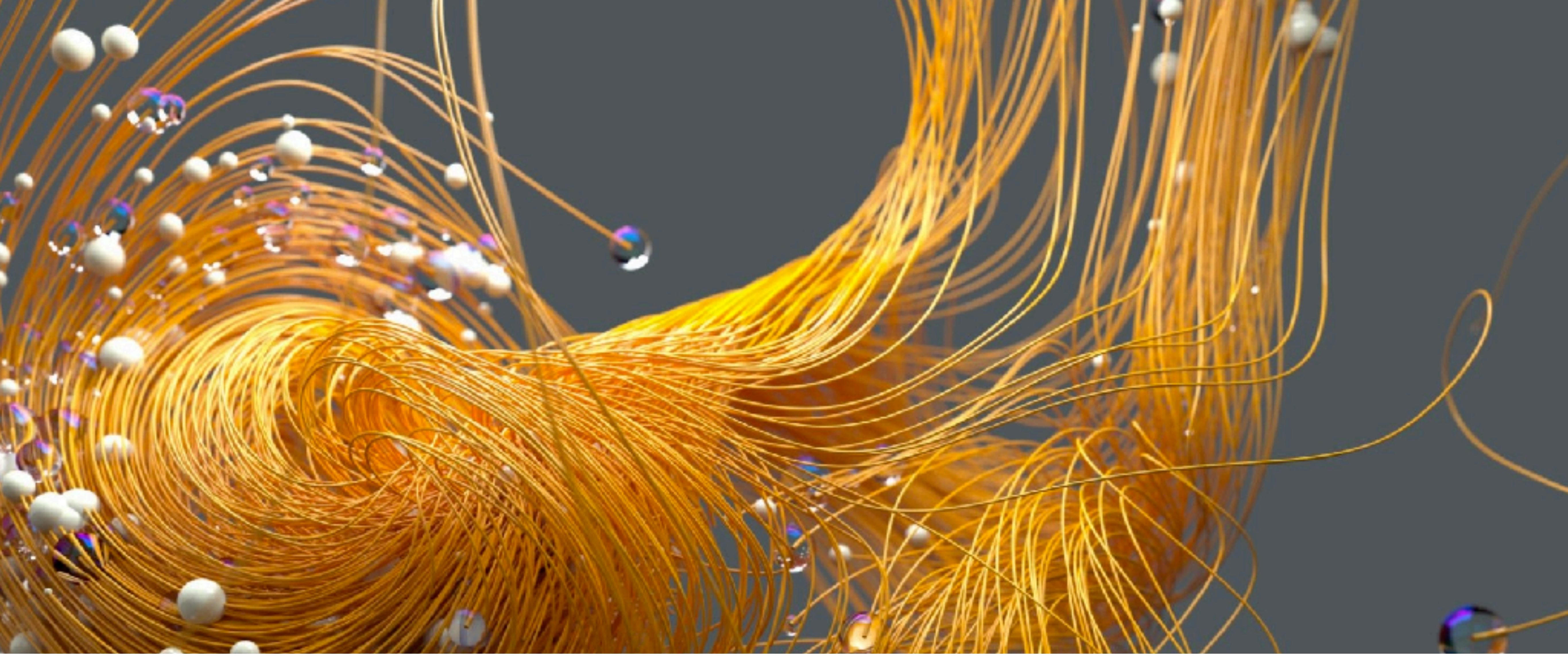


Psyc 60: Intro to Statistics

Prof. Judith Fan

Spring 2022

Last Time

Lab 4c: Effect Size

1 → 2 → 3

1. *General announcements*
2. *Break out into lab groups* — Stay in the same groups as last time!
3. *Complete daily feedback survey* — Link is in the course syllabus

Due This Week

8	May 16	**Confidence intervals and effect size** *Before:* Chapter 11 *During:* Lab 4C	**Review Session 3** *Before:* None *During:* Wrap-up Lab 4	Quiz 4; Project Milestone 4 Due (Model Fitting)

Due This Week

8	May 16	**Confidence intervals and effect size** *Before:* Chapter 11 *During:* Lab 4C	**Review Session 3** *Before:* None *During:* Wrap-up Lab 4	Quiz 4; Project Milestone 4 Due (Model Fitting)

Released Thursday at 5PM & due by 4:59PM on Friday

Due This Week

8	May 16	**Confidence intervals and effect size** *Before:* Chapter 11 *During:* Lab 4C	**Review Session 3** *Before:* None *During:* Wrap-up Lab 4	Quiz 4; Project Milestone 4 Due (Model Fitting)

Milestone 4A: Data Preprocessing AND Milestone 4B: Model Fitting both due this Friday by 11:59PM PT.

Reminders & Announcements

Please be mindful of UC San Diego campus policy on masking in instructional contexts. Even while masking is no longer required in many places on campus, everyone in PSYC 60 is still expected to continue masking in **lecture, section, and office hours**.

Reminders & Announcements

Please be mindful of UC San Diego campus policy on masking in **instructional contexts.** Even while masking is no longer required in many places on campus, everyone in PSYC 60 is still expected to continue masking in **lecture, section, and office hours**.

SONA credit hours are due by Wed., June 1 at 4PM! Sign up for studies now to make sure you can get these done on time!

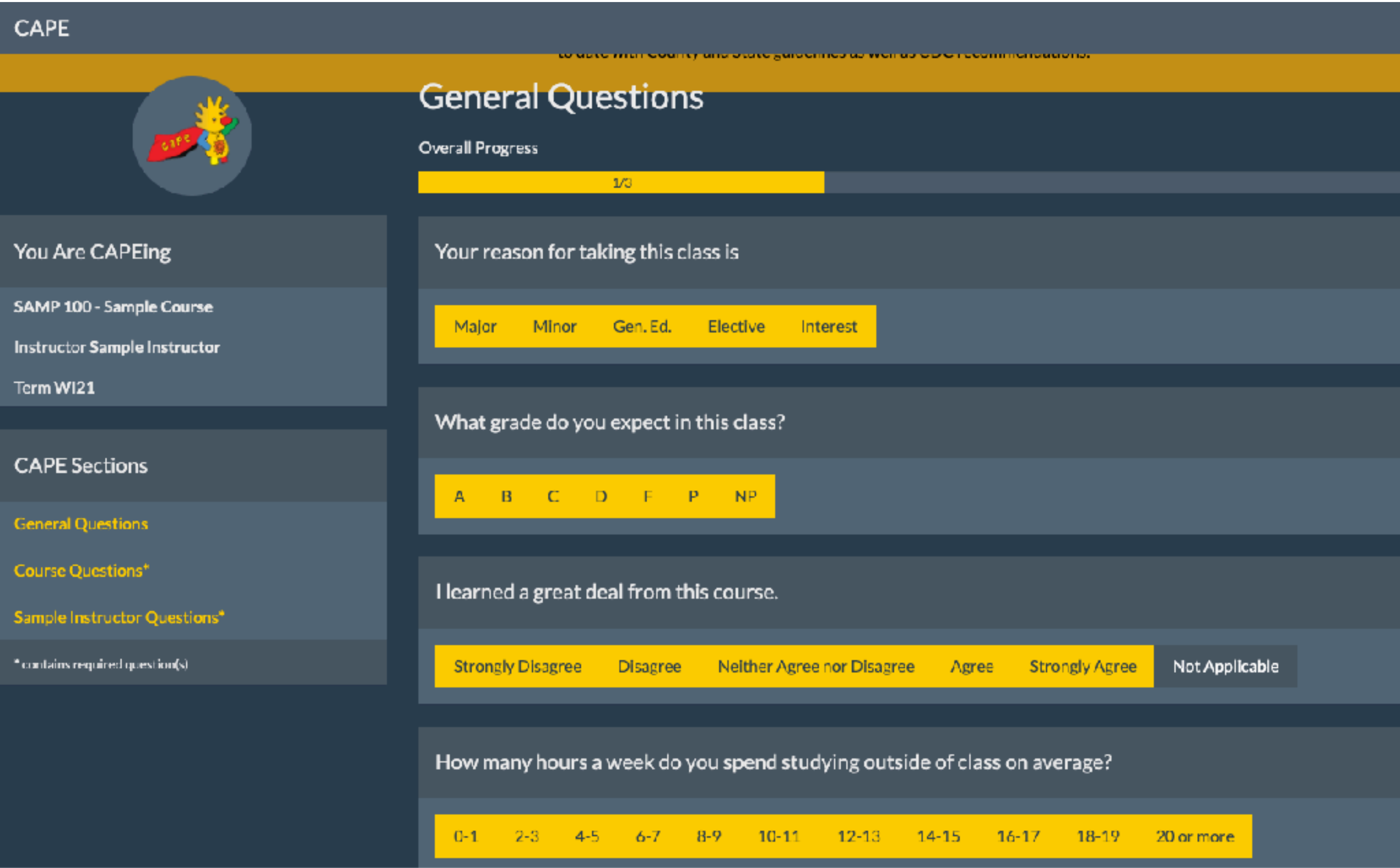
Reminders & Announcements

Please be mindful of UC San Diego campus policy on masking in instructional contexts. Even while masking is no longer required in many places on campus, everyone in PSYC 60 is still expected to continue masking in **lecture, section, and office hours**.

SONA credit hours are due by Wed., June 1 at 4PM! Sign up for studies now to make sure you can get these done on time!

Prof. Fan's Office Hours are today at 11am-12pm! Feel free to walk back to McGill Hall Room 5141 with me after class!

CAPES OPEN NEXT WEEK!



Capes open next week!

- **CAPEs (Course and Professor Evaluations) open next Monday, May 23!** Please do take the time to fill one out for PSYC 60!

- **This is your chance to share your <u>honest</u> feedback about your learning experience in this course and help me make this course better.** I read and take your comments seriously, and it means a lot to me to get an accurate picture of what you've gotten out of this course, so I know what I can do in the future to better support future students in this class. These evaluations are also the official ones used by UCSD to evaluate me as a professor.

- **I appreciate your courage & persistence throughout this quarter, at the end of a challenging year.** It has made me so proud to see each of you grow and learn, and I admire your willingness to embrace these intellectual challenges and also *<u>support one another with patience and compassion.</u>*

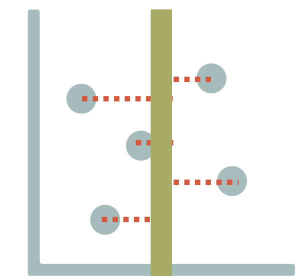
Today

Mini-Review Session #3

1 → 2 → 3

1. *Using an explanatory variable to model variation in an outcome variable*
2. *Quantifying effects using confidence intervals*
3. *What is the (Pearson) correlation coefficient?*

1 *Using an explanatory variable to model variation in an outcome variable*



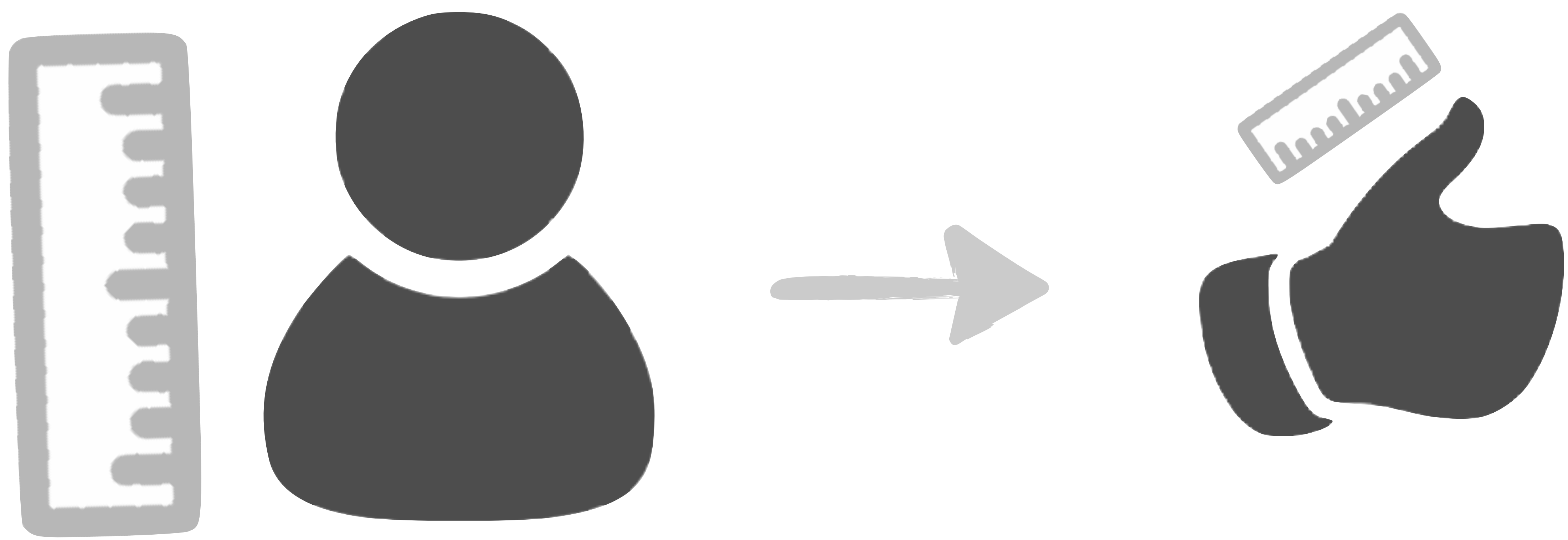
data = model + error

data	model	error
what we actually observe	what we expect to observe	difference between expected and observed

1

Using an explanatory variable to model variation in an outcome variable

Example: Predicting height from thumb length



1

Using an explanatory variable to model variation in an outcome variable

What is the General Linear Model (GLM)?

A general linear model is a specific type of statistical model in which the values of a dependent/outcome variable is determined by a linear combination of independent predictor variables that are each multiplied by a weight (often represented by the letter ***b*** or Greek letter "beta," β).

$$Y_i = b_0 + b_1 X_i + e_i$$

observed value of outcome variable (*e.g., thumb length*) — *intercept* — *slope* — *value of predictor variable* (*e.g., height*) — *error*

$\hat{Y}_i$ ***predicted*** *value of outcome variable*

Y_i ***observed*** *value of outcome variable*

1

Using an explanatory variable to model variation in an outcome variable

Using algebraic notation to specify linear regression model

*The little **i** is a counter that refers to each individual data point.*

$$Y_i = b_0 + b_1 X_i + e_i$$

Y_i: observed value of outcome variable (e.g., thumb length)

b_0: intercept

b_1: slope

X_i: value of predictor variable (e.g., height)

e_i: error

$\hat{Y}_i$ **predicted** value of outcome variable

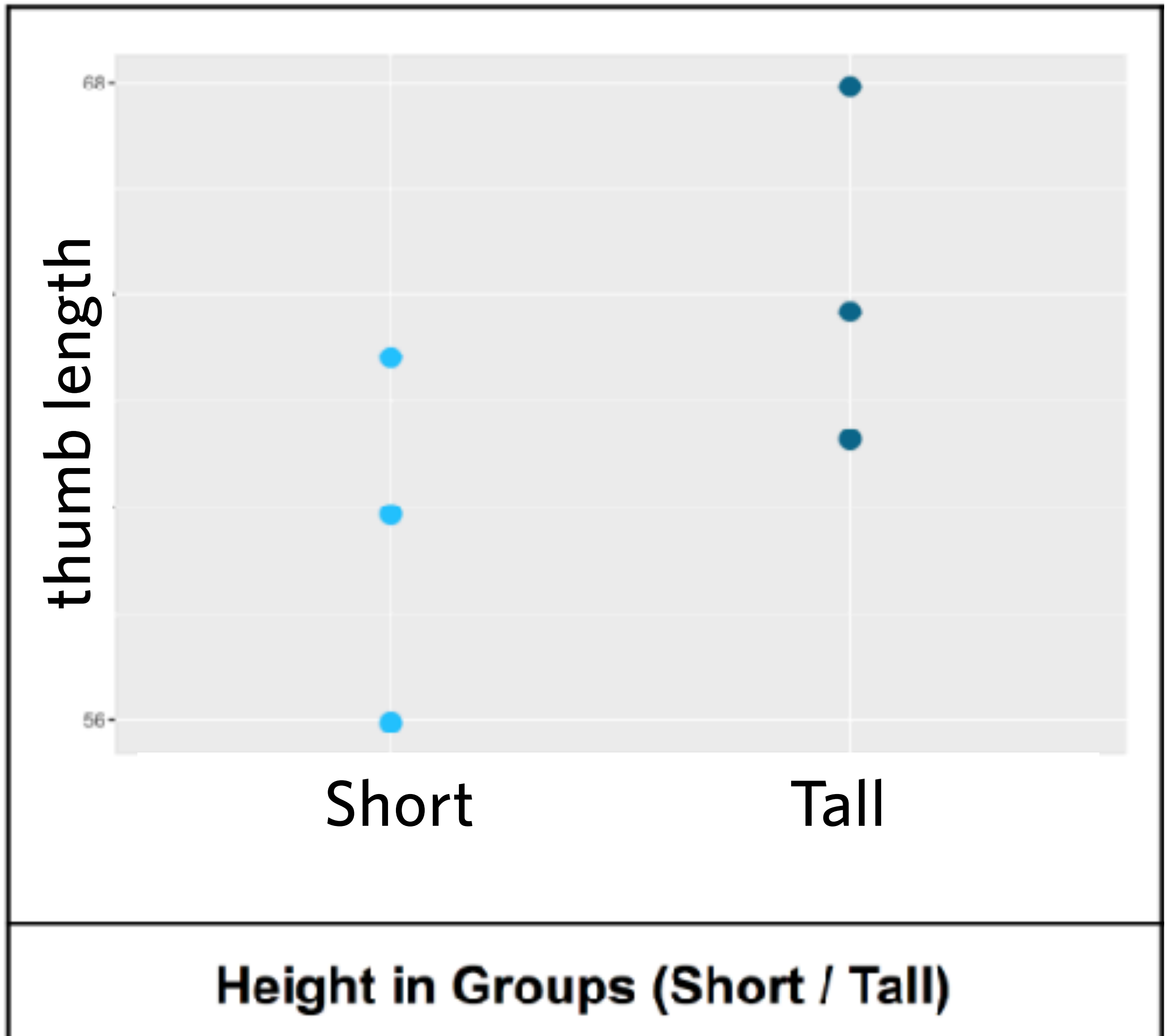
Y_i **observed** value of outcome variable

