Simple linear regression I & II

Quantitative Methods for International Politics

IPOL 3270 • Fall 2026

September 27, 2026



Plan for today

Relationships between variables

Correlation

Drawing lines

Lines and regression

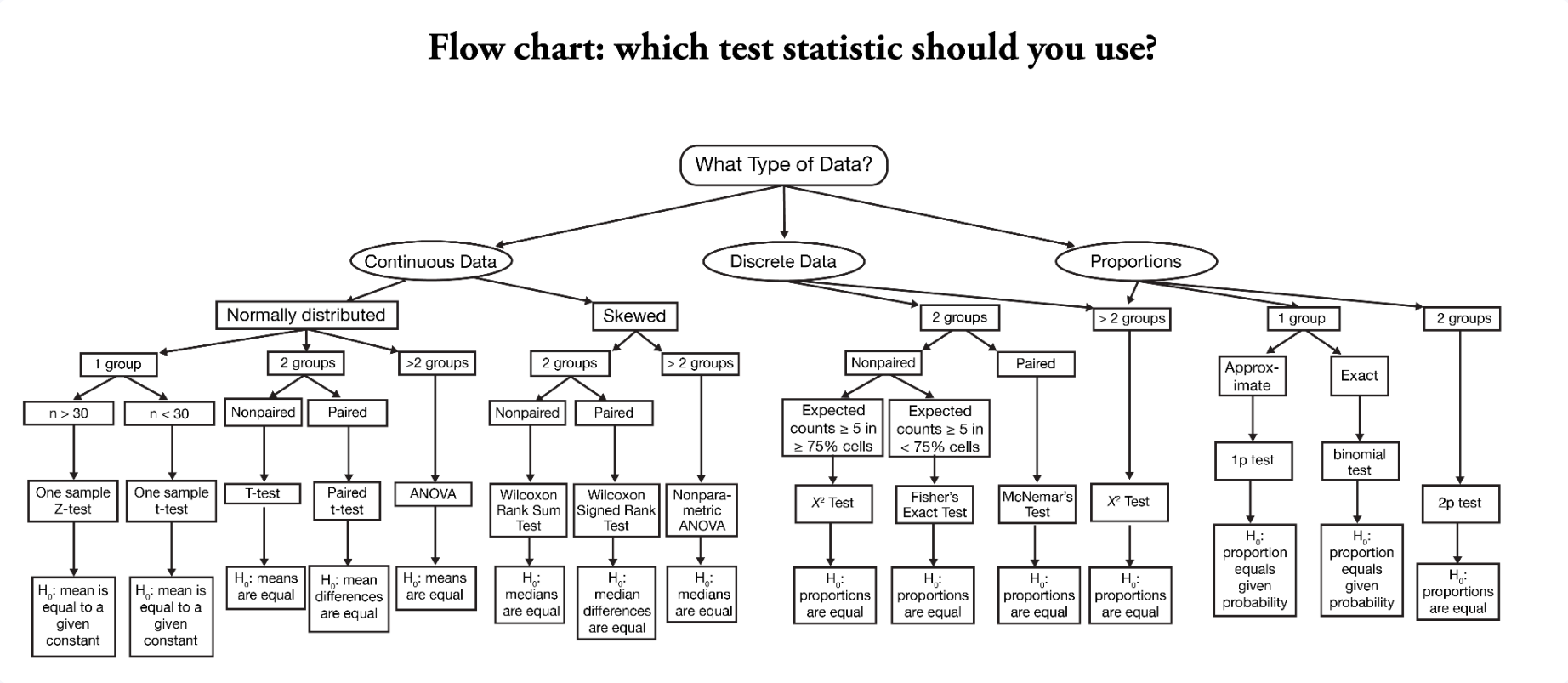
Categories and regression

Relationships between variables

Why statistics?

- We want to know the relationships and differences between variables!
- We want to know if those relationships and differences are “real”

Standard stats classes



Everything in statistics can be done with regression!

- Sample averages
- Differences in means
- Differences in proportions
- Slopes

Essential parts of regression

Y

Outcome variable

Response variable

Dependent variable

Thing you want to explain or predict

X

Explanatory variable

Predictor variable

Independent variable

Thing you use to explain or predict Y

Your turn #1: Identify variables

- A study examines the effect of smoking on lung cancer
- Researchers predict genocides by looking at negative media coverage, revolutions in neighboring countries, and economic growth
- You want to see if taking more honors classes in secondary school improves university grades
- Netflix uses your past viewing history, the day of the week, and the time of the day to guess which show you want to watch next

04:00

Two purposes of regression

Prediction

Forecast the future

Focus is on Y

Netflix trying to guess your next show

Predicting which countries will have a coup

Explanation

Explain effect of X on Y

Focus is on X

Netflix looking at the effect of the time of day on show selection

Measuring the effect of foreign aid on democratization

Basic steps for regression

1. Plot **X** and **Y**
2. Check how strongly they move together
3. Draw a line that approximates the relationship (and that would plausibly work for data not in the sample!)
4. Find mathy parts of the line
5. Interpret the math

Fertility and life expectancy

Do countries with longer life expectancy have lower fertility rates?

```
library(moderndive)

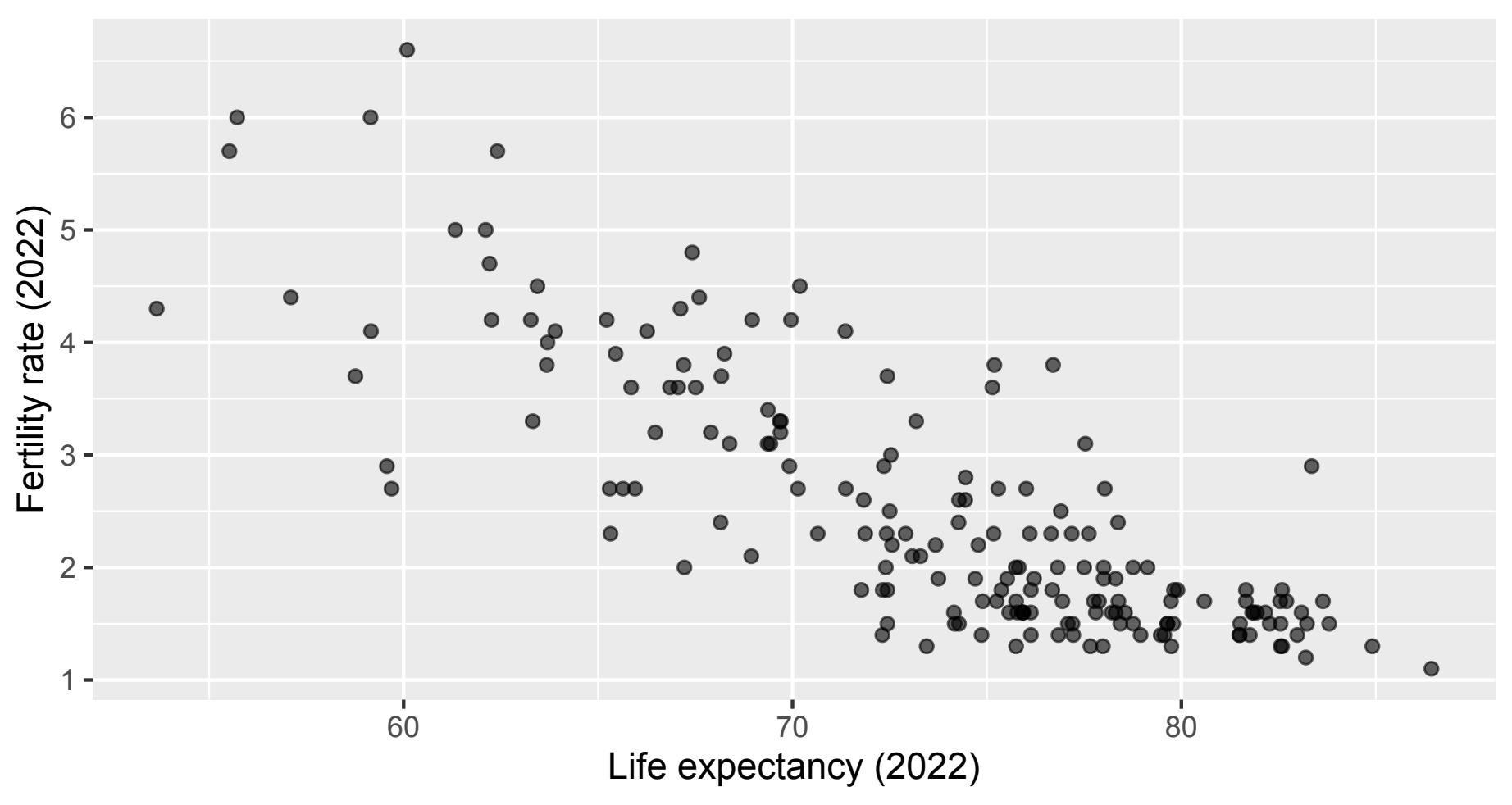
UN_data_ch5 <- un_member_states_2024 |>
  select(
    iso,
    country,
    life_exp = life_expectancy_2022,
    fert_rate = fertility_rate_2022,
    obes_rate = obesity_rate_2016
  ) |>
  na.omit()
```

- `fert_rate`: Expected number of children born per woman (**Y**)
- `life_exp`: Average life expectancy in years (**X**)
- `obes_rate`: % of adults who are obese
- 181 countries

Fertility and life expectancy

iso	country	life_exp	fert_rate
AFG	Afghanistan	53.65	4.3
ALB	Albania	79.47	1.4
DZA	Algeria	78.03	2.7
AGO	Angola	62.11	5.0
ATG	Antigua and Barbuda	77.80	1.6
ARG	Argentina	78.31	1.9
ARM	Armenia	76.13	1.6
AUS	Australia	83.09	1.6
AUT	Austria	82.27	1.5
AZE	Azerbaijan	74.15	1.6

Plot X and Y



Correlation

Univariate vs. bivariate

Univariate summaries	Bivariate summaries
Describe one variable	**Describe how two variables are related**
Mean, median, standard deviation, minimum, maximum, etc.	Correlation!

Correlation coefficient (r)

A number that measures the **strength** and **direction** of the **linear** relationship between two numeric variables

