CORRELATION AND BASIC REGRESSION

MPA 630: Data Science for Public Management

October 11, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

Revisiting correlation

Introduction to regression

Drawing the best lines

Lines and math

Translating lines to stats

REVISITING CORRELATION

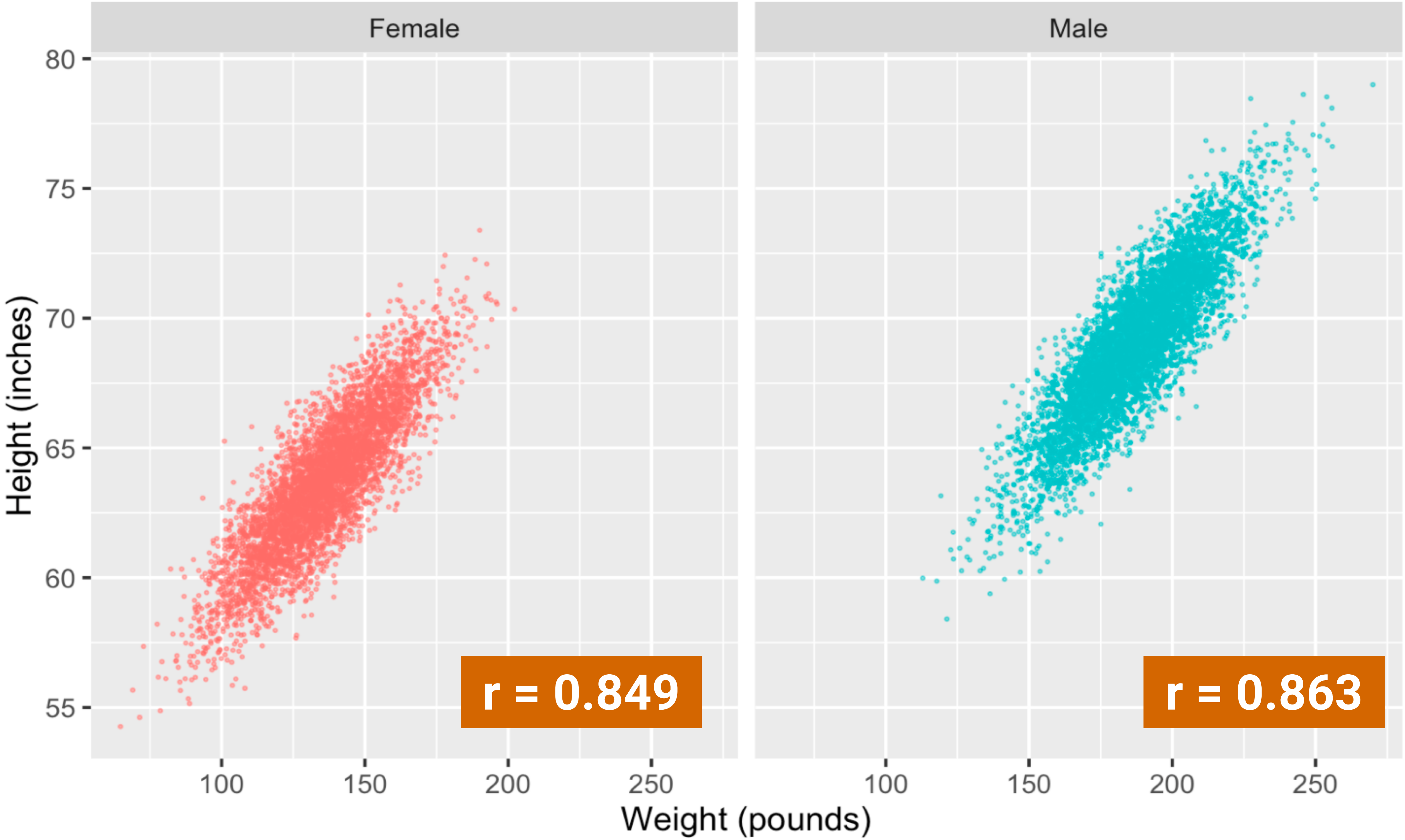
CORRELATION

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

How closely two variables are related + direction of relation

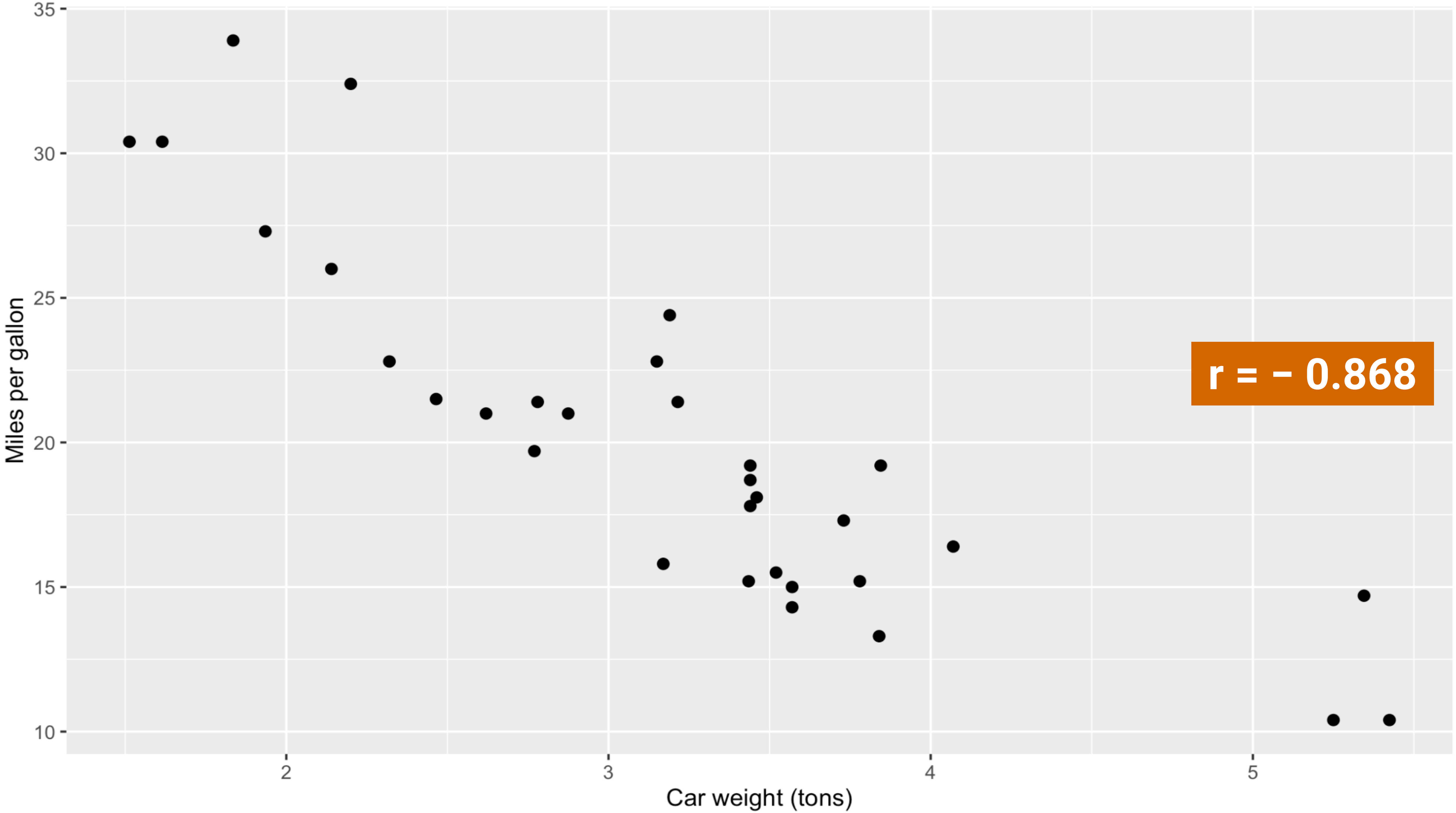
−1 to 1

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

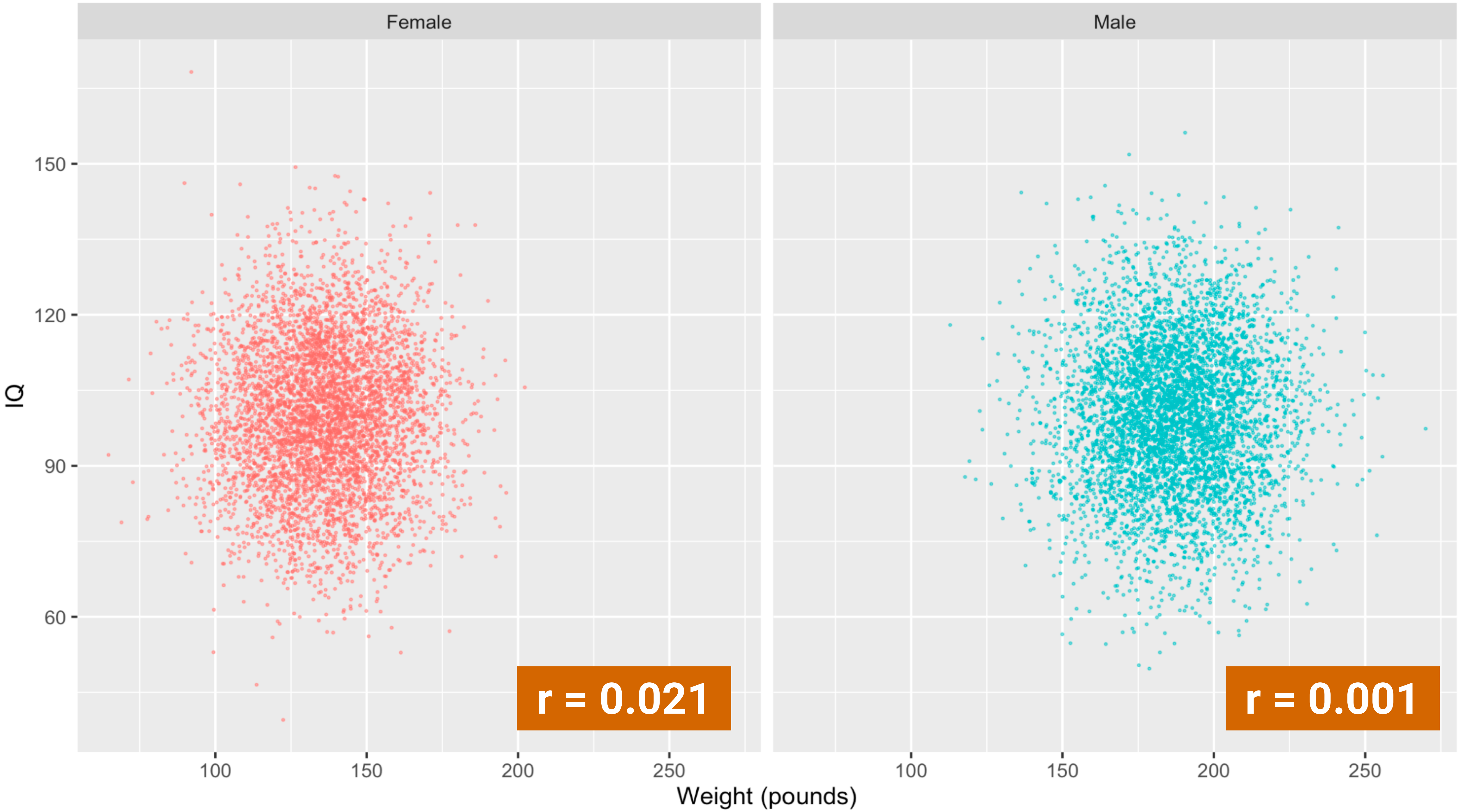

Female
