

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

REVISING CORRELATION

CORRELATION

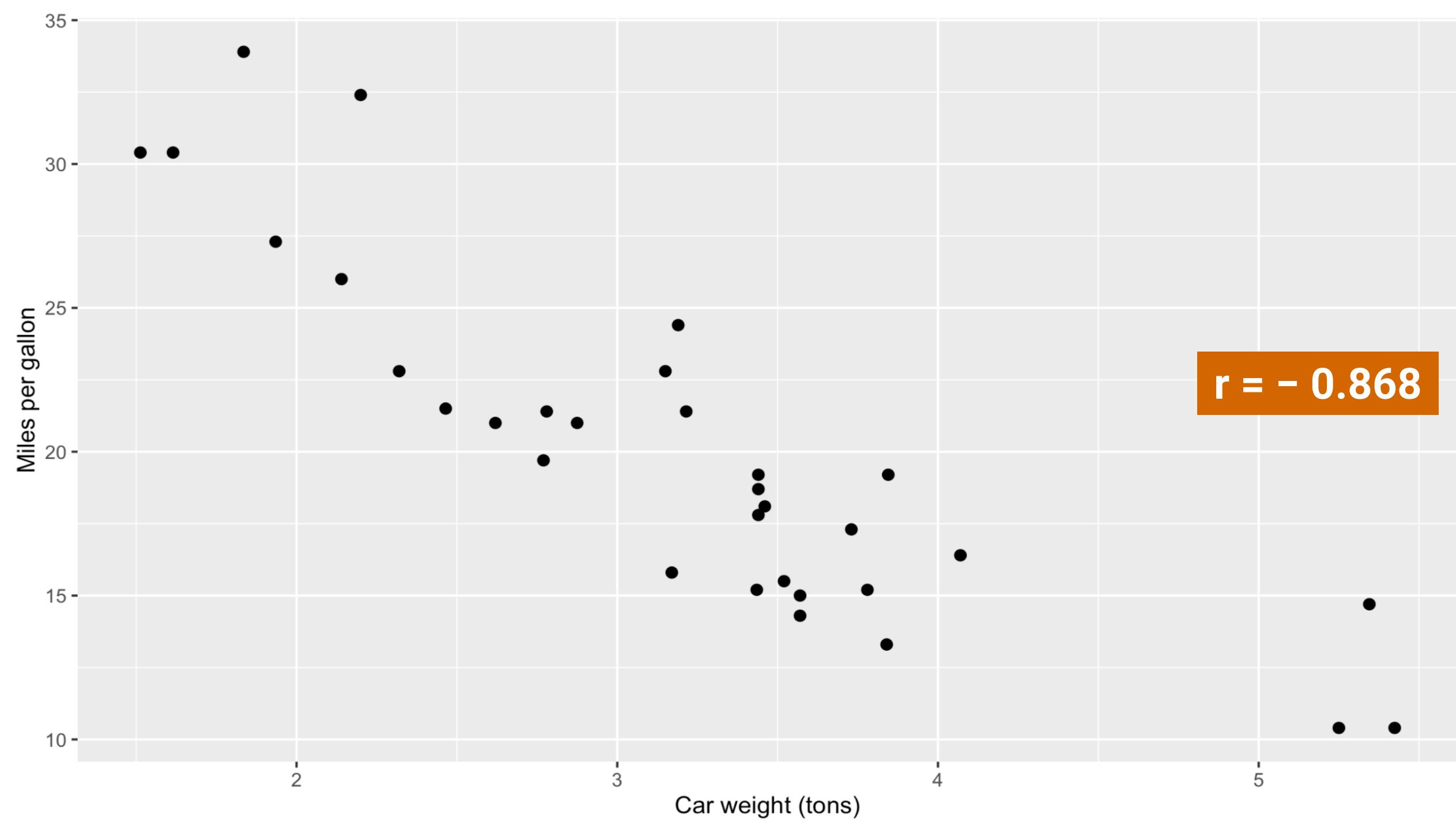
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