Written as r; ranges between –1 and 1

–1

Perfect negative relationship: as **X** goes up, **Y** goes down

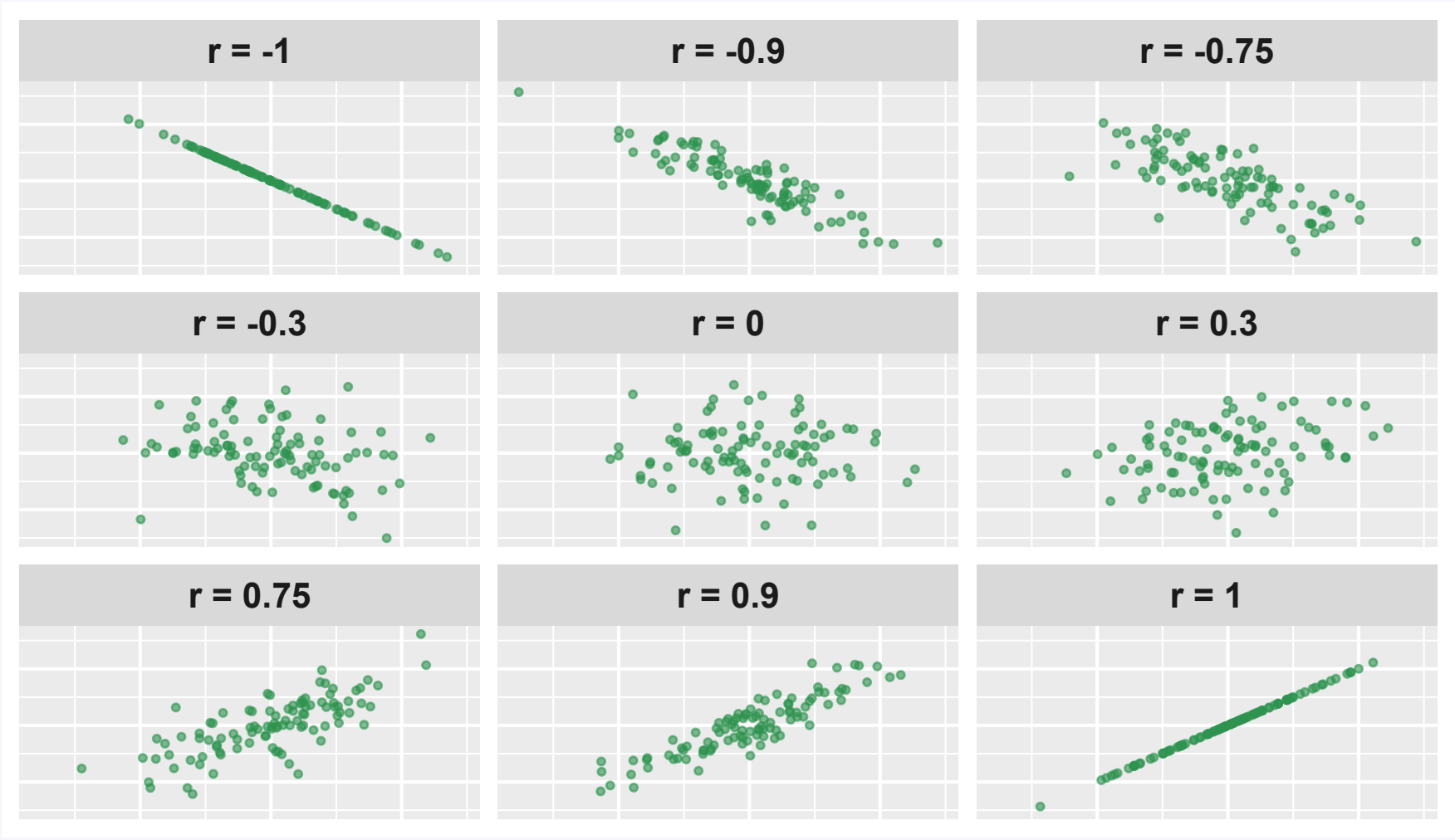
0

No relationship: **X** and **Y** move independently

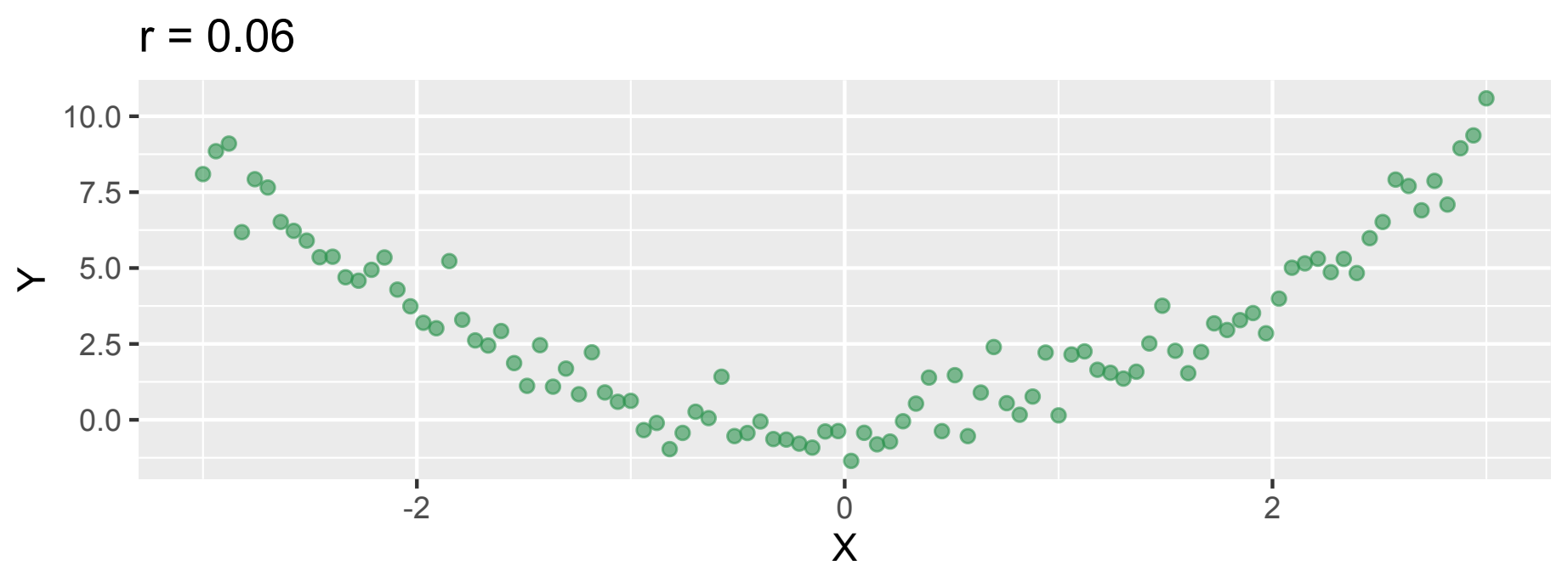
+1

Perfect positive relationship: as **X** goes up, **Y** goes up

What do correlations look like?



Correlation only measures *lines*



Clear relationship, but r ≈ 0
because it's not a straight line!

Guess the correlation!

https://www.guessthecorrelation.com/

How close can you get?

04:00

Fertility and life expectancy

What's your guess?

```
UN_data_ch5 |>
  get_correlation(
    fert_rate ~ life_exp
  )
```

```
# A tibble: 1 × 1
     cor
   <dbl>
1 -0.812
```

```
# Or with regular {dplyr}
UN_data_ch5 |>
  summarize(
    r = cor(fert_rate, life_exp)
  )
```



Interpreting correlation

r = -0.812

Fairly strong negative relationship!

Countries with higher life expectancy tend to have lower fertility rates

Rough guidelines (though these are subjective!):

- ±0.1–0.3: Weak
- ±0.3–0.7: Moderate
- ±0.7–1.0: Strong

Correlation doesn't tell us much

How much does fertility change when life expectancy goes up by one year?

Correlation only tells us direction and strength, not size!

What does life expectancy *do* to fertility?

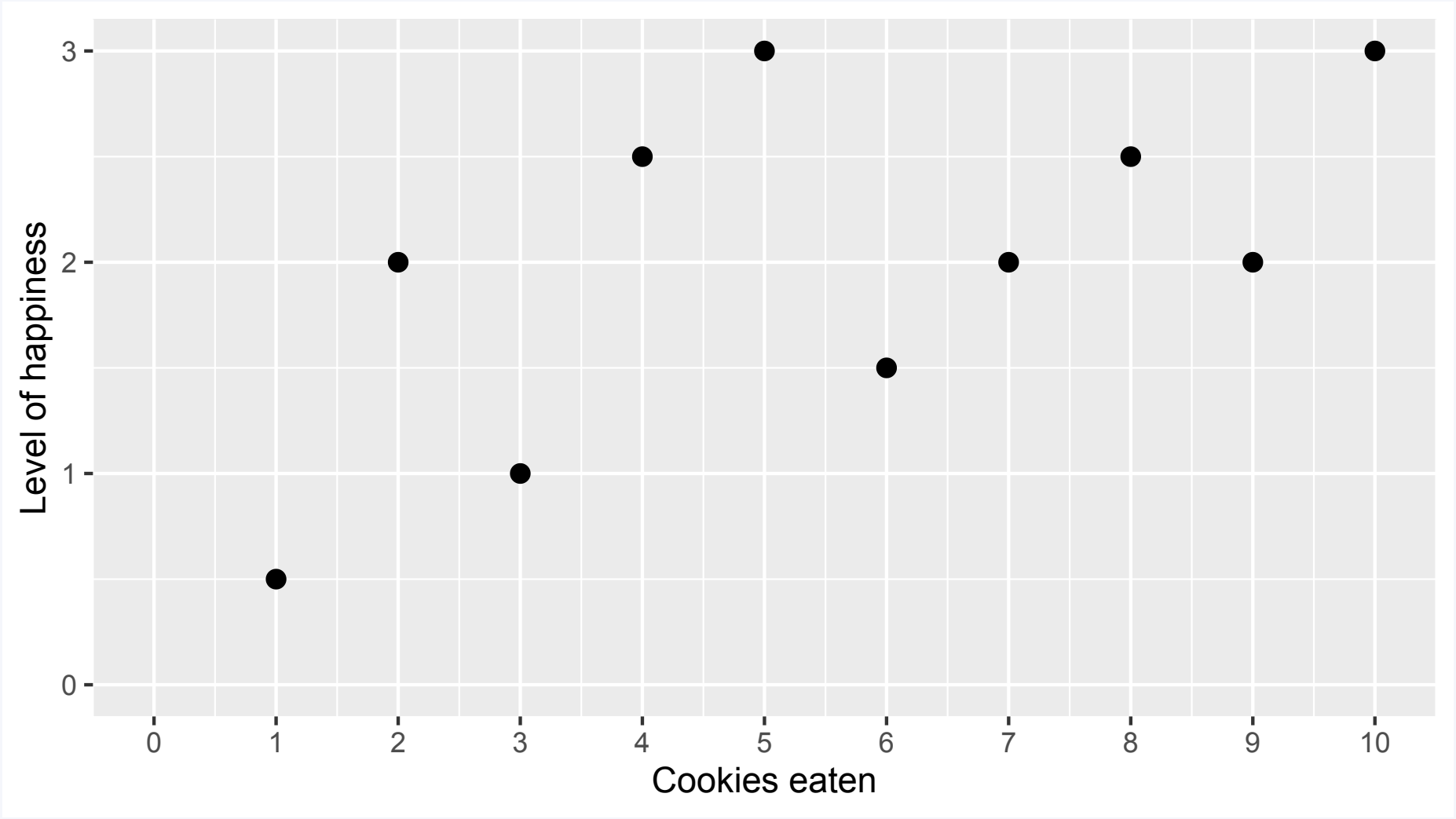
Correlation isn't causation! Life expectancy doesn't *cause* fertility—wealth, education, access to health care, etc. all influence both

For real details about the relationship, we need a line

Drawing lines

Cookies and happiness

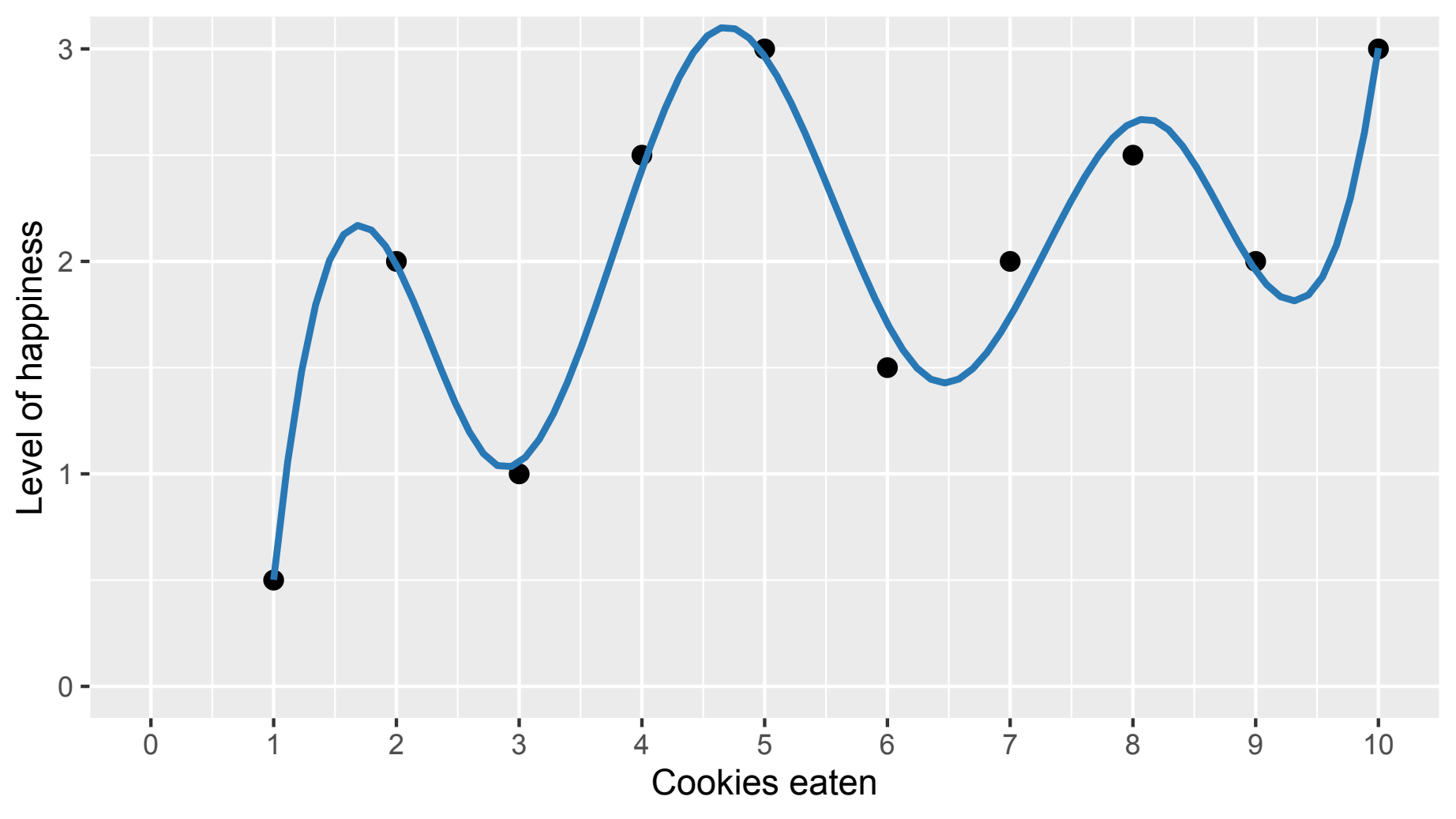
cookies	happiness
1	0.5
2	2.0
3	1.0
4	2.5
5	3.0
6	1.5
7	2.0
8	2.5
9	2.0
10	3.0