Male
80
75
70
65
60
55
Height (inches)
100
150
200
250
100
150
200
250
Weight (pounds)

r = 0.849

r = 0.863



r = − 0.868


Female
Male
IQ
150
120
90
60
100
150
200
250
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

GUESS THE CORRELATION

X	Y	
Vehicle velocity	Travel time	–
Salinity of water	Buoyancy	+
Alcohol consumed	Judgment	–
Income	Happiness	?
Age	Health	?
Hair length	Shampoo use	+
Tadpole age	Tadpole tail length	–

INTRODUCTION TO REGRESSION

WHY

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

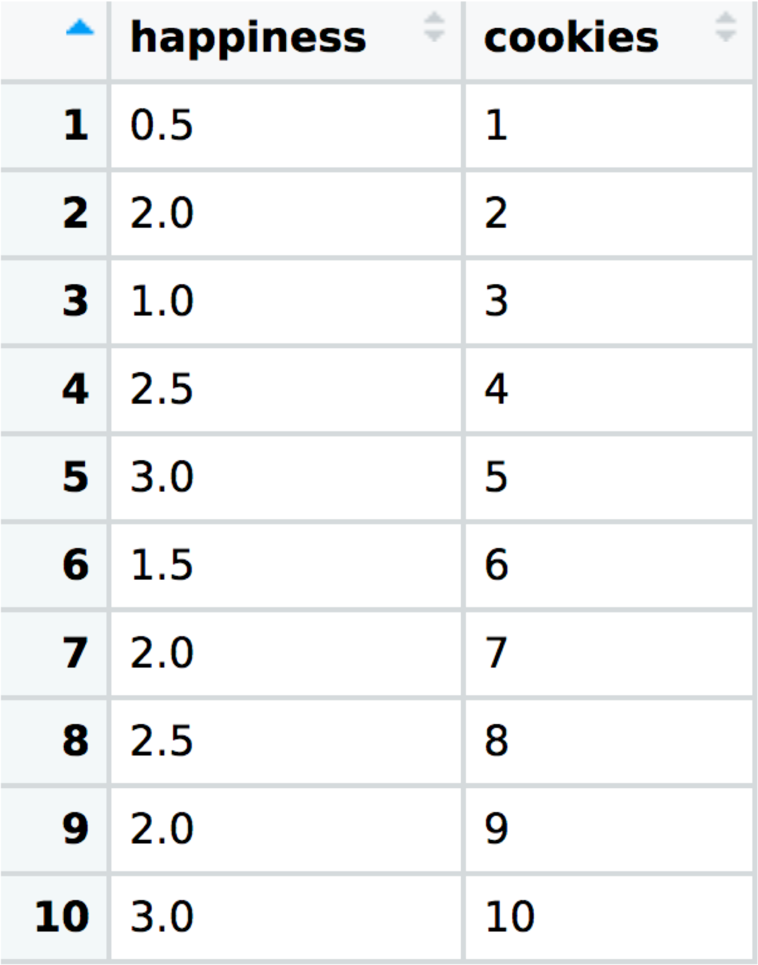
Draw a line that approximates the relationship