How closely two variables are related + direction of relation

-1 to 1

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





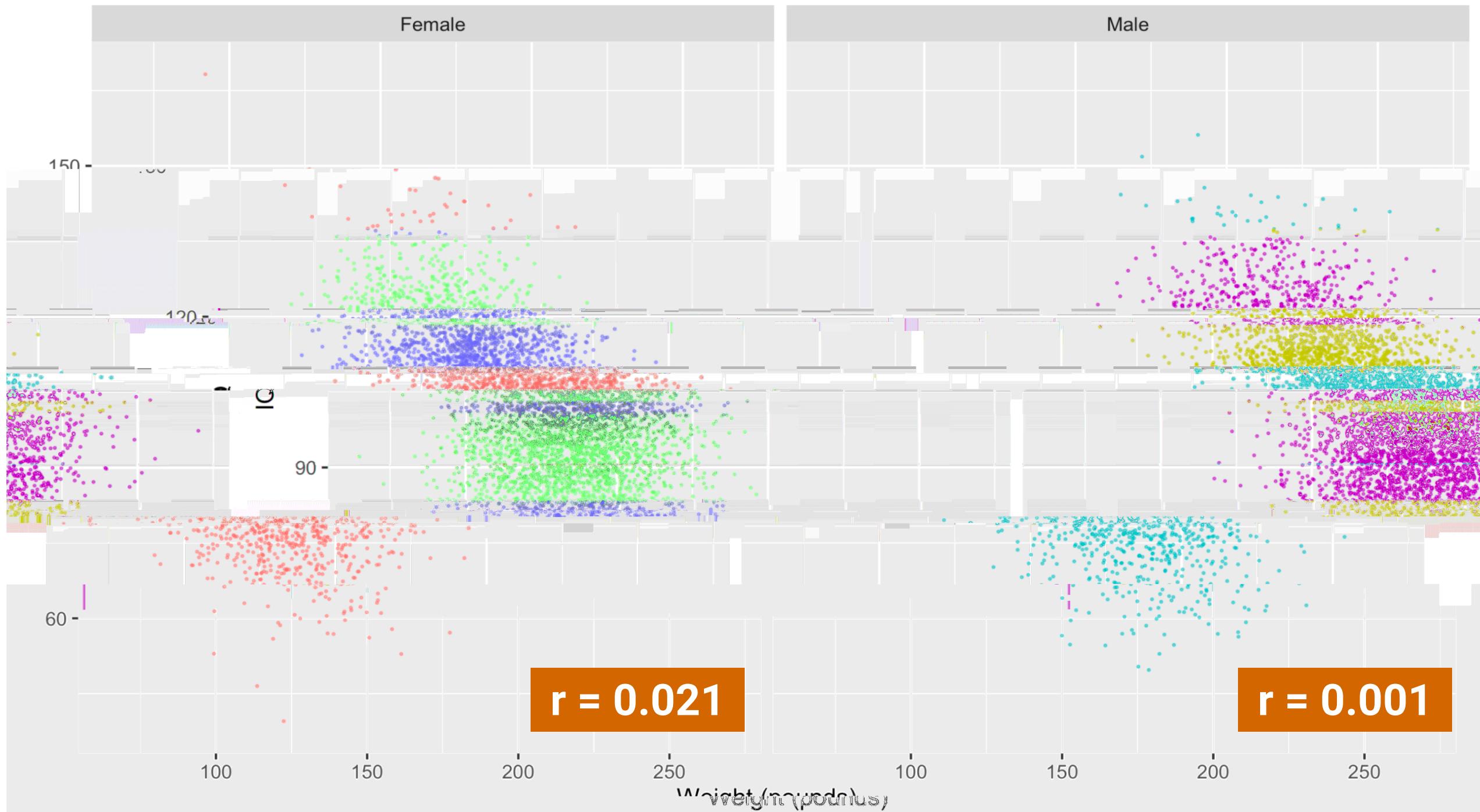
Female

Male

$r = 0.021$

$r = 0.001$

Wvirbty(m/podium)



GENERAL GUIDELINES

0	No relationship	Can be positive or negative
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	

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

Forecast the future

Focus is on Y

Netflix trying to
guess your next show

Predicting who will
escape poverty

Explanation

Explain effect of X on Y

Focus is on X

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

Looking at the effect of food
stamps on poverty reduction

HOW

Plot X and Y

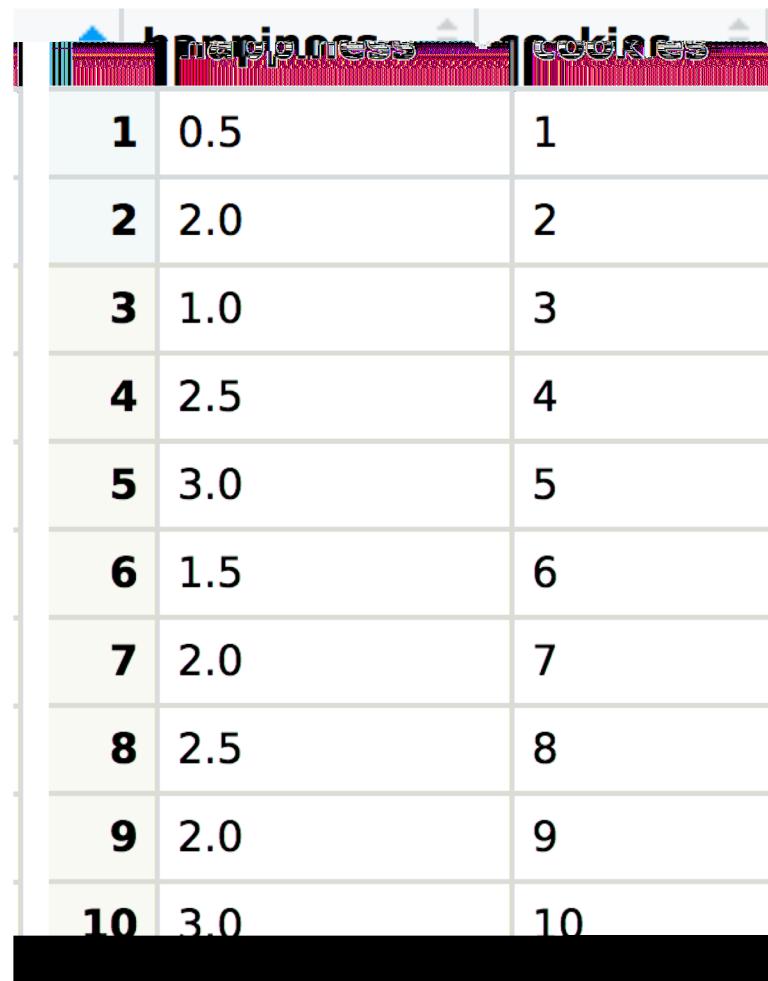
Draw a line that approximates
the relationship

