for simplicity that the relationship between $E(Y)$ and X is a straight line. Then our *statistical model* relating mean FDS use and hours of sleep can be written as:

$$E(Y \mid X = x) = \beta_0 + \beta_1 X$$

5a. statistical model

Let's assume for simplicity that the relationship between $E(Y)$ and X is a straight line. Then our *statistical model* relating mean FDS use and hours of sleep can be written as:

$$E(Y \mid X = x) = \beta_0 + \beta_1 X$$

This isn't really anything new -- it's the same as:

$$\begin{aligned} y &= mx + b \\ y &= b + mx \\ E(FDS \mid SleepHours) &= \beta_0 + \beta_1(SleepHours) \end{aligned}$$

Which you've probably seen already in other classes. The main point here is that this statistical model encodes an assumption that the relationship between mean FDS use to hours of sleep is governed by an intercept and a slope.

5a. statistical model

But how is this statistical model related to our question of interest? For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?

$$E(Y \mid X = (x+1)) - E(Y \mid X = x)$$

5a. statistical model

But how is this statistical model related to our question of interest? For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?

$$E(Y \mid X = (x+1)) - E(Y \mid X = x)$$

Well, we know that for any arbitrary value of $X = x$, the expected value of Y according to our model is

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \tag{1}$$

And for the next value $X = x + 1$, we can substitute $(x + 1)$ into this expression:

$$E(Y \mid X = (x+1)) = \beta_0 + \beta_1 (x + 1) \tag{2}$$

5a. statistical model

If we subtract equation (1) from equation (2), we get:

$$\begin{aligned} E(Y \mid X = (x+1)) - E(Y \mid X = x) &= \beta_0 + \beta_1(x+1) \\ &\quad - (\beta_0 + \beta_1 x) \\ &= \boxed{\beta_1} \end{aligned}$$

5a. statistical model

If we subtract equation (1) from equation (2), we get:

$$\begin{aligned} E(Y \mid X = (x+1)) - E(Y \mid X = x) &= \beta_0 + \beta_1(x+1) \\ &\quad - (\beta_0 + \beta_1 x) \\ &= \boxed{\beta_1} \end{aligned}$$

Which shows us that what we're interested in: the change in mean food delivery service use for a 1 hour increase in hours of sleep, is simply given by β_1, or the slope, from this statistical model!

5a. statistical model

If we subtract equation (1) from equation (2), we get:

$$\begin{aligned} E(Y \mid X = (x+1)) - E(Y \mid X = x) &= \beta_0 + \beta_1(x+1) \\ &\quad - (\beta_0 + \beta_1 x) \\ &= \boxed{\beta_1} \end{aligned}$$

Which shows us that what we're interested in: the change in mean food delivery service use for a 1 hour increase in hours of sleep, is simply given by β_1, or the slope, from this statistical model!

Given our data, the next question is how exactly to estimate β_1. In other words, what's the most likely value of the slope relating mean FDS to sleep hours, considering what we actually observe?