Find mathy parts of the line

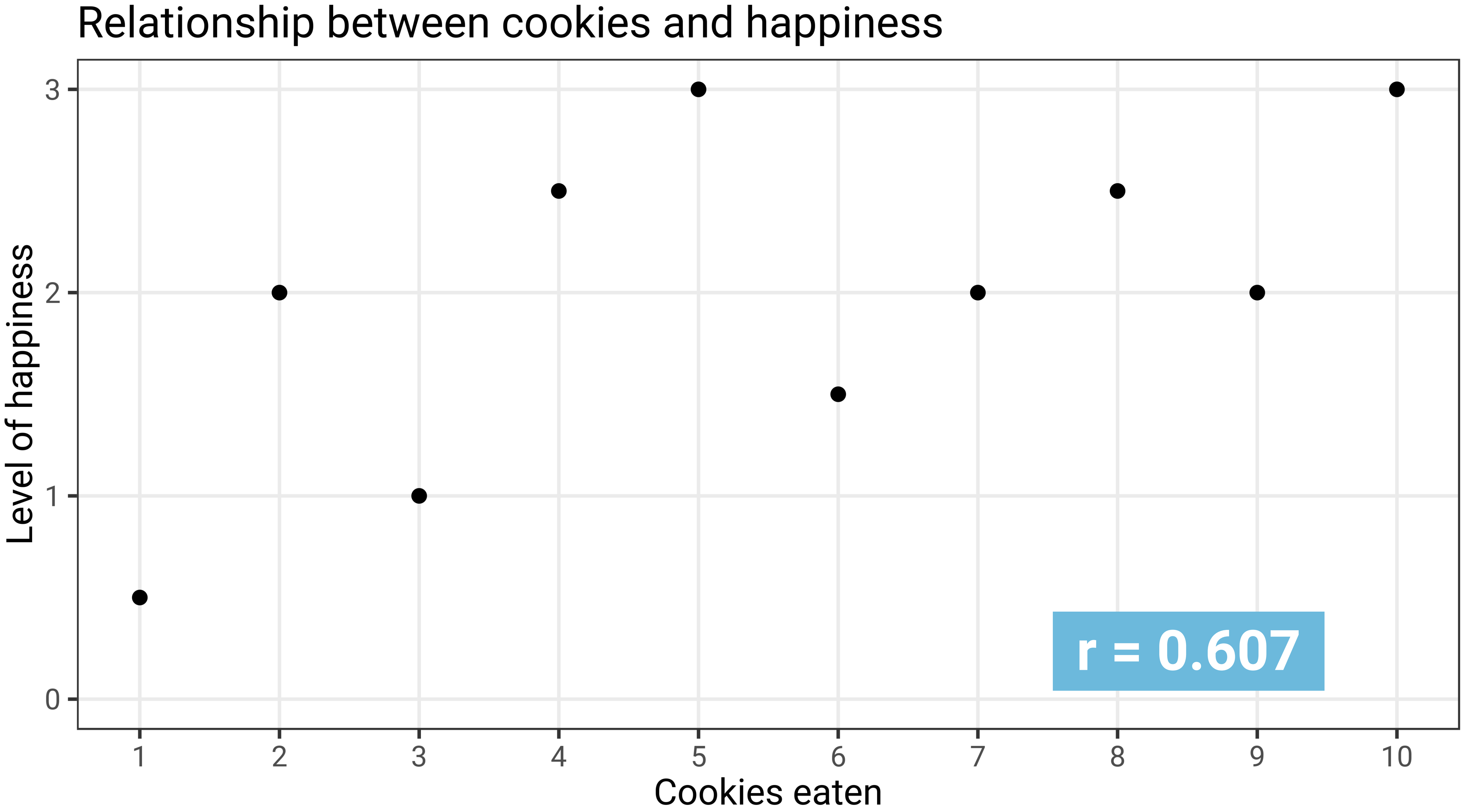
Interpret the math

DRAWING THE BEST LINES

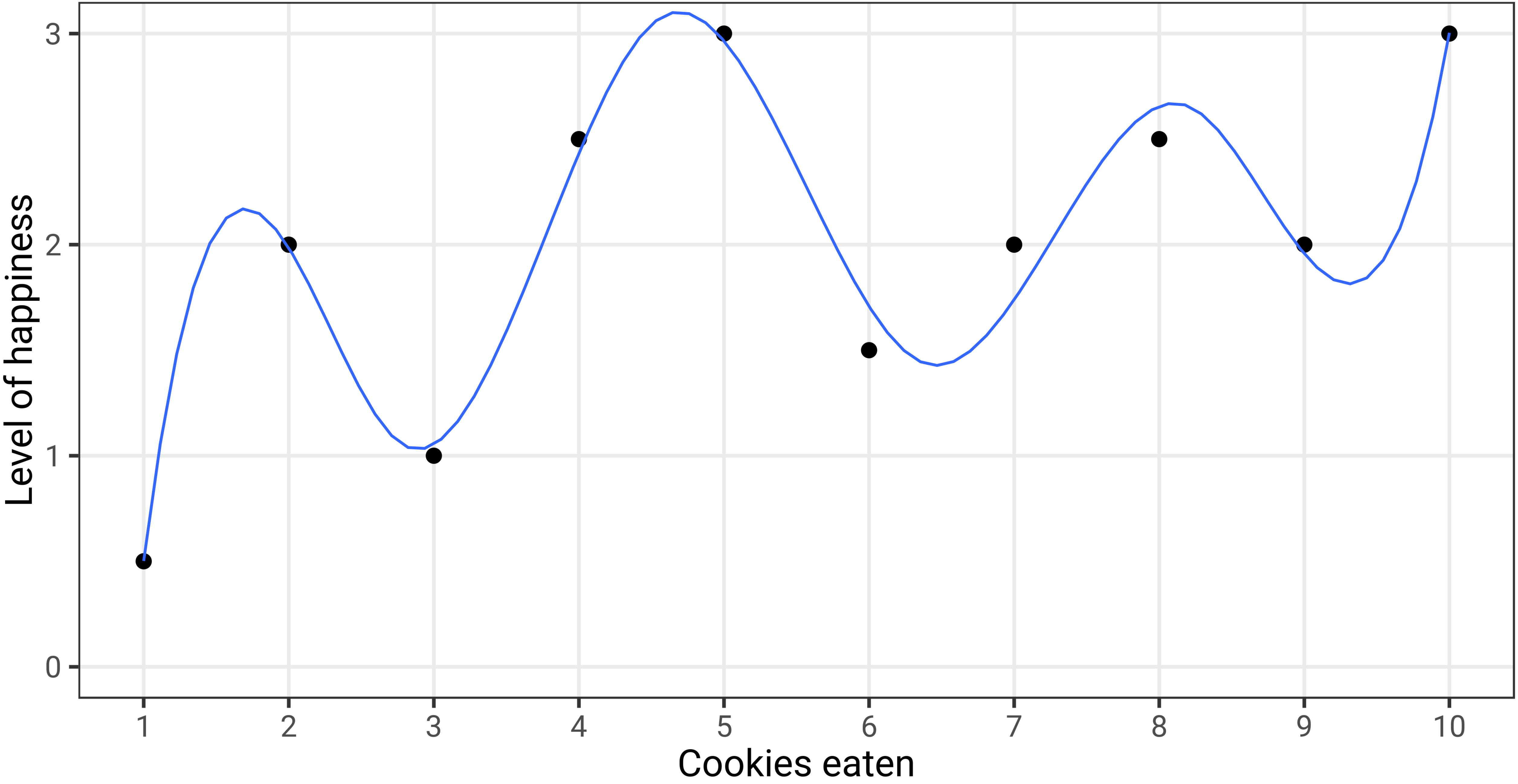
COOKIE CONSUMPTION AND HAPPINESS

