R Data Analysis Course – Week 6

Regressions and Correlation

Statistical Analysis

Ronnie Steinitz, Ph.D.
Dian Fossey Gorilla Fund

Common waxbill, Fossey Fund Campus, 2025

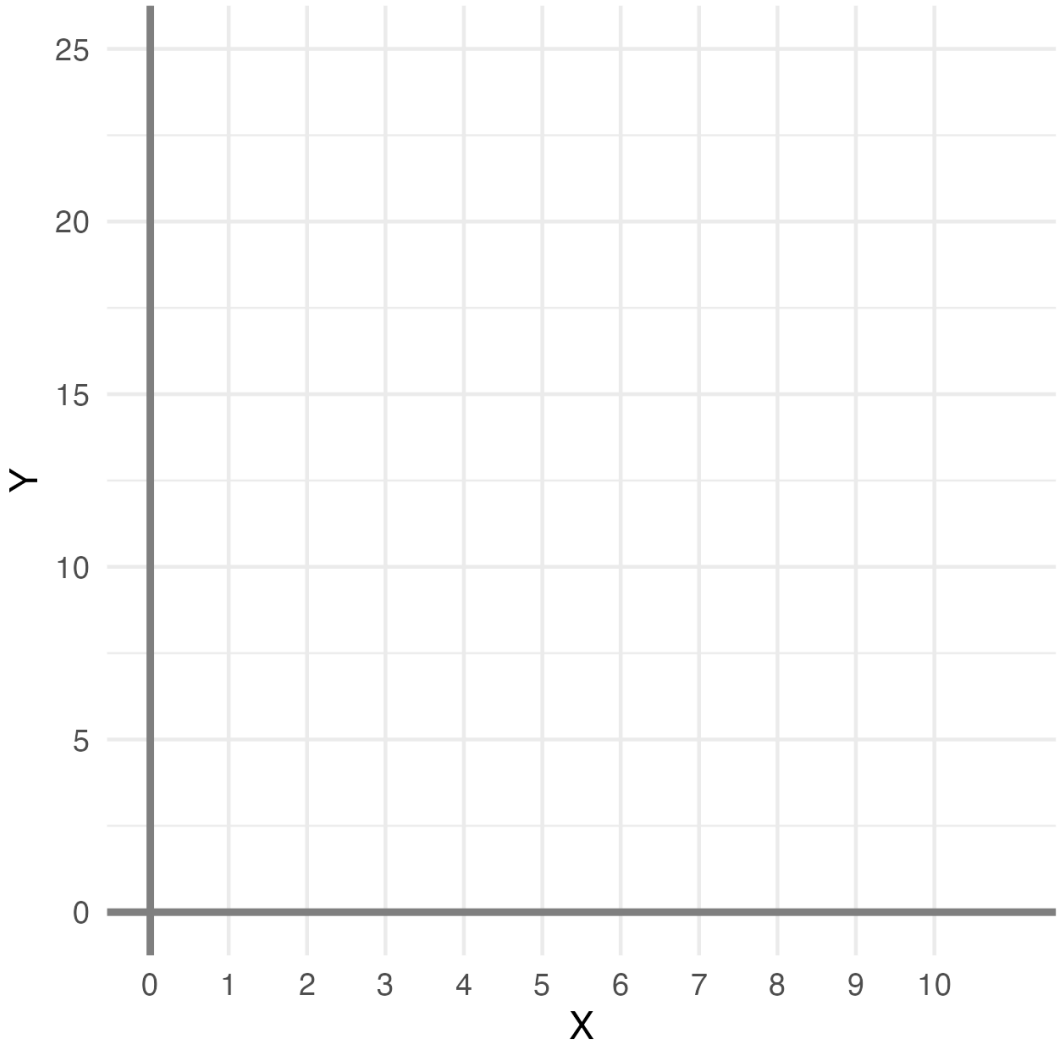
Resources to reinforce your understanding

The Google Drive folder includes:

- Every week: a **complete .Rmd file** with detailed in-line explanations and example code from the lecture
- General: an updated **"Functions Learned So Far"** reference sheet

Every week: I offer help sessions: extra support with your code, project setup, or any course concepts

What is a linear regression?