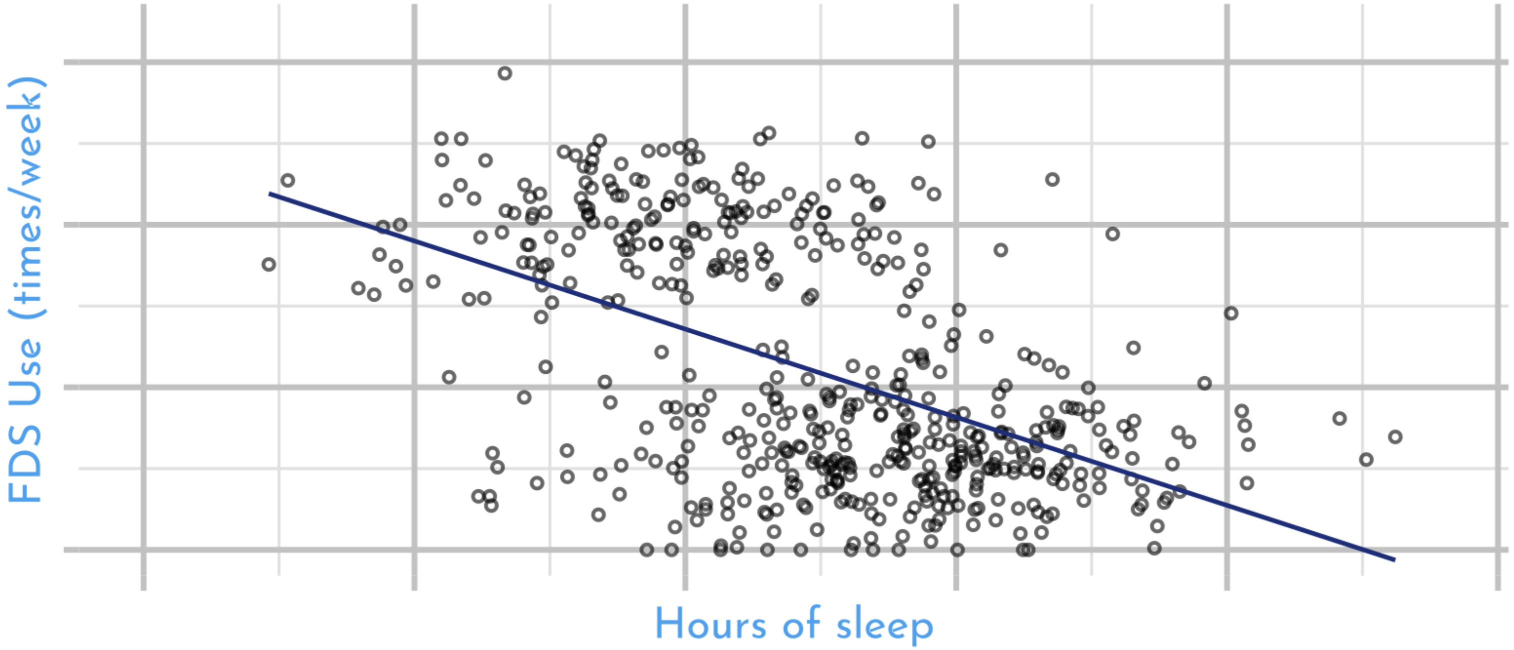
5b. estimation

The most common way to estimate parameters from a linear model like the one we've specified in our statistical model is an algorithm called ordinary least squares (OLS).

> Aside: naming conventions in statistics can be weird. Least squares: based on the mechanism of the algorithm. Ordinary: less complicated than methods developed later chronologically.