Find mathy parts of the line

Interpret the math

DRAWING THE BEST LINES

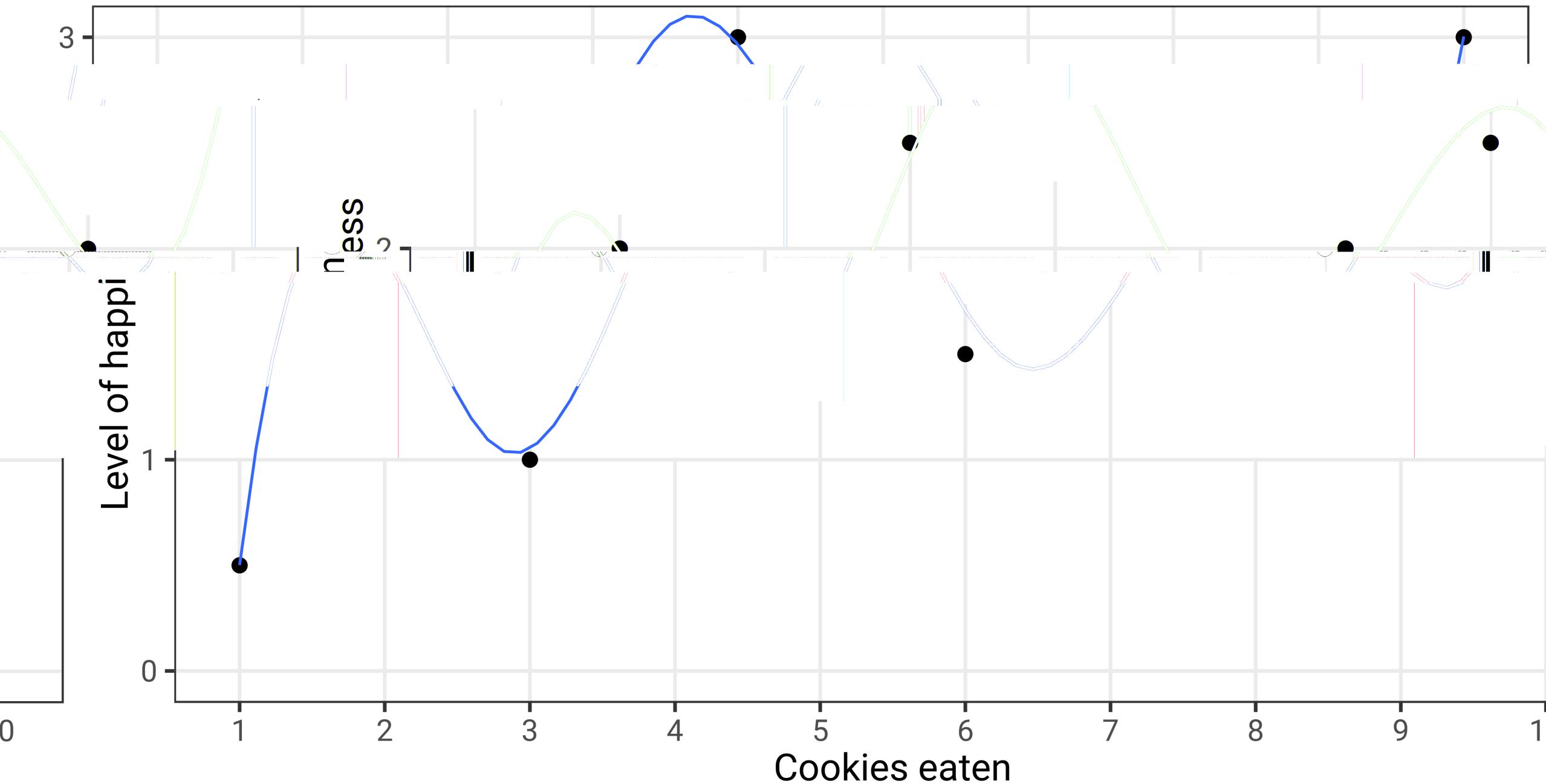
COOKIE CONSUMPTION AND HAPPINESS



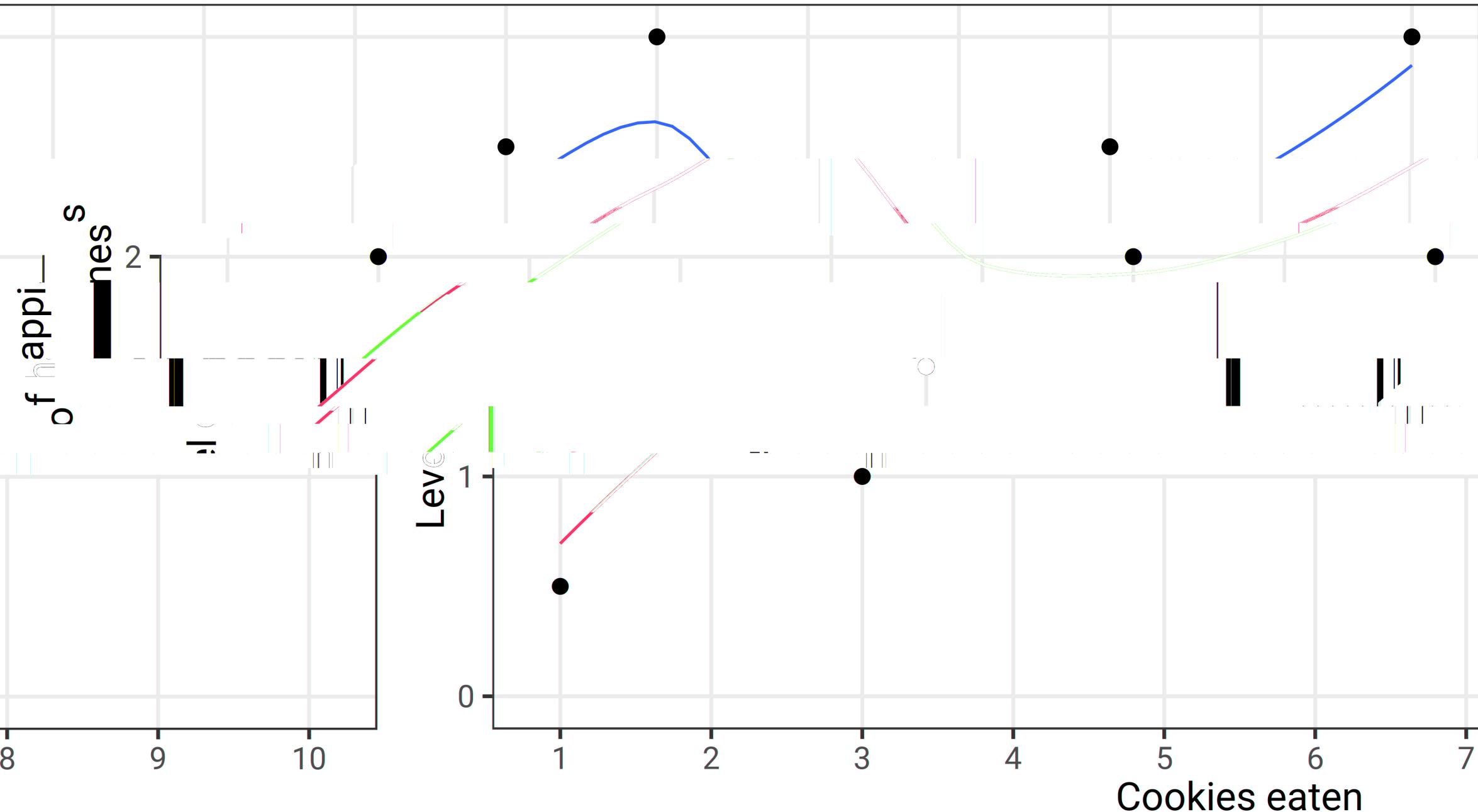
Relationship between cookies and happiness

