Comparing two numeric variables: Basics of linear regression

Research Methods for Human Inquiry

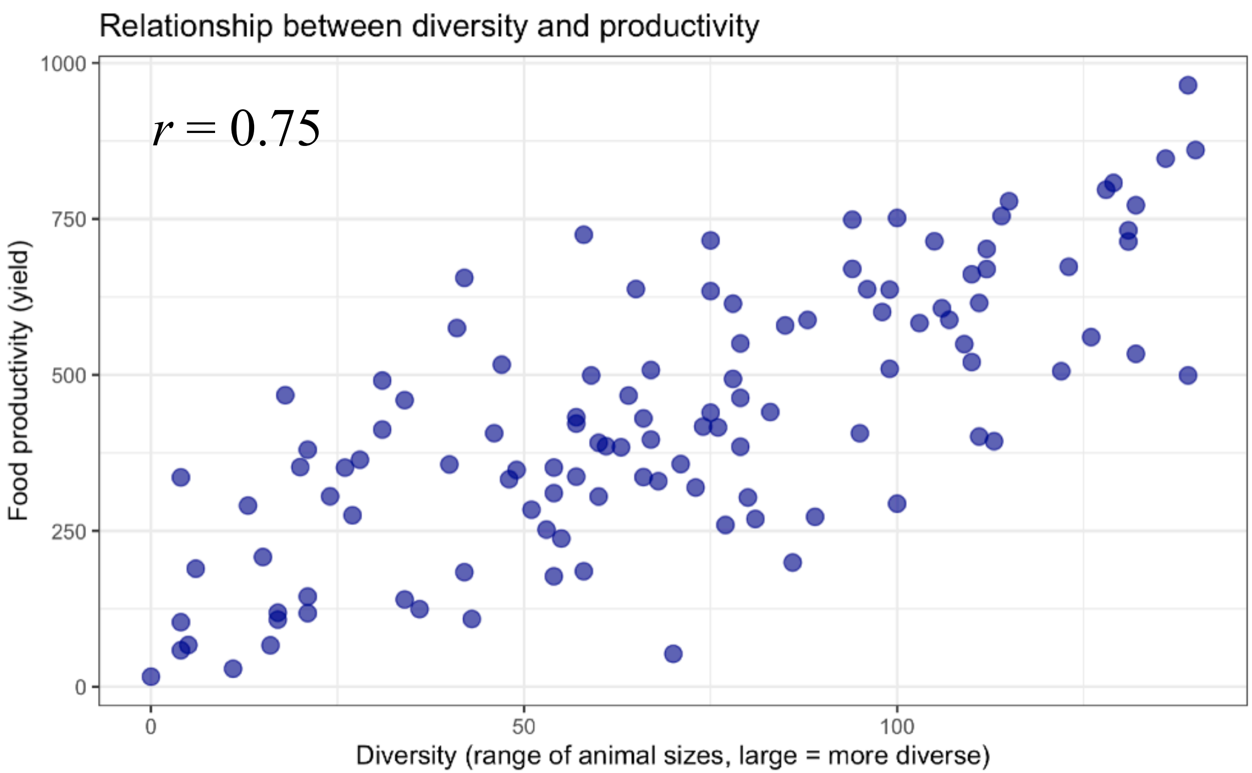
Andrew Perfors

What is linear regression?

- A tool for describing the **relationships** between multiple interval scale variables
- Outcome and predictor are BOTH numeric
  - we can have multiple predictors
  - (it can be generalised to handle other situations too, but we won't talk about them for now)

We've already seen this…

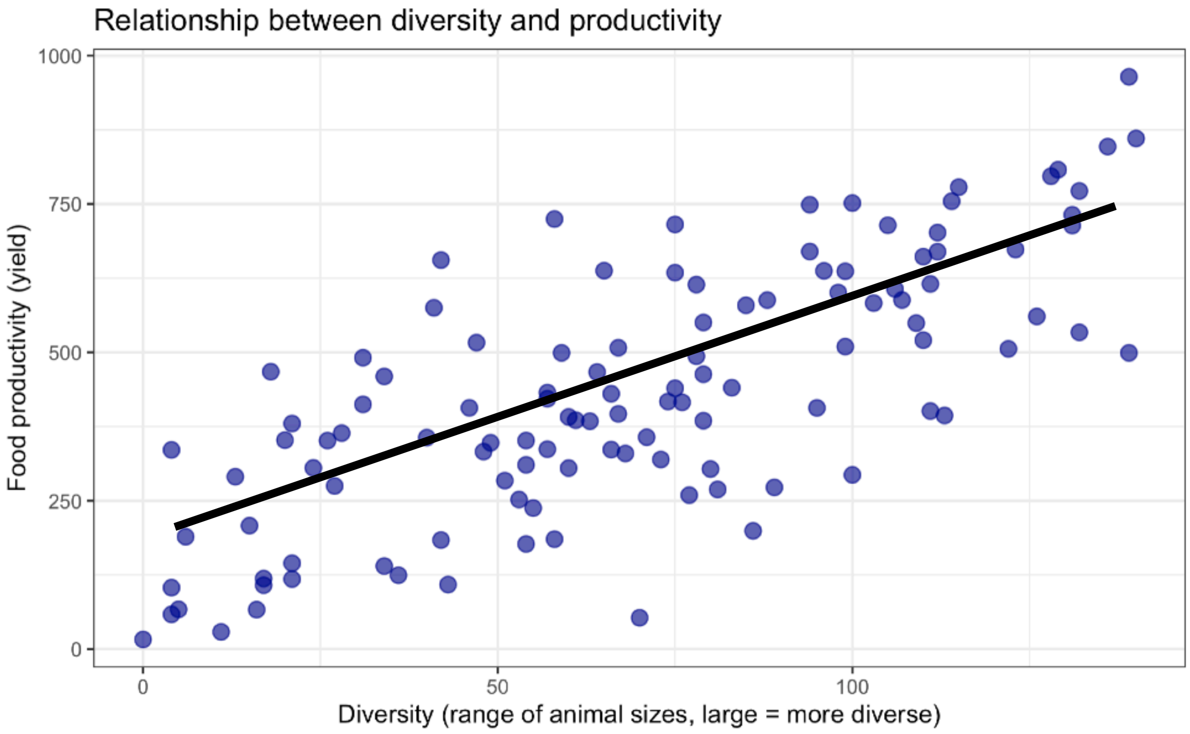
- A correlation is a relationship between two (or more) numeric variables



This was useful, but what if we want to compare multiple variables to see which is contributing most?