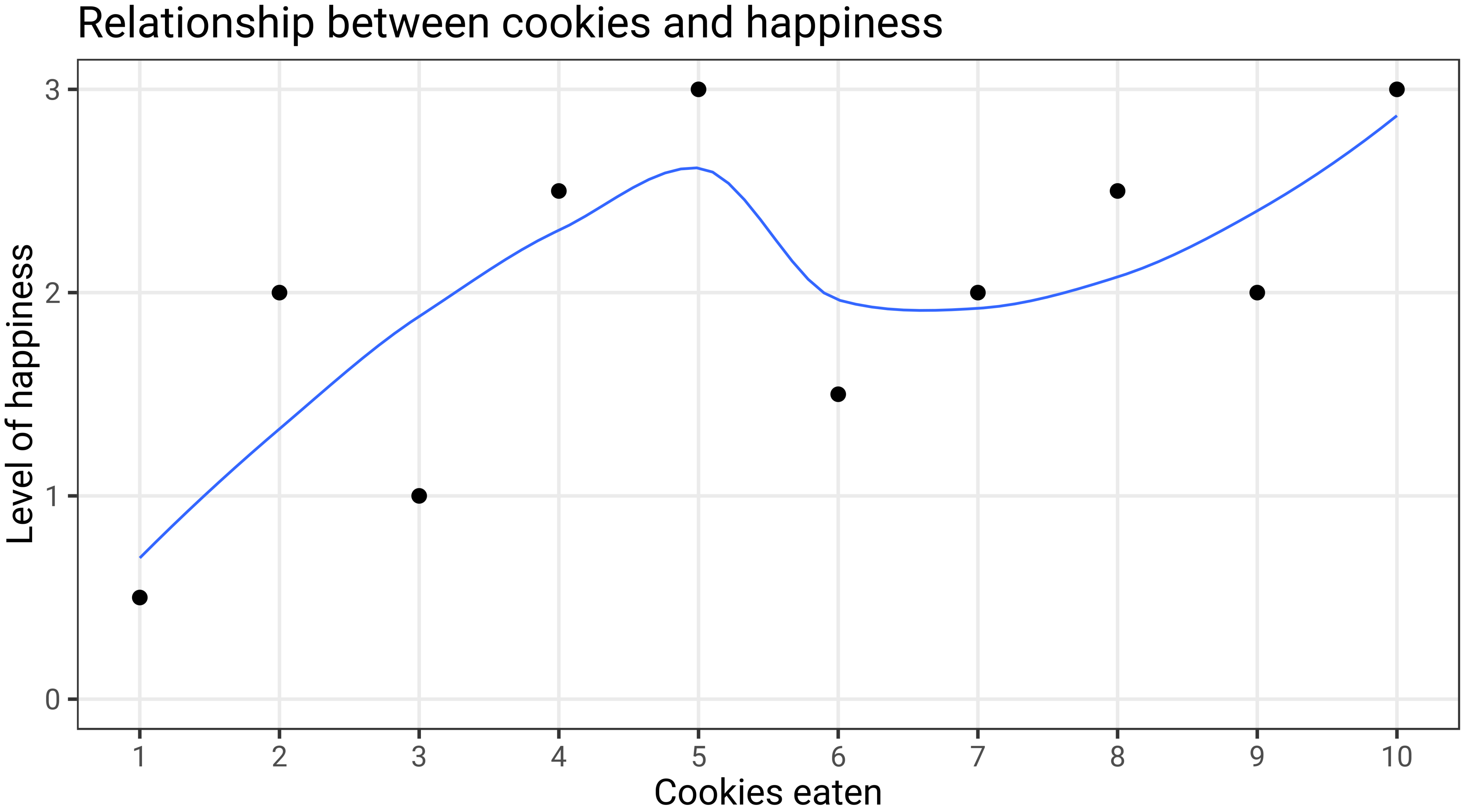


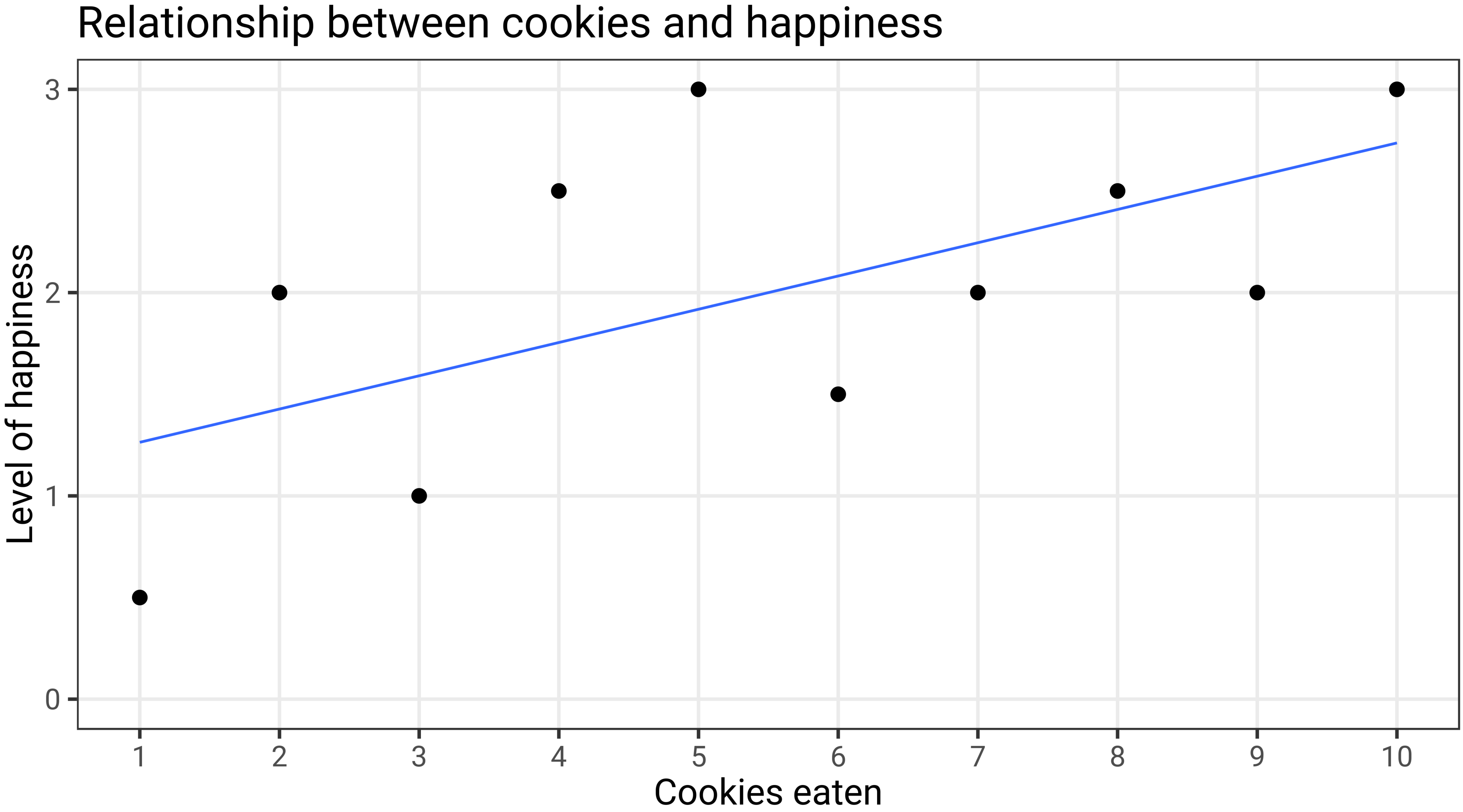
	happiness	cookies
1	0.5	1
2	2.0	2
3	1.0	3
4	2.5	4
5	3.0	5
6	1.5	6
7	2.0	7
8	2.5	8
9	2.0	9
10	3.0	10

