Regression and inference

Session 2

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

Drawing lines

Lines, Grεεκ, and regression

Null worlds and statistical significance

Drawing lines

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

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 AP classes in high school improves 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 enroll in SNAP

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 SNAP on poverty reduction

How?

Plot X and Y

Draw a line that approximates the relationship

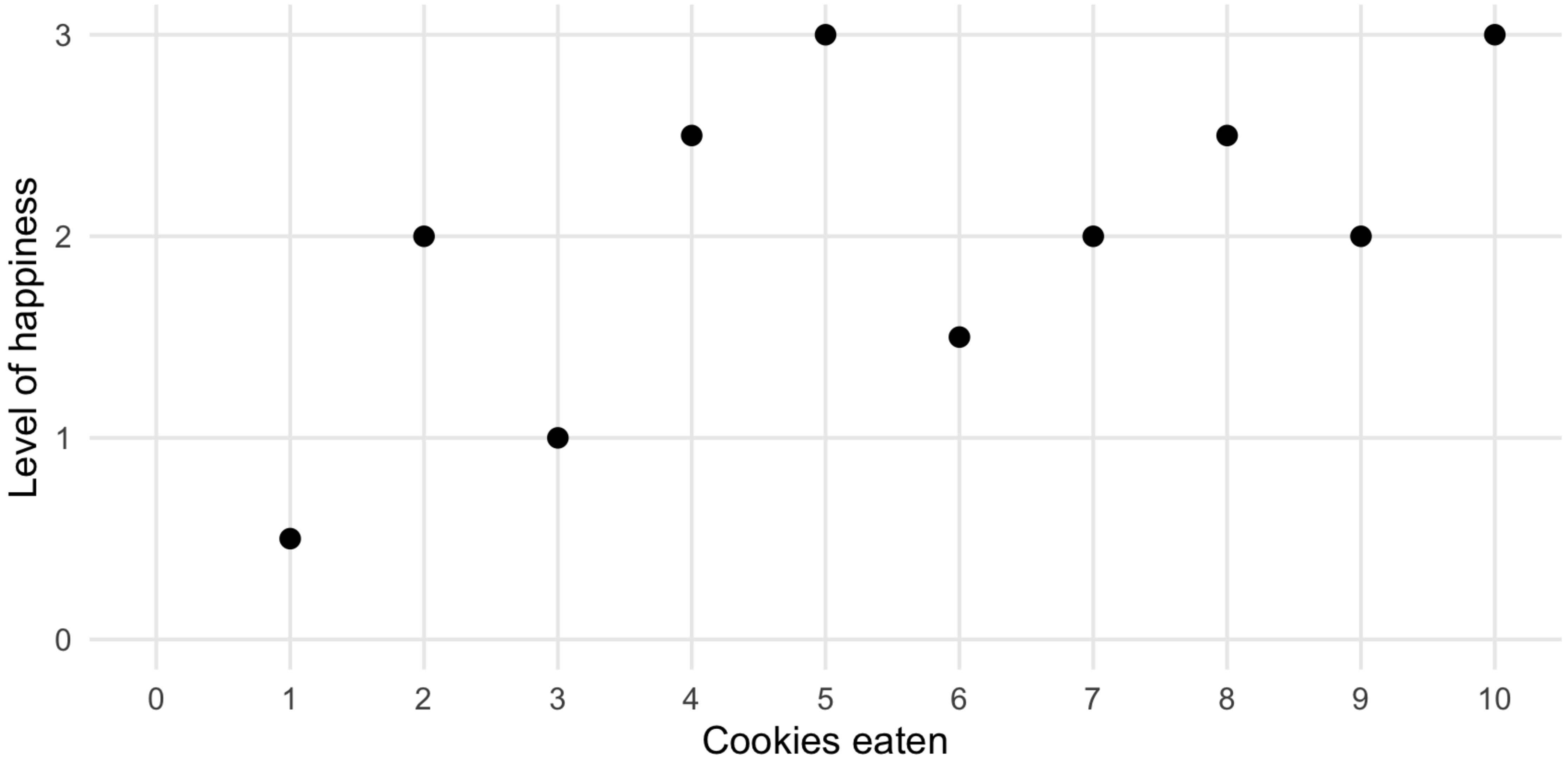
and that would plausibly work for data not in the sample!

