# SMI606: Week 7
# Bivariate Linear Regression

**Dr. Calum Webb**
Sheffield Methods Institute, the University of Sheffield.
c.j.webb@sheffield.ac.uk

# Sign In

# Learning Objectives

What will I learn? | How does this week fit into my course?

By the end of this week you will:

- Understand how to interpret and report simple bivariate linear regression models.
- Be able to estimate simple bivariate linear regression models in `R` using the `lm()` function.
- Be able to plot a simple linear regression line using the `geom_smooth()` function in `ggplot()` (with `method = "lm"` argument).
- Be able to check whether data meets the assumptions of simple linear regression.

# Learning Objectives

What will I learn? | How does this week fit into my course?

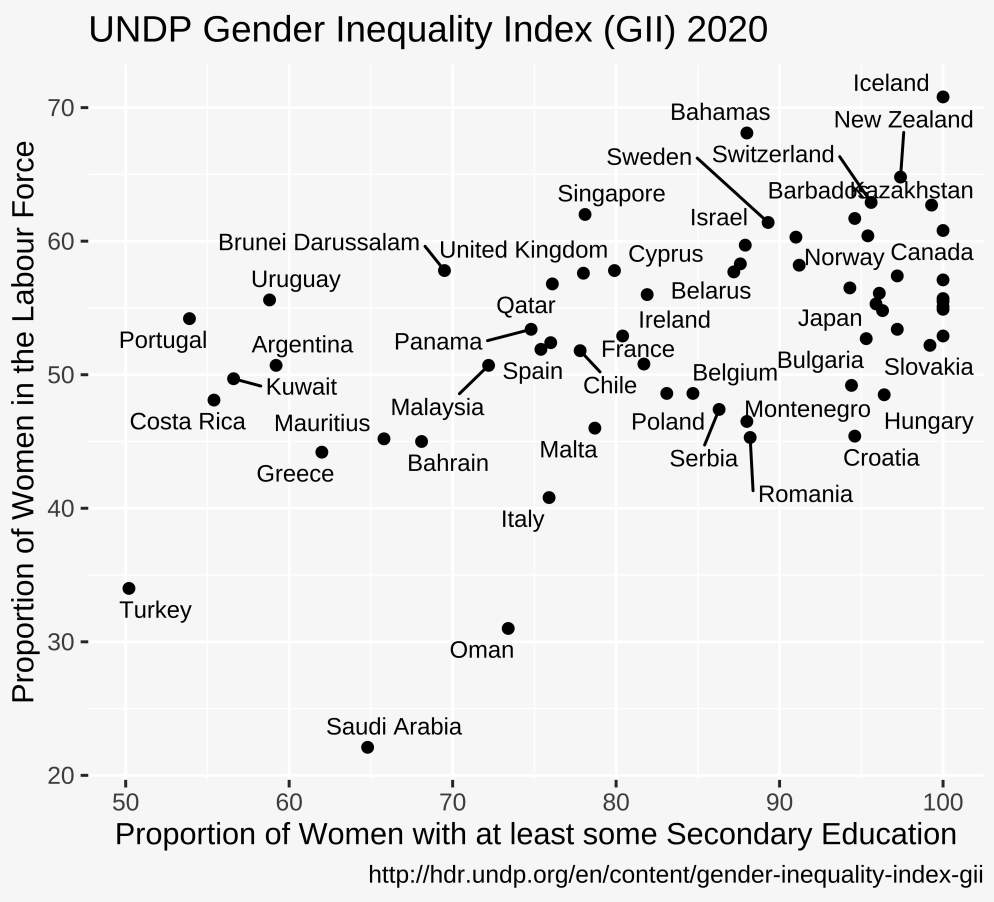
- Regression is the workhorse of contemporary quantitative social science research, and incorporates within it many of the concepts we've learned.
- This week will prepare you for extending bivariate linear regression to multiple linear regression (week 8), and logistic regression (week 9), to answer more complex (and more interesting!) social science research questions.
- A good understanding of regression will create a foundation for understanding cutting-edge advanced methods for longitudinal and multilevel models (In Advanced Quants module).

# What is linear regression?

# What is linear regression?

### Code for UN Data plot

```
library(tidyverse)
library(ggrepel)

un_data <- read_csv("un-gender-extract.csv")

un_data %>%
  ggplot() +
  geom_point(
    aes(x = sec_ed_women, y = lab_force_women)
    ) +
  geom_text_repel(
    aes(x = sec_ed_women, y = lab_force_women, label = country)
    size = 3) +
  ylab("Proportion of Women in the Labour Force") +
  xlab("Proportion of Women with at least some Secondary Educat
  labs(title = "UNDP Gender Inequality Index (GII) 2020",
       caption = "http://hdr.undp.org/en/content/gender-inequal
```



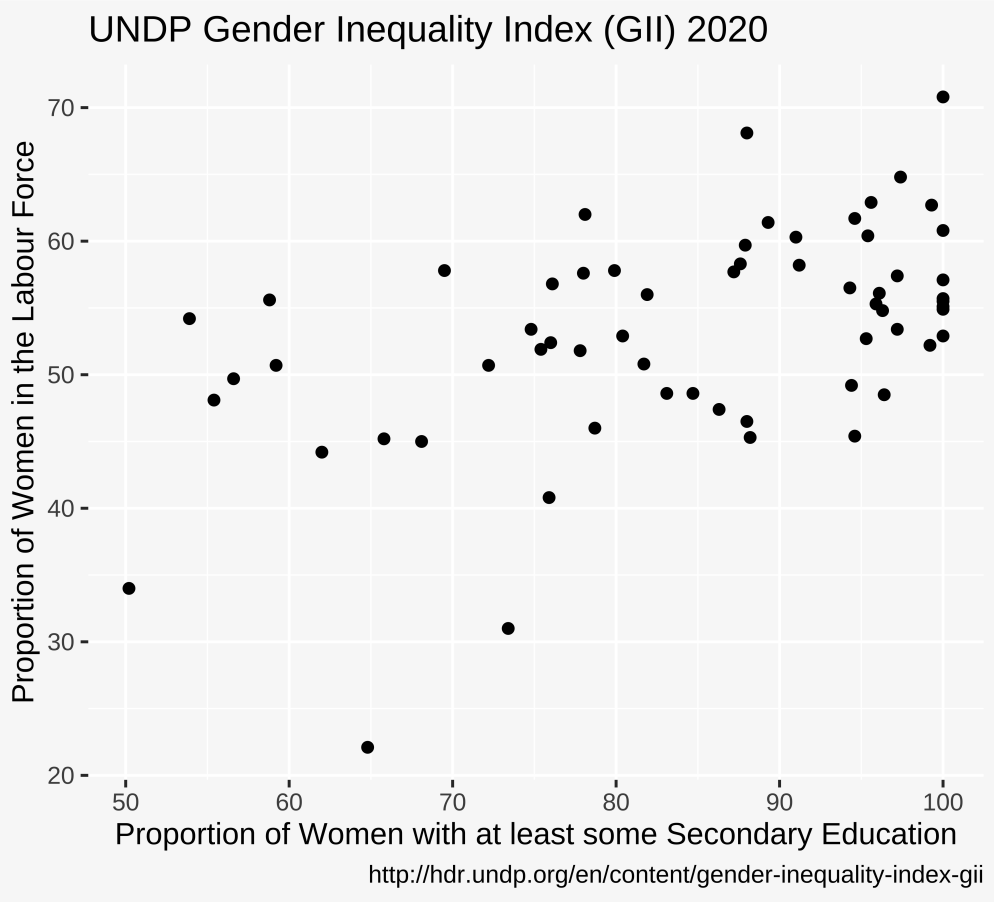
# What is linear regression?

**Which of the following statements is more informative?**

- There was a moderate correlation of R = 0.5 between rates of women completing at least some secondary education and women's labour market participation.

*Or*

- Every additional 1 percentage point of women completing at least some secondary education was associated with an increase in women's labour market participation of 0.3 percentage points.



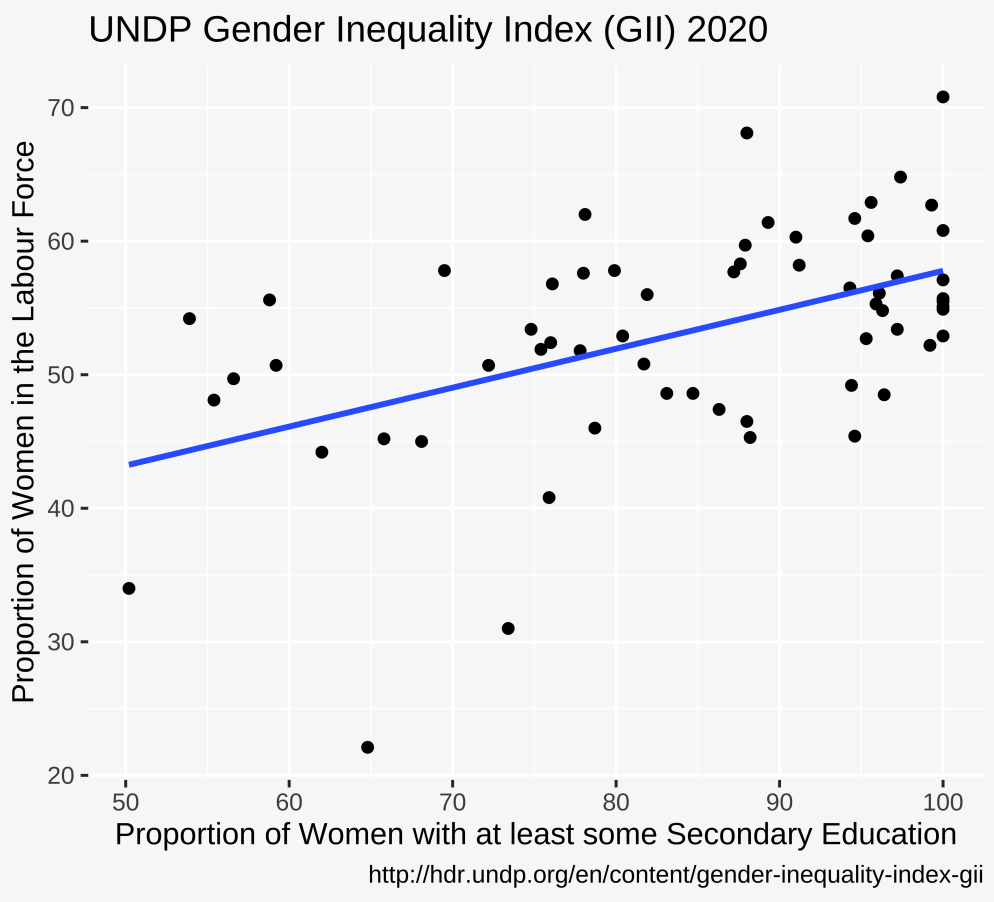
# What is linear regression?

**Which of the following statements is more informative?**

- There was a moderate correlation of R = 0.5 between rates of women completing at least some secondary education and women's labour market participation.

*Or*

- Every additional 1 percentage point of women completing at least some secondary education was associated with an increase in women's labour market participation of 0.3 percentage points.



$$y = mx + b$$

*Equation for a Straight line*

*"y is equal to m times by x (the slope) plus b (the intercept)"*

$$y = b + mx$$

*Shift the intercept over*

*"y is equal to b (the intercept) plus m times by x (the slope)"*

$$y = b_0 + b_1 x$$

*Change all estimates to b with a subscript to differentiate*

*"y is equal to $b_0$ plus $b_1$ times by x"*

$$\bar{y} = b_0 + b_1 x$$

*Put a bar over Y to indicate that it's the mean of Y, not the exact value.*

*"The mean of y is equal to $b_0$ plus $b_1$ times by x"*

$$\bar{y} = b_0 + b_1 x$$

*"The mean of y is equal to $b_0$ plus $b_1$ times by x"*

**Substitute our variables of interest**

Every additional 1 percent of women completing at least some secondary education was associated with an increase in women's labour market participation of 0.3 percentage points.

$$\bar{\text{women}}\text{LMP} = b_0 + b_1 x$$

*Change Y to be our dependent variable.*

*"The mean of women's labour market participation is equal to $b_0$ plus $b_1$ times by x"*

**Identify our dependent (y) variable**

Every additional 1 percent of women completing at least some secondary education was associated with an increase in **women's labour market participation** of 0.3 percentage points.

$$\bar{\text{women}}\text{LMP} = b_0 + b_1 \text{womenSER}$$

*Change X to be our independent variable.*

*"The mean of women's labour market participation is equal to $b_0$ plus $b_1$ times by women's secondary education rate"*

**Identify our independent (x) variable**

Every additional 1 percent of **women completing at least some secondary education** was associated with an increase in women's labour market participation of 0.3 percentage points.

$$\bar{\text{women}}\text{LMP} = b_0 + b_1\text{womenSER}$$

*Change X to be our independent variable.*

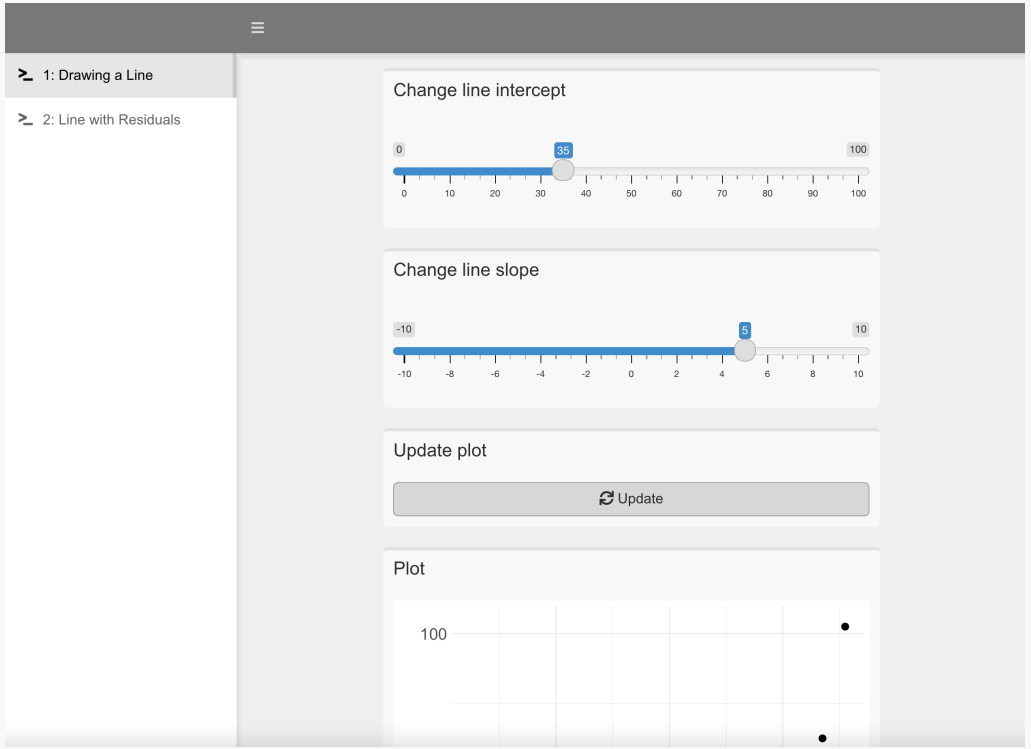
*"The mean of women's labour market participation is equal to $b_0$ plus $b_1$ times by women's secondary education rate"*

- Find $b_0$ ?
- Find $b_1$ ?

# Finding the best fitting line

**How do we find the best fitting line? A competition!**

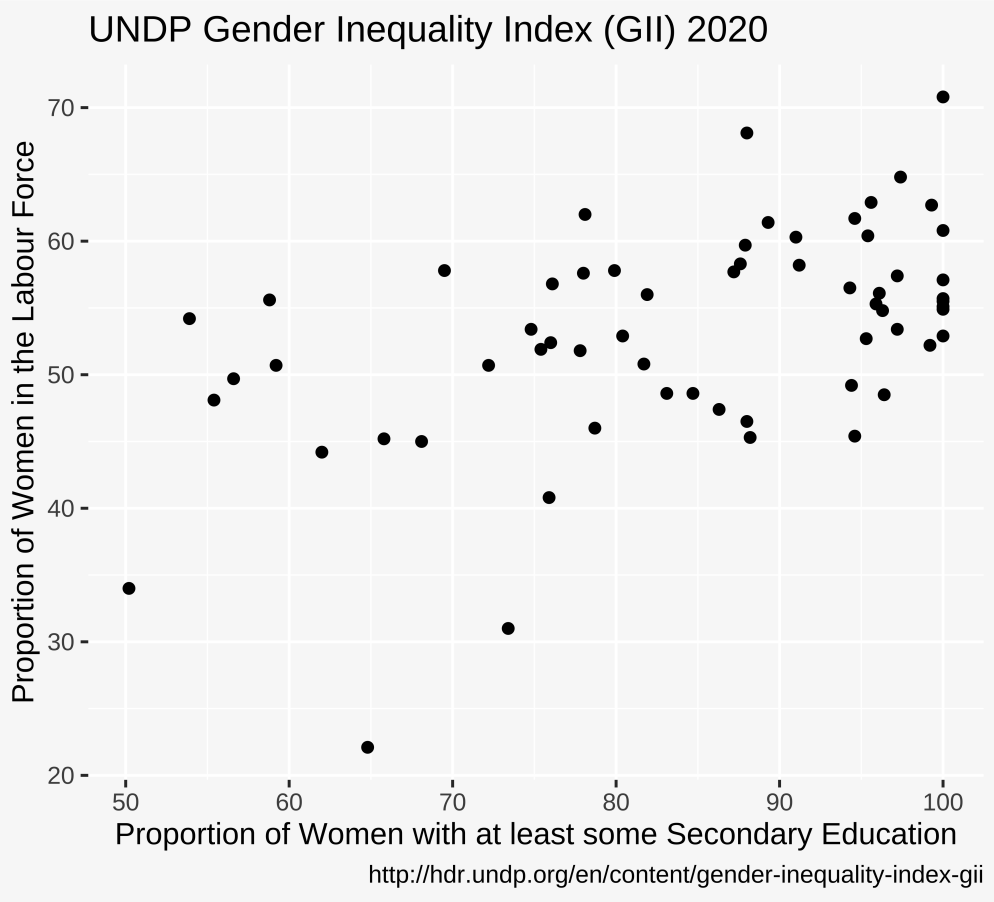
- Honour system - no cheating!
- Open the Shiny app in your web browser, or download and run the source code in R
- In pairs, your task is to get the best fitting line to the data that the app has generated for you by manipulating the intercept and slope values with the sliders.
- Once you've decided on the best fitting line, check the actual results from the linear regression and see how far off you were. Check the residuals plot and see how it changes between your estimate and the linear regression one.
- Whoever gets the closest guess wins!




https://webb.shinyapps.io/find-line/

# Finding the best fitting line with Ordinary Least Squares

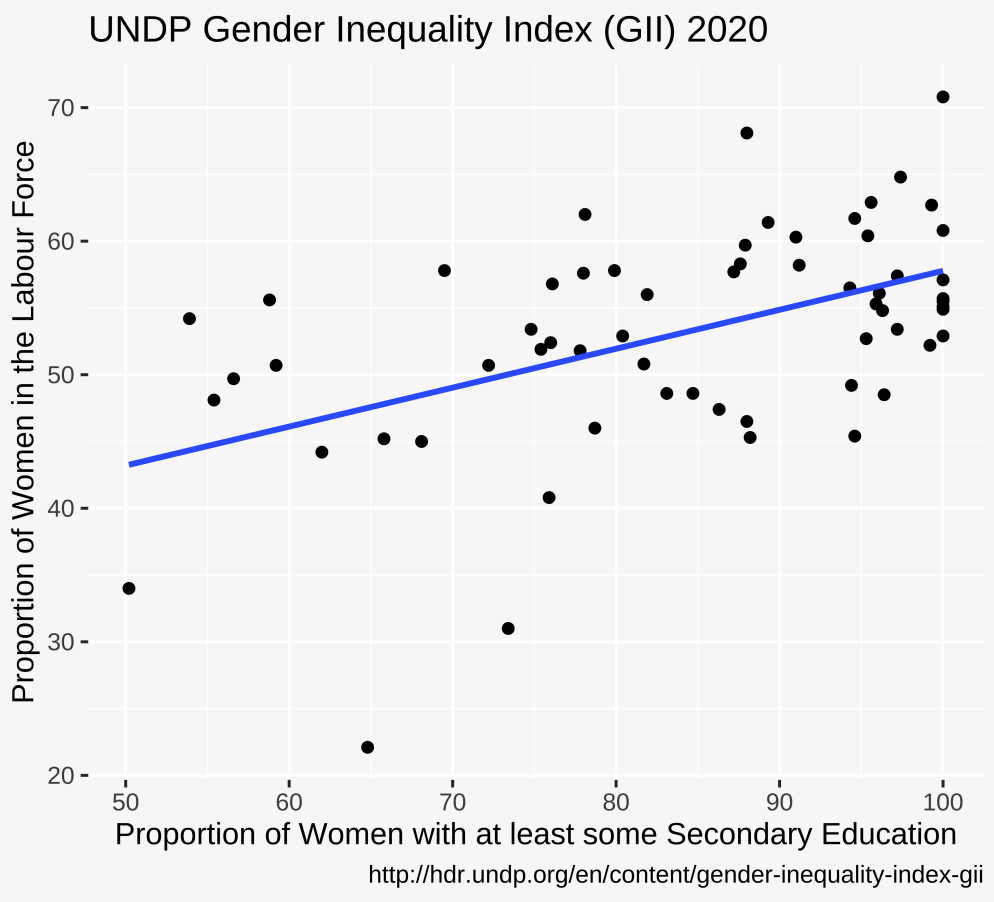
The slope and intercept of the line can be found using the **Ordinary Least Squares regression estimator**, which **minimizes the sum of the squared residuals**.