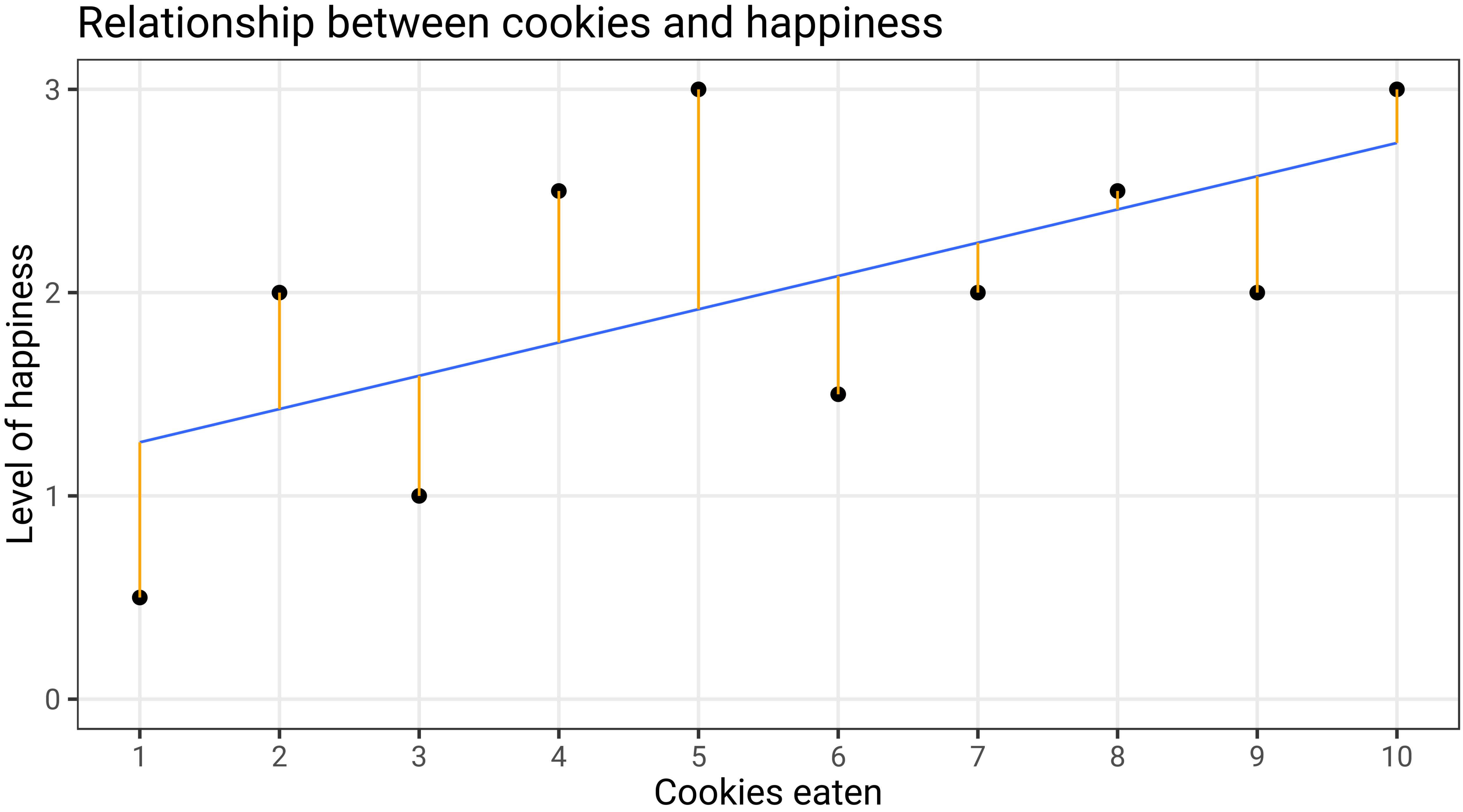

Relationship between cookies and happiness
Level of happiness
0
1
2
3
Cookies eaten
1
2
3
4
5
6
7
8
9
10
r = 0.607