Find mathy parts of the line

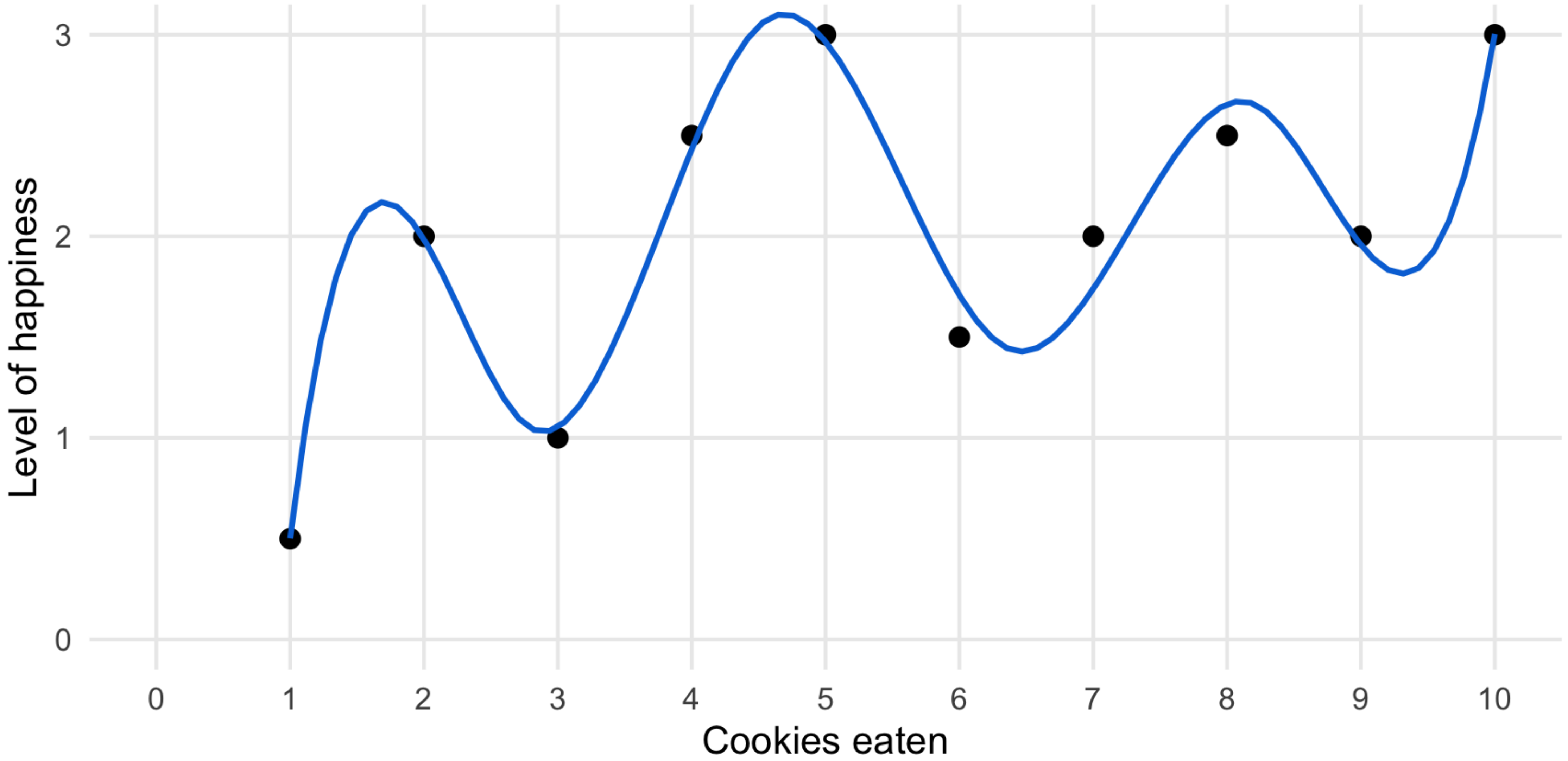
Interpret the math

Cookies and happiness

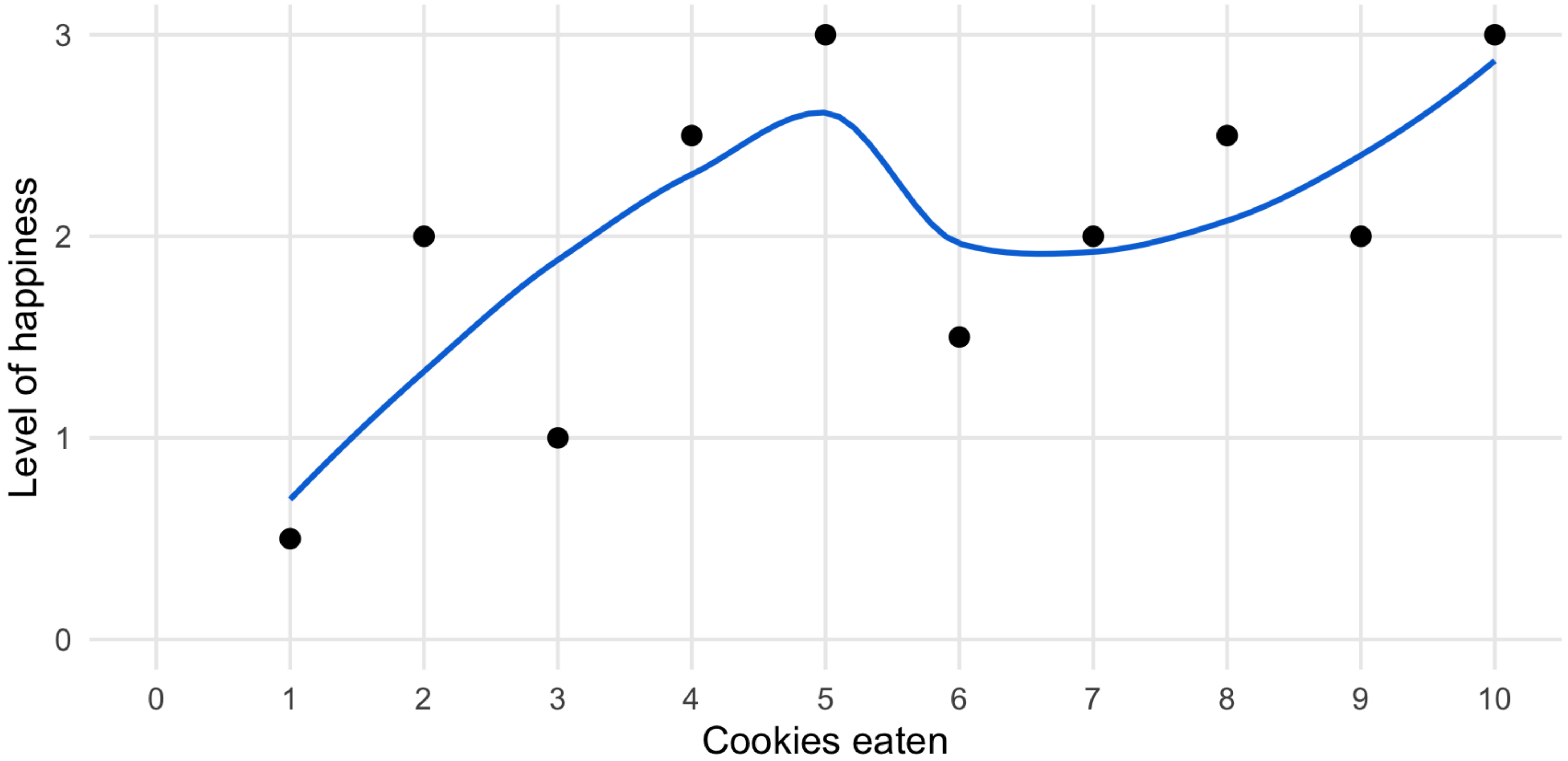
```
## # A tibble: 10 × 2
##    happiness cookies
##        <dbl>   <int>
##  1       0.5       1
##  2       2         2
##  3       1         3
##  4       2.5       4
##  5       3         5
##  6       1.5       6
##  7       2         7
##  8       2.5       8
##  9       2         9
## 10       3        10
```



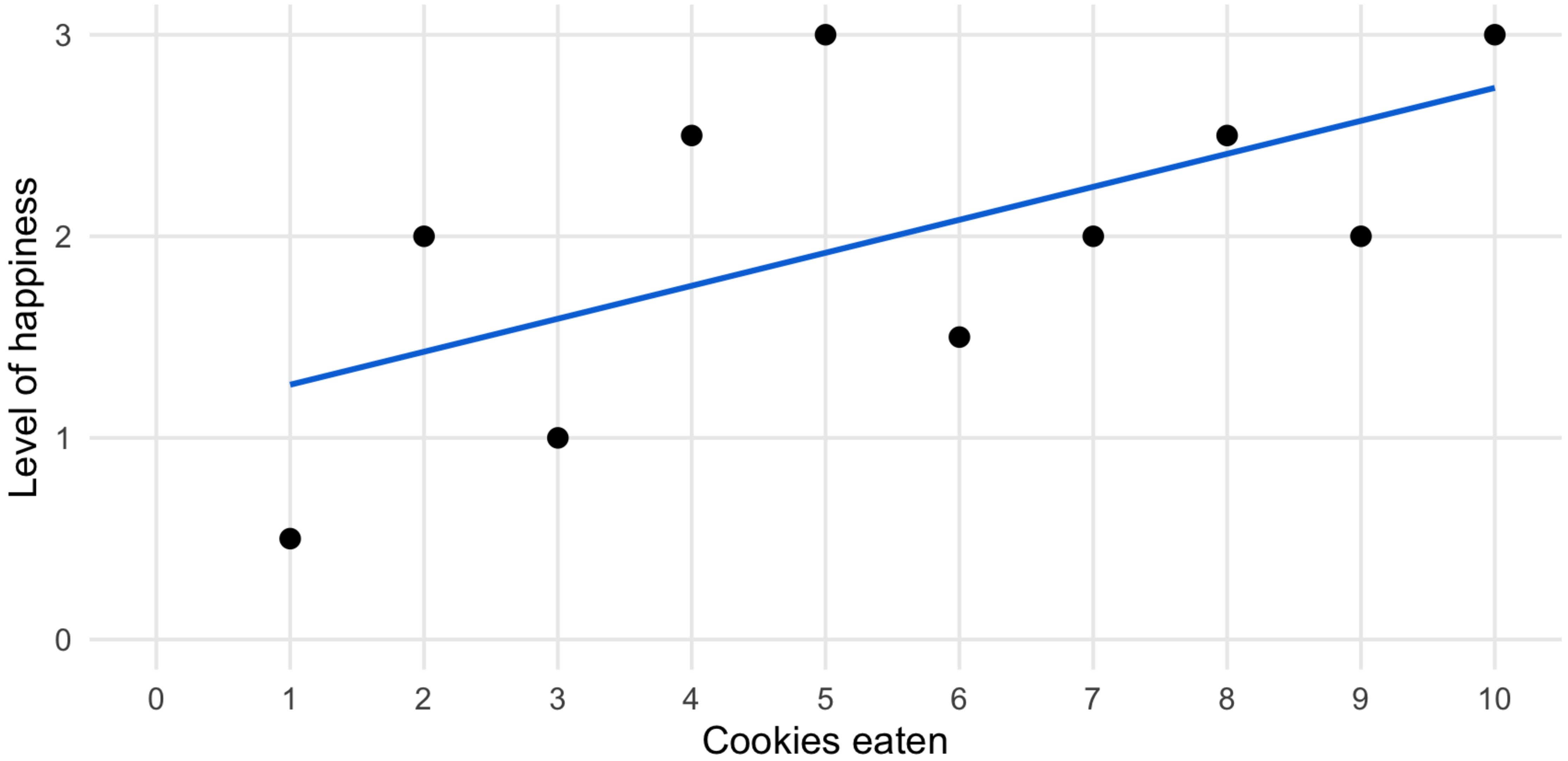
Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten

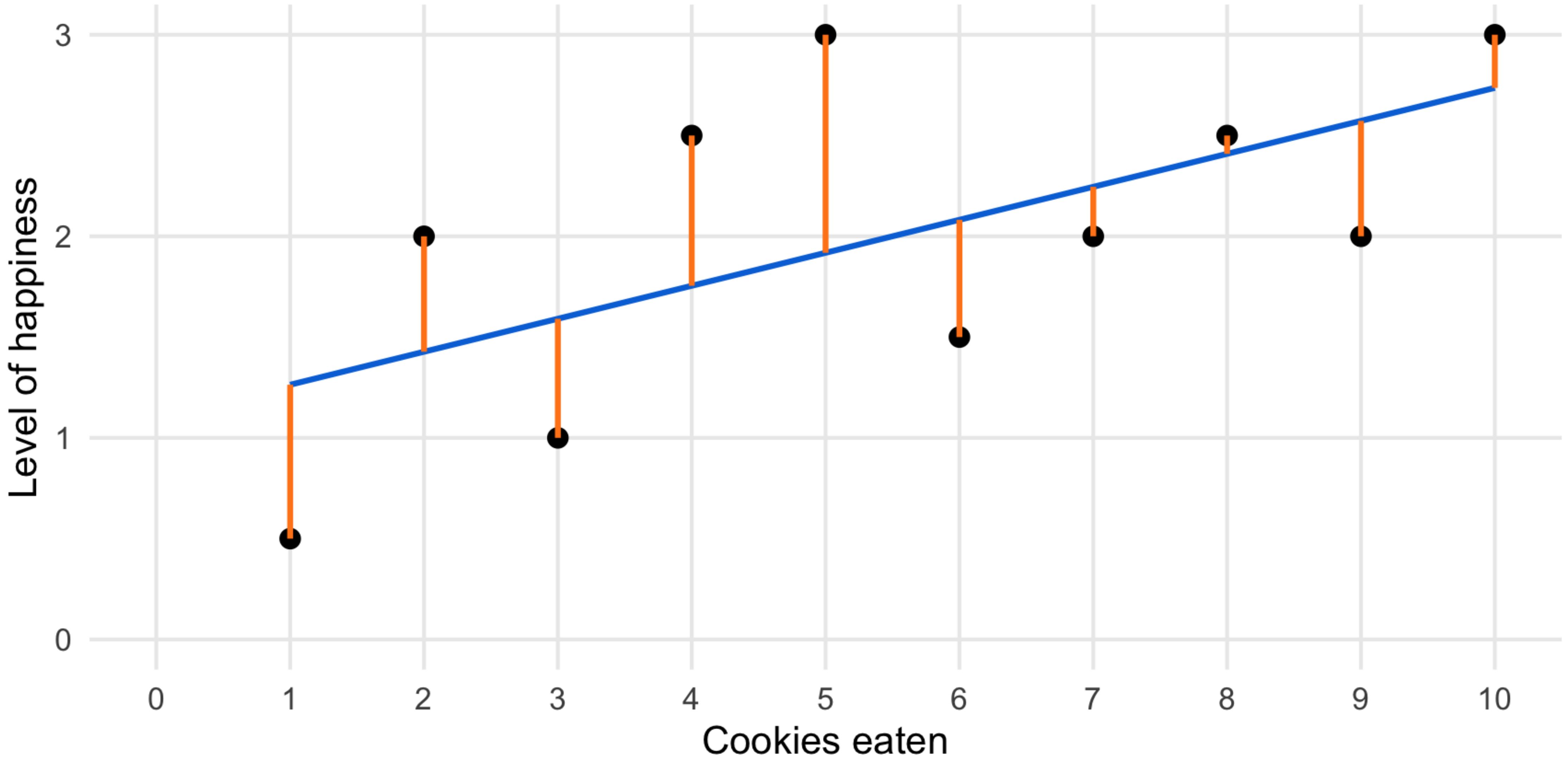


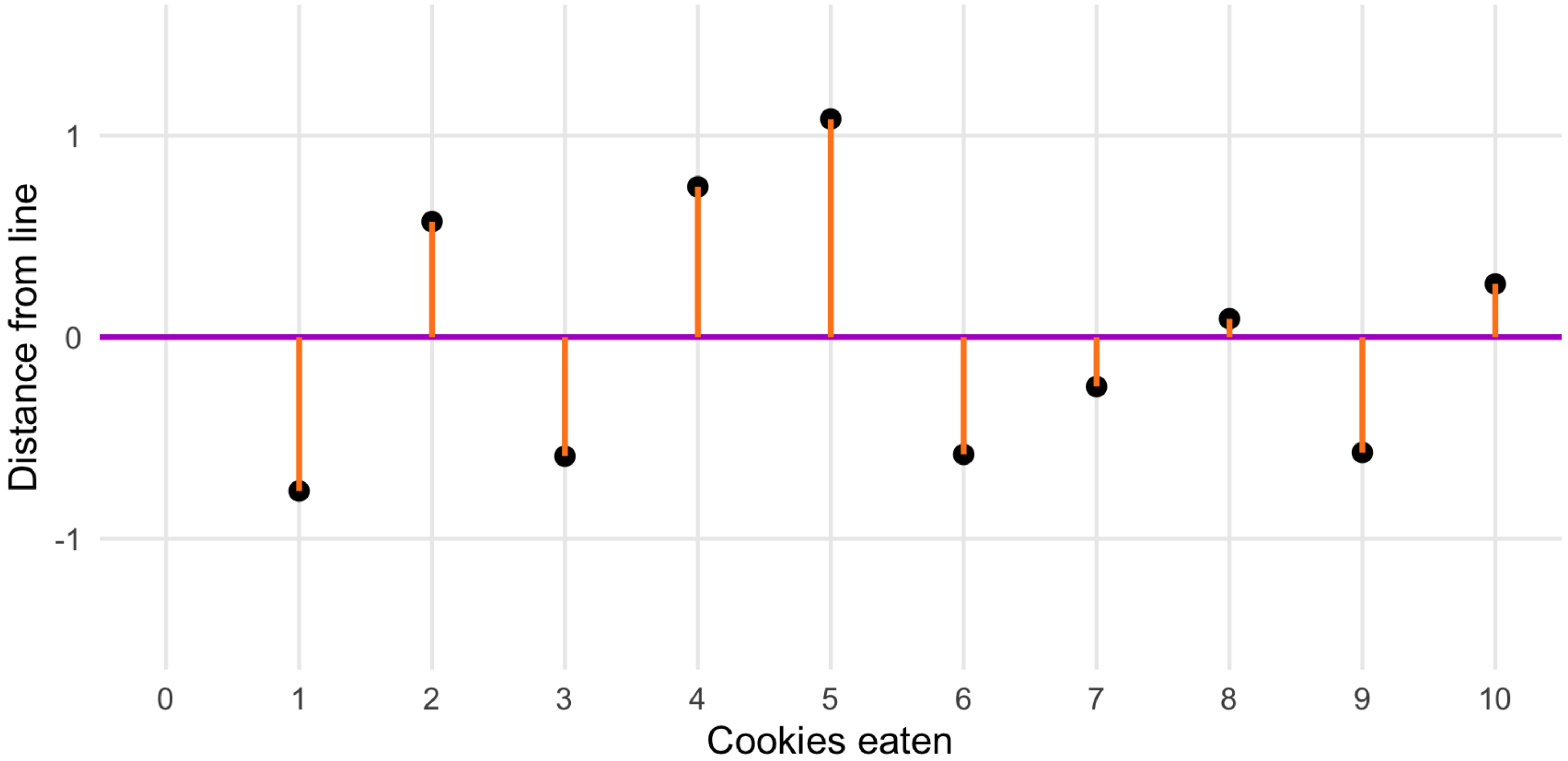
Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten

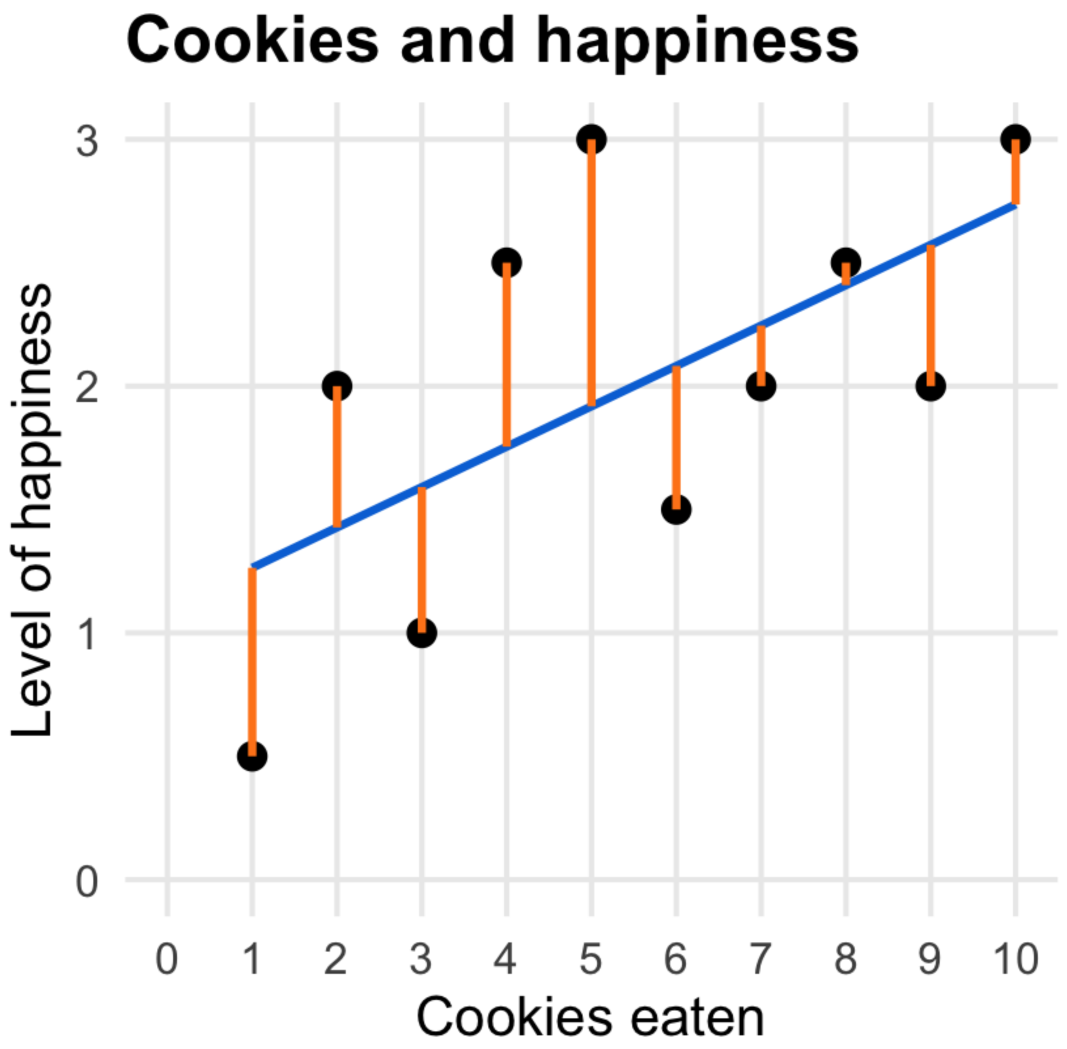


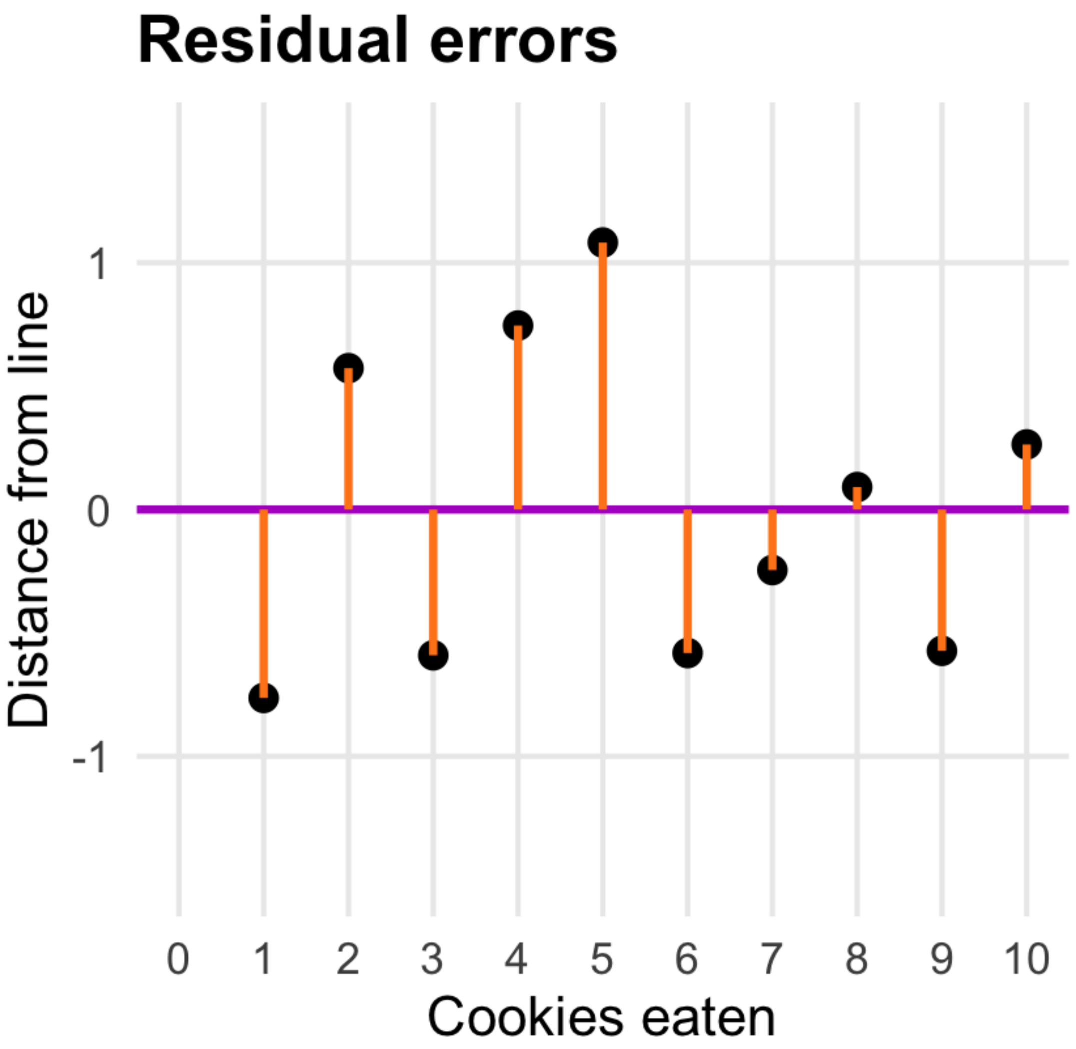
Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


