REGRESSION AND INFERENCE

PMAP 8521: Program Evaluation for Public Service

September 9, 2019

Fill out your reading report on iCollege!

PLAN FOR TODAY

Revisiting R Markdown

Correlation, regression, and drawing lines

Lines, math, and Greek

Multiple regression

Regression and inference

REVISITING R MARKDOWN

CORRELATION, REGRESSION, & DRAWING LINES

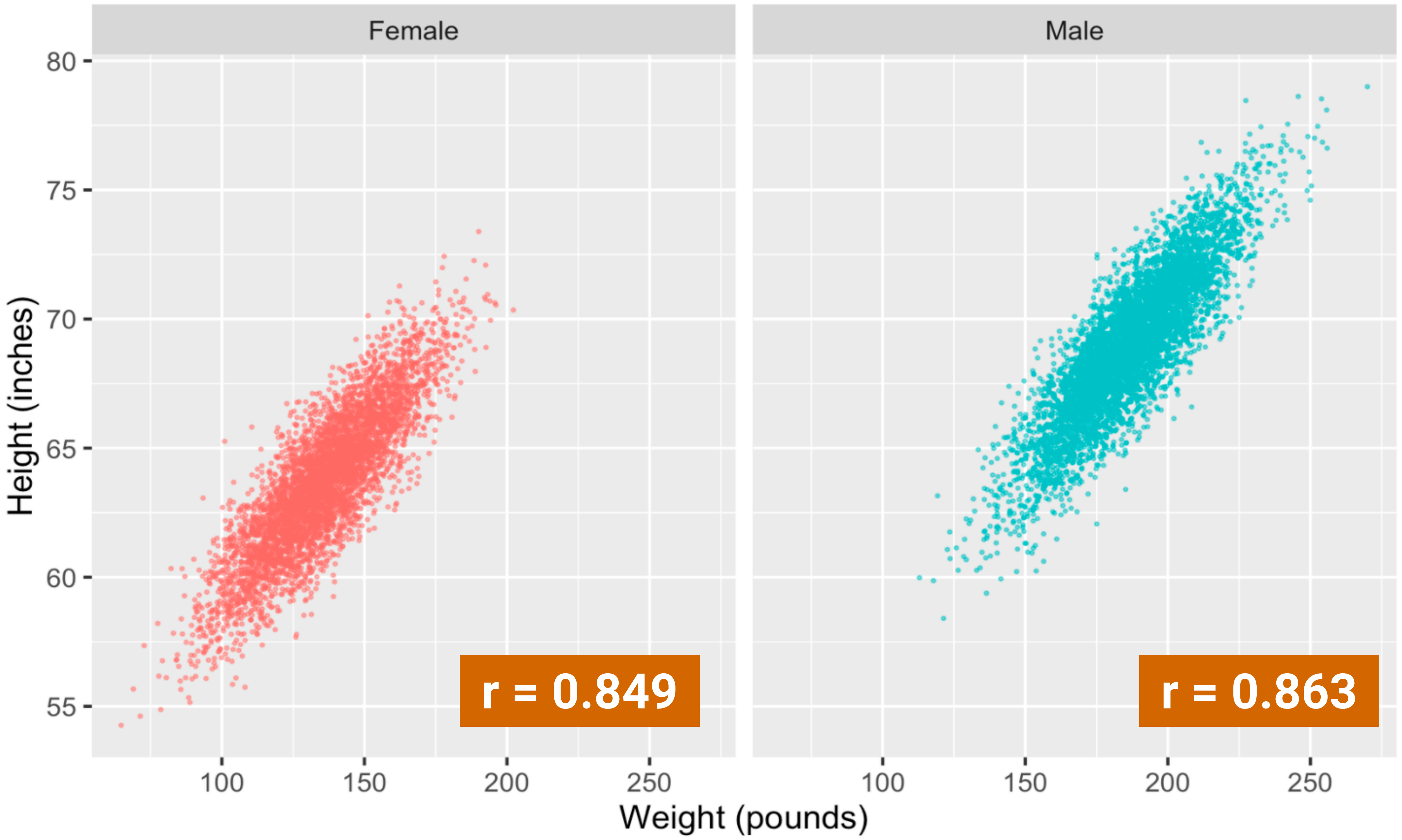
CORRELATION

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

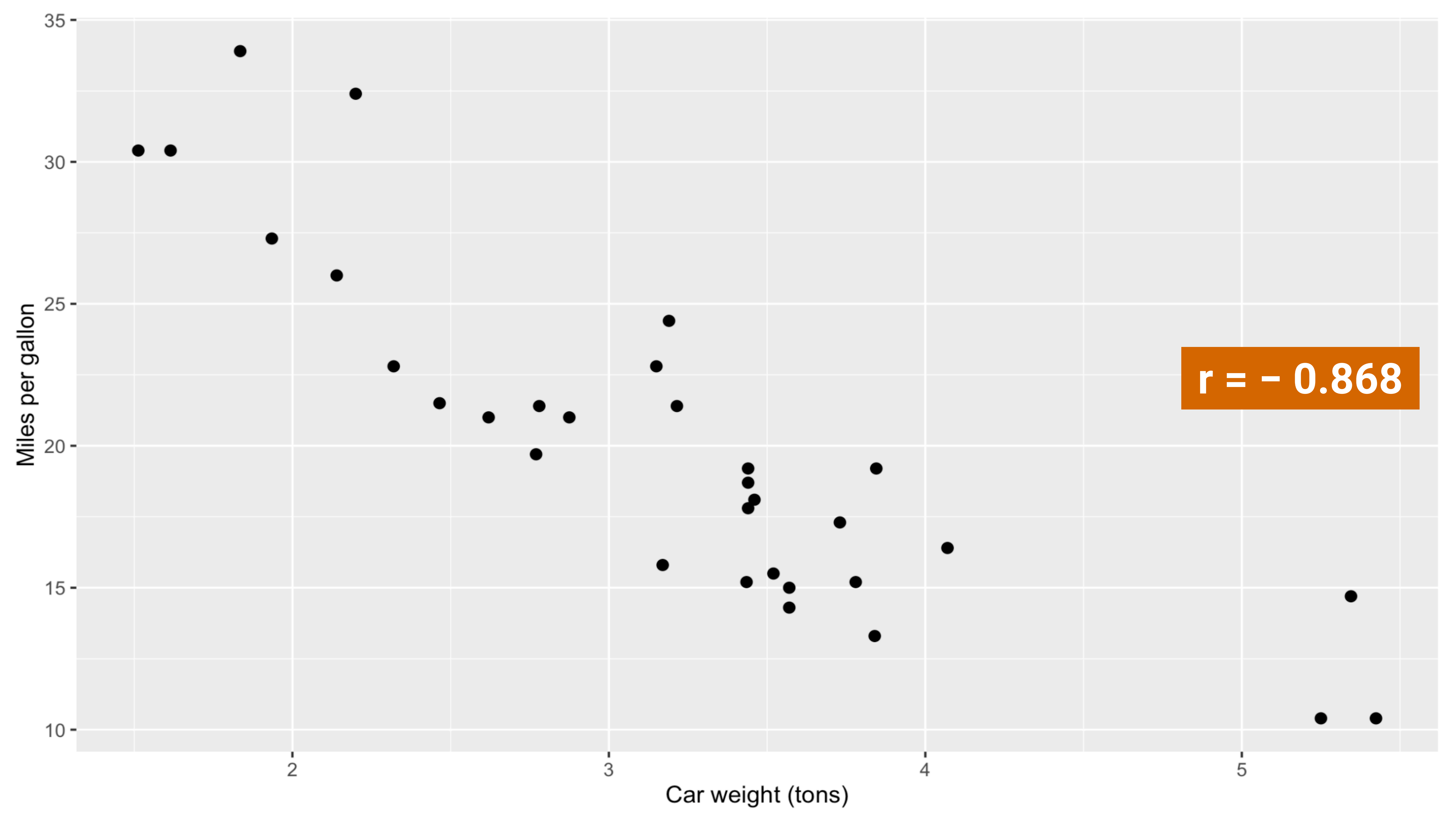
How closely two variables are related + direction of relation

−1 to 1

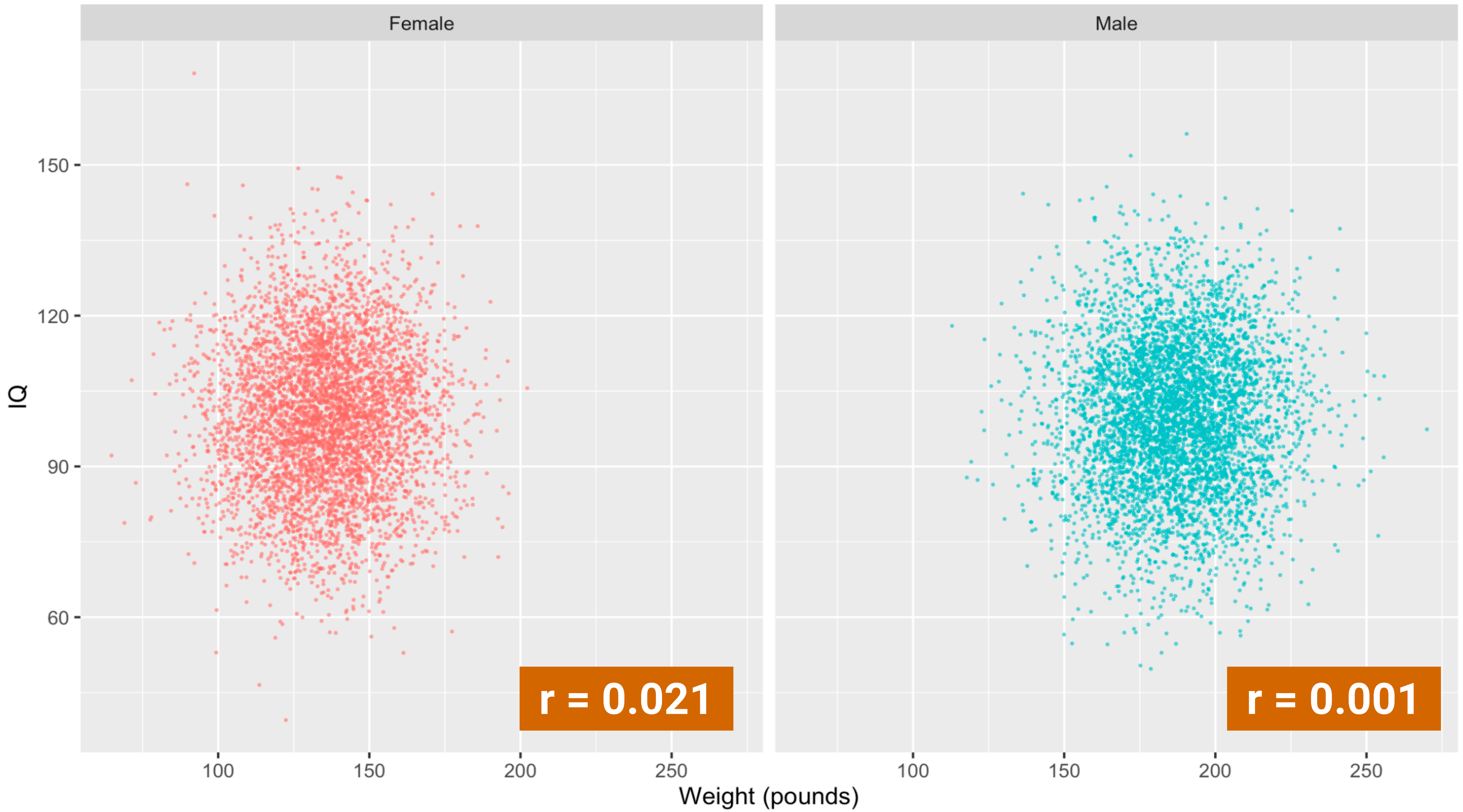
−1 and 1 = perfectly correlated; 0 = perfectly uncorrelated



Female
Male
80
75
70
65
60
55
Height (inches)
100
150
200
250
Weight (pounds)
r = 0.849
r = 0.863



r = – 0.868
Miles per gallon
35
30
25
20
15
10
2
3
4
5
Car weight (tons)



Female
Male
IQ
150
120
90
60
100
150
200
250
Weight (pounds)
r = 0.021
r = 0.001

GENERAL GUIDELINES

0	No relationship
0.01–0.19	Little to no relationship
0.20–0.29	Weak relationship
0.30–0.39	Moderate relationship
0.40–0.69	Strong relationship
0.70–0.99	Very strong relationship
1	Perfect relationship

Can be positive or negative

TEMPLATE

As the value of X goes up, Y tends to go up (or down) a lot/a little/not at all

WHY REGRESSION?