$r = 0.607$

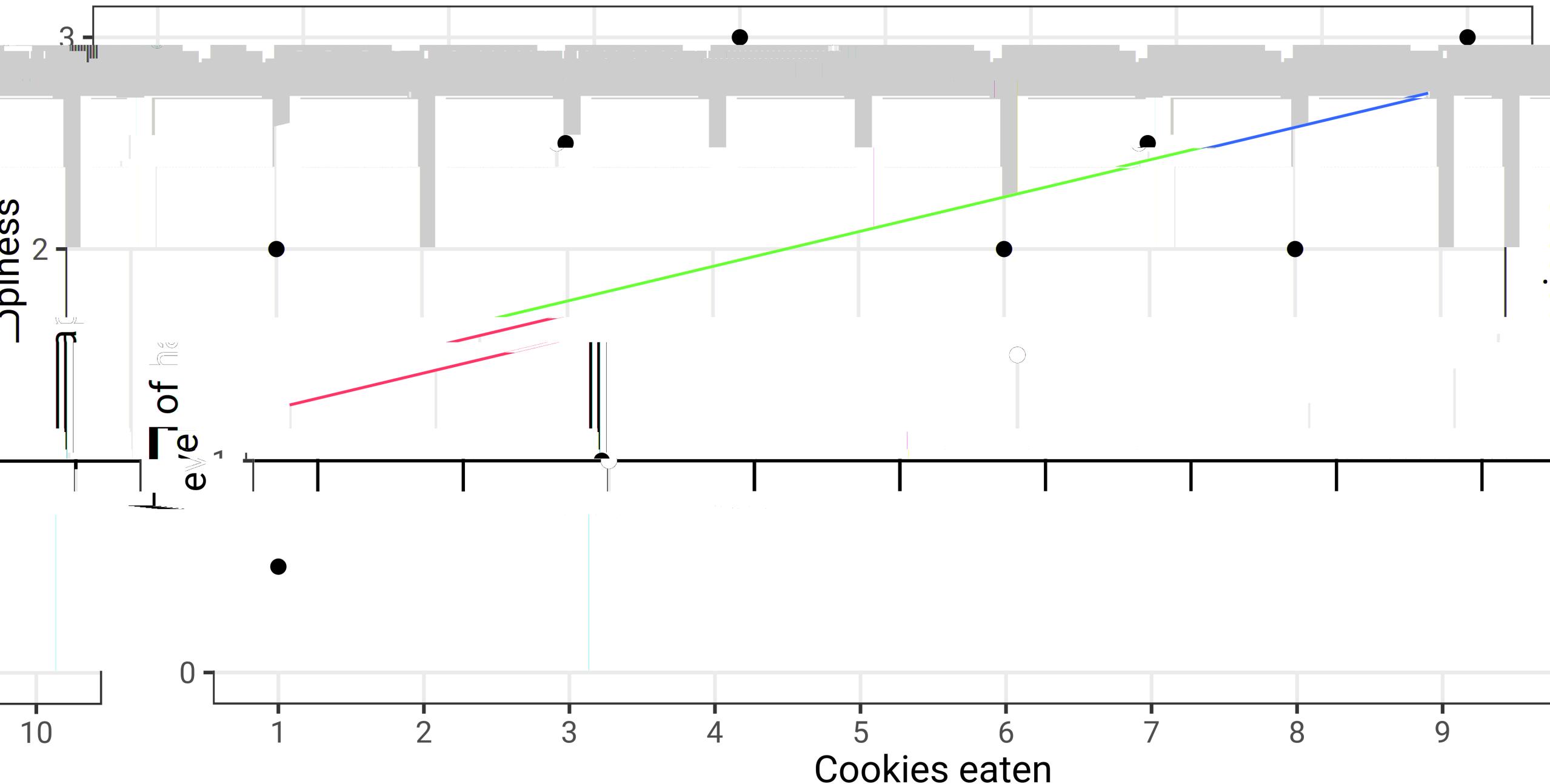
Relationship between cookies and happiness