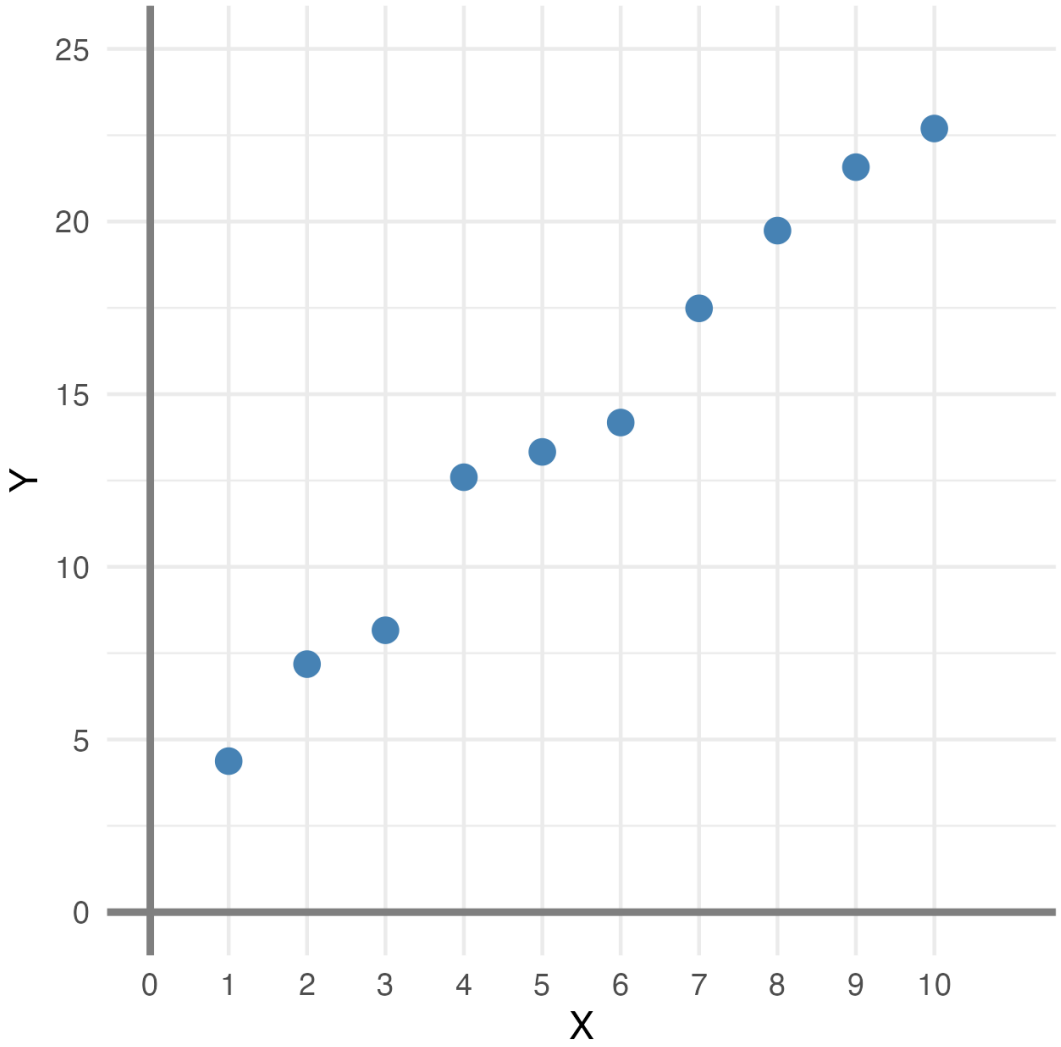


Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

What is a linear regression?

