The Simple Regression Model

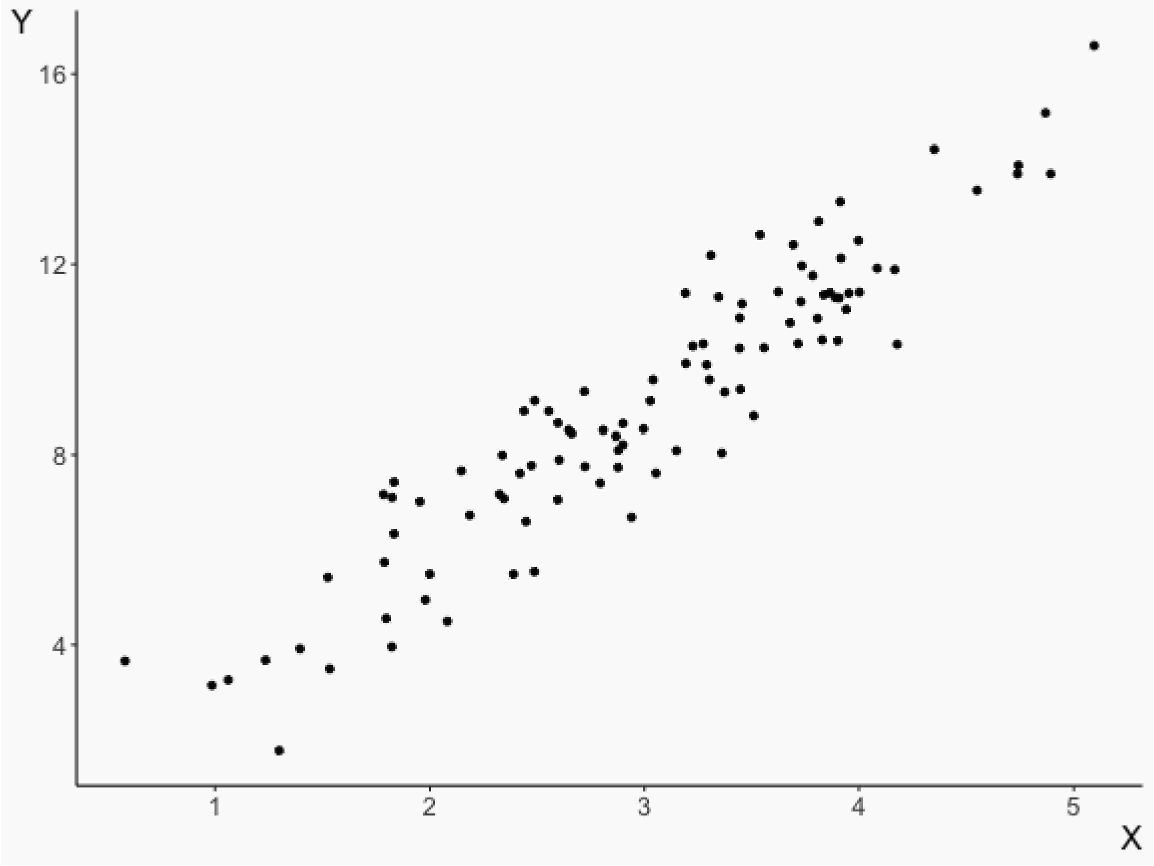
Introduction

What is Regression?