# Finding the best fitting line with Ordinary Least Squares

The slope and intercept of the line can be found using the **Ordinary Least Squares regression estimator**, which **minimizes the sum of the squared residuals**.

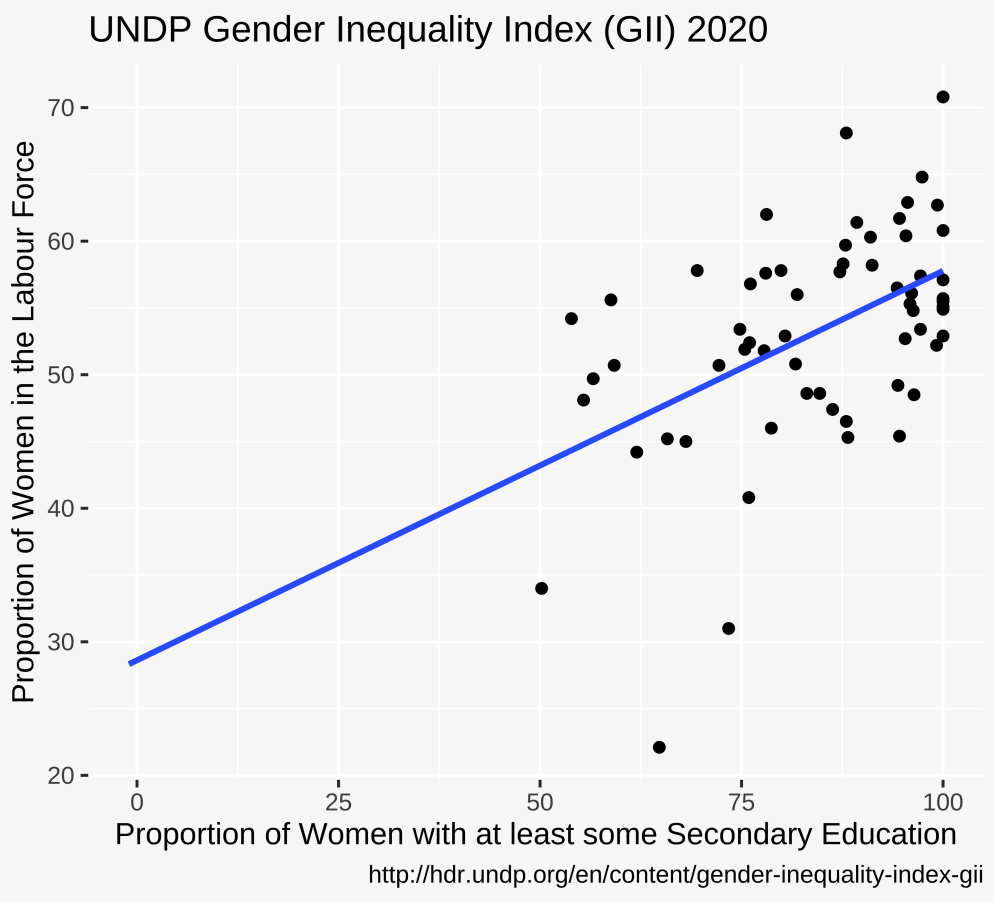
- This gives us the **slope** of the line (how much, on average, does Y change for every one-unit increase in X)



# Finding the best fitting line with Ordinary Least Squares

The slope and intercept of the line can be found using the **Ordinary Least Squares regression estimator**, which **minimizes the sum of the squared residuals**.

- This gives us the **slope** of the line (how much, on average, does Y change for every one-unit increase in X)
- It also tells us the **intercept** of the line (the point at which the line crosses the Y-axis)



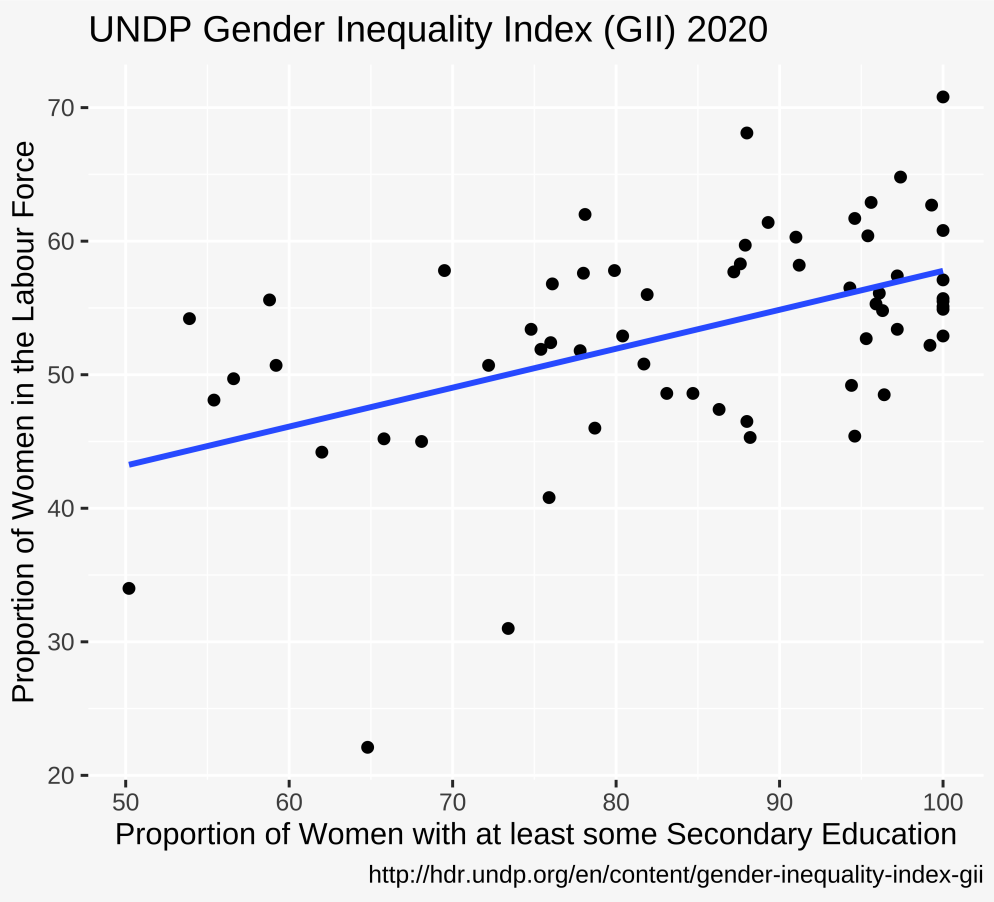
# Finding the best fitting line with Ordinary Least Squares

The slope and intercept of the line can be found using the **Ordinary Least Squares regression estimator**, which **minimizes the sum of the squared residuals**.

- This gives us the **slope** of the line (how much, on average, does Y change for every one-unit increase in X)
- It also tells us the **intercept** of the line (the point at which the line crosses the Y-axis)

We can estimate a **linear regression model** in `R` using the inbuilt `lm()` function.