Relationship between cookies and happiness

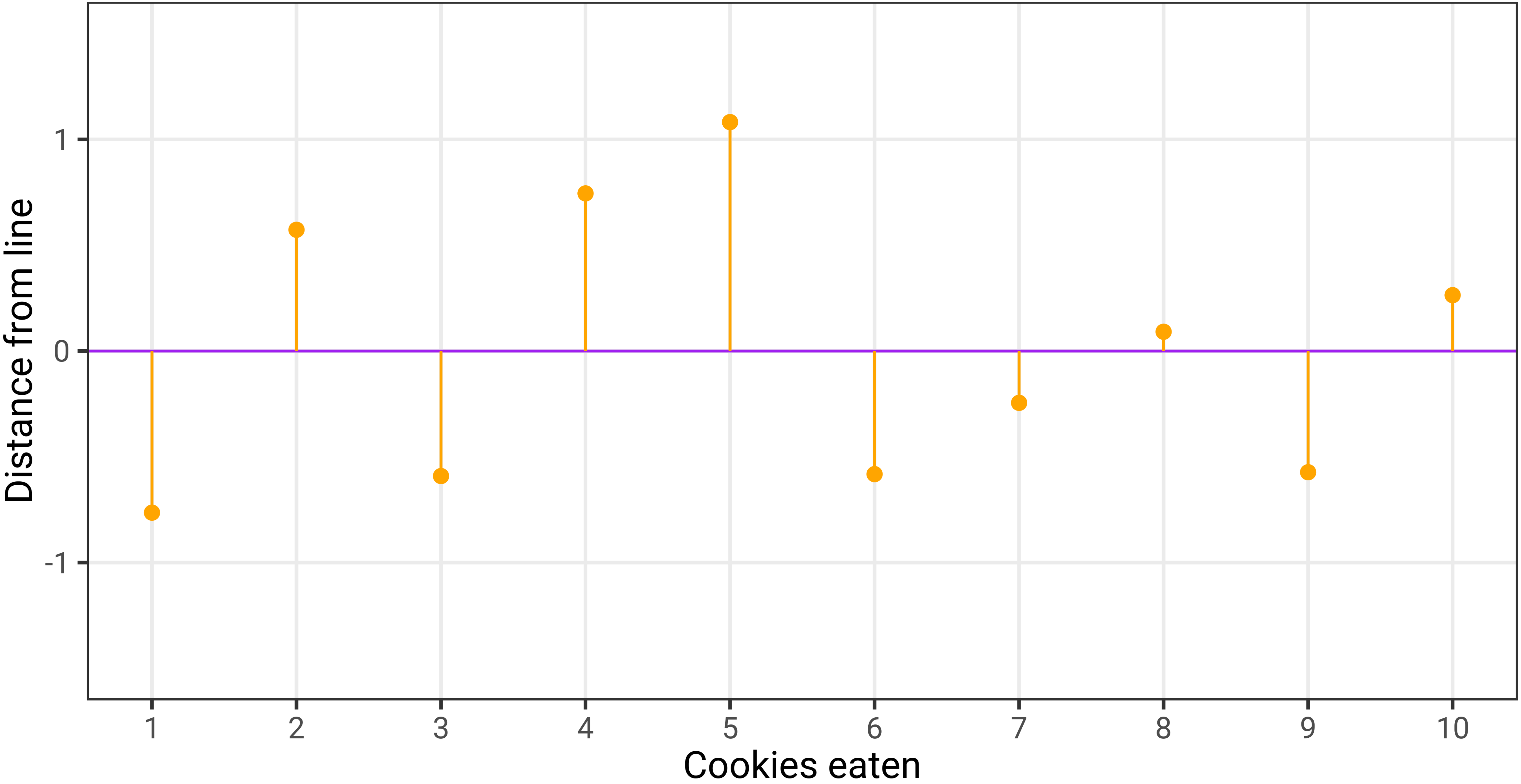


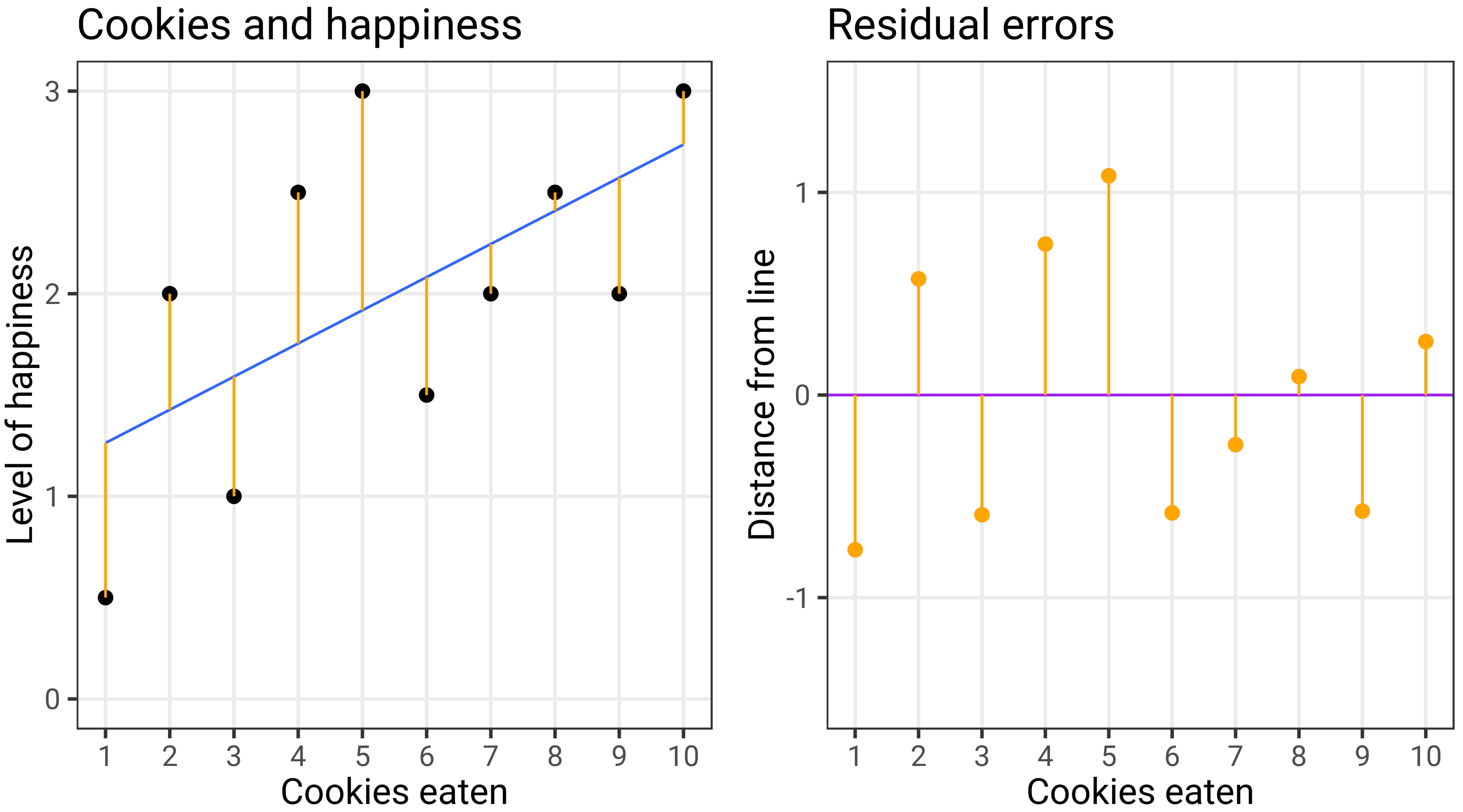

Relationship between cookies and happiness
Level of happiness
0
1
2
3
Cookies eaten
1
2
3
4
5
6
7
8
9
10