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

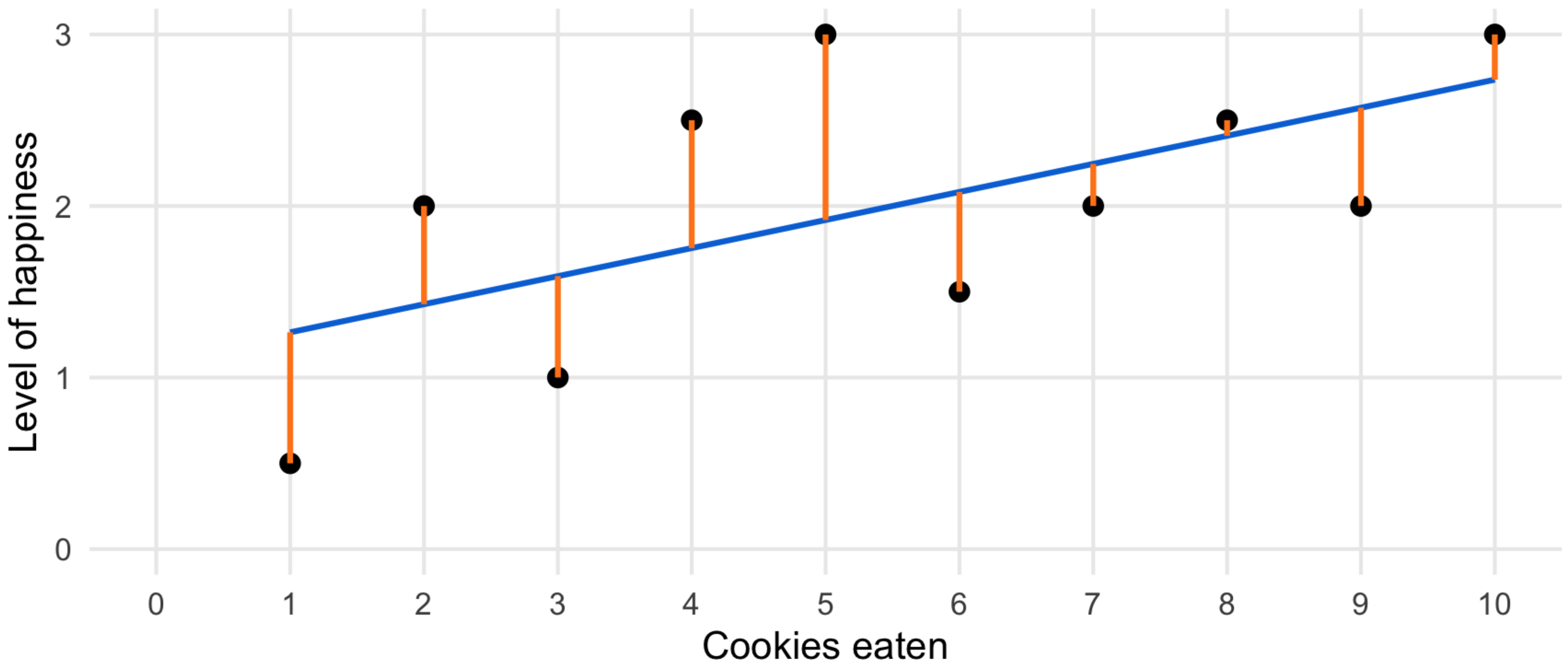

Level of happiness
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


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


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

Ordinary least squares (OLS) regression



Lines, Greek, and regression

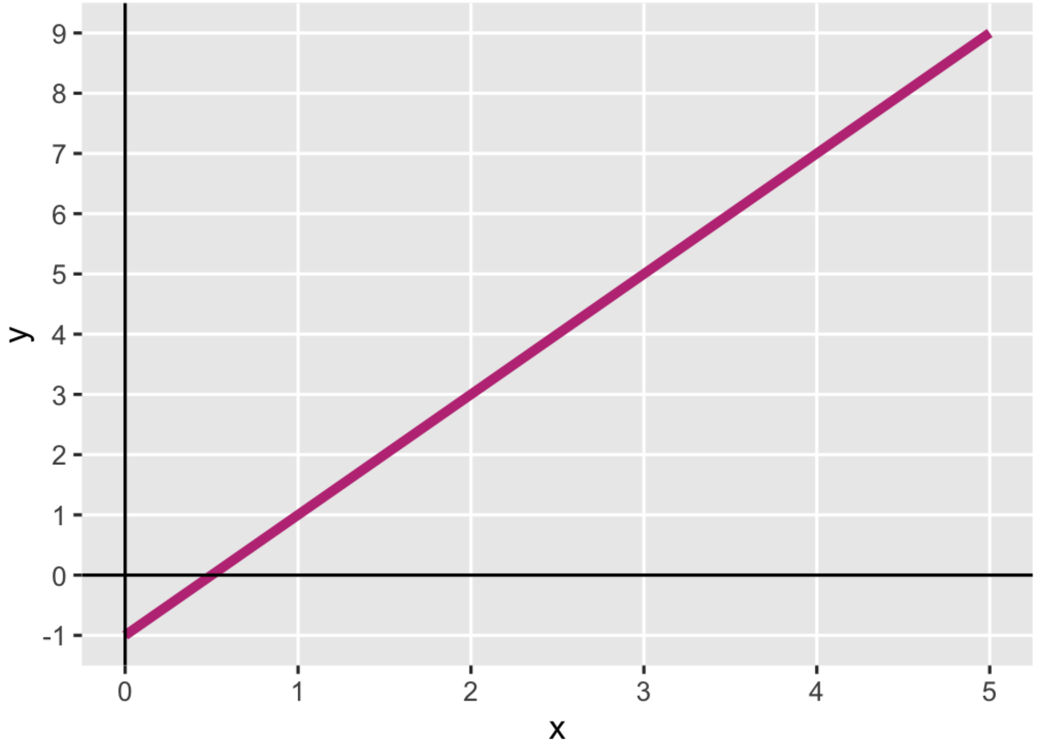
Drawing lines with math

$$y = mx + b$$

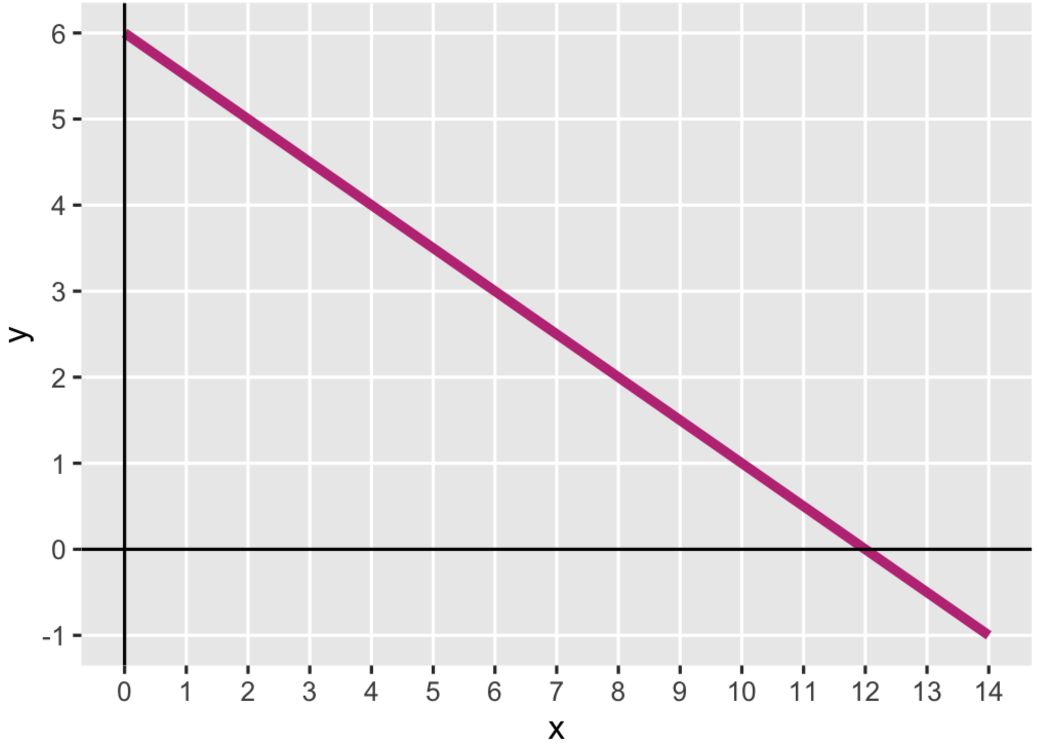
y	A number
x	A number
m	Slope ($\frac{\text{rise}}{\text{run}}$)
b	y-intercept

Slopes and intercepts

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Greek, Latin, and extra markings

Statistics: use a sample to make inferences about a population

Greek

Letters like β_1 are the ***truth***

Letters with extra markings like $\hat{\beta}_1$ are our ***estimate*** of the truth based on our sample

Latin

Letters like X are ***actual data*** from our sample

Letters with extra markings like $\bar{X}$ are ***calculations*** from our sample

Estimating truth

Data → Calculation → Estimate → Truth

Data	X
Calculation	$\bar{X} = \frac{\sum X}{N}$
Estimate	$\hat{\mu}$
Truth	μ

$$\bar{X} = \hat{\mu}$$

$$X \rightarrow \bar{X} \rightarrow \hat{\mu} \xrightarrow{\text{🤞 hopefully 🤞}} \mu$$