5b. estimation

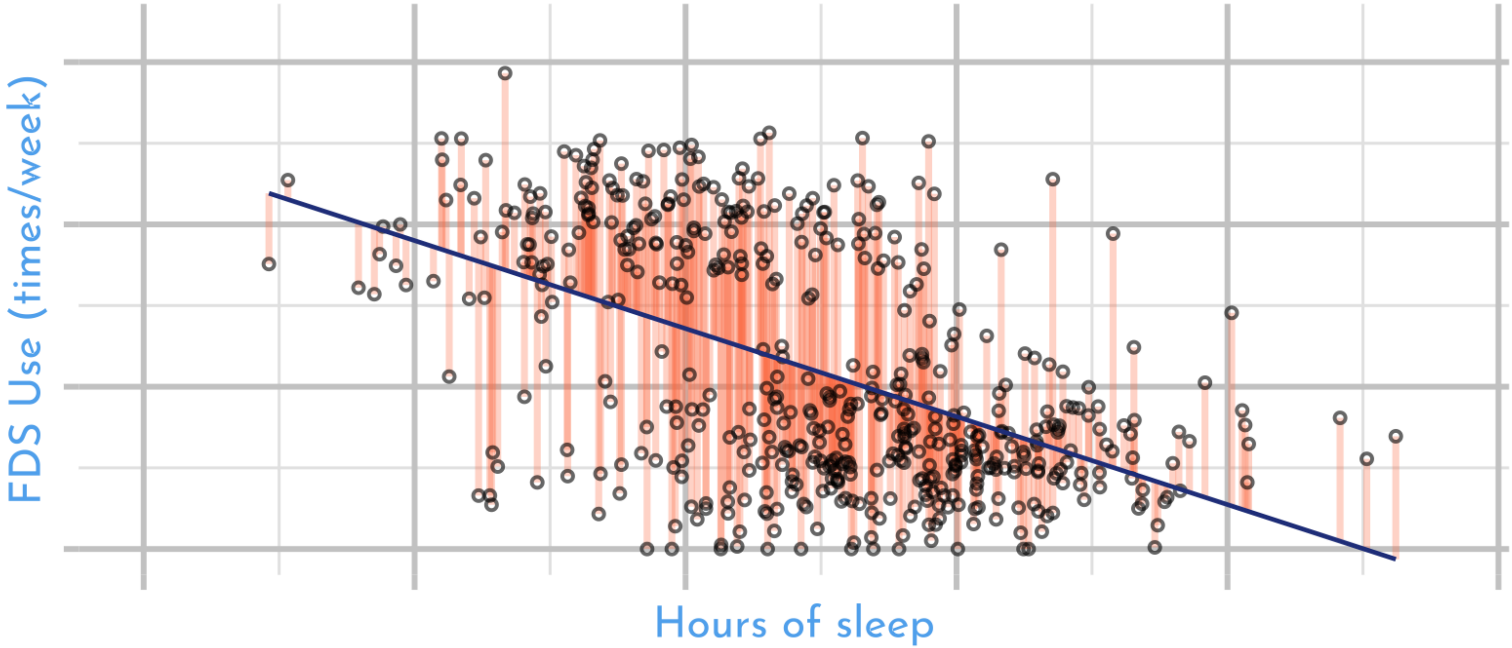
Let's look at the steps in the OLS algorithm:



1. Pick a line, any line, by defining candidate values of β_0 and β_1

5b. estimation

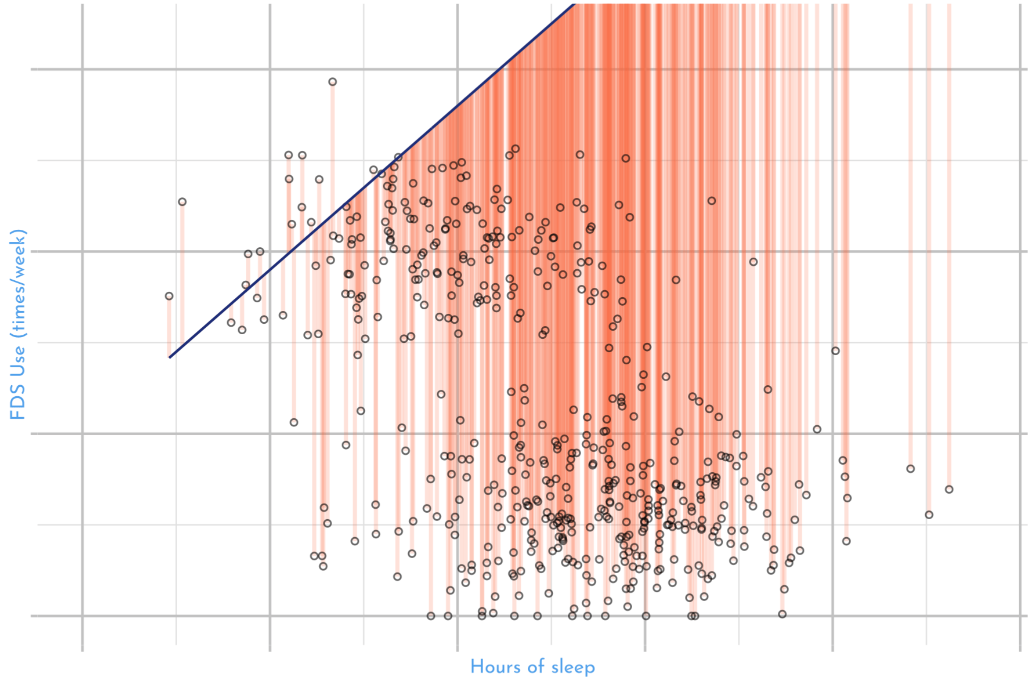
Let's look at the steps in the OLS algorithm:



2. Calculate the difference between the observed points and the value of E(Y) predicted by our candidate model, and square it.

5b. estimation

Let's look at the steps in the OLS algorithm:



3. Find the combination of intercept and slope that **minimizes the average of the squared distances**

5b. estimation

When we only have two variables, it turns out the OLS solution to our question: what is the most likely slope given the data I observe is given by:

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

If we divide both the numerator by $n - 1$, we have:

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

Which is cool because:

$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$: the sample covariance between X and Y

$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$: the sample variance of X.

5b. estimation

Let's do this with the FUN dataset in `R` to illustrate:

```
ols.fit <- lm(Y.con ~ A.con, data = big.FUN)
summary(ols.fit)

Call:
lm(formula = Y.con ~ A.con, data = big.FUN)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0173 -2.3592 -0.3798  2.2430 11.2263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.53031    0.11610   99.31   <2e-16 ***
A.con       -0.98526    0.01761  -55.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.145 on 9998 degrees of freedom
Multiple R-squared:  0.2385,    Adjusted R-squared:  0.2384
F-statistic:  3131 on 1 and 9998 DF,  p-value: < 2.2e-16
```

5b. estimation

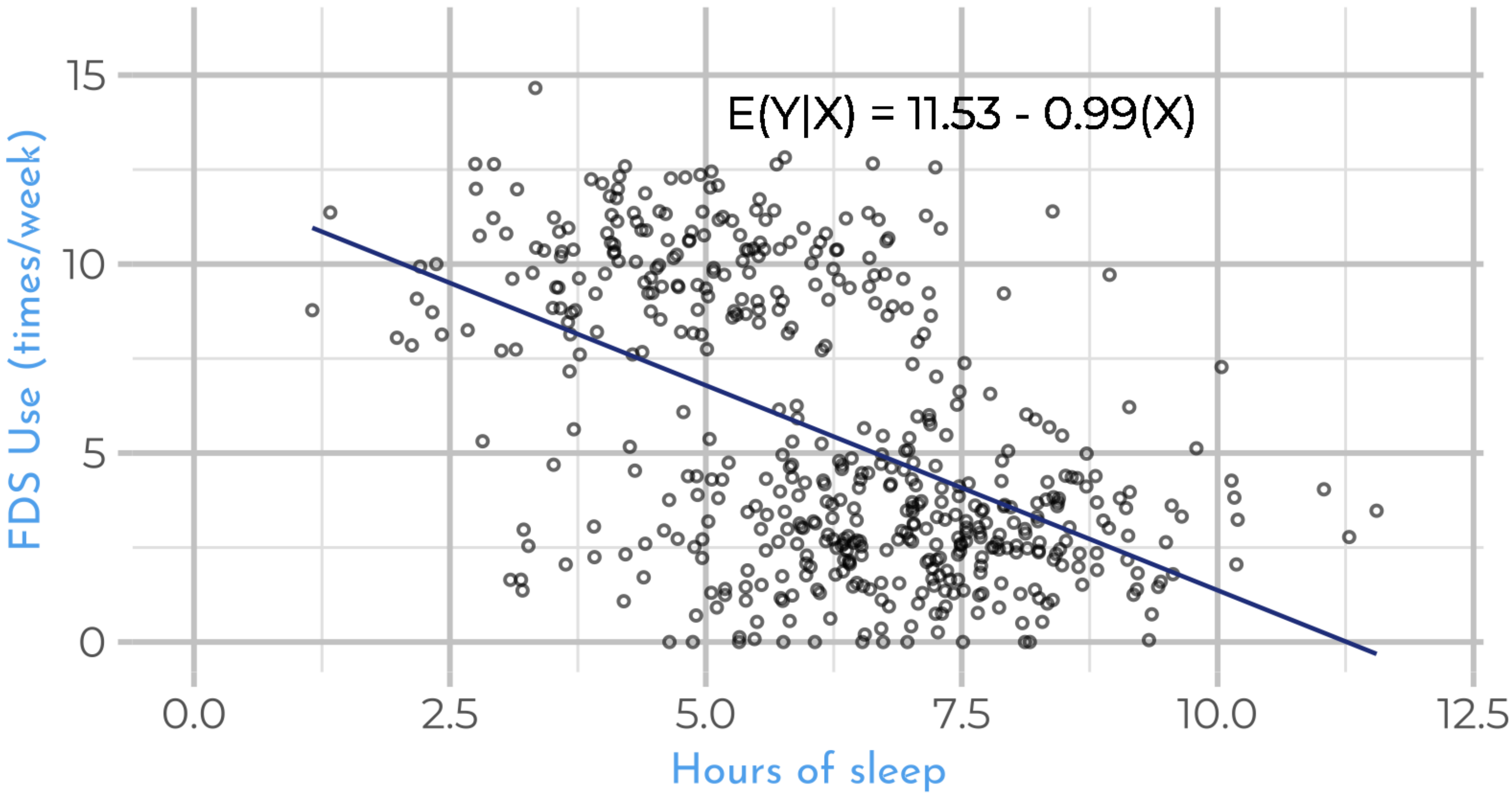
Let's do this with the FUN dataset in `R` to illustrate:

```
ols.fit <- lm(Y.con ~ A.con, data = big.FUN)
summary(ols.fit)

Call:
lm(formula = Y.con ~ A.con, data = big.FUN)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0173 -2.3592 -0.3798  2.2430 11.2263 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.53031    0.11610   99.31   <2e-16 ***
A.con       -0.98526    0.01761  -55.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.145 on 9998 degrees of freedom
Multiple R-squared:  0.2385,    Adjusted R-squared:  0.2384 
F-statistic:  3131 on 1 and 9998 DF,  p-value: < 2.2e-16
```

The estimated line according to OLS is therefore given by:

$$E(Y \mid X = x) = 11.53 - 0.99(X)$$

5b. estimation



6. interpretation

How do we interpret this? Remember that our research question, for every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?, is represented by the expression $E(Y \mid X = (x+1)) - E(Y \mid X = x)$, which we determined was equal to β_1 according to our statistical model.

6. interpretation

How do we interpret this? Remember that our research question, for every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?, is represented by the expression $\boldsymbol{E}(Y \mid X = (x+1)) - \boldsymbol{E}(Y \mid X = x)$, which we determined was equal to β_1 according to our statistical model.

We used OLS to estimate the intercept and slope parameters, and found that $\widehat{\beta_1} = -\mathbf{0.99}$

6. interpretation

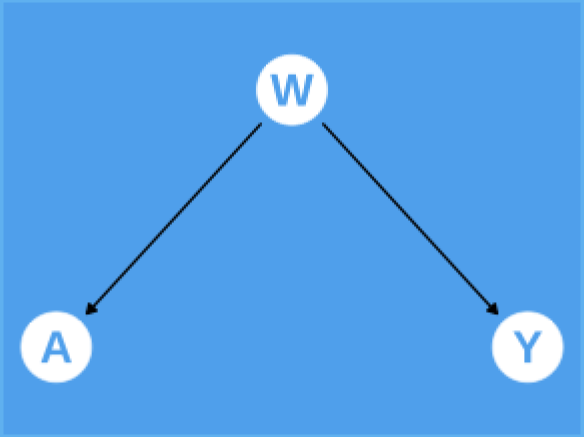
How do we interpret this? Remember that our research question, for every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?, is represented by the expression $E(Y \mid X = (x+1)) - E(Y \mid X = x)$, which we determined was equal to β_1 according to our statistical model.