Correlation between car weight and mileage (MPG) is −0.868

If you shave 1 ton off the weight of a car, how much will the car's mileage improve?

Correlation shows direction and magnitude. That's all.

ESSENTIAL PARTS

Y	~	X (or lots of Xs)
Outcome variable		Explanatory variable
Response variable		Predictor variable
Dependent variable		Independent variable
Thing you want to explain or predict		Thing you use to explain changes in Y

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 students taking more AP classes in high school improves their college 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

TWO PURPOSES OF REGRESSION

Prediction	Explanation
Forecast the future	Explain effect of X on Y
Focus is on Y	Focus is on X
Netflix trying to guess your next show	Netflix looking at the effect of time of day on show selection
Predicting who will escape poverty	Looking at the effect of food stamps on poverty reduction

HOW

Plot X and Y

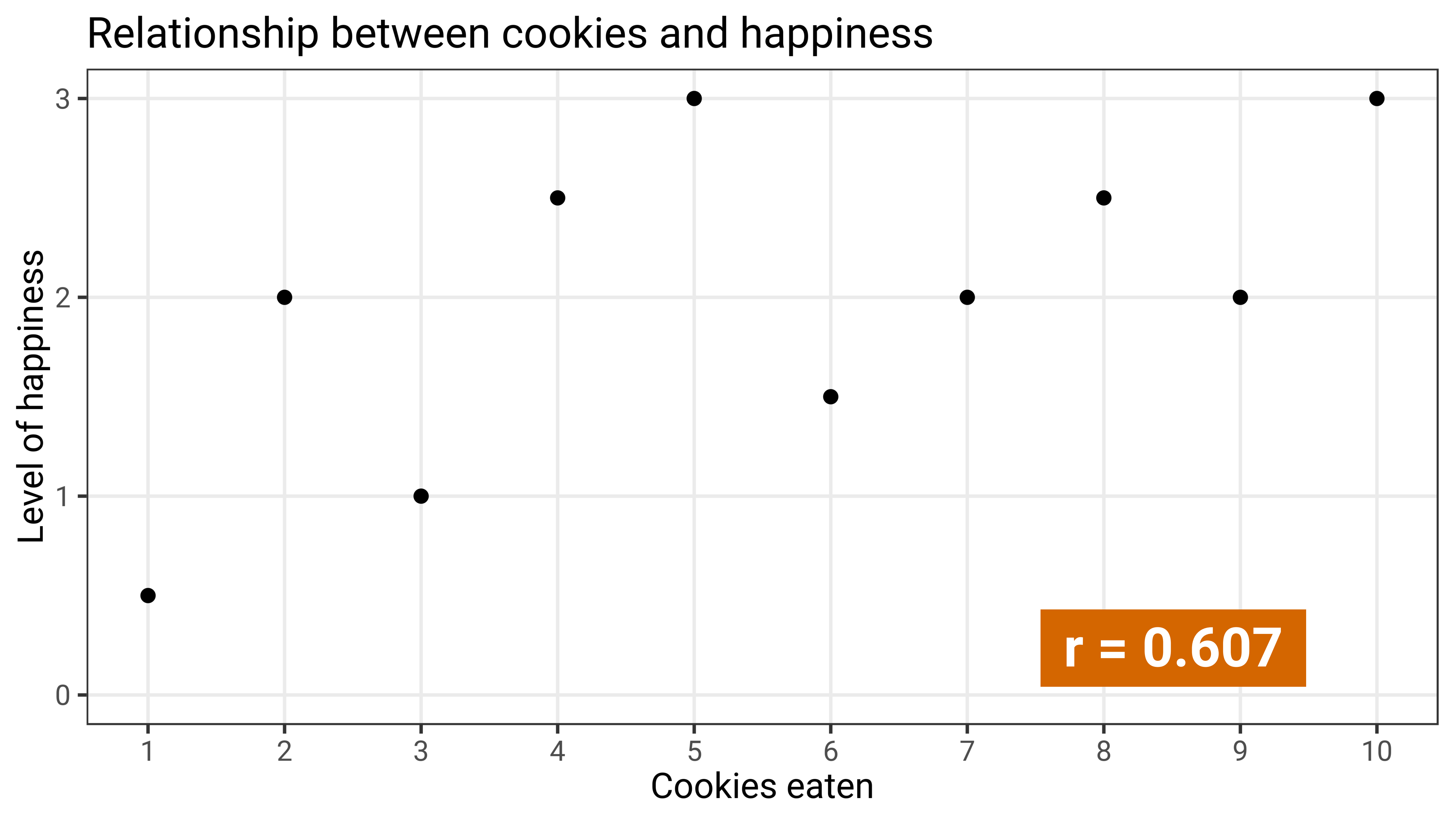
Draw a line that approximates the relationship