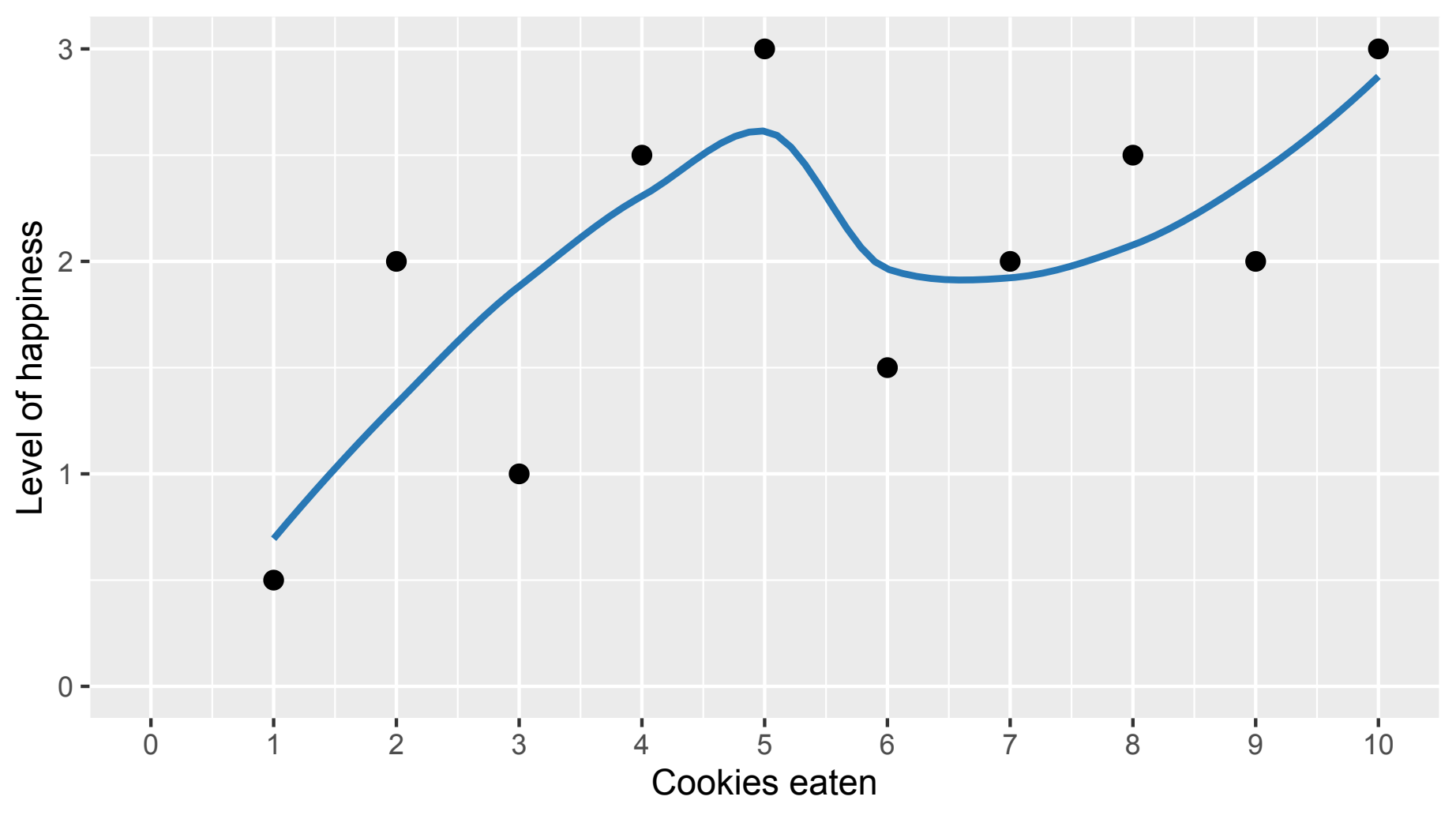

Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten

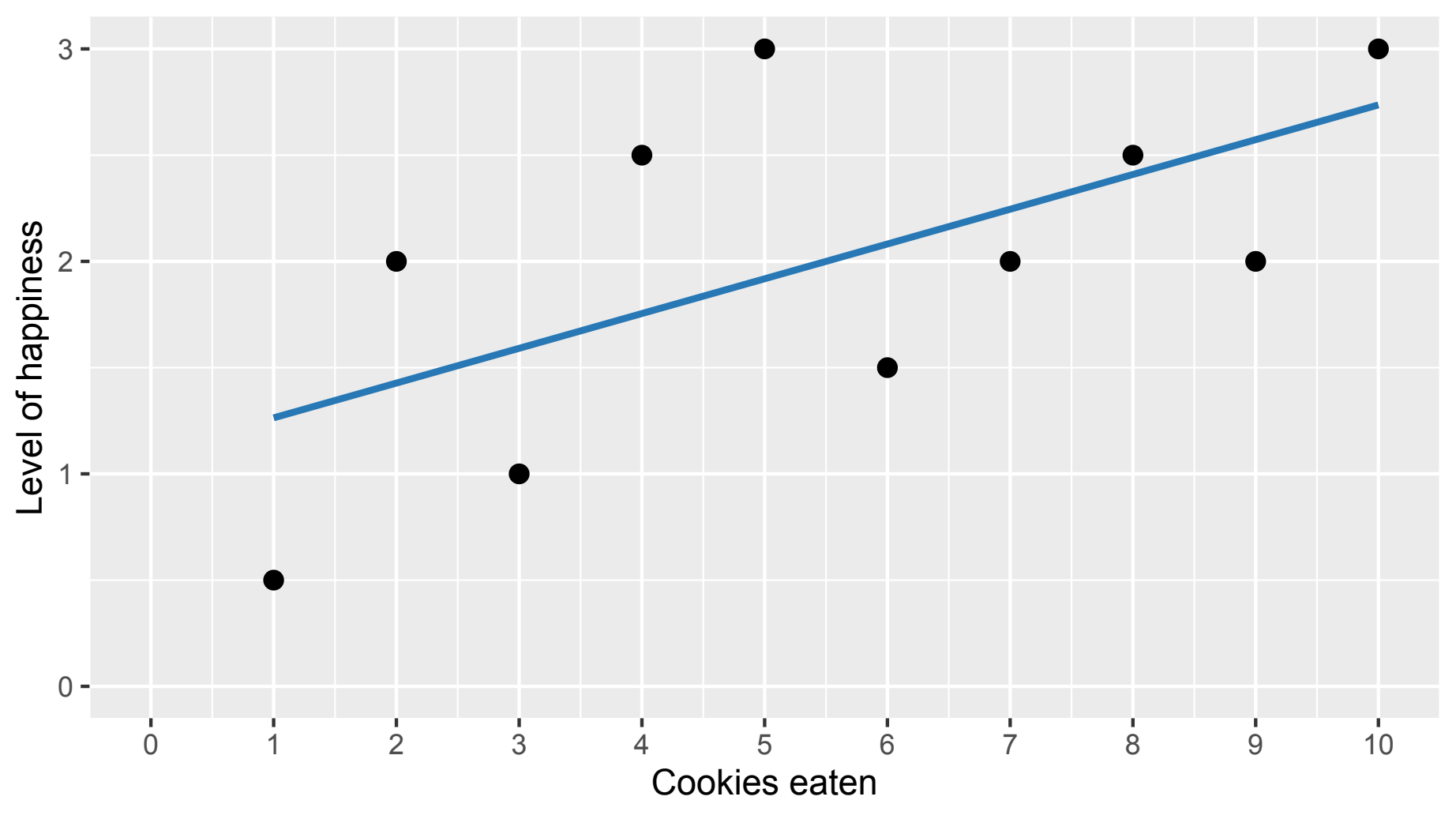
Draw your best line!

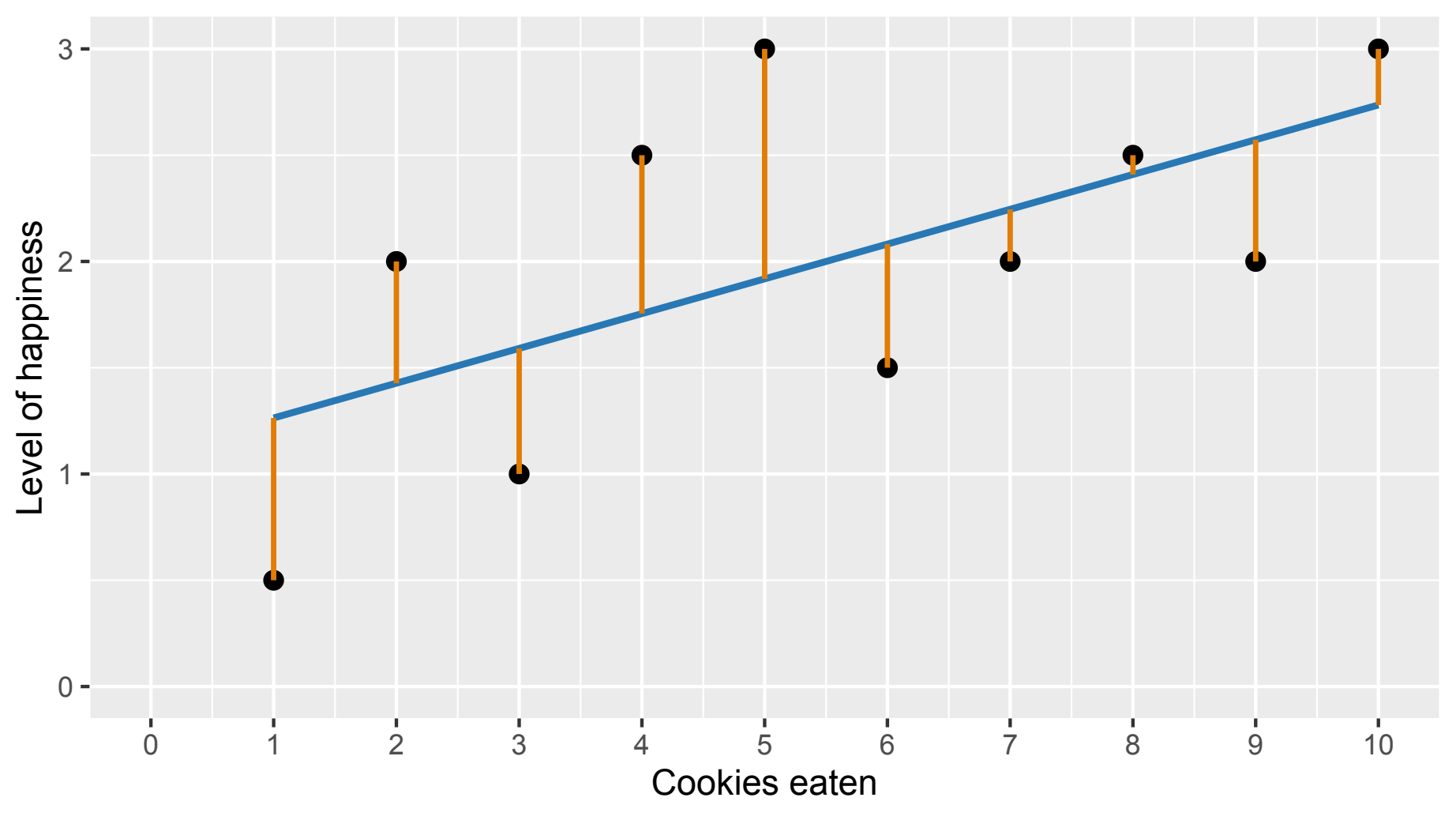
Draw a *straight* line that fits the points the best