Drawing lines with stats

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \varepsilon$$

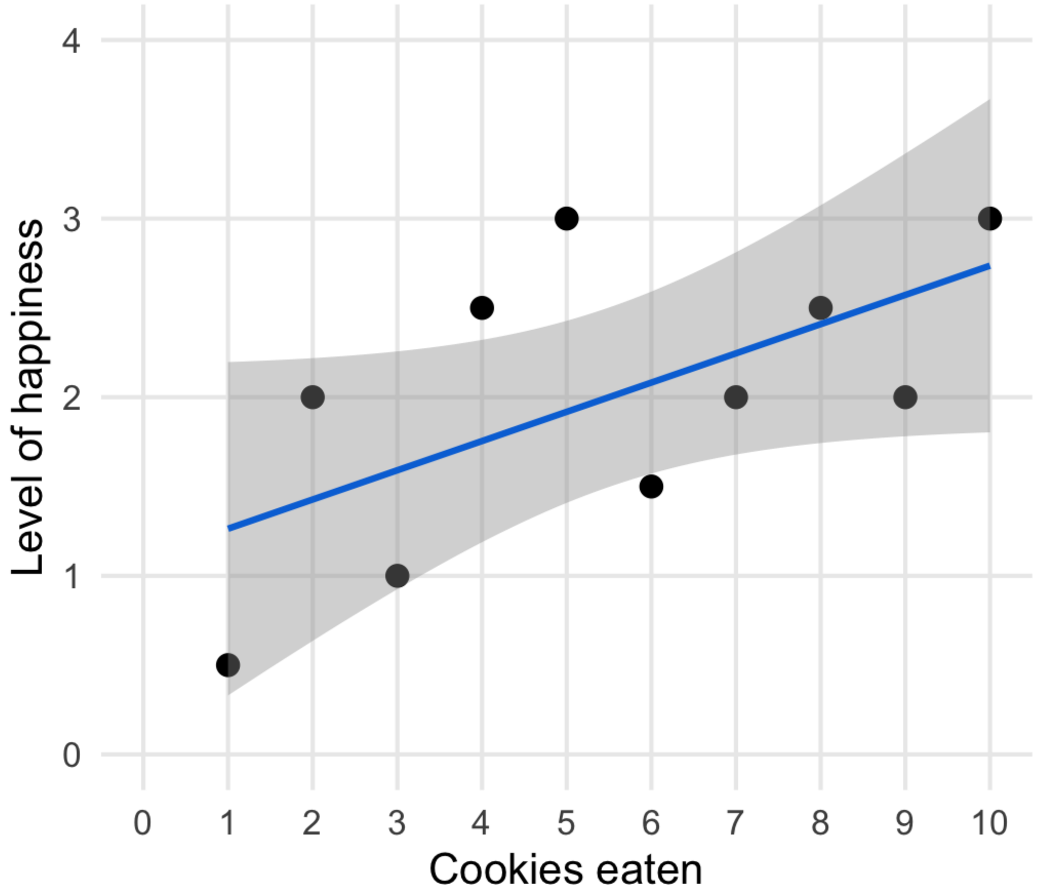
y	$\hat{y}$	Outcome variable (DV)
x	x_1	Explanatory variable (IV)
m	$\hat{\beta}_1$	Slope
b	$\hat{\beta}_0$	y-intercept
	ε	Error (residuals)

(most of the time we can get rid of markings on Greek and just use β)

Modeling cookies and happiness

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

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



Building models in R

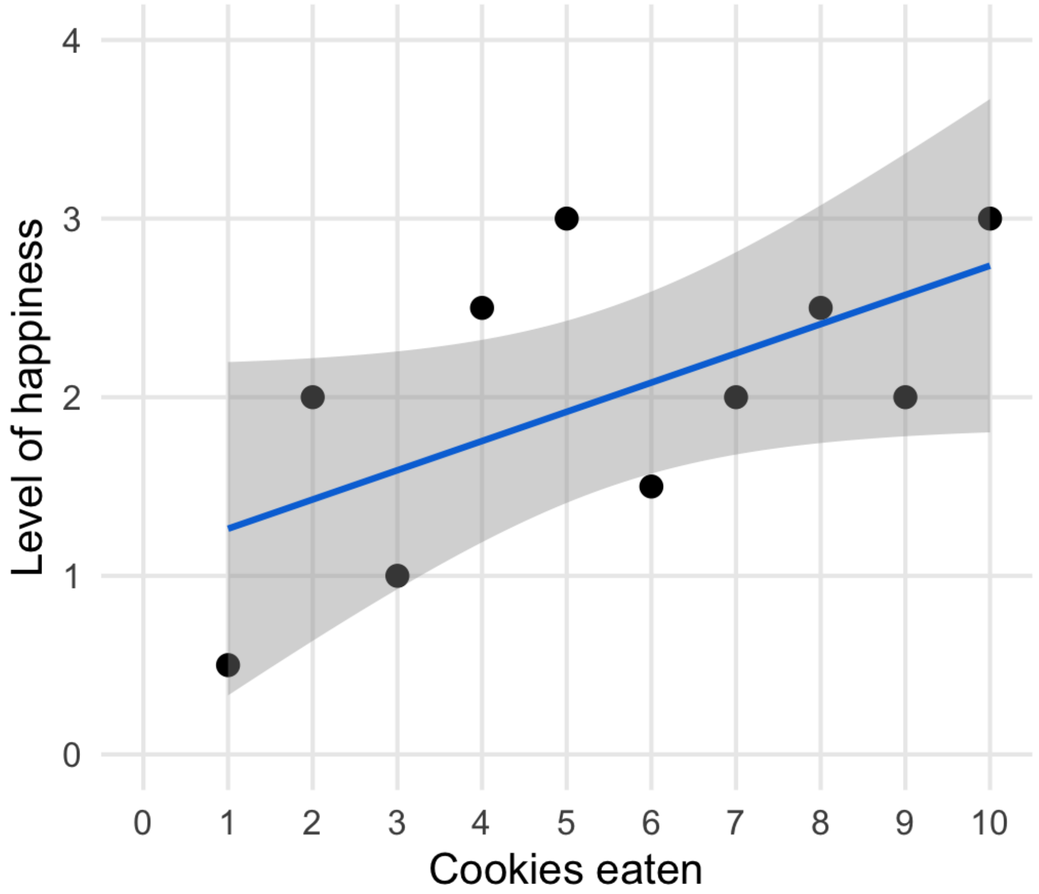
```r
name_of_model <- lm(<Y> ~ <X>, data = <DATA>)

summary(name_of_model)  # See model details
```

```r
library(broom)

# Convert model results to a data frame for plotting
tidy(name_of_model)

# Convert model diagnostics to a data frame
glance(name_of_model)
```

Modeling cookies and happiness

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

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



Modeling cookies and happiness

```
tidy(happiness_model, conf.int = TRUE)
```

```
## # A tibble: 2 × 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)    1.1      0.470       2.34  0.0475   0.0156     2.18
## 2 cookies        0.164    0.0758      2.16  0.0629  -0.0111     0.338
```

```
glance(happiness_model)
```