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)
```



# How does the Ordinary Least Squares estimator work?

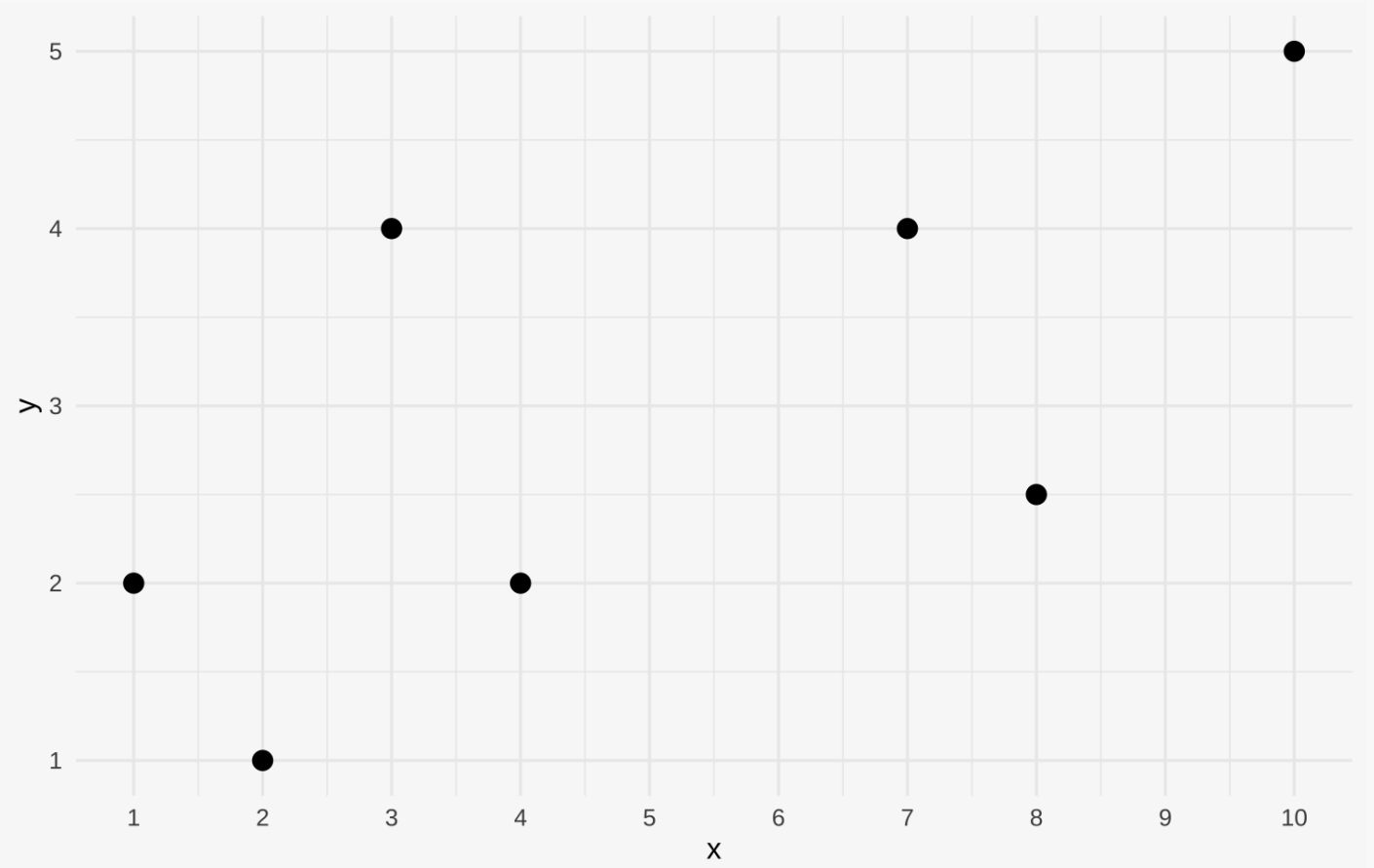
*You don't need to know this for any assessments, but you should at least understand generally how it works!*

# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Let's break down these equations into each of their steps so we can get an understanding of what they are doing.

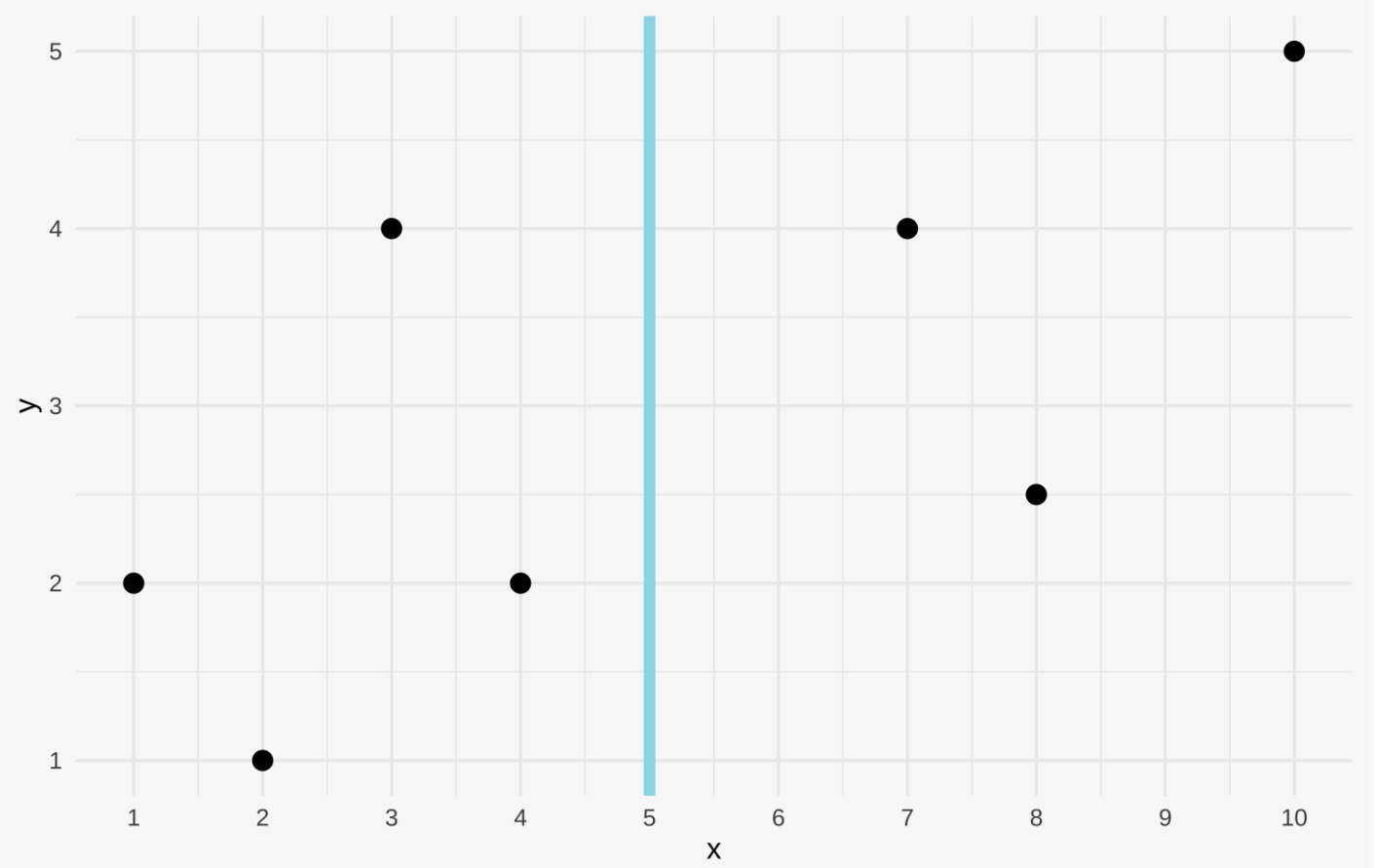


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

One of the terms in the equation is $\bar{x}$ (x bar or bar x), which is referring to the sample mean of the X variable.

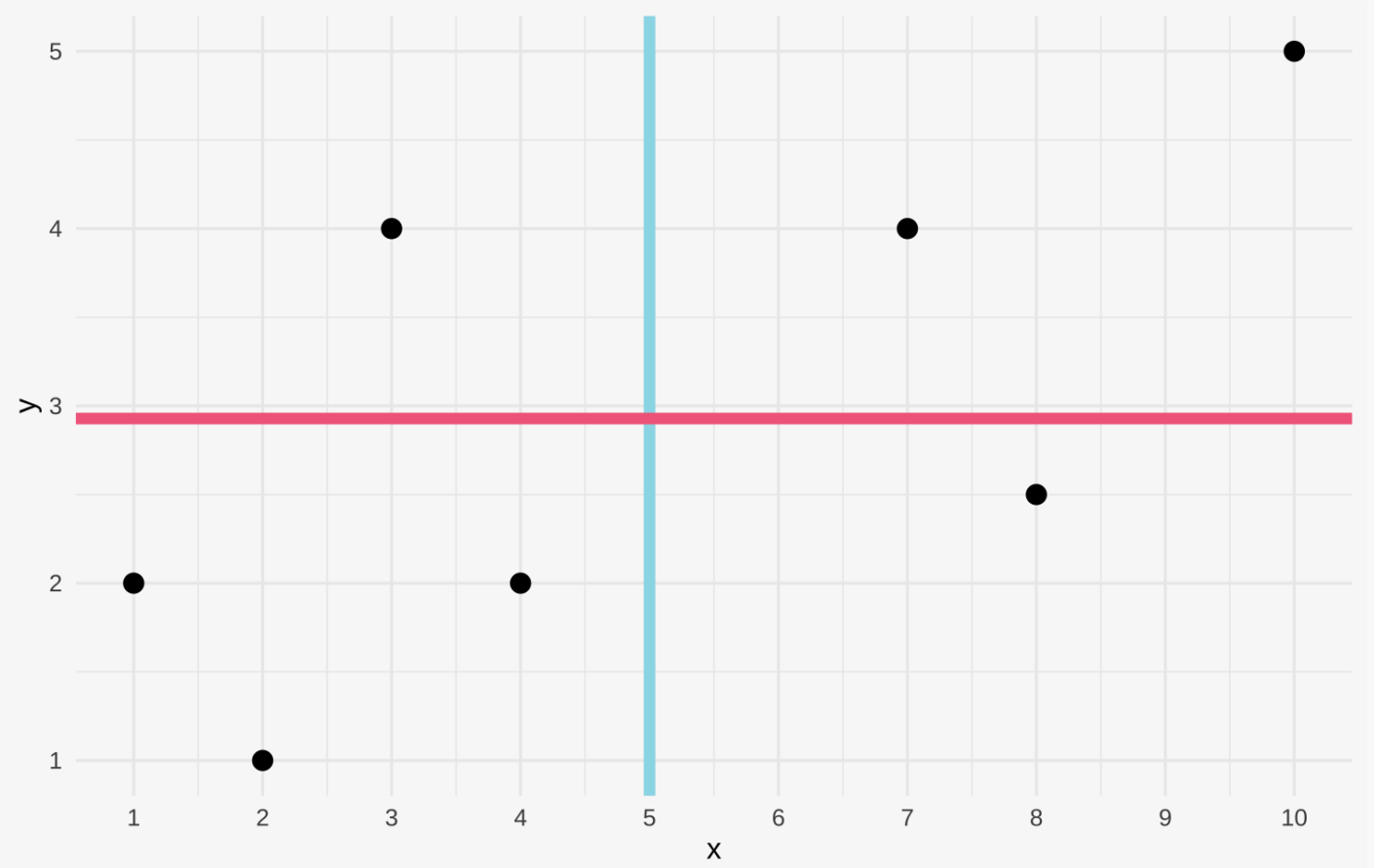


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The other bar, $\bar{y}$ (y bar), refers to the mean of the y variable. Our regression line will have to pass through the point where $\bar{x}$ and $\bar{y}$ meet.

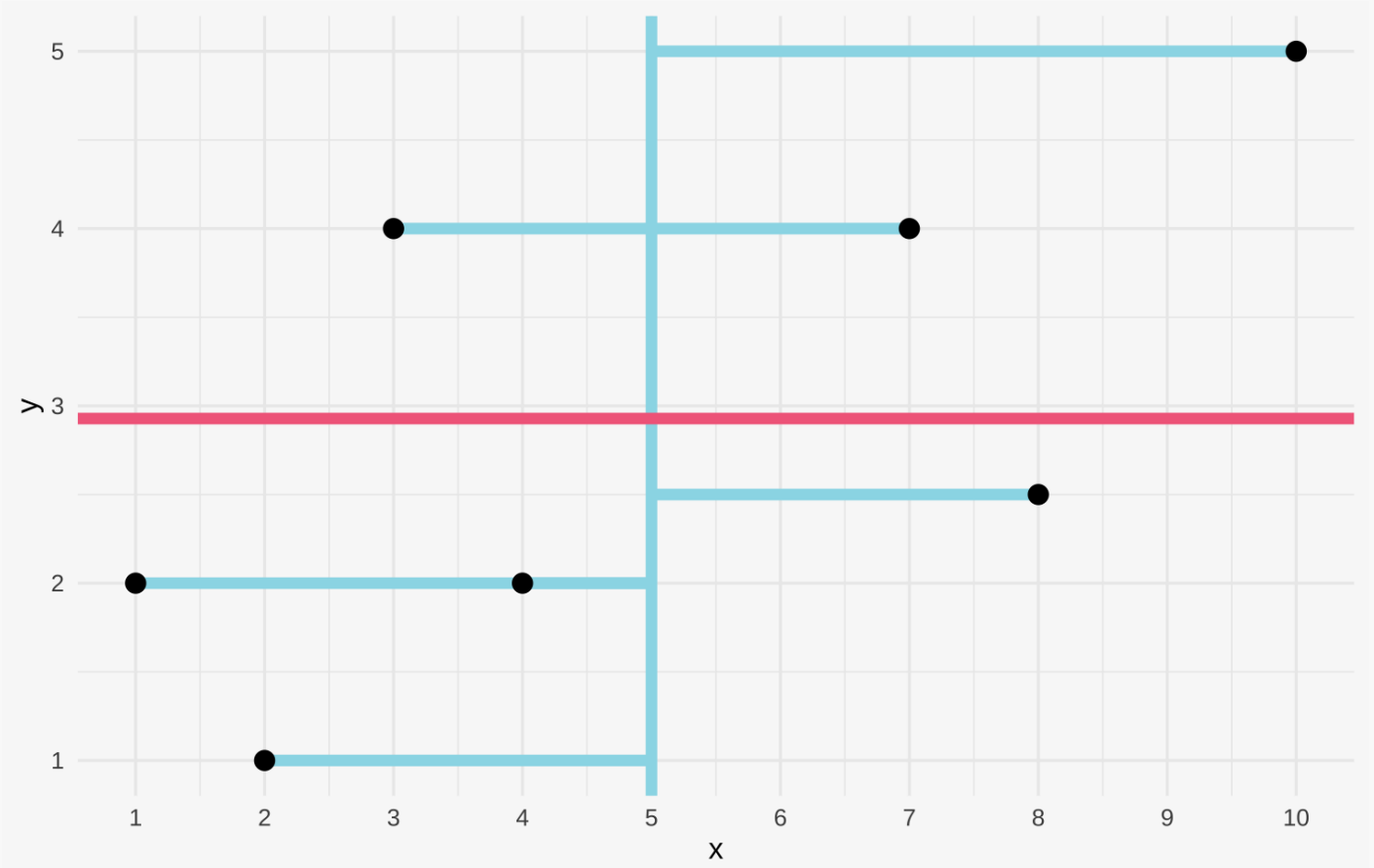


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We can see that in the numerator one of the terms is the sum ( $\sum_{i=1}^{n}$ ) of the difference between the mean of x ( $\bar{x}$ ) and the observed values of x $x_i$. This term is also squared in the denominator.

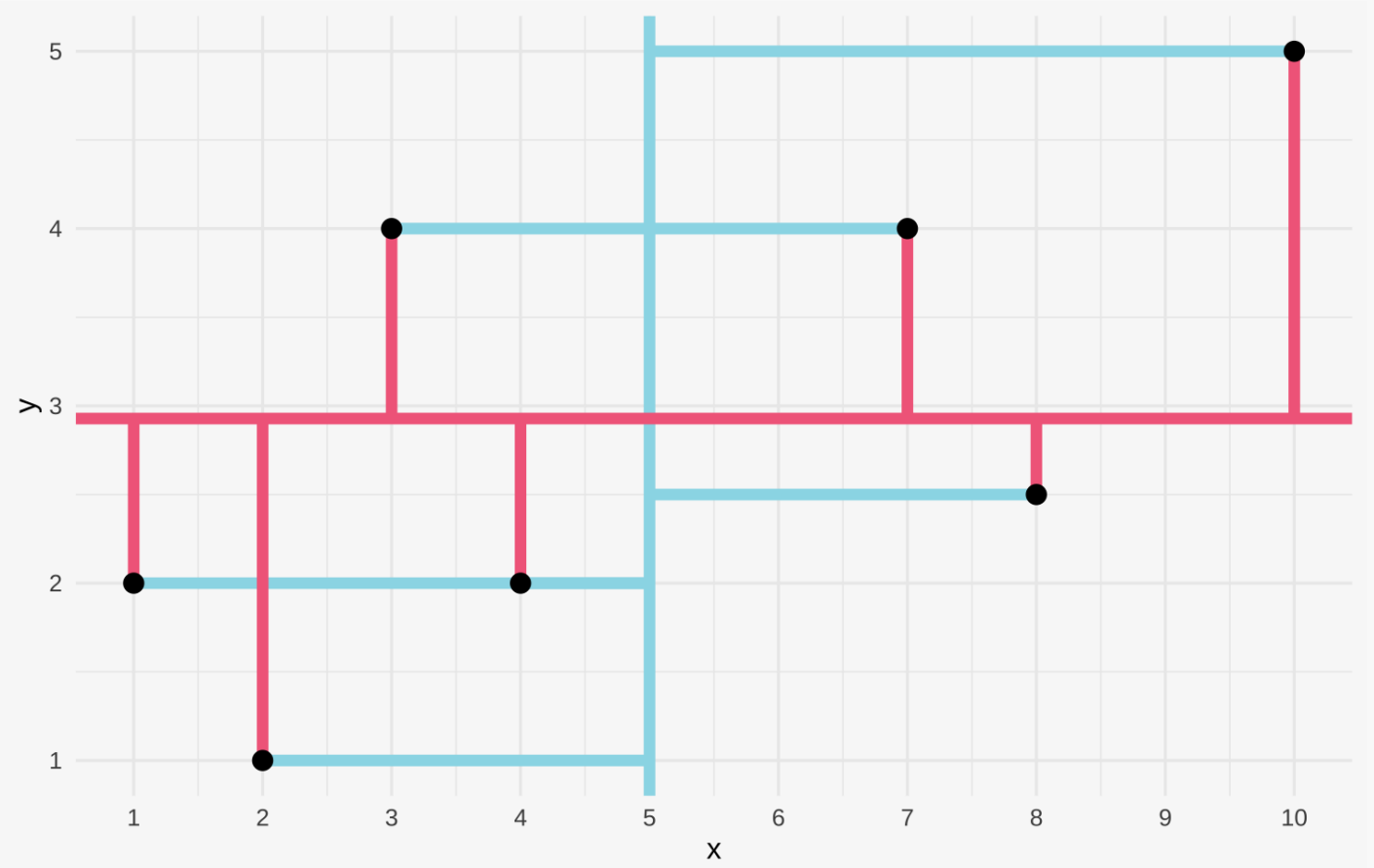


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Similarly, we also have a term for the equivalent for y - the difference between the mean of y $\bar{y}$ and the observed values of y $y_i$.

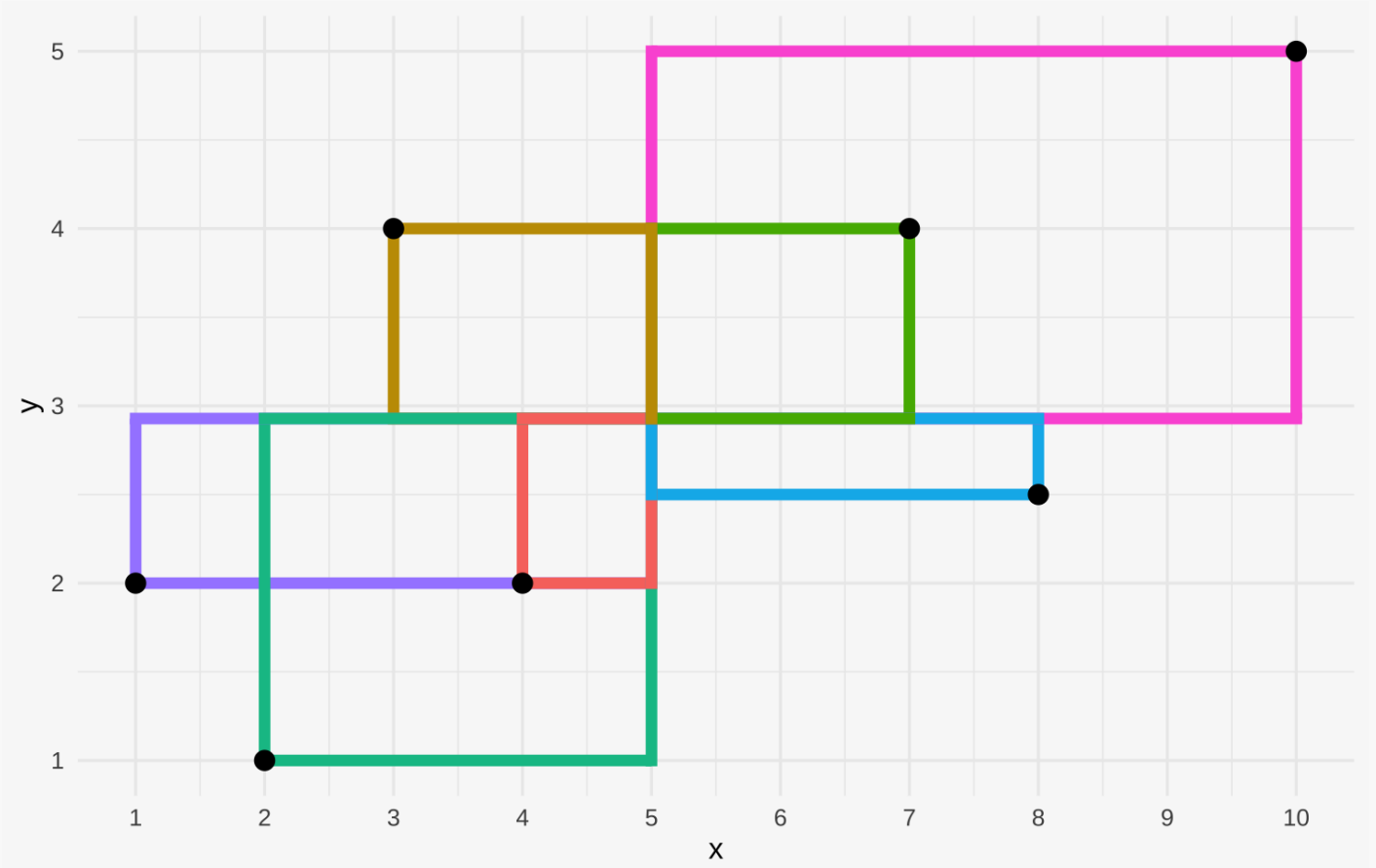


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Notice that when we take the product of $(y_i - \bar{y})$ and $(x_i - \bar{x})$ it gives us the area of a rectangle.

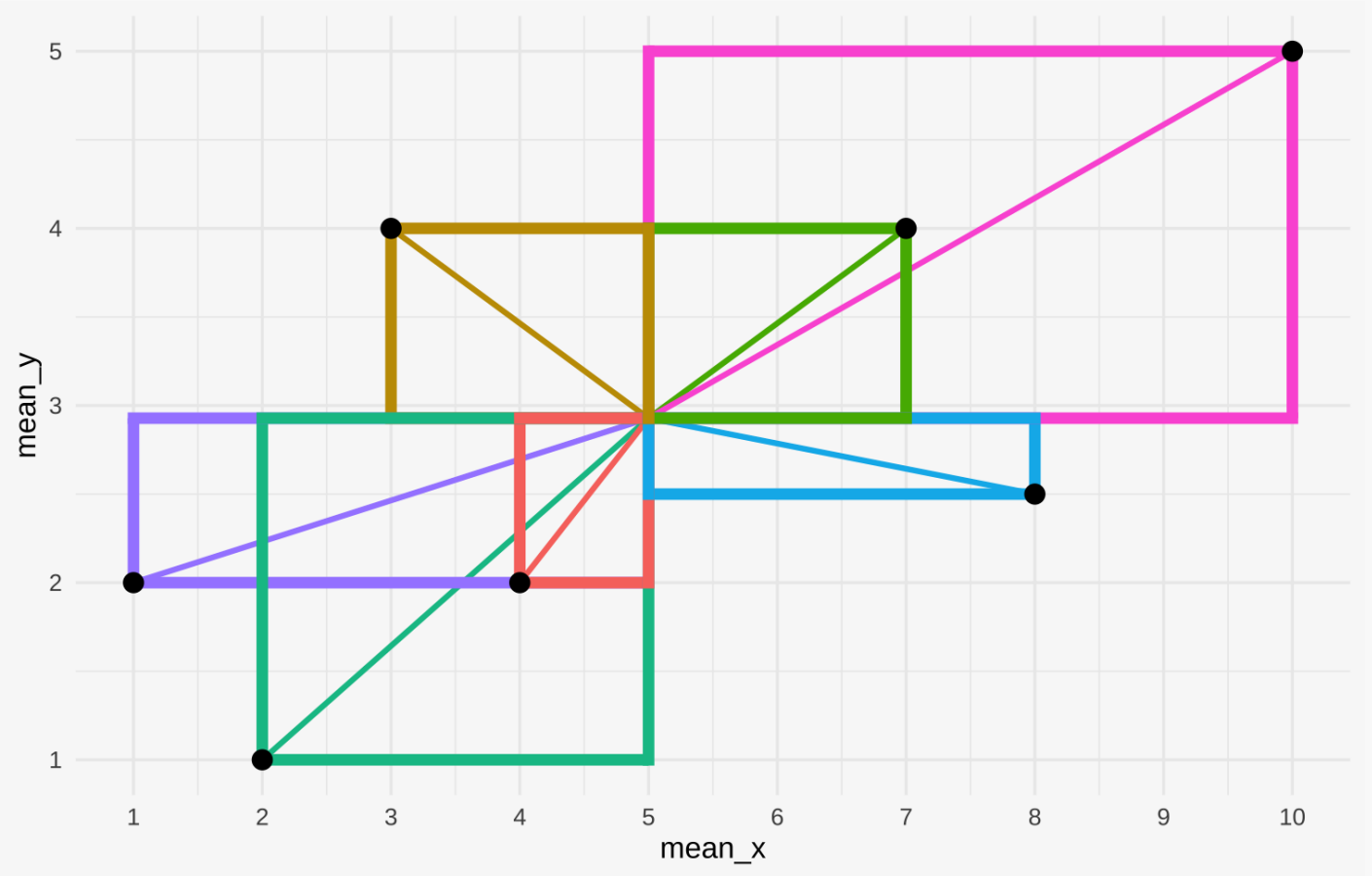


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

As a result, we can think about the distance from the mean of y and each specific point as a unique incidence of the change in y, relative to its mean, for an increase or decrease in X of however far away it is from the mean of X.

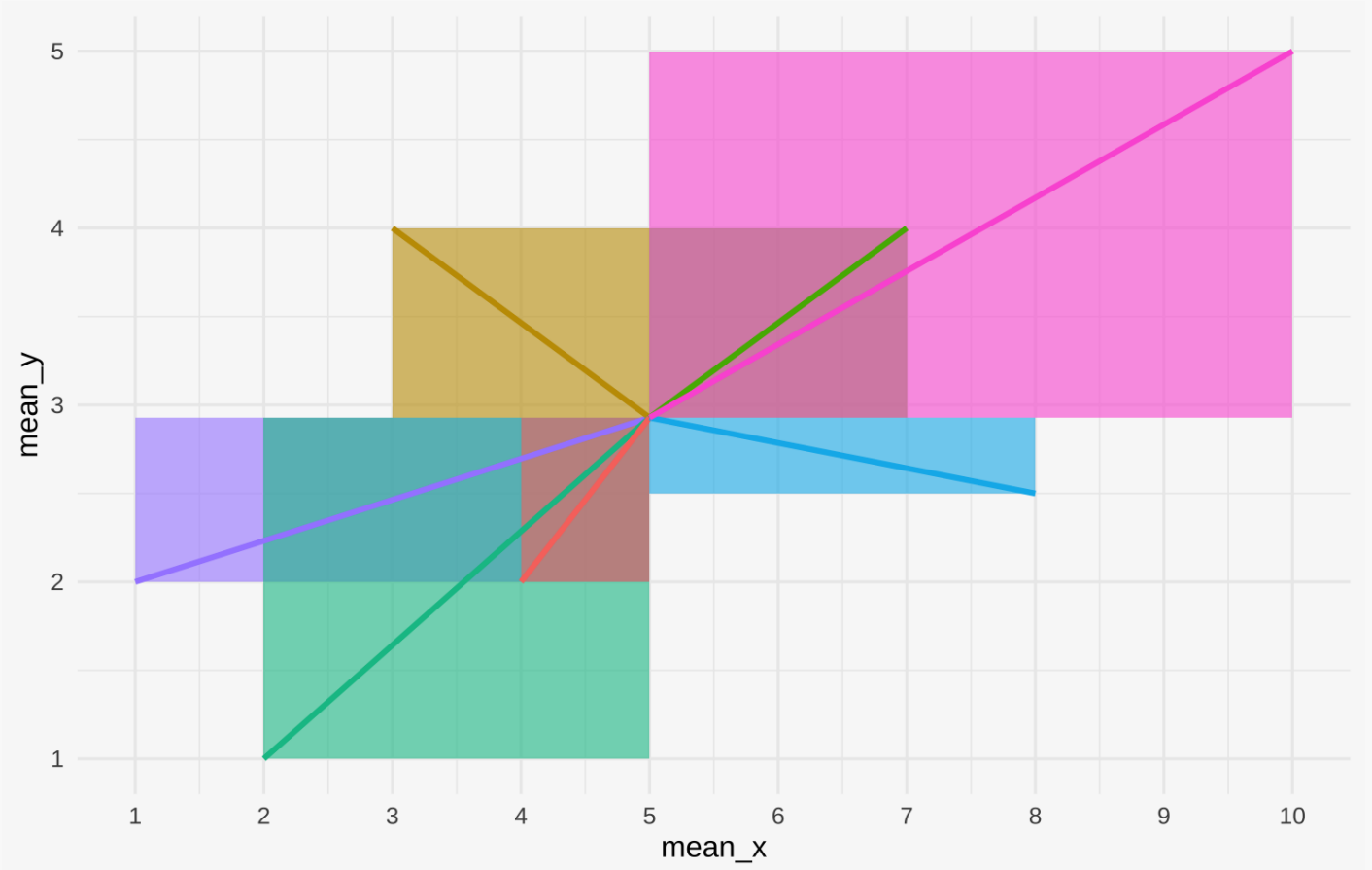


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Because the rectangles are overlapping, I'm going to seperate them out now but keep them to scale to make what happens when we work through the fraction clearer.

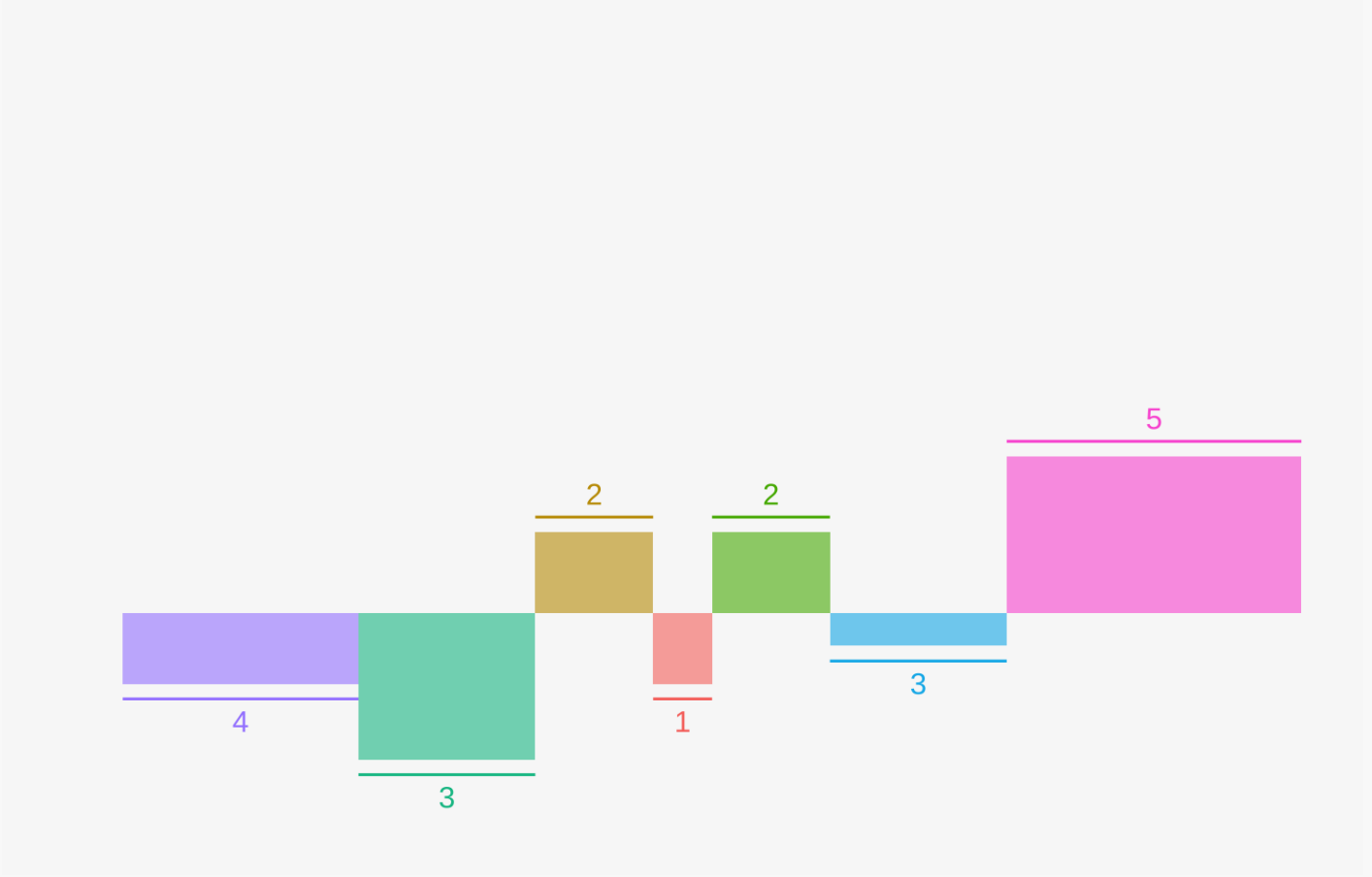


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now we have each of the squares and the distances of X that they cover (keep in mind that half of these will be negative, but will cover the same distance in the other direction).

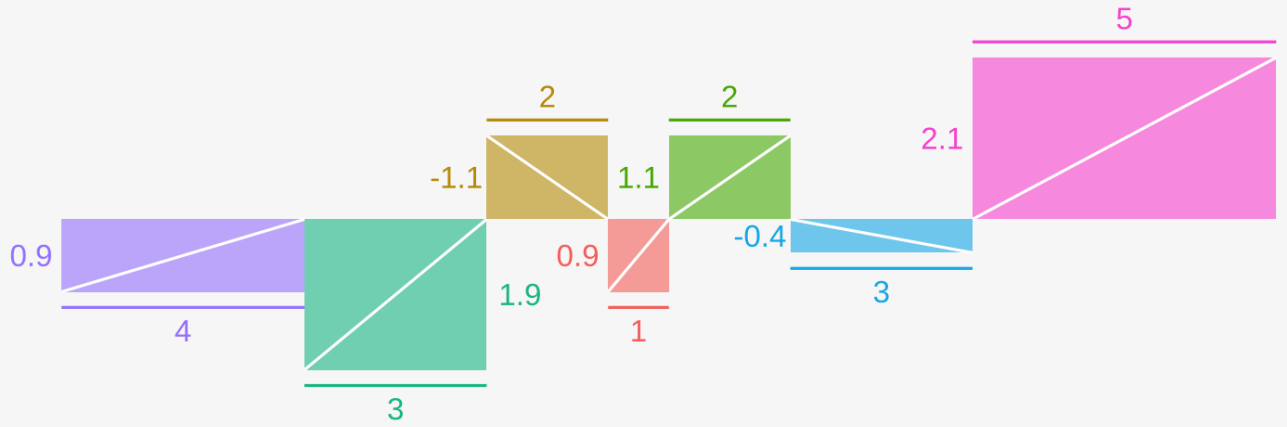


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We also have the changes from the mean of Y that each of those distances correspond with

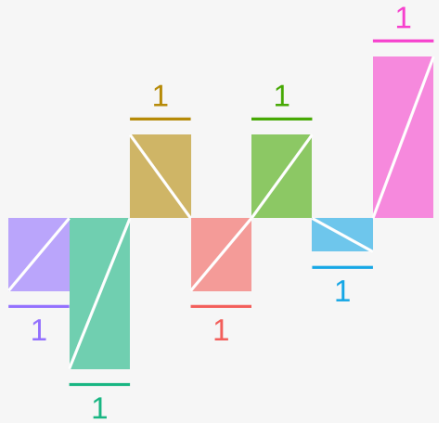