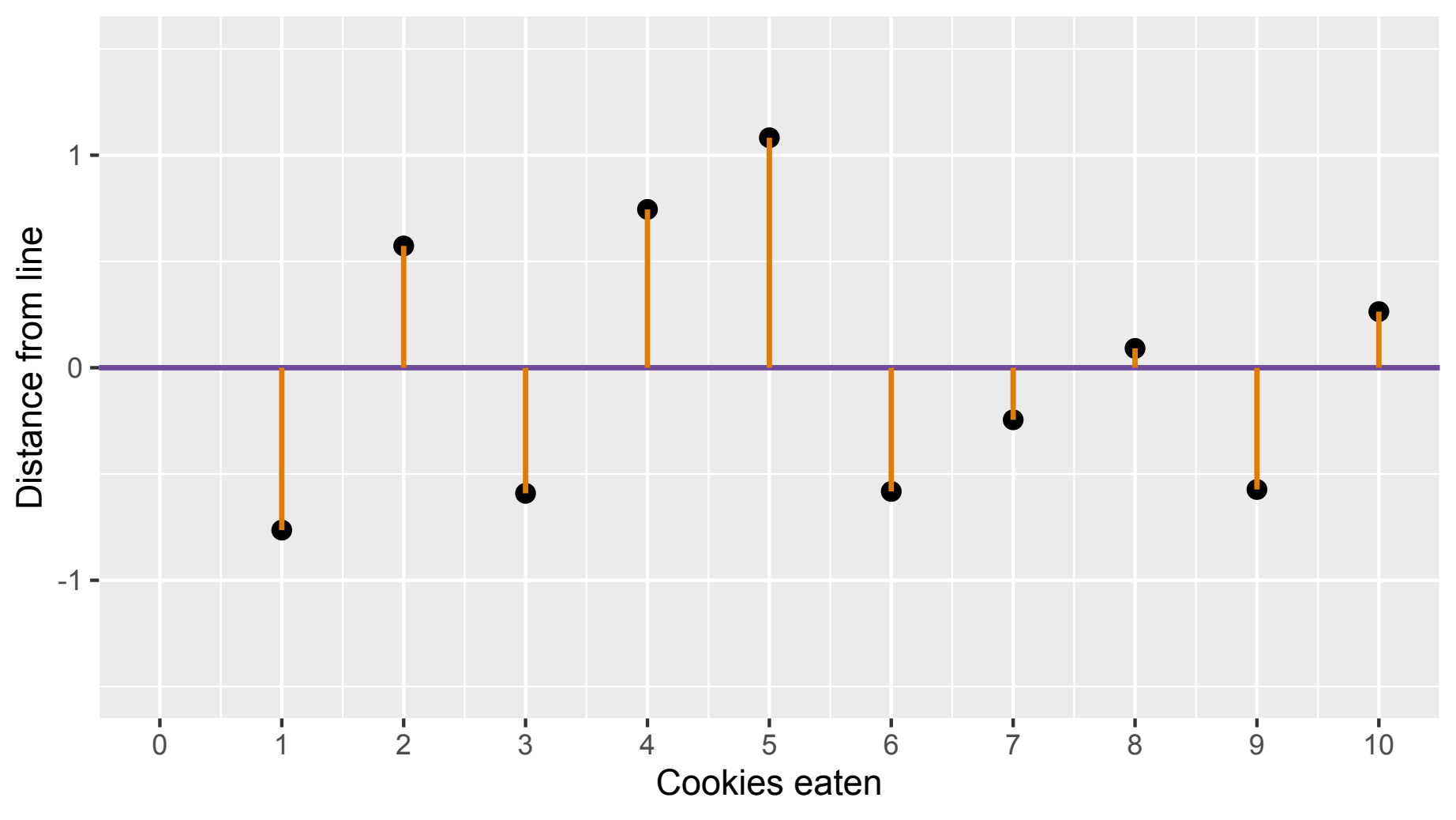
05:00

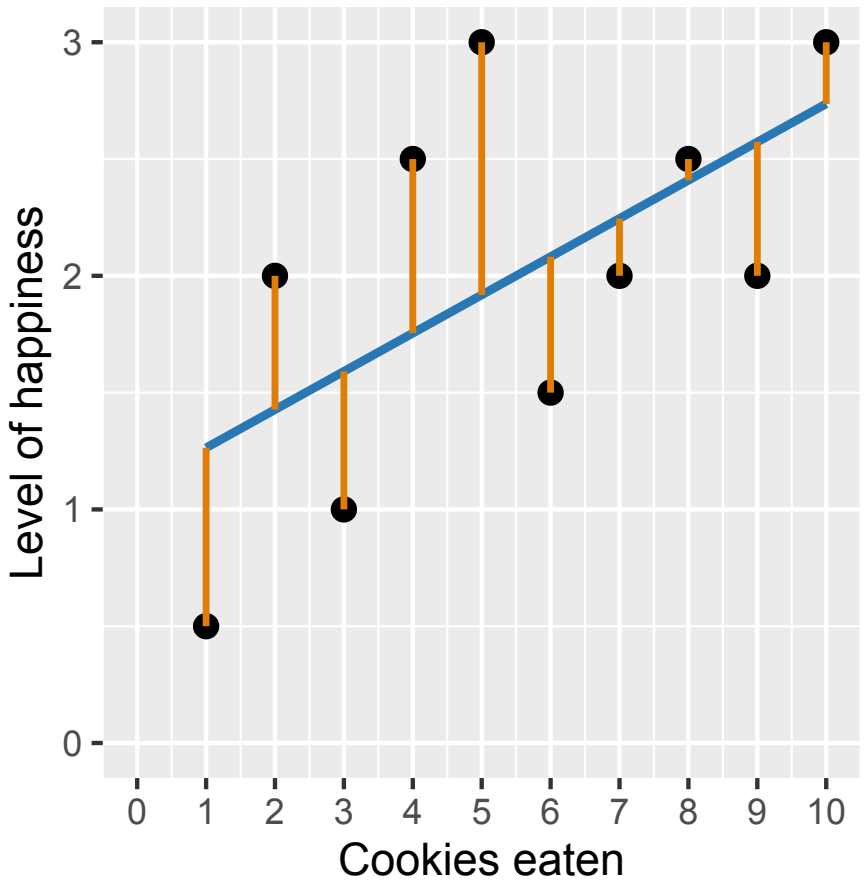

Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten

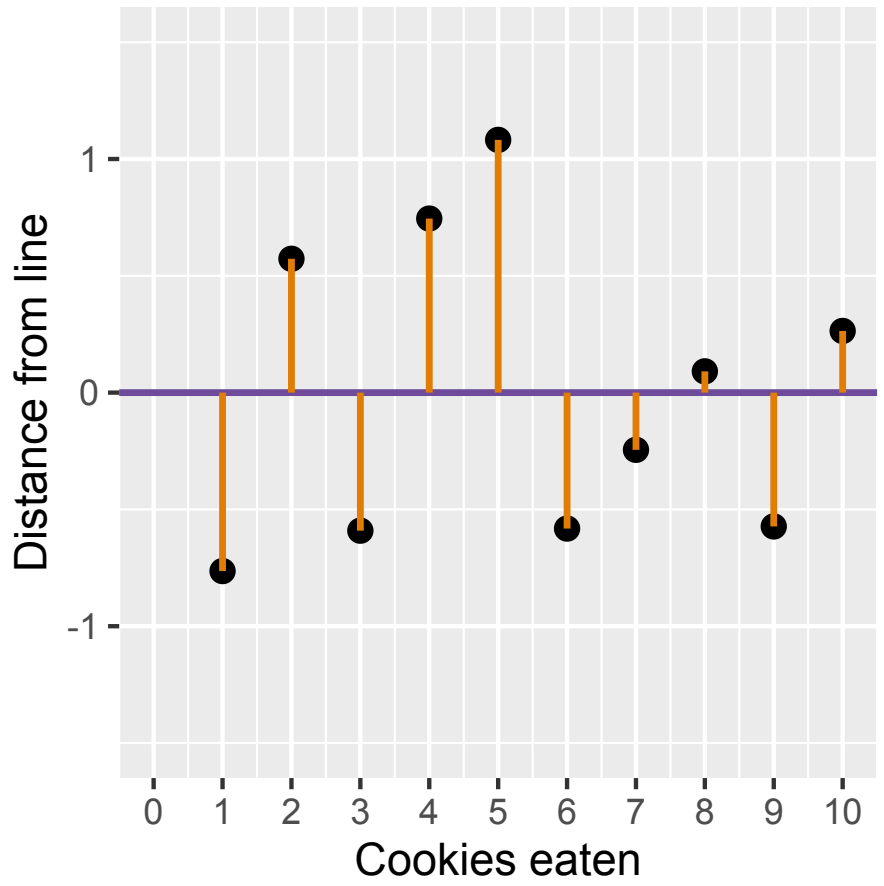

Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Distance from line
1
0
-1
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Cookies and happiness
Level of happiness
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
Cookies eaten


Residuals
Distance from line
1
0
-1
0 1 2 3 4 5 6 7 8 9 10
Cookies eaten

Residuals

Residual = observed value – value on the line

$$\text{residual} = y - \hat{y}$$

The line's "error" or "lack of fit" for each observation

Calculate differences

Calculate how far each point is from your line and square it

05:00

Residuals

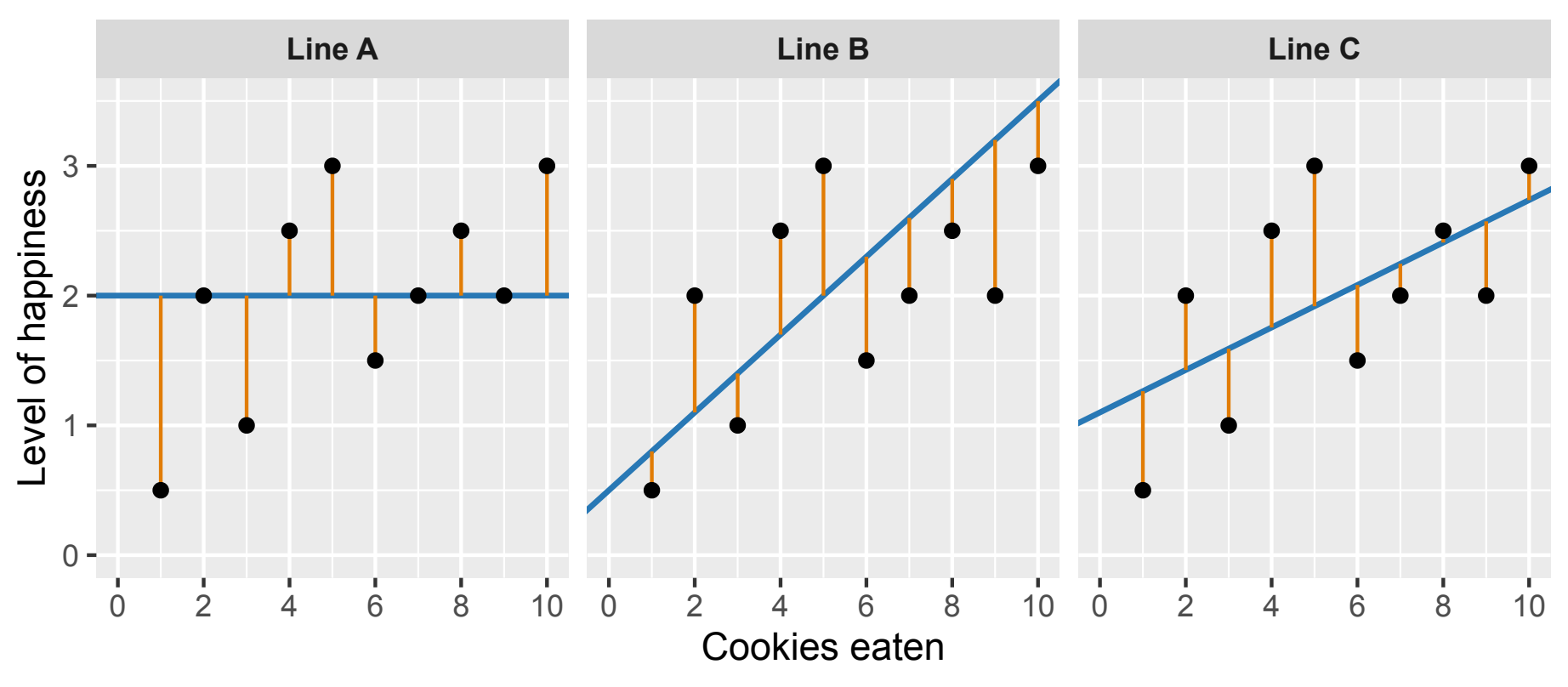
Residual = observed value – value on the line

$$\text{residual} = y - \hat{y}$$

The line's "error" or "lack of fit" for each observation

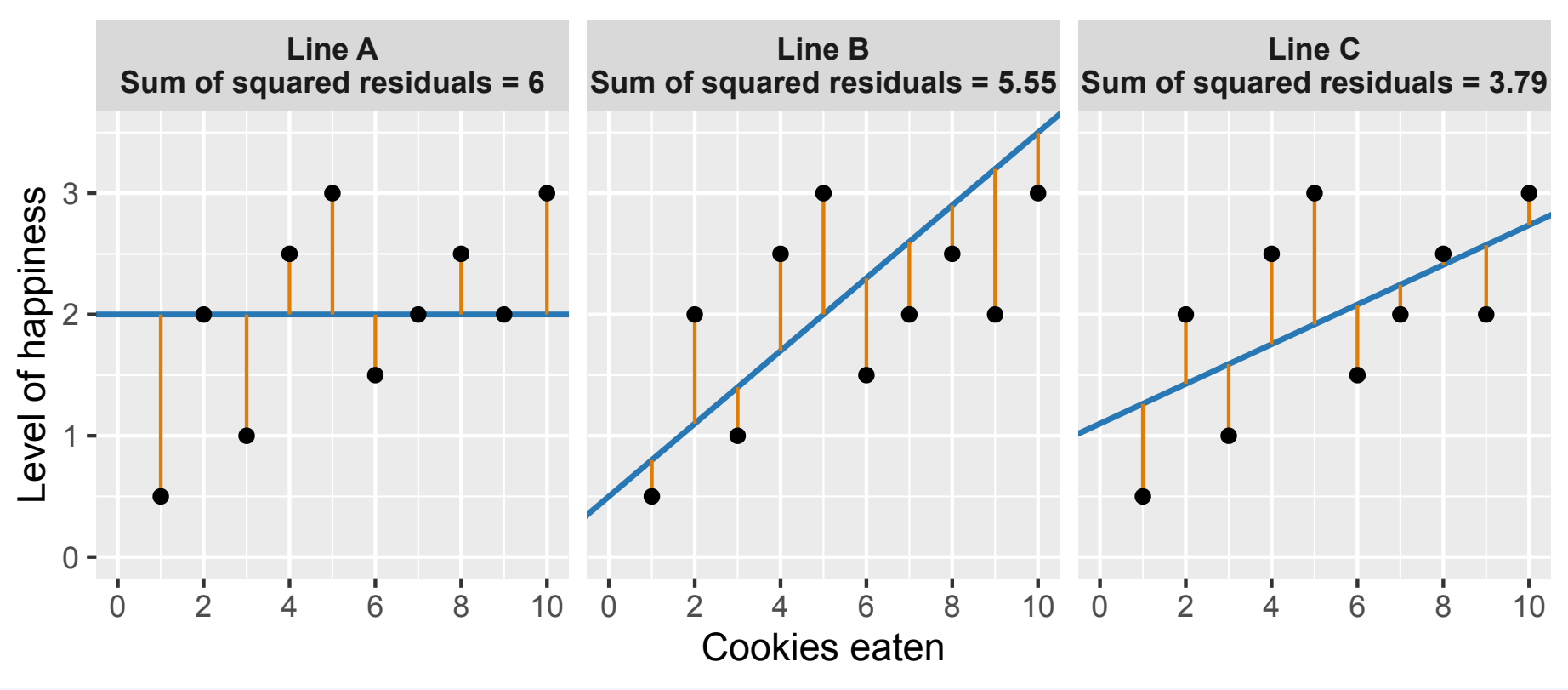
Good lines have small residuals

Which line is best?