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


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

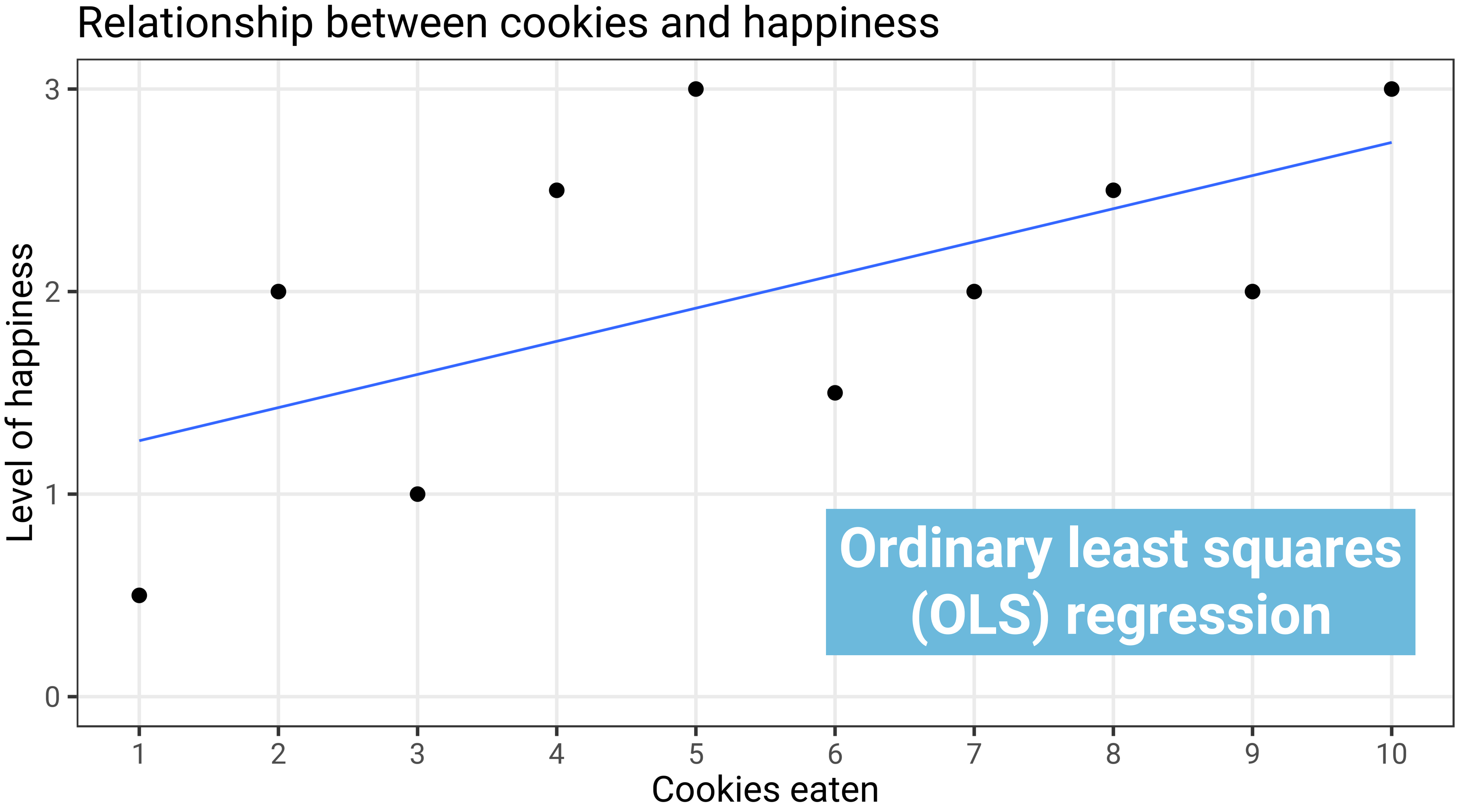
Residual errors (distance from line)





Cookies and happiness

Residual errors



Relationship between cookies and happiness

Ordinary least squares (OLS) regression

LINES AND MATH

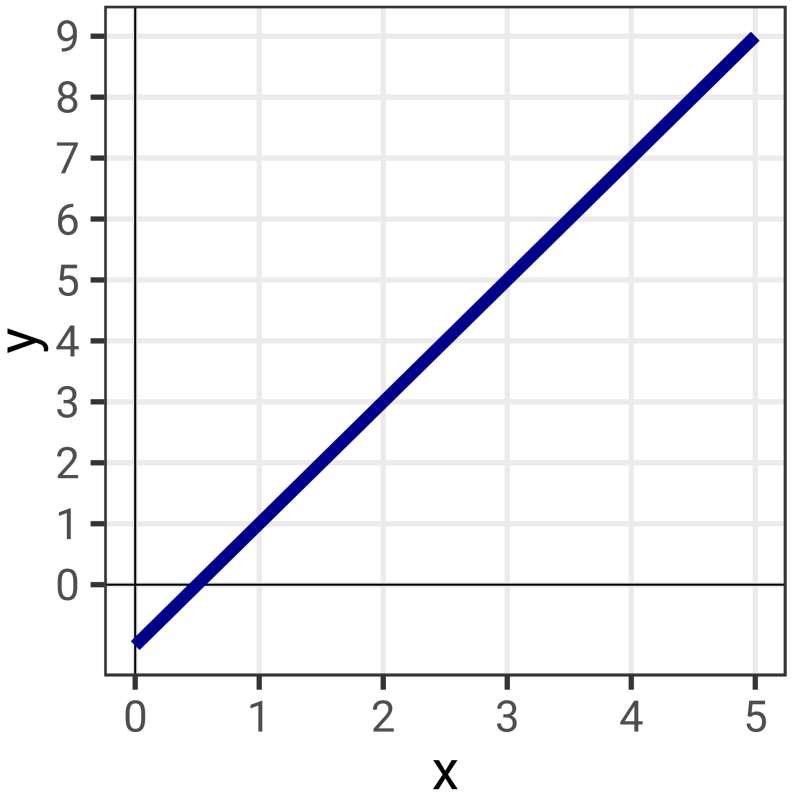
DRAWING LINES WITH MATH

$$y = mx + b$$

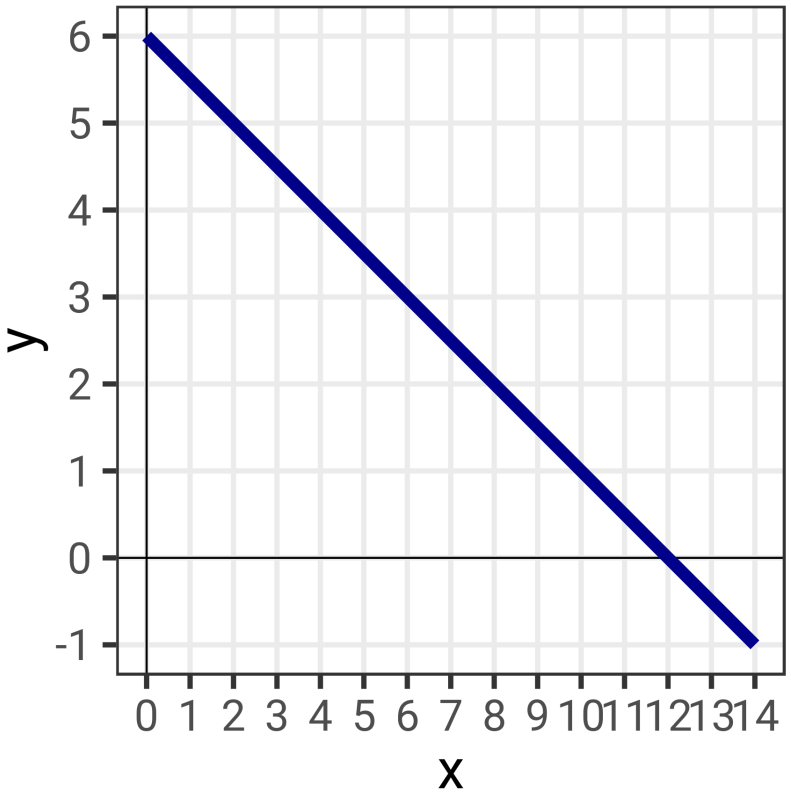
y	A number	
x	A number	
m	Slope	$\frac{rise}{run}$
b	y intercept	