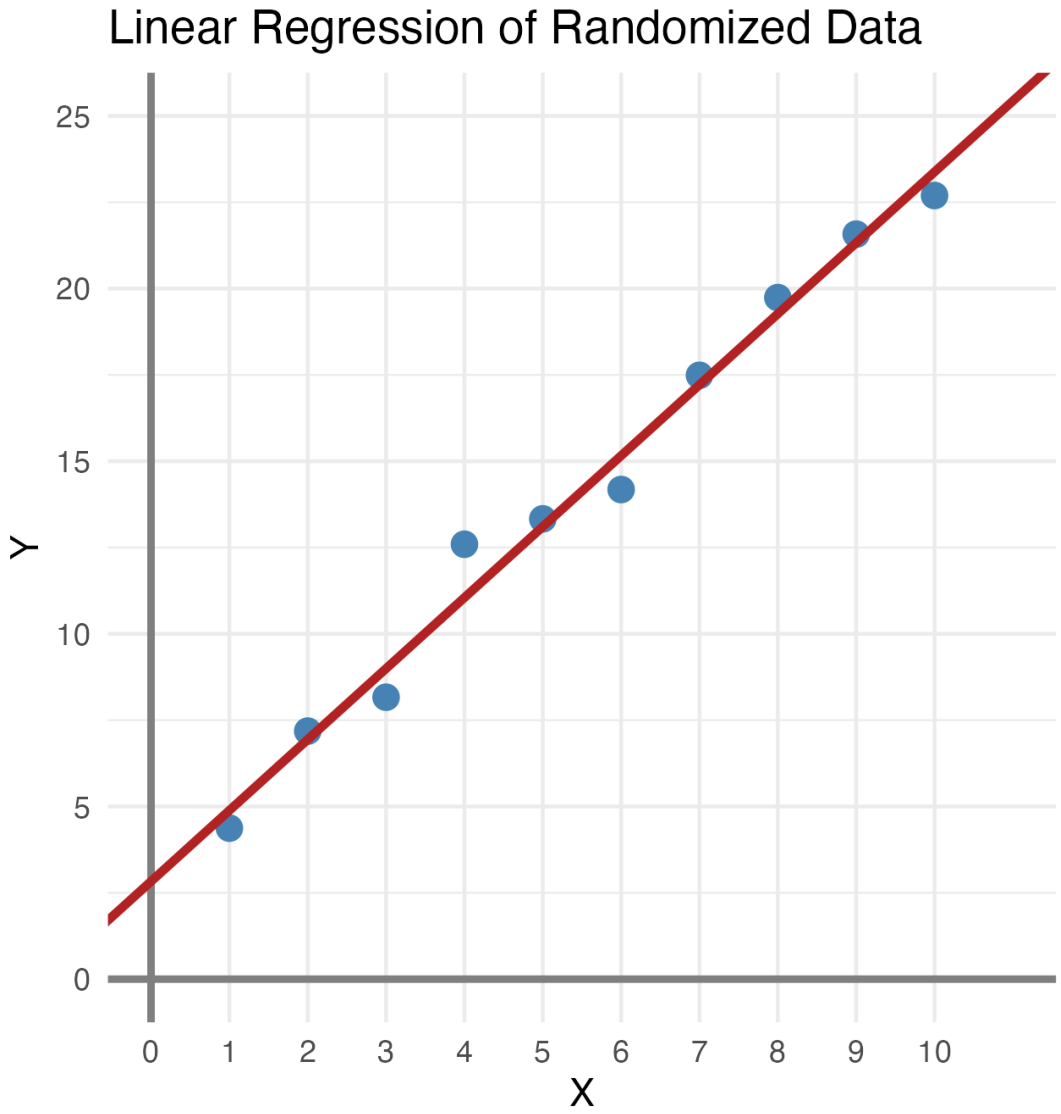
Points only = *scatterplot*



Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

What is a linear regression?



Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

Linear Regressions

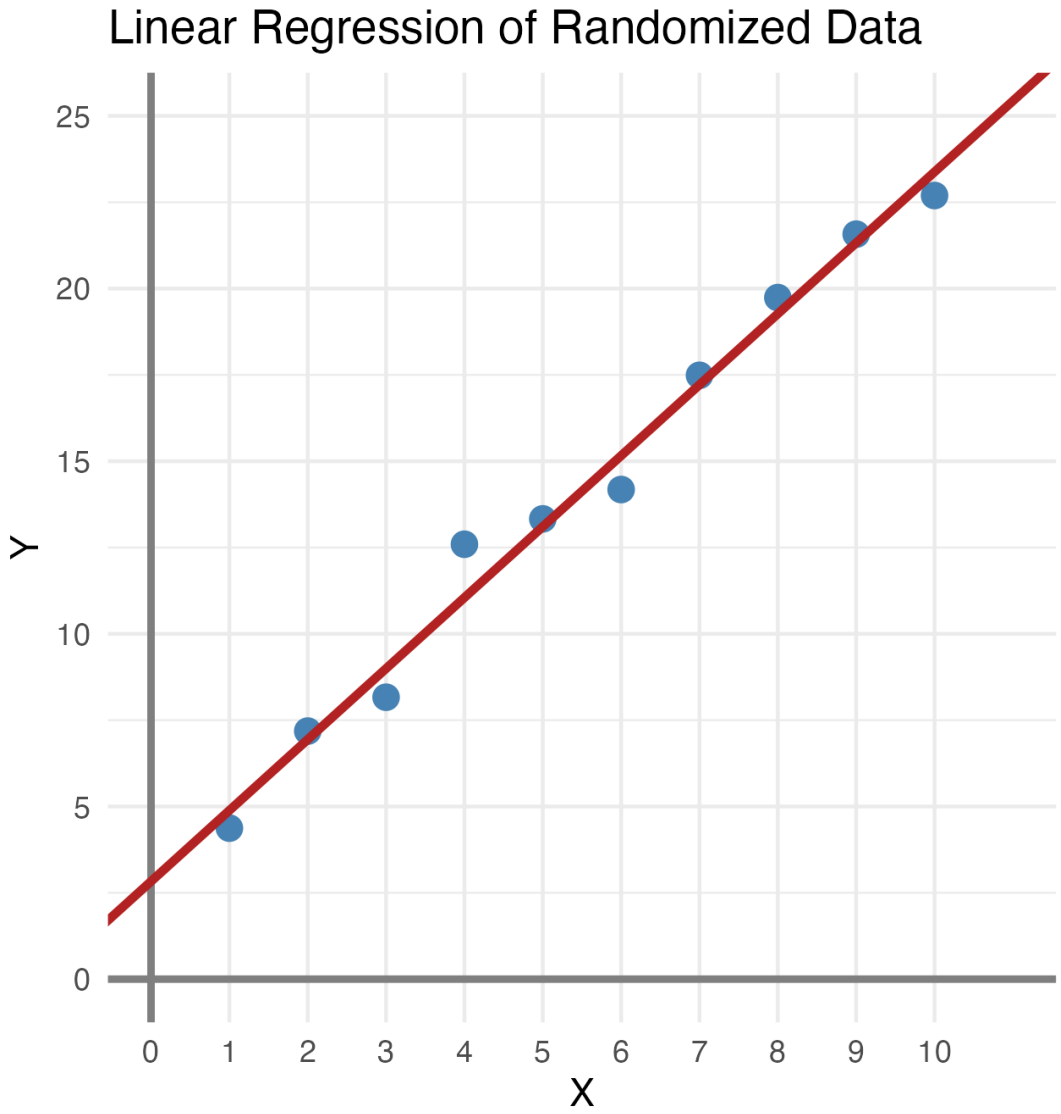
- A method for **quantifying relationships** between two continuous variables