Regression: lets you hypothesis test and check *which* variables most influence an outcome. Can also characterise more about the relationship between variables

Regression



Fundamental idea: fit the best **regression line** to the data, and then try to understand that line

Formula for a regression line

$$Y = b_1 X + b_0$$

A regression line

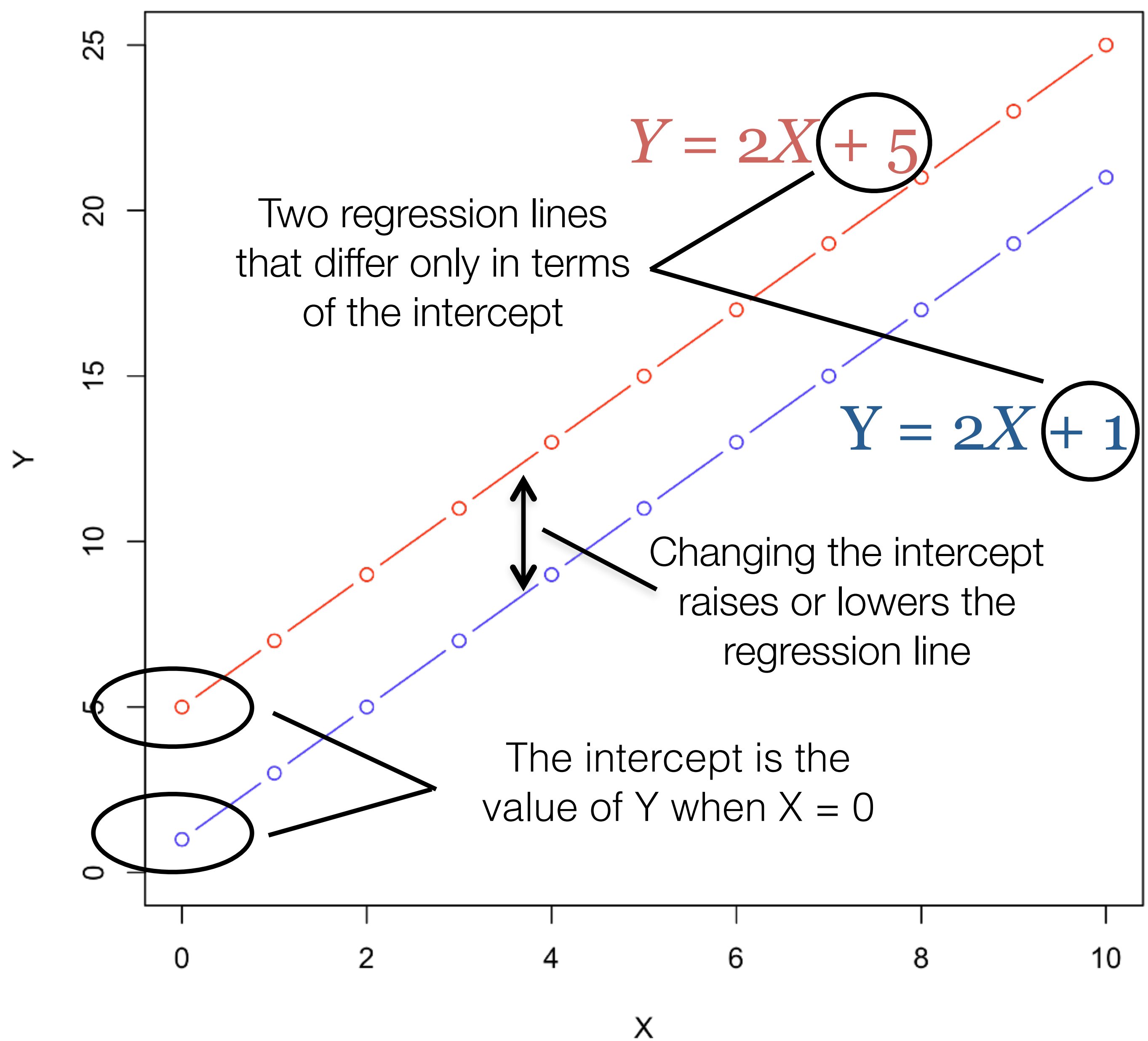
b_1 is the <u>slope</u> of the regression line