SLOPES AND INTERCEPTS

$y = 2x - 1$



$y = -0.5x + 6$



GRAPH THESE

$y = 5x + 2$

$y = x - 1$

$y = -2x + 11$

$y = 6 - 2x$

$y = 0.33x - 1$

$y = 0.75x - 3$

TRANSLATING LINES TO STATISTICS

DRAWING LINES WITH STATS

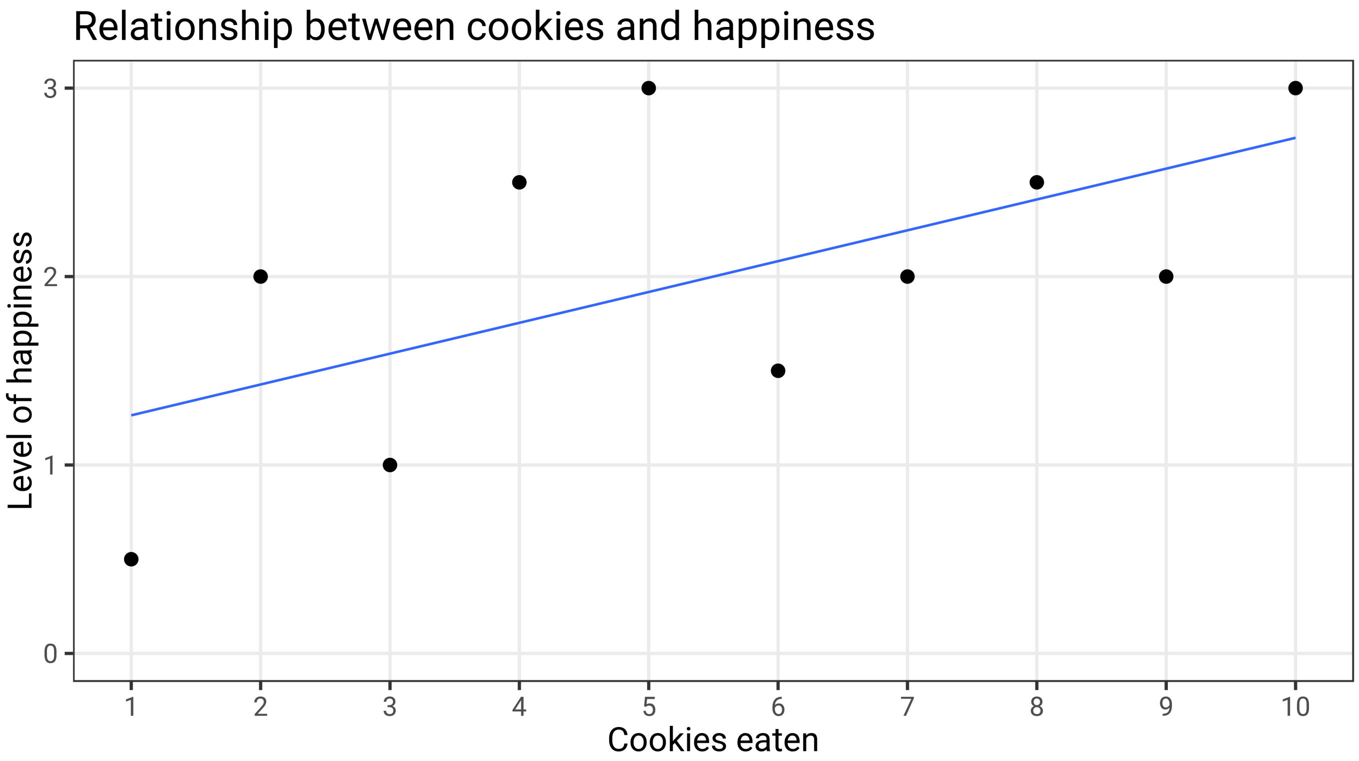
$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

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

MODELING COOKIES AND HAPPINESS

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

$$\widehat{happiness} = \beta_0 + \beta_1 cookies + \varepsilon$$



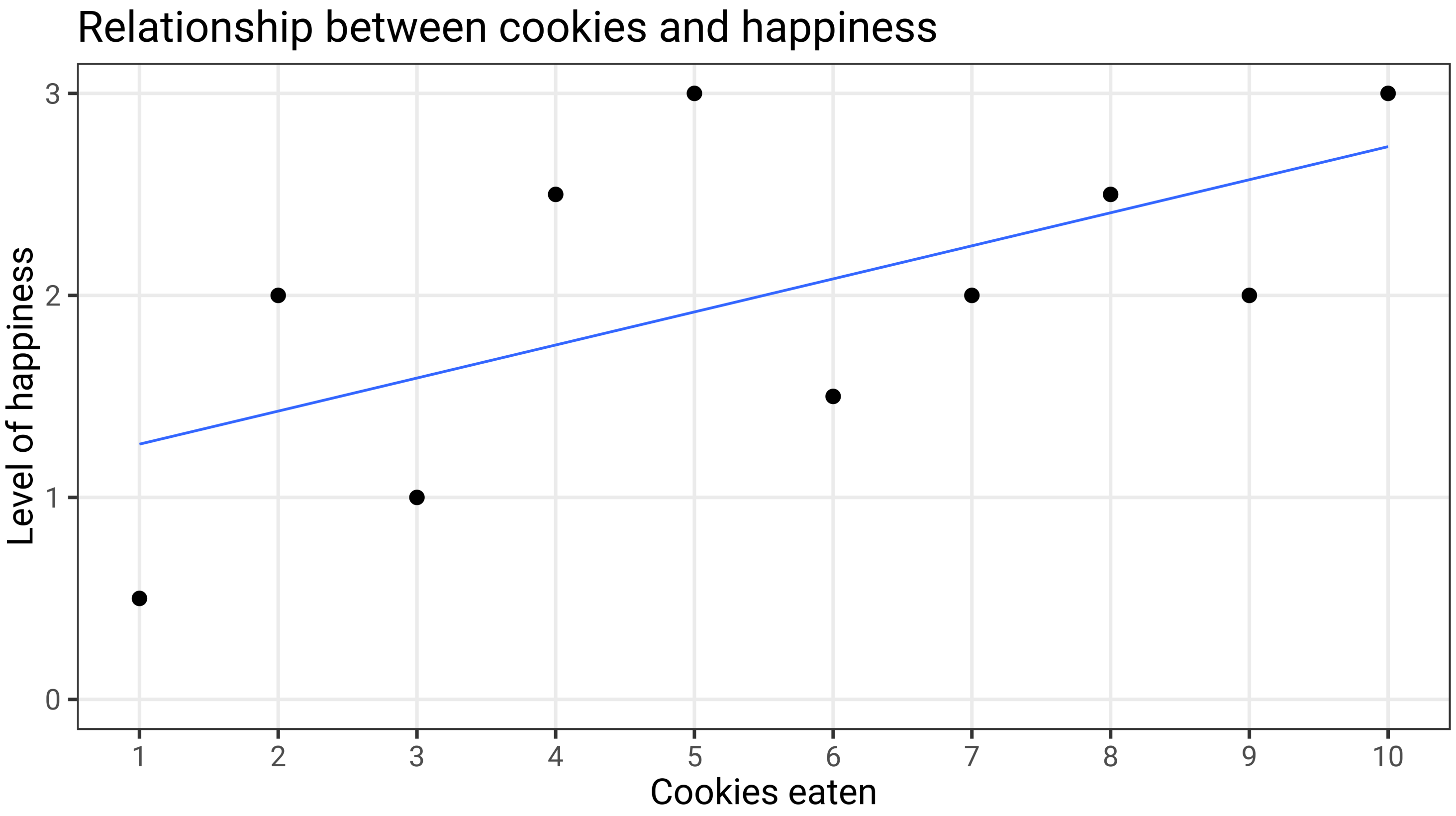
MODELING COOKIES AND HAPPINESS

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

cookies_model %>%
  get_regression_table()
```

```
# 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
```

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



term	estimate	std_er	stati	p_val	lower	uppe
intercept	1.1	0	2 9	0	0.	2 4
cookies	0.164	76	159	63	11	338

Chapter 10

Chapter 11

Chapter 11

Chapter 9

Chapter 9

TEMPLATE

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

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

REAL LIFE EXAMPLE