Find mathy parts of the line

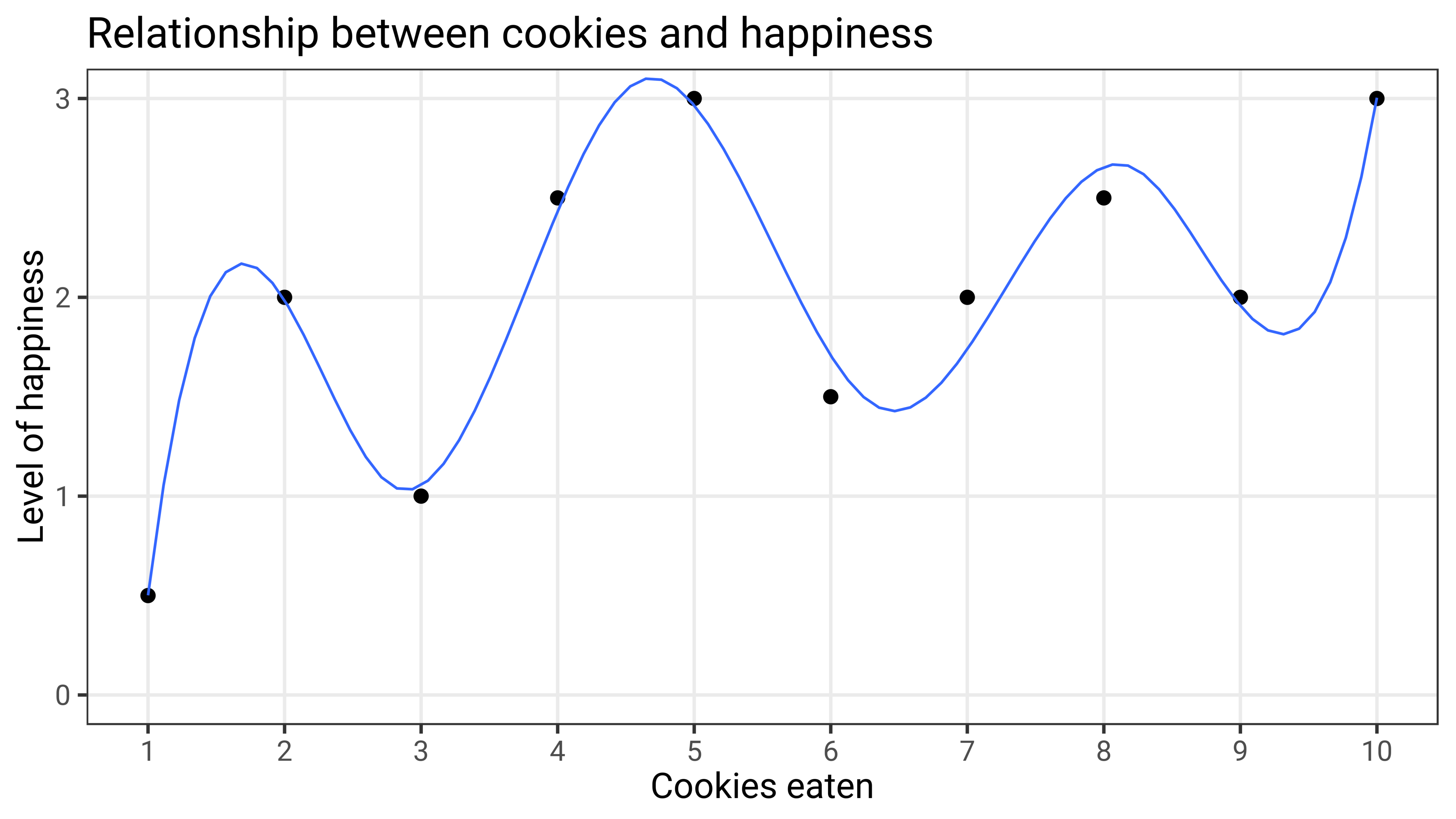
Interpret the math

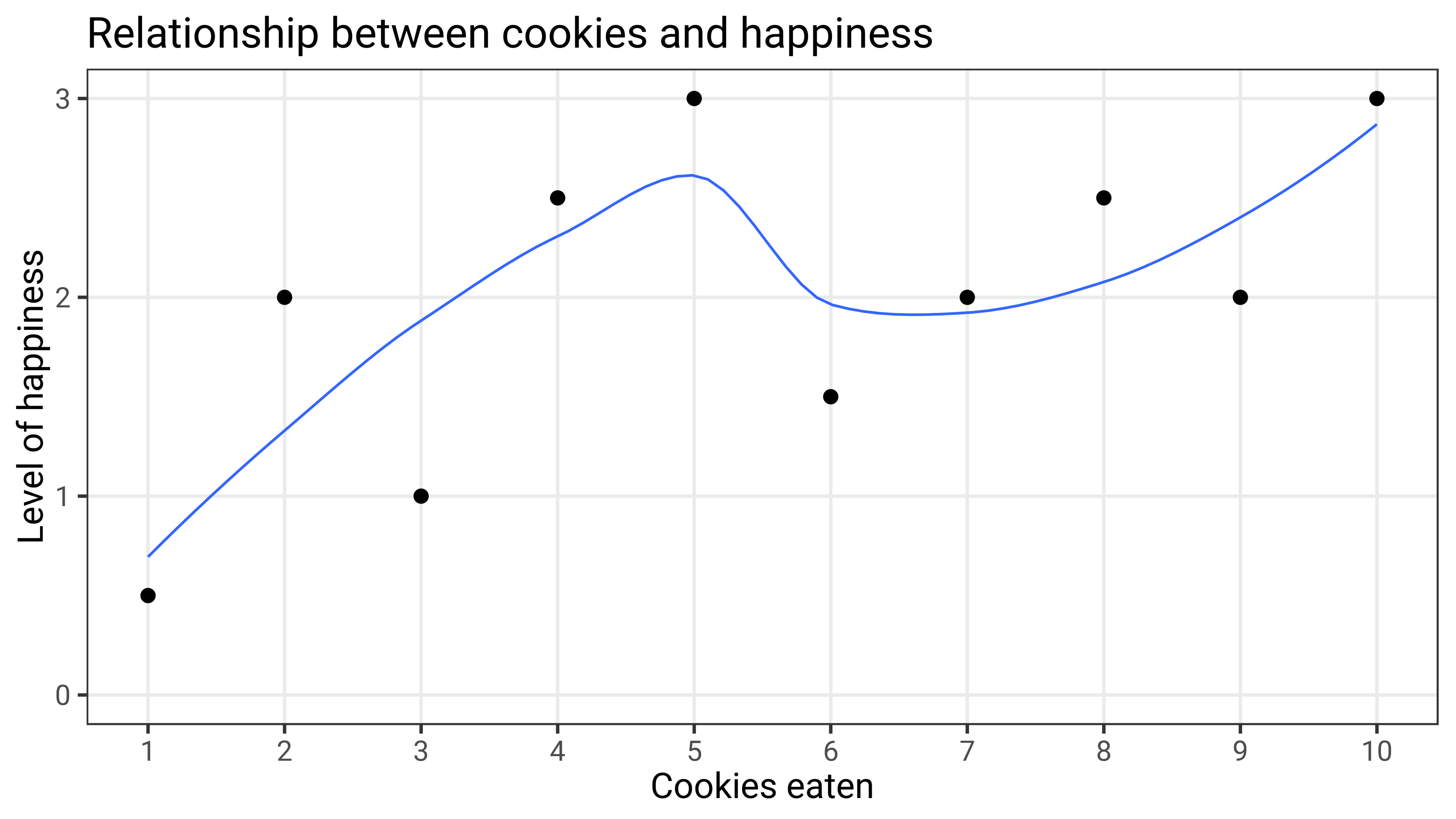
COOKIE CONSUMPTION AND HAPPINESS

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

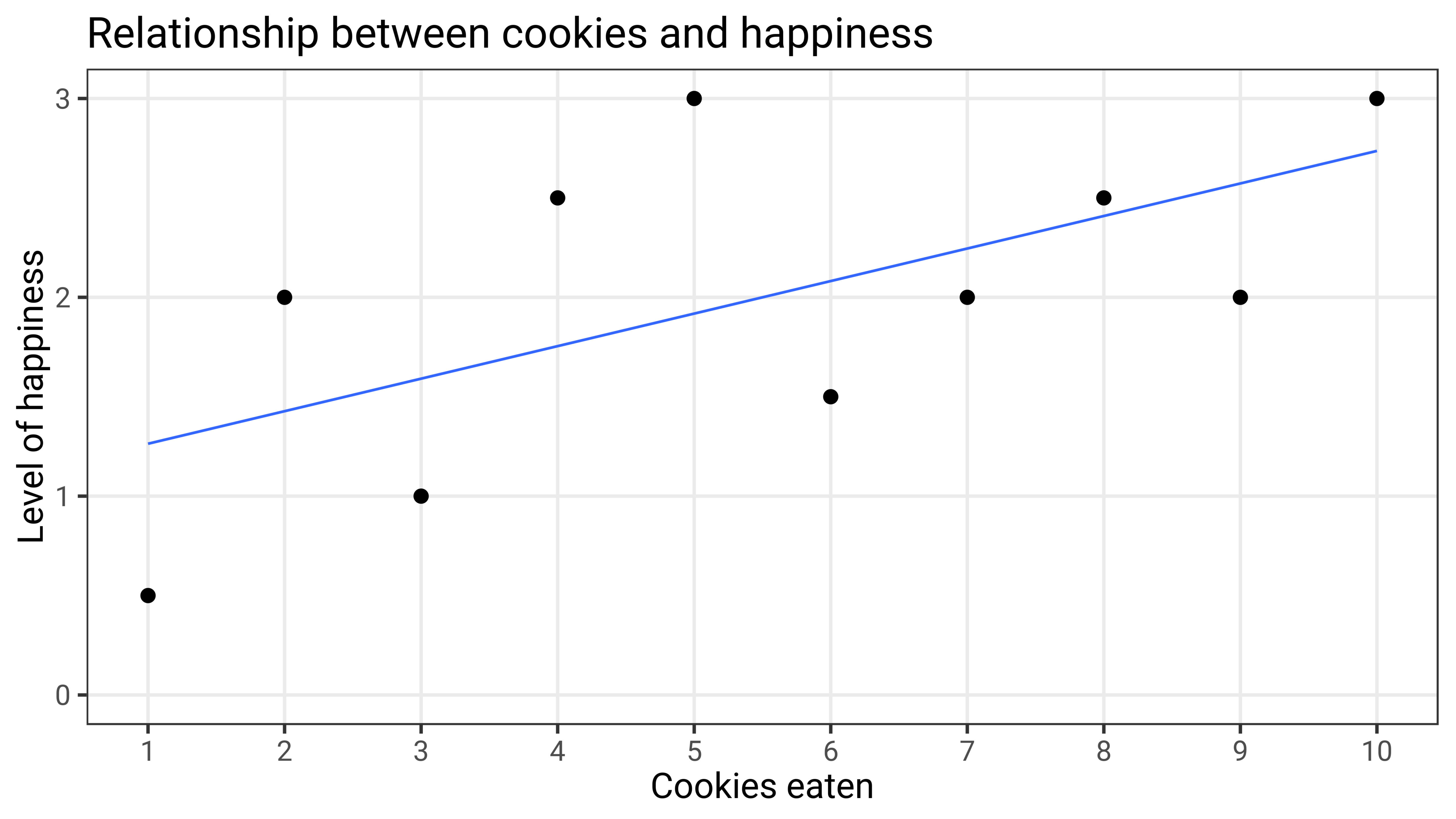


Relationship between cookies and happiness
Level of happiness
0
1
2
3
1
2
3
4
5
6
7
8
9
10
Cookies eaten
r = 0.607