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now, watch what happens to the distances when we divide by one of the $x - \bar{x}$ terms, removing it completely from the numerator and removing one of the powers of the denominator.



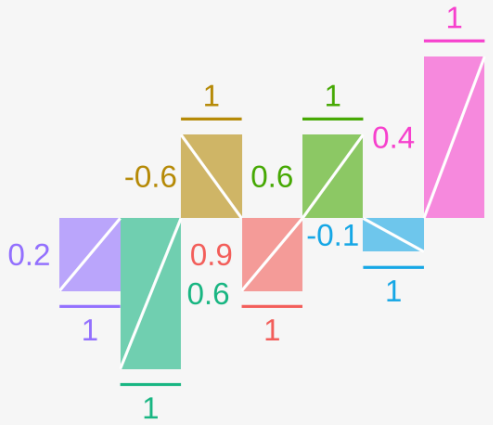
# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Of course, we now also need to re-scale the distances in Y; if the X distance previously was 4 then the Y distance for a 1-unit increase of X will be whatever it originally was divided by 4.

Now all of our changes are on the same scale.



# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

If we take the average of all of these scaled changes, we end up with the average change in **Y for a one-unit increase in X**: the definition of a slope in a linear equation!

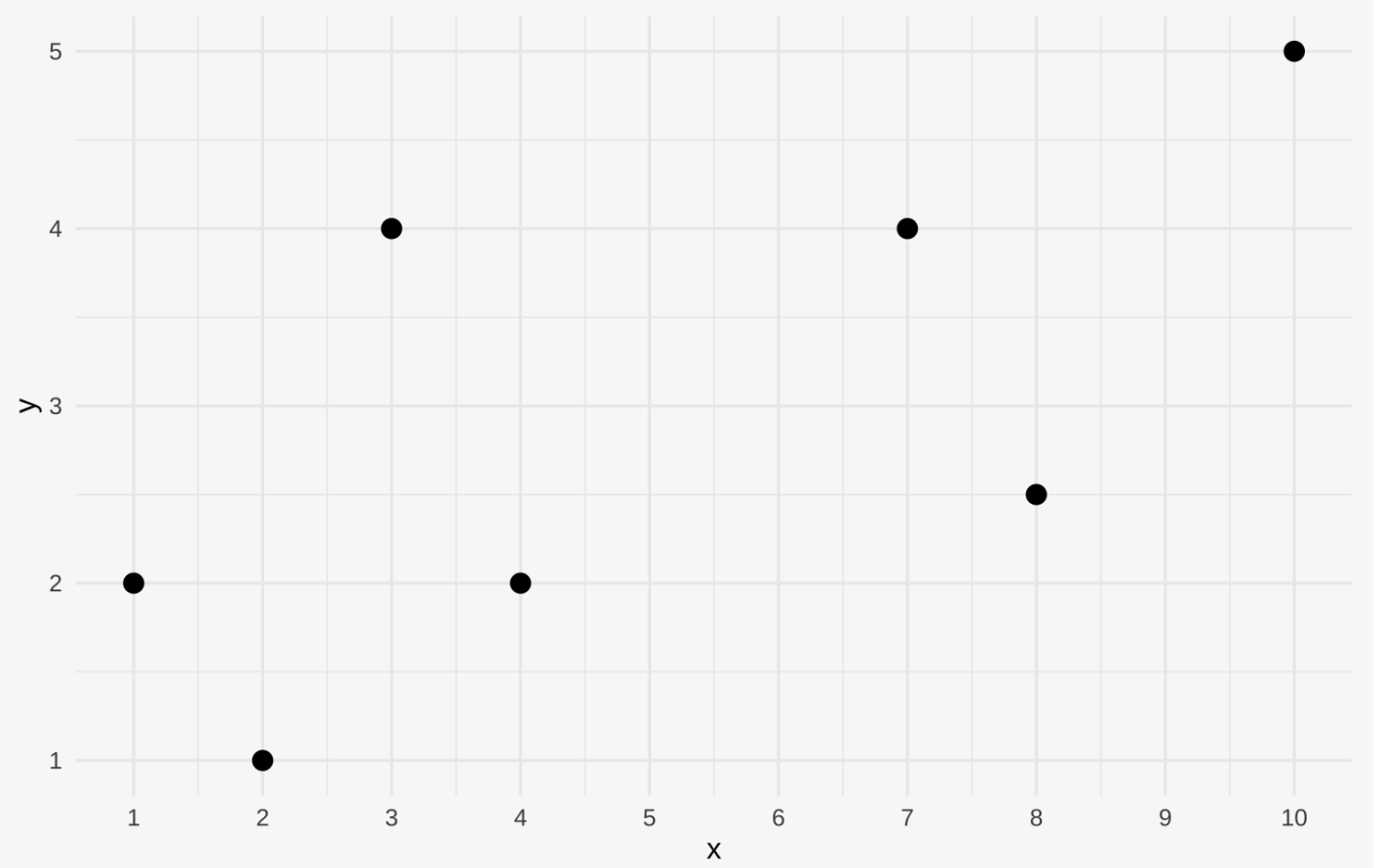
$$\frac{1}{} \quad 0.29$$

# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

But the slope is only one part of the equation! How does the second equation get us the intercept (the point at which we can start drawing the line). Without this we could draw the line...

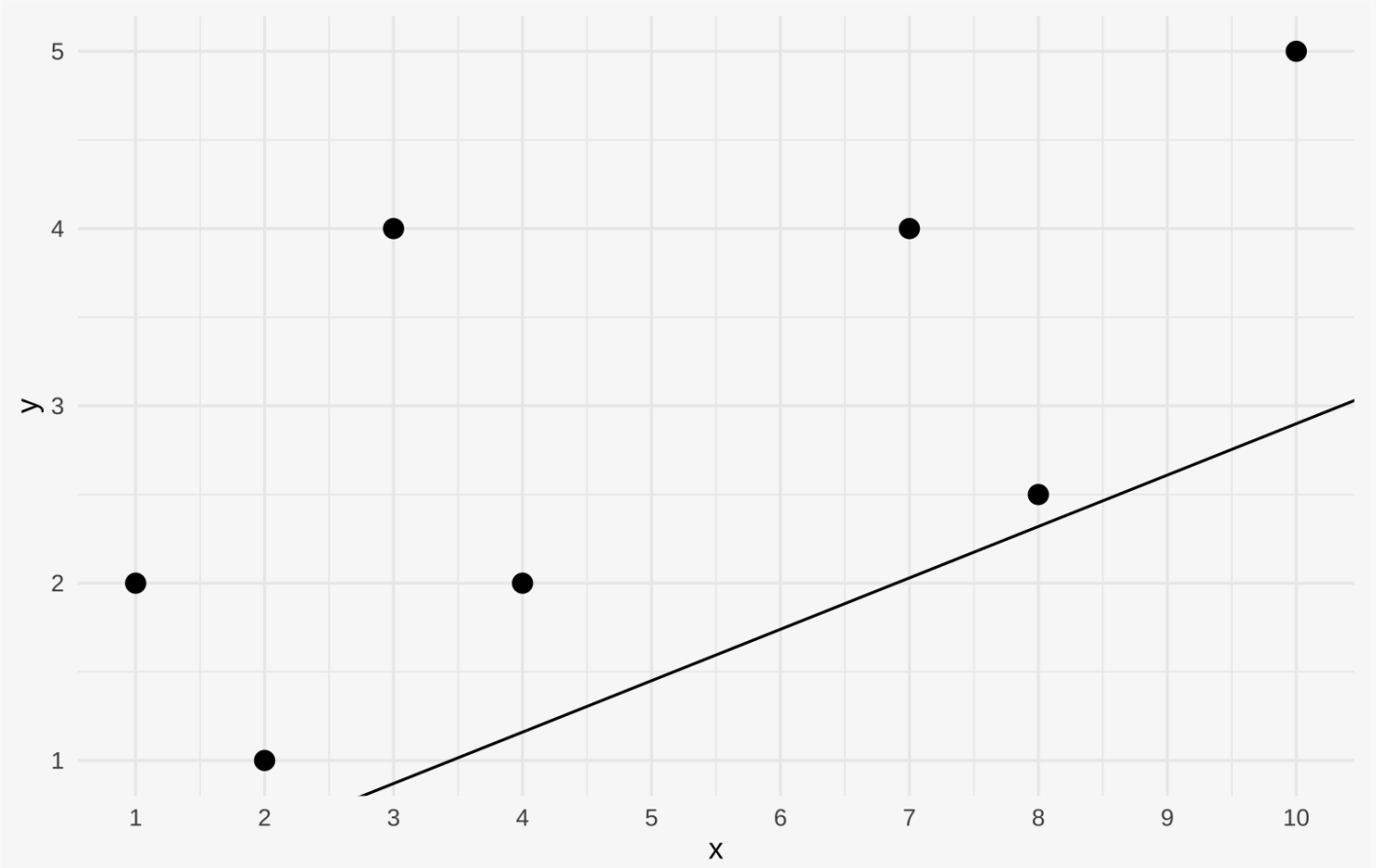


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Here?

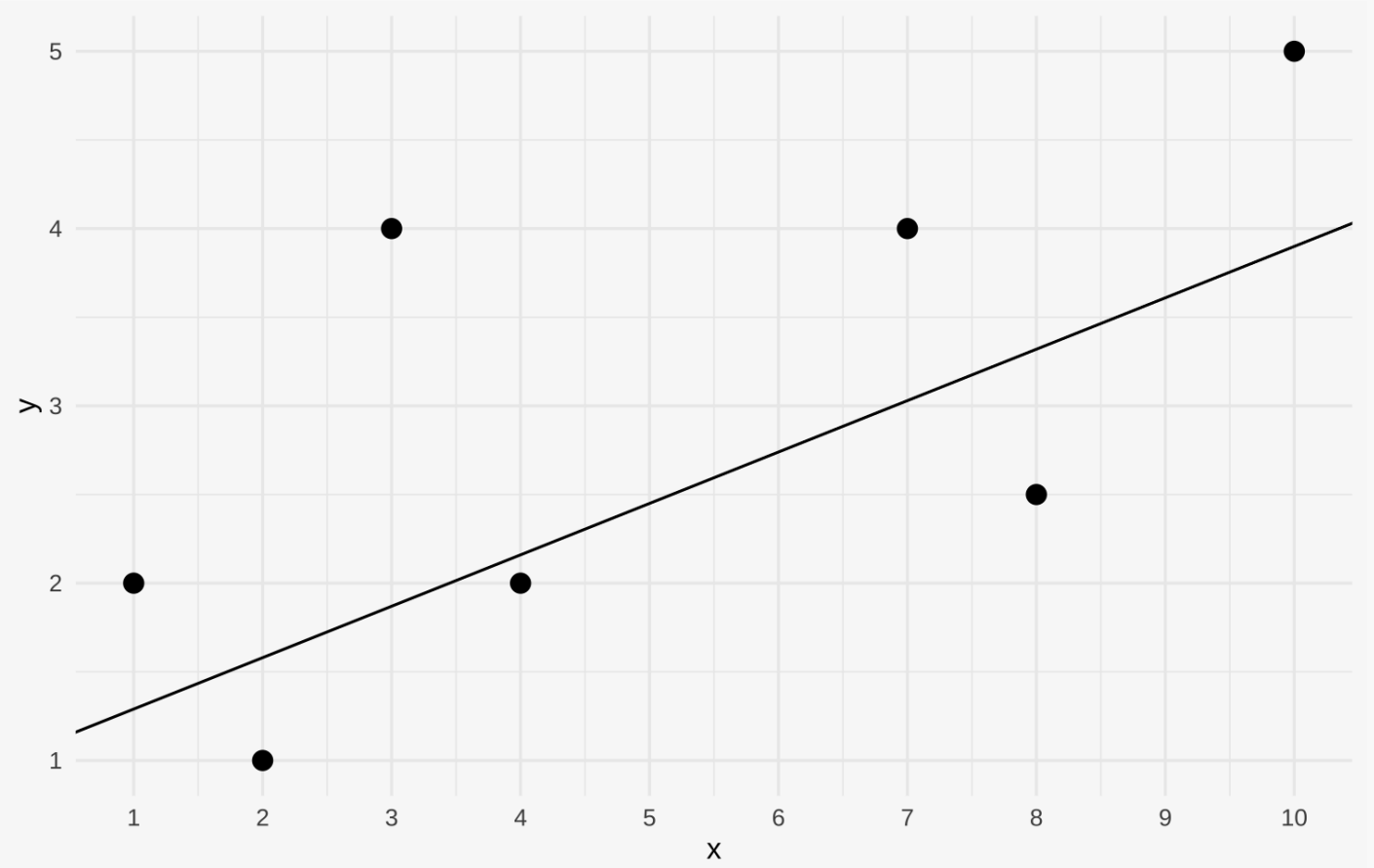


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Or here?

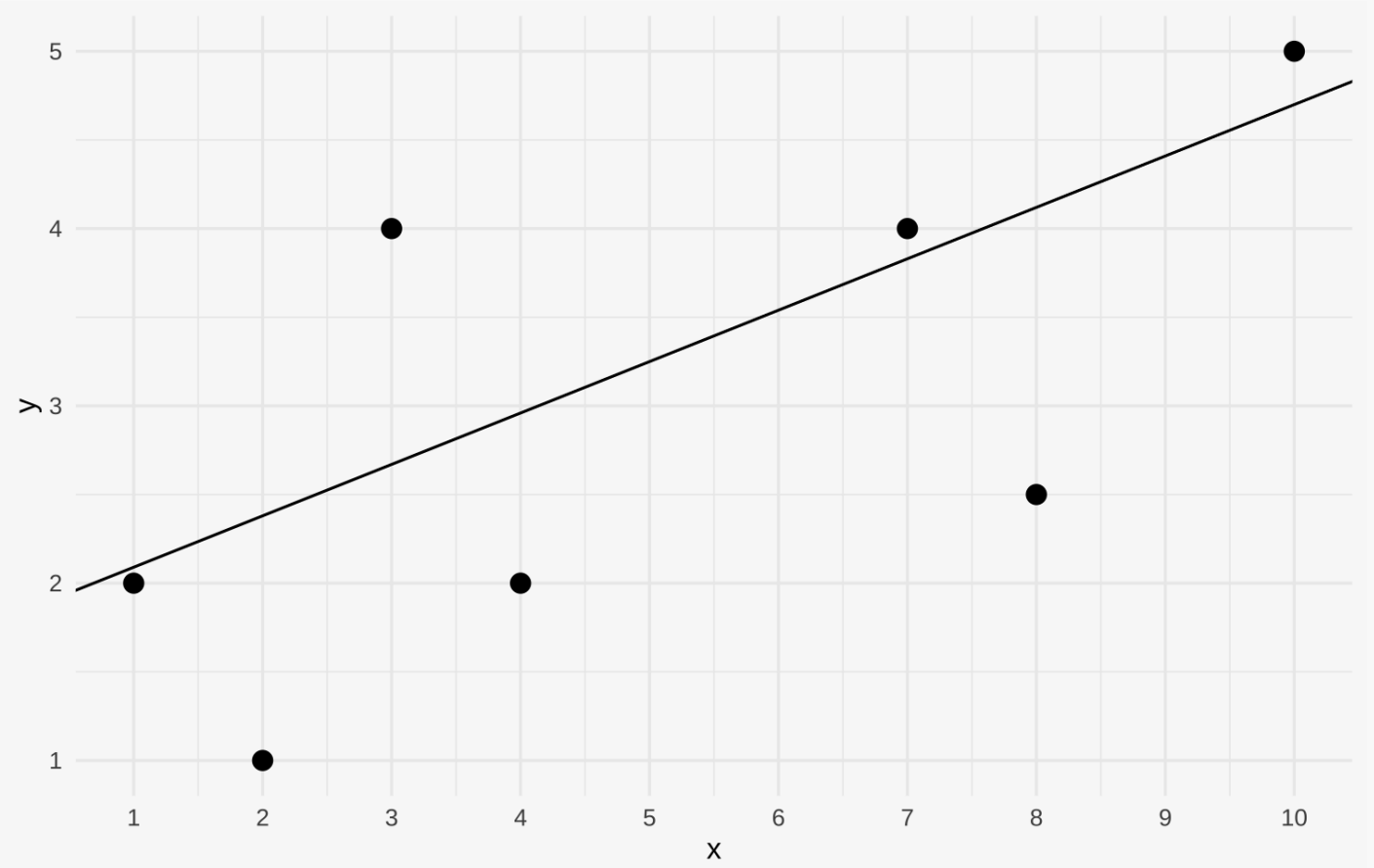


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Or here??

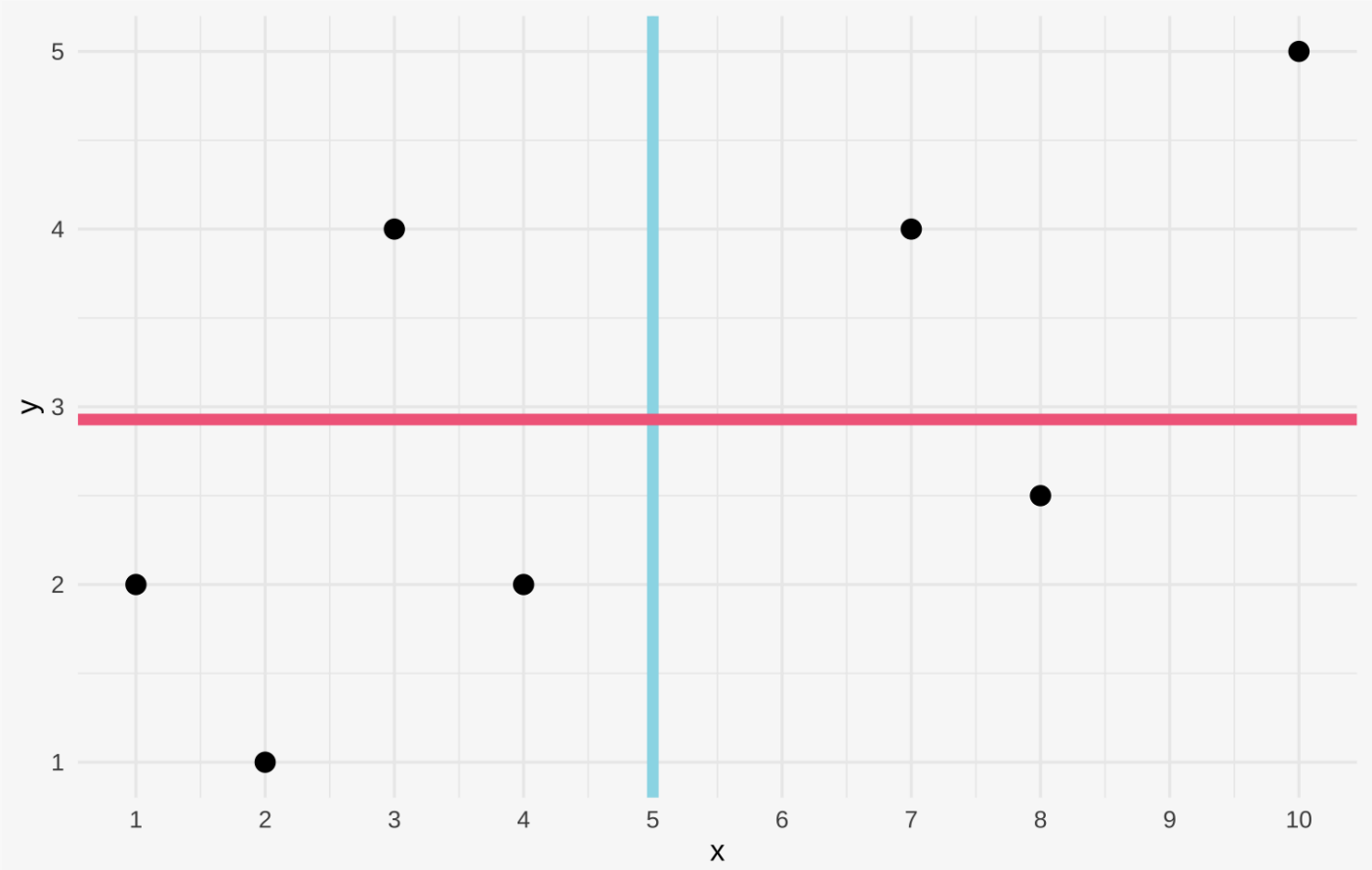


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Well, if you remember from earlier, we said that we know the line *must* cross the point where the mean of x and the mean of y meet.

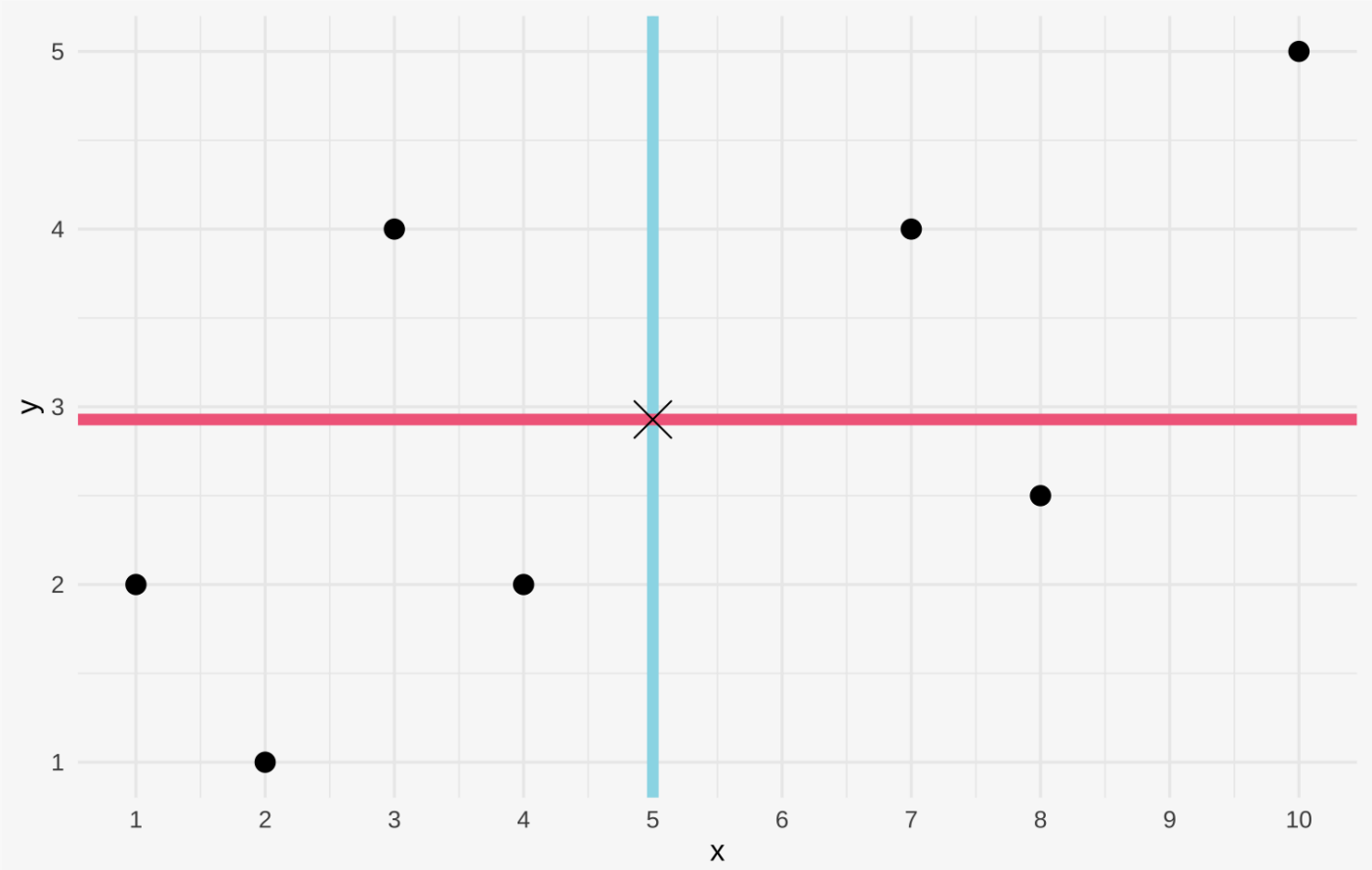


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

So what if we take that as a starting point and then **work backwards from the mean until we get to X = 0, each time also subtracting the slope value from the mean of Y**.

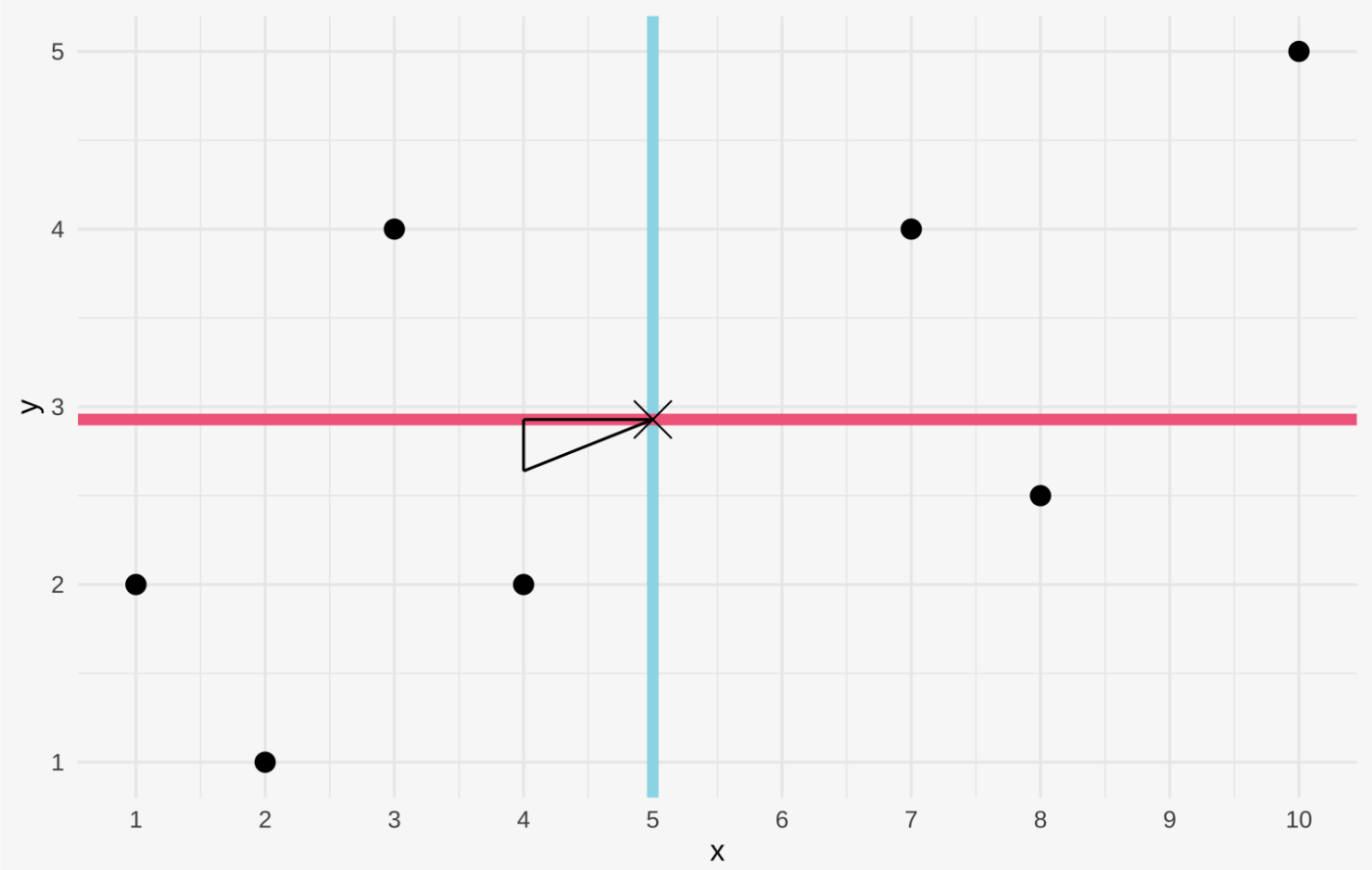


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

One step...

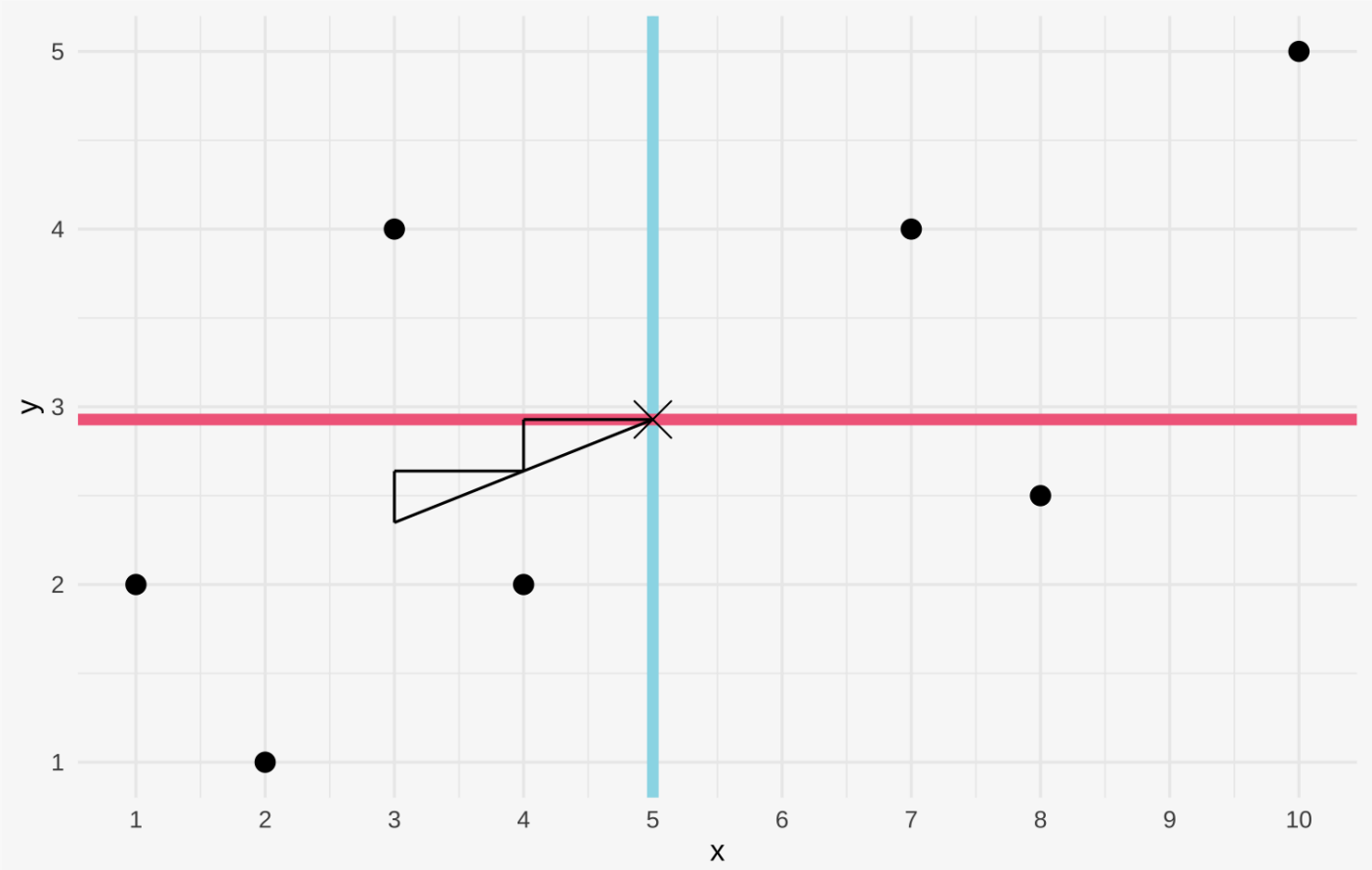


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Two steps...

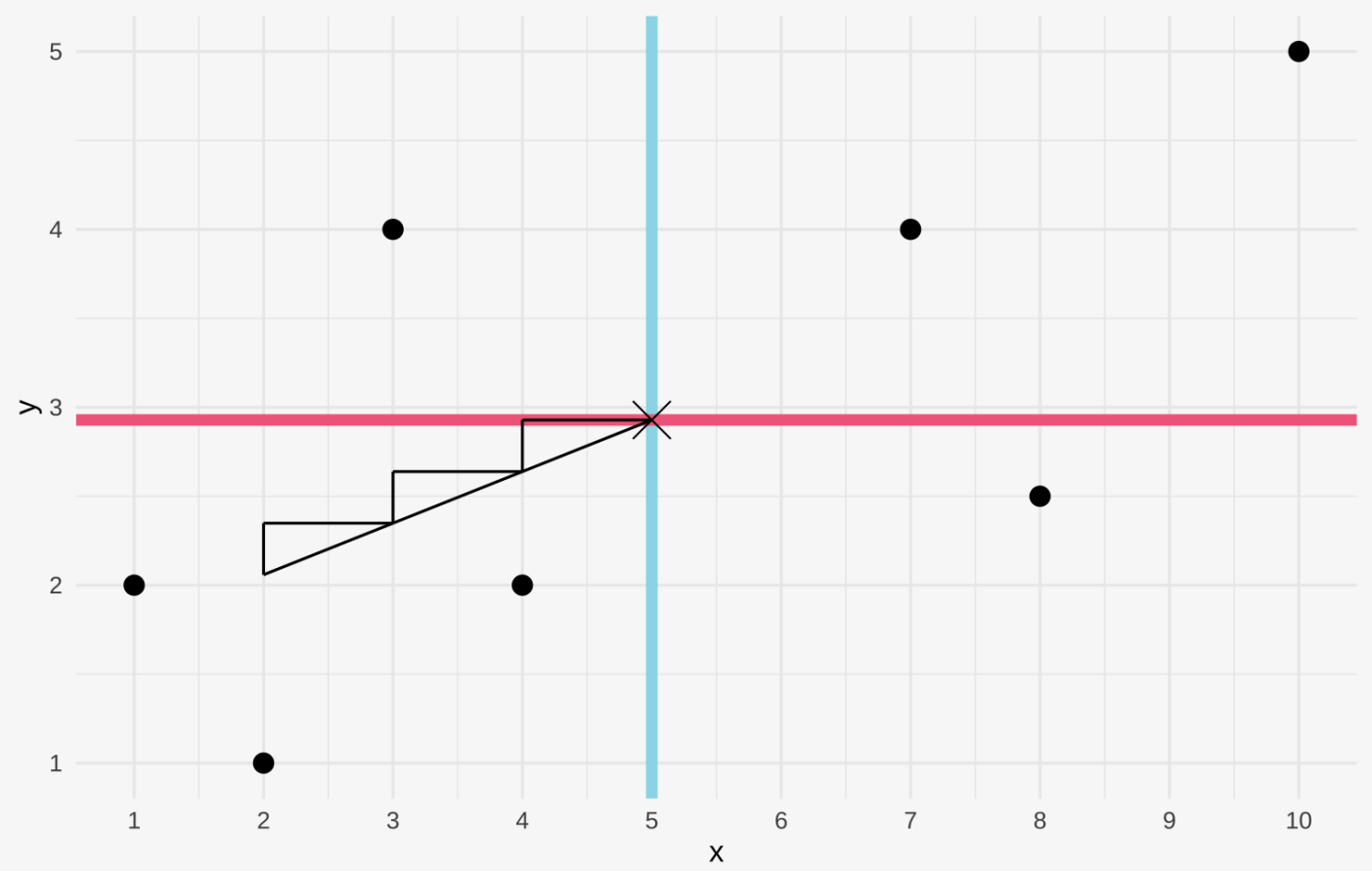


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Three steps...

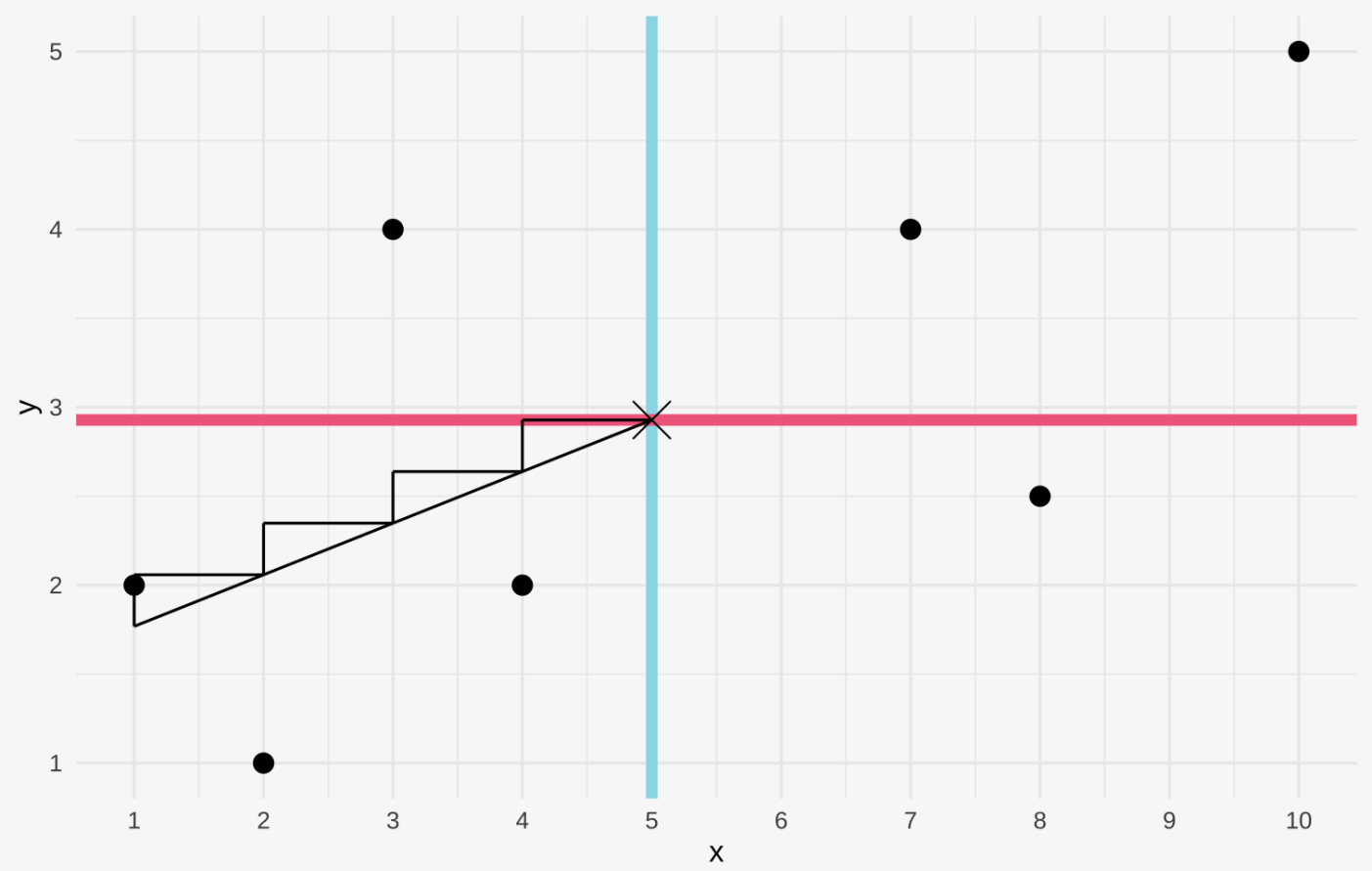


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Four steps...

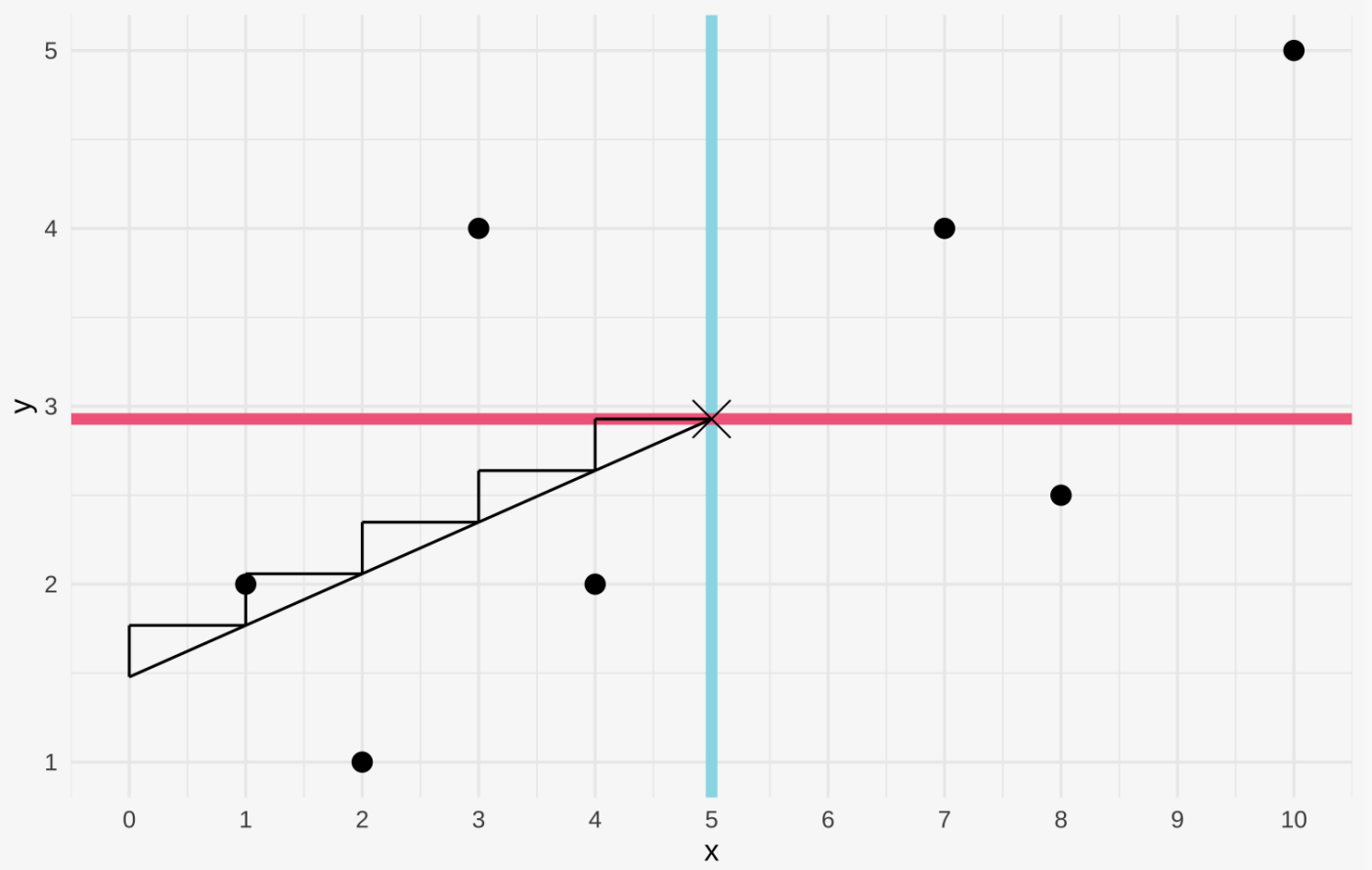


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And now our regression line is in contact the point at which X = 0: this gives us the intercept, the other part of the equation, of around 1.48.

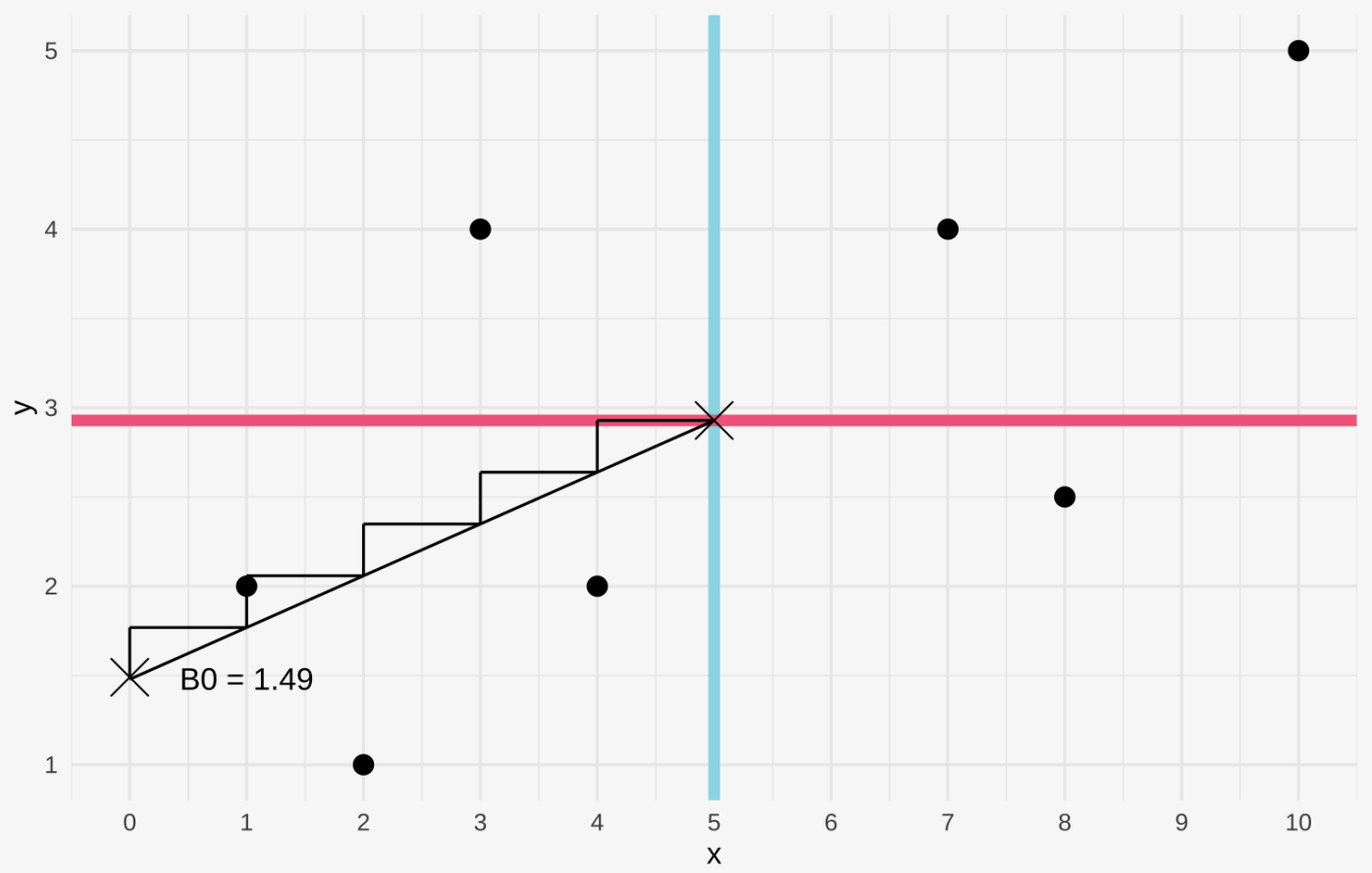


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And now our regression line is in contact the point at which X = 0: this gives us the intercept, the other part of the equation, of around 1.48.