b_0 is the <u>intercept</u> of the regression line

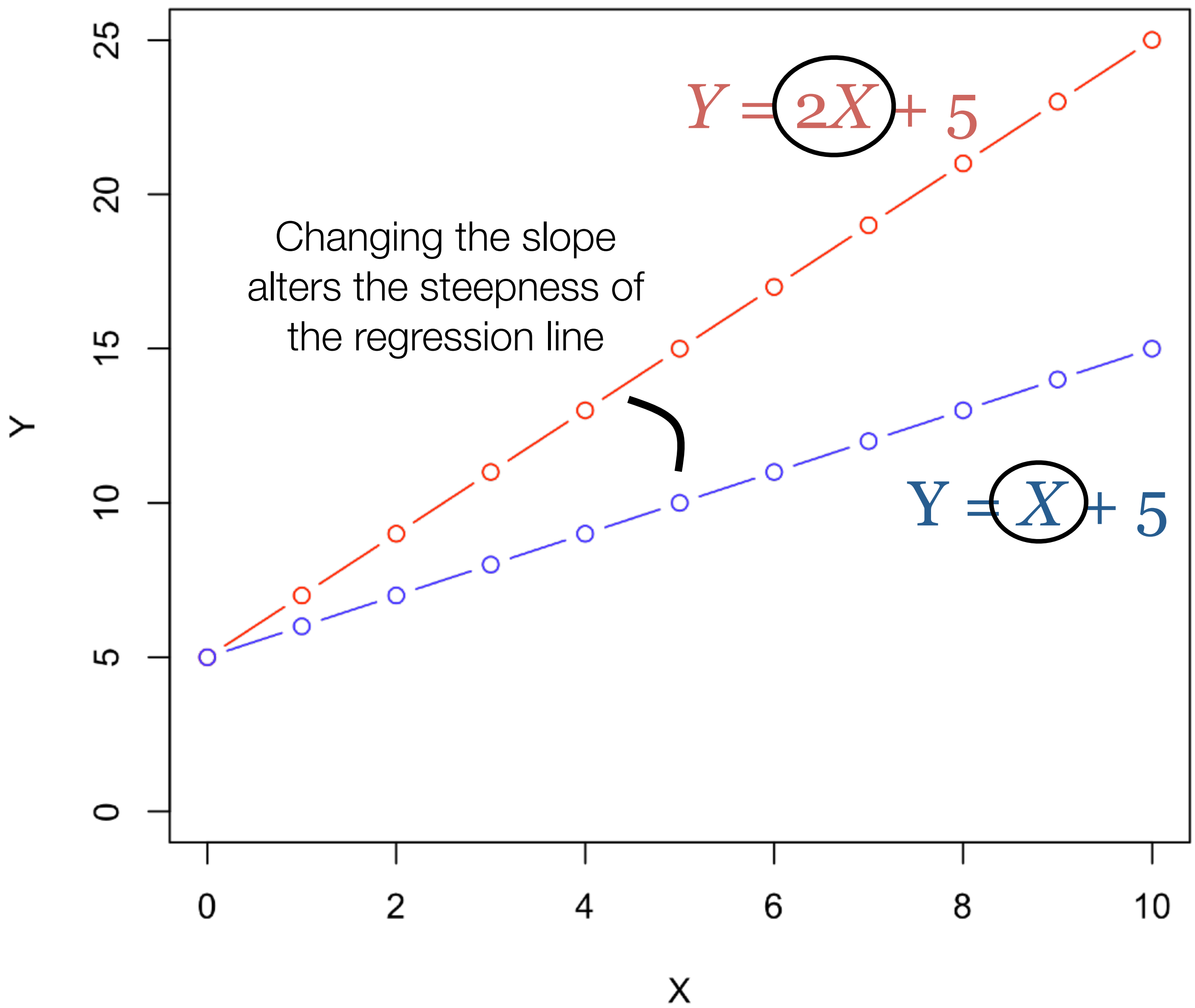
$$Y = b_1 X + b_0$$

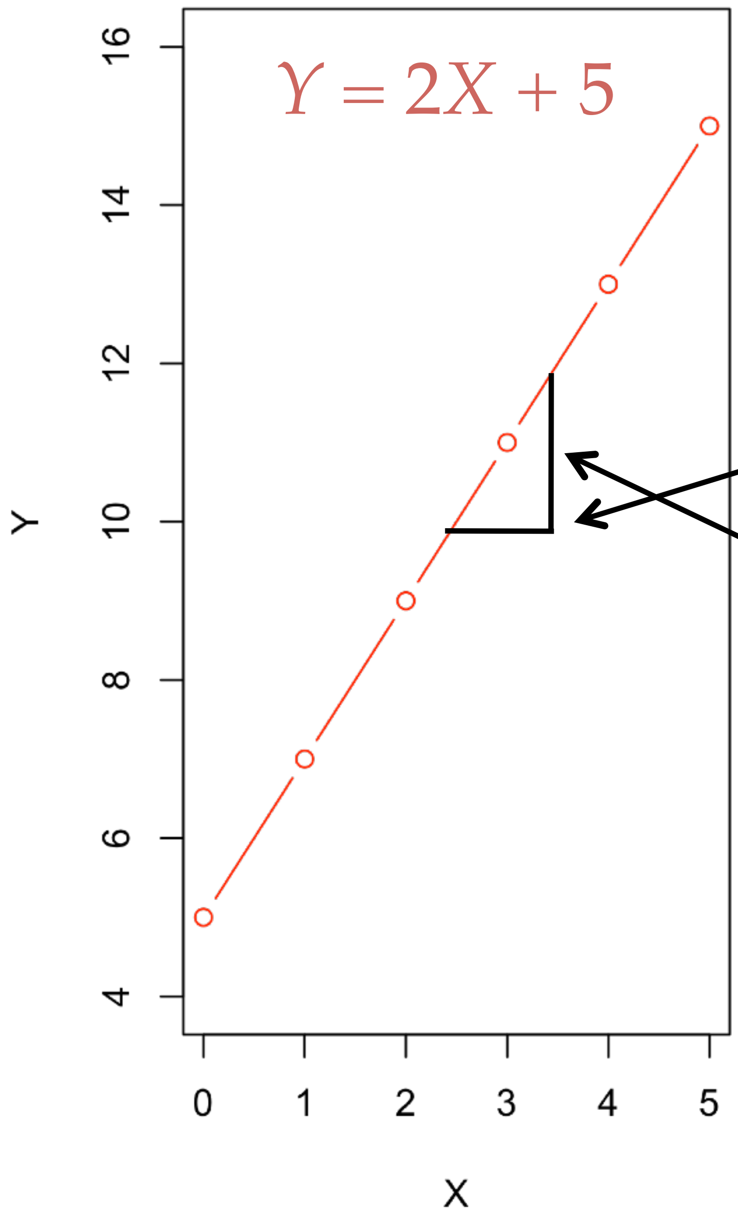
Y is the outcome variable (yield)

X is the predictor variable (diversity)



$Y = 2X + 5$
Two regression lines
that differ only in terms
of the intercept
$Y = 2X + 1$
Changing the intercept
raises or lowers the
regression line
The intercept is the
value of Y when X = 0
Y
X
0
5
10
15
20
25
0
2
4
6
8
10


Y = 2X + 5
Changing the slope alters the steepness of the regression line
Y = X + 5
X
Y
0
2
4
6
8
10
0
5
10
15
20
25



Suppose we have a slope of 2.