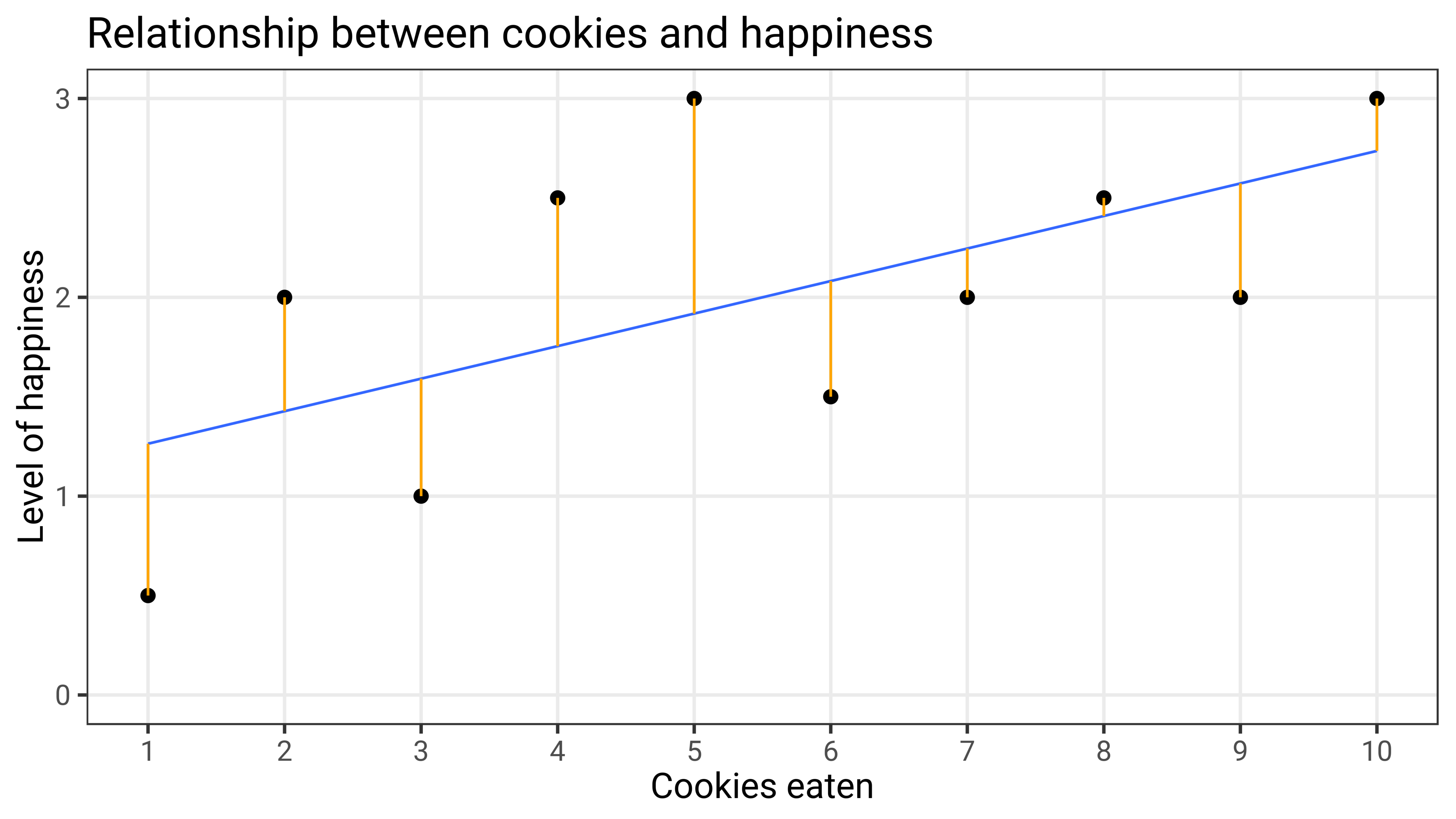

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10

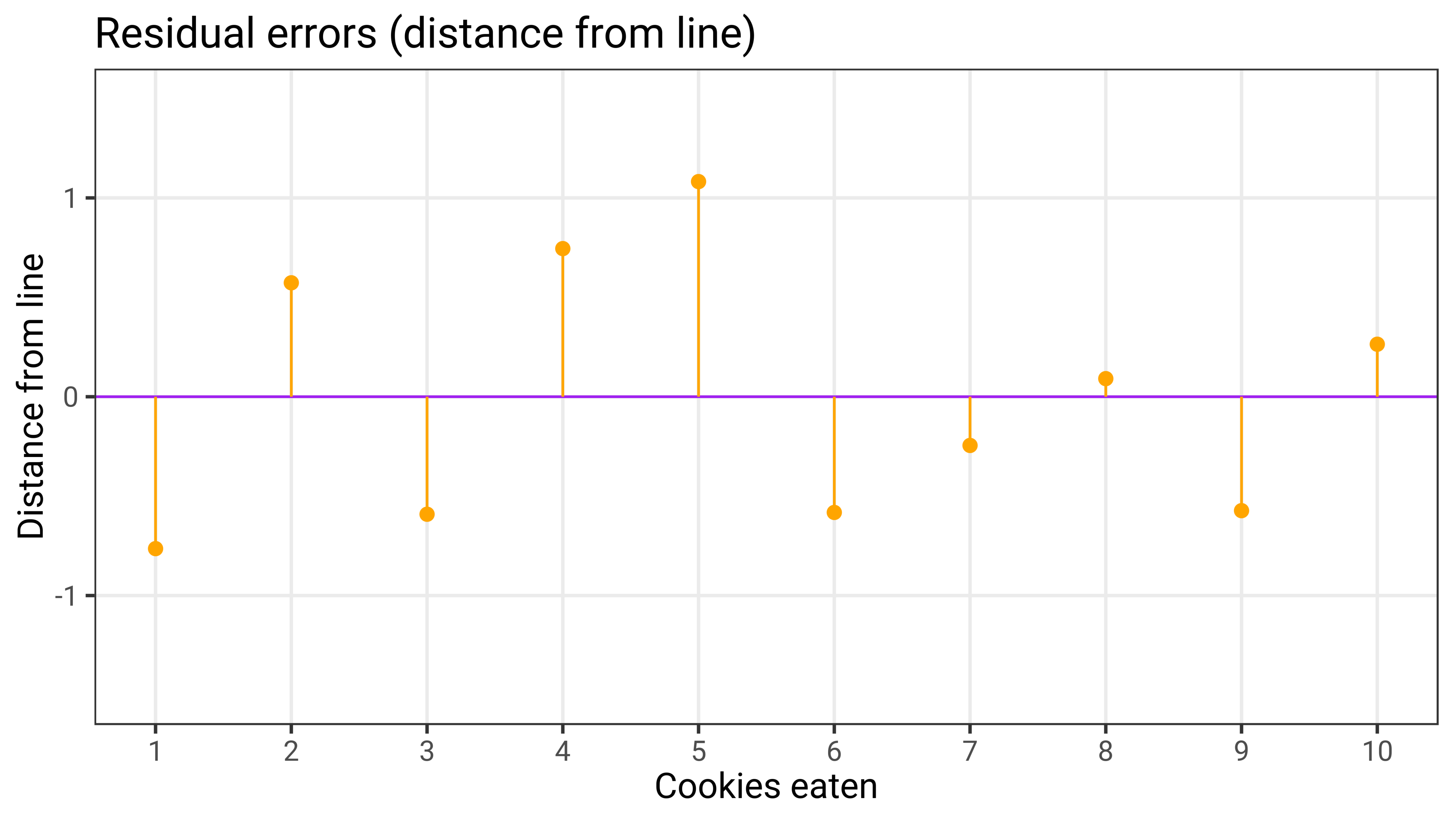


Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10

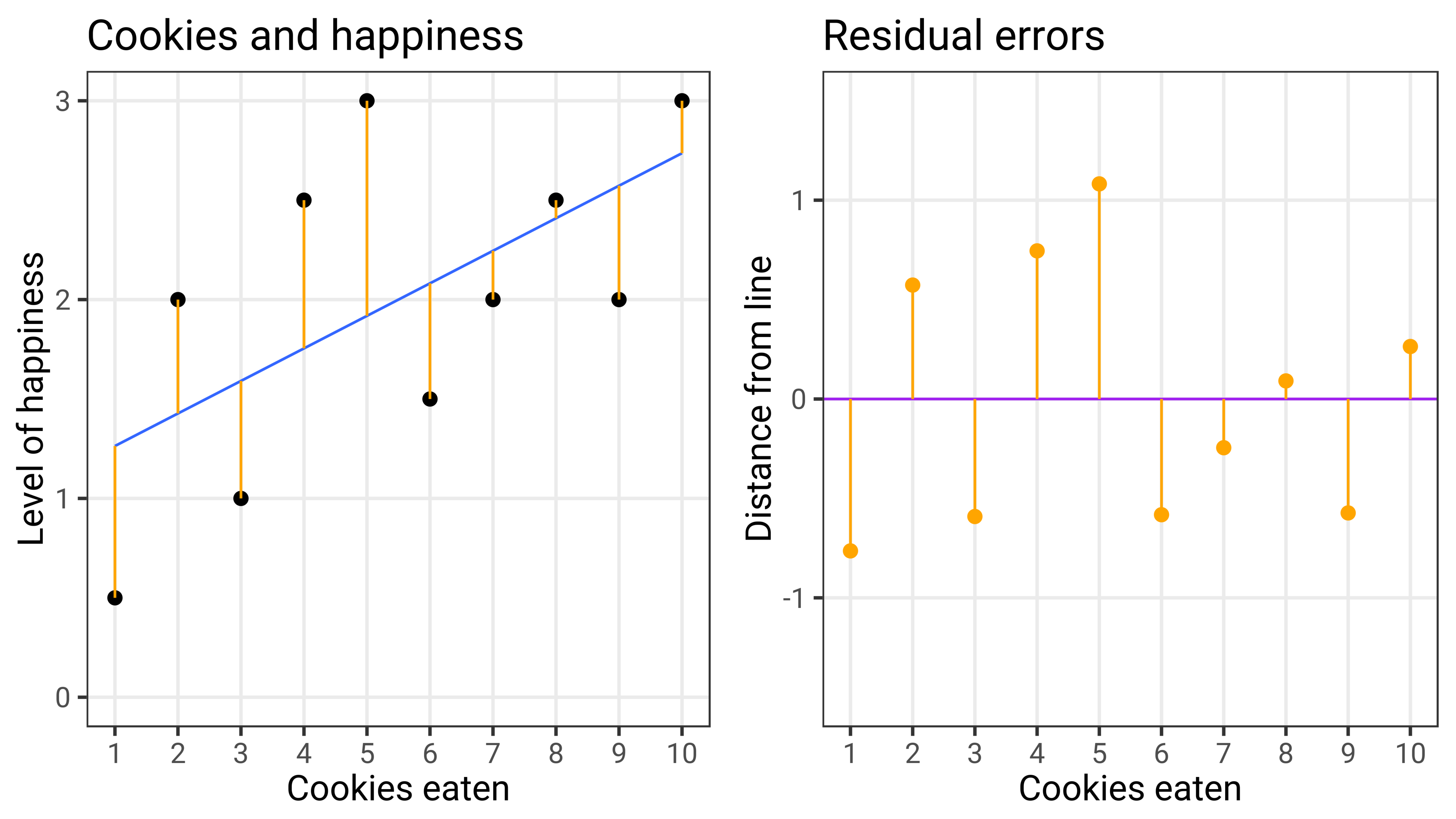


Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10

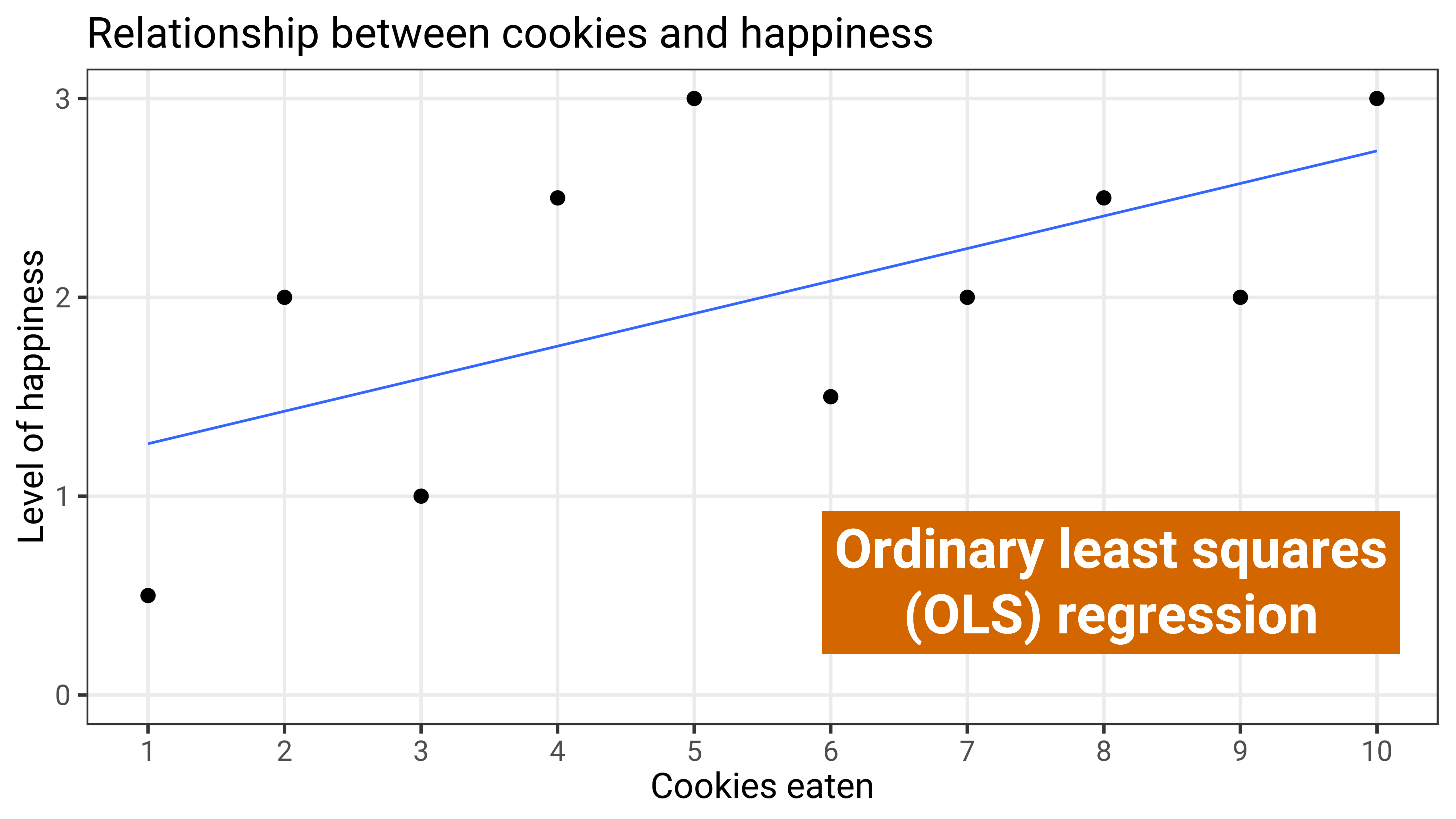

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10



Residual errors (distance from line)
Distance from line
1
0
-1
1
2
3
4
5
6
7
8
9
10
Cookies eaten



Cookies and happiness
Level of happiness
0
1
2
3
1
2
3
4
5
6
7
8
9
10
Cookies eaten
Residual errors
Distance from line
-1
0
1
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10
Ordinary least squares (OLS) regression