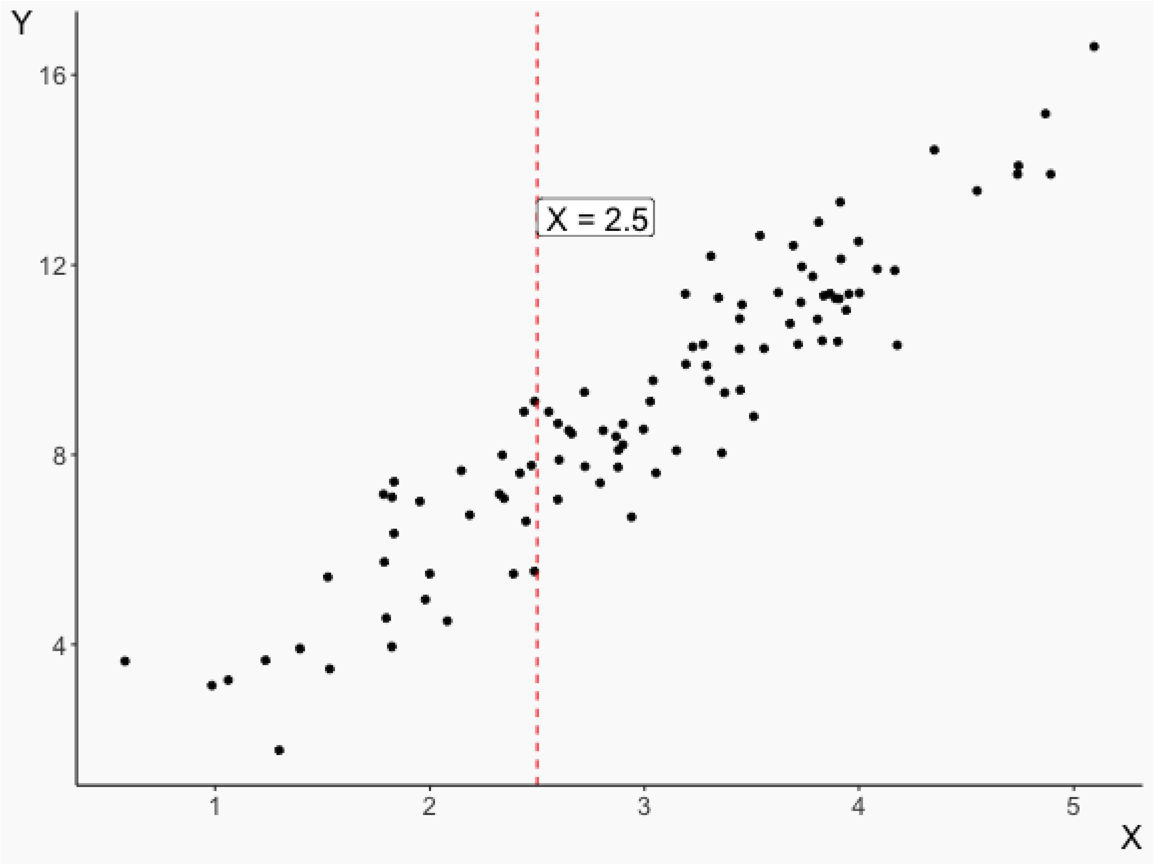
- In statistics, regression is the practice of *line-fitting*
- We want to *use one variable to predict another*
- Let's say using X to predict Y
- We'd refer to X as the "independent variable", and Y as the "dependent variable"
- **Regression is the idea that we should characterize the relationship between X and Y as a *line*, and use that line to predict Y**