What that means is…

If we increase the value of X by 1 …

Then the regression line will predict an increase in Y of 2

From regression lines to regression models

- The regression line is what we've just seen

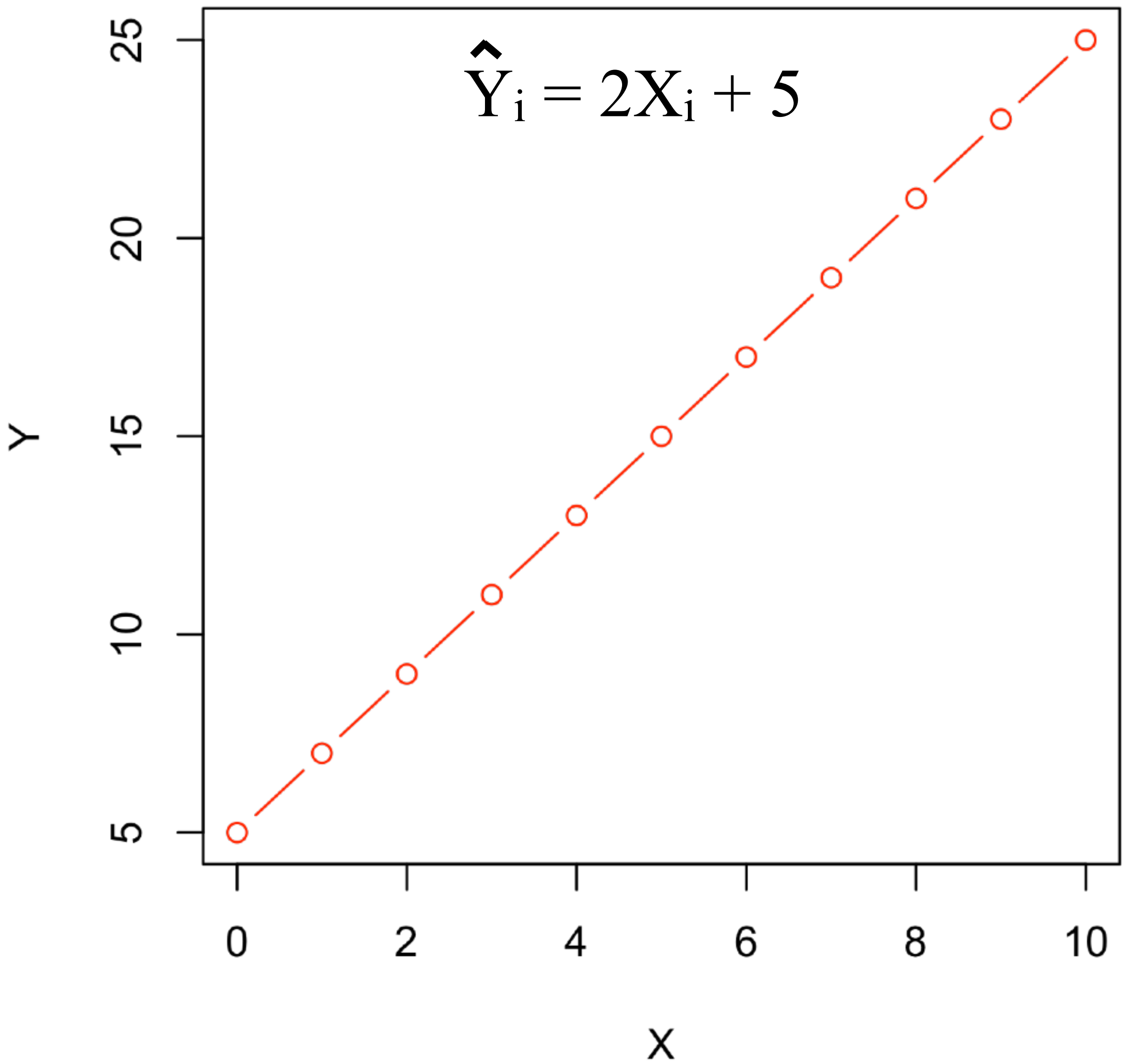
$$Y = b_1 X + b_0$$

- A regression model acknowledges the existence of random variation in the data

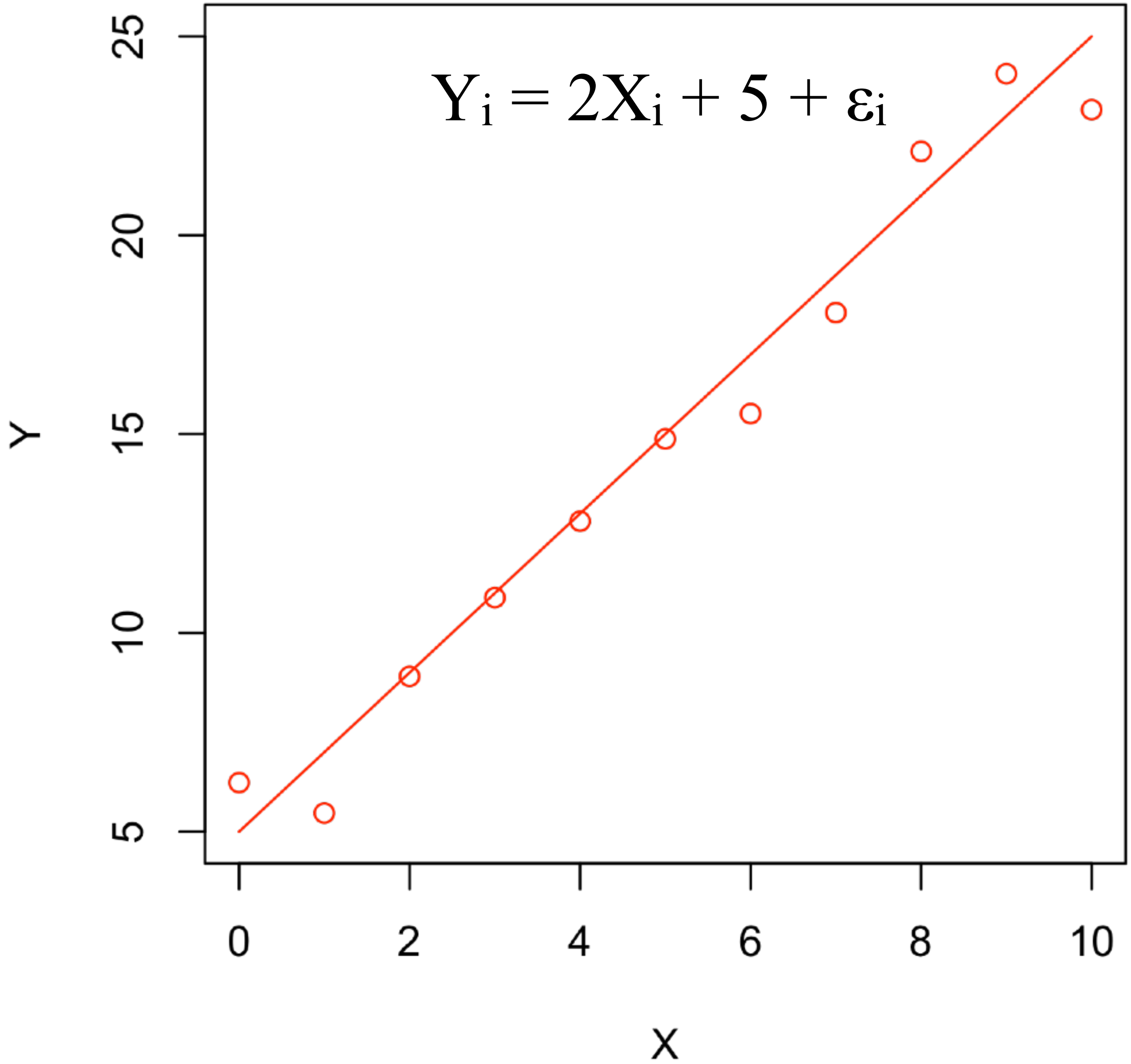
$$Y_i = b_1 X_i + b_0 + \varepsilon_i$$

  - The "i" subscript indicates we're talking about data here, specifically the i-th observation in the data set
  - The "epsilon" term ε_i is a "residual"... a deviation from the regression line