1

Using an explanatory variable to model variation in an outcome variable

Example: Predicting thumb length from height

Approach 1: Compare difference in mean thumb length between two groups

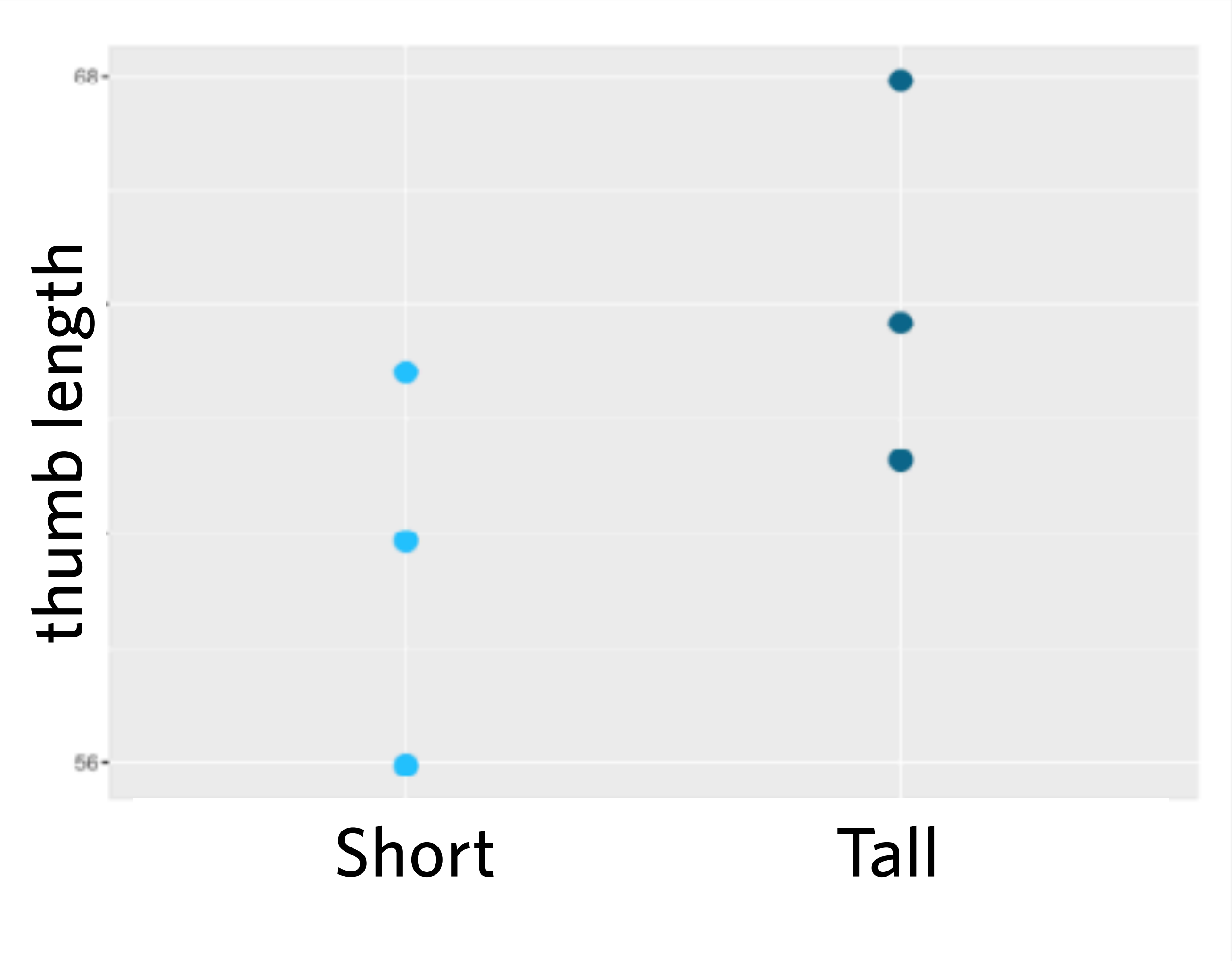


1 *Using an explanatory variable to model variation in an outcome variable*

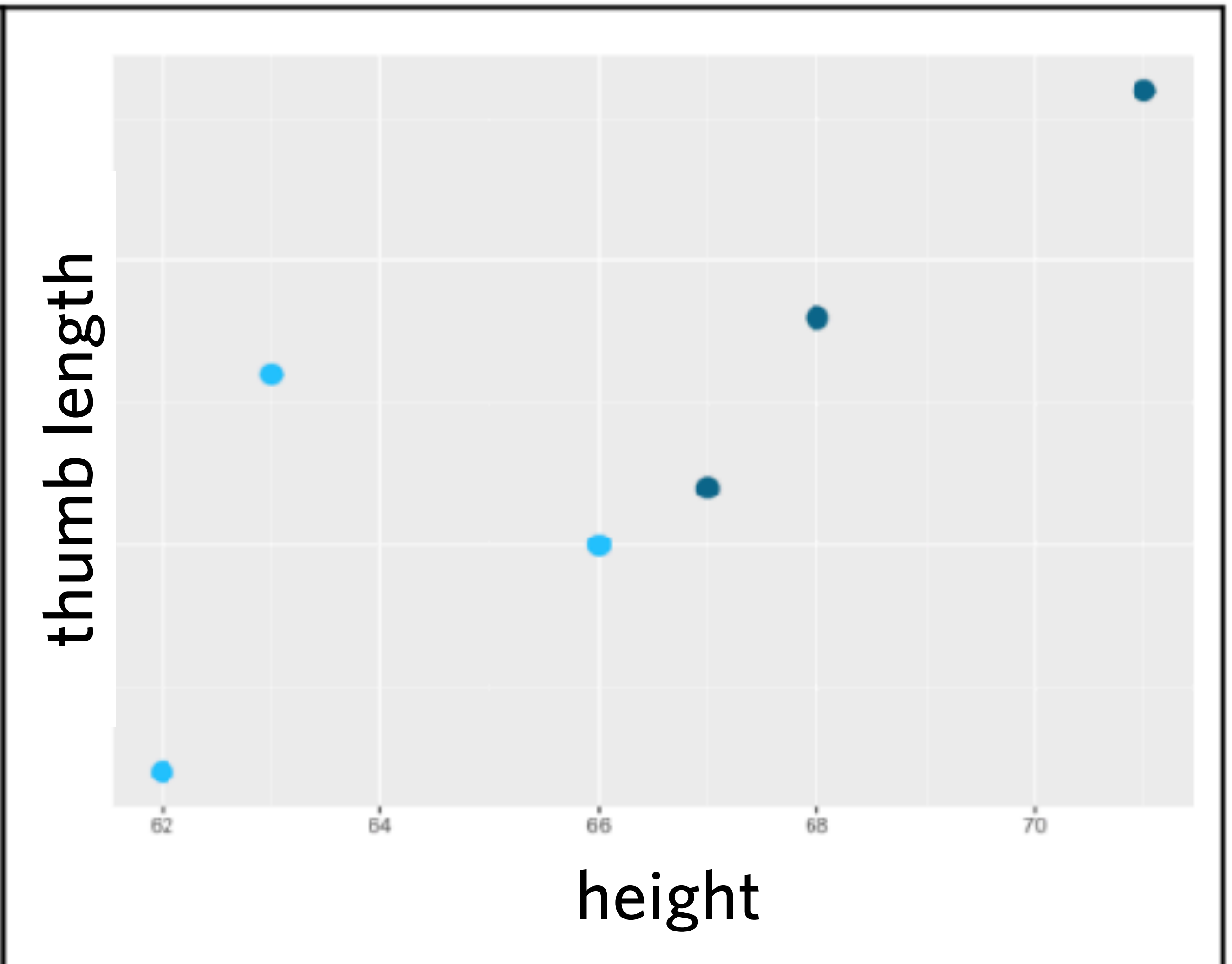
Example: Predicting thumb length from height

Approach 1: Compare difference in mean thumb length between two groups

Approach 2: Use linear regression to model continu relationship between height and thumb length



Height in Groups (Short / Tall)



Height in Inches

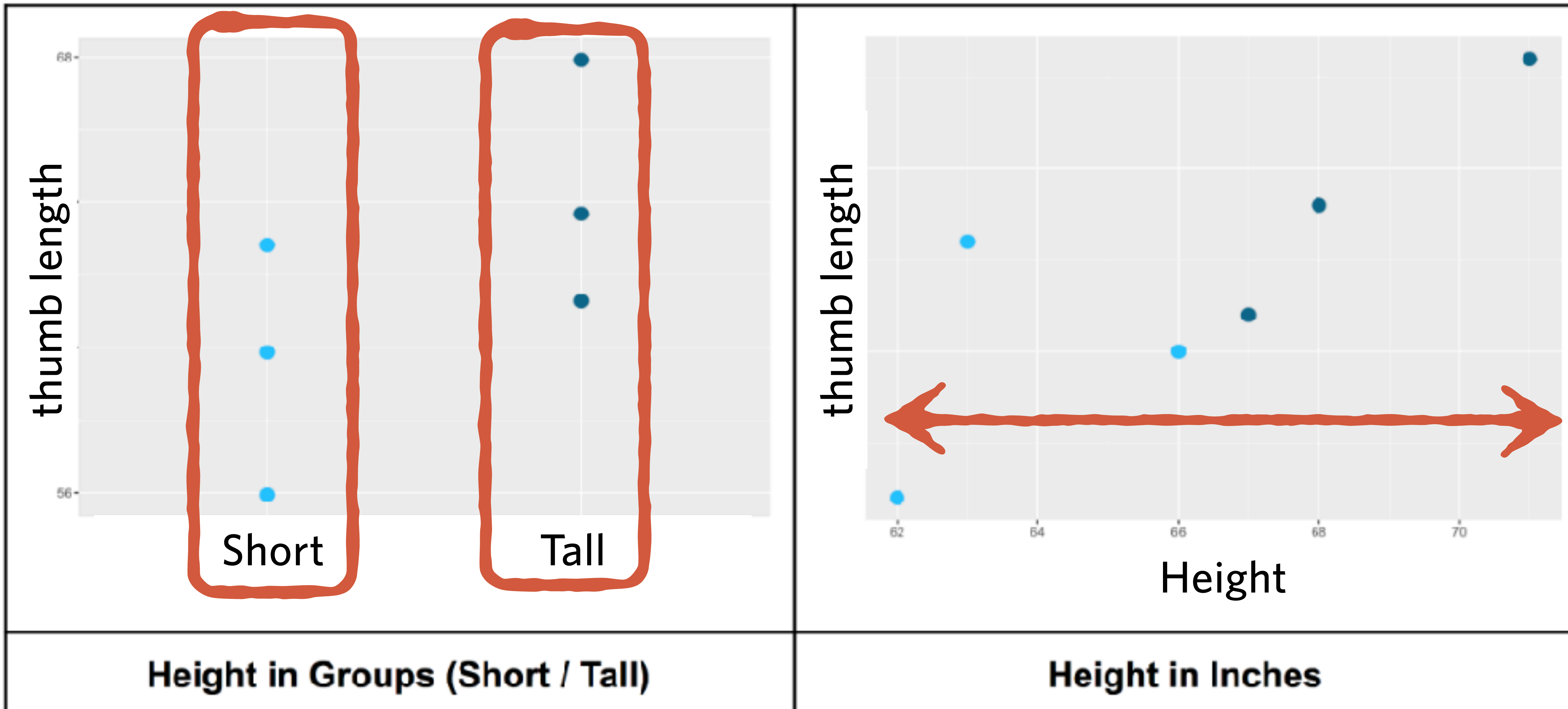
1

Using an explanatory variable to model variation in an outcome variable

Example: Predicting thumb length from height

Approach 1: Compare difference in mean thumb length between two groups

Approach 2: Use linear regression to model continu relationship between height and thumb length



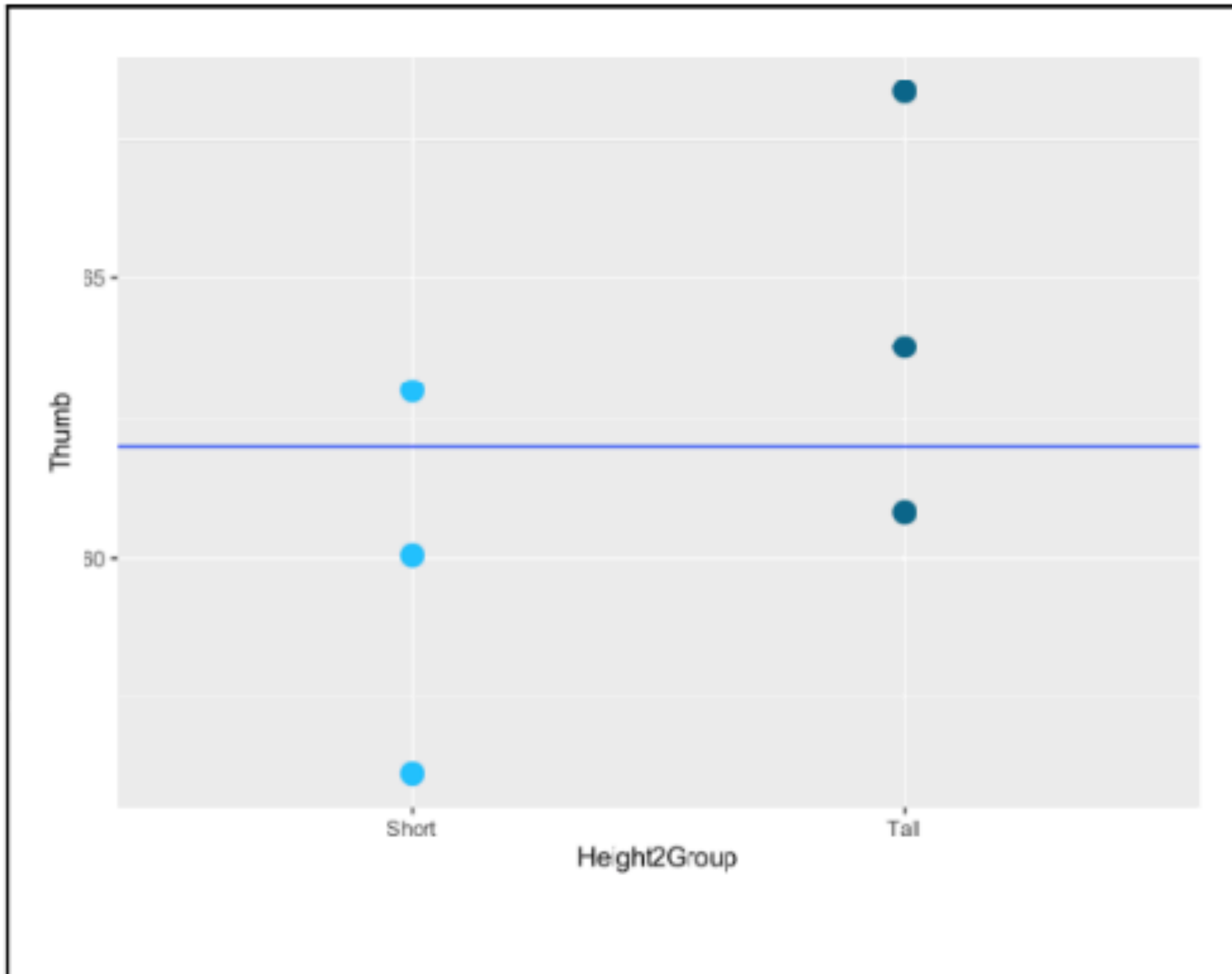
1

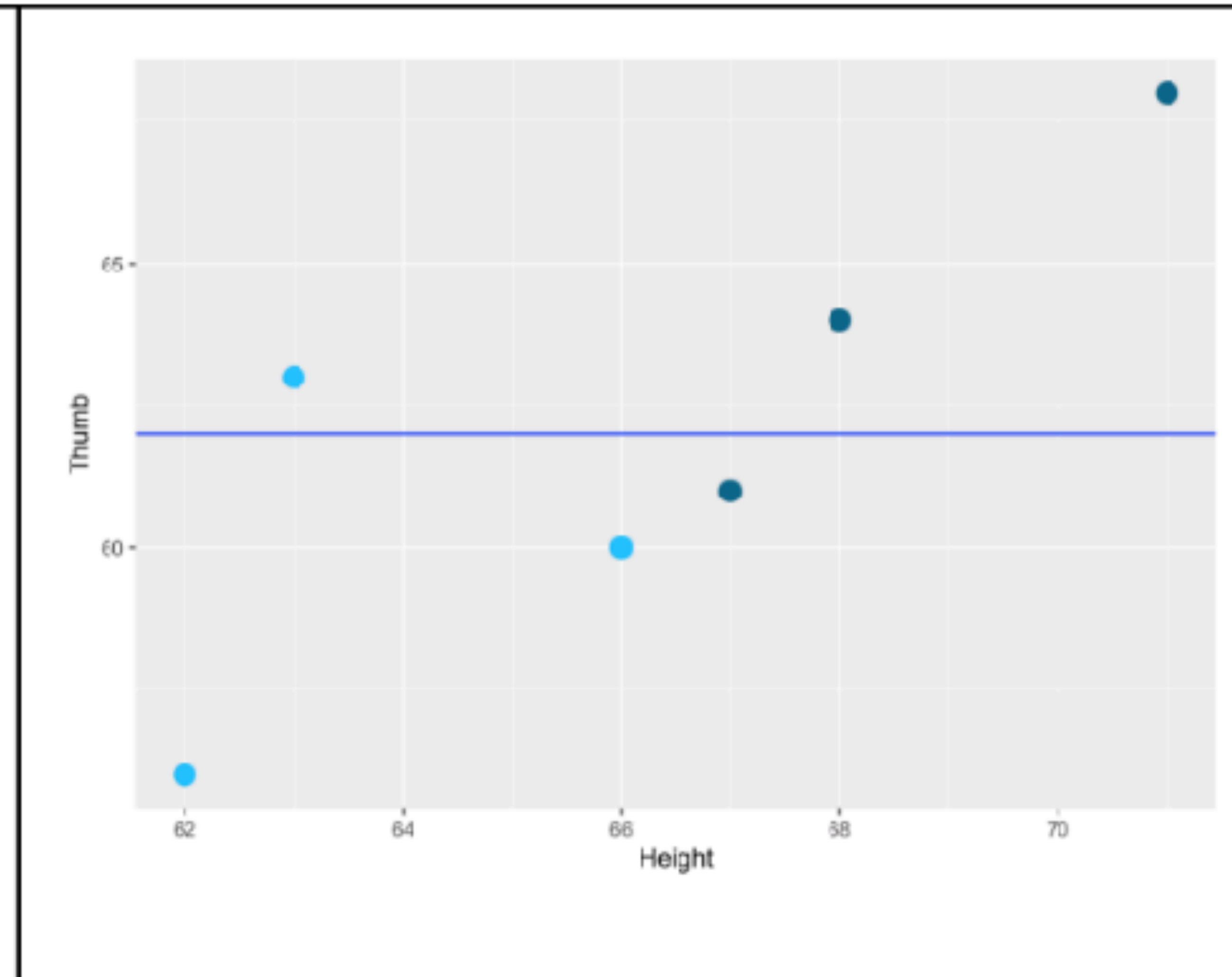
Using an explanatory variable to model variation in an outcome variable

Example: Predicting thumb length from height

Approach 1: Compare difference in mean thumb length between two groups

Approach 2: Use linear regression to model continu relationship between height and thumb length

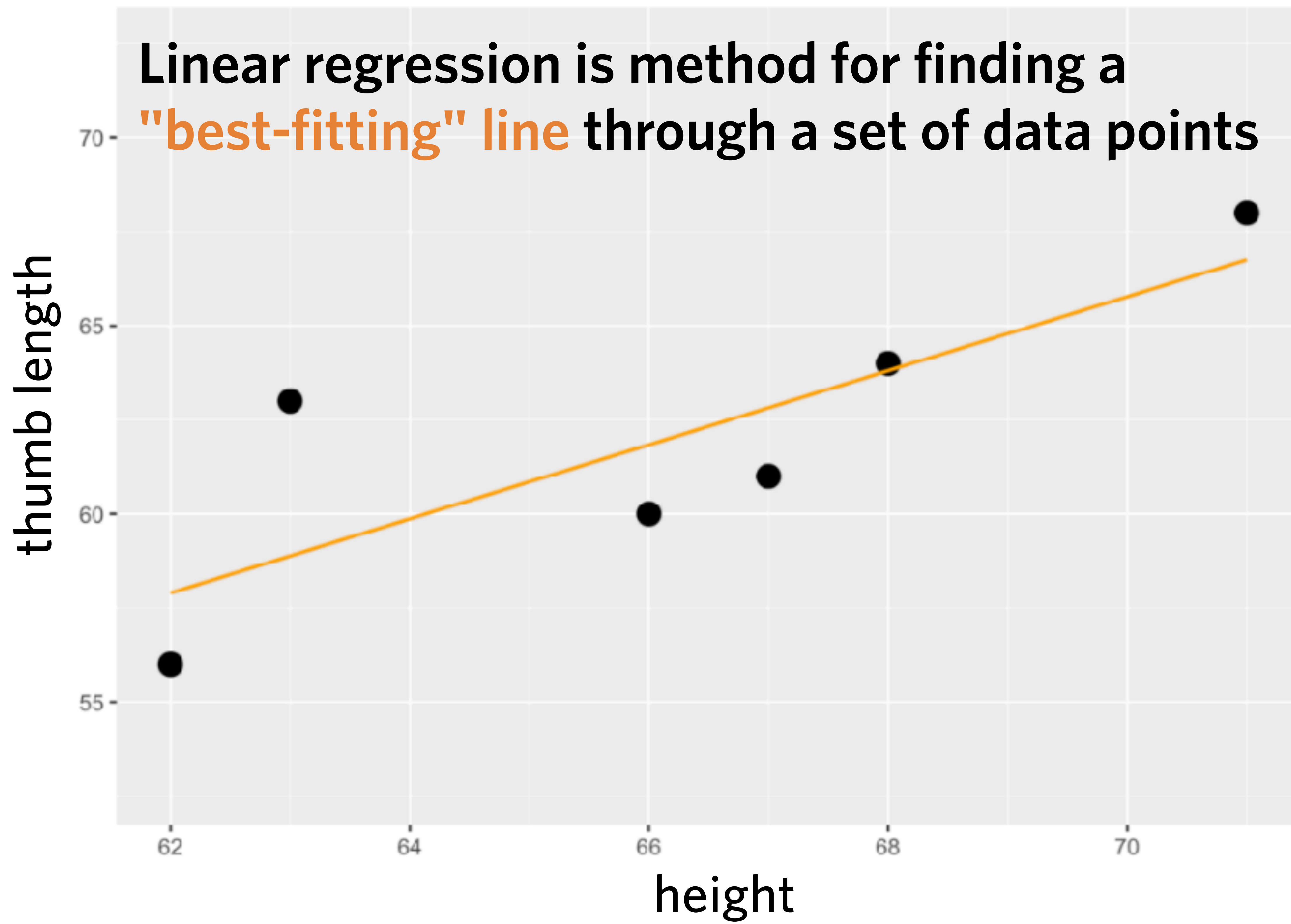




Empty model is the same regardless.