$$Y = a + bX$$

- a = intercept (predicted Y when X = 0)
- b = slope (change in Y for each 1-unit change in X)

Regression line: a straight line that best represents the data, positioned so that the points are, on average, as close to the line as possible.

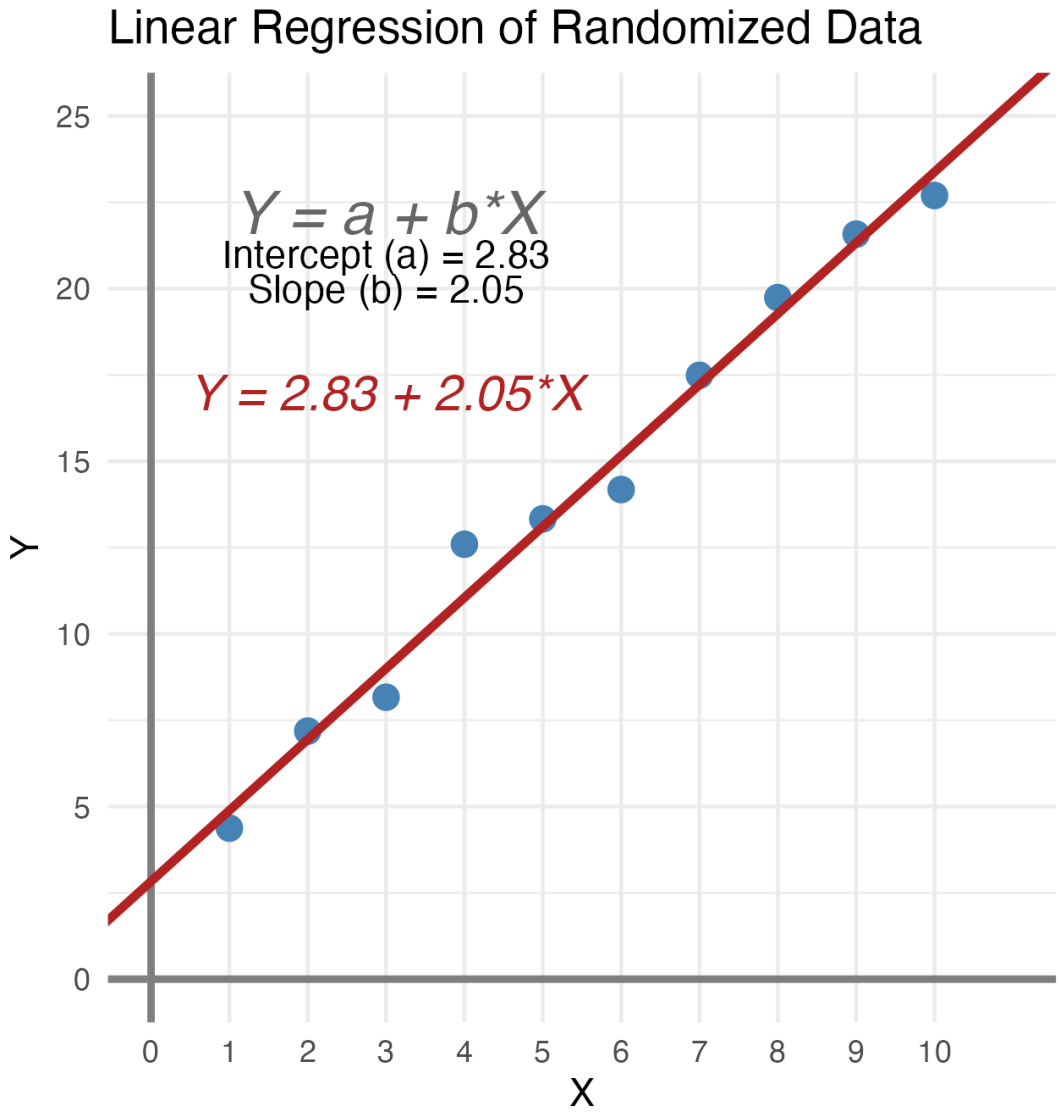
What is a linear regression?



Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

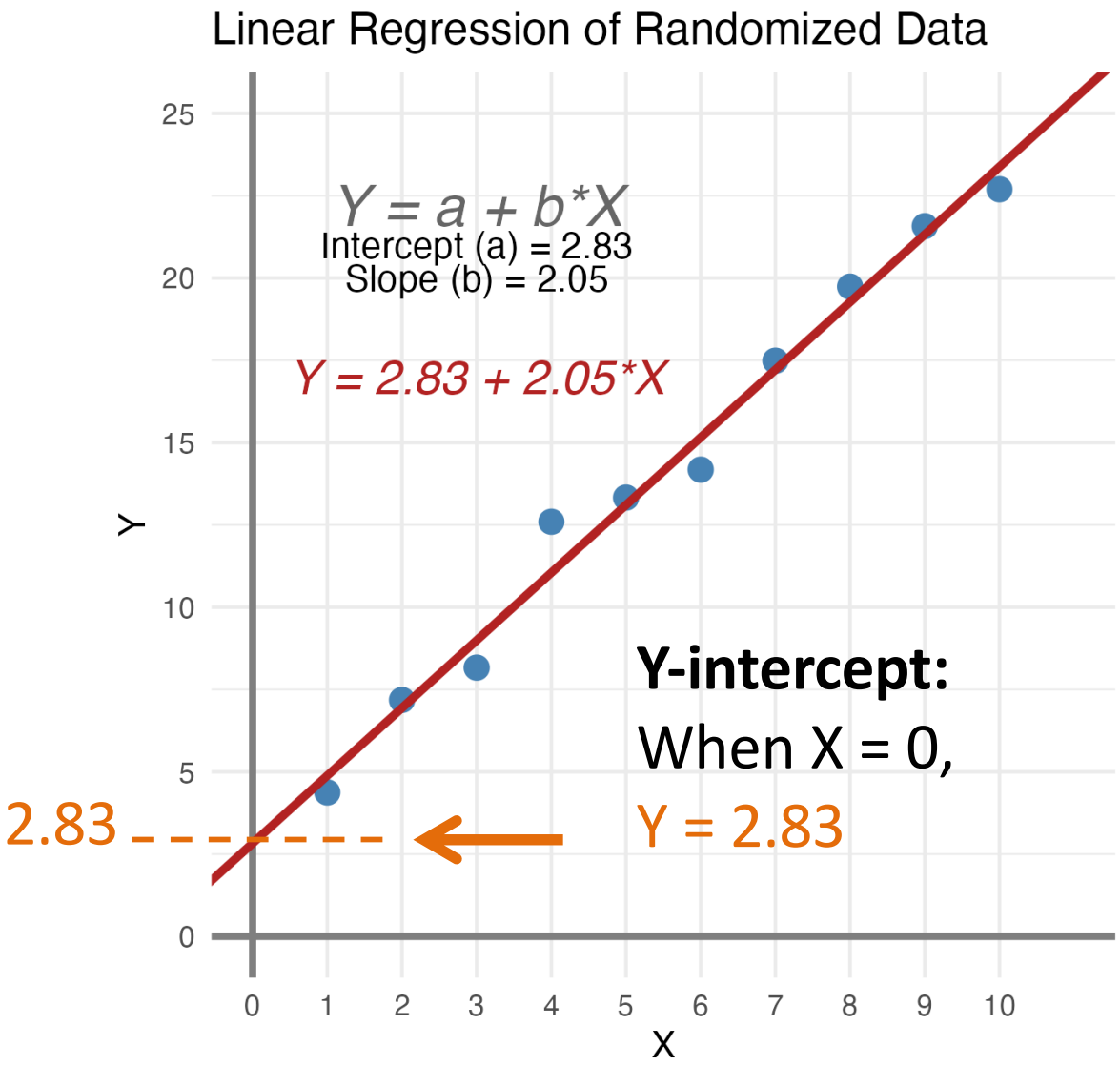
What is a linear regression?



Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

What is a linear regression?



Data

X	Y
9	21.575781
1	4.373546
6	14.179532
7	17.487429
5	13.329508
2	7.183643
3	8.164371
8	19.738325
10	22.694612
4	12.595281

Regression Statistics

`lm(data, y ~ x)`

"given X, what is Y?"

Model outputs:

- **Coefficients** → slope and intercept
- **Standard errors** → uncertainty around estimates
- **p-values** → are relationships statistically significant?
- **R^2** → how much of Y's variation is explained by X
  - i.e., a "goodness of fit" statistic

```
model <- lm(y ~ x, data = my_dataset)
summary(model)
```

"given x, what is y?"

```
Call:
lm(formula = y ~ x, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9800 -0.6410  0.2338  0.2678  1.5452

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  2.83118    0.55270   5.122     0.000904 ***
x            2.05473    0.08908  23.067 0.0000000132 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8091 on 8 degrees of freedom
Multiple R-squared:  0.9852,    Adjusted R-squared:  0.9833
F-statistic: 532.1 on 1 and 8 DF,  p-value: 0.00000001324
```

```
model <- lm(y ~ x, data = my_dataset)
summary(model)
```

"given x, what is y?"

Our model, evaluated as ***Y = a + bX***

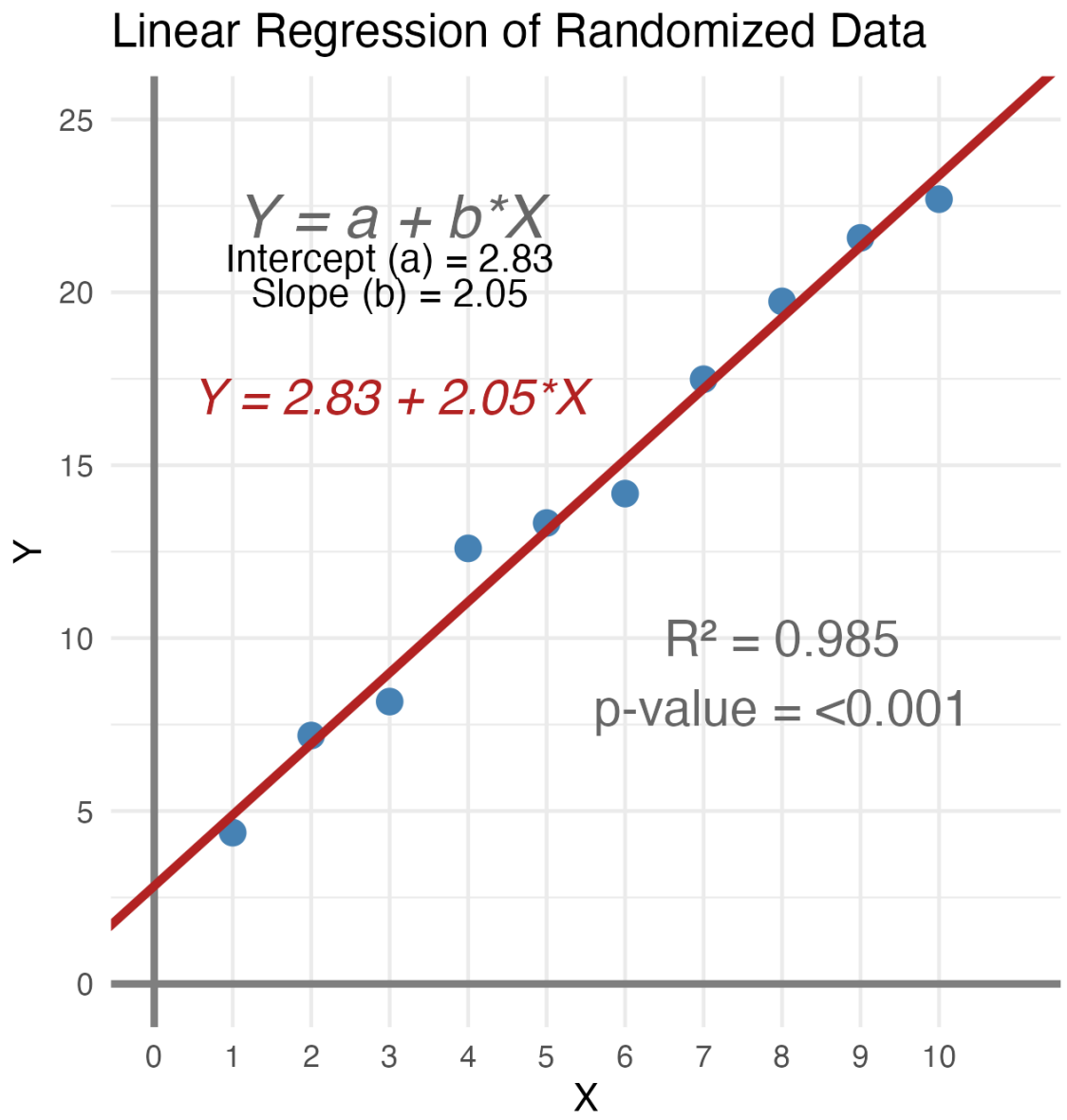
```
Call:
lm(formula = y ~ x, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9800 -0.6410  0.2338  0.2678  1.5452

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  2.83118    0.55270   5.122     0.000904 ***
x            2.05473    0.08908  23.067 0.0000000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8091 on 8 degrees of freedom
Multiple R-squared:  0.9852,    Adjusted R-squared:  0.9833
F-statistic: 532.1 on 1 and 8 DF,  p-value: 0.00000001324
```