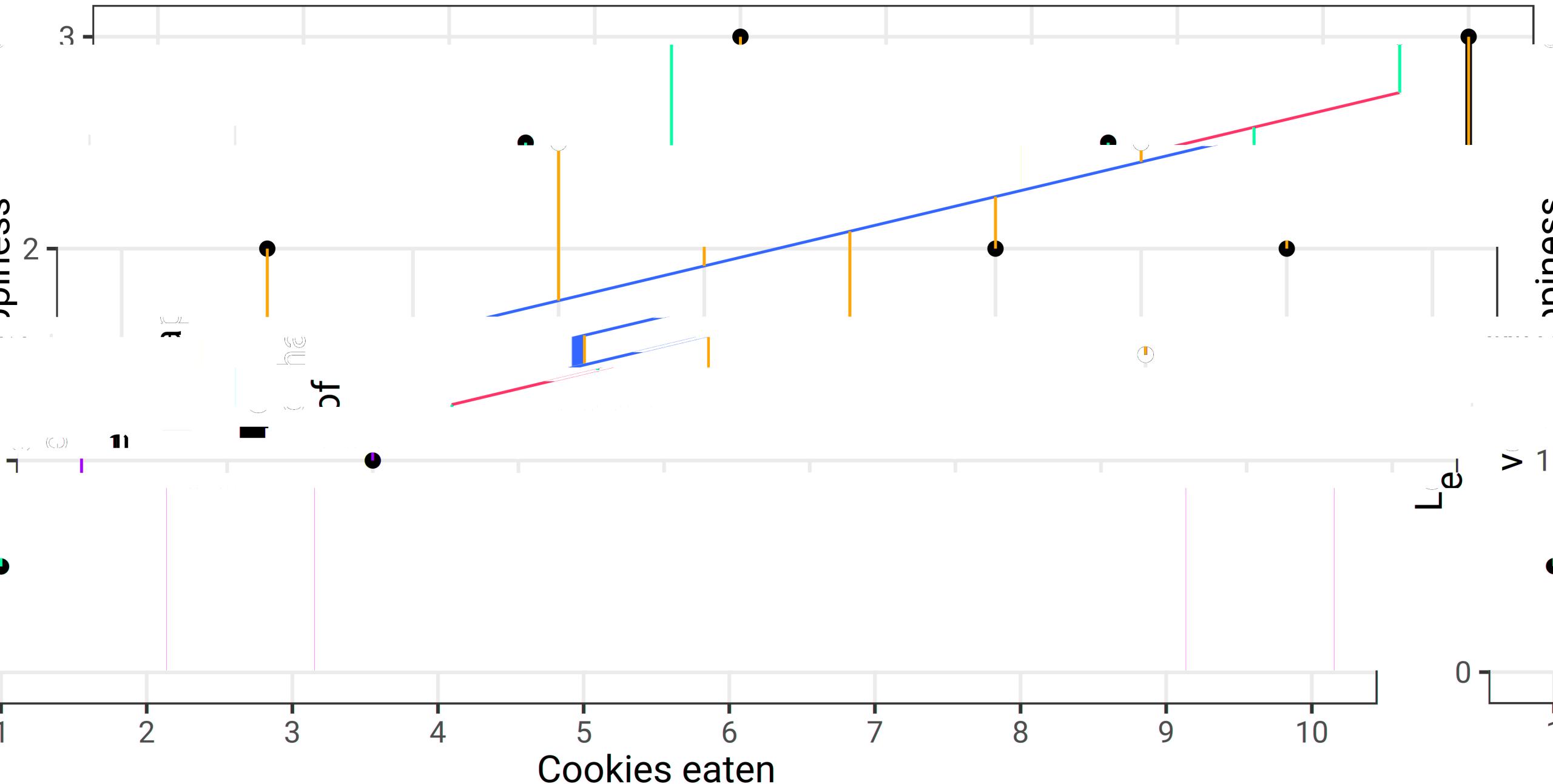
Relationship between cookies and happiness



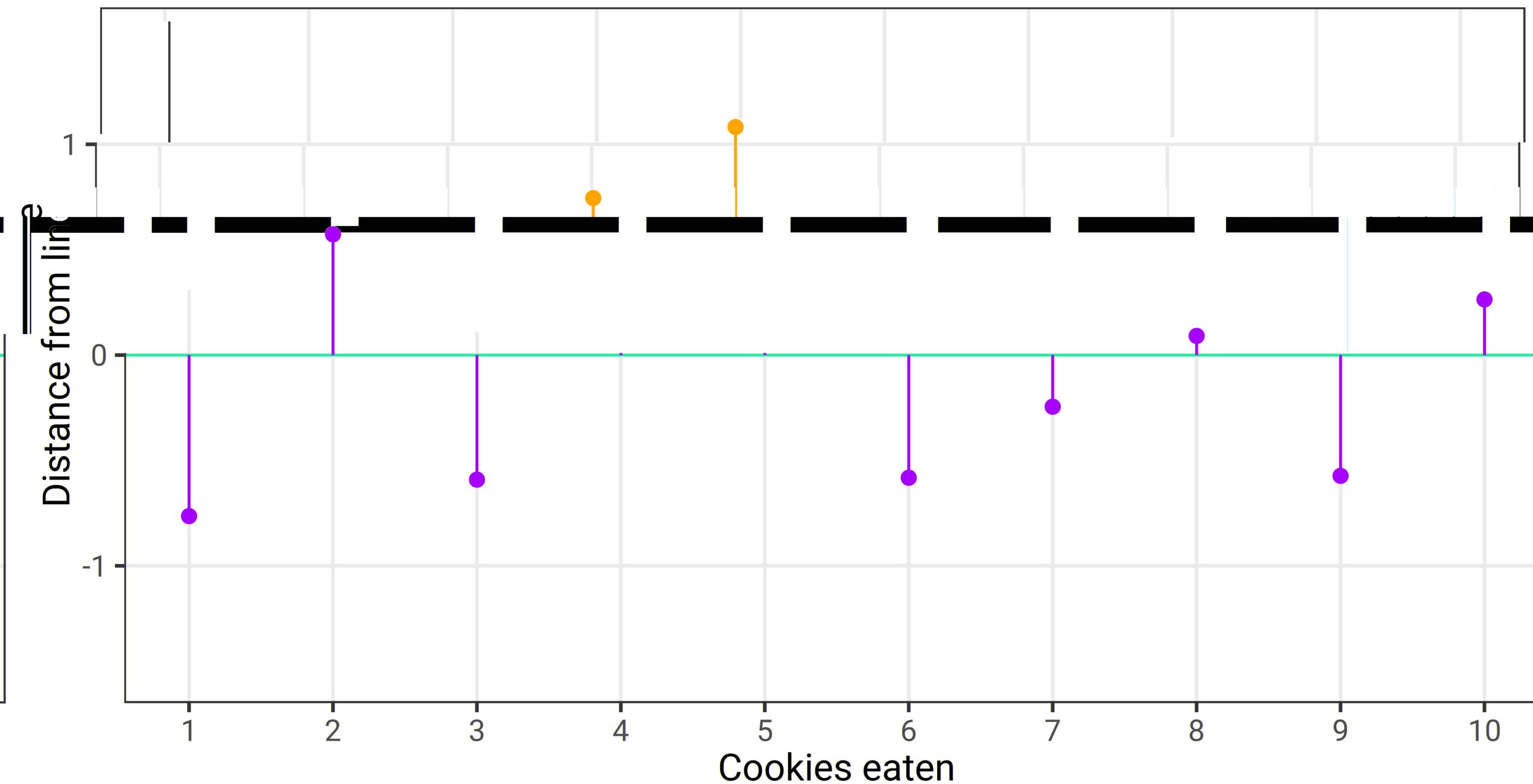
Relationship between cookies and happiness