We used OLS to estimate the intercept and slope parameters, and found that $\widehat{\beta_1} = -\mathbf{0.99}$

Working backwards using this logic, then, we can conclude that for every additional hour of sleep a student gets per night, food delivery service decreases by 0.99 times per week on average.

beta estimation, maybe causation?

But wait! Remember our true data-generating process for the FUN study. In reality, sleep hours **does not** affect delivery service use.



We have the same issue of confounding, just as we did when calculating the prevalence ratio, covariance, and correlation! Linear regression **doesn't magically solve these issues**

beta estimation, maybe causation?

However, it does help us move forward more efficiently. While we could just as easily re-run our regression models within strata of our covariate (dissertation writing status) another solution is to simply include this confounder as a covariate in our statistical model.

beta estimation, maybe causation?

However, it does help us move forward more efficiently. While we could just as easily re-run our regression models within strata of our covariate (dissertation writing status) another solution is to simply include this confounder as a covariate in our statistical model.

Instead of our previous model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

beta estimation, maybe causation?

However, it does help us move forward more efficiently. While we could just as easily re-run our regression models within strata of our covariate (dissertation writing status) another solution is to simply include this confounder as a covariate in our statistical model.

Instead of our previous model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

We can include dissertation writing status (let's call it W) into the mix:

$$E(Y \mid X = x, W = w) = \beta_0 + \beta_1 x + \beta_2 w$$

beta estimation, maybe causation?

However, it does help us move forward more efficiently. While we could just as easily re-run our regression models within strata of our covariate (dissertation writing status) another solution is to simply include this confounder as a covariate in our statistical model.

Instead of our previous model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

We can include dissertation writing status (let's call it W) into the mix:

$$E(Y \mid X = x, W = w) = \beta_0 + \beta_1 x + \beta_2 w$$

What new assumption does this encode?

beta estimation, maybe causation?

However, it does help us move forward more efficiently. While we could just as easily re-run our regression models within strata of our covariate (dissertation writing status) another solution is to simply include this confounder as a covariate in our statistical model.

Instead of our previous model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

We can include dissertation writing status (let's call it W) into the mix:

$$E(Y \mid X = x, W = w) = \beta_0 + \beta_1 x + \beta_2 w$$

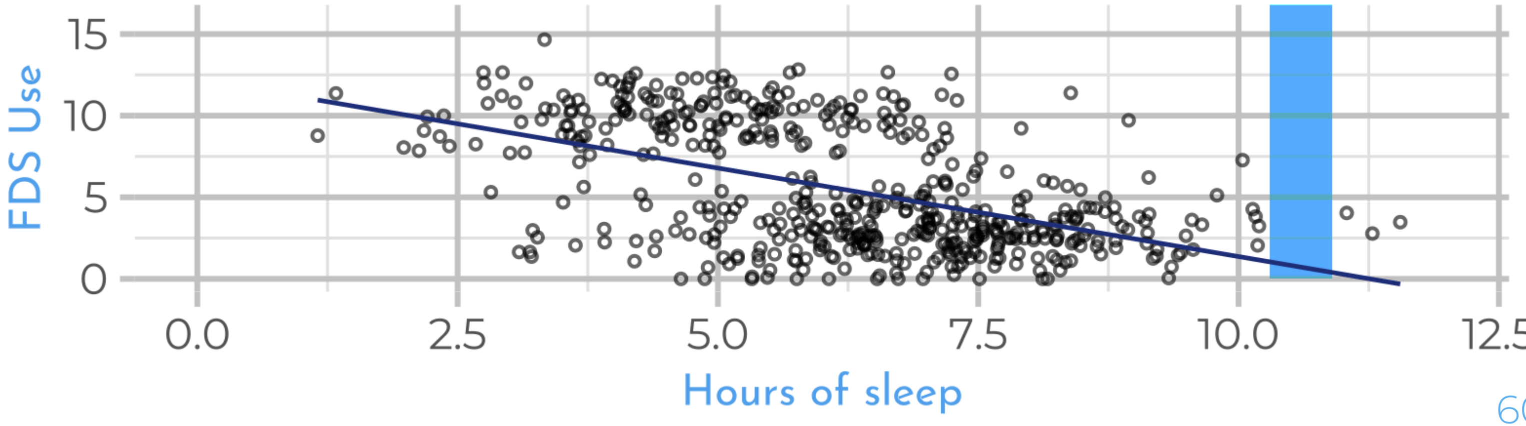
What new assumption does this encode?