X and Y

- Data for X and Y

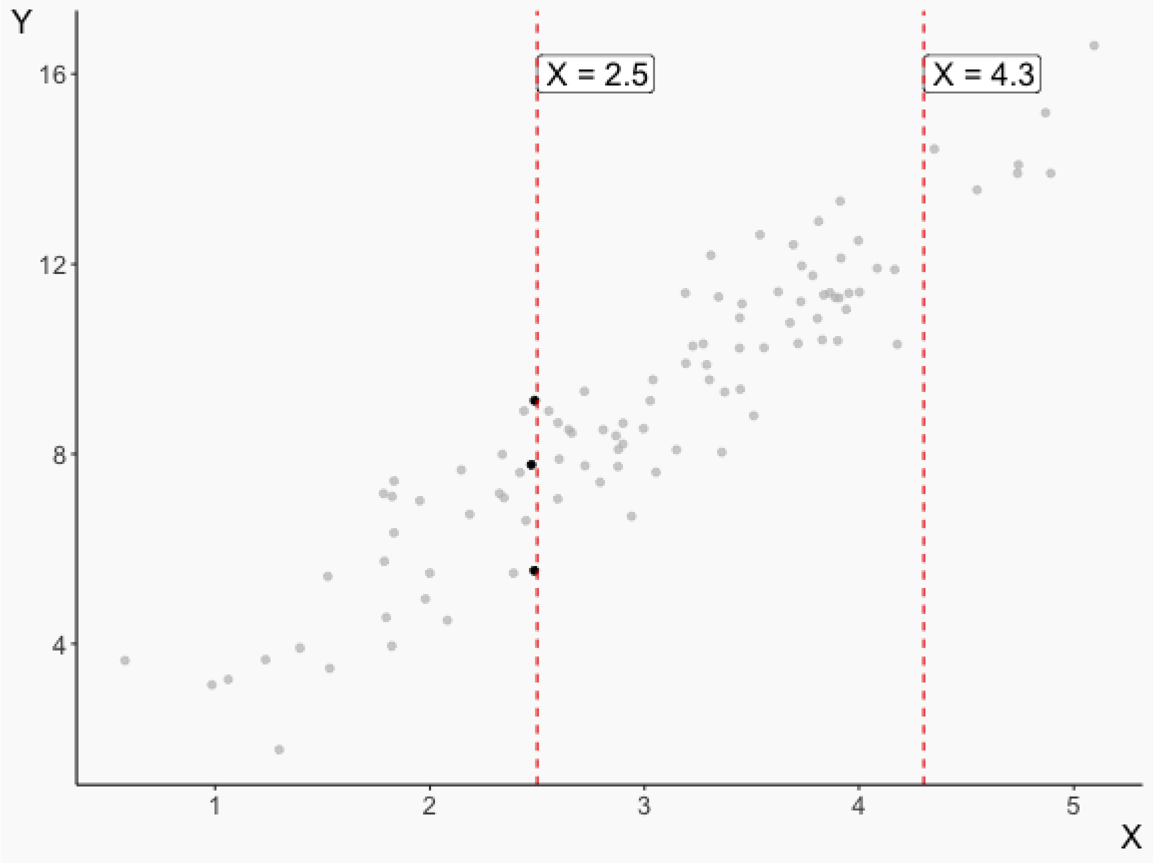


X and Y



- X is 2.5 and want to predict Y.
- If we look carefully there are multiple of values of Y for $X = 2.5$. What would the predicted value of Y be?

X and Y



- What if we want to predict Y for a value we DON'T have any actual observations of, like $X = 4.3$?

Data is Granular

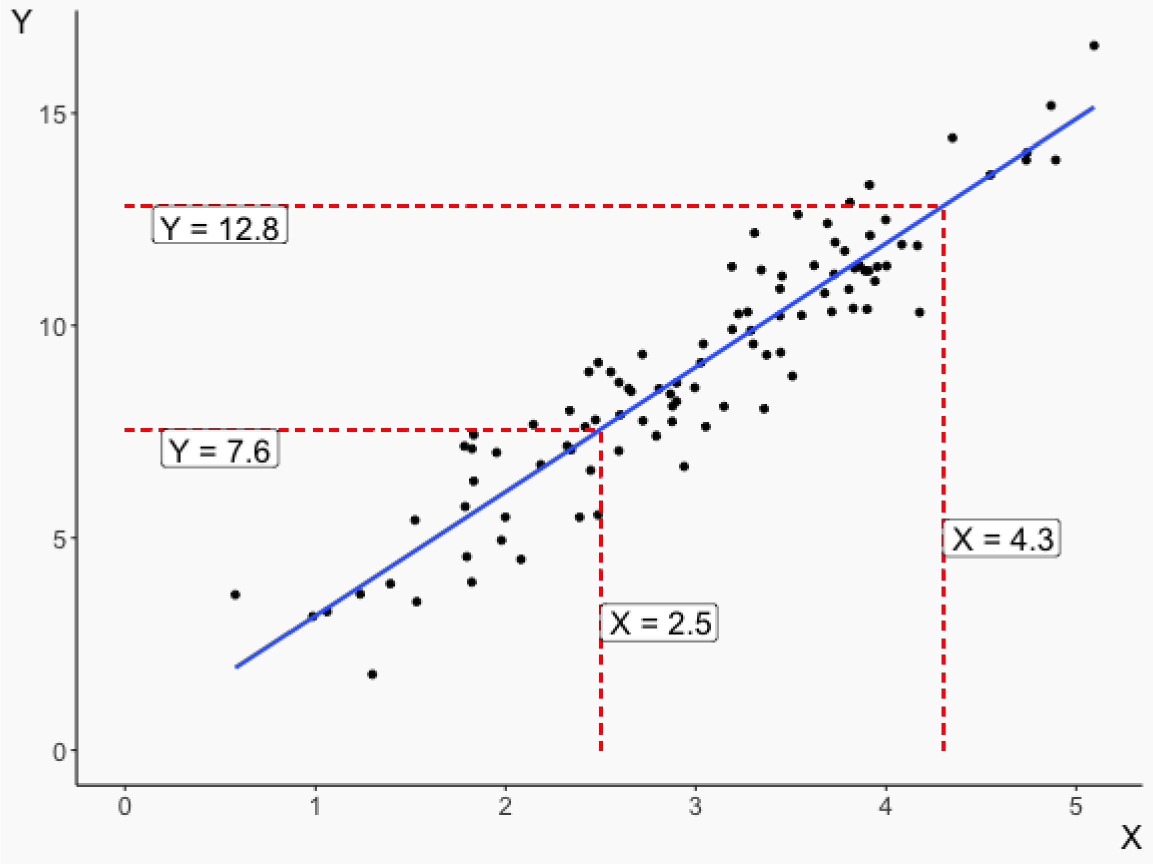
- If we try to fit *every point*, we will get a mess that won't really tell us the relationship between X and Y
- So, we *simplify* the relationship into a *shape*: a line! The line smooths out those three points around $X = 2.5$ and fills in that gap around $X = 4.3$.

Isn't This Worse?