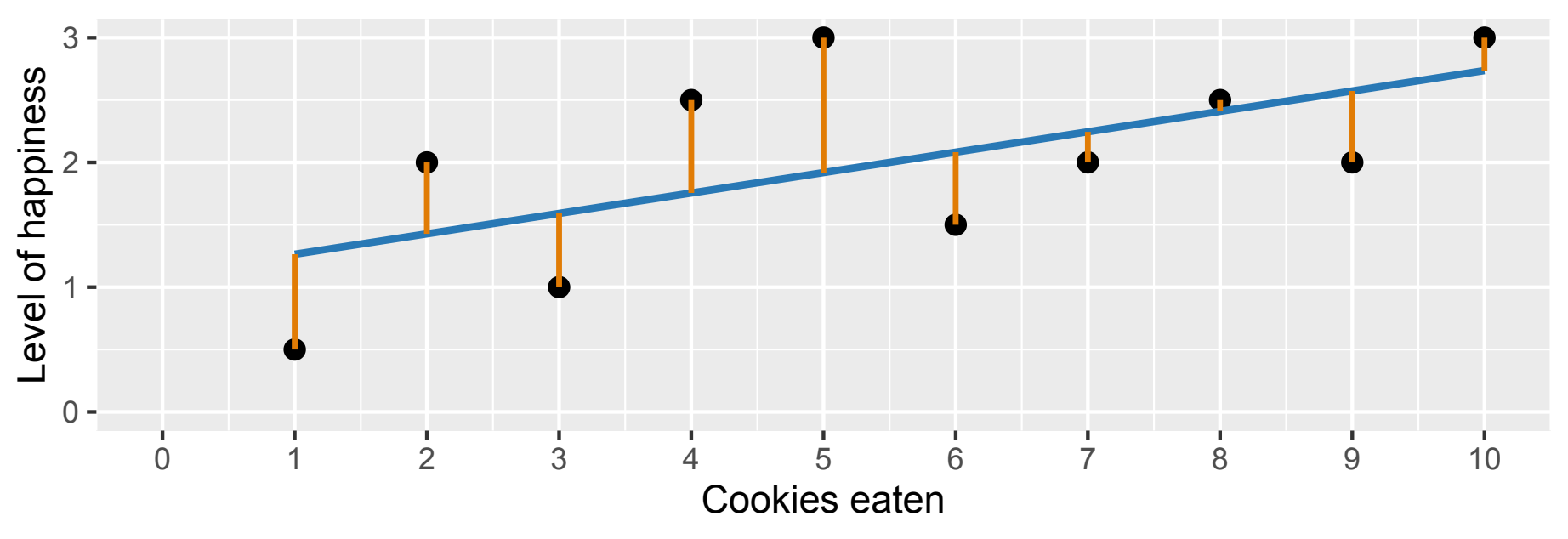
Ordinary least squares (OLS)

1. Square each residual
2. Add them all up



Ordinary least squares (OLS)

The “best-fitting” line is the one with the **smallest sum of squared residuals**



Lines and regression

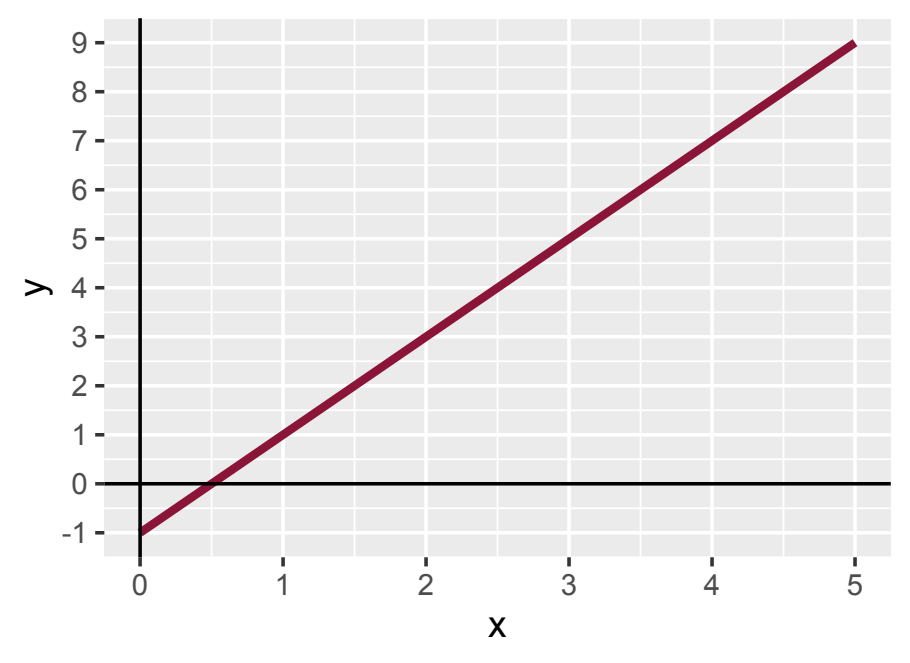
Drawing lines with math

$$y = mx + b$$

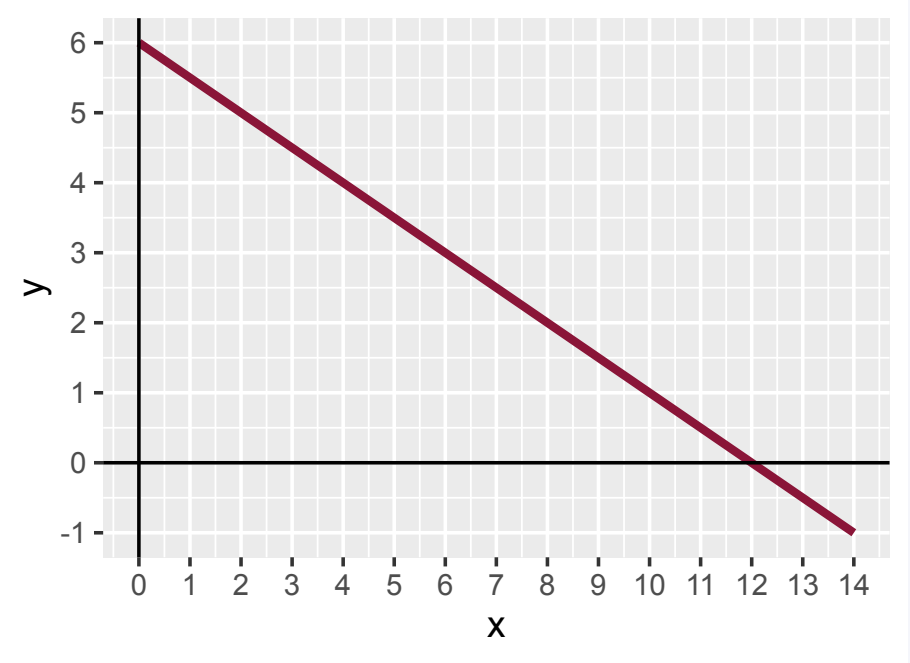
y	A number
x	A number
m	Slope ($\frac{\text{rise}}{\text{run}}$)
b	y-intercept

Slopes and intercepts

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Draw some lines!

$y = 2x - 7$

$y = -0.5x + 1$

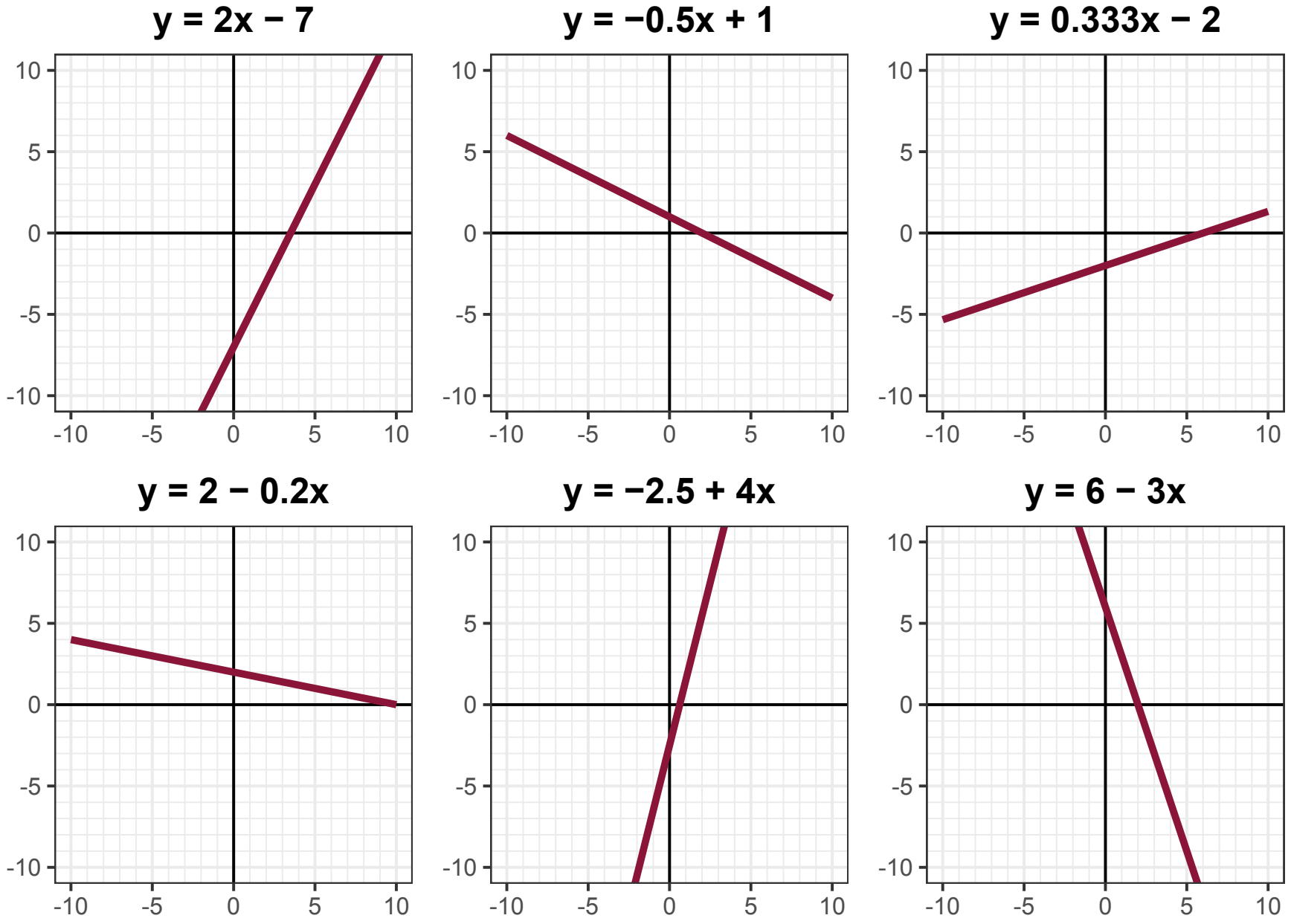
$y = 0.333x - 2$

$y = 2 - 0.2x$

$y = -2.5 + 4x$

$y = 6 - 3x$

05:00


y = 2x − 7
y = −0.5x + 1
y = 0.333x − 2
y = 2 − 0.2x
y = −2.5 + 4x
y = 6 − 3x
10
5
0
-5
-10
-10
-5
0
5
10

Drawing lines with stats

$y = mx + b$		$\hat{y} = b_0 + b_1 x$
y	Outcome variable (**Y**)	$\hat{y}$
x	Explanatory variable (**X**)	x
m	Slope	b_1
b	y-intercept	b_0

The ^ hat on $\hat{y}$ is called a "hat"; read it like "y hat"

$\hat{y}$ means it's a *fitted value*, or the value on the line

Building models in R

```
name_of_model <- lm(<Y> ~ <X>, data = <DATA>)
```

- `lm()` = **l**inear **m**odel
- `<Y> ~ <X>` is a **model formula** (same as in `get_correlation()`!)

```
library(moderndive)

# Make the model
name_of_model <- lm(..., data = BLAH)

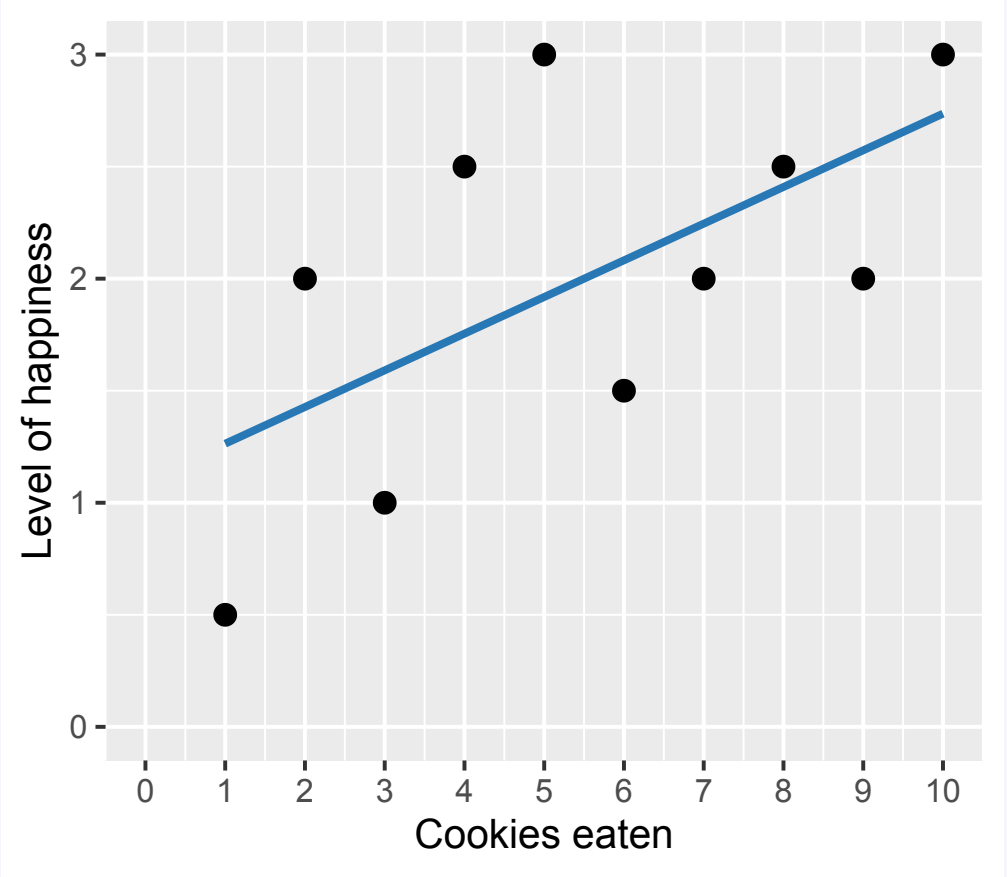
# Coefficients in a nice table
get_regression_table(name_of_model)

# Fitted values and residuals for
# every observation
get_regression_points(name_of_model)
```