1

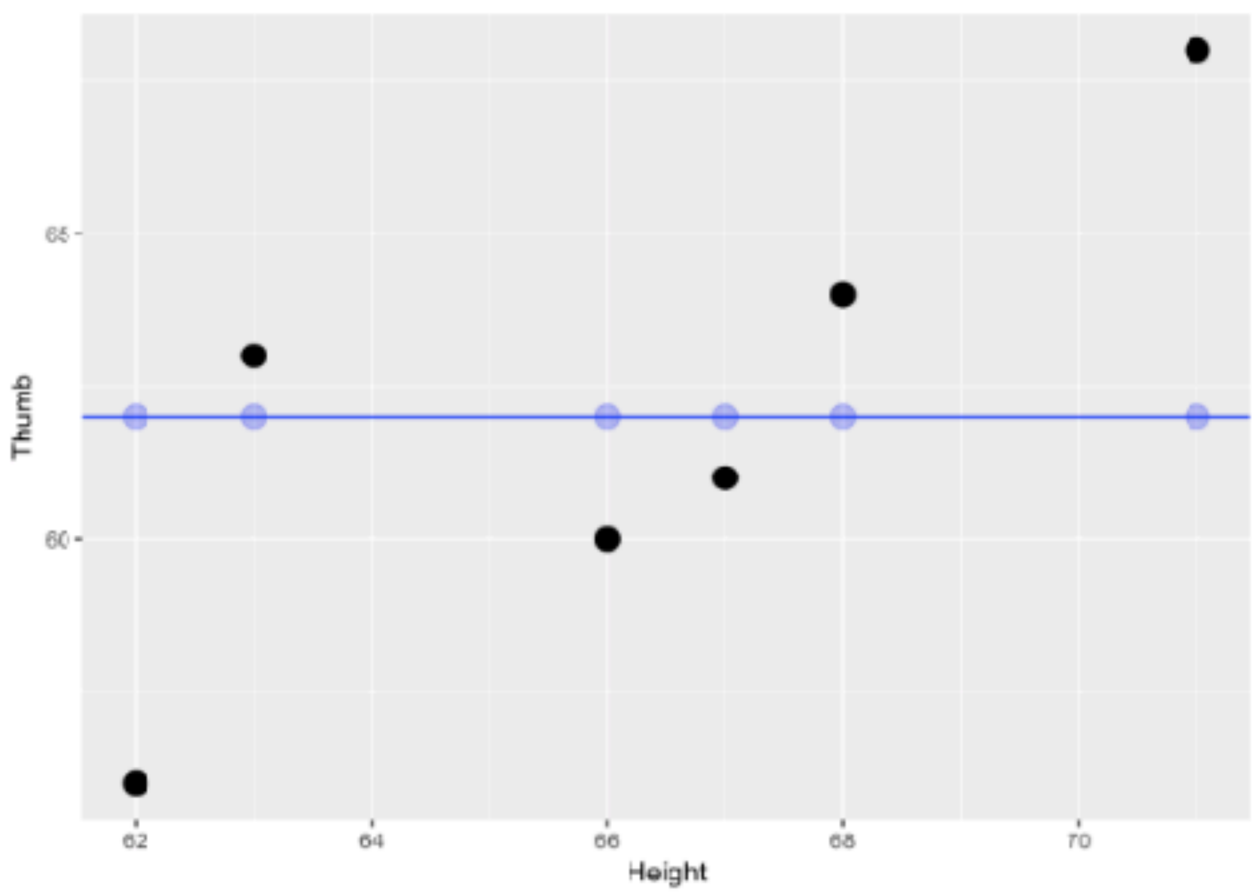
Using an explanatory variable to model variation in an outcome variable



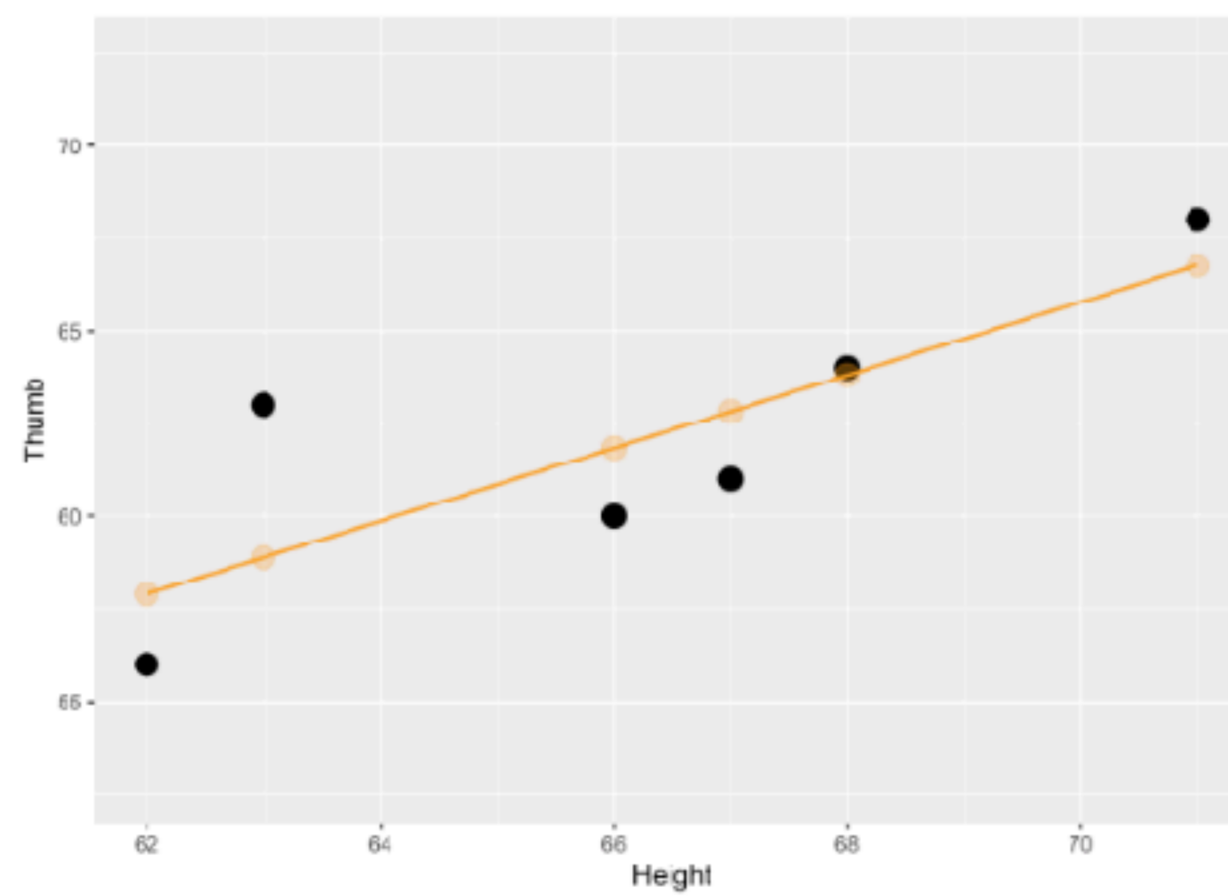
1 *Using an explanatory variable to model variation in an outcome variable*

Empty model could be visualized as a horizontal line (zero slope) drawn through the mean of outcome variable (e.g., mean thumb length)

Linear regression is method for finding a "best-fitting" line through a set of data points that takes the explanatory variable into account.

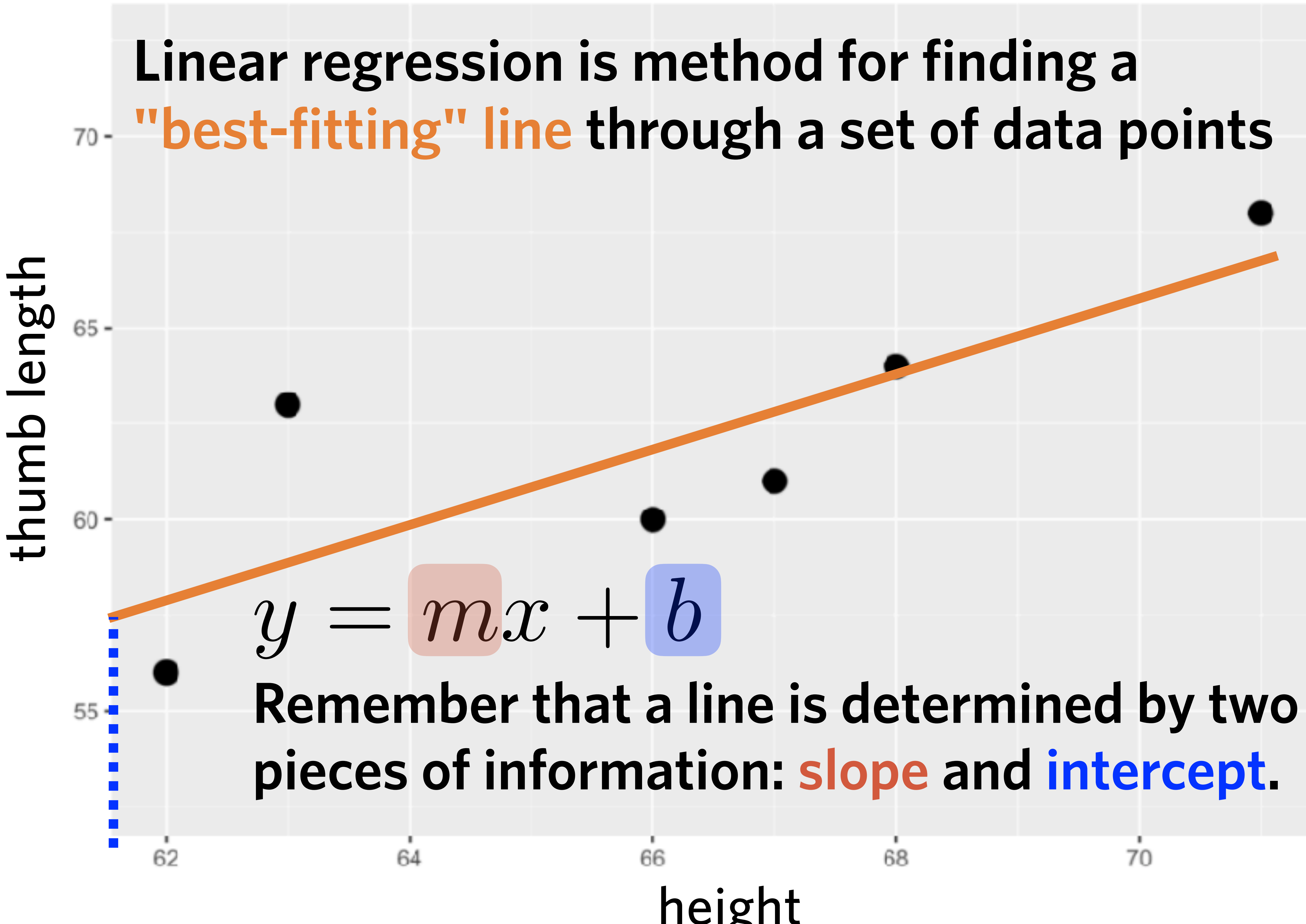


Empty Model



Height Model

1 *Using an explanatory variable to model variation in an outcome variable*

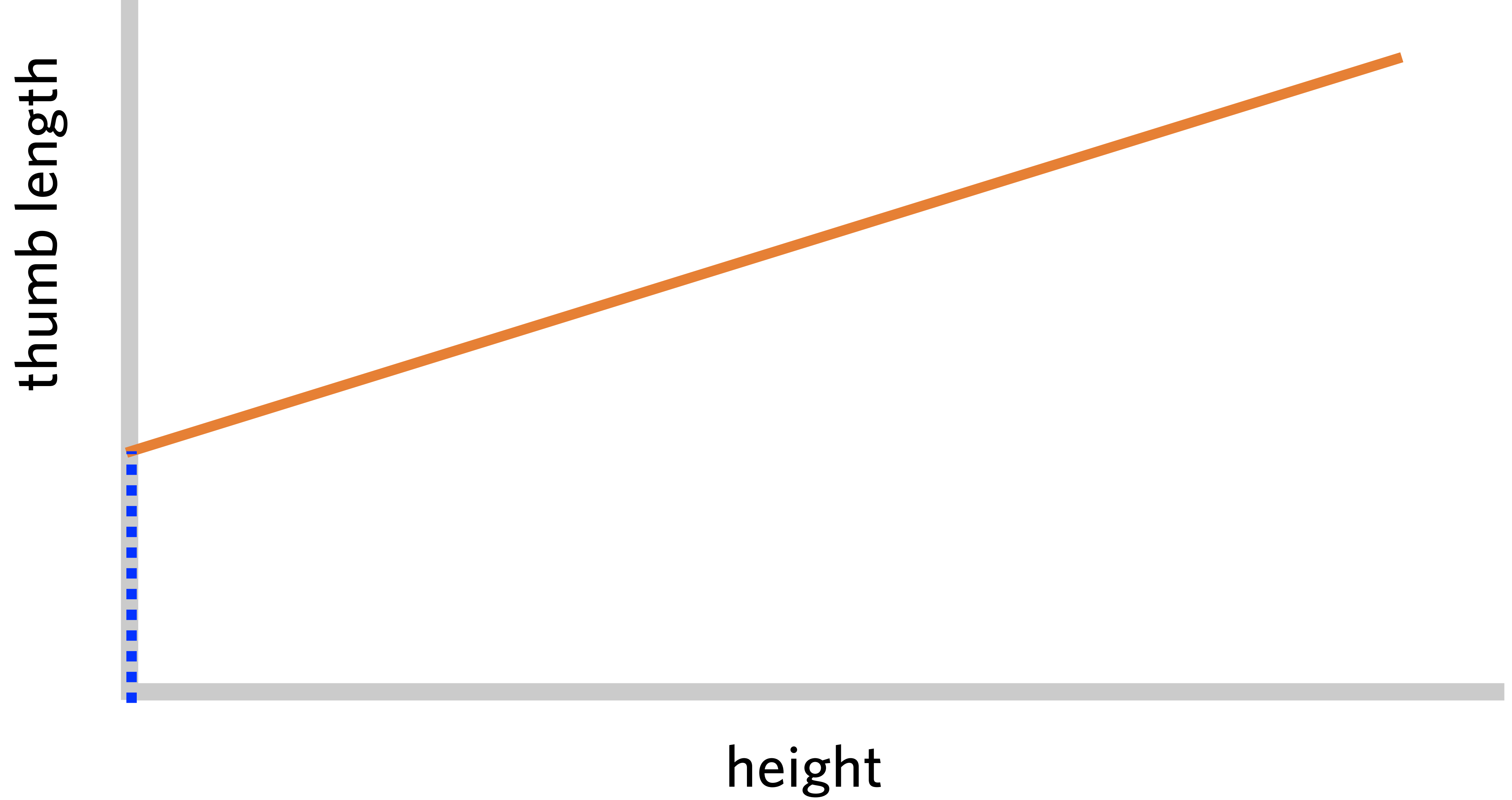


1

Using an explanatory variable to model variation in an outcome variable

$$y = mx + b$$

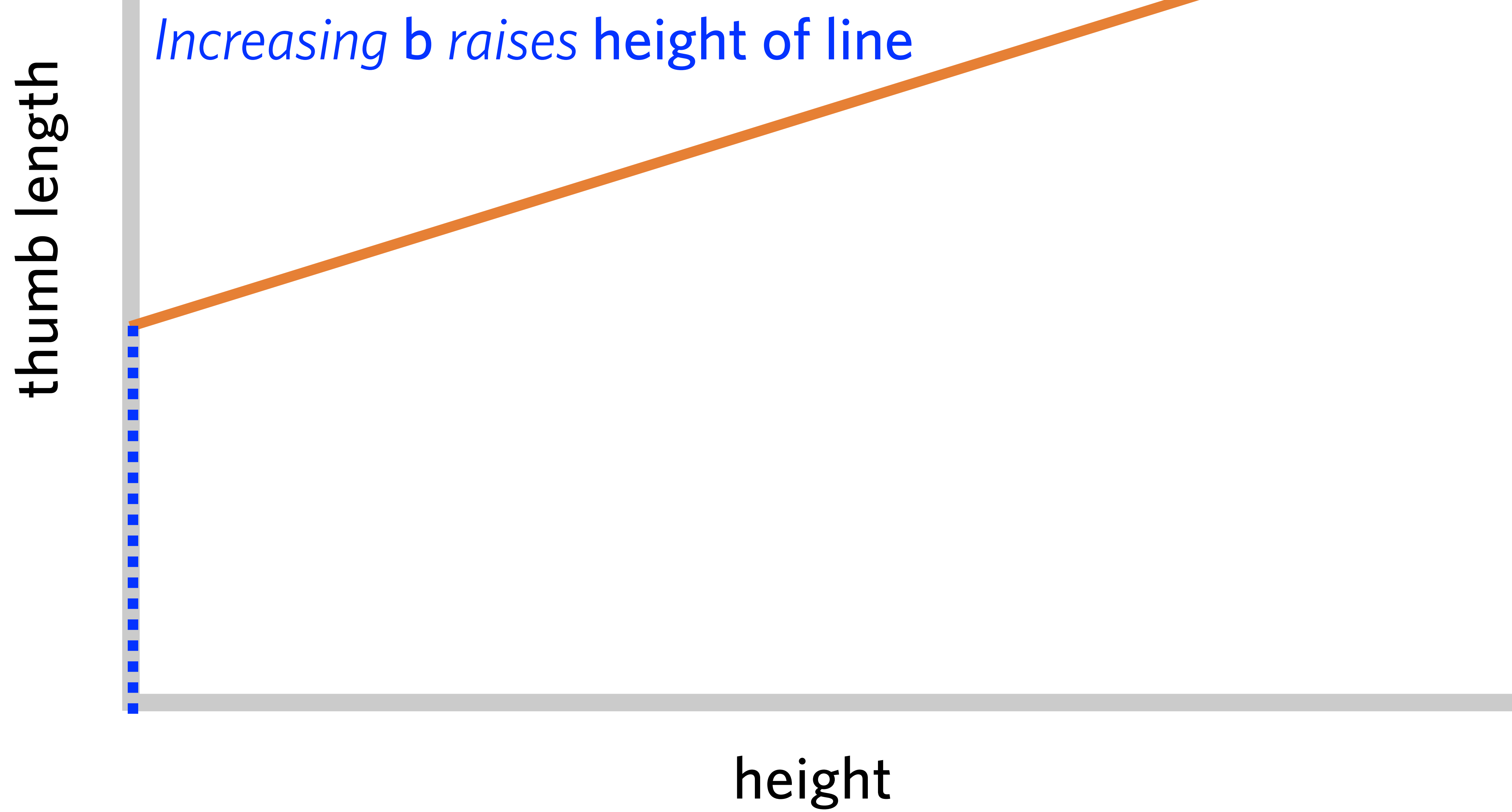
Remember that a line is determined by two pieces of information: slope and intercept.



1 *Using an explanatory variable to model variation in an outcome variable*

$$y = mx + b$$

Remember that a line is determined by two pieces of information: slope and intercept.

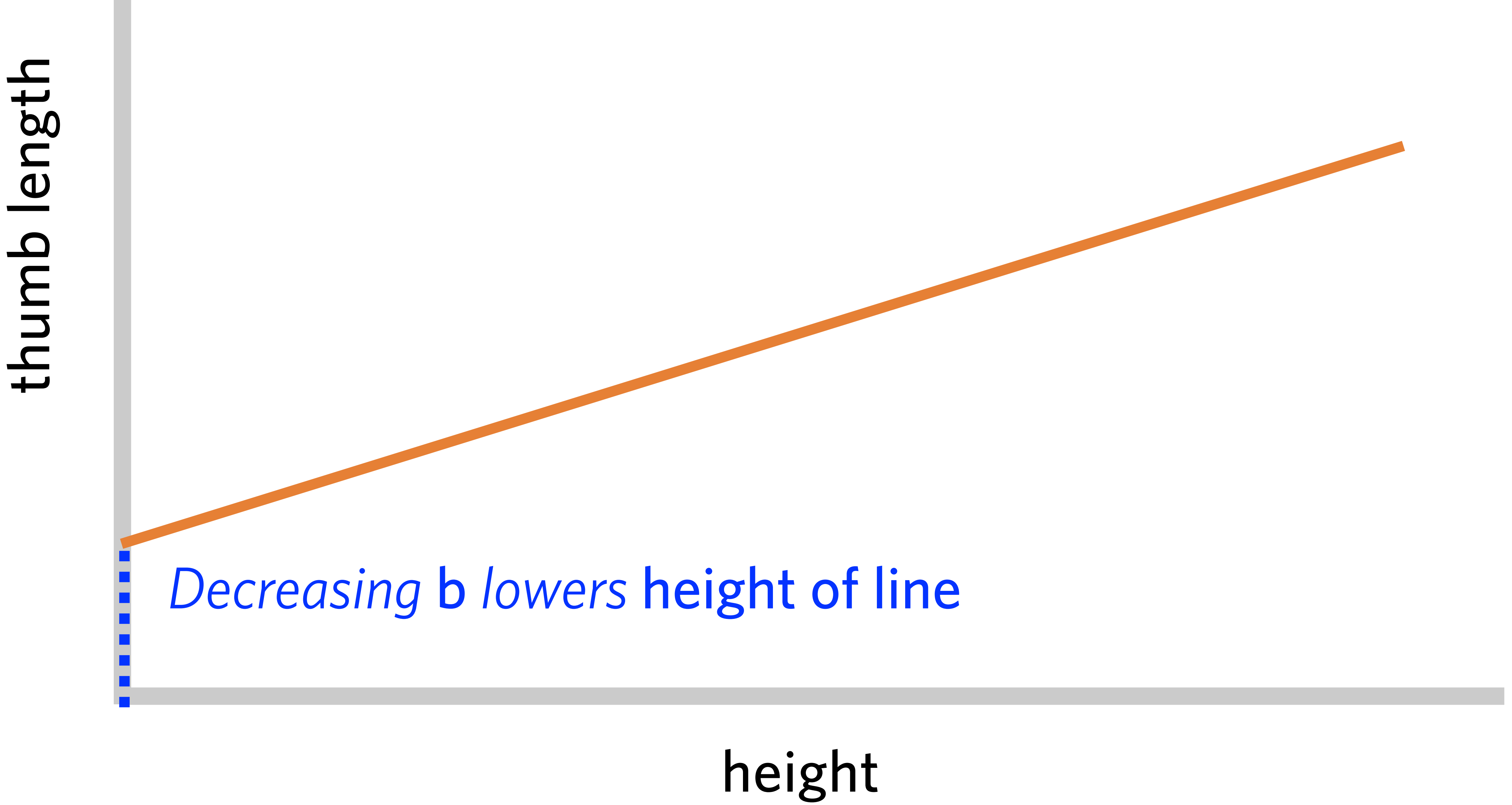


1

Using an explanatory variable to model variation in an outcome variable

$$y = mx + b$$

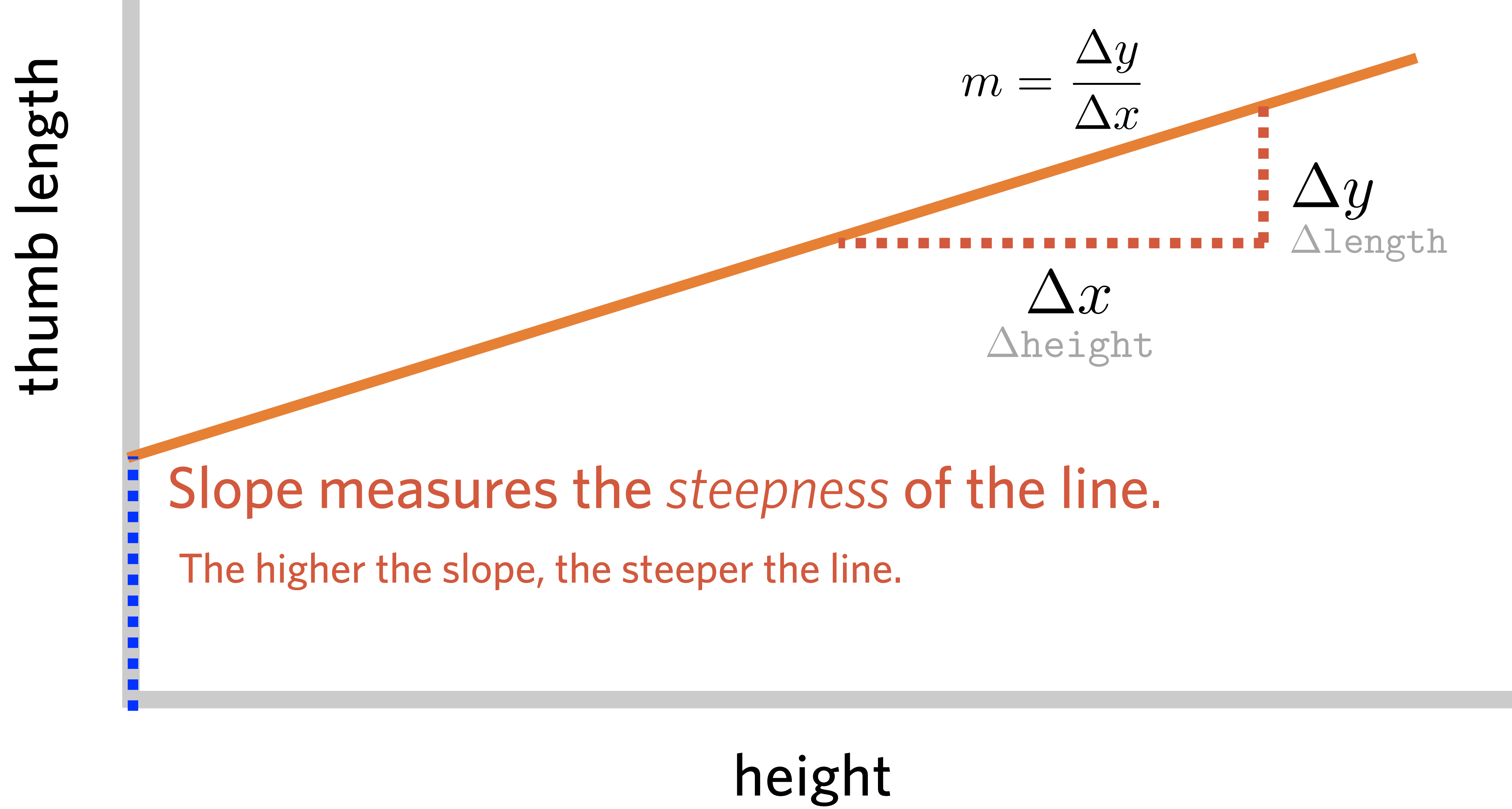
Remember that a line is determined by two pieces of information: slope and intercept.