- By adding a line, we are necessarily *simplifying* our presentation of the data.
- Our prediction of the *data we have* will be less accurate than if we just make predictions point-by-point
- However, we'll do a better job predicting *other* data (avoiding "overfitting")
- And, since a *shape* is something we can interpret, as opposed to a long list of predictions, which we can't really, the line will do a better job of telling us about the *true underlying relationship*

The Line Does a Few Things:

- We can get a *prediction* of Y for a given value of X (If we follow $X = 2.5$ up to our line we get $Y = 7.6$)
- We see the *relationship*: the line slopes up, telling us that "more X means more Y too!"



Lines

- The line we get is the *fit* of our model
- A model "fit" means we've taken a *shape* (our line) and picked the one that best fits our data
- All forms of regression do this
- Ordinary Least Squares specifically uses a **straight line** as its shape
- The resulting line we get can also be written out as an actual line, i.e.

$$Y = intercept + slope * X$$

Lines

- We can use the *line* to plug in a value of X and get a prediction for Y
- Because these Y values are predictions, we'll give them a hat $\hat{Y}$

$$Y = 3 + 4 * X$$

$$\hat{Y} = 3 + 4 * (3.2)$$

$$\hat{Y} = 15.8$$

Lines

- The intercept is the prediction of Y when $X = 0$

$$Y = 3 + 4 * X$$

$$\hat{Y} = 3 + 4 * 0$$

$$\hat{Y} = 3$$

Lines

- And as X increases, we know how much we expect Y to increase because of the slope

$$Y = 3 + 4 * X$$

$$\hat{Y} = 3 + 4 * 3 = 15$$

- How much does Y increases when we increase X by 1?

$$\hat{Y} = 3 + 4 * 4 = 19$$

- In increase by the **slope** (which is 4 here)

Ordinary Least Squares. Review

- Regression fits a *shape* to the data
- Ordinary least squares specifically fits a *straight line* to the data
- The straight line is described using an $intercept$ and a $slope$
- When we plug an X into the line, we get a prediction for Y, which we call $\hat{Y}$
- When $X = 0$, we predict $\hat{Y} = intercept$
- When X increases by 1, our prediction of Y increases by the $slope$
- If $slope > 0$, X and Y are positively $(+)$ related/correlated
- If $slope < 0$, X and Y are negatively $(-)$ related/correlated

Concept Checks

- How does producing a *line* let us use X to predict Y?
- If our line is $Y = 5 - 2 * X$, explain what the -2 means in a sentence
- Not all of the points are exactly on the line, meaning some of our predictions will be wrong! Should we be concerned? Why or why not?
- We know that regression fits a line. But **how does it do that** exactly?

Predictions and Residuals

- Whenever you make a prediction of any kind, you rarely get it *exactly right*
- The difference between your prediction and the actual data is the *residual*

$$Y = 3 + 4 * X$$

If we have a data point where $X = 4$ and $Y = 18$, then

$$\hat{Y} = 3 + 4 * 4 = 19$$

Then the *residual* is $Y - \hat{Y} = 18 - 19 = -1$.

Predictions and Residuals

Our relationship doesn't look like this...

$$Y = intercept + slope * X$$

Instead, it's...