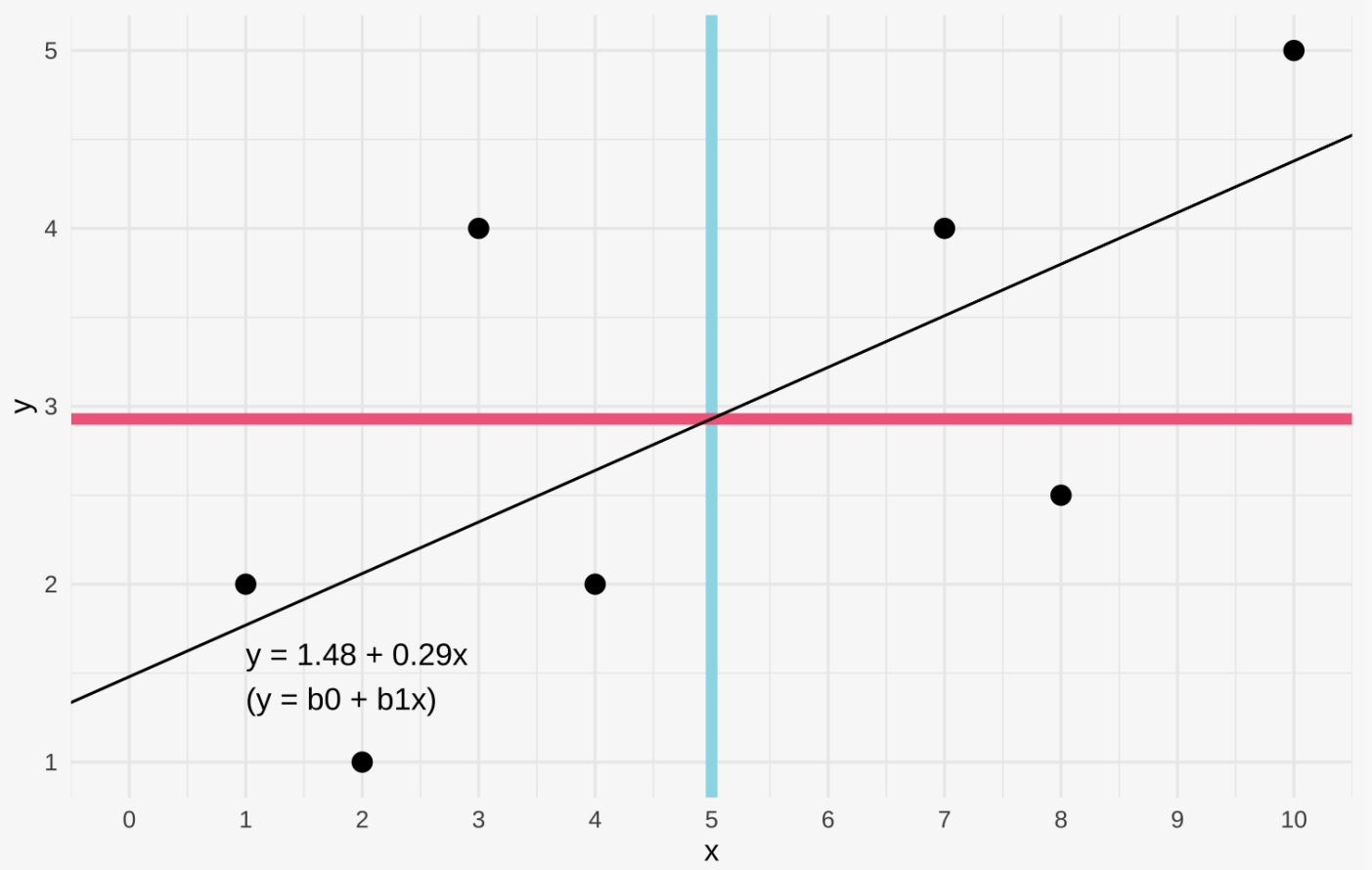


# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Even though the equations above look quite overwhelming, we can see visually exactly what they are doing to calculate the regression line. The result from those equations minimises the squared residuals.



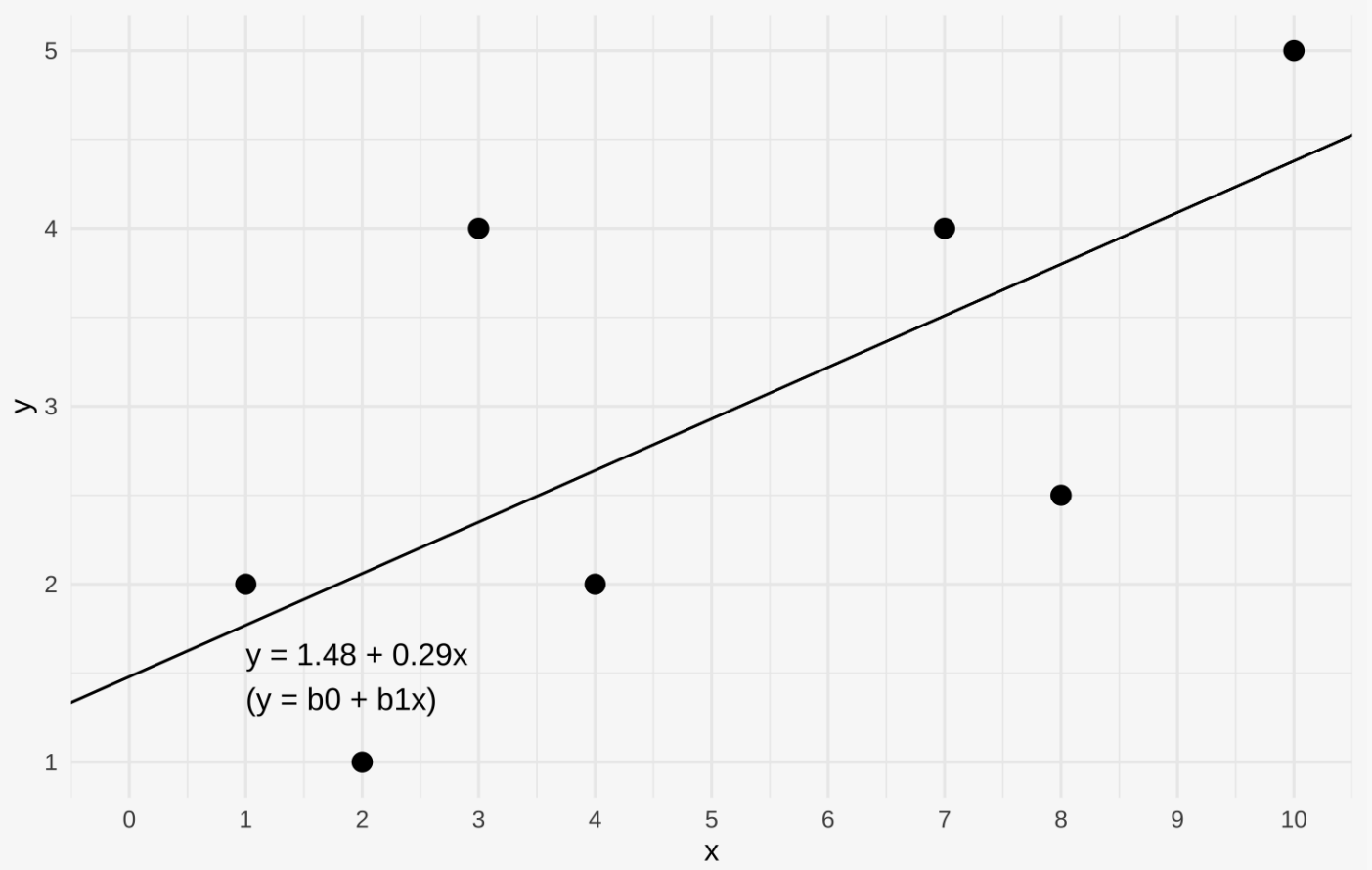
# Aside: OLS: how does it work?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = 0.29$$

$$\hat{\beta}_0 = 1.48$$



# How to report a regression model

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.
- **p-values (Pr(>|t|))**: If appropriate, whether the associations between X and Y were statistically significant.

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.
- **p-values (Pr(>|t|))**: If appropriate, whether the associations between X and Y were statistically significant.
- **Intercept/slope (Estimate)**: The strength and direction of the relationship.

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.
- **p-values (Pr(>|t|))**: If appropriate, whether the associations between X and Y were statistically significant.
- **Intercept/slope (Estimate)**: The strength and direction of the relationship.
  - **Direction**: Is the estimate a positive number (as X increases, Y also increases), or a negative number (as X increases, Y decreases).

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.
- **p-values (Pr(>|t|))**: If appropriate, whether the associations between X and Y were statistically significant.
- **Intercept/slope (Estimate)**: The strength and direction of the relationship.
  - **Direction**: Is the estimate a positive number (as X increases, Y also increases), or a negative number (as X increases, Y decreases).
  - **Effect size**: Exactly how much change in Y would be expect to see on average for a 1-unit increase in X?

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Using the `summary()` function then gives us our model output. This gives us a range of results, from very general statements to very specific ones.

- **R-squared**: Total variance in Y that can be explained by variance in X.
- **p-values (Pr(>|t|))**: If appropriate, whether the associations between X and Y were statistically significant.
- **Intercept/slope (Estimate)**: The strength and direction of the relationship.
    - **Direction**: Is the estimate a positive number (as X increases, Y also increases), or a negative number (as X increases, Y decreases).
    - **Effect size**: Exactly how much change in Y would be expect to see on average for a 1-unit increase in X?
    - **Confidence Intervals**: A 95% confidence interval around the effect size can be calculated by adding and subtracting 1.96 times the standard error to the estimate.

# Reporting a regression model