LINES, MATH, AND GREEK

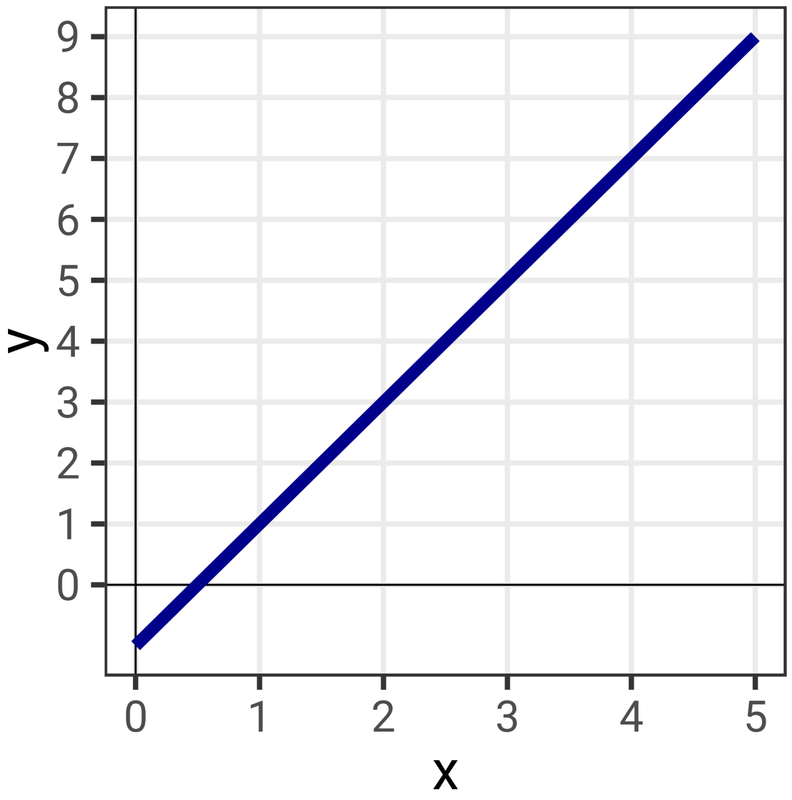
DRAWING LINES WITH MATH

$$y = mx + b$$

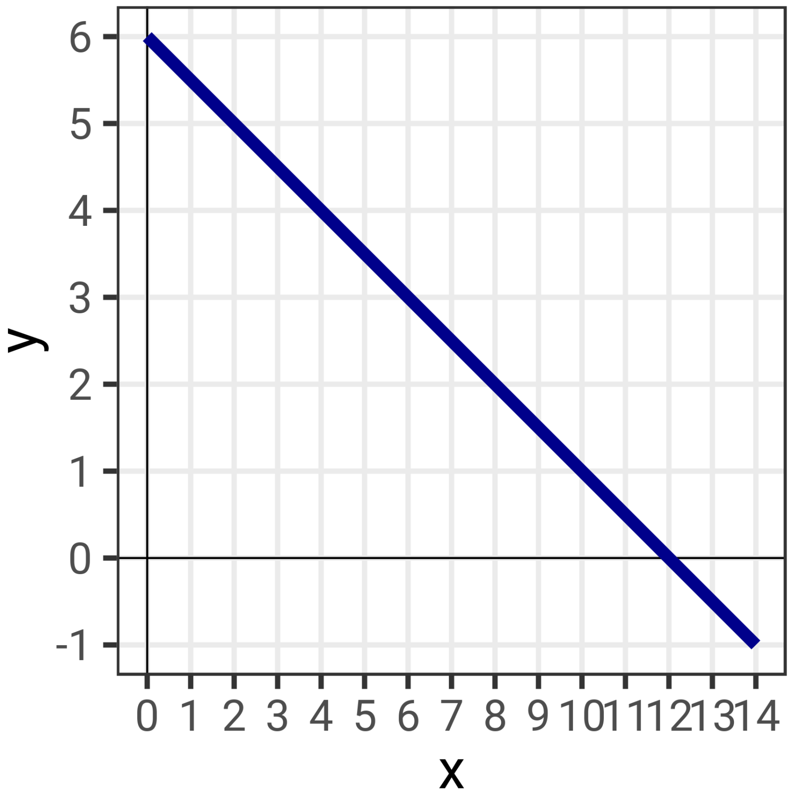
y	A number	
x	A number	
m	Slope	$\frac{rise}{run}$
b	y intercept	

SLOPES AND INTERCEPTS

$$y = 2x - 1$$



$$y = -0.5x + 6$$



GRAPH THESE

$$y = 5x + 2$$

$$y = x - 1$$

$$y = -2x + 11$$

$$y = 6 - 2x$$

$$y = 0.33x - 1$$

$$y = 0.75x - 3$$

DRAWING LINES WITH STATS

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

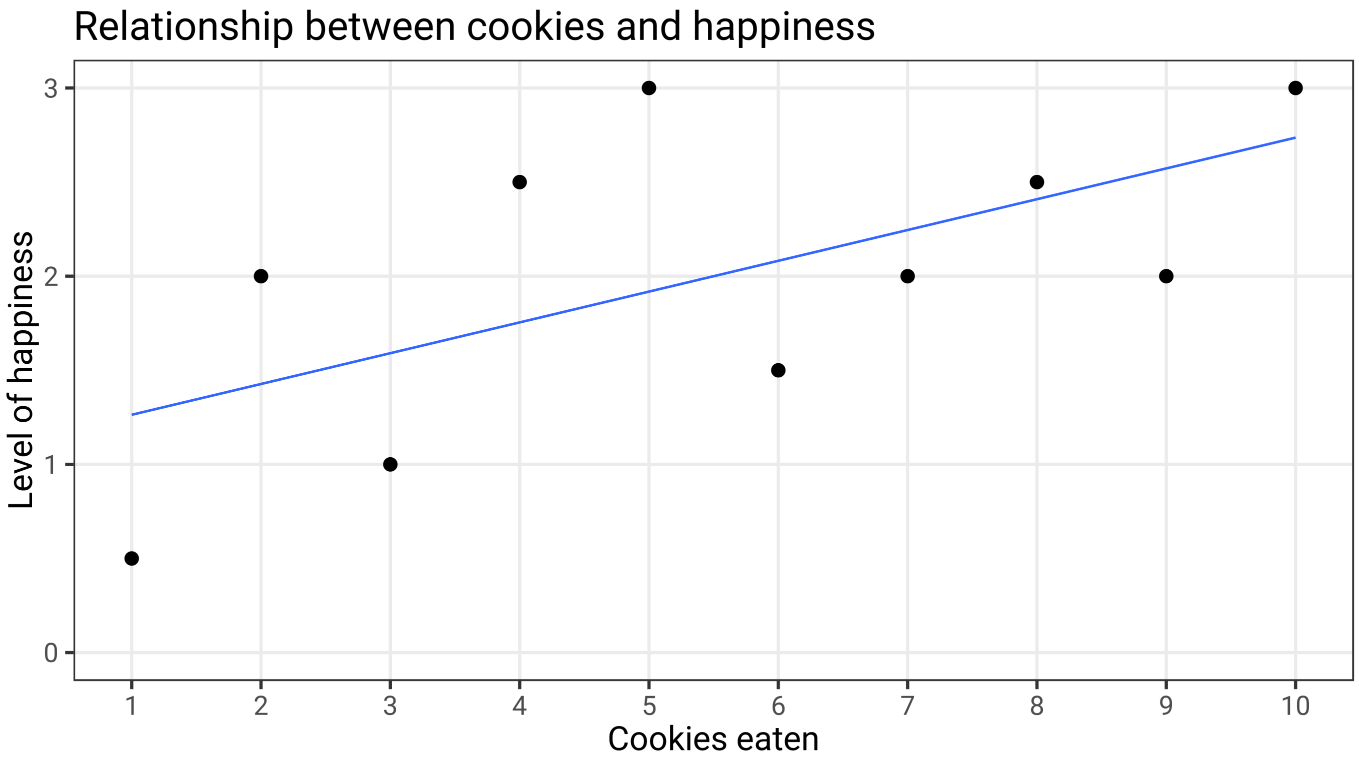
$y = mx + b$

y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
m	β_1	Slope
b	β_0 (α)	y-intercept
	ε	Error (residuals)

MODELING COOKIES AND HAPPINESS

$$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\hat{\text{happiness}} =$$
$$\beta_0 + \beta_1 \text{cookies} + \epsilon$$



MODELING COOKIES AND HAPPINESS

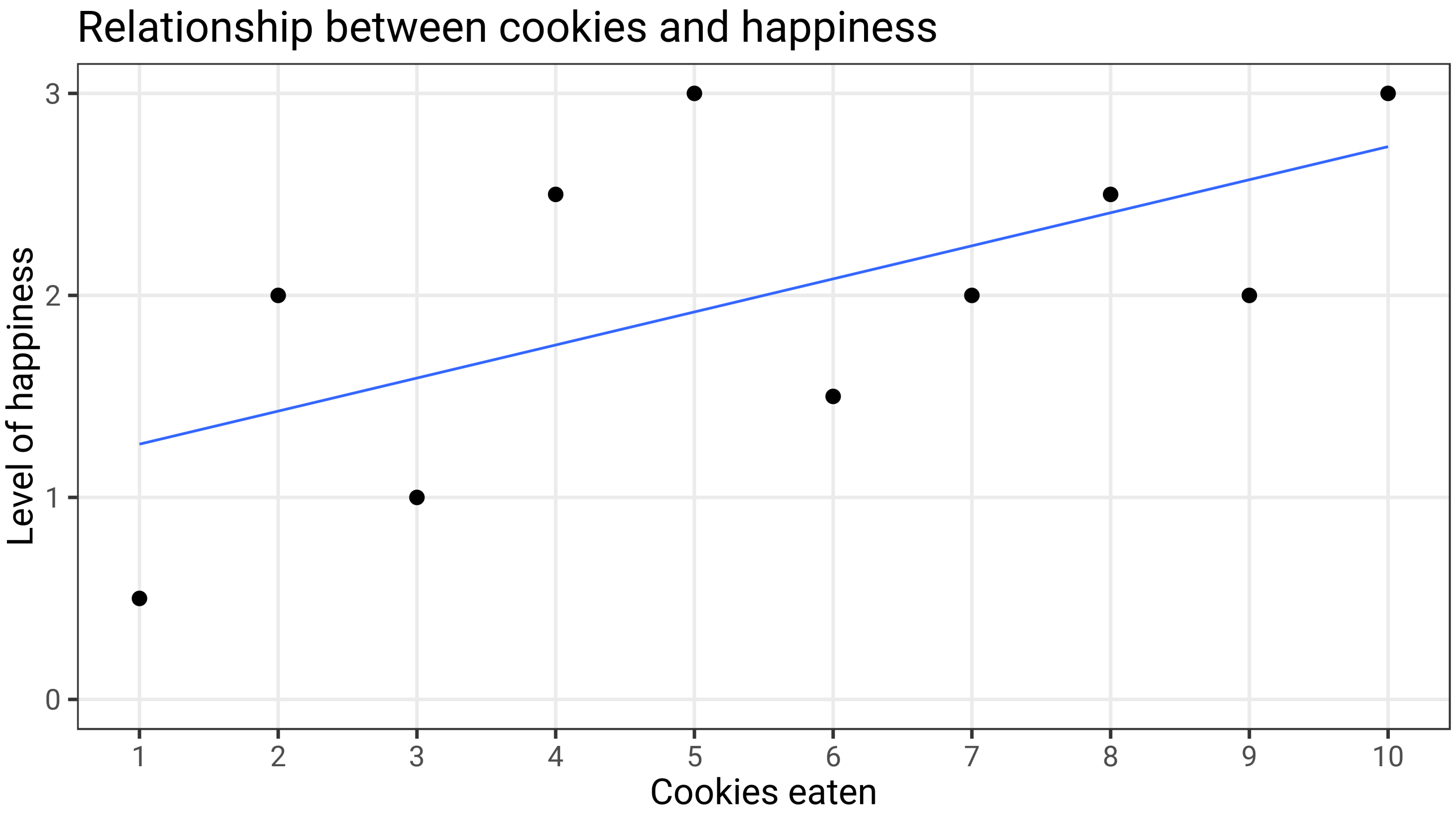
```
cookies_model <- lm(happiness ~ cookies,
                    data = cookies_data)

tidy(cookies_model)
```

```
# A tibble: 2 x 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
```

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{cookies} + \epsilon$$

$$\hat{\text{happiness}} = 1.1 + (0.164 \times \text{cookies}) + \epsilon$$



term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	1.1	0.47	2.339	0.047	0.016	2.184
cookies	0.164	0.076	2.159	0.063	-0.011	0.338