What the regression line predicts is $\hat{Y}_i$

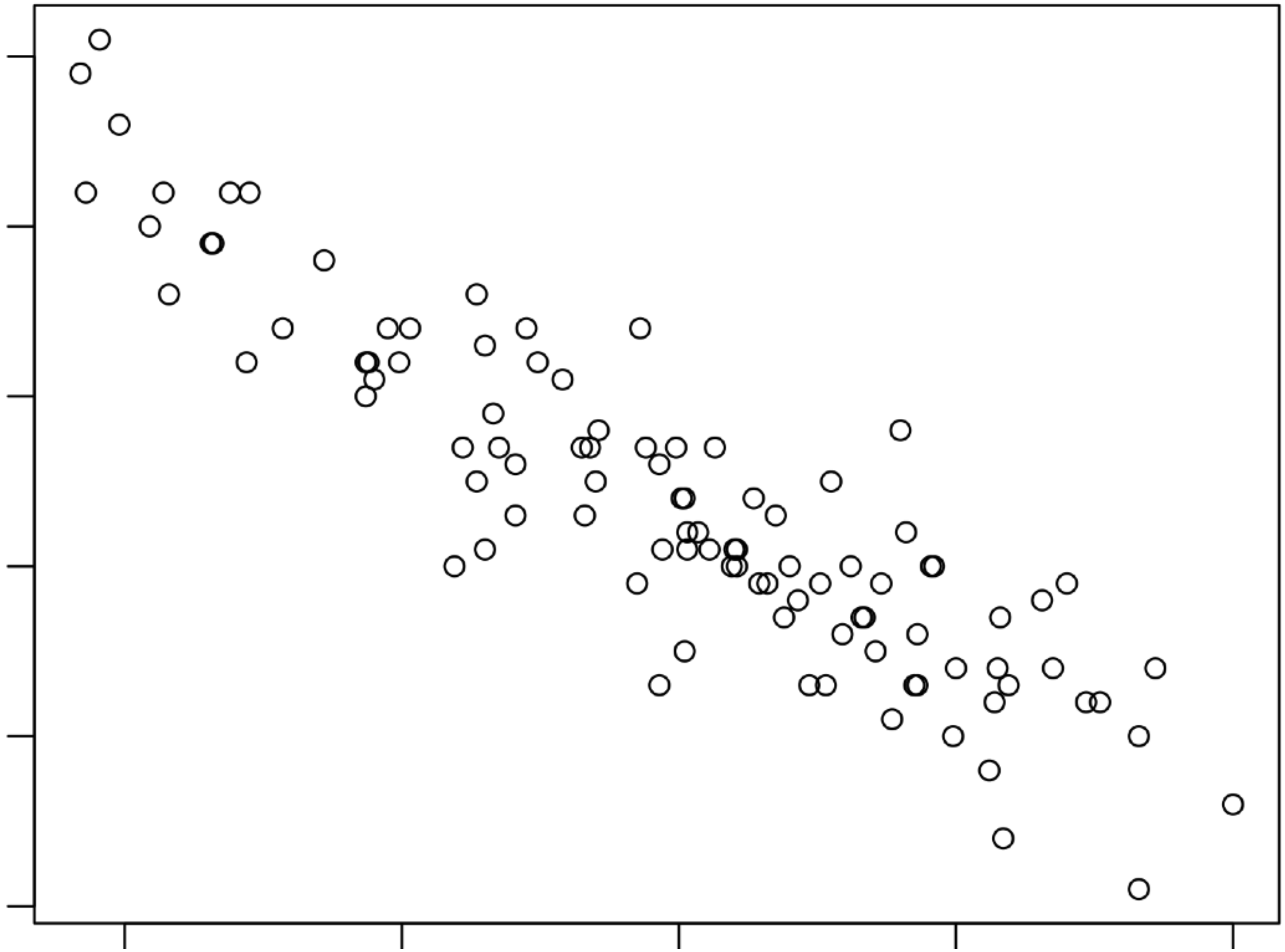


What we actually observe is Y_i

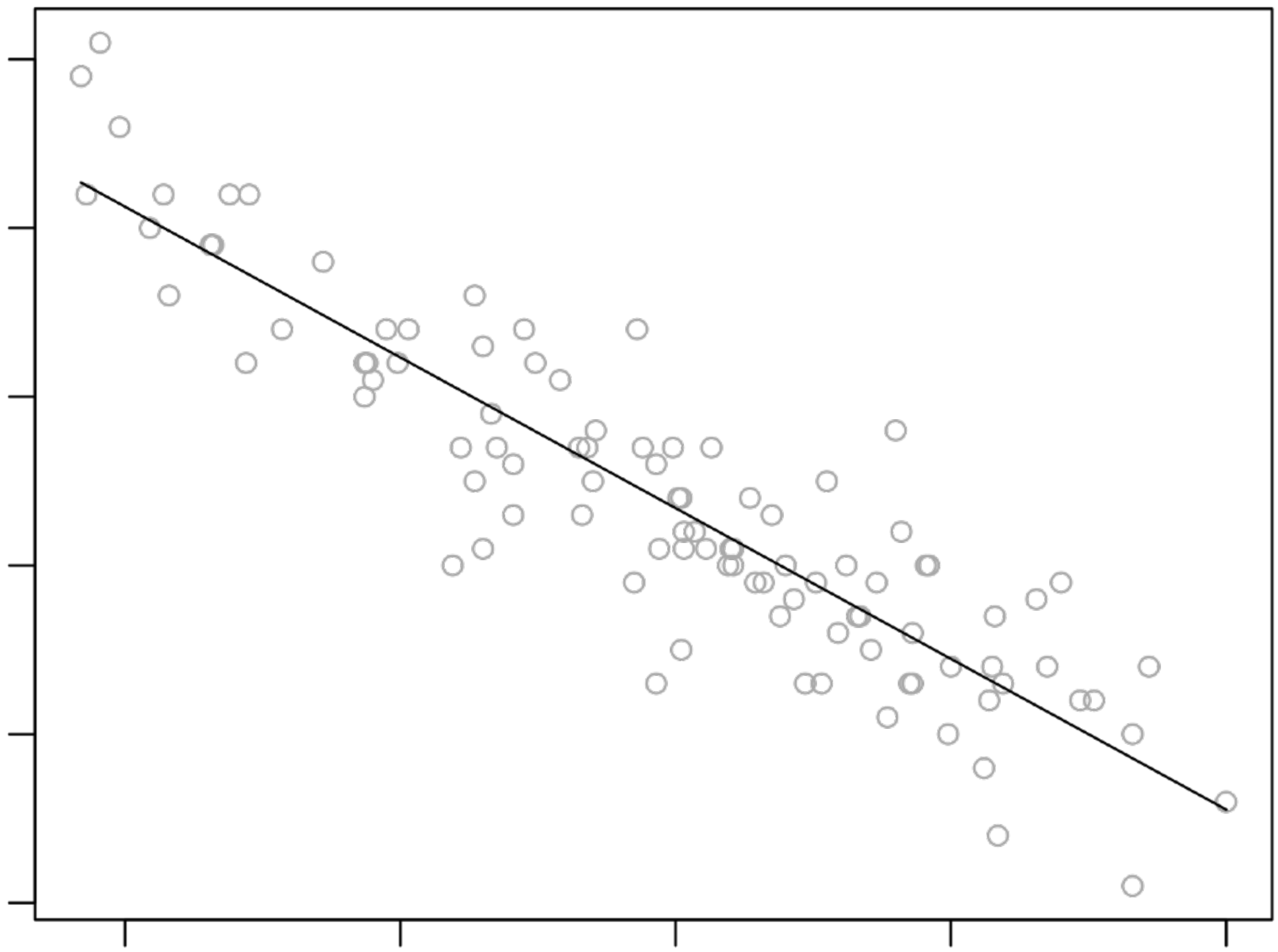


How do we **estimate** a regression line?