$$Y = intercept + slope * X + residual$$

We still use $intercept + slope * X$ to predict Y though, so this is also

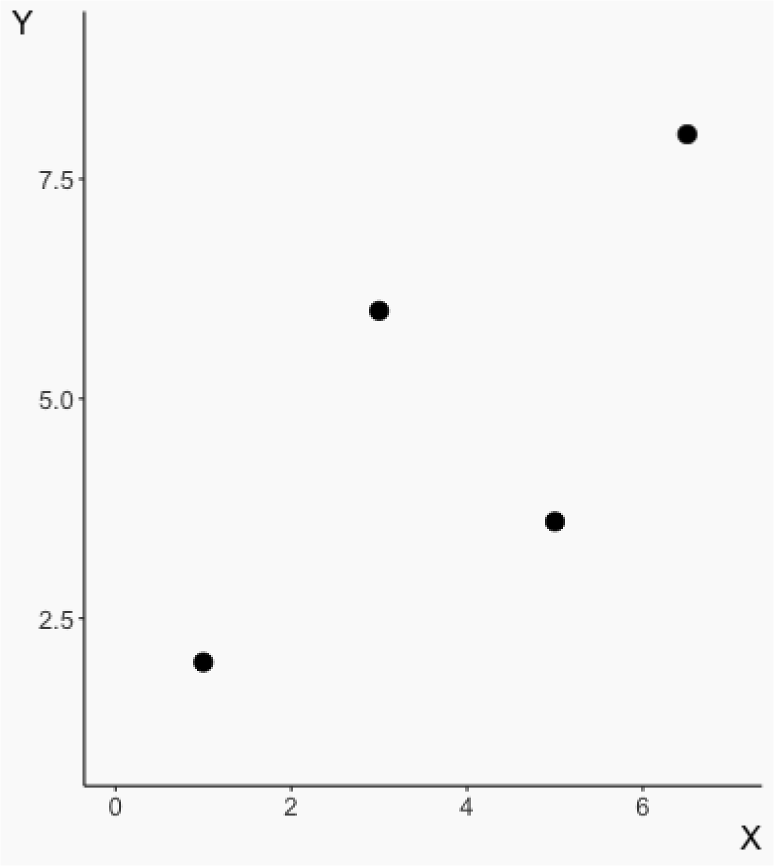
$$Y = \hat{Y} + residual$$

Ordinary Least Squares

- A good prediction should make the residuals as small as possible
- We are going to *square* those residuals, so the really-big residuals count even more. We really don't want to have points that are super far away from the line!
- Then, we pick a line to minimize those squared residuals ("least squares")