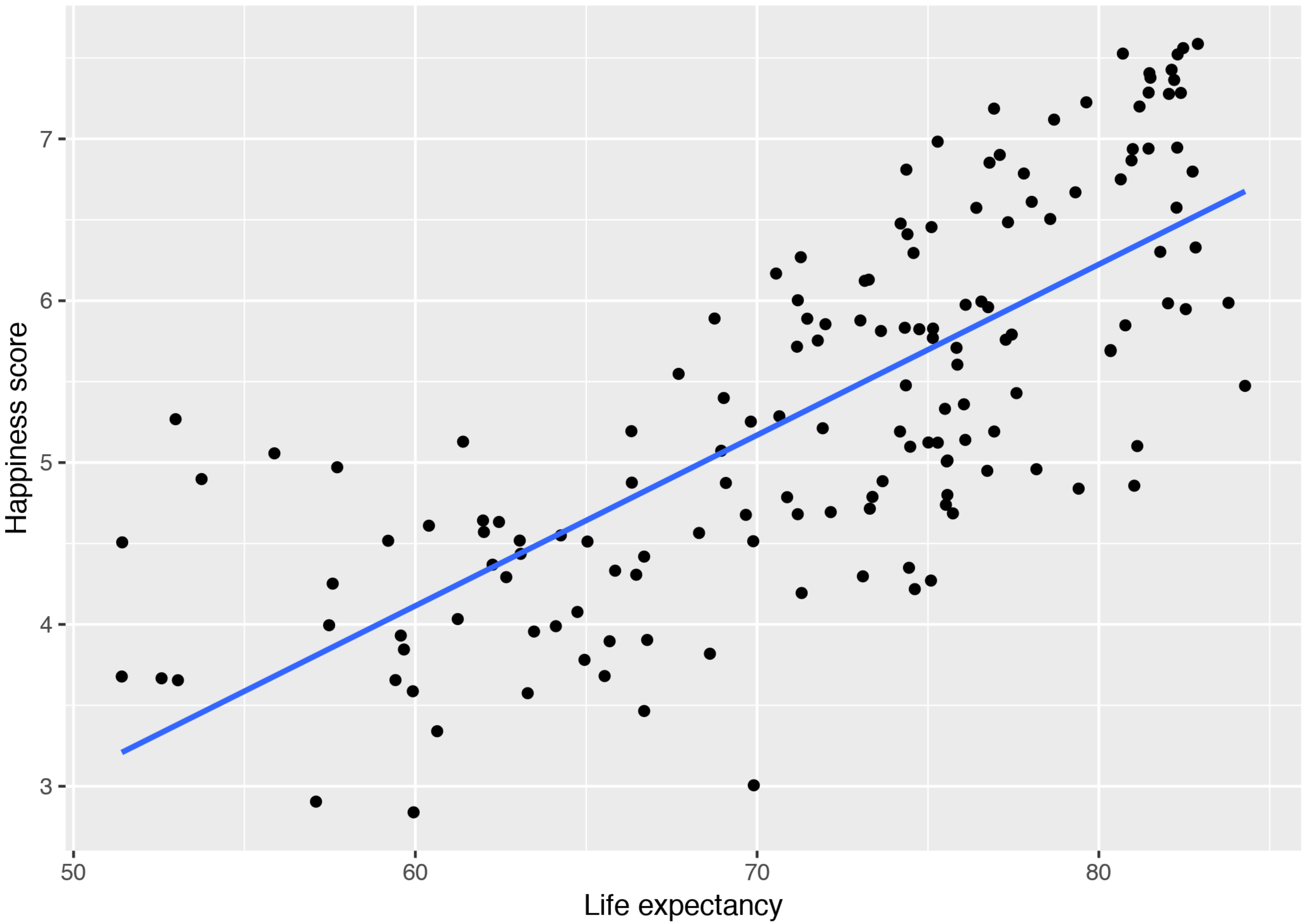
TEMPLATE

A one unit increase in X is *associated* with a β_1 increase (or decrease) in Y, on average

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

MULTIPLE REGRESSION

WORLD HAPPINESS



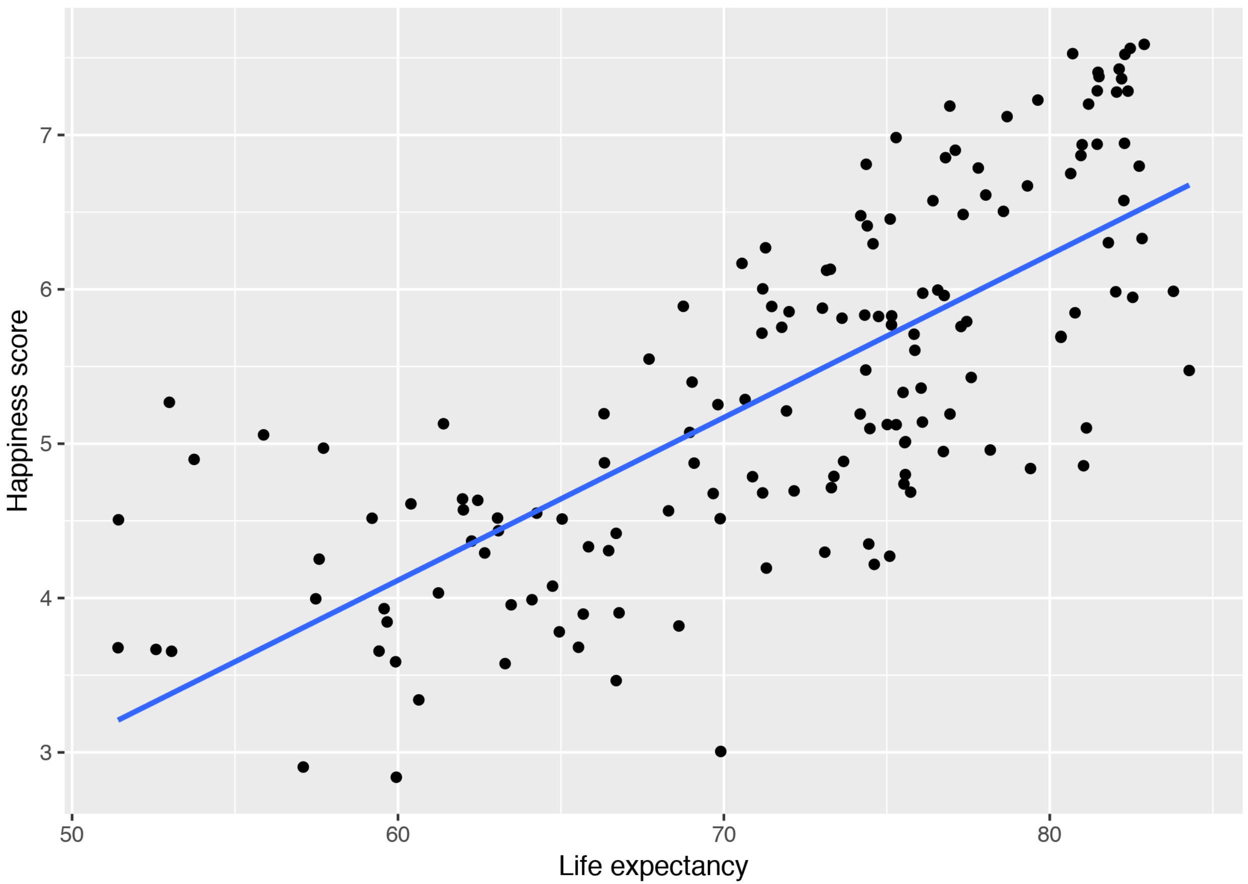
```r
model1 <- lm(happiness_score ~ life_expectancy,
             data = world_happiness)
tidy(model1)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.215	0.556	-3.983	0	-3.313	-1.116
life_expectancy	0.105	0.008	13.73	0	0.09	0.121

$$\widehat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$

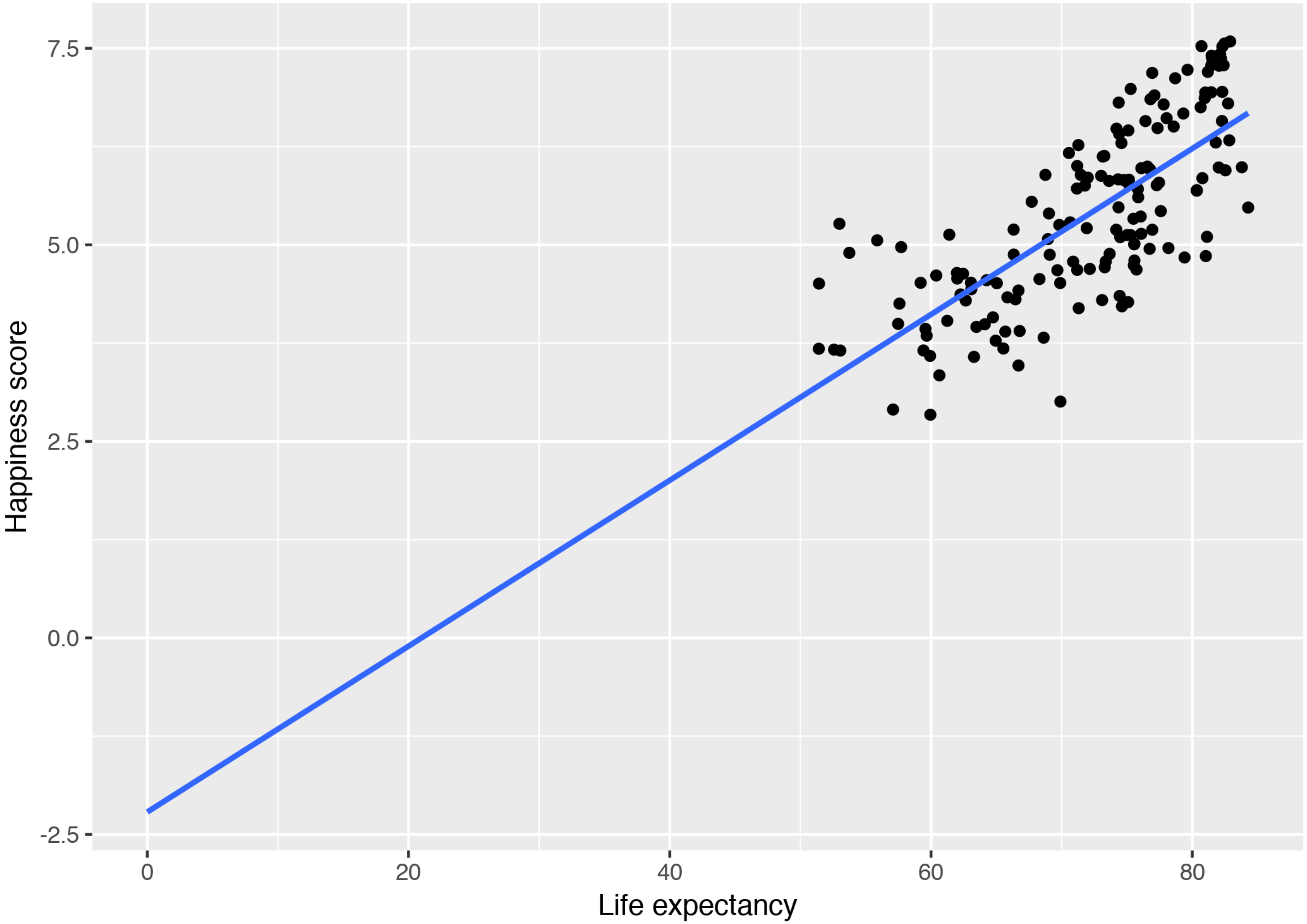
$$\widehat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

WORLD HAPPINESS



$$\hat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

WORLD HAPPINESS



VARIABLE TYPES

Numeric variables	Categorical variables
(Continuous)	(Factors)
Numbers	**Not numbers**

NUMERIC OR CATEGORICAL?

Income

True/false

- 18–25
- 26–34
- 35–44
- 45–54

State

Weight

Tax rates

Political party

Gender

- Strongly agree
- Agree
- Disagree
- Strongly disagree

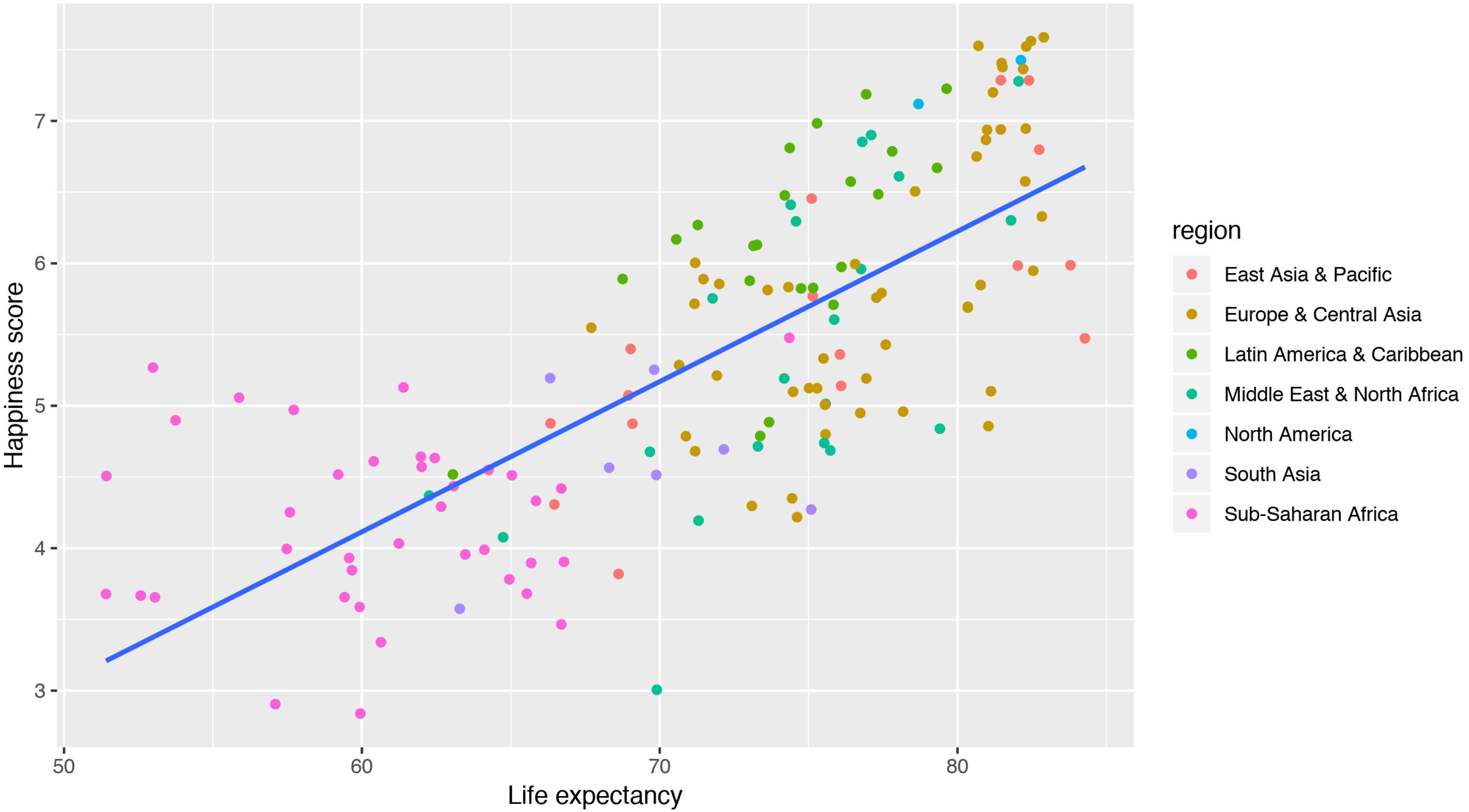
Year

Happiness

Age

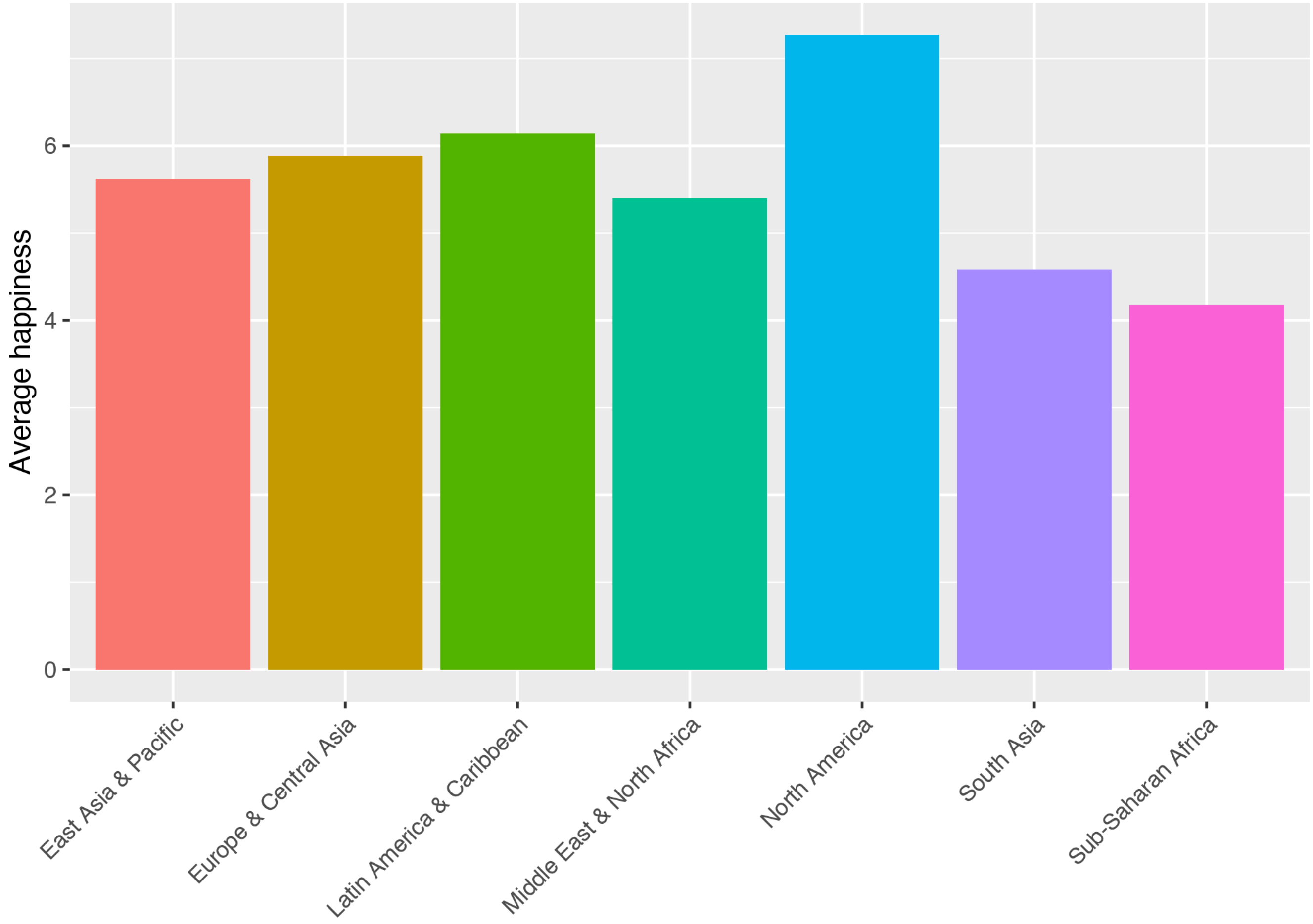
Day of the week

LIFE EXPECTANCY IS NOT THE FULL STORY



REGIONAL DIFFERENCES

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181