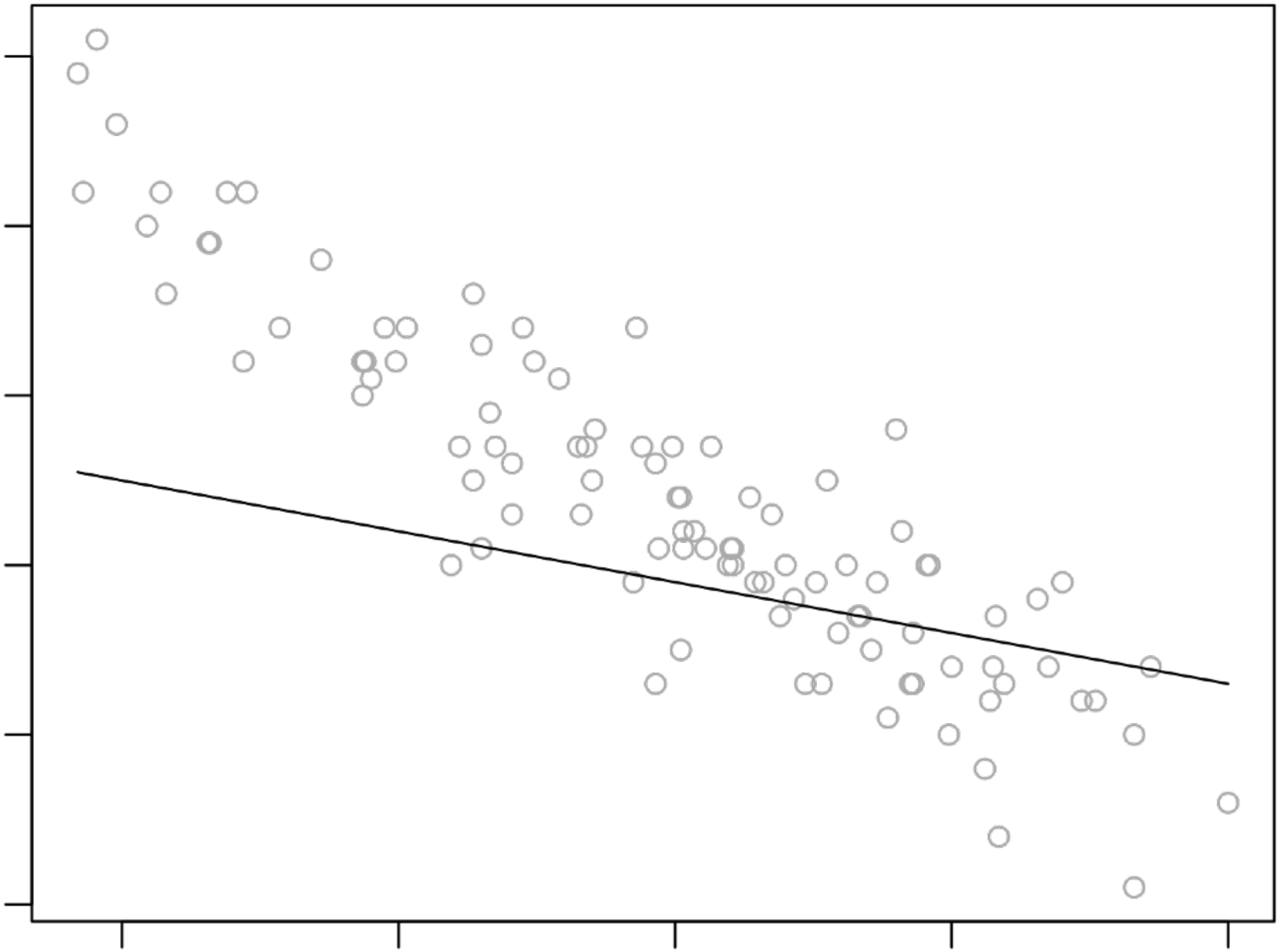
Imagine the following data



The best-fitting regression line



NOT the best-fitting regression line

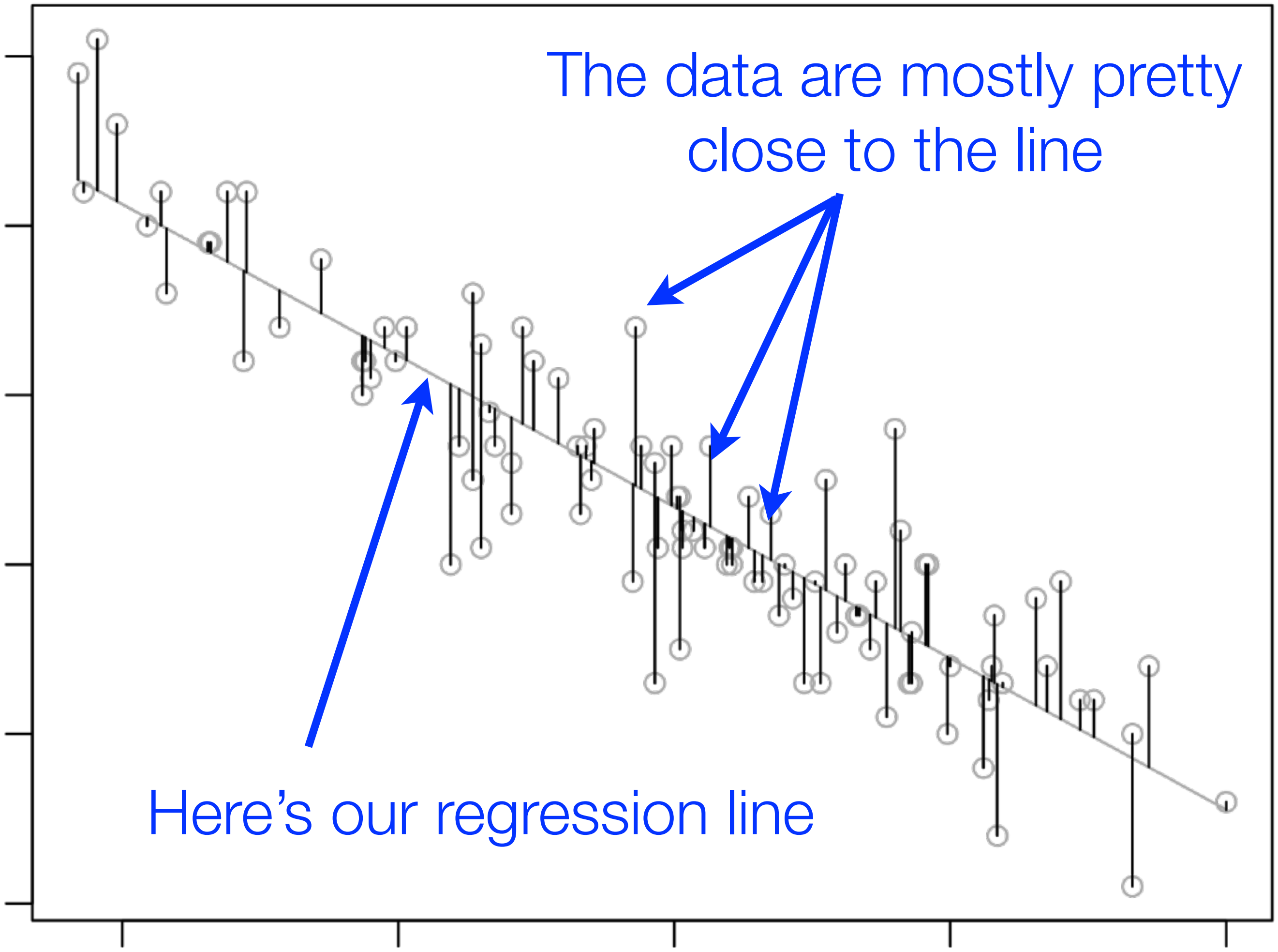


How do we know what the best fitting regression line is?