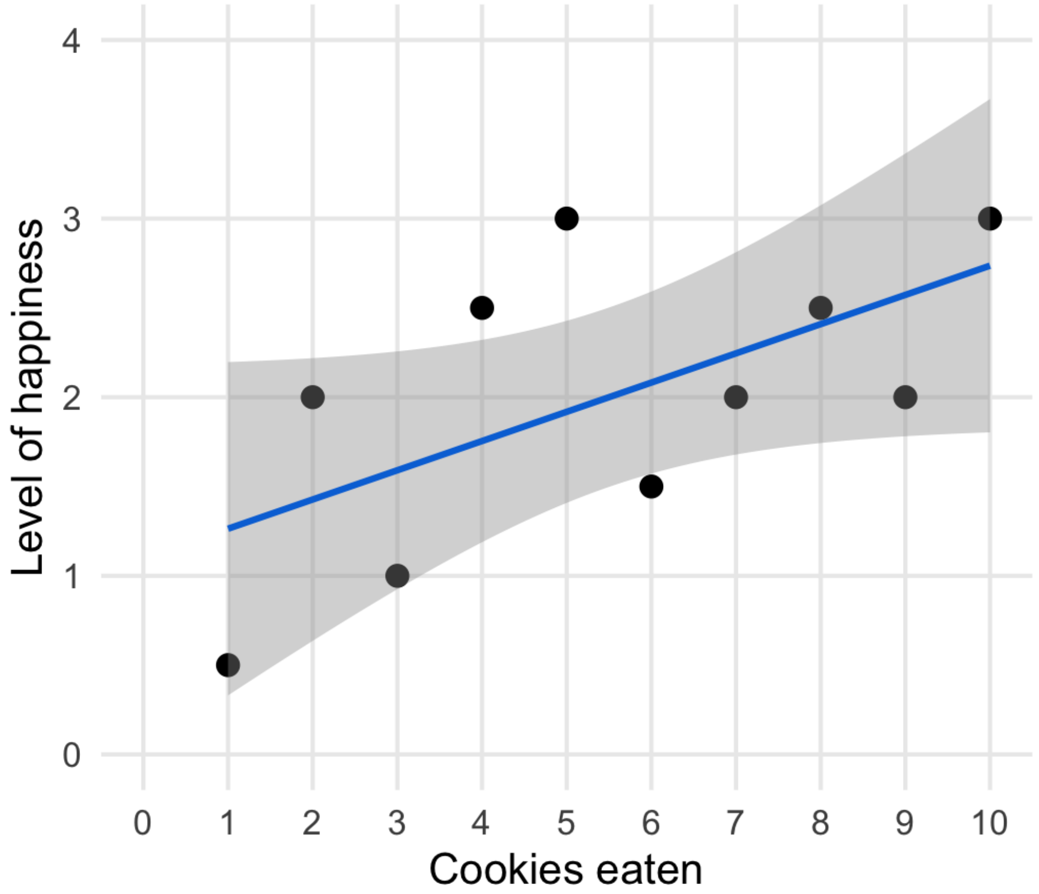
```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.368         0.289 0.688      4.66  0.0629     1  -9.34  24.7  25.6
## # ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Translating results to math

```
## # A tibble: 2 × 2
##   term        estimate
##   <chr>          <dbl>
## 1 (Intercept)    1.1
## 2 cookies        0.164
```

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

$$\widehat{\text{happiness}} = 1.1 + 0.16 \times \text{cookies} + \varepsilon$$



Template for single variables

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

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

$$\widehat{\text{happiness}} = 1.1 + 0.16 \times \text{cookies} + \varepsilon$$

Multiple regression

We're not limited to just one explanatory variable!

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

```
car_model <- lm(hwy ~ displ + cyl + drv,
                data = mpg)
```

$$\widehat{\text{hwy}} = \beta_0 + \beta_1 \text{displ} + \beta_2 \text{cyl} + \beta_3 \text{drv:f} + \beta_4 \text{drv:r} + \varepsilon$$

Modeling lots of things and MPG

```
tidy(car_model, conf.int = TRUE)
```

```
## # A tibble: 5 × 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    33.1      1.03       32.1  9.49e-87    31.1     35.1
## 2 displl         -1.12     0.461     -2.44 1.56e- 2    -2.03     -0.215
## 3 cyl            -1.45     0.333     -4.36 1.99e- 5    -2.11     -0.796
## 4 drvf            5.04     0.513      9.83 3.07e-19     4.03      6.06
## 5 drvr            4.89     0.712      6.86 6.20e-11     3.48      6.29
```

$$\widehat{\text{hwy}} = 33.1 + (-1.12) \times \text{displ} + (-1.45) \times \text{cyl} + (5.04) \times \text{drv:f} + (4.89) \times \text{drv:r} + \varepsilon$$

Sliders and switches

Sliders and switches

Categorical variables

Continuous variables

Filtering out variation

Each X in the model explains some portion of the variation in Y

Interpretation is a little trickier, since you can only ever move one switch or slider at at time

Template for continuous variables

Holding everything else constant, a one unit increase in **X** is ***associated*** with a β_n increase (or decrease) in **Y**, on average

$$\widehat{\text{hwy}} = 33.1 + (-1.12) \times \text{displ} + (-1.45) \times \text{cyl} + (5.04) \times \text{drv:f} + (4.89) \times \text{drv:r} + \varepsilon$$

On average, a one unit increase in cylinders is associated with 1.45 lower highway MPG, holding everything else constant

Template for categorical variables

Holding everything else constant, **Y** is β_n units larger (or smaller) in X_n, compared to $X_{omitted}$, on average

$$\widehat{\text{hwy}} = 33.1 + (-1.12) \times \text{displ} + (-1.45) \times \text{cyl} + (5.04) \times \text{drv:f} + (4.89) \times \text{drv:r} + \varepsilon$$

On average, front-wheel drive cars have 5.04 higher highway MPG than 4-wheel-drive cars, holding everything else constant

Economists and Greek letters

Economists like to assign all sorts of Greek letters to their different coefficients

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

Equation 2.1 on p. 57 in *Mastering 'Metrics*

i = an individual

α ("alpha") = intercept

β ("beta") = coefficient just for *treatment*, or the causal effect

γ ("gamma") = coefficient for the *identifying variable* (being in Group A or not)

Economists and Greek letters

$$\ln Y_i = \alpha + \beta P_i + \gamma A_i + \delta_1 \text{SAT}_i + \delta_2 \text{PI}_i + e_i$$

Equation 2.2 on p. 61 in *Mastering 'Metrics*

i = an individual

α ("alpha") = intercept

β ("beta") = coefficient just for *treatment*, or the causal effect

γ ("gamma") = coefficient for the *identifying variable* (being in Group A or not)

δ ("delta") = coefficient for *control variables*

These are all the same thing!

$$\ln Y_i = \alpha + \beta P_i + \gamma A_i + \delta_1 \text{SAT}_i + \delta_2 \text{PI}_i + e_i$$

$$\ln Y_i = \beta_0 + \beta_1 P_i + \beta_2 A_i + \beta_3 \text{SAT}_i + \beta_4 \text{PI}_i + e_i$$

```
lm(log(income) ~ private + group_a + sat + parental_income,
   data = income_data)
```

(I personally like the all-β version instead of using like the entire Greek alphabet, but you'll see both varieties in the real world)

Null worlds and statistical significance

"hopefully"

How do we know if our estimate is the truth?

$$X \rightarrow \bar{X} \rightarrow \hat{\mu} \xrightarrow{\text{🤞 hopefully 🤞}} \mu$$

Are action movies rated higher than comedies?

Data	IMDB ratings	D
Calculation	Average action rating - average comedy rating	$\bar{D} = \frac{\sum D_{\text{Action}}}{N} - \frac{\sum D_{\text{Comedy}}}{N}$
Estimate	$\bar{D}$ in a sample of movies	$\hat{\delta}$
Truth	Difference in rating for *all* movies	δ

```r
head(movie_data)
```

```
## # A tibble: 6 × 4
##   title                year rating genre
##   <chr>               <int>  <dbl> <fct>
## 1 Tarzan Finds a Son!  1939    6.4 Action
## 2 Silmido              2003    7.1 Action
## 3 Stagecoach           1939    8   Action
## 4 Diamondbacks         1998    1.9 Action
## 5 Chaos Factor, The    2000    4.5 Action
## 6 Secret Command       1944    7   Action
```

```r
movie_data |>
  group_by(genre) |>
  summarize(avg_rating = mean(rating))
```