```r
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\widehat{\text{happiness}} = \beta_0 + \beta_1 \text{Europe} + \beta_2 \text{Latin America} + \\ \beta_3 \text{MENA} + \beta_4 \text{North America} + \\ \beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon$$

```r
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\begin{aligned}\widehat{\text{happiness}} = {} & 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + \\ & (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + \\ & (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon\end{aligned}$$

HAPPINESS IN EAST ASIA

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 0) + (0.527 \times 0) + (-0.214 \times 0) + (1.655 \times 0) + (-1.037 \times 0) + (-1.437 \times 0) + \epsilon$$

$$\hat{\text{happiness}} = 5.618$$

HAPPINESS IN EUROPE

$$
\begin{aligned}
\hat{\text{happiness}} = 5.618 &+ (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + \\
&(-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + \\
&(-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon
\end{aligned}
$$

$$
\begin{aligned}
\hat{\text{happiness}} = 5.618 &+ (0.271 \times 1) + (0.527 \times 0) + \\
&(-0.214 \times 0) + (1.655 \times 0) + \\
&(-1.037 \times 0) + (-1.437 \times 0) + \epsilon
\end{aligned}
$$

$$
\begin{aligned}
\hat{\text{happiness}} &= 5.618 + (0.271 \times 1) \\
&= 5.889
\end{aligned}
$$

Regression coefficients

term	estimate
intercept	5.618
regionEurope & Central Asia	0.271
regionLatin America & Caribbean	0.527
regionMiddle East & North Africa	-0.214
regionNorth America	1.655
regionSouth Asia	-1.037
regionSub-Saharan Africa	-1.437

Averages

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181

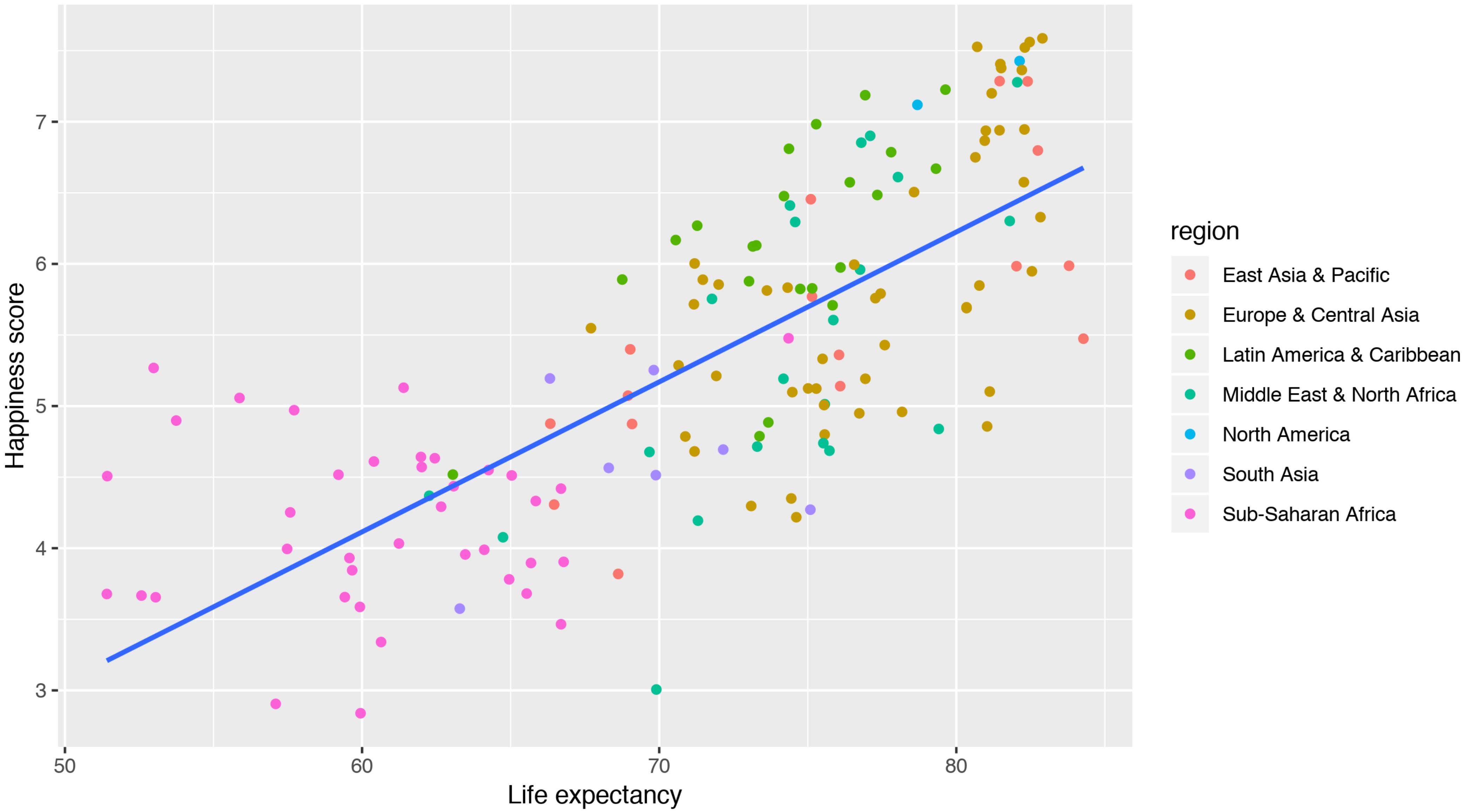
TEMPLATE

On average, y is β_n units larger (or smaller) in x_n, compared to x_0

On average, national happiness is 1.65 points higher in North America than in East Asia

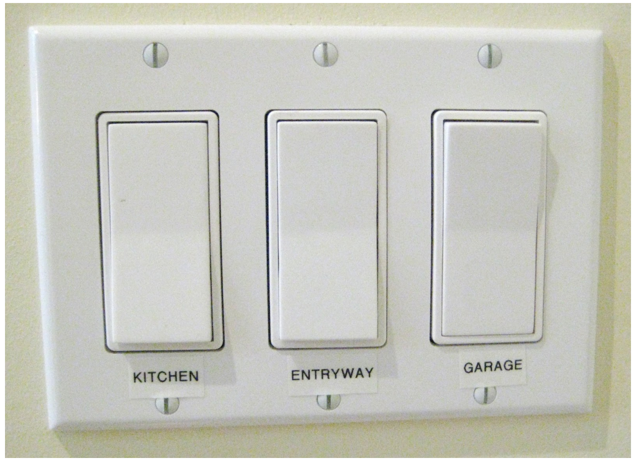
On average, compared to East Asia, national happiness is 1.44 points lower in Sub Saharan Africa

GETTING CLOSER



SLIDERS AND SWITCHES

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$

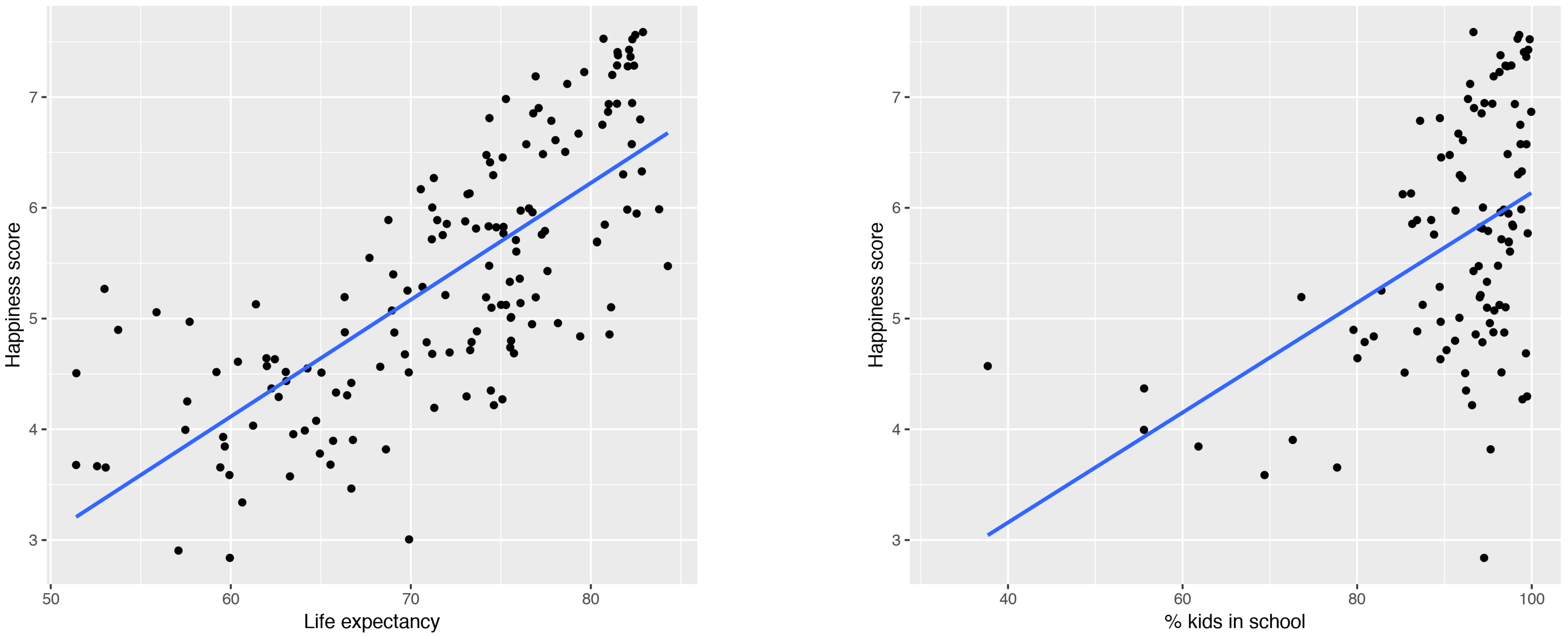


$$\begin{aligned}\hat{\text{happiness}} = \beta_0 + \beta_1 \text{Europe} + \beta_2 \text{Latin America} + \\ \beta_3 \text{MENA} + \beta_4 \text{North America} + \\ \beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon\end{aligned}$$

ALL AT ONCE!

$$\widehat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} + \\ \beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} + \\ \beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

HAPPINESS ~ LIFE + SCHOOL