Cookies and happiness

$$\widehat{\text{happiness}} = b_0 + b_1 \times \text{cookies}$$

```
cookies_model <- lm(
  happiness ~ cookies,
  data = cookies
)
```



Cookies and happiness

```
get_regression_table(cookies_model)
```

```
# A tibble: 2 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    1.1       0.47       2.34   0.047    0.016    2.18 
2 cookies      0.164     0.076      2.16   0.063   -0.011    0.338
```

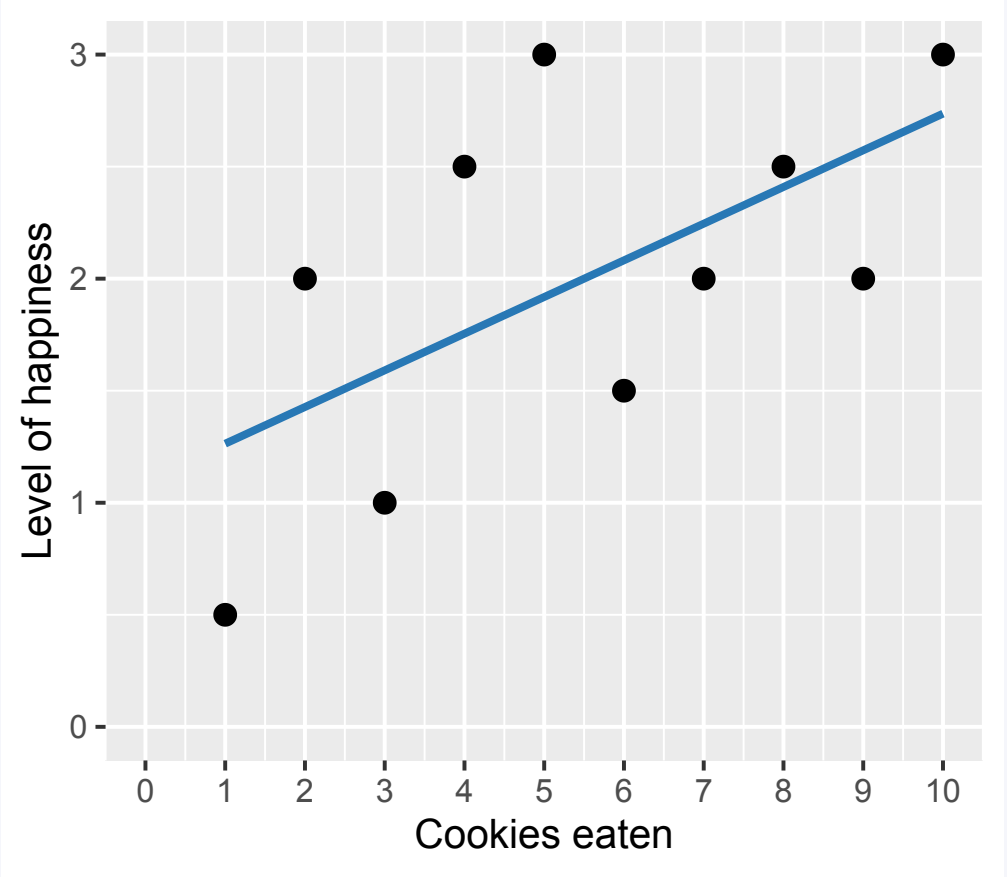
(We'll learn what all those other columns mean later in the semester!)

Translating results to math

term	estimate
intercept	1.100
cookies	0.164

$$\widehat{\text{happiness}} = b_0 + b_1 \times \text{cookies}$$

$$\widehat{\text{happiness}} = 1.1 + 0.164 \times \text{cookies}$$



Template for the slope

A one unit increase in X is *associated* with a b_1 increase (or decrease) in Y, on average

$$\widehat{\text{happiness}} = 1.1 + 0.164 \times \text{cookies}$$

On average, eating one more cookie is associated with a 0.164 increase in happiness

Template for the intercept

b_0 is the average value of Y when X is 0

$$\widehat{\text{happiness}} = 1.1 + 0.164 \times \text{cookies}$$

Average happiness is 1.1 for someone who eats 0 cookies

Sliders

term	estimate
intercept	1.100
cookies	0.164

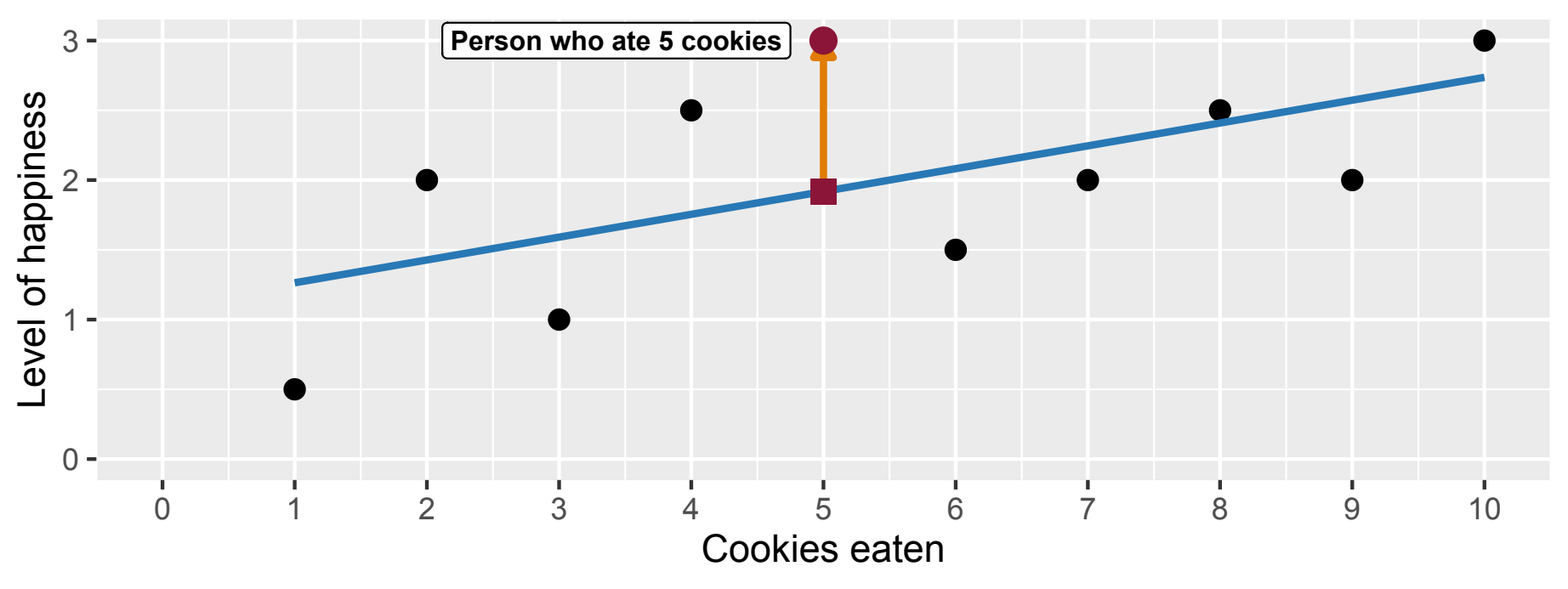
$$\widehat{\text{happiness}} = b_0 + b_1 \times \text{cookies}$$

$$\widehat{\text{happiness}} = 1.1 + 0.164 \times \text{cookies}$$



Continuous variable

Observed vs. fitted values



- **Observed value** (y): Actual happiness = 3
- **Fitted value** ($\hat{y}$): $1.1 + 0.164 \times 5 = 1.918$
- **Residual** ($y - \hat{y}$): $3 - 1.918 = 1.082$

Residuals for every person

```
get_regression_points(cookies_model)
```

ID	happiness	cookies	happiness_hat	residual
1	0.5	1	1.264	-0.764
2	2.0	2	1.427	0.573
3	1.0	3	1.591	-0.591
4	2.5	4	1.755	0.745
5	3.0	5	1.918	1.082
6	1.5	6	2.082	-0.582
7	2.0	7	2.245	-0.245
8	2.5	8	2.409	0.091
9	2.0	9	2.573	-0.573
10	3.0	10	2.736	0.264

Fertility and life expectancy

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

$$\widehat{\text{fert_rate}} = b_0 + b_1 \times \text{life_exp}$$



Fertility and life expectancy

$$\widehat{\text{fert_rate}} = b_0 + b_1 \times \text{life_exp}$$

```
demographics_model <- lm(
  fert_rate ~ life_exp,
  data = UN_data_ch5
)
```



Fertility and life expectancy

```
get_regression_table(demographics_model)
```

```
# A tibble: 2 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept   12.6       0.545      23.1       0   11.5      13.7  
2 life_exp    -0.137     0.007     -18.6       0   -0.152    -0.123
```

Translating results to math

term	estimate
intercept	12.599
life_exp	-0.137

$$\widehat{\text{fert_rate}} = b_0 + b_1 \times \text{life_exp}$$

$$\widehat{\text{fert_rate}} = 12.6 + (-0.137) \times \text{life_exp}$$



Template for the slope

A one unit increase in X is *associated* with a b_1 increase (or decrease) in Y, on average

$$\widehat{\text{fert_rate}} = 12.6 + (-0.137) \times \text{life_exp}$$

On average, a one-year increase in life expectancy is associated with a 0.137 decrease in the fertility rate

Use the right language!

“Associated”

Correlation isn’t causation! Adding a year of life to everyone in a country won’t *make* people have fewer kids

“On average”

Two countries with life expectancies one year apart won’t have fertility rates exactly 0.137 apart. Some are above the line, some below

Correlation vs. slope

Correlation

r = -0.812

Strength and direction of the relationship

Always between –1 and 1; no units

Slope

b_1 **= -0.137**

Size of the relationship

Measured in units of **Y** per one unit of **X**

They'll always have the same sign, but usually not the same value!

Template for the intercept

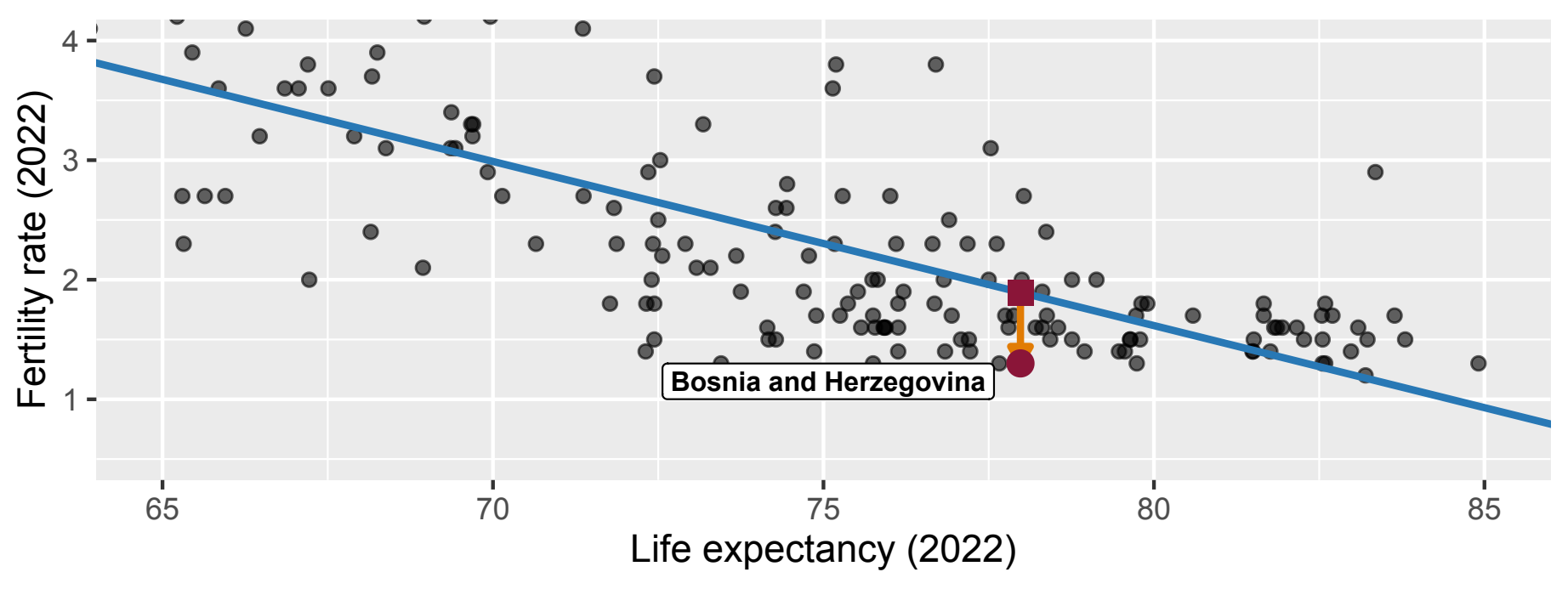
b_0 is the average value of Y when X is 0

$$\widehat{\text{fert_rate}} = 12.6 + (-0.137) \times \text{life_exp}$$

In a country where life expectancy is 0 years, the average fertility rate would be 12.6

A life expectancy of 0 is actually impossible, so this intercept has no practical meaning—it's just where the line crosses the y-axis!