1 *Using an explanatory variable to model variation in an outcome variable*

$$y = mx + b$$

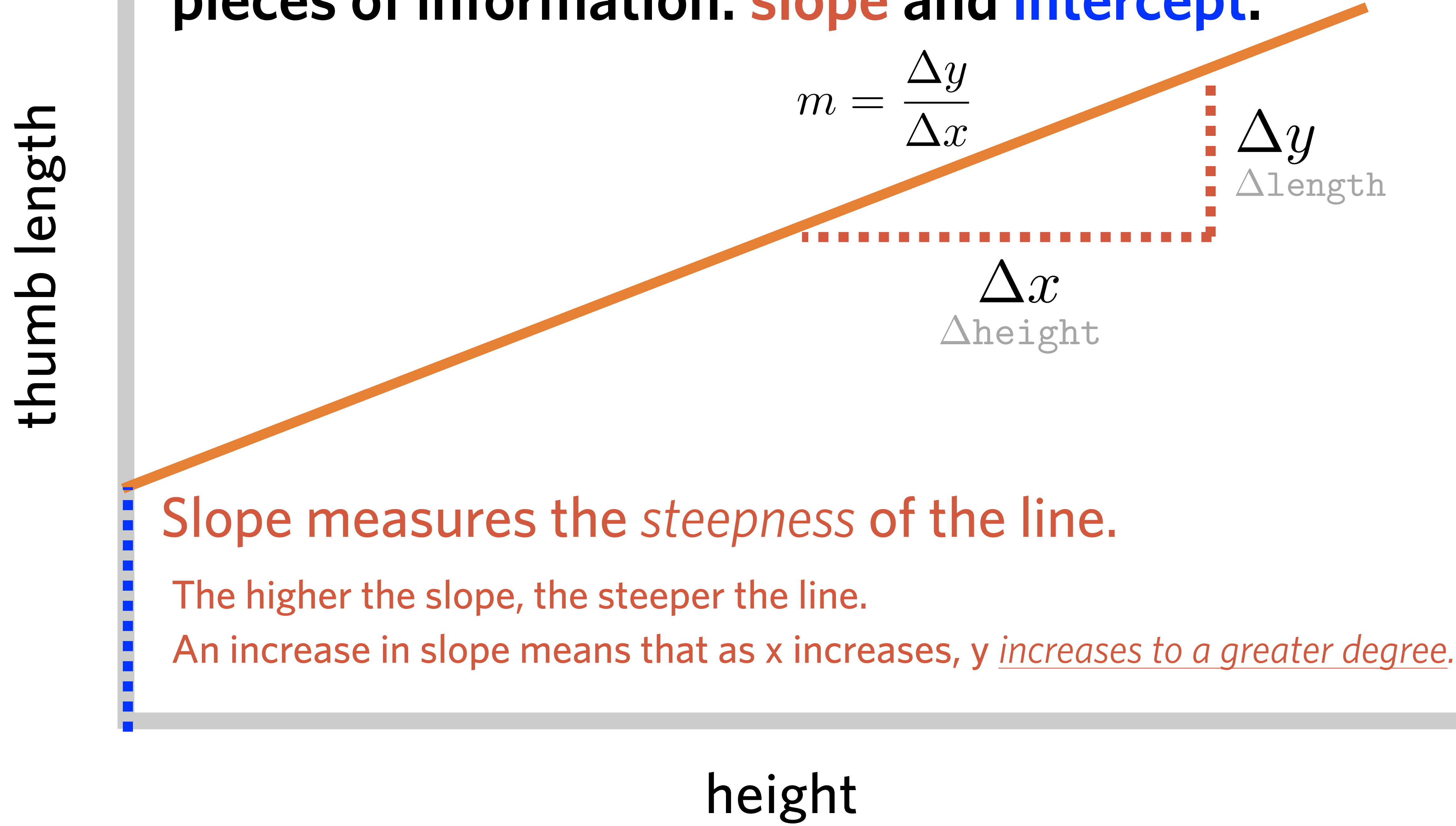
Remember that a line is determined by two pieces of information: slope and intercept.



1 *Using an explanatory variable to model variation in an outcome variable*

$$y = mx + b$$

Remember that a line is determined by two pieces of information: slope and intercept.



Slope measures the *steepness* of the line.

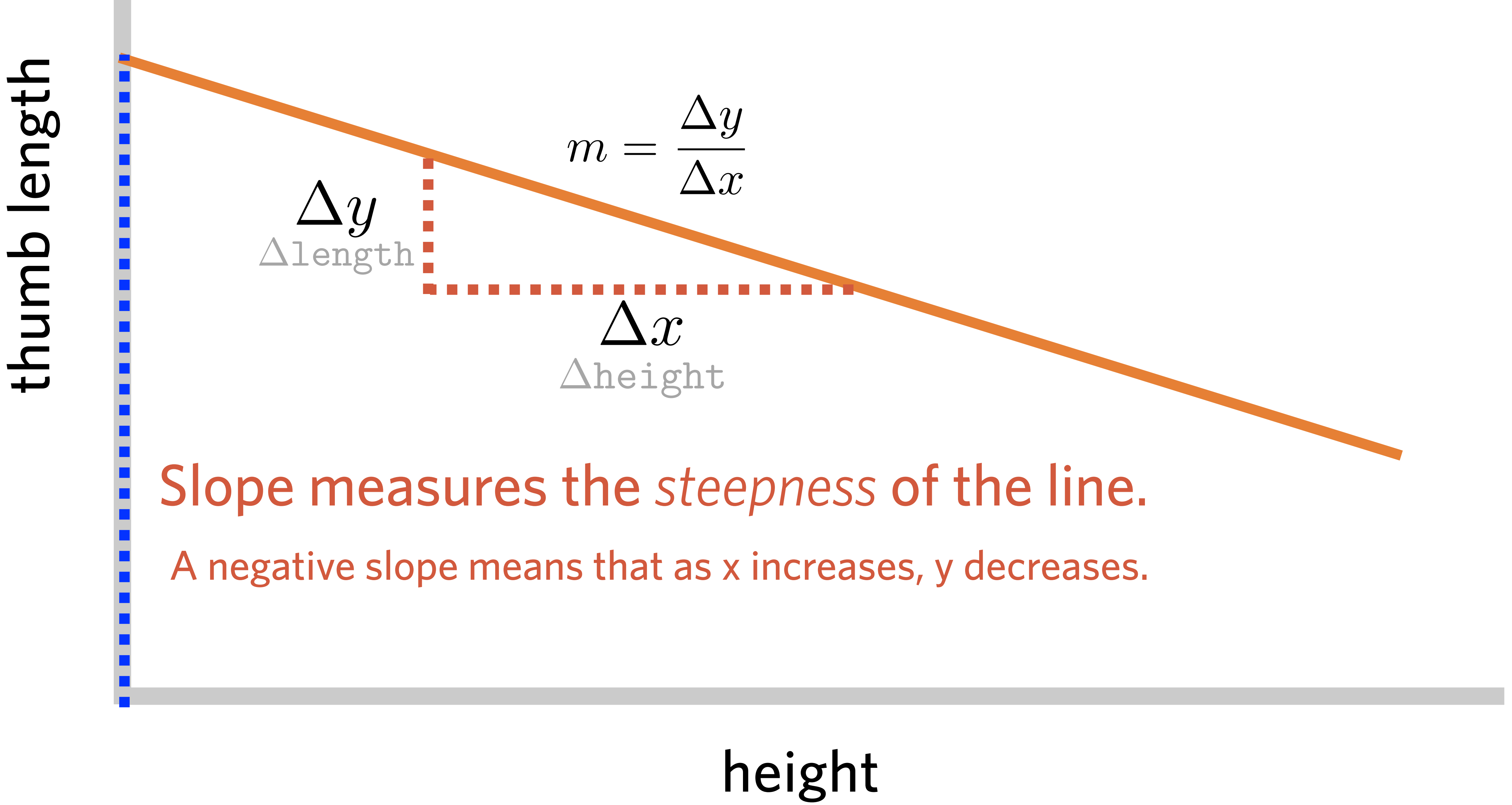
The higher the slope, the steeper the line.

An increase in slope means that as x increases, y *increases to a greater degree*.

1 *Using an explanatory variable to model variation in an outcome variable*

$$y = mx + b$$

Remember that a line is determined by two pieces of information: slope and intercept.



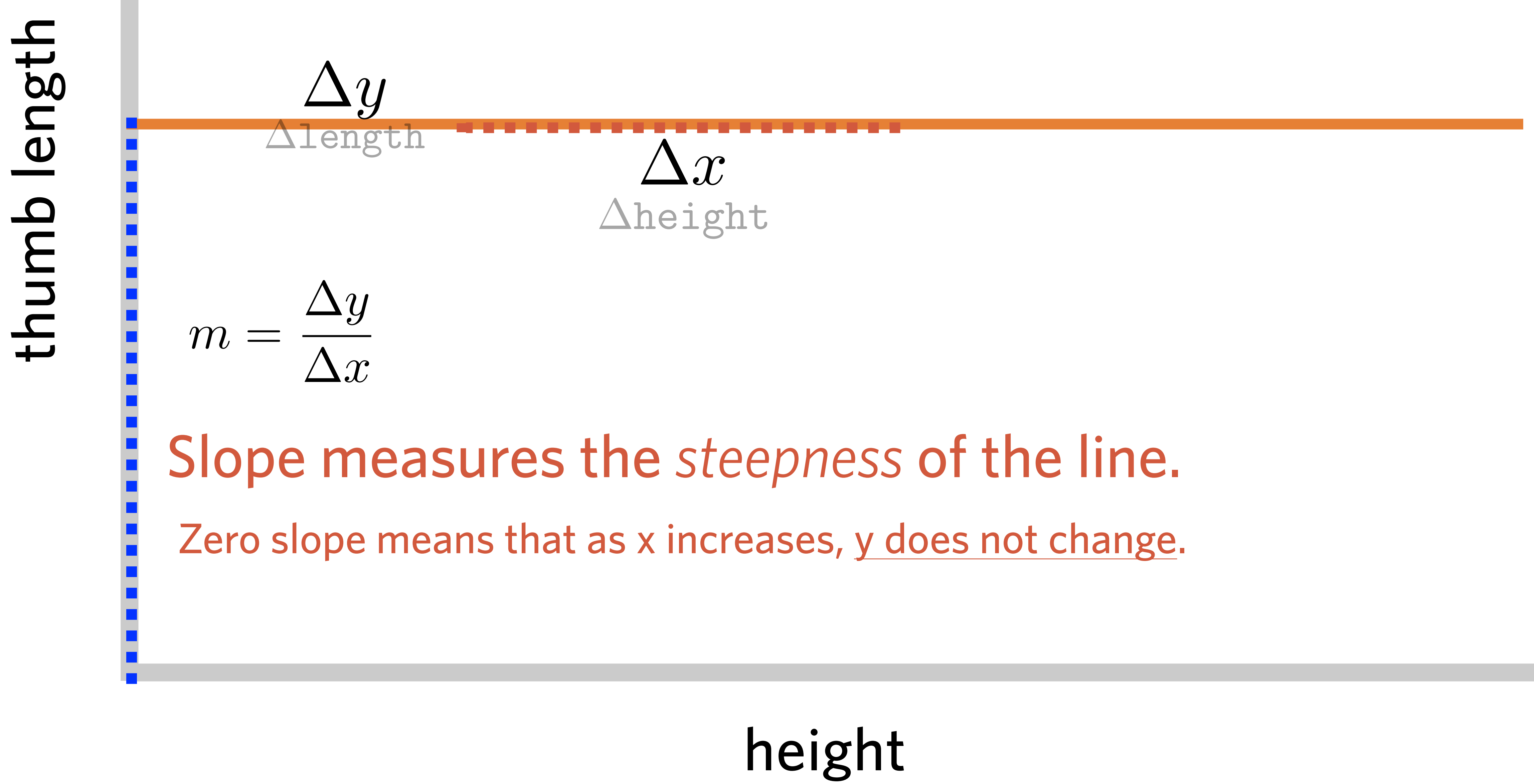
Slope measures the *steepness* of the line.

A negative slope means that as x increases, y decreases.

1 Using an explanatory variable to model variation in an outcome variable

$$y = mx + b$$

Remember that a line is determined by two pieces of information: slope and intercept.



$$m = \frac{\Delta y}{\Delta x}$$

Slope measures the *steepness* of the line.

Zero slope means that as x increases, y does not change.

1 *Using an explanatory variable to model variation in an outcome variable*

$$y = mx + b$$

Remember that a line is determined by two pieces of information: slope and intercept.

Δy

Δlength

Δx

Δheight

$$m = \frac{\Delta y}{\Delta x}$$

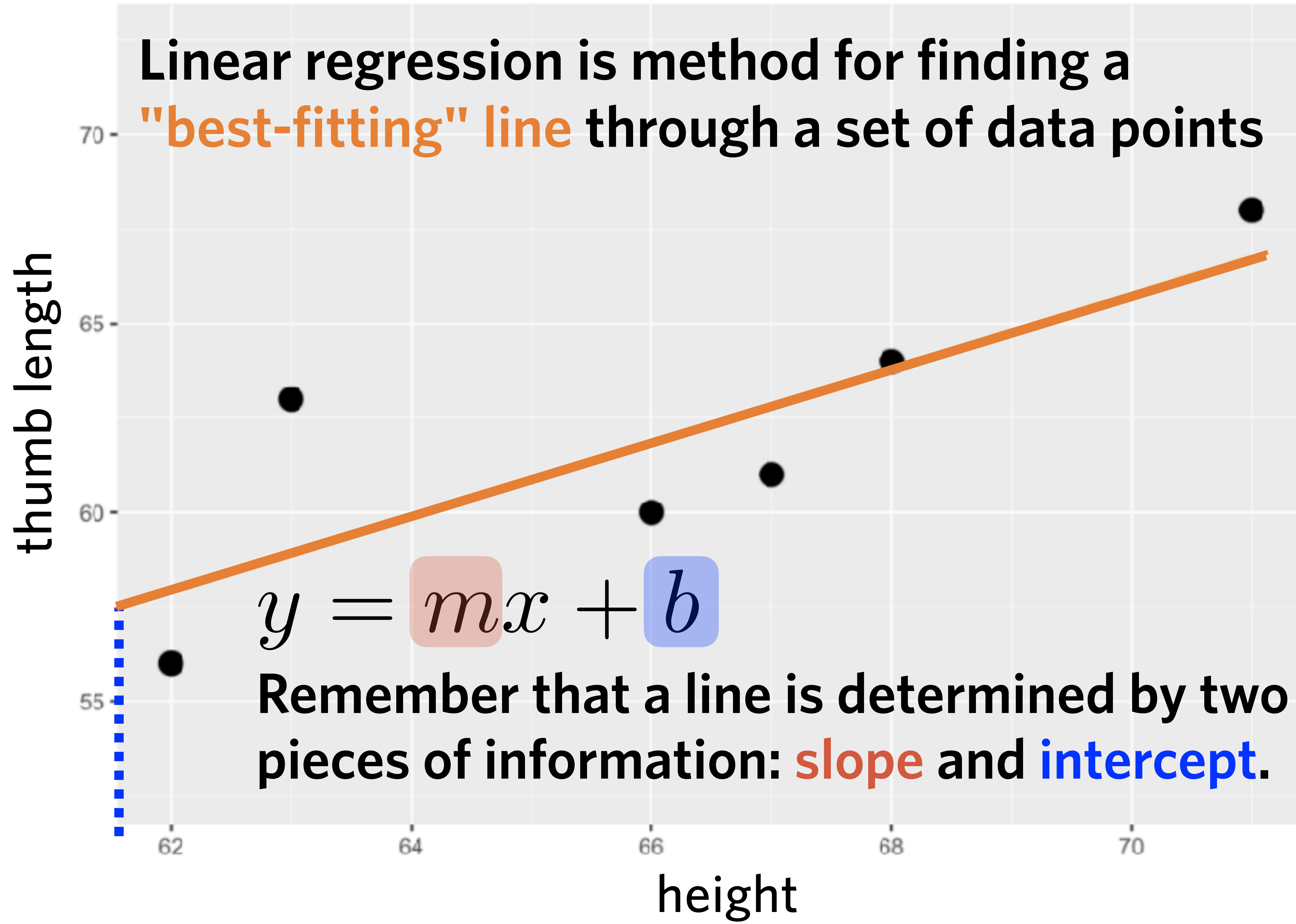
Slope measures the *steepness* of the line.

Zero slope means that as x increases, y does not change.