```
lab_model <- lm(data = un_data,
                formula = lab_force_women ~ sec_ed_women)

summary(lab_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ sec_ed_women, data = un_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4142  -3.6709   0.4669   5.0493  13.8211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.61977    5.58156   5.128 3.31e-06 ***
## sec_ed_women  0.29158    0.06545   4.455 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.164 on 60 degrees of freedom
## Multiple R-squared:  0.2485,    Adjusted R-squared:  0.236
## F-statistic: 19.84 on 1 and 60 DF,  p-value: 3.727e-05
```

Overall, the percentage of women with at least some secondary school education could explain approximately 25 per cent of the variance in women's labour market participation. The relationship between the two variables was statistically significant ($p<0.05$). As the percentage of women with at least some secondary school education increased, the percentage of women participating in the labour market also increased. A 1 percentage point increase in women with secondary school education was associated with a 0.3 percentage point increase in women's labour market participation, on average.

# Note: Pretty regression output

If using Rmarkdown, you can make prettier regression results tables using the `stargazer` package. For example:

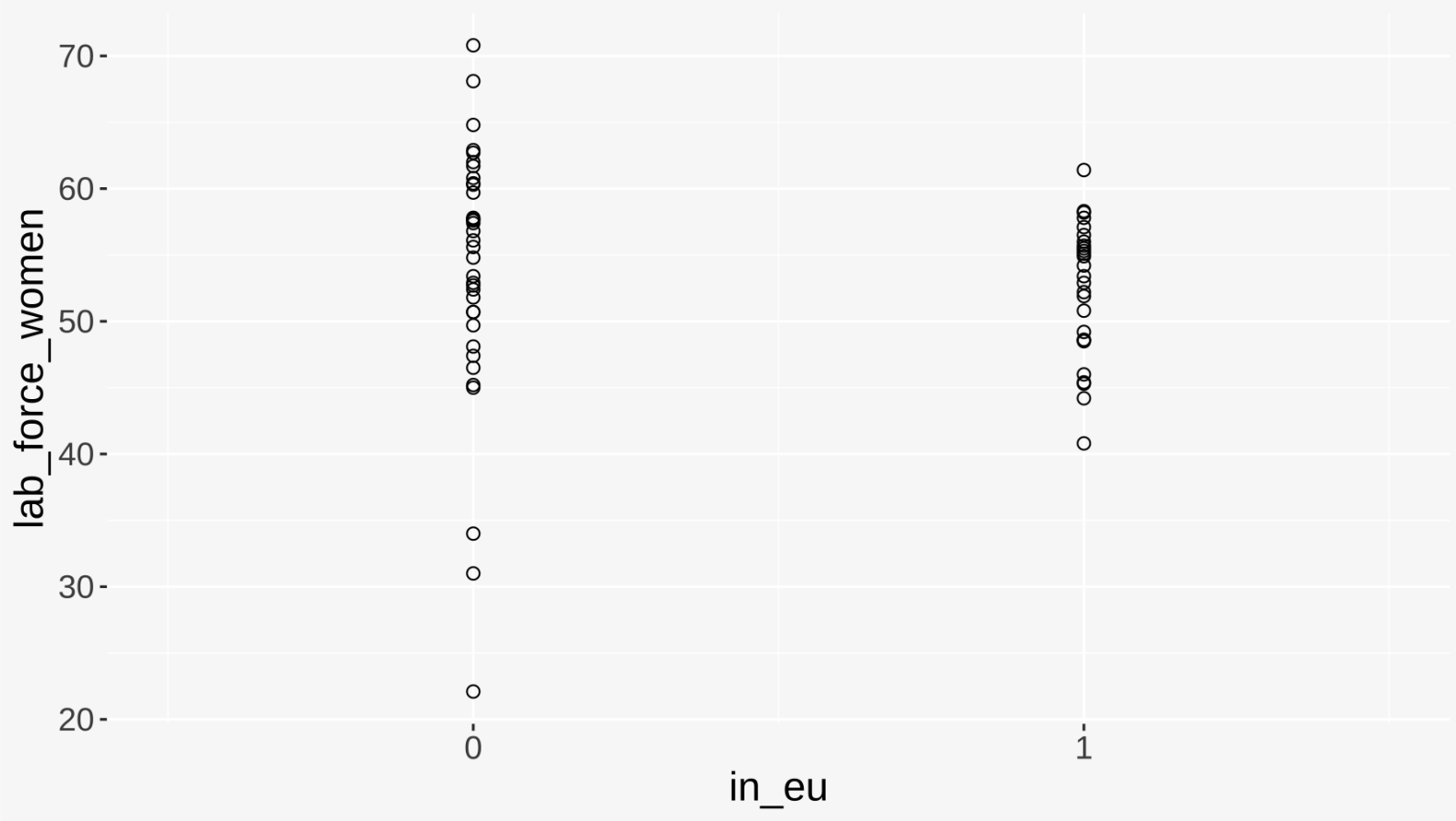
```
library(stargazer)