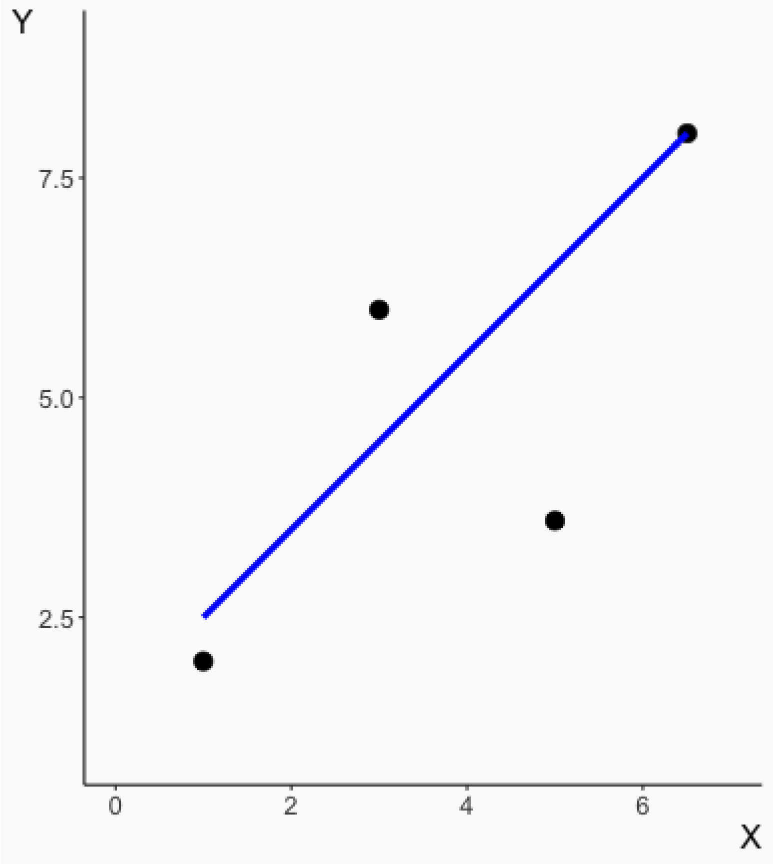
Ordinary Least Squares

- Start with our data



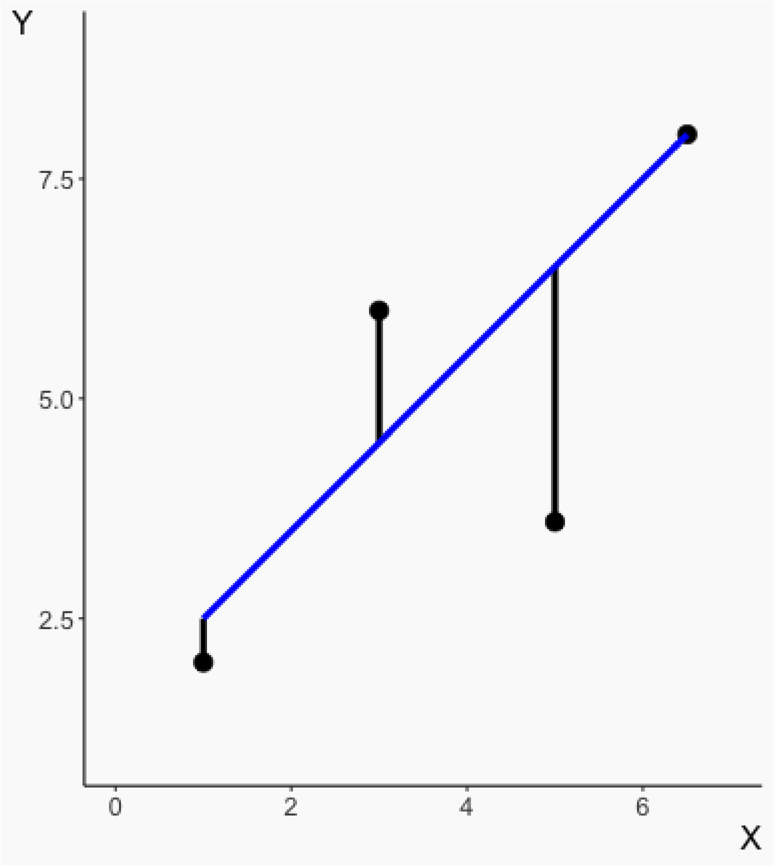
Ordinary Least Squares

- Let's just pick a line at random, not necessarily from OLS



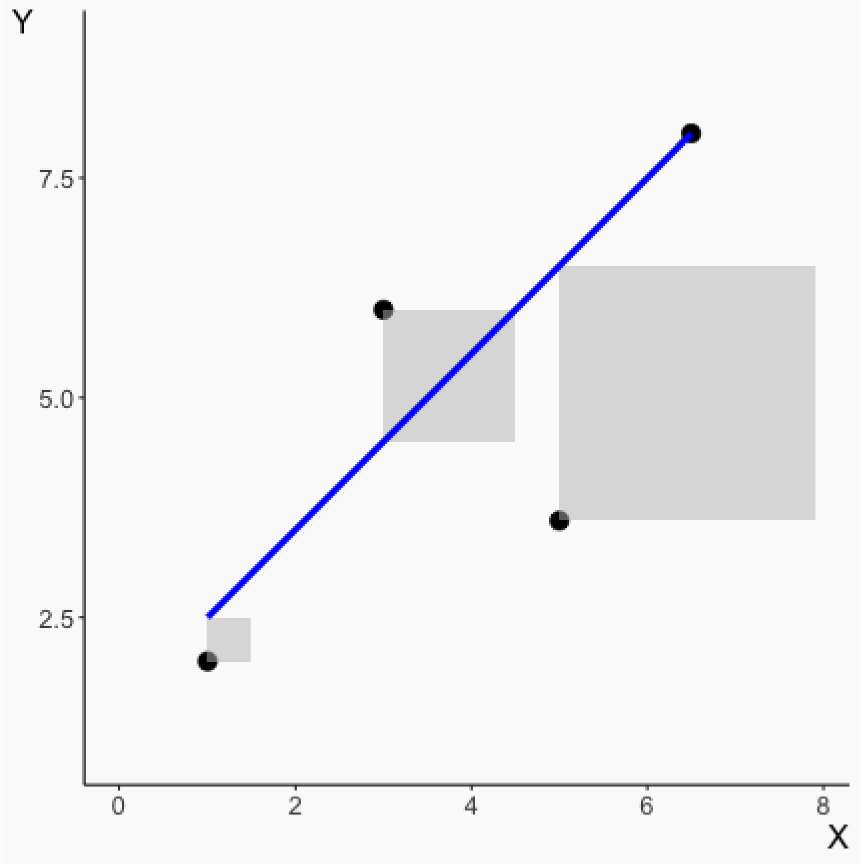
Ordinary Least Squares

- The vertical distance from point to line is the residual



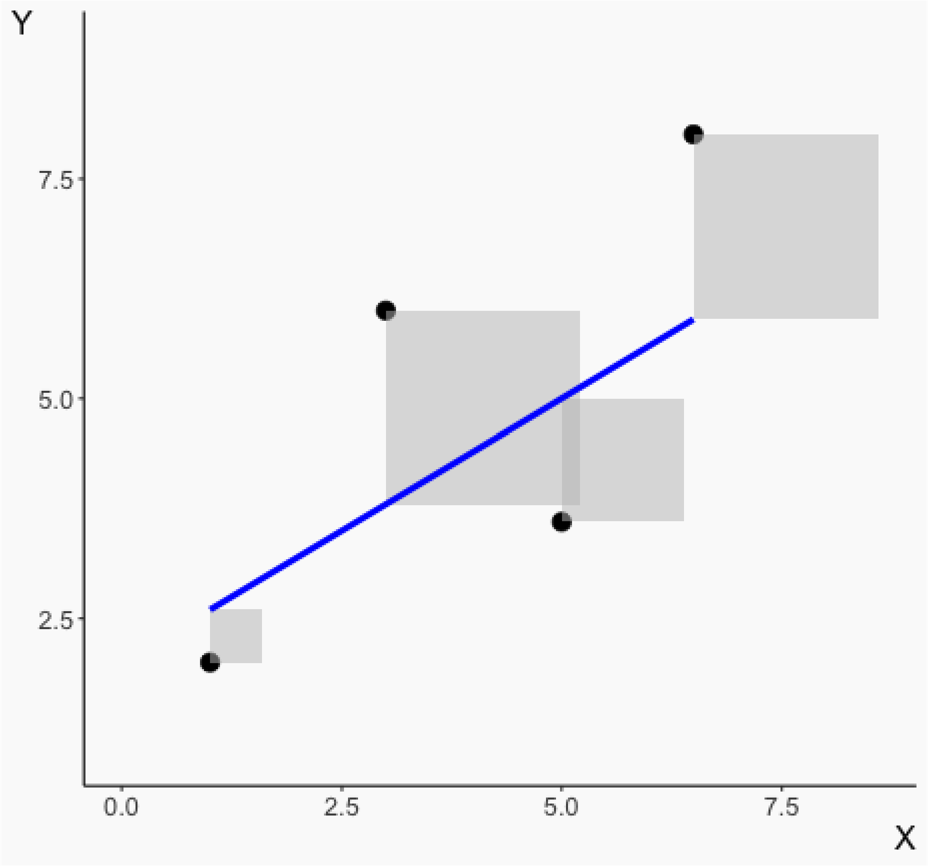
Ordinary Least Squares

- Now square those residuals



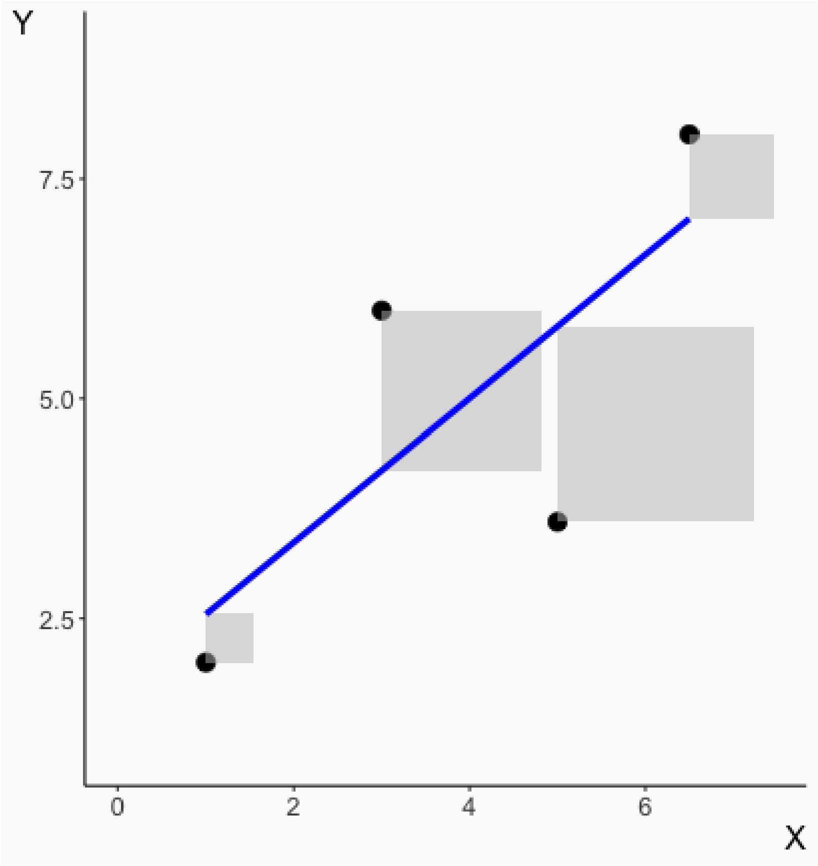
Ordinary Least Squares

- Can we get the total area in the squares smaller with a different line?



Ordinary Least Squares

- Ordinary Least Squares, I can promise you, gets it the smallest



Ordinary Least Squares

- How does it figure out which line makes the smallest squares?
- There's a mathematical formula for that!
- We will derive the formula in the next lecture
- Let's Play Now: **Guess the regression line- SIMULATION**

Concept Checks

- If I have the below OLS-fitted line from a dataset of children:

$$Height(Inches) = 18 + 2 * Age$$

And we have the kids Darryl who is 10 years old and 40 inches tall, and Bijetri who is 9 years old and 37 inches tall, what are each of their:

- predicted values
- residuals
- sum of their squared residuals

Ordinary Least Squares in R

- Ordinary Least Squares is built in to R using the `lm` function
- Let's run a regression on the `Orange` data set of tree age and circumference

```
data(Orange)
lm(circumference ~ age, data = Orange)
```

```
##
## Call:
## lm(formula = circumference ~ age, data = Orange)
##
## Coefficients:
## (Intercept)          age
##     17.3997       0.1068
```