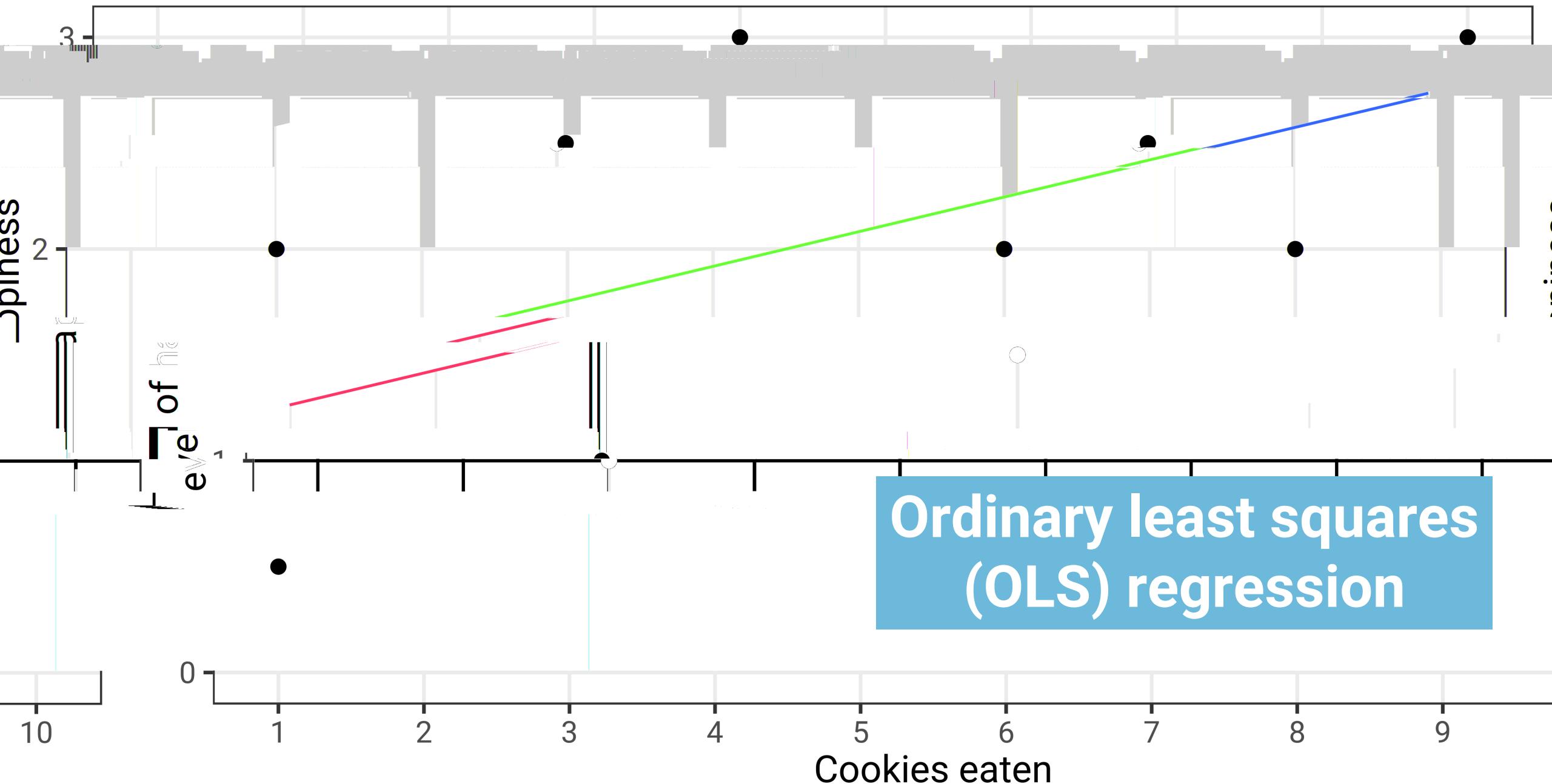
Relationship between cookies and happiness



Residual errors (distance from line)