```r
model_life <- lm(happiness_score ~ life_expectancy,
                 data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.215	0.556	-3.983	0	-3.313	-1.116
life_expectancy	0.105	0.008	13.73	0	0.09	0.121

```r
model_school <- lm(happiness_score ~ school_enrollment,
                   data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci
intercept	1.173	0.879	1.334	0.185	-0.571
school_enrollment	0.05	0.01	5.19	0	0.031

BOTH AT THE SAME TIME

Life expectancy and school enrollment both explain some variation in happiness

On its own, a 1 year increase in school enrollment is associated with a 0.105 point increase in happiness, on average

On its own, a 1% increase in school enrollment is associated with a 0.05 point increase in happiness, on average

Some of that explanation is shared!

```r
model_life_school <- lm(happiness_score ~ life_expectancy +
                          school_enrollment,
                        data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci
intercept	-2.111	0.835	-2.529	0.013	-3.767
life_expectancy	0.101	0.014	7.447	0	0.074
school_enrollment	0.003	0.01	0.331	0.741	-0.016

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} + \epsilon$$

$$\hat{\text{happiness}} = -2.11 + (0.101 \times \text{life expectancy}) + (0.003 \times \text{school enrollment}) + \epsilon$$

FILTERING OUT VARIATION

Each x in the model explains some portion of the variation in y

This will often change the simple regression coefficients

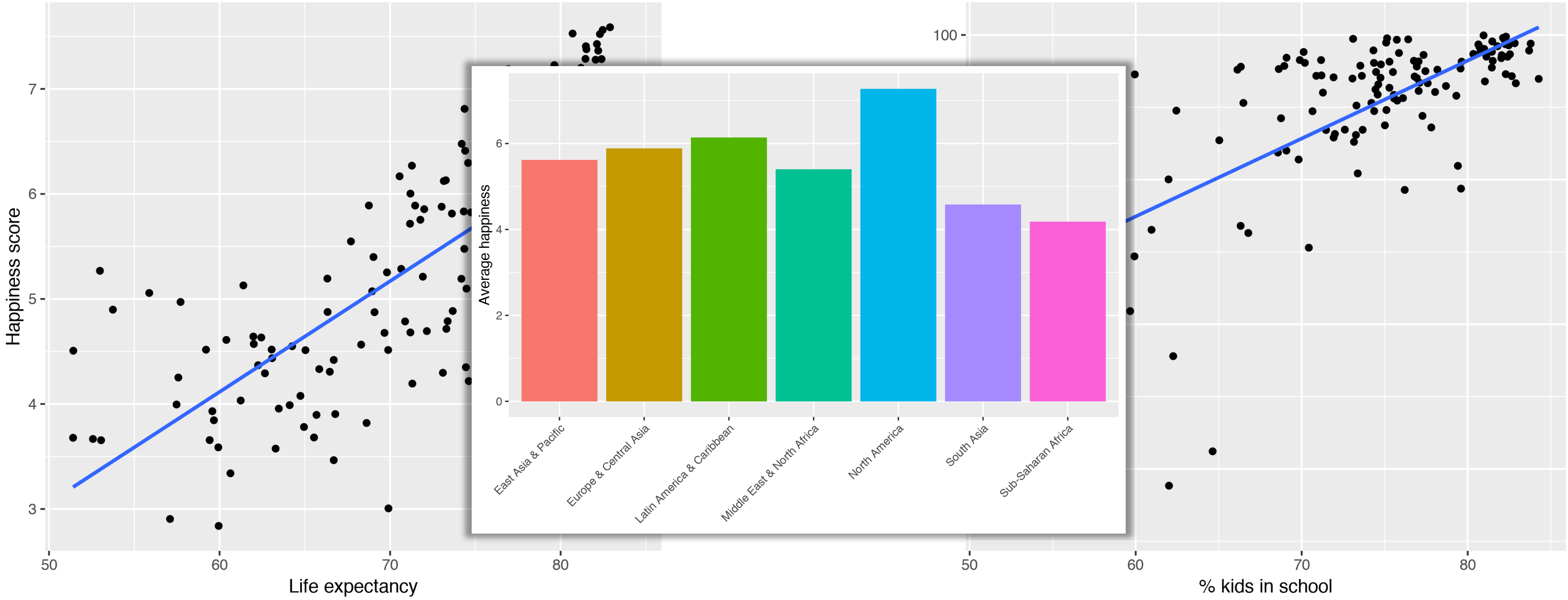
Interpretation is a little trickier, since you can only ever move **one** switch or slider (or variable)

TEMPLATE

Taking all other variables in the model into account, a one unit increase in x_n is associated with a β_n increase (or decrease) in y, on average

Controlling for school enrollment, a 1 year increase in life expectancy is associated with a 0.1 point increase in national happiness, on average

HAPPINESS ~ LIFE + SCHOOL + REGION



```r
model_life_school_region <-
  lm(happiness_score ~ life_expectancy + school_enrollment + region,
     data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	-2.821	1.355	-2.083	0.04
life_expectancy	0.102	0.017	5.894	0
school_enrollment	0.008	0.01	0.785	0.435
regionEurope & Central Asia	0.031	0.255	0.123	0.902
regionLatin America & Caribbean	0.732	0.294	2.489	0.015
regionMiddle East & North Africa	0.189	0.317	0.597	0.552
regionNorth America	1.114	0.581	1.917	0.058
regionSouth Asia	-0.249	0.45	-0.553	0.582
regionSub-Saharan Africa	0.326	0.407	0.802	0.425

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

REGRESSION AND INFERENCE

Does attending a private university cause an increase in earnings?

How can we create fake treatment and control groups?

TABLE 2.1
The college matching matrix

Applicant group	Student	Private: Ivy	Private: Leafy	Private: Smart	Public: All State	Public: Tall State	Public: Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Why can’t we just calculate mean(private) − mean(public)

The people in groups A and B aren’t the same

Table 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Private – public

−$5,000

$30,000

???

???

REGRESSION AND CONTROLS

$$y_i = \alpha + \beta P_i + \gamma A_i + \epsilon_i$$

$$\text{earnings} = \alpha + \beta_1 \text{Private} + \beta_2 \text{Group A} + \epsilon$$

```
model_earnings <- lm(Earnings ~ Private + Group A, data = schools)
```

term	estimate	std_error	statistic	p_value
Intercept	40000	11952.29	3.3467	0.08
Private	10000	13093.07	0.7638	0.52
Group A	60000	13093.07	4.5826	0.04