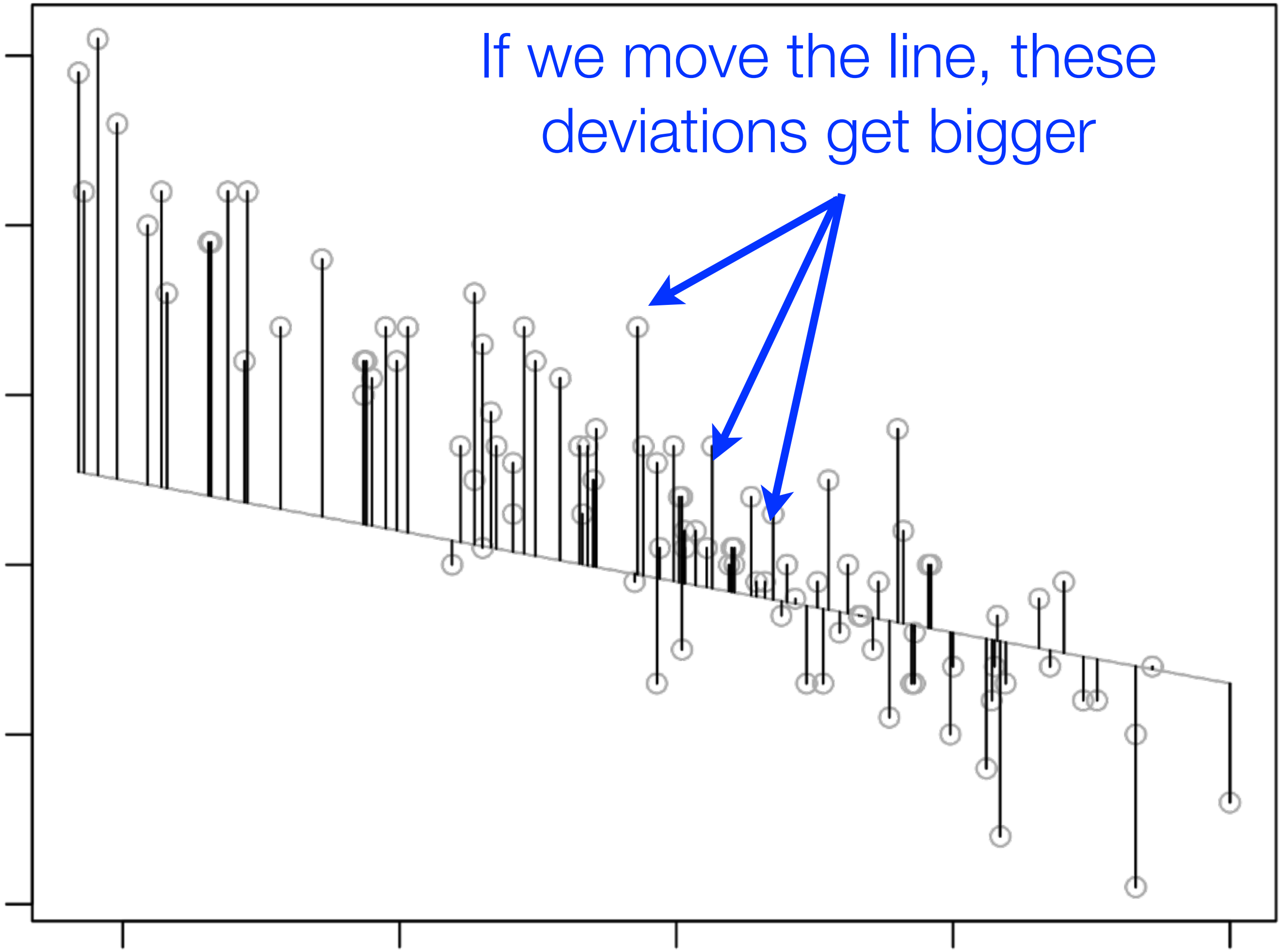
- In this case it's visually obvious:
  - It's a nice simple problem with one predictor X and one outcome Y, so the scatter plot makes it easy
  - Real life problems are rarely this helpful.