Relationship between cookies and happiness



LINES AND MATH

DRAWING LINES WITH MATH

$$y = mx + b$$

y

A number

x

A number

m

Slope

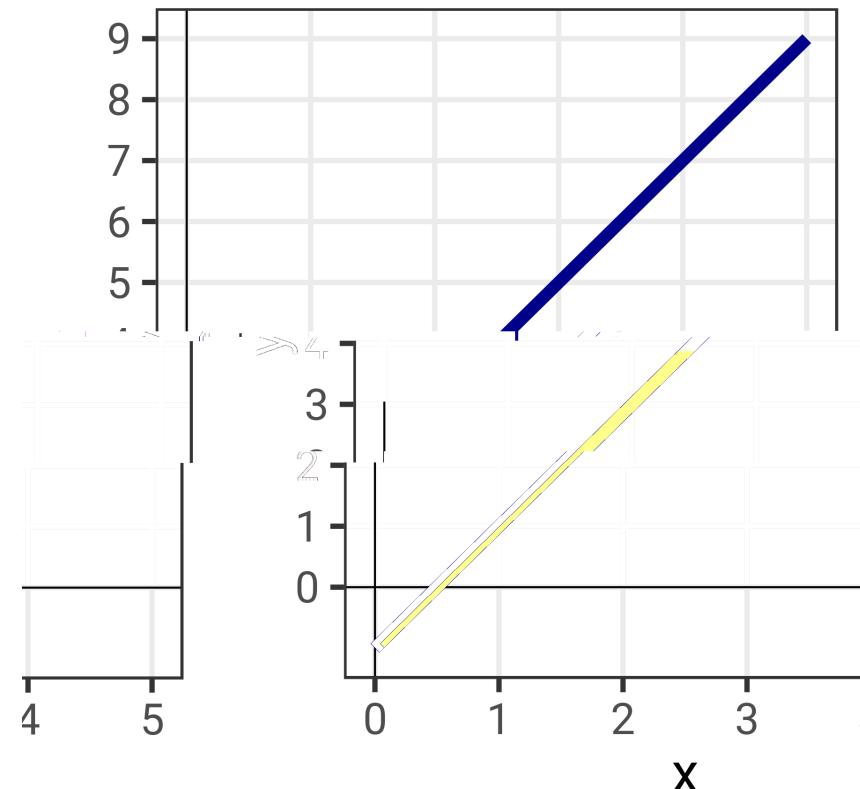
$\frac{\text{rise}}{\text{run}}$

b

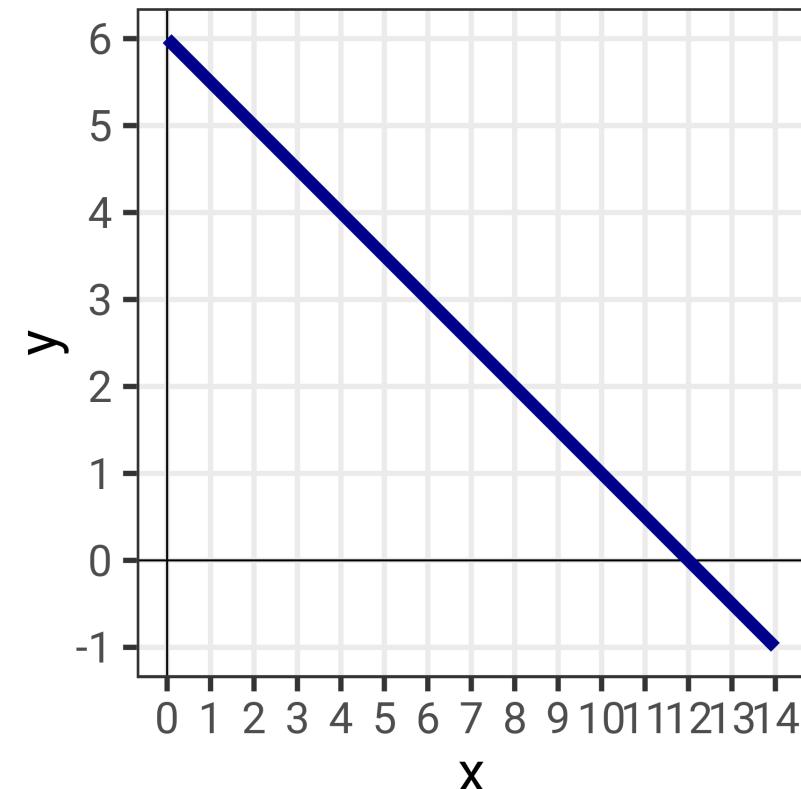
y intercept

SLOPES AND INTERCEPTS

$$y = 2x - 1$$



$$y = -0.5x + 6$$



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

$$y = mx + b$$

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

y

\hat{y}

Outcome variable

x

x_1

Explanatory variable

m

β_1

Slope

b

β_0

y intercept

ϵ

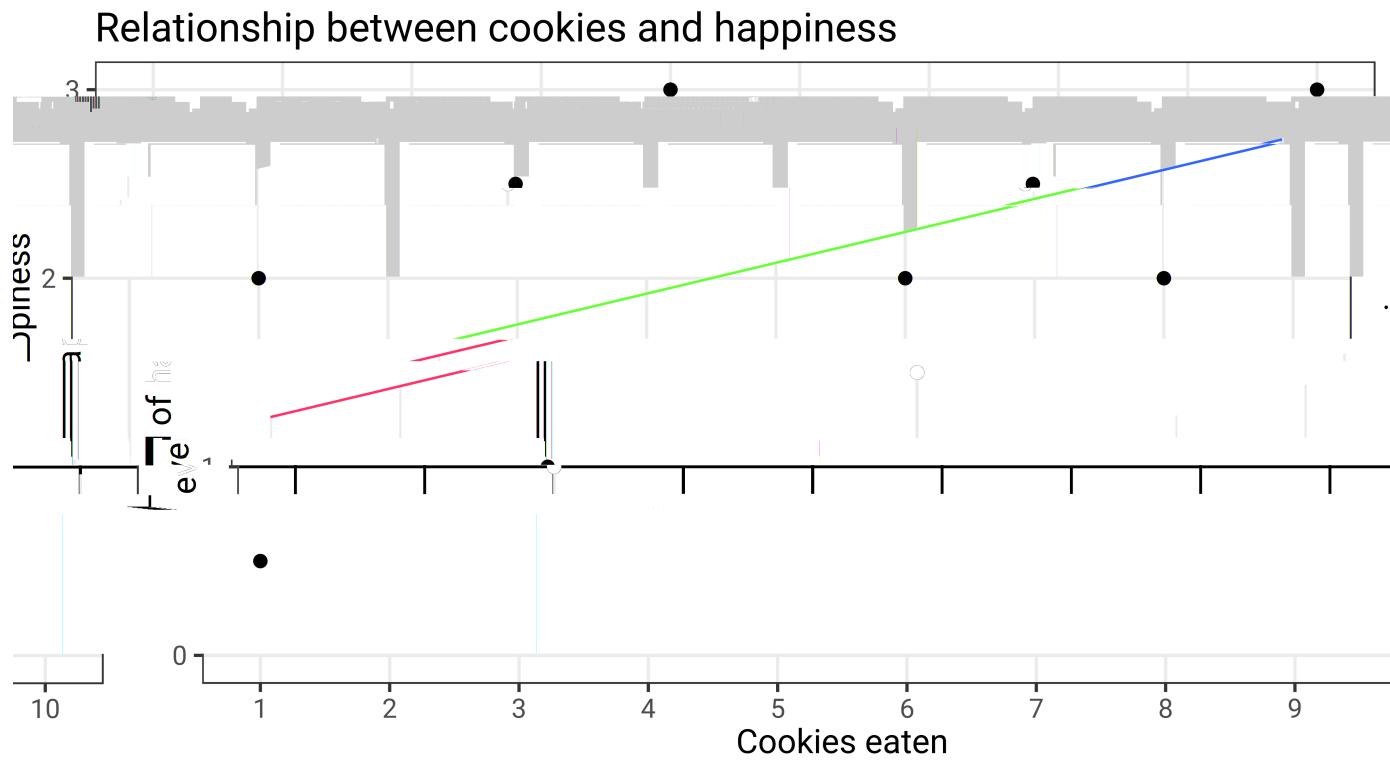
Error (residuals)