thumb length

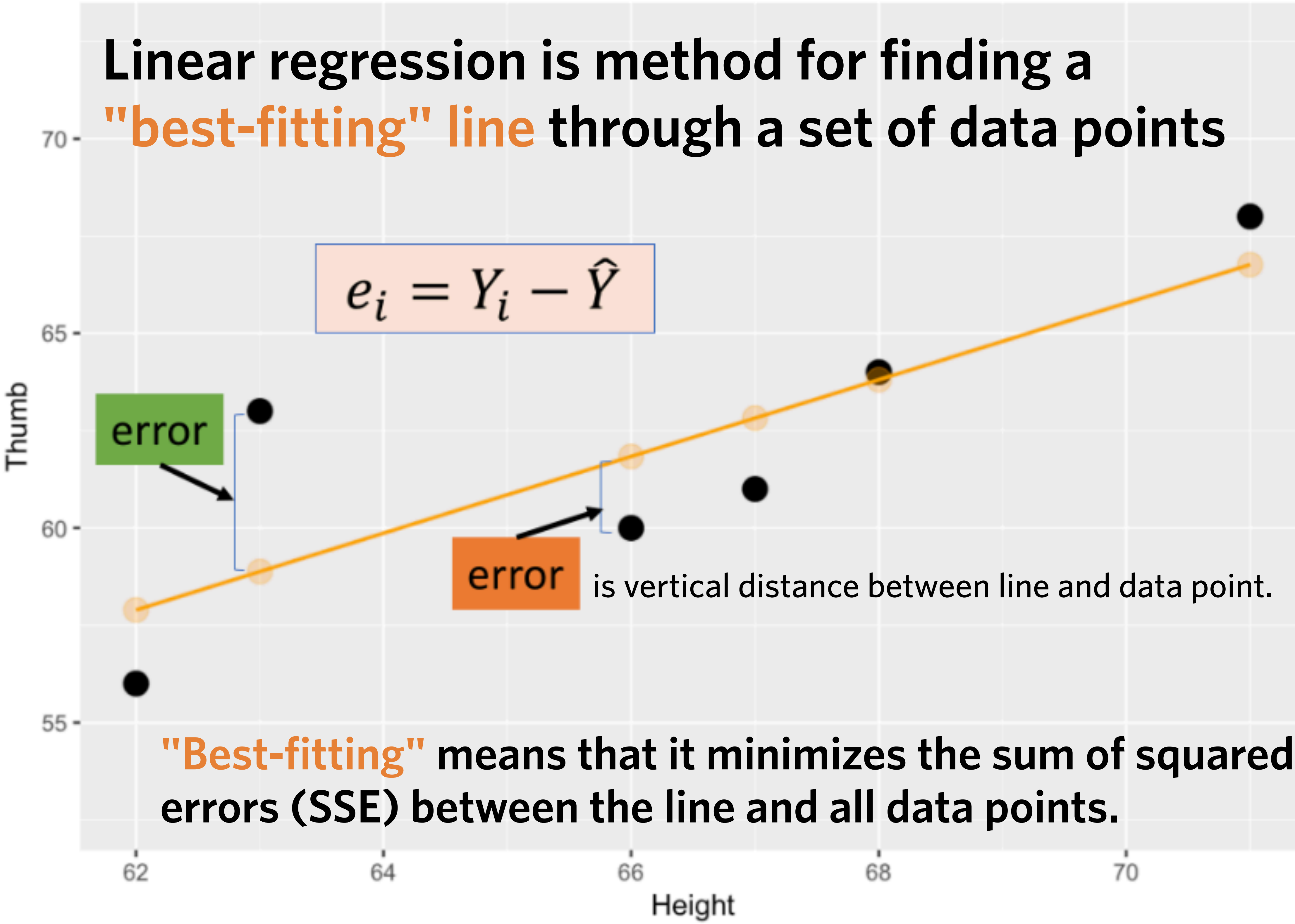
height

1 *Using an explanatory variable to model variation in an outcome variable*



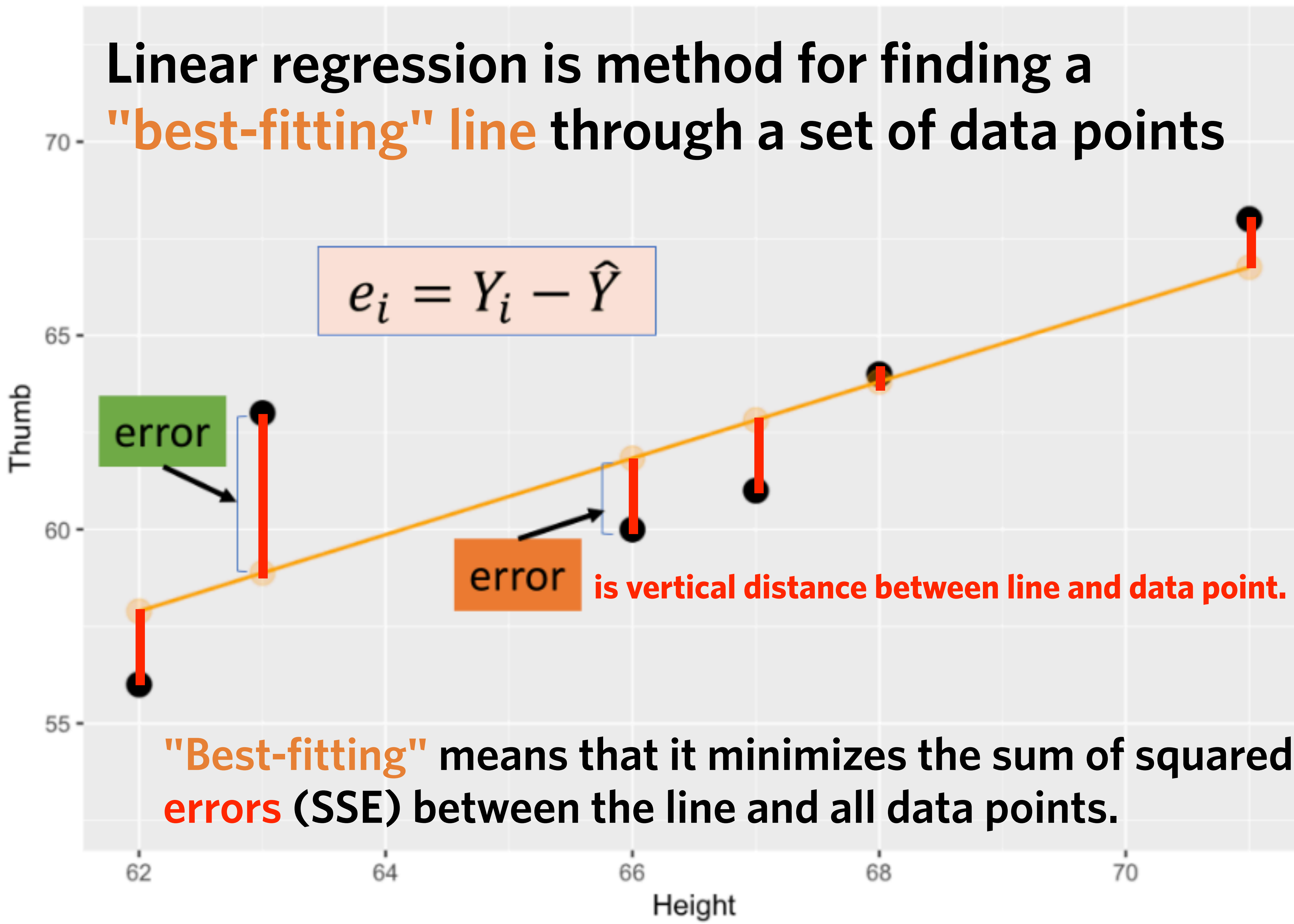
1

Using an explanatory variable to model variation in an outcome variable

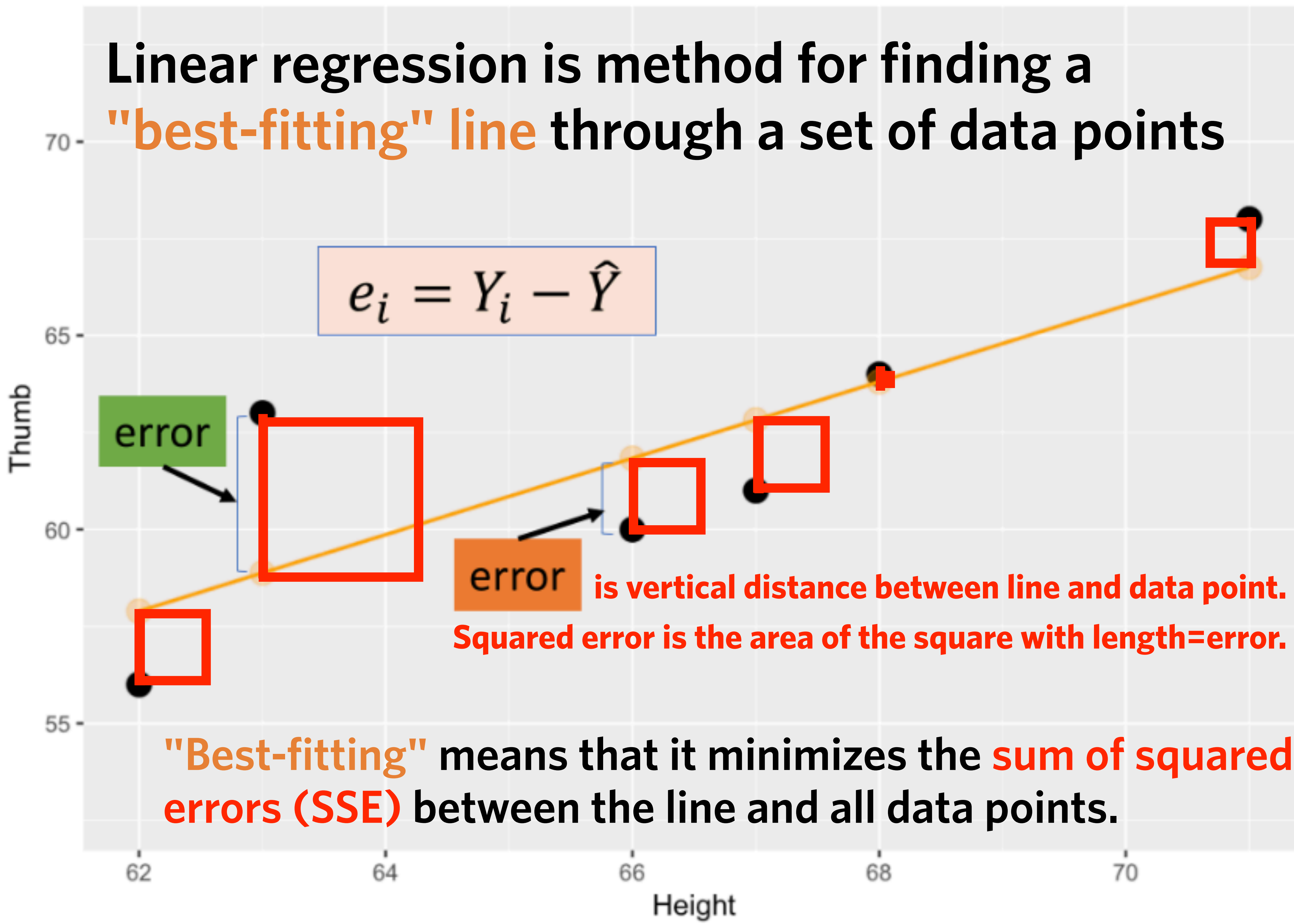


1

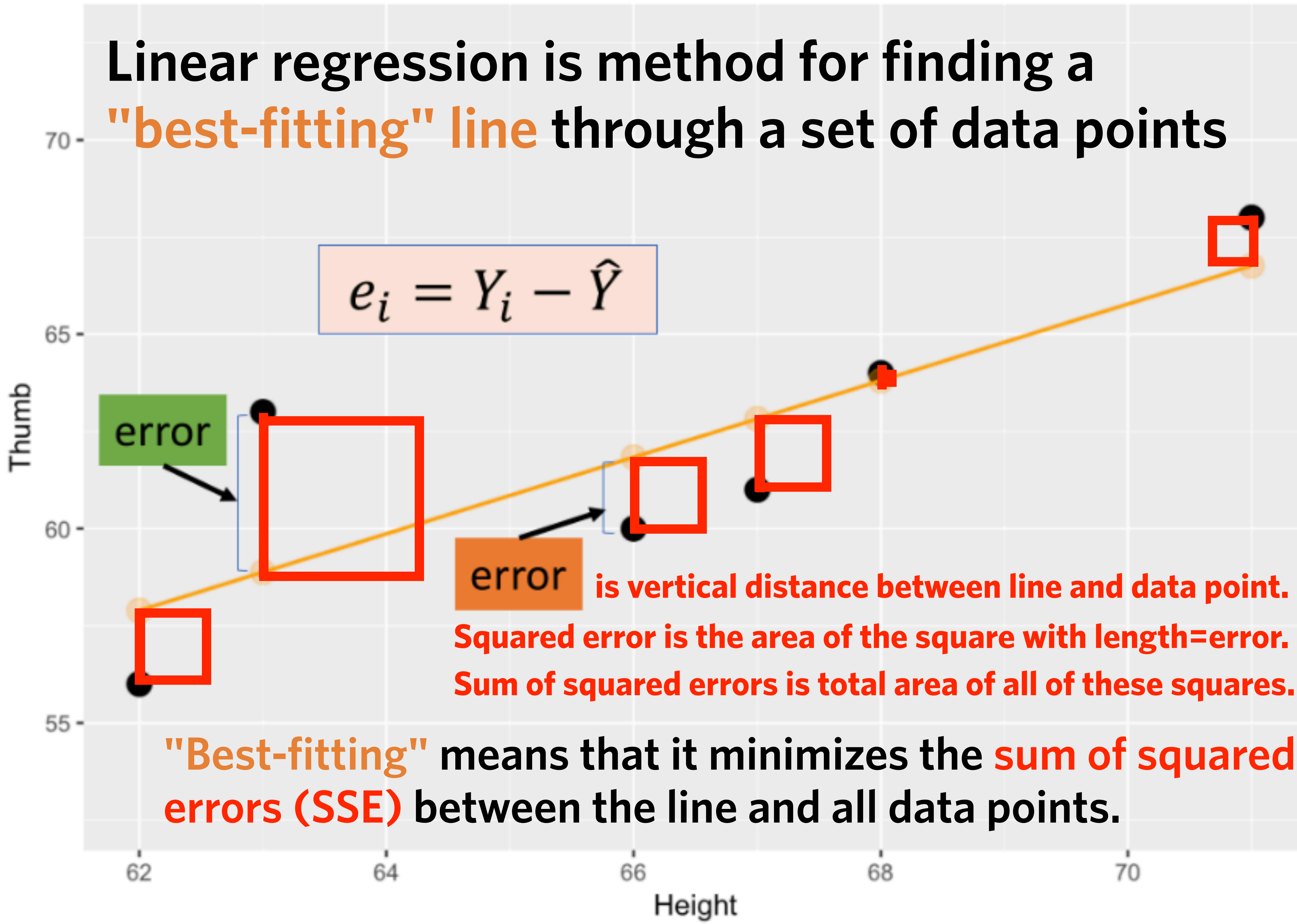
Using an explanatory variable to model variation in an outcome variable



1 *Using an explanatory variable to model variation in an outcome variable*



1 *Using an explanatory variable to model variation in an outcome variable*



1

Using an explanatory variable to model variation in an outcome variable

Symbol	Group Mean Model	Regression Model
Y_i	Person *i* thumb length	Person *i* thumb length
b_0	Mean thumb length for short people (59 in the tiny dataset)	y-intercept for regression line (predicted thumb length when Height = 0)
b_1	Increment between short and tall group means for thumb length (6 in the tiny dataset)	Slope of the regression line (increment in thumb length for each one-inch increase in height)
X_i	Height for person *i* coded as short=0, tall=1	Height for person *i* measured in inches
e_i	Error for person *i* $(Y_i - \hat{Y}_i)$	Error for person *i* $(Y_i - \hat{Y}_i)$

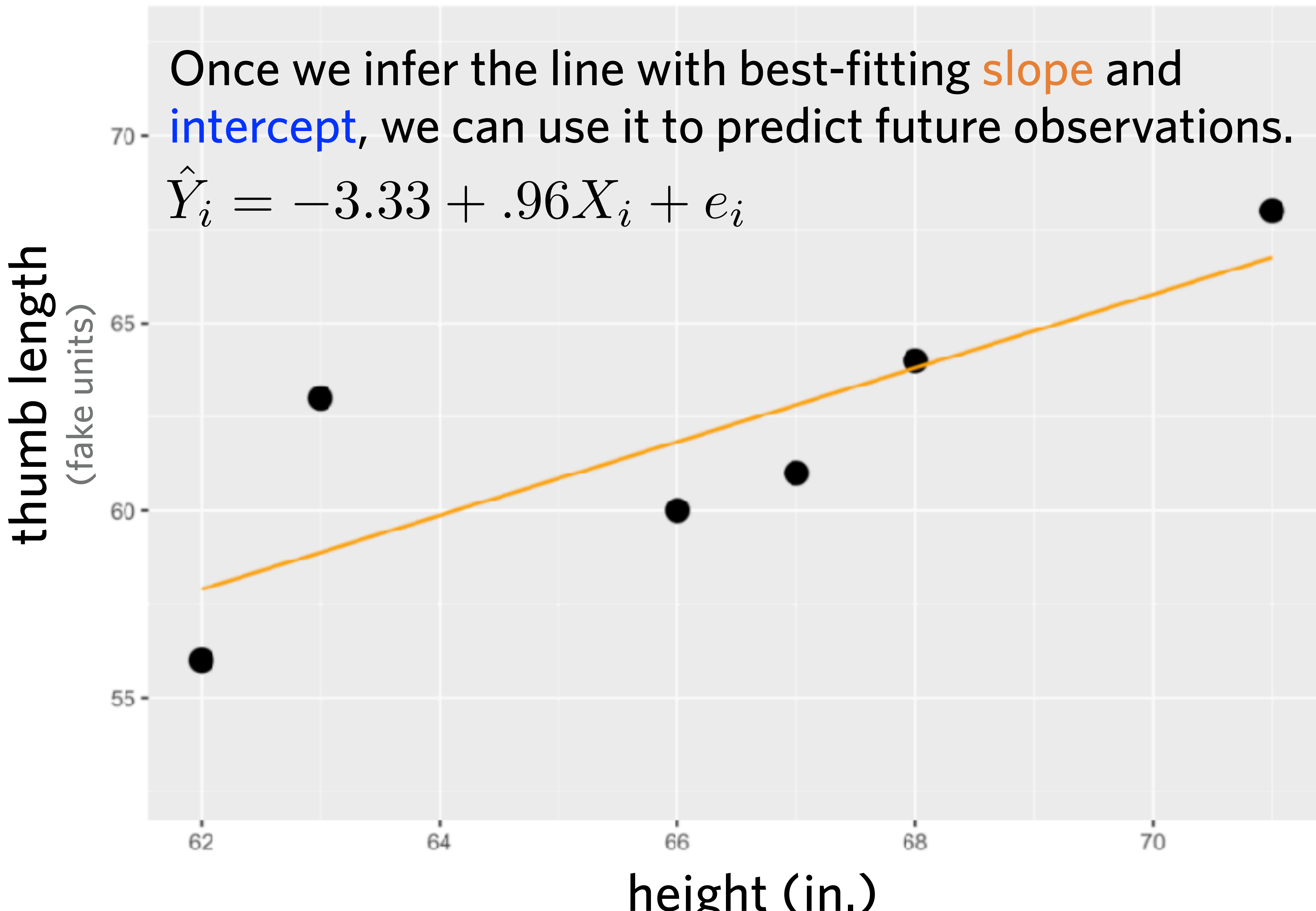
1

Using an explanatory variable to model variation in an outcome variable

Using regression models to make predictions

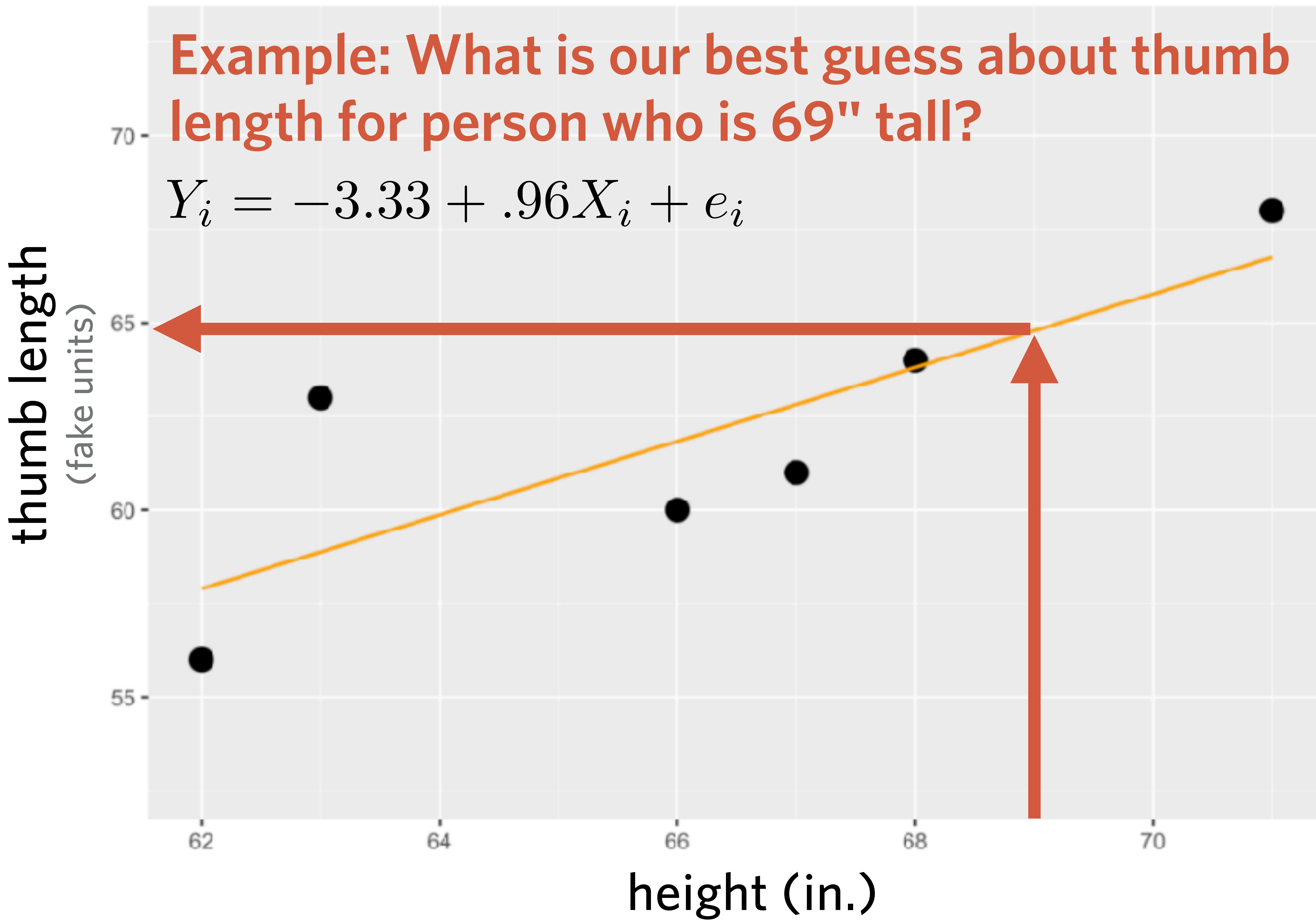
Once we infer the line with best-fitting slope and intercept, we can use it to predict future observations.

$$\hat{Y}_i = -3.33 + .96X_i + e_i$$



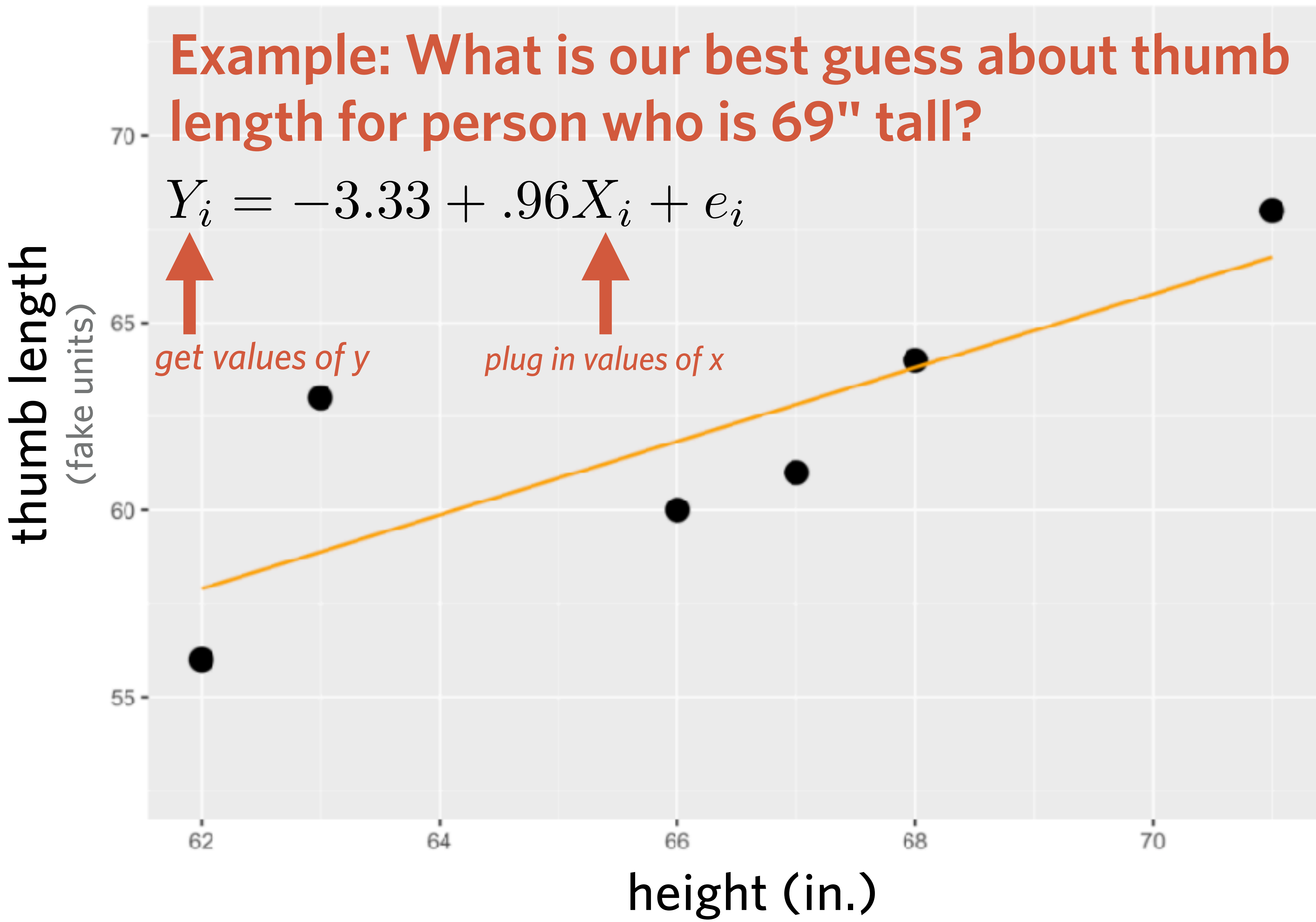
1 *Using an explanatory variable to model variation in an outcome variable*