stargazer::stargazer(lab_model, type = "html")
```

$$\bar{\text{women}}\text{LMP} = 28.6 + 0.3\text{womenSER}$$

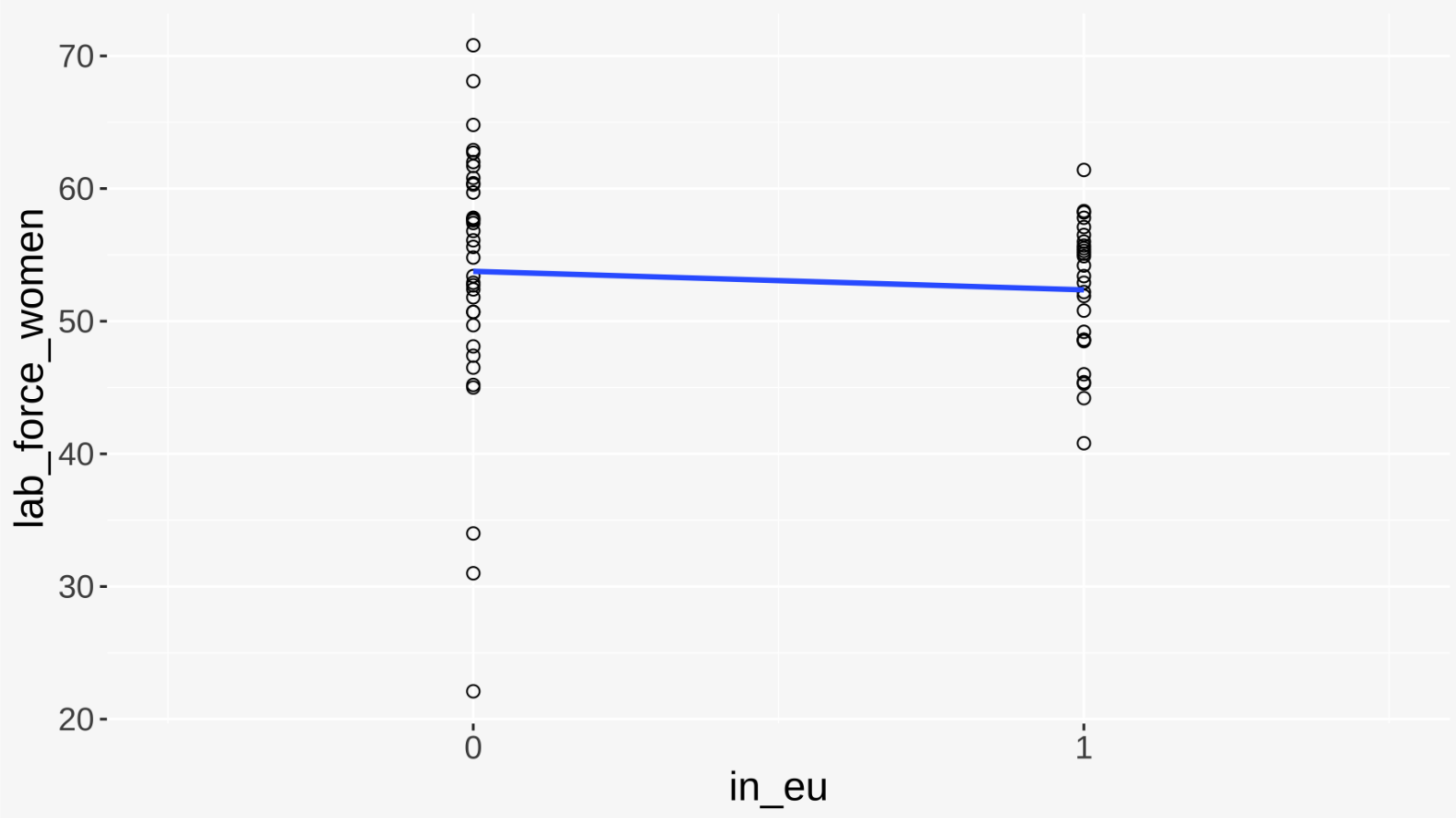
| | *Dependent variable:* |
|---|---|
| | lab_force_women |
| sec_ed_women | 0.292*** |
| | (0.065) |
| Constant | 28.620*** |
| | (5.582) |
| Observations | 62 |
| $R^2$ | 0.249 |
| Adjusted $R^2$ | 0.236 |
| Residual Std. Error | 7.164 (df = 60) |
| F Statistic | 19.845*** (df = 1; 60) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

# Binary categorical predictors

# Binary categorical predictors



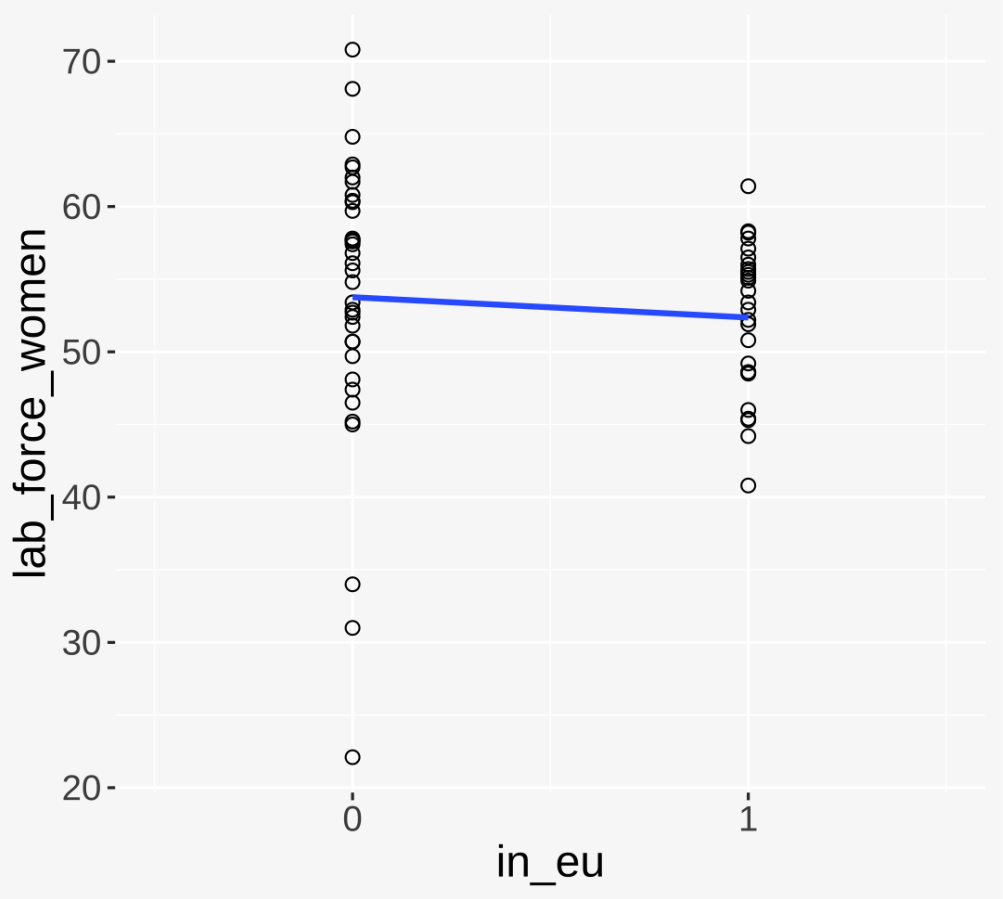
# Binary categorical predictors



# Binary categorical predictors