Observed vs. fitted values



- **Observed value** (y): Bosnia's actual fertility rate = 1.3
- **Fitted value** ($\hat{y}$): $12.599 + (-0.137) \times 77.98 = 1.894$
- **Residual** ($y - \hat{y}$): $1.3 - 1.894 = -0.594$

Residuals for every country

```
get_regression_points(demographics_model)
```

iso	fert_rate	life_exp	fert_rate_hat	residual
AFG	4.3	53.65	5.234	-0.934
ALB	1.4	79.47	1.689	-0.289
DZA	2.7	78.03	1.887	0.813
AGO	5.0	62.11	4.072	0.928
ATG	1.6	77.80	1.918	-0.318
ARG	1.9	78.31	1.848	0.052
ARM	1.6	76.13	2.148	-0.548
AUS	1.6	83.09	1.192	0.408

Your turn #2

Does a country's life expectancy predict its happiness?

Use `simple-regression.qmd` in the "Simple regression" project

1. Make a scatterplot of `happiness` and `life_expectancy`
2. Find the correlation between the two with `get_correlation()`
3. Fit a model with `lm()` and look at the coefficients with `get_regression_table()`
4. Interpret the slope and the intercept

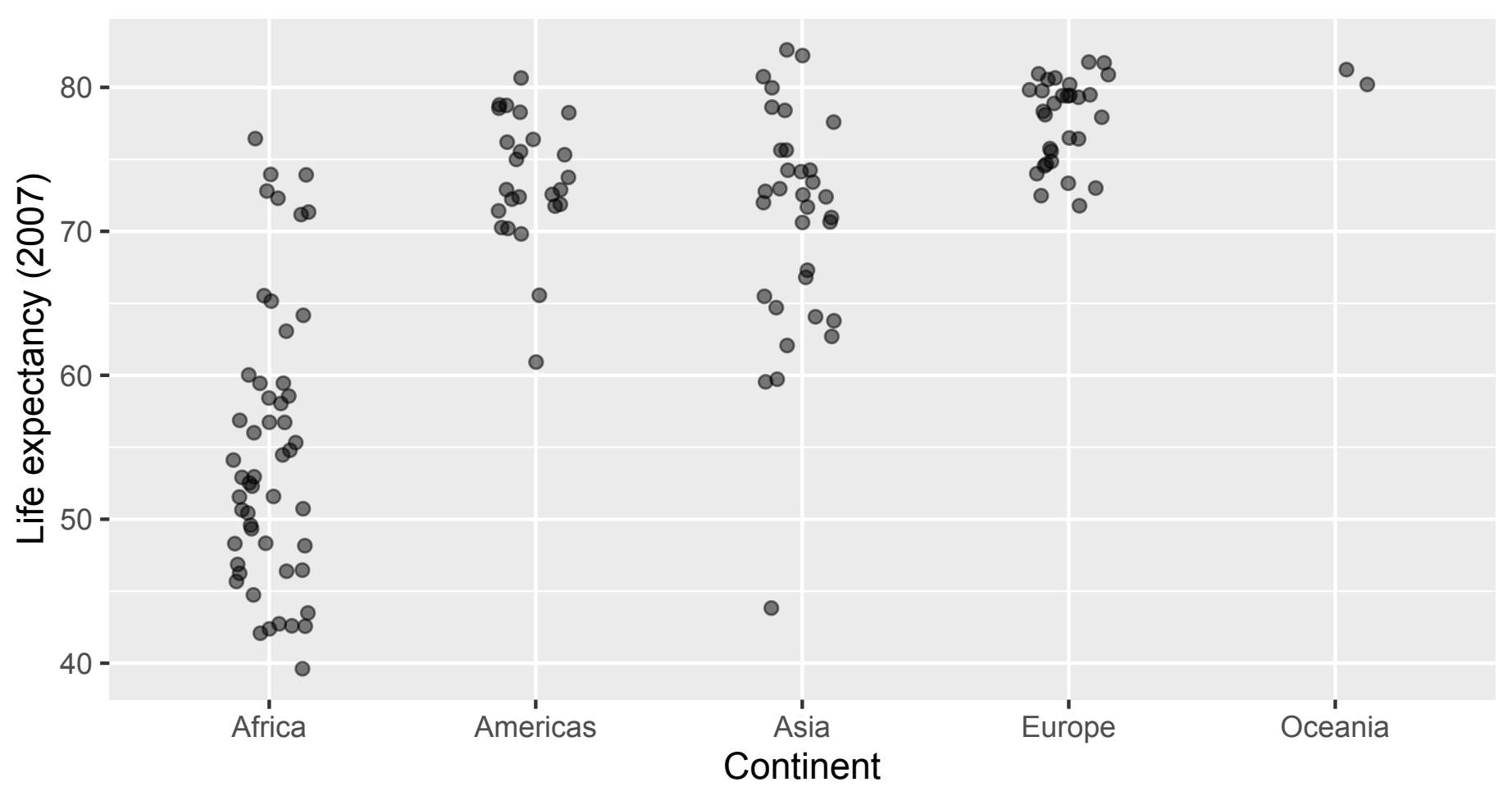
08:00

Categories and regression

Life expectancy & continents (2007)

country	continent	lifeExp	gdpPercap
Afghanistan	Asia	43.828	975
Albania	Europe	76.423	5,937
Algeria	Africa	72.301	6,223
Angola	Africa	42.731	4,797
Argentina	Americas	75.320	12,779
Australia	Oceania	81.235	34,435
Austria	Europe	79.829	36,126
Bahrain	Asia	75.635	29,796
Bangladesh	Asia	64.062	1,391
Belgium	Europe	79.441	33,693

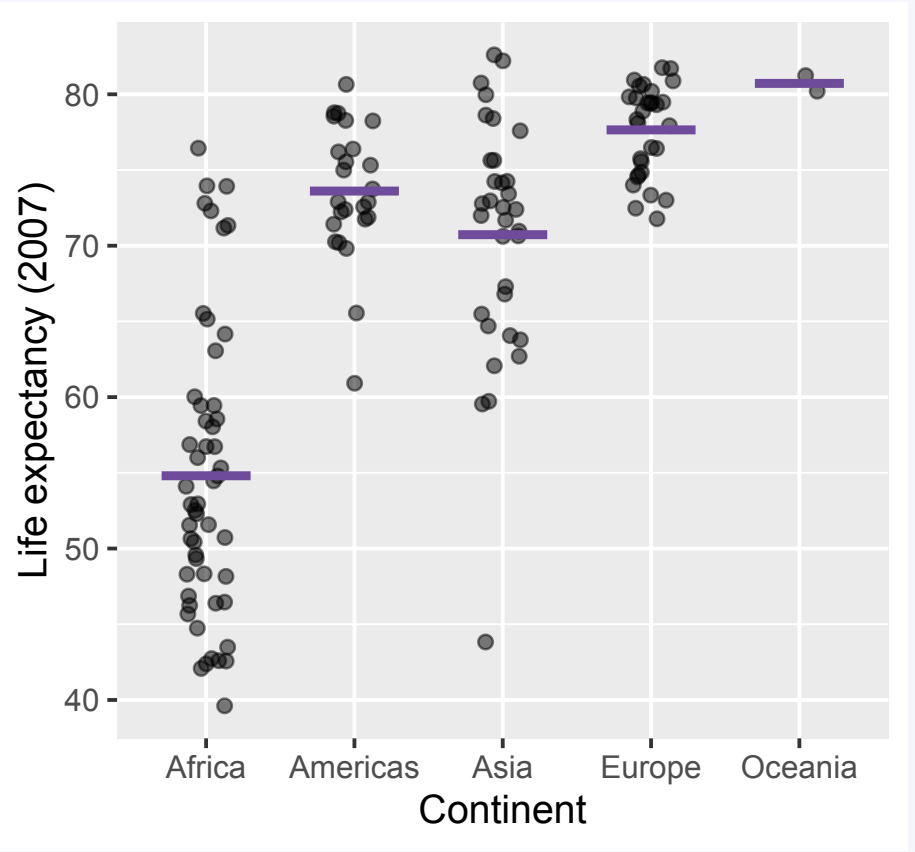
Plot X and Y


80
70
60
50
40
Life expectancy (2007)
Africa
Americas
Asia
Europe
Oceania
Continent

Averages for each continent

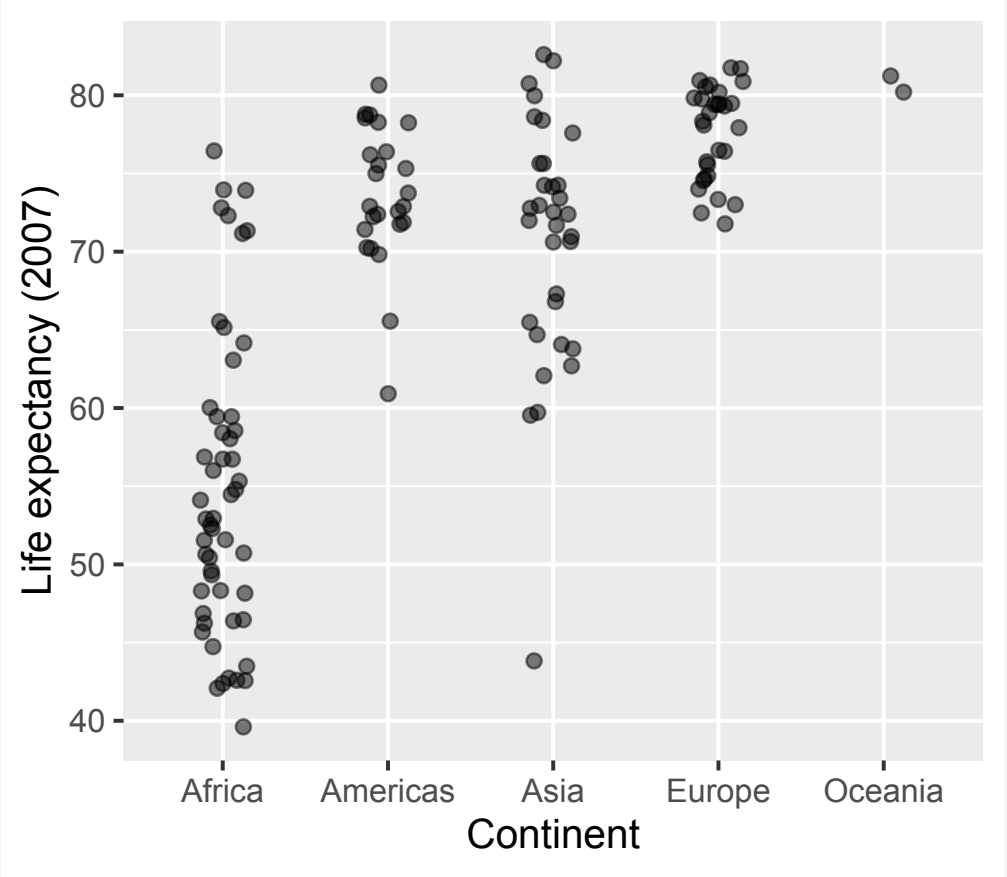
```
gapminder_2007 |>
  group_by(continent) |>
  summarize(
    n = n(),
    mean = mean(lifeExp),
    median = median(lifeExp)
  )
```

continent	n	mean	median
Africa	52	54.81	52.93
Americas	25	73.61	72.9
Asia	33	70.73	72.4
Europe	30	77.65	78.61
Oceania	2	80.72	80.72



Can we draw a line through this?

- **X** isn't a number!
- There's no "one unit increase" in continent
- ...but `lm()` works anyway!



Building the model

```
lifeexp_model <- lm(lifeExp ~ continent, data = gapminder_2007)
get_regression_table(lifeexp_model)
```

```
# A tibble: 5 × 7
  term                estimate std_error statistic p_value lower_ci upper_ci
  <chr>                  <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept               54.8      1.02     53.4         0     52.8     56.8
2 continent-Americas      18.8      1.8      10.4         0     15.2     22.4
3 continent-Asia          15.9      1.65      9.68        0     12.7     19.2
4 continent-Europe        22.8      1.70     13.5         0     19.5     26.2
5 continent-Oceania       25.9      5.33      4.86        0     15.4     36.4
```

Where'd Africa go?!