Using regression models to make predictions



1 *Using an explanatory variable to model variation in an outcome variable*

Using regression models to make predictions



Using an explanatory variable to model variation in an outcome variable

Assessing model fit with sum of squared errors

empty model

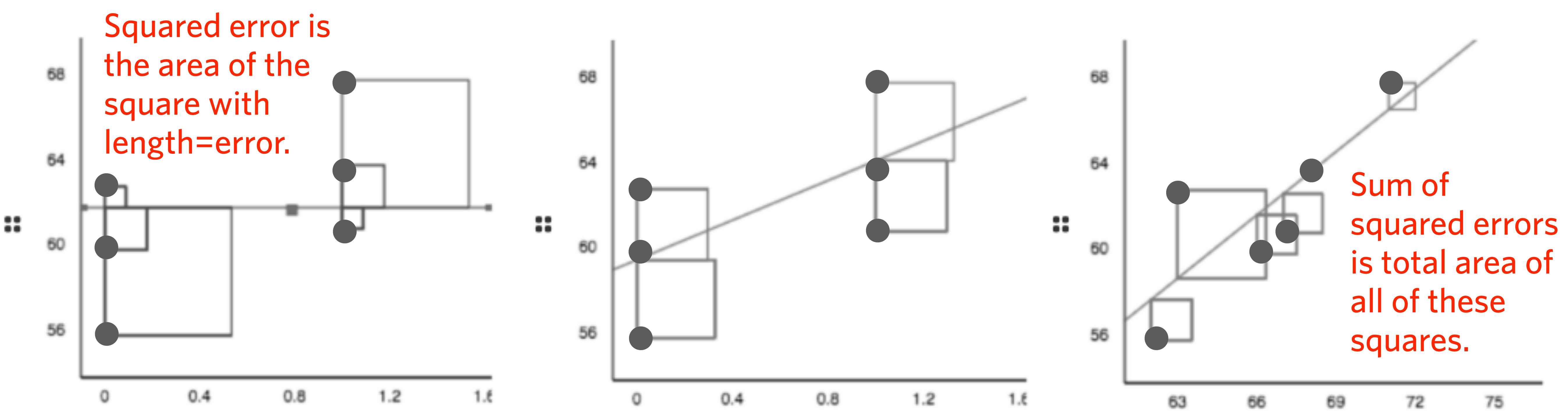
uses grand mean of entire dataset to make predictions

group model

models differences between group means

linear regression model

models continuous variation in outcome variable as a linear function of predictor variable



Remember when we first introduced Sum of Squared Errors (SSE)?

We can use this quantity to evaluate how well different models fit the data. The lower the SSE, the better the fit.

1

Using an explanatory variable to model variation in an outcome variable

Assessing model fit with sum of squares

We can use the same F-statistic we used for ANOVAs to compare how much more variation one model explains than another (simpler version of that) model

$$F = \frac{\frac{SSE_{null} - SSE_{linear}}{k_{linear} - k_{null}}}{\frac{SSE_{linear}}{n - k_{linear}}}$$

The more variation explained by linear regression model, the smaller the SSE_{linear}, and thus the larger the F stat.

$\boldsymbol{k}$ is number of fitted parameters.

For the null model, the only parameter is the mean.
For the linear model, the parameters are slope & intercept.

```
This is computed for you when you use the lm() function in R.
```

Today

Mini-Review Session #3

1 → 2 → 3

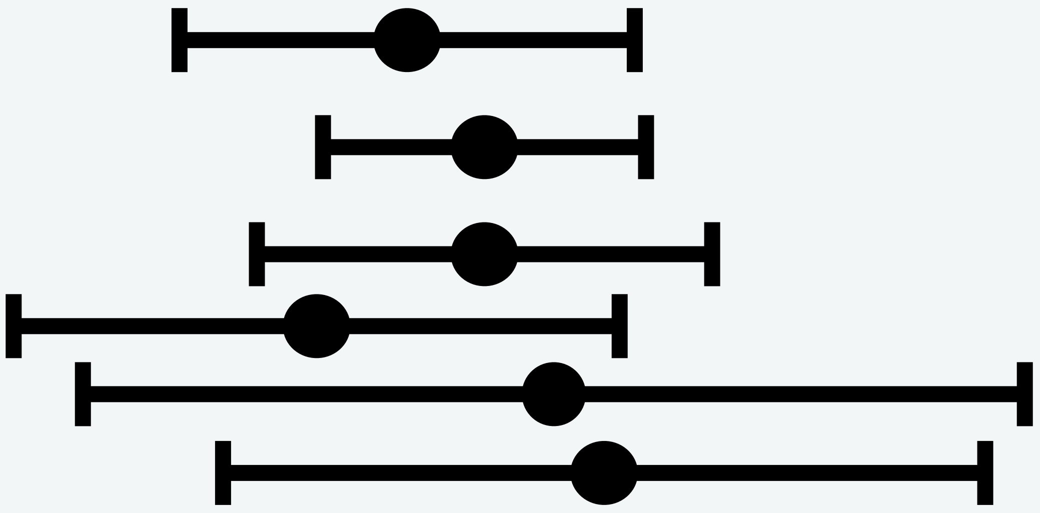
1. *Using an explanatory variable to model variation in an outcome variable*
2. *Quantifying effects using confidence intervals*
3. *What is the (Pearson) correlation coefficient?*

2 *Quantifying effects using confidence intervals*

Confidence Intervals

- A confidence interval is an interval (defined by a lower bound and an upper bound) that will contain the true population parameter (e.g., y-intercept, slope) with a given probability.
- For example, the **95% confidence interval** for the mean is an interval that will capture the true population mean 95% of the time.
- Any particular confidence interval either does or does not contain the true parameter.
- But in the long run, if we imagine repeating the study many times and constructing a 95% CI each time, these CIs will contain the true population mean 95% of the time.

95% confidence interval is about capturing the true population parameter in the long term across many samples



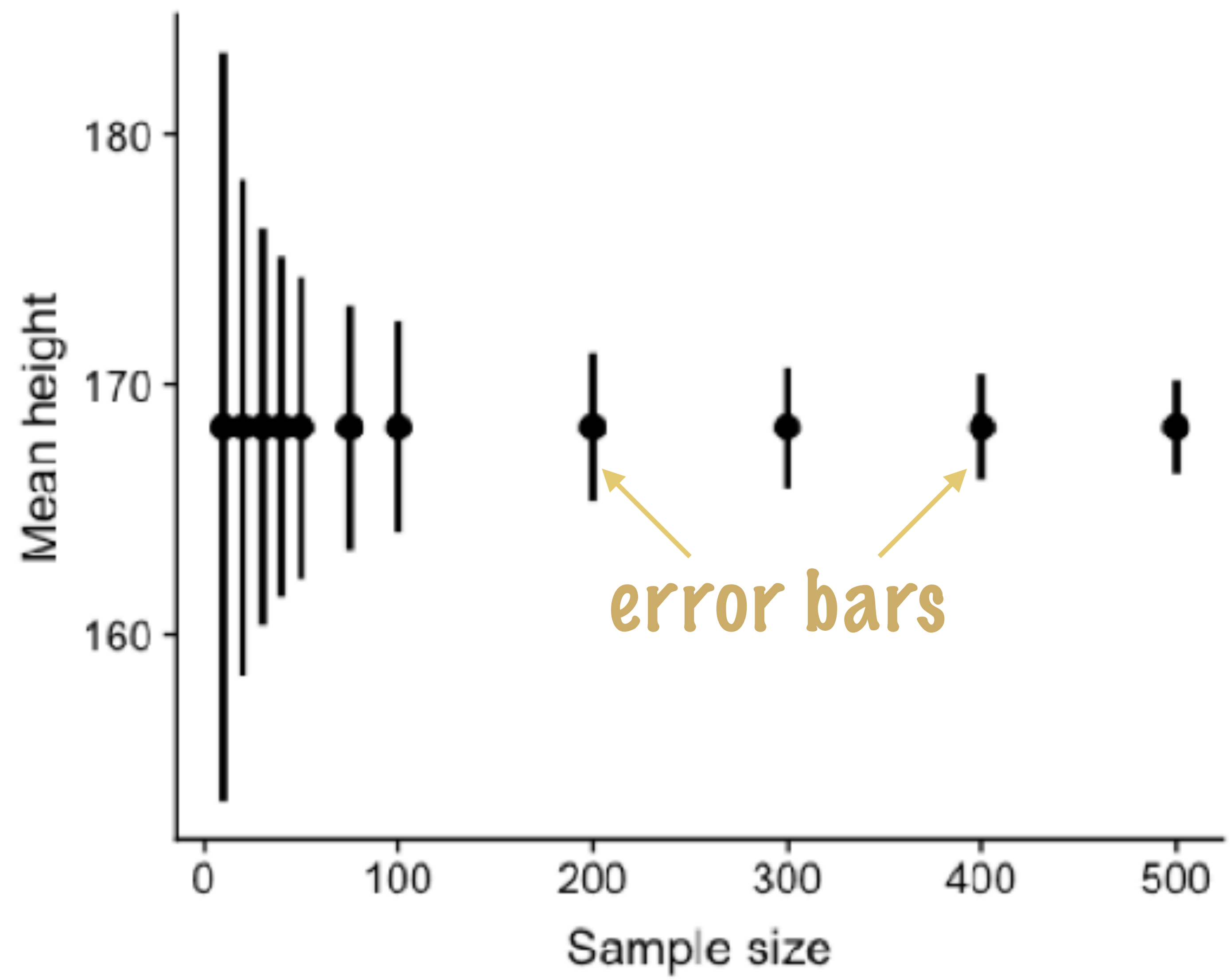
2

Quantifying effects using confidence intervals

Confidence interval length depends on sample size

- Because the standard error decreases with sample size, the CI gets narrower as the sample size increases

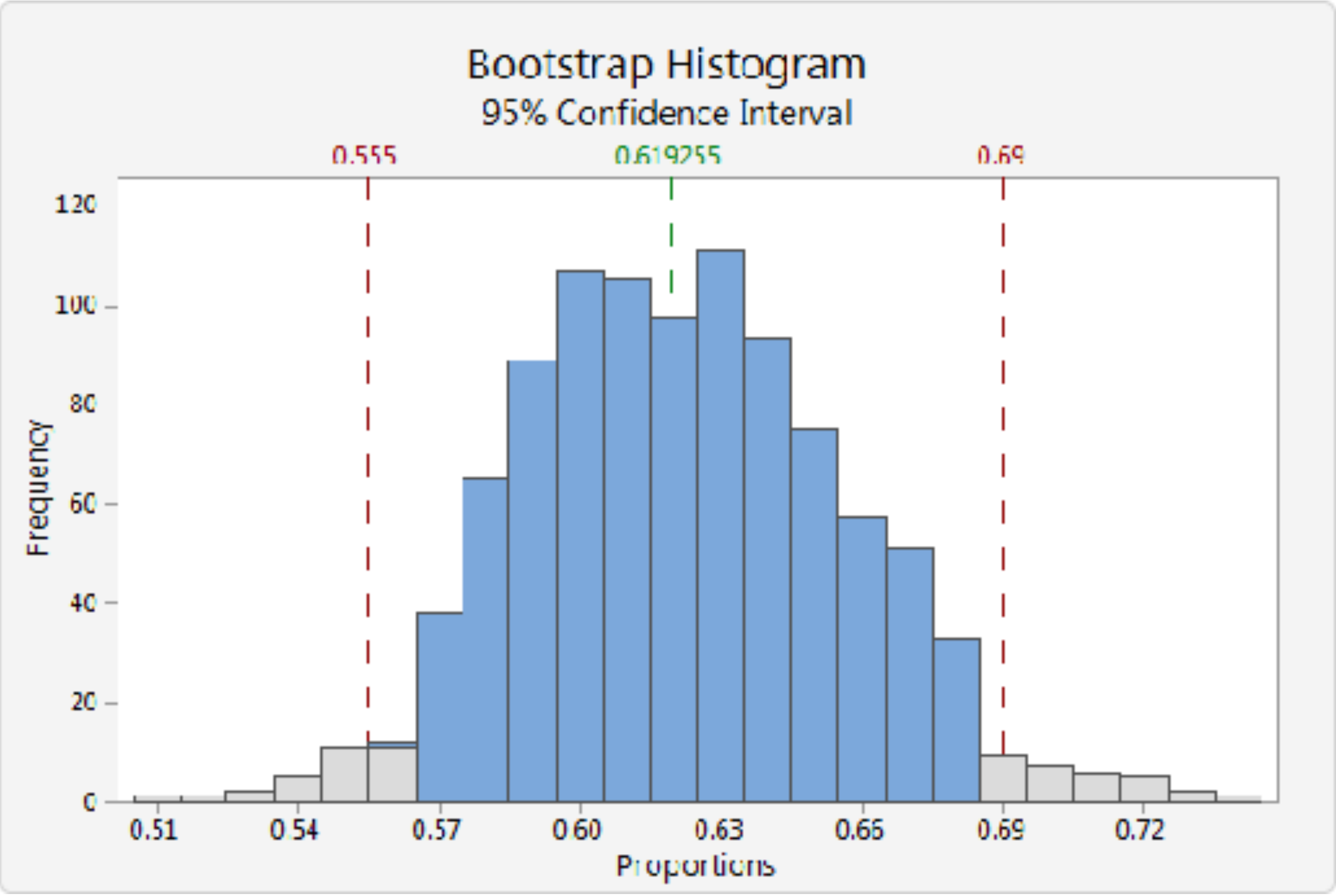
- The confidence interval becomes increasingly tighter as the sample size increases, but increasing samples provide diminishing returns (just like SEM b/c we are taking the square root of sample size).



An example of the effect of sample size on the width of the confidence interval for the mean.

Constructing confidence intervals by resampling

- Repeatedly resampling the data with replacement (a.k.a. "bootstrap resampling") and computing your statistic of interest (e.g., mean, b1) is a way of generating a **sampling distribution** of that statistic.
- The lower bound of a 95% CI is the 2.5th percentile of the sampling distribution, and the upper bound of a 95% CI is the 97.5th percentile of the sampling distribution.



Confidence intervals

- Bottom line: confidence intervals are extremely useful for simultaneously communicating effect size (i.e., the estimate of mean) and uncertainty in your estimate (i.e., due to sampling variability).
- It is good practice to report 95% CIs in addition to test-statistics and p-values when reporting the findings of your statistical analyses.

Today

Mini-Review Session #3

1 → 2 → 3

1. *Using an explanatory variable to model variation in an outcome variable*
2. *Quantifying effects using confidence intervals*
3. *What is the (Pearson) correlation coefficient?*

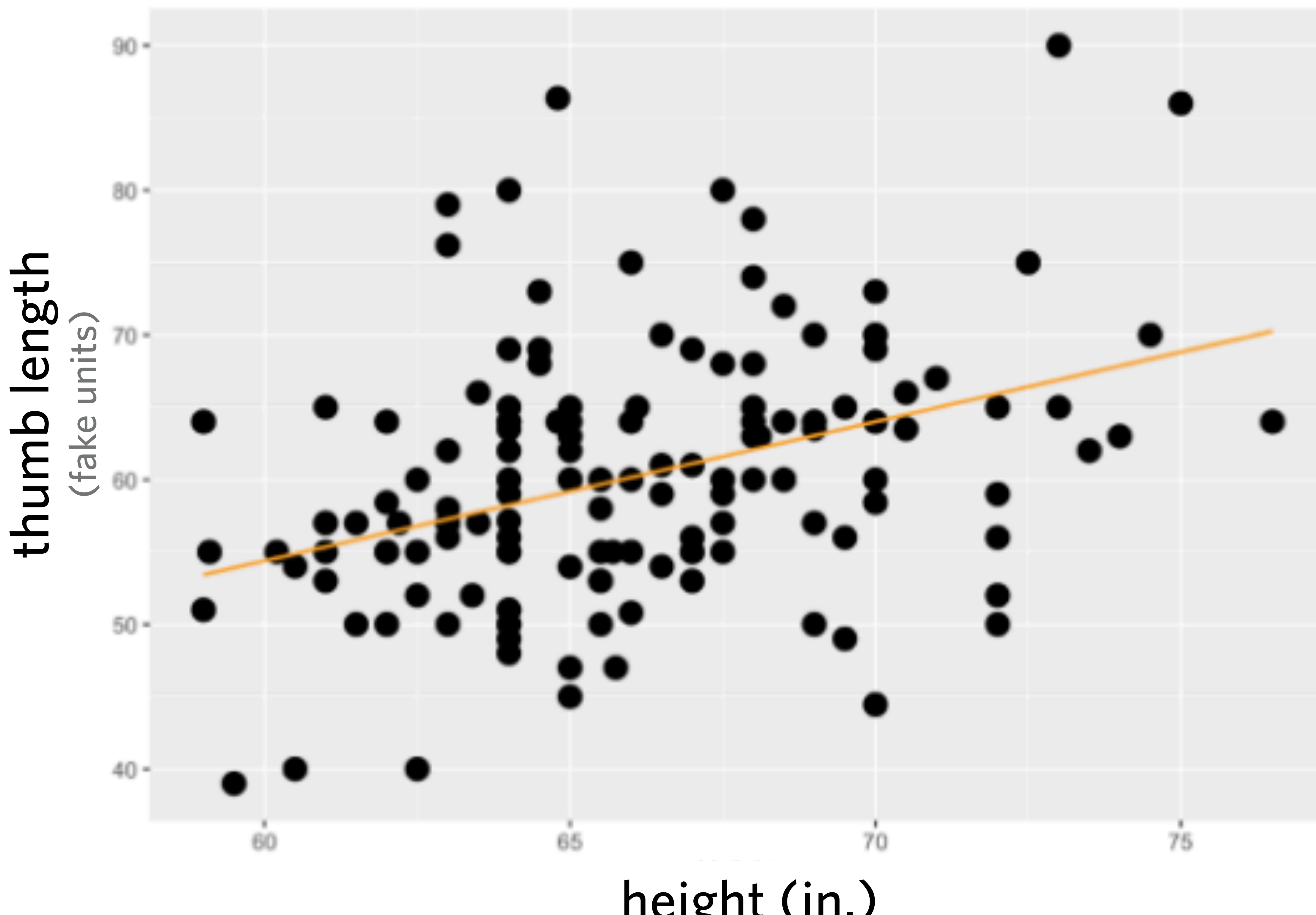
3

What is the (Pearson) correlation coefficient?

What does the word "correlation" mean to you?