Mean FDS use per week is a function of sleep hours and/or dissertation writing, or neither

beta estimation, maybe causation?

Linear regression, compared to correlation and covariance, is:

- Flexible:
  - I can add 1, 10, or 50 covariates (assuming I have the data)
  - I can easily change my assumptions about the functional form of relationships between variables (e.g. line? curve? other?)
- Efficient: When the data are sparse, I can borrow information from neighboring observations (this is particularly helpful when we have multiple continuous explanatory variables)



beta estimation, maybe causation?

We can fit an updated model, with dissertation status, in `R`:

```
ols.fit2 <- lm(Y.con ~ A.con + W, data = big.FUN)
summary(ols.fit2)

Call:
lm(formula = Y.con ~ A.con + W, data = big.FUN)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3039 -1.0379 -0.0136  1.0038  6.5108 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.90385    0.07199   40.34   <2e-16 ***
A.con        0.01151    0.00992    1.16    0.246    
W            7.01219    0.03774  185.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 9997 degrees of freedom
Multiple R-squared:  0.829,	Adjusted R-squared:  0.829 
F-statistic: 2.423e+04 on 2 and 9997 DF,  p-value: < 2.2e-16
```

This gives us the right answer: There is no relationship between sleep hours and food delivery service use independent of dissertation writing status

Questions?

Complete this R exercise on regression basics here

the plan for today

1. Introduce the FÜN Study
2. Relationships between variables
3. Intro to linear regression
4. Wrapping up + conclusions

4. identify assumptions needed

Linear regression allows to relate one random variable to another. If we're interested in causal relationships, i.e.:

4. identify assumptions needed

Linear regression allows to relate one random variable to another. If we're interested in causal relationships, i.e.:

> For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?

vs.

4. identify assumptions needed

Linear regression allows to relate one random variable to another. If we're interested in causal relationships, i.e.:

> For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?

vs.

> If students sleep an additional hour per night, what is the absolute change in food delivery service use compared to if students did not sleep an additional hour per night?

4. identify assumptions needed

Linear regression allows to relate one random variable to another. If we're interested in causal relationships, i.e.:

> For every additional hour of sleep a student gets per night, what is the absolute change in food delivery service use on average?

vs.

> If students sleep an additional hour per night, what is the absolute change in food delivery service use compared to if students did not sleep an additional hour per night?

Then we need to think carefully about what covariates we need to adjust for (i.e. include in our regression model) in order to isolate the effect of sleeping on food delivery service use.

4. identify assumptions needed

This is to say that linear regression, like correlation or covariance or other methods, is just a tool we can use to answer questions. But these tools can be mis-used and without the proper subject matter knowledge, can lead us astray.

i.e. regression is dumb, you are not!

you made it!

You are here for a reason!

You can learn!

You are capable!

We believe in you!