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

happiness
= $\beta_0 + \beta_1$ cookies + ε

A

A



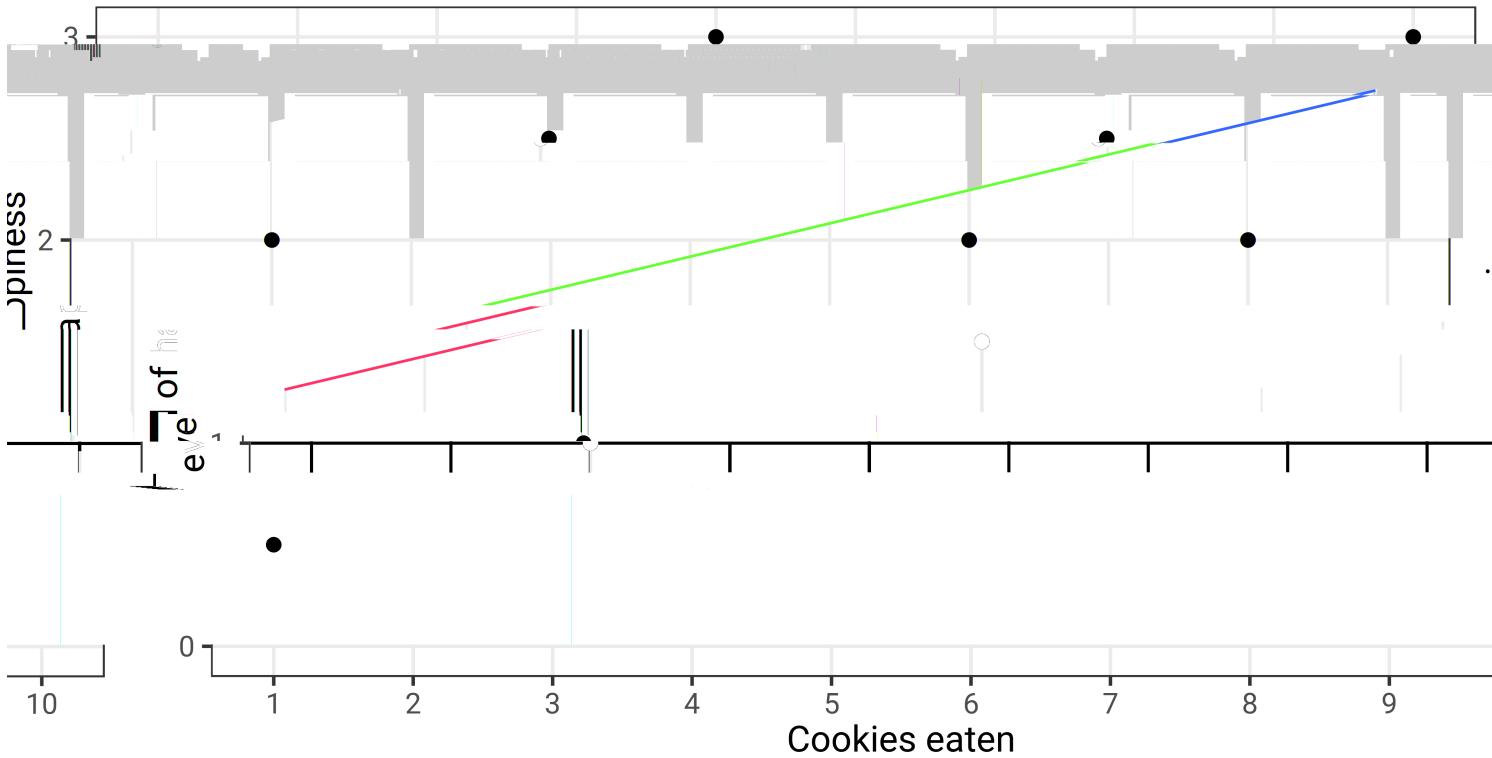
MODELING COOKIES AND HAPPINESS



```
# 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
```

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

Relationship between cookies and happiness



term	estimate	std_error	statistic	p_value	lower	upper
intercept	1.1	0.0229	2.29	0.0263	0.91	2.24
cookies	0.164	0.0159	1.059	0.2963	0.11	0.338

Chapter 10 / 16

Chapter 11 / 159

Chapter 11 / 163

Chapter 9 / 11

Chapter 9 / 338

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