The base case

One category gets left out and becomes the base case

Also called the baseline, reference category, or omitted category

R uses the first category (alphabetically, by default) as the base case, so here it's Africa

Every other coefficient is measured relative to the base case

Translating results to math

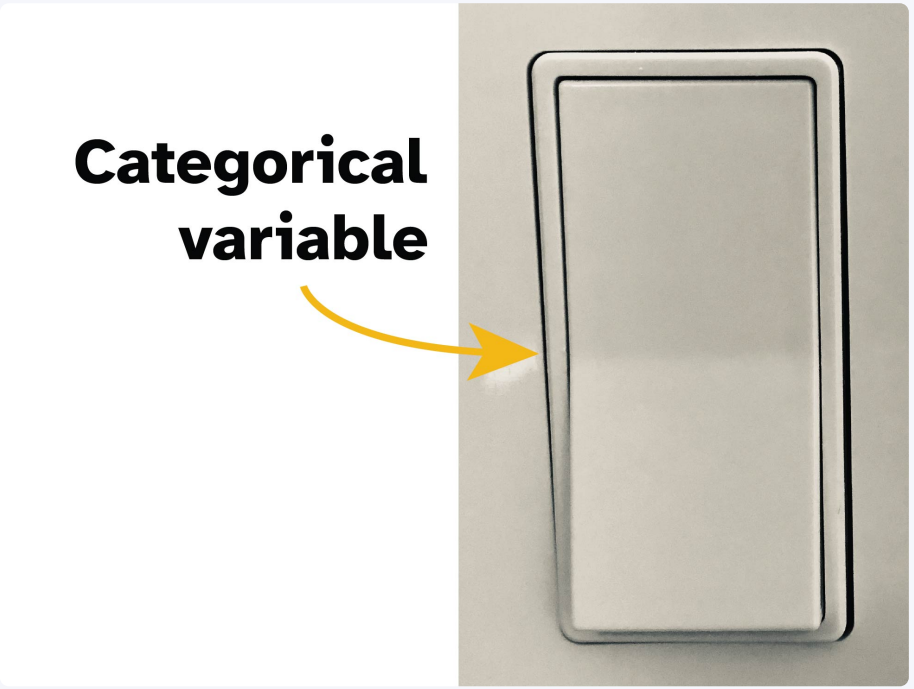
term	estimate
intercept	54.806
continent-Americas	18.802
continent-Asia	15.922
continent-Europe	22.843
continent-Oceania	25.913

$$\widehat{\text{lifeExp}} = b_0 + b_1 \times \text{Americas} + b_2 \times \text{Asia} + b_3 \times \text{Europe} + b_4 \times \text{Oceania}$$

$$\widehat{\text{lifeExp}} = 54.81 + 18.8 \times \text{Americas} + 15.92 \times \text{Asia} + 22.84 \times \text{Europe} + 25.91 \times \text{Oceania}$$

Switches

term	estimate
intercept	54.806
continent-Americas	18.802
continent-Asia	15.922
continent-Europe	22.843
continent-Oceania	25.913



Template for the intercept

b_0 is the average value of Y for the base case

$$\widehat{\text{lifeExp}} = 54.81 + 18.8 \times \text{Americas} + \ldots$$

Average life expectancy in Africa is 54.81 years

All switches are off, so this is just the average of the base case

Template for the other coefficients

On average, Y is b higher (or lower) in this category, compared to the base case

$$\widehat{\text{lifeExp}} = 54.81 + 18.8 \times \text{Americas} + \ldots$$

On average, life expectancy in the Americas is 18.8 years higher than in Africa

It's **not** "a one unit increase in continent"!

It's the offset from the base case

Flip the switches!

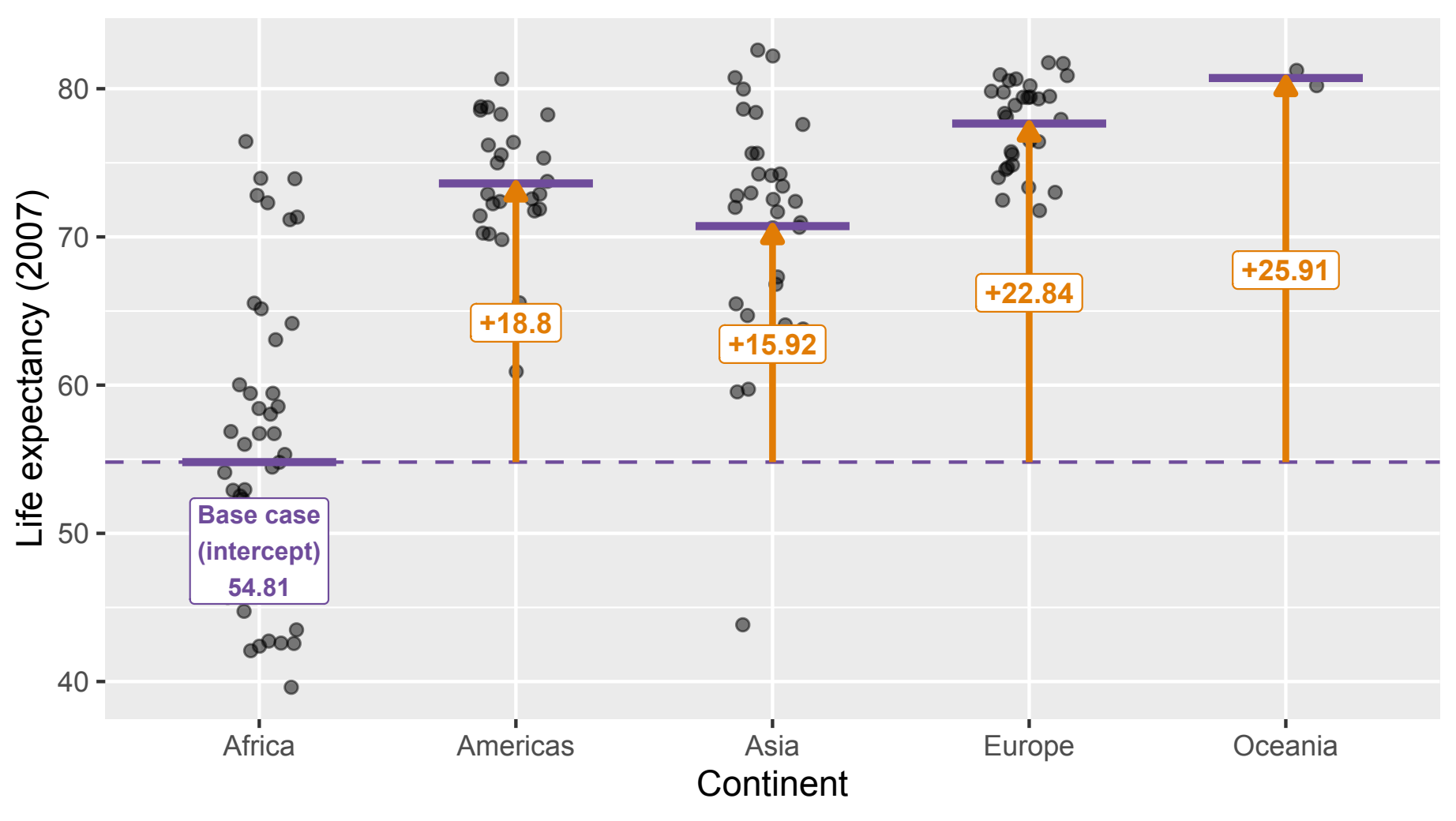
Switch on	$\widehat{\text{lifeExp}}$
None (Africa)	54.81
Americas	$54.81 + 18.8 = 73.61$
Asia	$54.81 + 15.92 = 70.73$
Europe	$54.81 + 22.84 = 77.65$
Oceania	$54.81 + 25.91 = 80.72$

```
gapminder |>
  filter(year == 2007) |>
  group_by(continent) |>
  summarise(avg = mean(lifeExp))
```

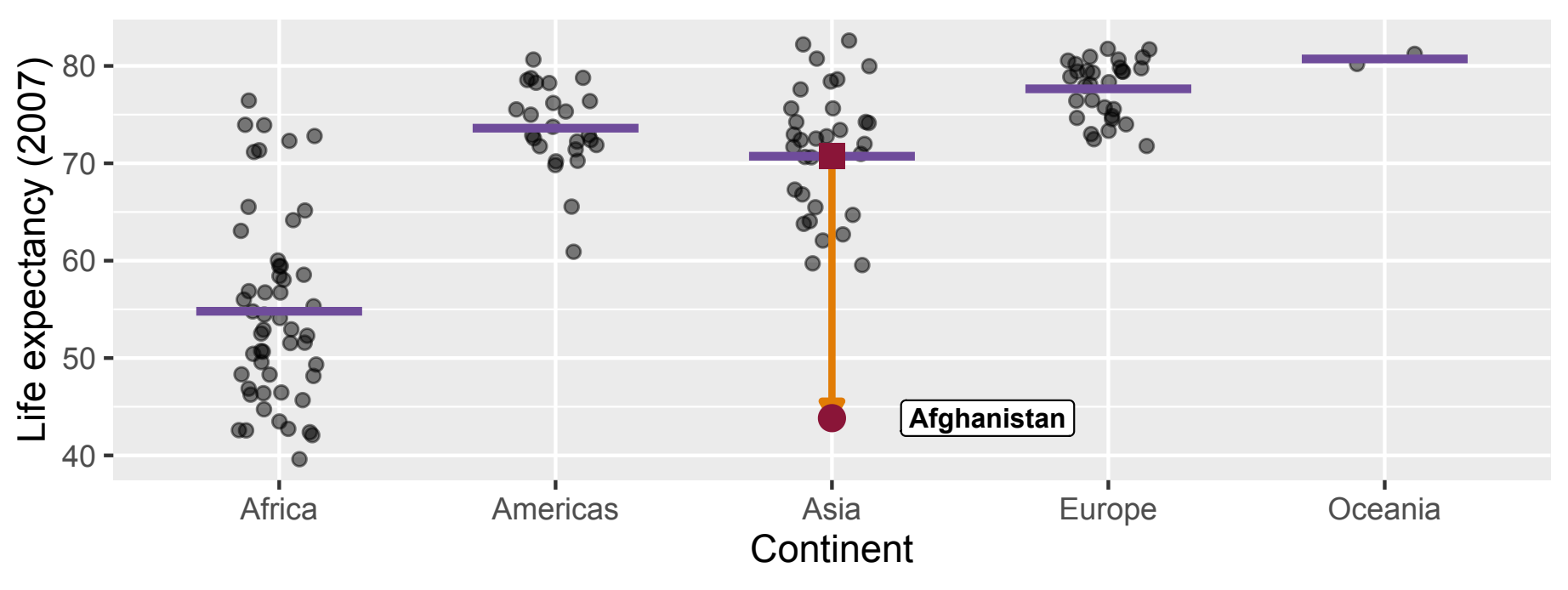
```
# A tibble: 5 × 2
  continent   avg
  <fct>     <dbl>
1 Africa     54.8
2 Americas   73.6
3 Asia       70.7
4 Europe     77.6
5 Oceania    80.7
```

These are just the averages for each continent!

Offsets from the base case



Observed vs. fitted values



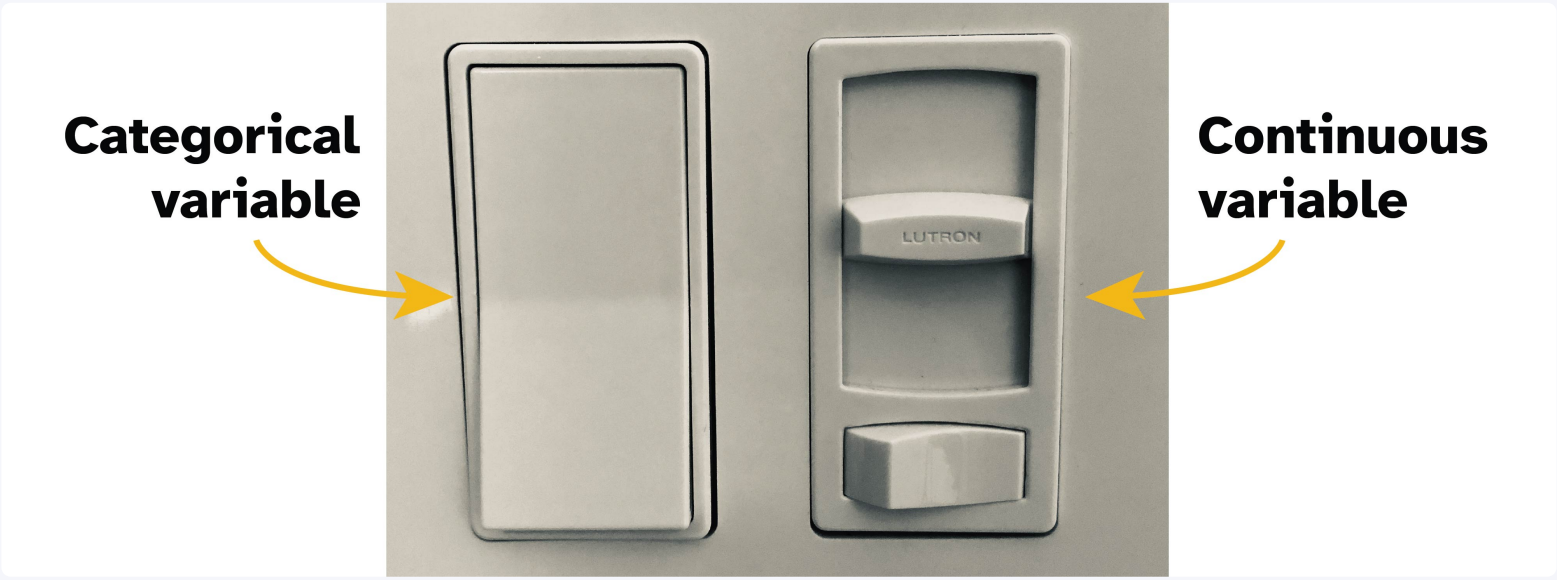
● **Observed value** (y): Afghanistan's actual life expectancy = 43.83 ■ **Fitted value** ($\hat{y}$): $54.81 + 15.92 \times 1 = 70.73$ (the Asia average) ↓ **Residual** ($y - \hat{y}$): $43.83 - 70.73 = -26.9$

Residuals for every country

```
get_regression_points(lifeexp_model, ID = "country")
```

country	lifeExp	continent	lifeExp_hat	residual
Afghanistan	43.828	Asia	70.728	-26.900
Albania	76.423	Europe	77.649	-1.226
Algeria	72.301	Africa	54.806	17.495
Angola	42.731	Africa	54.806	-12.075
Argentina	75.320	Americas	73.608	1.712
Australia	81.235	Oceania	80.720	0.516
Austria	79.829	Europe	77.649	2.180
Bahrain	75.635	Asia	70.728	4.907

Sliders and switches



Categorical X

Y shifts by b_n

compared to the base case

Continuous X

Y changes by b_n

Your turn #3

Is happiness different across continents?

Keep using `simple-regression.qmd`

1. Make a plot of `happiness` across `continent`
2. Find the average `happiness` in each continent with `group_by()` and `summarize()`
3. Fit a model with `lm()` and look at the coefficients with `get_regression_table()`
4. Interpret the intercept and the coefficient for Europe

08:00