Regression coefficients (Intercept) = ***a***, x = ***b***

p-value

Adj. R^2 - the proportion of variance in ***Y*** explained by the predictor (***X***)

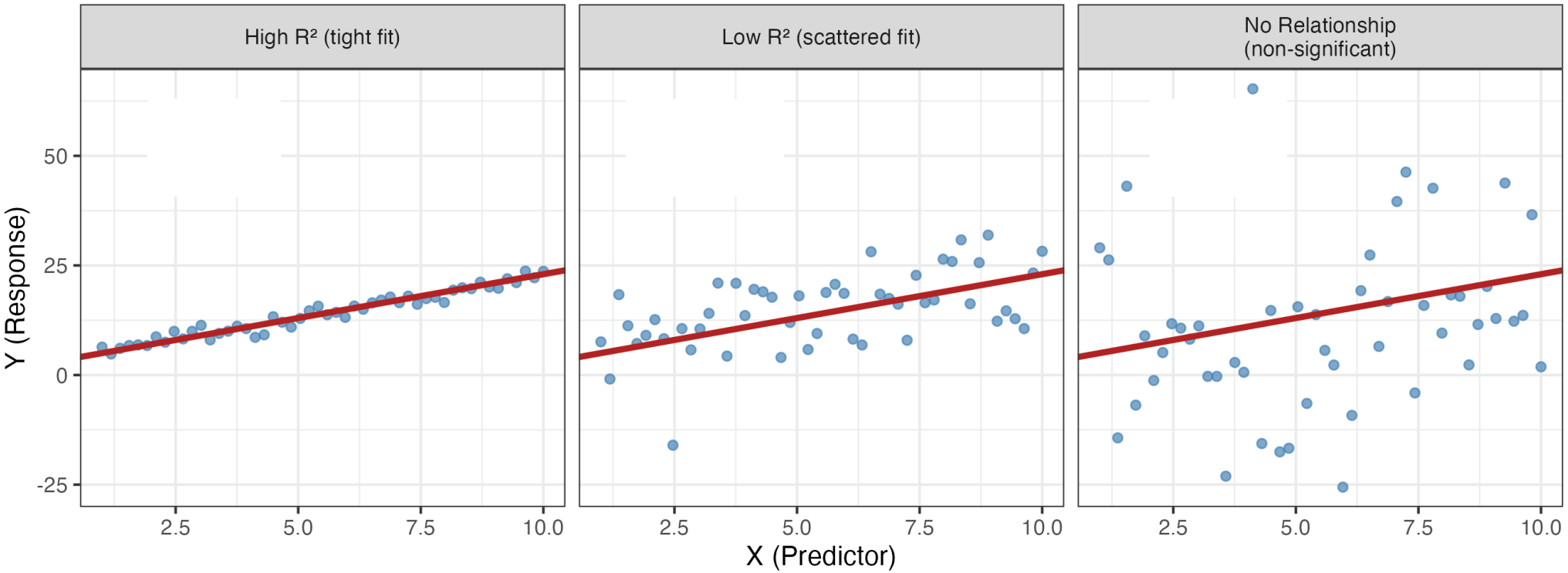
Regression Statistics



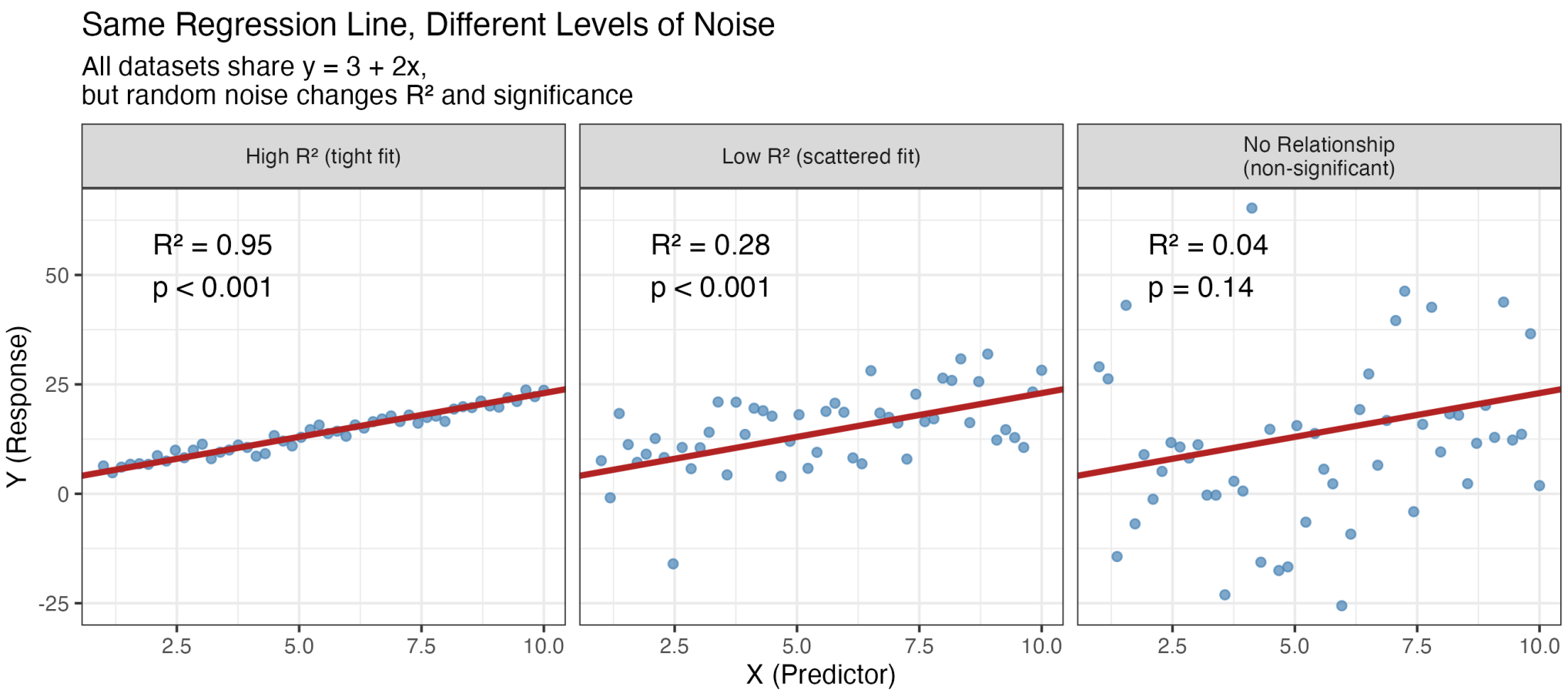
When $R^2 = 1$, every point falls exactly on the modeled line

A ***p-value*** < 0.05 means that ***X*** has a statistically significant relationship with ***Y***, and the result is unlikely to be caused by random chance alone

How well do the data fit?



How well do the data fit?



Highly significant, Very good fit

Highly significant, Slightly weaker fit

Not significant, Poor fit

Skills Learning

code-along lecture

Skills Application

Laboratory exercise

Week 6 - Skills Application

Download **Skills Application Instructions** from *Google Drive folder* > *Week 6*

If you are working on **your own** dataset:

- Keep working on your ***Lastname_Firstname_Data.Rmd***

If you are working on **sample** dataset:

- create a Markdown file just for this week.
- Name it: ***Lastname_Firstname_Week6.Rmd*** **in `/Week 6`**

In your new Markdown file: Use code blocks (Ctrl+Alt+i) to load packages, import dataset into the environment, and save data as an object (e.g., `*data <-* `). Then follow ***instructions***.