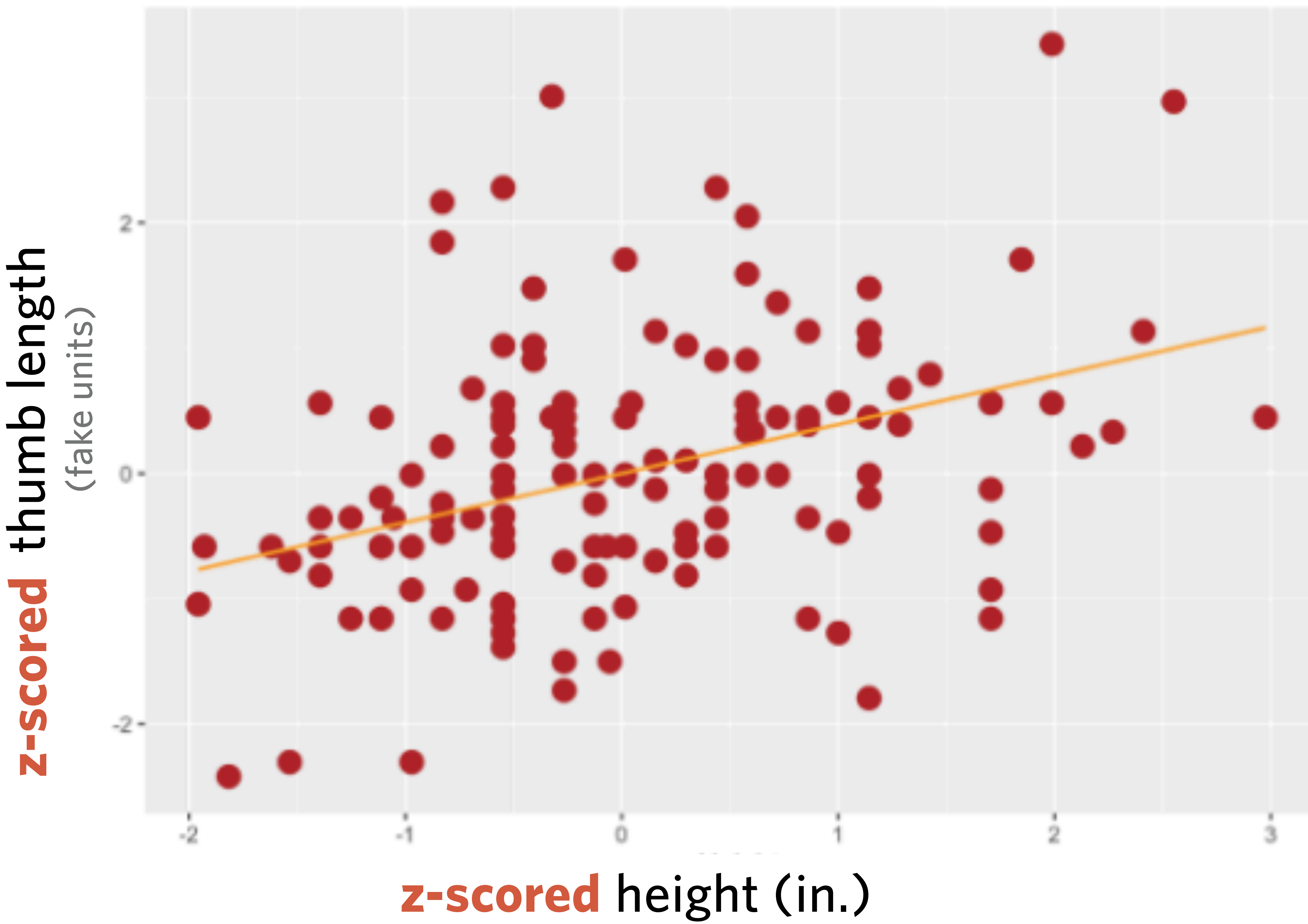
3 *What is the (Pearson) correlation coefficient?*

Pearson's correlation coefficient



3 *What is the (Pearson) correlation coefficient?*

Pearson's correlation coefficient

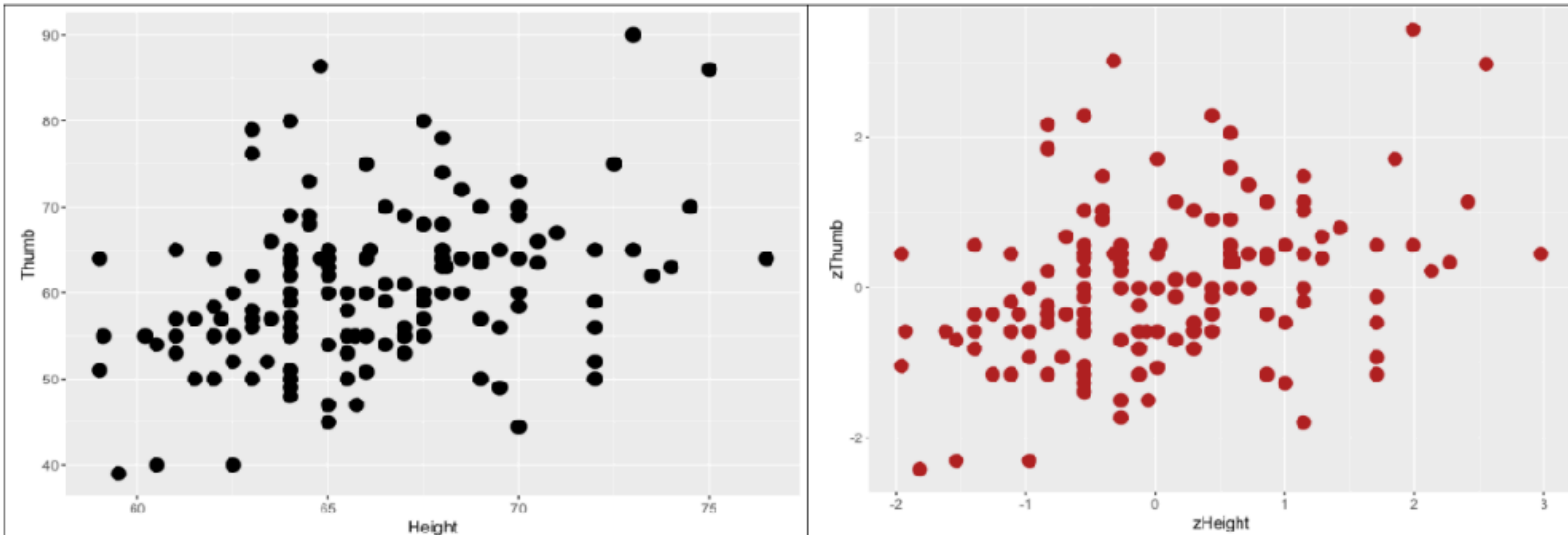


3

What is the (Pearson) correlation coefficient?

Pearson's correlation coefficient

Compare these two scatter plots. How are they similar? How are they different?



Thumb by Height	zThumb by zHeight

Calculating Pearson's correlation coefficient

- Variance for a single variable

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

- Covariance between two variables:

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

3 *What is the (Pearson) correlation coefficient?*

Calculating Pearson's correlation coefficient

- Pearson's correlation coefficient (*r*) scales the covariance so that it has a standard scale (ranging between -1 and +1).

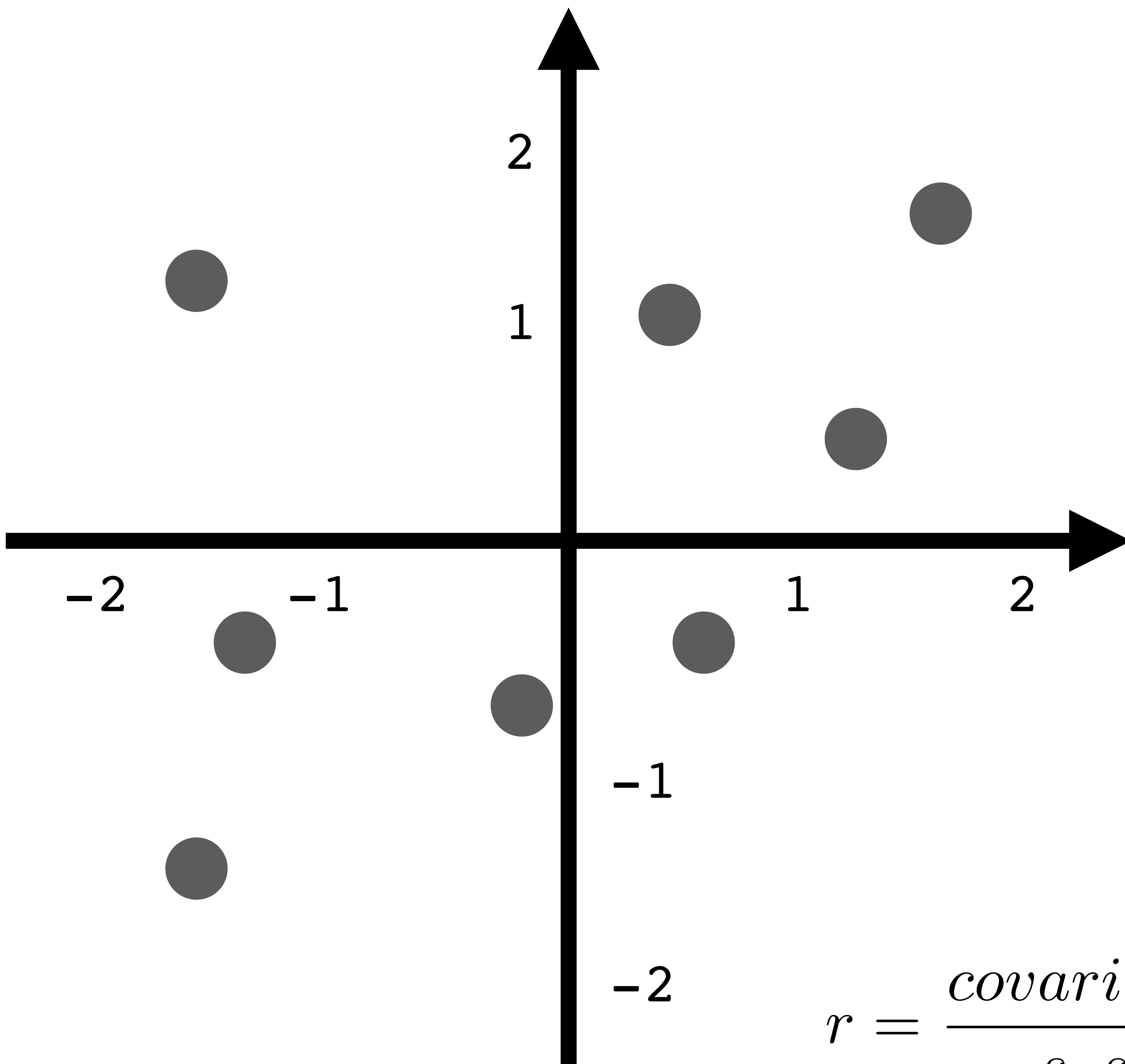
$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

- Pearson's correlation coefficient (*r*) measures the covariance between z-scored data (since the std deviation of z-scored data is 1)

3 *What is the (Pearson) correlation coefficient?*

Calculating Pearson's correlation coefficient



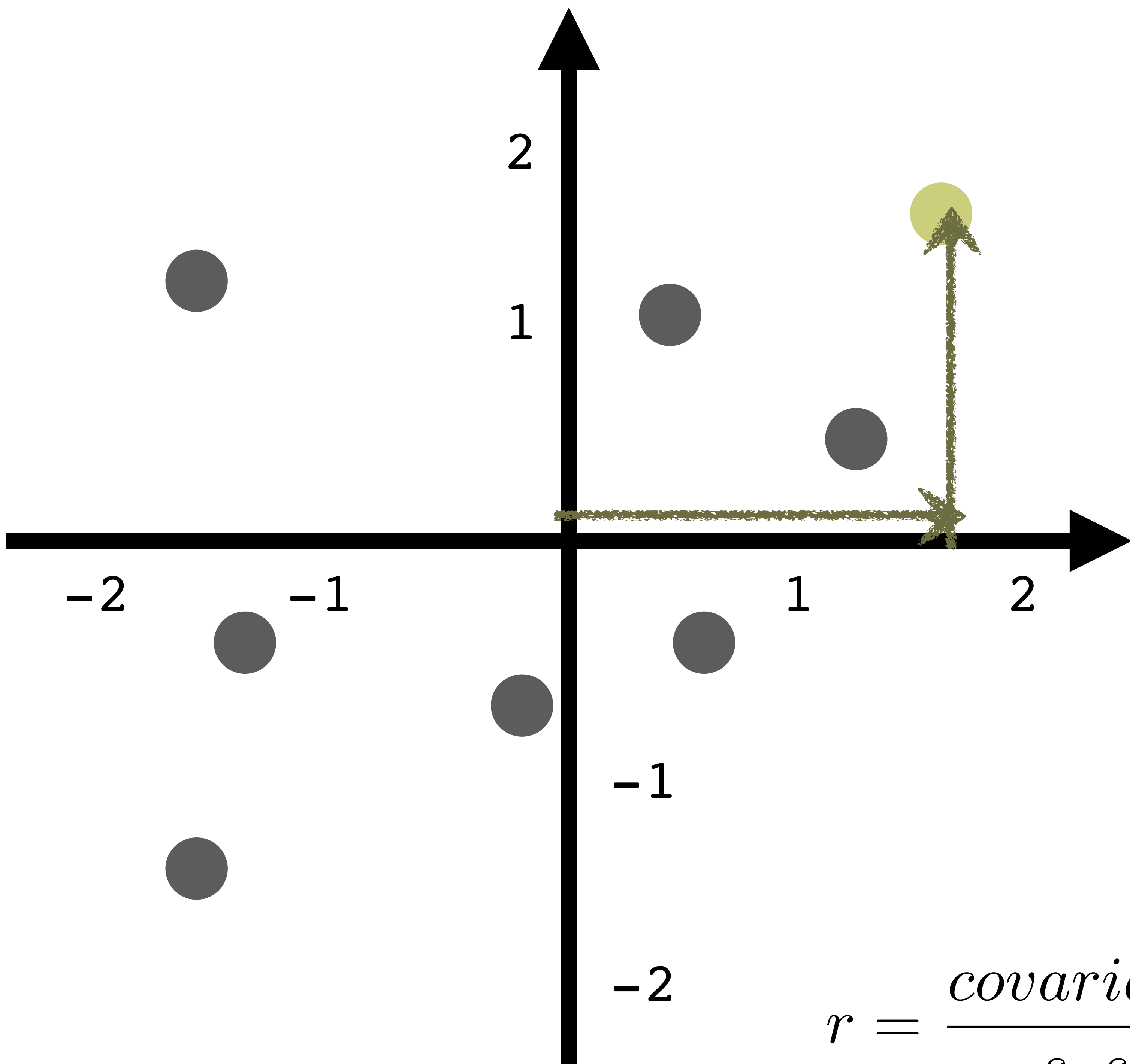
- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
-

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

3

What is the (Pearson) correlation coefficient?

Calculating Pearson's correlation coefficient



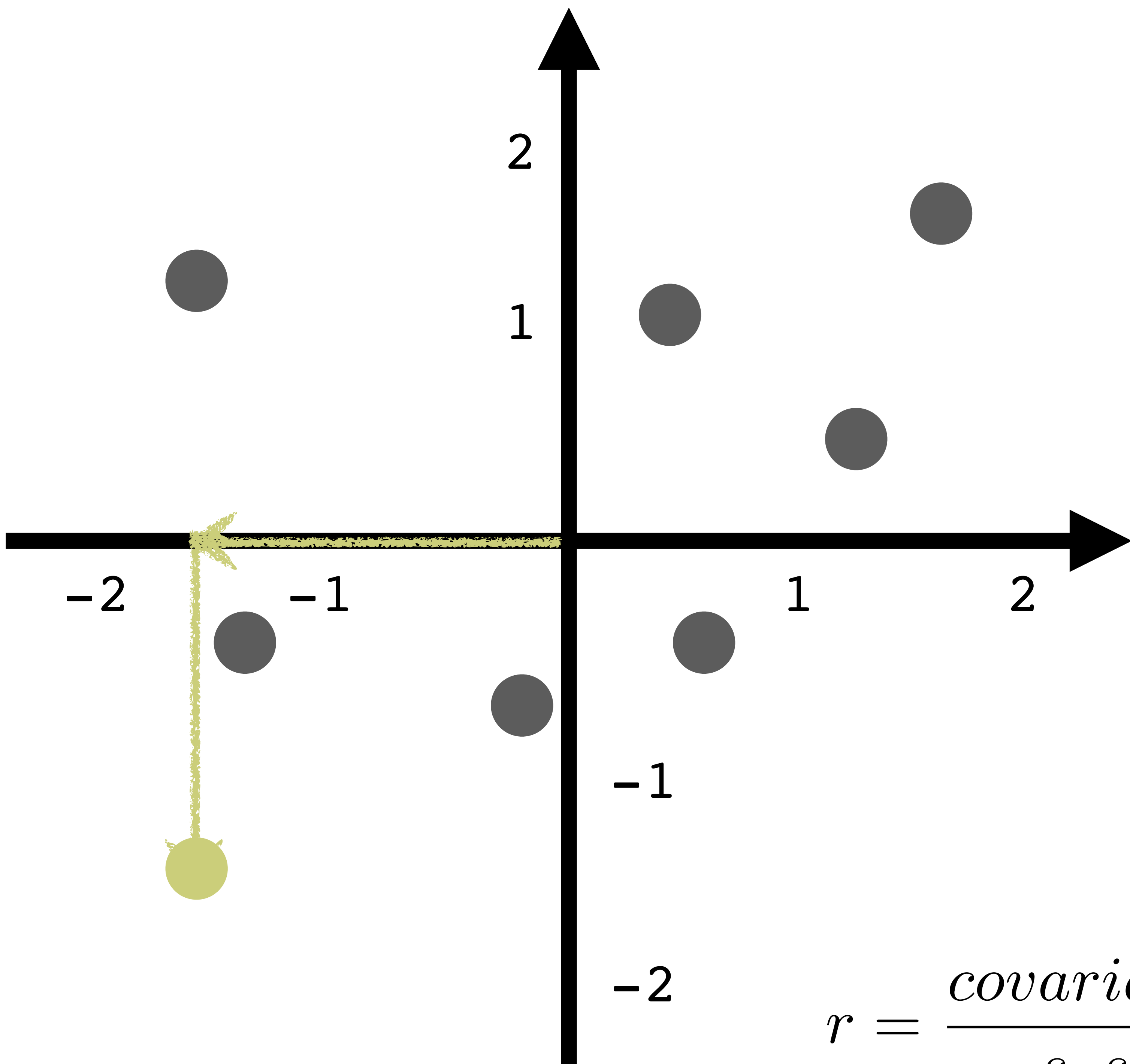
- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- A data point that is in the top right of this z-scored scatter plot will increase Pearson's r.
  - Its x-value > mean(x) and y-value > mean(y).

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3

What is the (Pearson) correlation coefficient?

Calculating Pearson's correlation coefficient



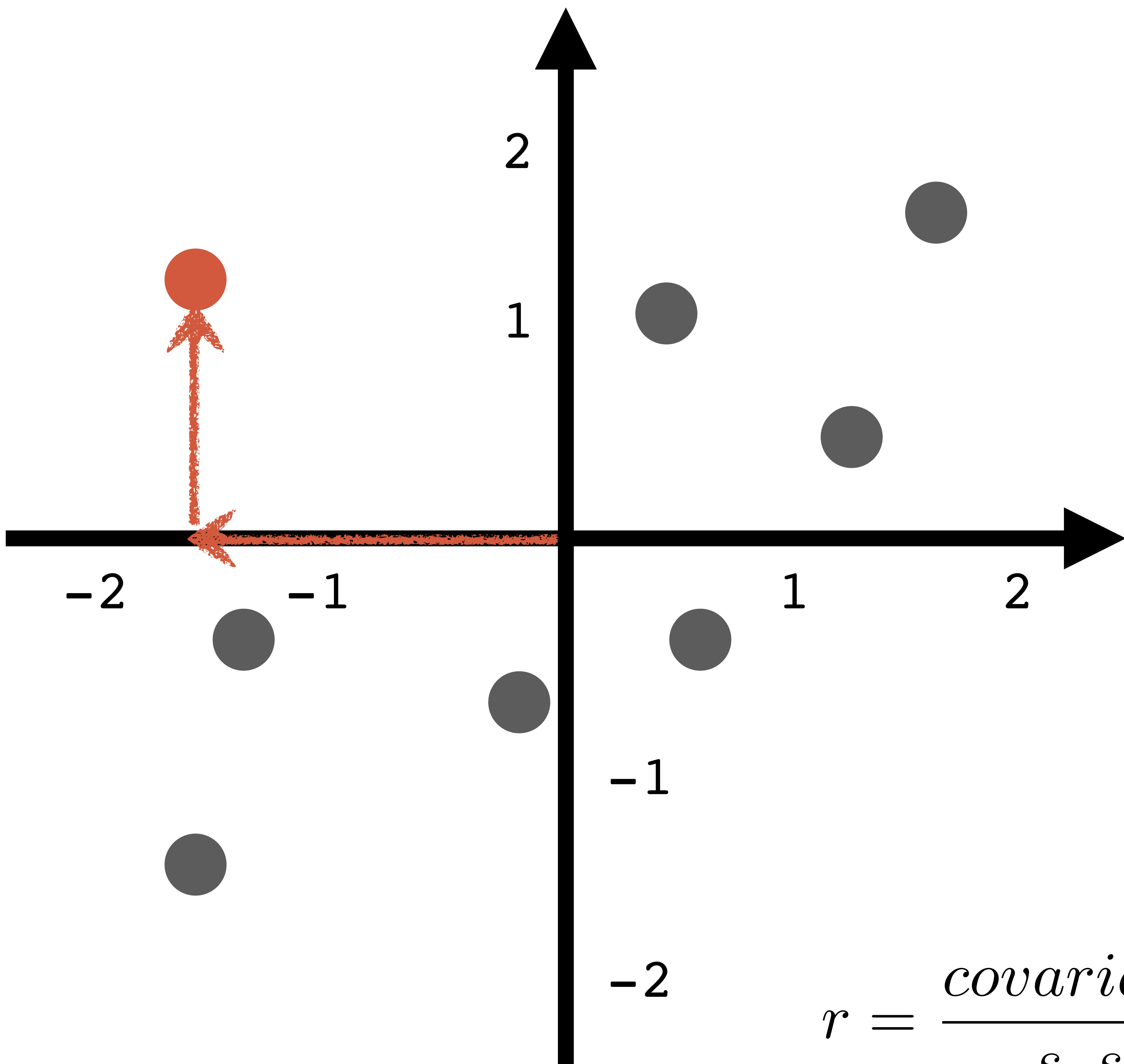
- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- A data point that is in the top right of this z-scored scatter plot will increase Pearson's r.
  - Its x-value > mean(x) and y-value > mean(y).
- A data point in the bottom left will also increase Pearson's r.
  - Its x-value < mean(x) and y-value < mean(y), so product is positive.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3

What is the (Pearson) correlation coefficient?