- We're going to need something a bit fancier than "just looking at it"

The best-fitting regression line



NOT the best-fitting regression line



The principle of "least squares"

- The best regression line for data (X,Y) is the one that minimises the sum squared deviation between the predictions $\hat{Y}_i$ and the actual values Y_i

$$\sum_i (Y_i - \hat{Y}_i)^2$$

What the actual value was for the i-th observation

What the regression model predicts for the i-th observation

The principle of "least squares"

- The best regression line for data (X,Y) is the one that minimises the sum squared deviation between the predictions $\hat{Y}_i$ and the actual values Y_i

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

- This is referred to as the <u>residual</u> sum of squares, and it's analogous to the within groups sum of squares (residuals) in ANOVA.
- Our goal is to estimate the values of b_0 and b_1 that minimise SS_{res}

And how do we do that?

- By using an ugly looking bit of matrix algebra, which is implemented using blah blah blah magic blah QR blah.
- You don't need to know it for this class (or ever) so I haven't included it.
- Here's a picture of a kitten instead:

Estimating a regression model in R

Regression in R

- Like ANOVA, regression is done in stages
  - 1. `lm()` estimates the values of b_0, b_1 etc
  - 2. `summary()` runs some hypothesis tests
  - 3. other functions to pull out things of interest

- The `lm()` function
  - This is the main "workhorse" function
  - It creates an "`lm`" object (i.e. variable), which contains lots of quantities of interest relating to regressions
  - Let's see how this works in practice...

Using the `lm()` function

- `lm()` is a very powerful function, with many arguments that you can play with
- We only need two:
  - `formula` : a formula specifying the regression model
  - `data` : the data frame

```
lm( formula = yield ~ diversity, data = d )
```

Running the regression

```
> model1 <- lm(yield ~ diversity, data=d )
```

The formula uses an outcome variable of `yield` and a predictor variable of `diversity` (i.e., how much is the yield of a plot of land affected by the diversity of the species doing the farming?)

The dataset is called `d`

This command asks R to estimate the regression model, and store the results in a variable called `model1`

Running the regression

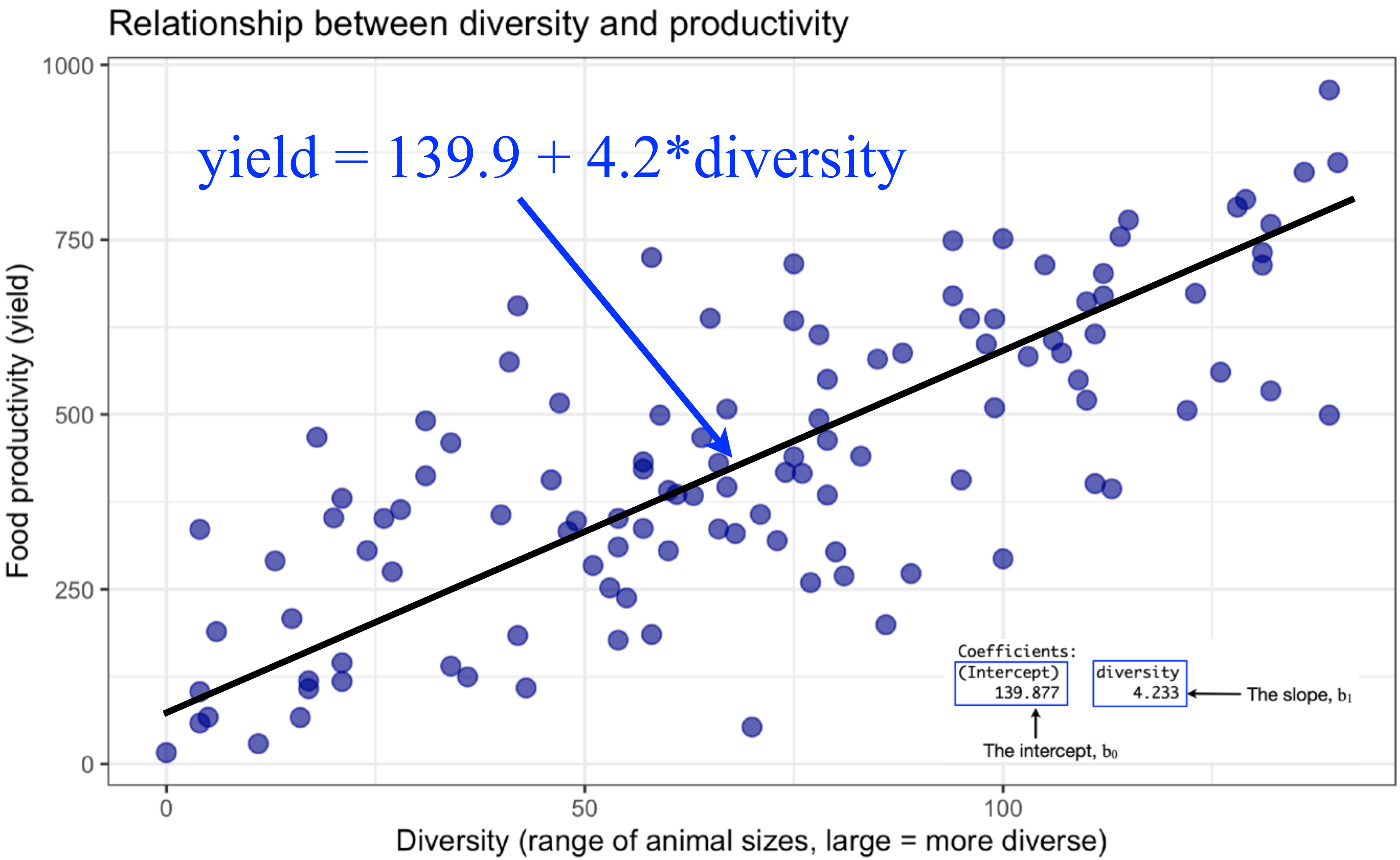
```
> model1 <- lm(yield ~ diversity, data=d )
> model1

Call:
lm(formula = yield ~ diversity, data = d)

Coefficients:
(Intercept)    diversity
    139.877        4.233
```

The slope, b_1

The intercept, b_0



- **Intercept:** If there was no diversity (i.e., max-min size was zero, everyone was the same size), you could expect about 139.9 units of food from that land
- **Slope:** For every additional unit of increase in diversity of range, you can expect about 4.2 more units of food from that land

So it really does look like having a range of sizes of species is associated with an increase in the productivity of the land! Often quite substantially!

Exercises are in w9day1exercises.Rmd