Ordinary Least Squares in R

- In this class, we'll be using `feols()` from the **fixest** package instead of `lm()`
- This doesn't make too much difference now, and the code looks the same so far, but this will help us easily connect to other stuff

```
library(fixest)
feols(circumference ~ age, data = Orange)
```

```
## OLS estimation, Dep. Var.: circumference
## Observations: 35
## Standard-errors: IID
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 17.39965   8.622660  2.01790 5.1793e-02 .
## age          0.10677   0.008277 12.90023 1.9306e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 23.0   Adj. R2: 0.829502
```

Ordinary Least Squares in R

- What's going on here?
- `circumference ~ age` is a *formula* object. It says to take the `circumference` variable is the dependent variable (Y). Have it vary (`~`) according to `age` (independent variable, X)
- `data = Orange` tells R in which data set to look for those variables
- The output shows the `(Intercept)` $(\hat{\beta}_0)$ as well as the slope $(\hat{\beta}_1)$ on `age` (why doesn't it just say `slope`? Because later we'll have more than one slope!)

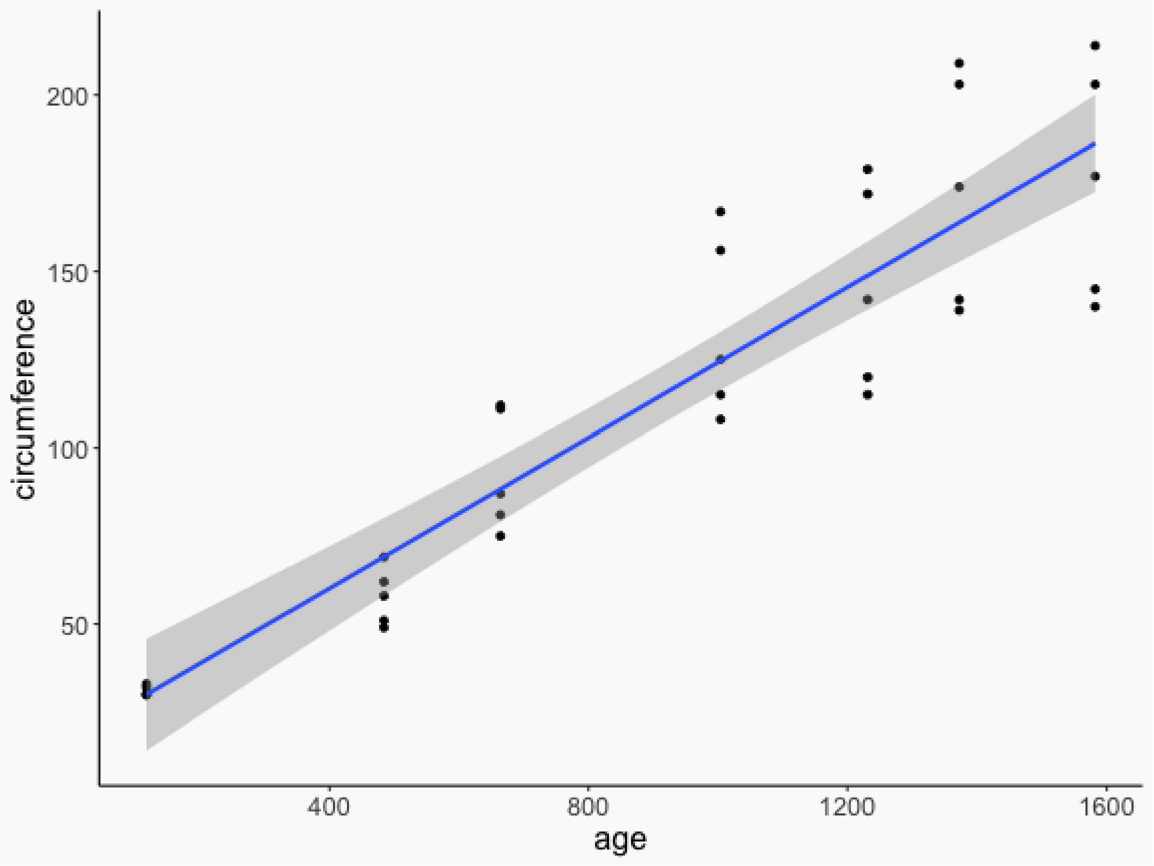
Ordinary Least Squares in R

- There's lots more information we can get from our regression, but that will wait for later
- For now, let's just make a nice graph of it using the **ggplot2** library (which you already got installed when you installed the **tidyverse**)

```
library(tidyverse) # This loads ggplot2 as well
ggplot(Orange, aes(x = age, y = circumference)) +
  geom_point() + # Draw points
  geom_smooth(method = 'lm') # add OLS line
```

Ordinary Least Squares in R

- The result:



Reference

The Effect: An Introduction to Research Design and Causality, by Nick Huntington-Klein

Teaching materials