```
eu_model <- lm(data = un_data,
               formula = lab_force_women ~ in_eu)

summary(eu_model)
```

```
##
## Call:
## lm(formula = lab_force_women ~ in_eu, data = un_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.660  -3.763   1.439   4.587  17.040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.760      1.392  38.623   <2e-16 ***
## in_eu         -1.397      2.109  -0.662     0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 8.235 on 60 degrees of freedom
## Multiple R-squared:  0.007258,    Adjusted R-squared:  -0.00
## F-statistic: 0.4387 on 1 and 60 DF,  p-value: 0.5103
```




"An increase of 1 in the 'in_eu' variable [meaning, the country is in the EU] was associated with an average decrease of 1.4 points in the percentage of women in the labour force."

# Assumptions of linear regression

Important! You will need to know how to check these for your assessments!

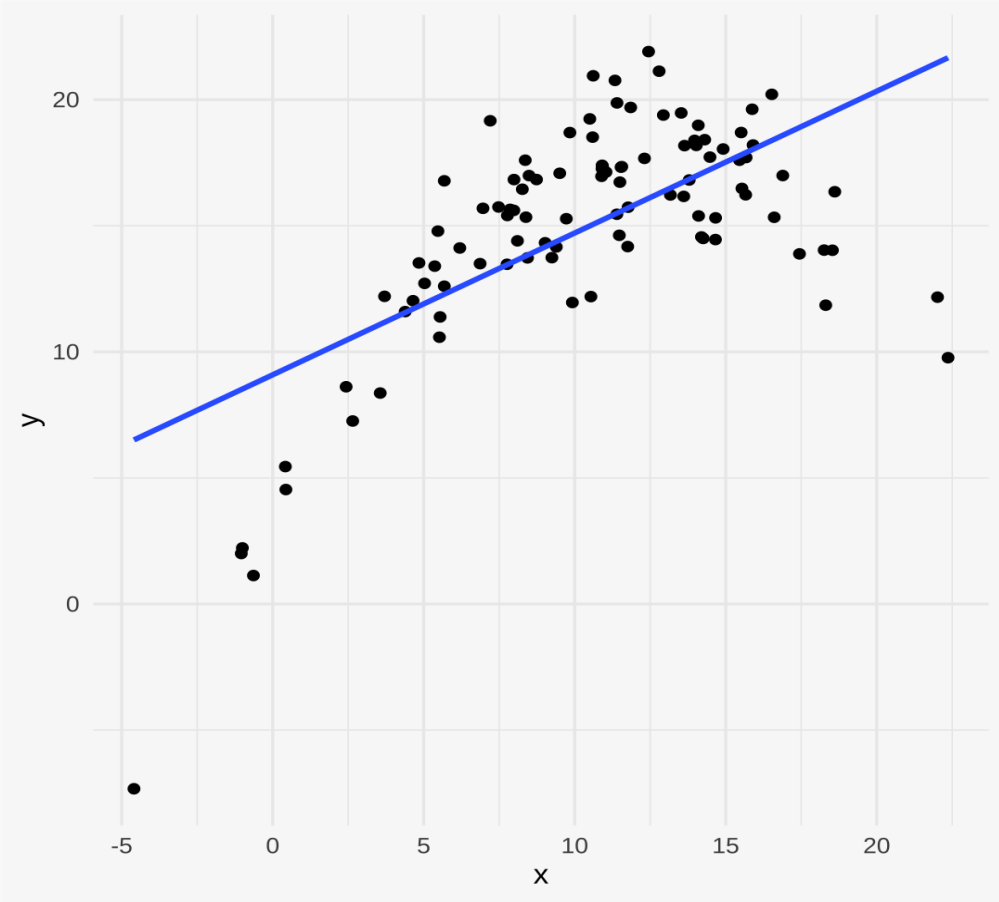
# Assumptions

- **Linearity**

The most appropriate approximation of the relationship in the data is a straight line.

Checked with: Scatterplot (bivariate); residuals versus fitted values (multiple regression, next week).

- **Homoscedasticity**
- **Outliers and leverage points**
- **Normality of residuals**



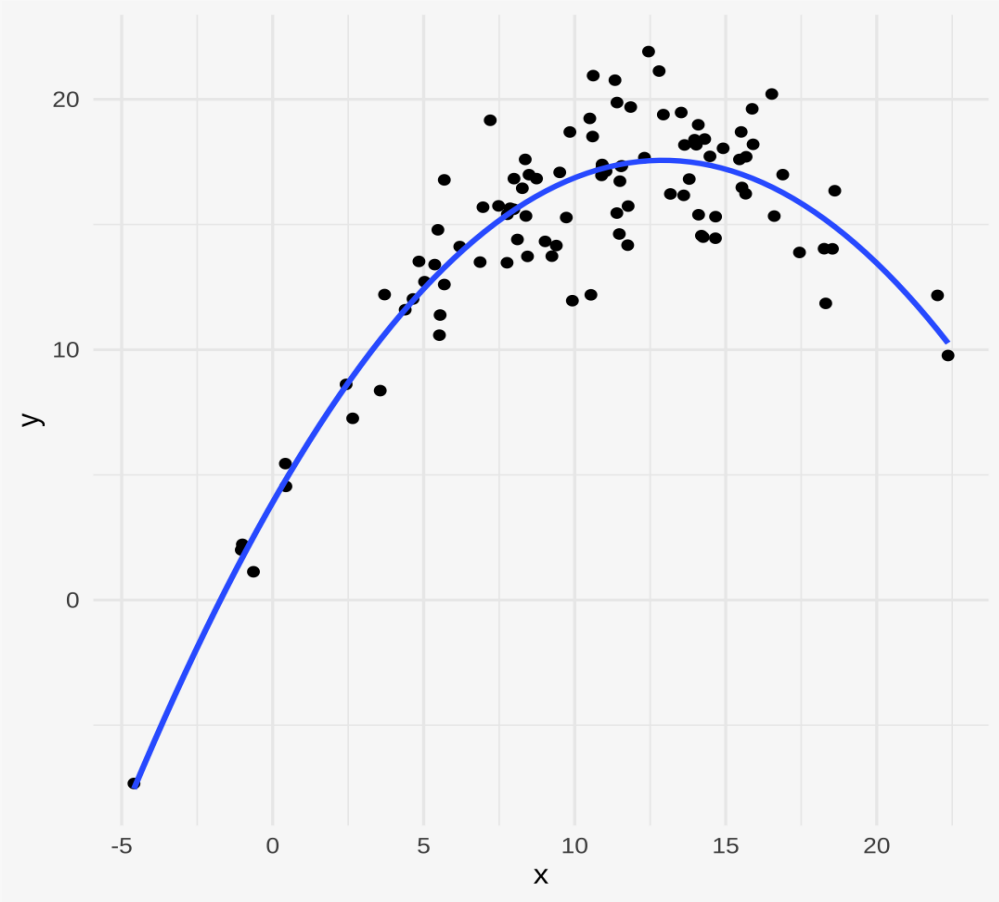
# Assumptions

- **Linearity**

The most appropriate approximation of the relationship in the data is a straight line.

Checked with: Scatterplot (bivariate); residuals versus fitted values (multiple regression, next week).

- **Homoscedasticity**
- **Outliers and leverage points**
- **Normality of residuals**



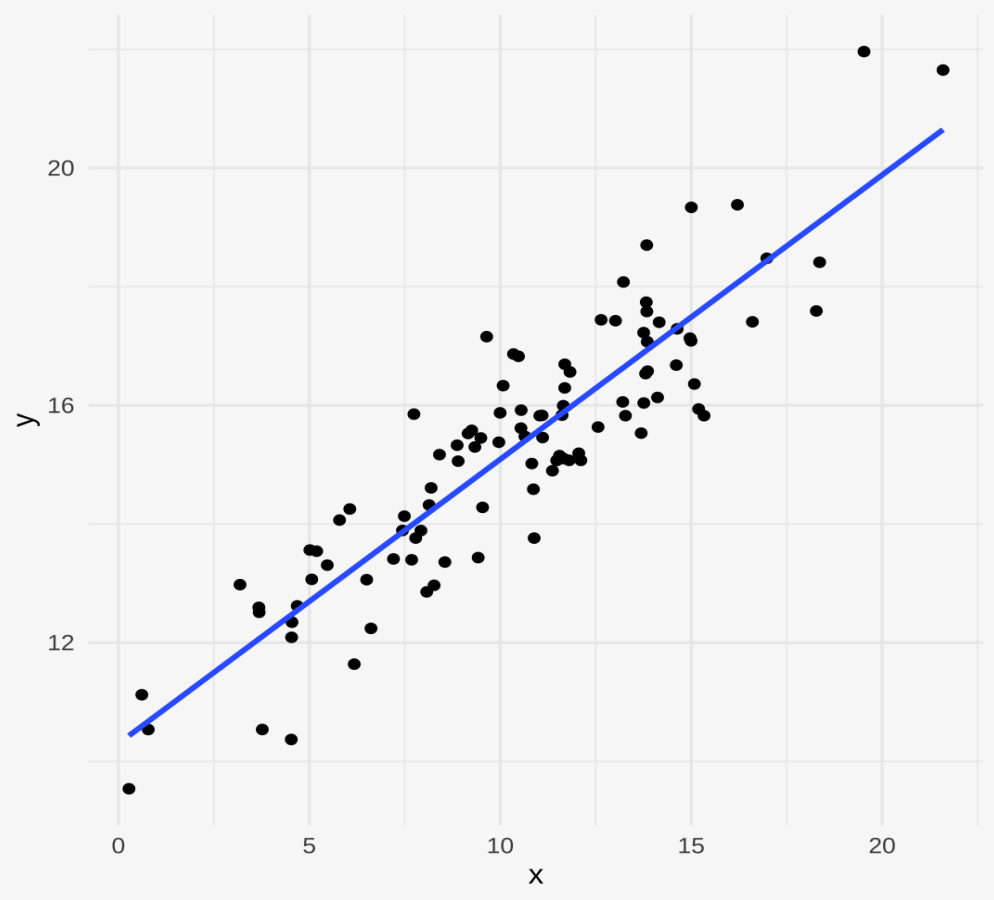
# Assumptions

- **Linearity**
- **Homoscedasticity**

The spread of points around the regression line is approximately the same size at all points in the regression line.

Checked with: Scatterplot (bivariate); spread-location plot (multiple regression, next week).

- **Outliers and leverage points**
- **Normality of residuals**



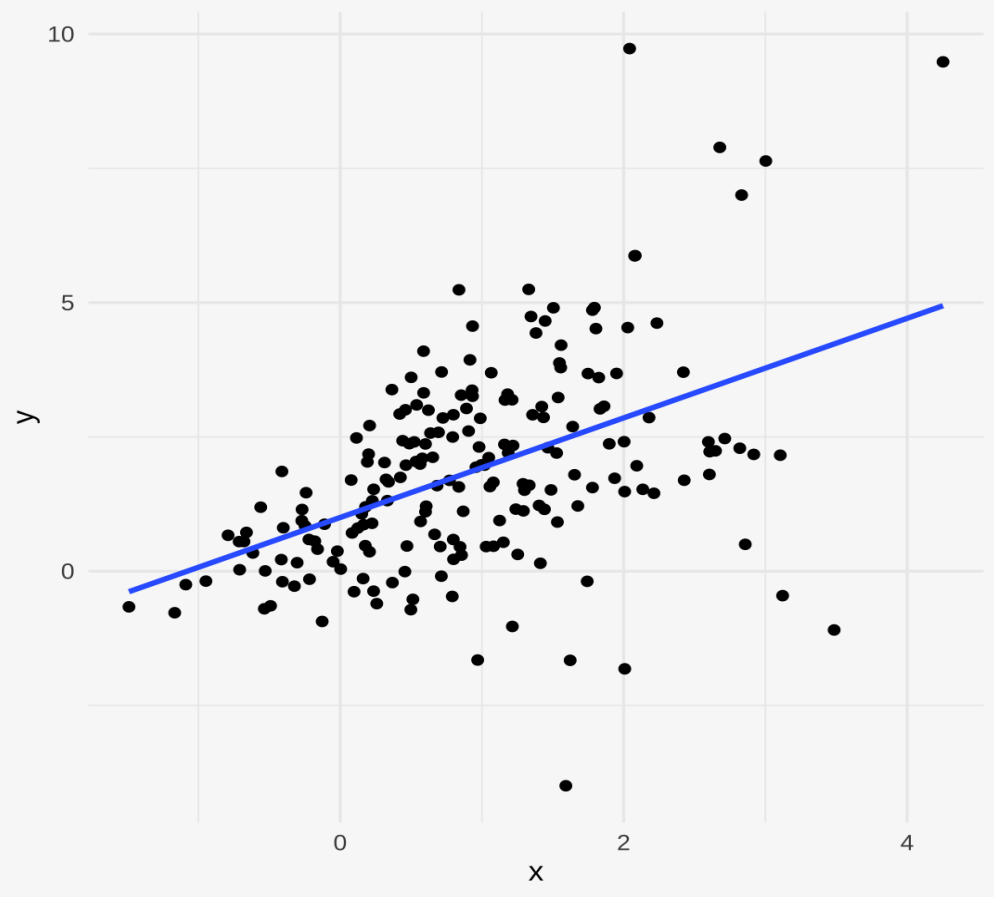
# Assumptions

- **Linearity**
- **Homoscedasticity**

The spread of points around the regression line is approximately the same size at all points in the regression line.

Checked with: Scatterplot (bivariate); spread-location plot (multiple regression, next week).

- **Outliers and leverage points**
- **Normality of residuals**



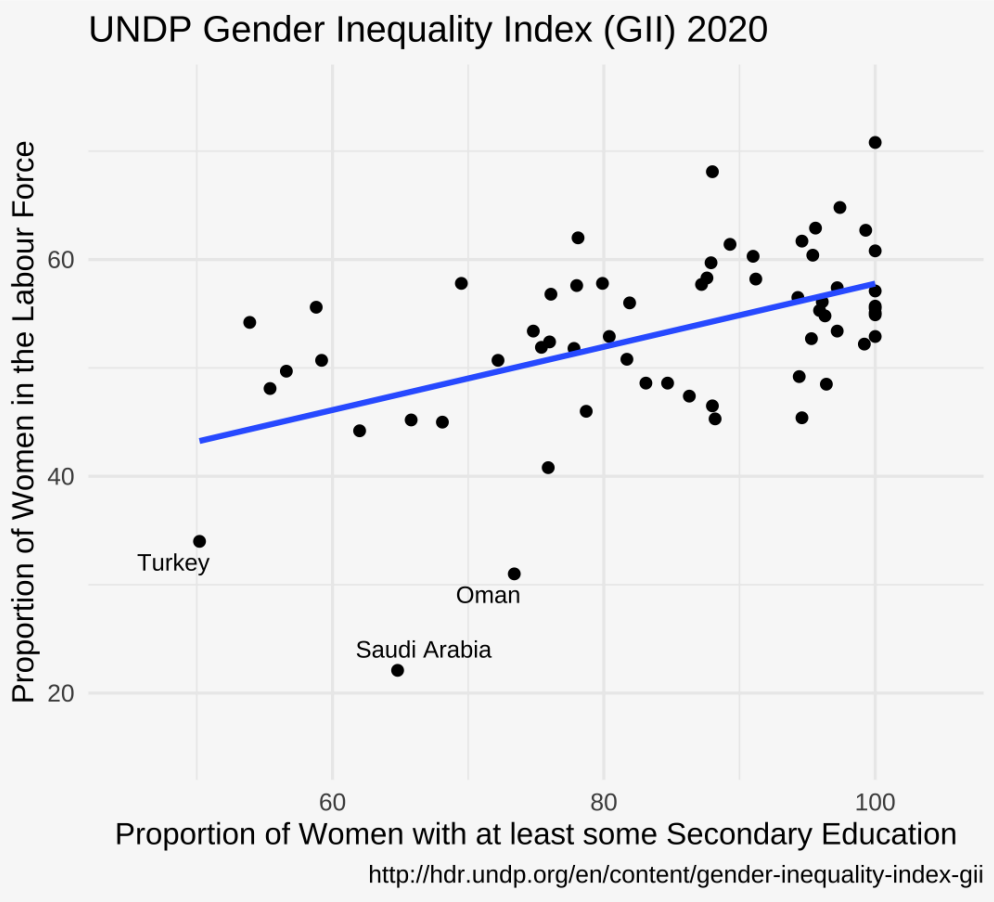
# Assumptions

- **Linearity**
- **Homoscedasticity**
- **Outliers and leverage points**

Outliers and leverage points can have an undue influence on the slope of the line. Outliers deviate far from the general pattern whereas leverage points tend to follow the trend (somewhat) but be far away from the general cluster. Can be **true** outliers, or artefacts/errors.

Checked with: Scatterplot (bivariate); leverage plot (multiple regression)

- **Normality of residuals**




$$Y = 28.62 + 0.29X$$

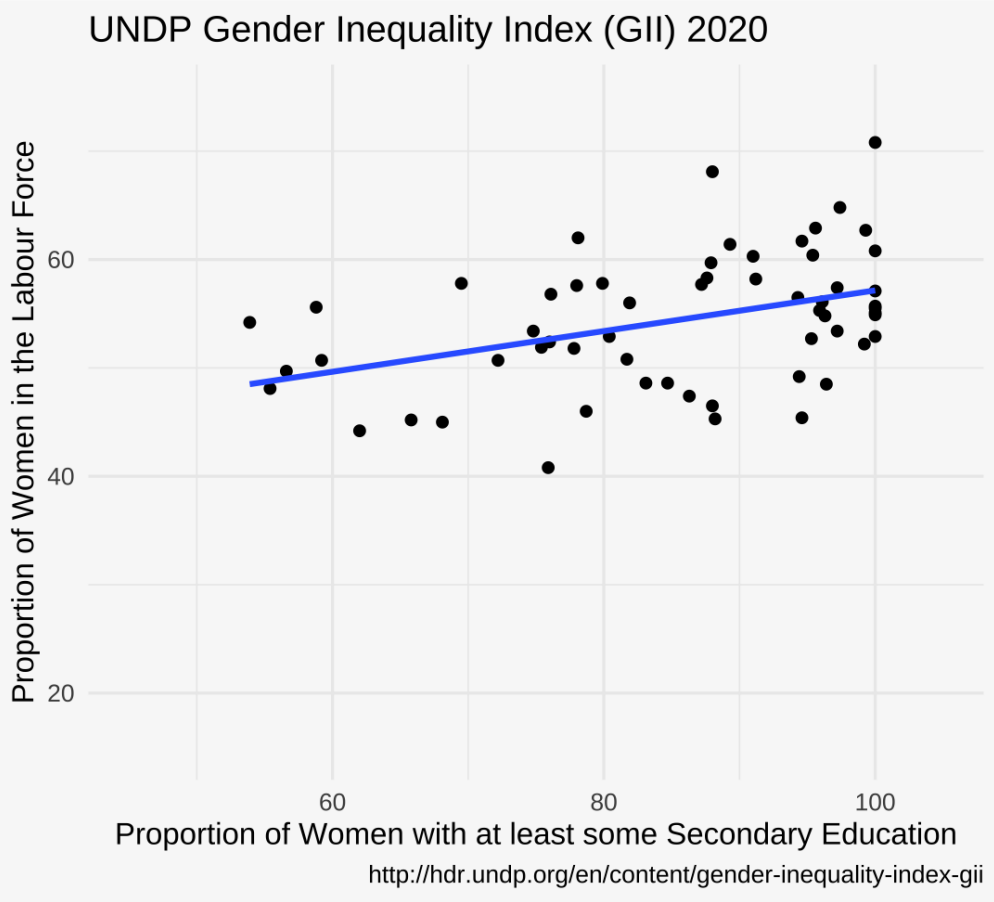
# Assumptions

- **Linearity**
- **Homoscedasticity**
- **Outliers and leverage points**

Outliers and leverage points can have an undue influence on the slope of the line. Outliers deviate far from the general pattern whereas leverage points tend to follow the trend (somewhat) but be far away from the general cluster. Can be **true** outliers, or artefacts/errors.

Checked with: Scatterplot (bivariate); leverage plot (multiple regression)

- **Normality of residuals**



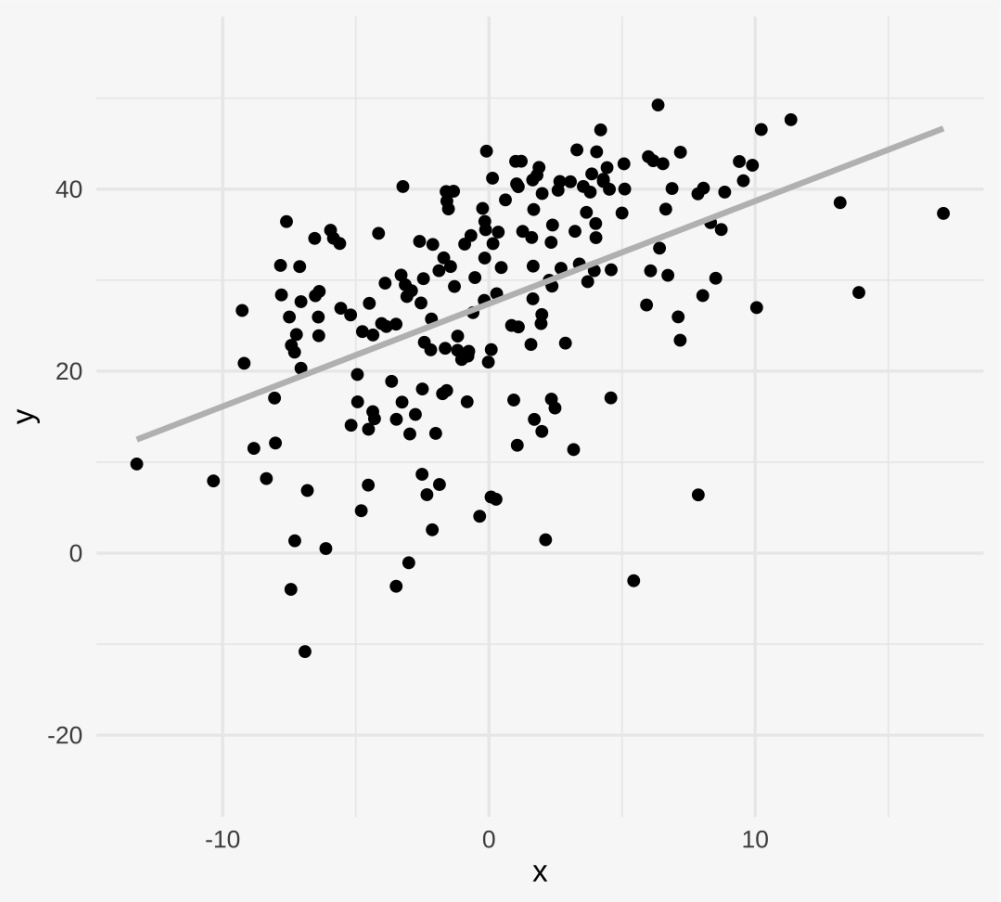

$$Y = 38.39 + 0.19X$$

# Assumptions

- **Linearity**
- **Homoscedasticity**
- **Outliers and leverage points**
- **Normality of residuals**

The residuals from the regression line have an approximately normal distribution around the line (e.g. there are not more points closer together on one side of the line than there are on the other).

Checked with: Scatterplot with regression line (bivariate); Q-Q plot (multiple regression)

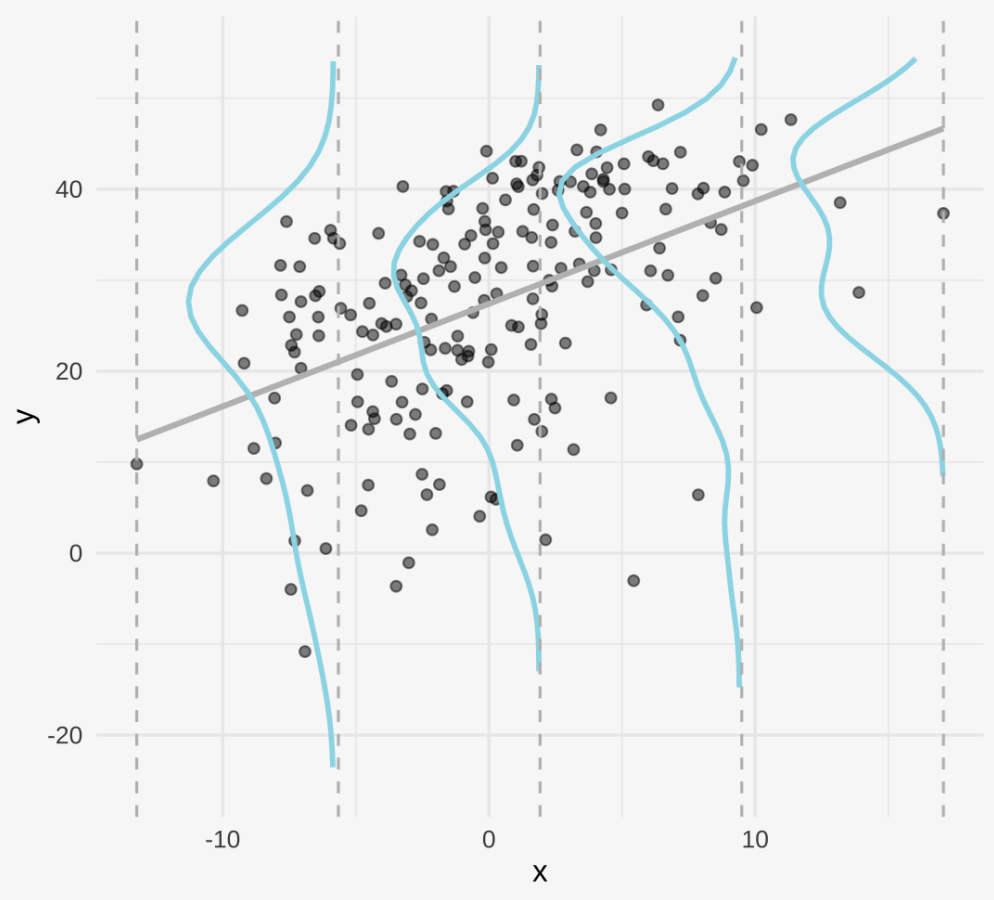


# Assumptions

- **Linearity**
- **Homoscedasticity**
- **Outliers and leverage points**
- **Normality of residuals**

The residuals from the regression line have an approximately normal distribution around the line (e.g. there are not more points closer together on one side of the line than there are on the other).

Checked with: Scatterplot with regression line (bivariate); Q-Q plot (multiple regression)



# Diagnostics: What happens and what to do when assumptions are violated? Cheat sheet.

- **Linearity**

This means that your estimates will be over- and under-estimated at different parts of the line. If your dependent variable is right-skewed (long tail of observations in the + direction), you can transform it to its log value using `log()` . Alternatively, you can fit a curvilinear model — bit more advanced but not too difficult!

- **Homoscedasticity**

Heteroscedasticity means that the standard errors will be biased (in different ways depending on the shape) — this can be resolved using robust standard errors and/or a weighted least squares estimator.

- **Outliers and leverage points**

Outliers and leverage points can generally have an undue influence on the slope in a regression. Remove any error-based outliers (e.g. someone enters their age as 400 instead of 40). Compare results with 'true' outliers included and excluded to see how they differ, present both.

- **Normality of residuals**

Non-normality of residuals can mean that standard errors are smaller than they should be, which can effect decisions made in hypothesis testing in marginal cases — but this is not generally a big problem outside of very small studies (N < 10 per variable) (Schmidt & Finan, 2018). Make this clear if either is true.

# Summary
# Bivariate Linear Regression

- Simple bivariate linear regression can be expressed as a line through the data points in our scatterplot.
- This line can be expressed very efficiently in the form of an equation ( $\bar{y} = b_0 + b_1 x$ ), and can be interpreted in an often meaningful way (as $X$ increases by 1, the mean value of $Y$ increases by $b_1$ )
- Regression models include inferential statistics (where $H_0$ = the slope of the regression line is not significantly different from what we would expect if it were 0 in the entire population, a.k.a. a flat line).
- Regression provides us with a highly flexible and descriptive form of statistical modelling, able to incorporate binary and non-normally distributed predictors of a continuous dependent variable.
- As we will see next week, regression models are even more useful as they allow us to include more than one predictor to explore conditional associations.

# R Exercise
# Bivariate Linear Regression

**Research question: To what extent are women's secondary school education rates, women's workforce participation rates, and societal bias against women associated with the proportion of seats in parliament held by women?**

- Download `week-7-exercise.zip` and open the `week-7-exercise` Rproject file and R script.