```
## # A tibble: 2 × 2
##   genre  avg_rating
##   <fct>       <dbl>
## 1 Action       5.41
## 2 Comedy       5.84
```

$$\hat{\delta} = \bar{D} = 5.41 - 5.84 = 0.43$$

Is the true δ 0.43?

Null worlds

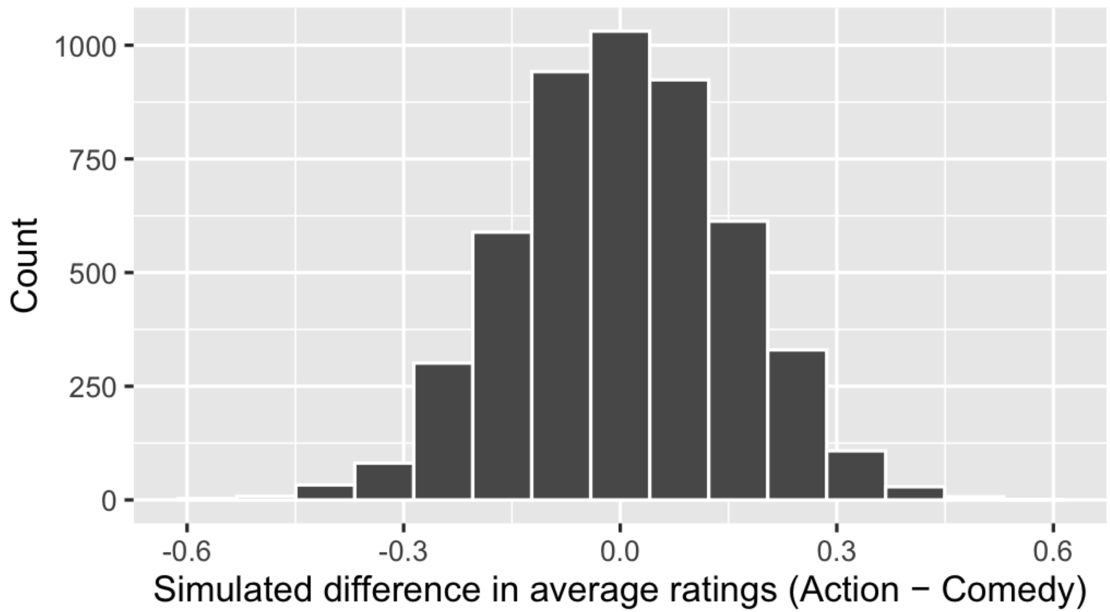
What would the world look like if the true δ was really 0?

Action movies and comedies wouldn't all have the same rating, but on average there'd be no difference

Simulated null world

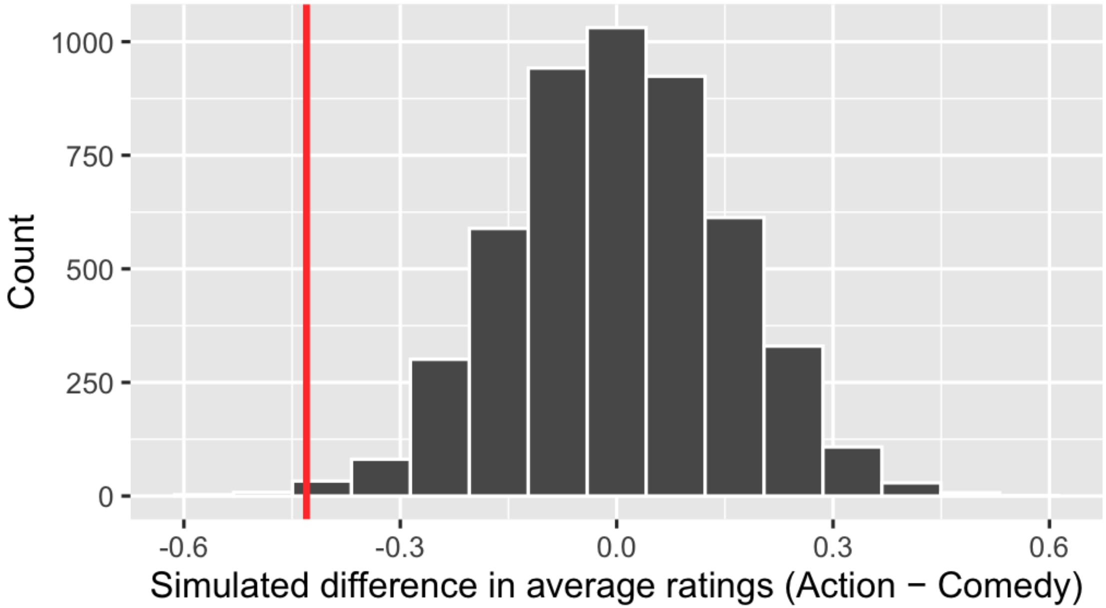
Shuffle the `rating` and `genre` columns and calculate the difference in ratings across genres

Do that ↑ 5,000 times



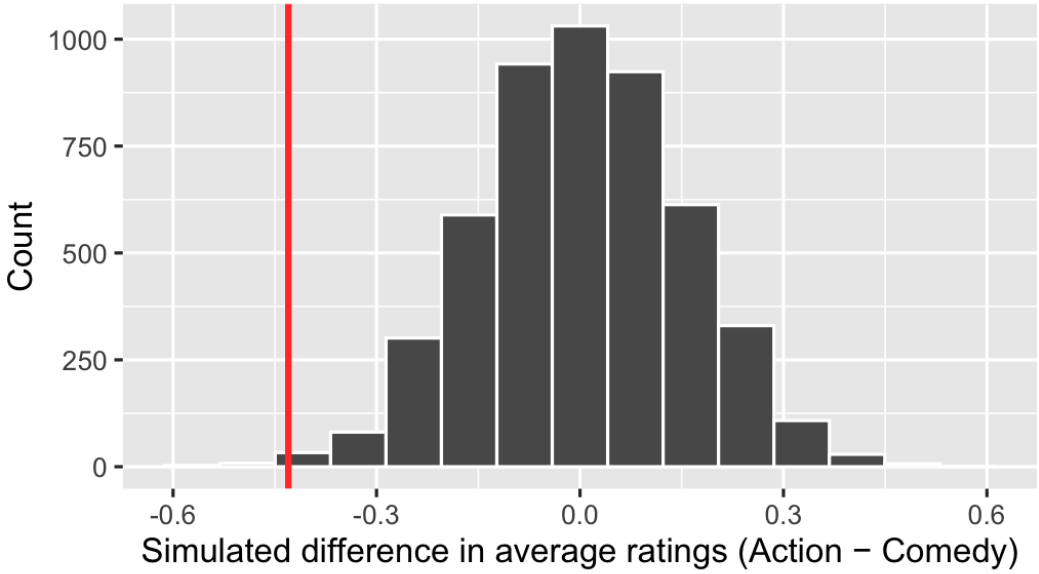
Check δ in the null world

Does the δ we observed fit well in the world where it's actually 0?



That seems fairly rare for a null world!

How likely is that δ in the null world?



What's the chance that we'd see that red line in a world where there's no difference?

p = 0.005

That's really really low!

p-values

That 0.005 is a p-value

p-value = probability of seeing something
in a world where the effect is 0

The δ we measured doesn't fit well
in the null world, so it's most likely not 0

We can safely say that there's a difference between the two groups. Action movies are rated lower, on average, than comedies

Significance

If $p < 0.05$, there's a good chance the estimate is not zero and is "real"

If $p > 0.05$, we can't say anything

That doesn't mean that there's no effect! It just means we can't tell if there is.

No need for all that simulation

This simulation stuff is helpful for the intuition behind a p-value, but you can also just interpret p-values in the wild

```
t.test(rating ~ genre, data = movie_data)
```

```
##
##      Welch Two Sample t-test
##
## data:  rating by genre
## t = -2.8992, df = 388.75, p-value = 0.003953
## alternative hypothesis: true difference in means between group Action and group Comedy is not e
## 95 percent confidence interval:
##  -0.7299913 -0.1400087
## sample estimates:
## mean in group Action mean in group Comedy
##                5.407                5.842
```

Slopes and coefficients

You can find a p-value for any Greek letter estimate, like β from a regression

$$\hat{\beta} \xrightarrow{\text{🤞 hopefully 🤞}} \beta$$

In a null world, the slope (β) would be zero

p-value shows us if β=hat would fit in a world where β is zero

Regression and p-values

```
tidy(car_model, conf.int = TRUE)
```

```
## # A tibble: 5 × 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)    33.1      1.03      32.1  9.49e-87    31.1     35.1
## 2 displ          -1.12     0.461     -2.44 1.56e- 2    -2.03    -0.215
## 3 cyl            -1.45     0.333     -4.36 1.99e- 5    -2.11    -0.796
## 4 drvf            5.04     0.513      9.83 3.07e-19     4.03     6.06
## 5 drvr            4.89     0.712      6.86 6.20e-11     3.48     6.29
```