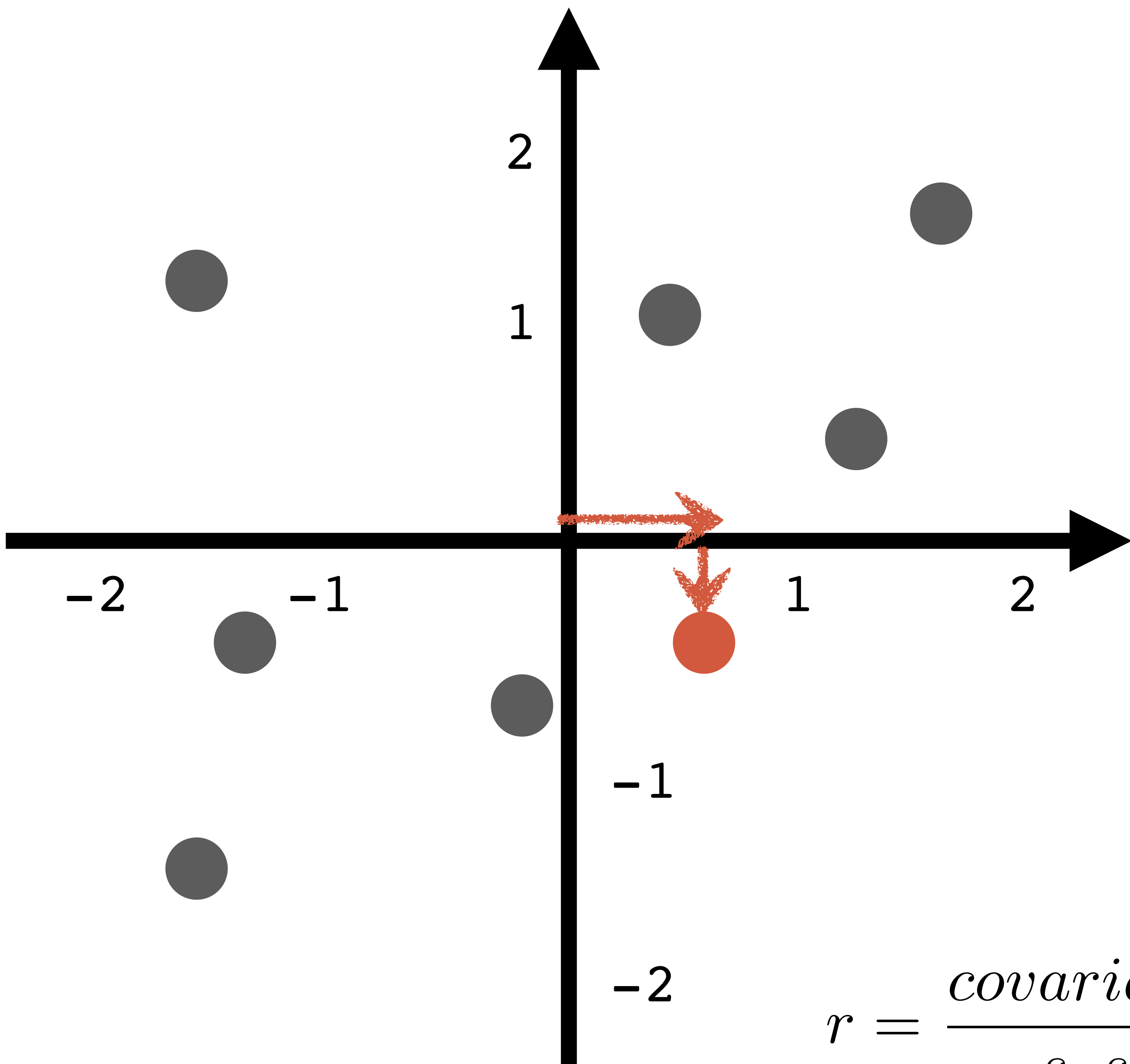
Calculating Pearson's correlation coefficient



- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- BUT a data point in the top left will *decrease* Pearson's *r*.
  - Its x-value < mean(x) but its y-value > mean(y), so product is negative.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Calculating Pearson's correlation coefficient

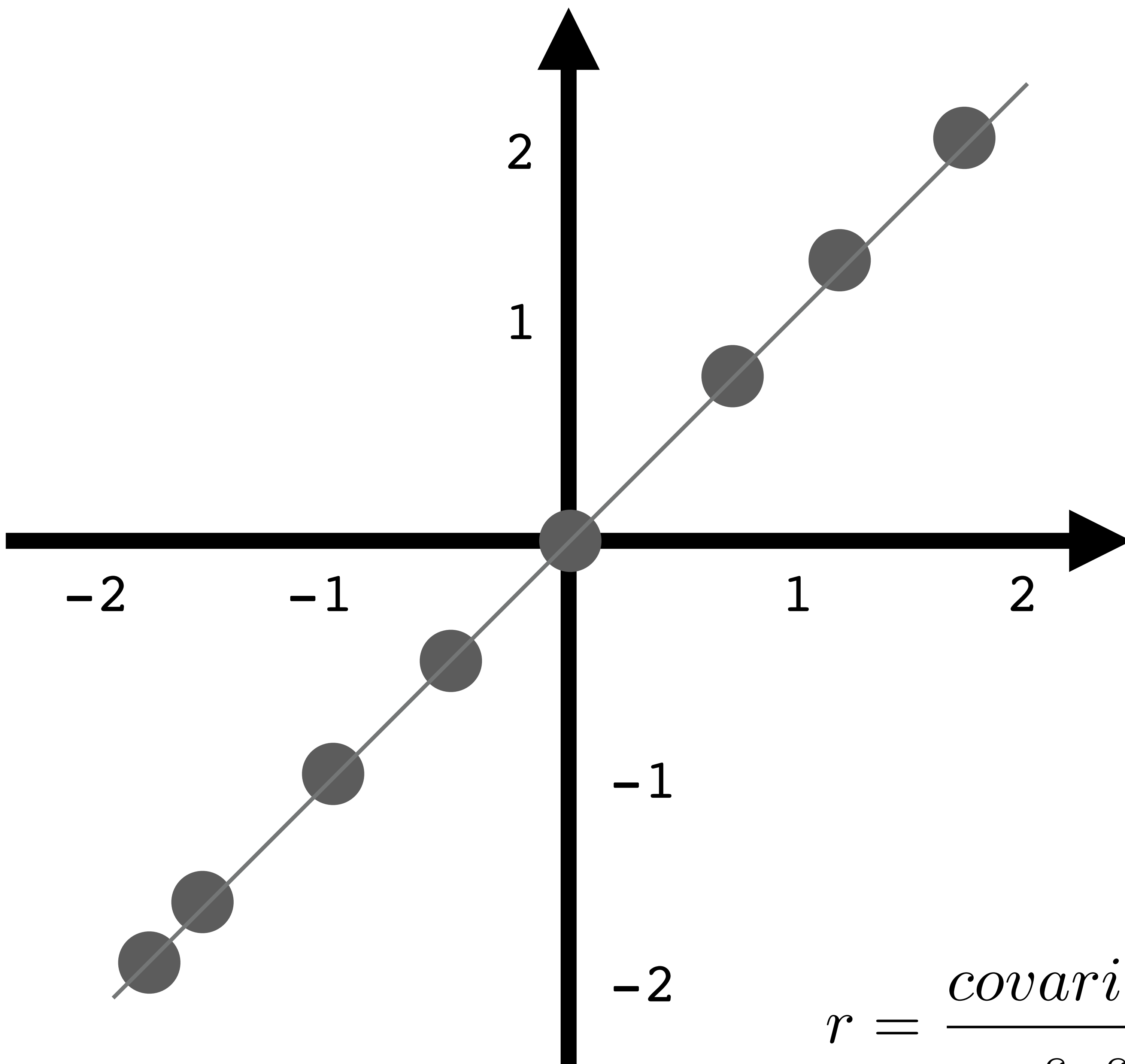


- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- BUT a data point in the top left will *decrease* Pearson's *r*.
  - Its x-value < mean(x) but its y-value > mean(y), so product is negative.
- Same goes for a data point in the bottom right.
  - Its x-value > mean(x) but its y-value < mean(y), so product is negative.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3 *What is the (Pearson) correlation coefficient?*

Calculating Pearson's correlation coefficient



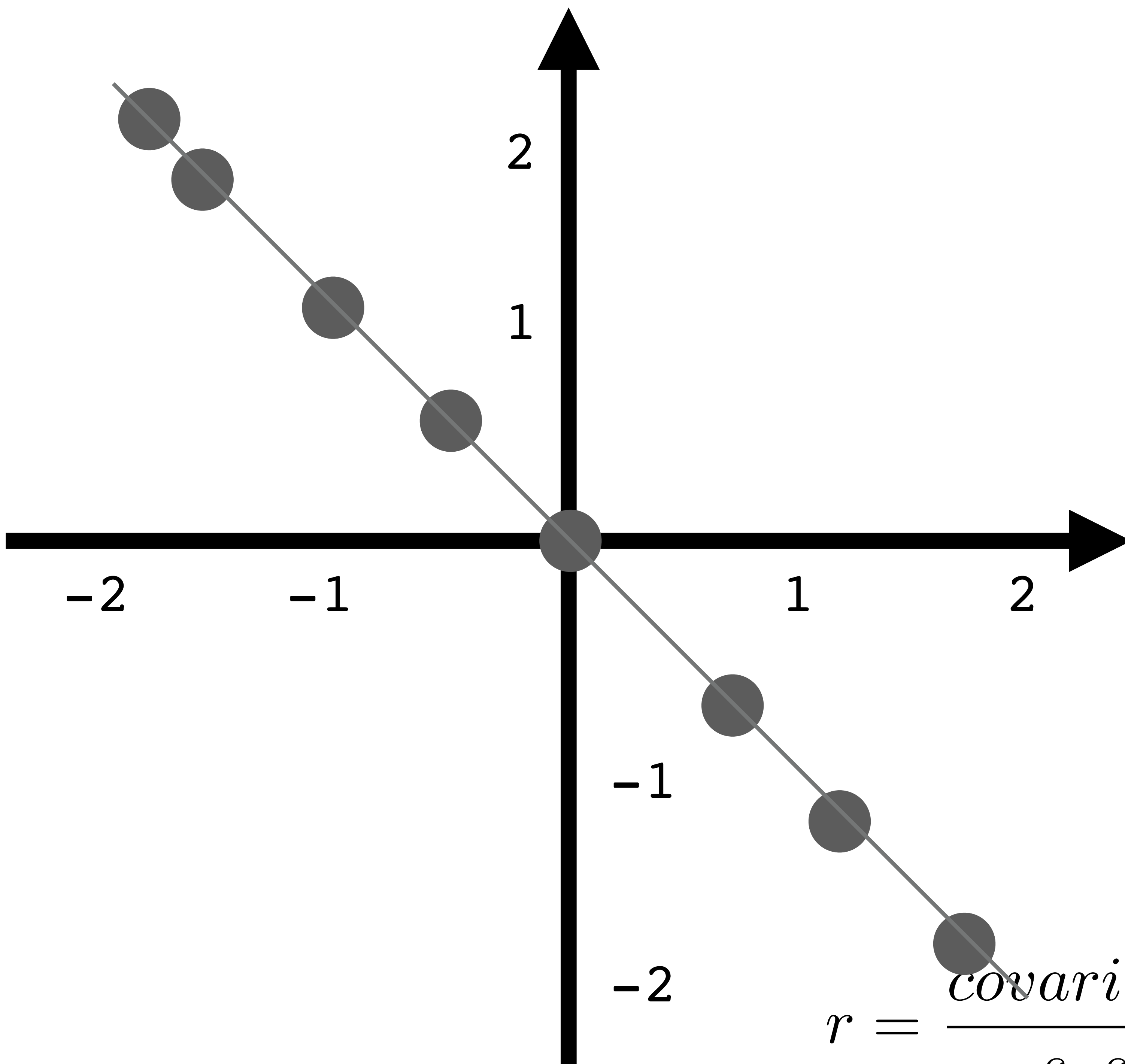
- Pearson's r is **maximized** ($r = +1$) when all data points fall perfectly onto a straight line with positive slope.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3

What is the (Pearson) correlation coefficient?

Calculating Pearson's correlation coefficient

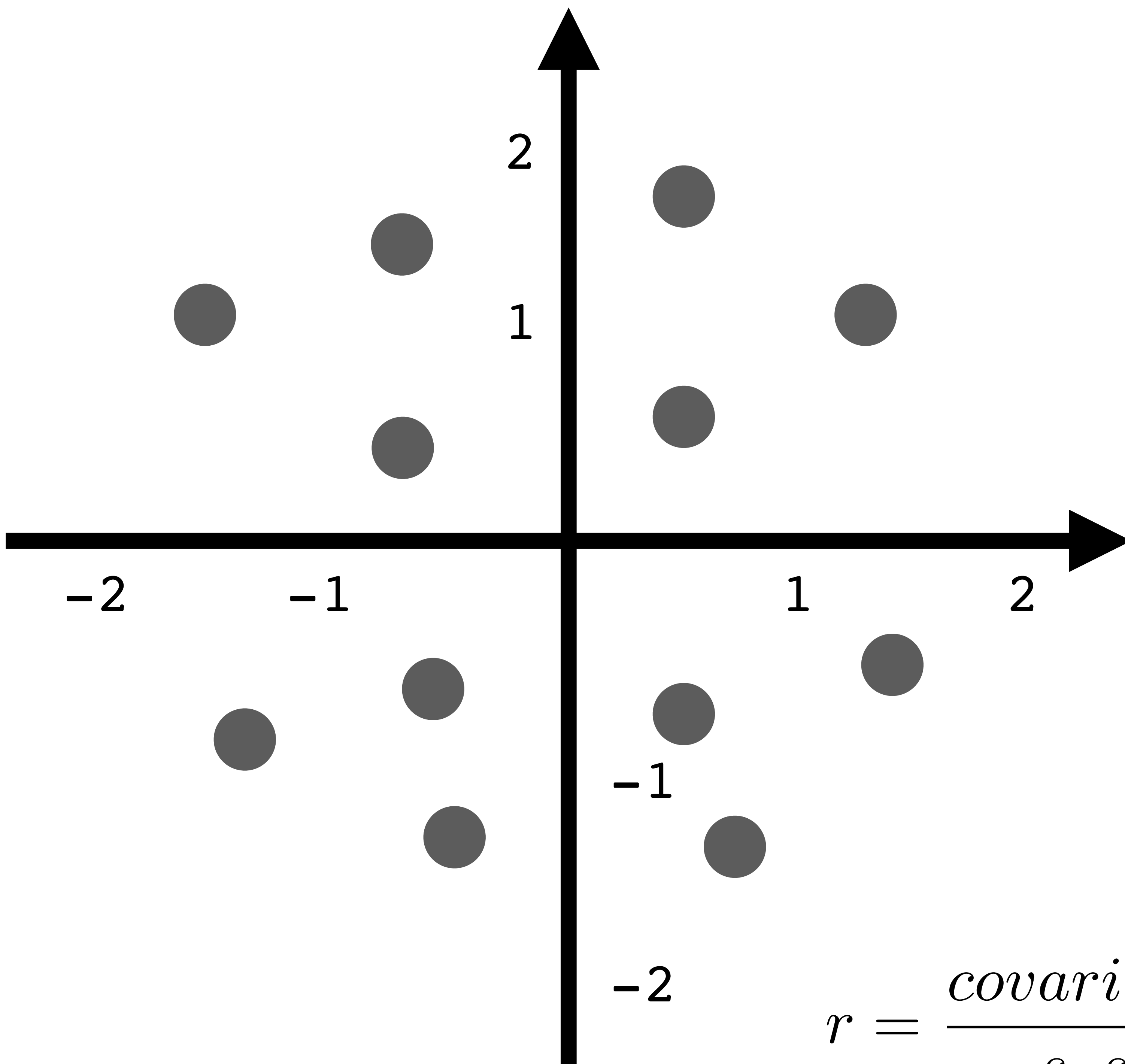


- Pearson's r is **maximized** (*r* = +1) when all data points fall perfectly onto a straight line with positive slope.
- Pearson's r is **minimized** (*r* = -1) when all data points fall perfectly onto a straight line with negative slope.

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3 *What is the (Pearson) correlation coefficient?*

Calculating Pearson's correlation coefficient

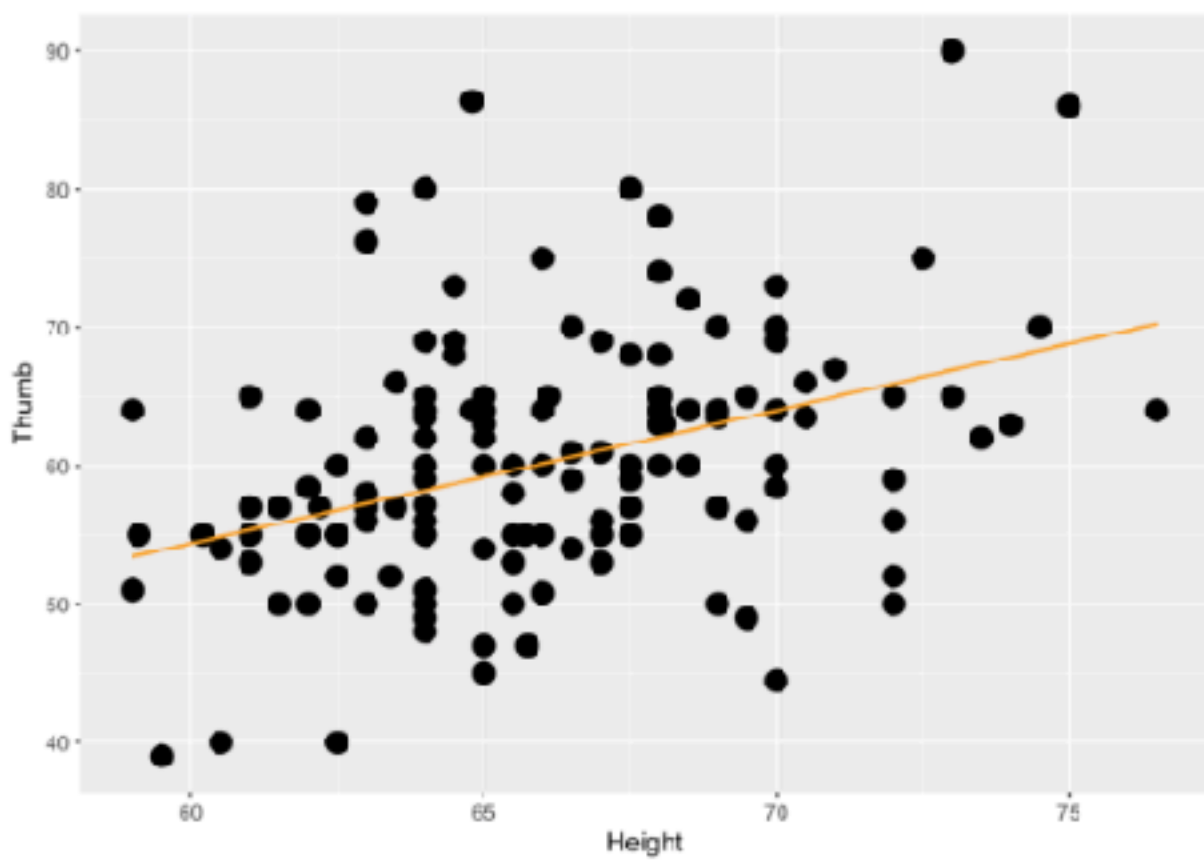


- Pearson's r is **maximized** (r = +1) when all data points fall perfectly onto a straight line with positive slope.
- Pearson's r is **minimized** (r = -1) when all data points fall perfectly onto a straight line with negative slope.
- Pearson's r is **zero** (r=0) when data points form a ball-shaped cloud with no apparent tendency toward positive or negative slope.

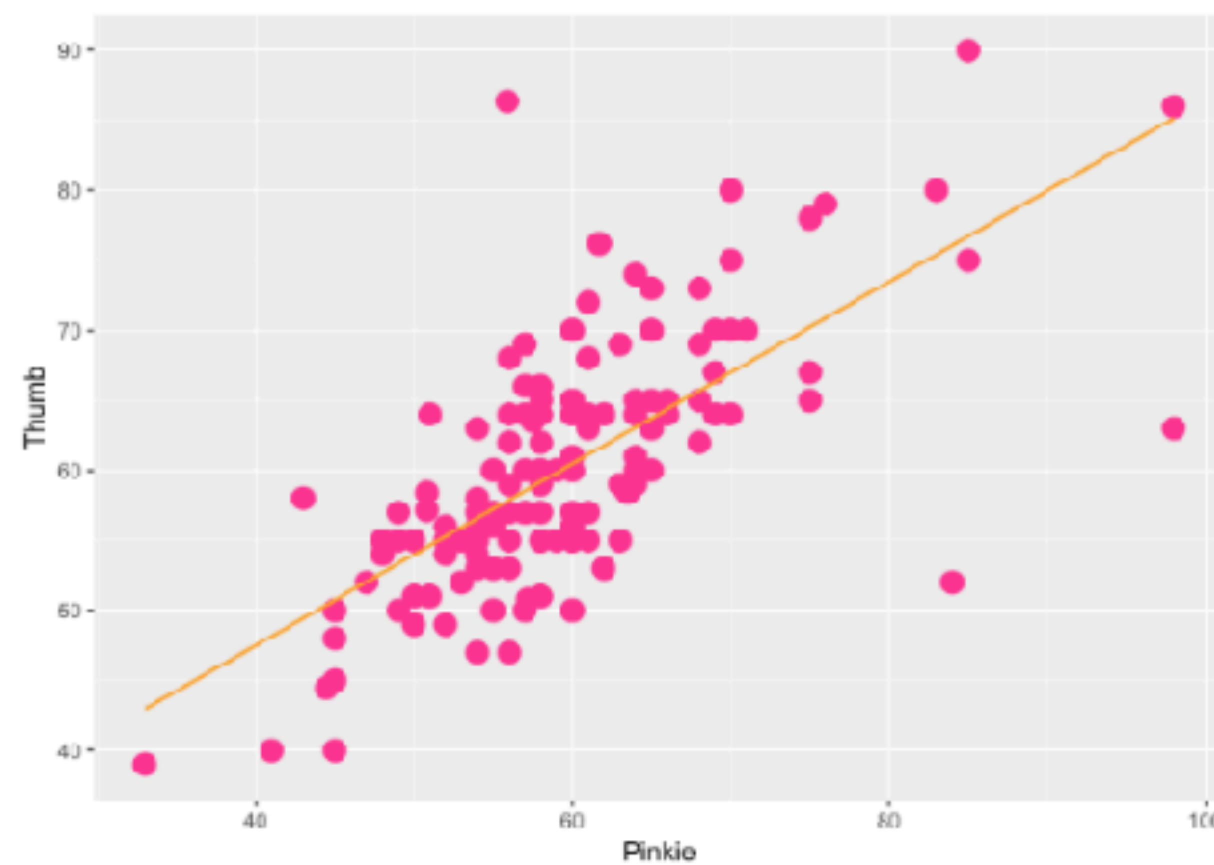
$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

3 *What is the (Pearson) correlation coefficient?*

Pearson's correlation coefficient



Thumb by Height



Thumb by Pinkie

"Correlation between thumb and pinkie is stronger than correlation between thumb and height."

3

What is the (Pearson) correlation coefficient?

How is Pearson's correlation coefficient related to the slope of a linear regression model?

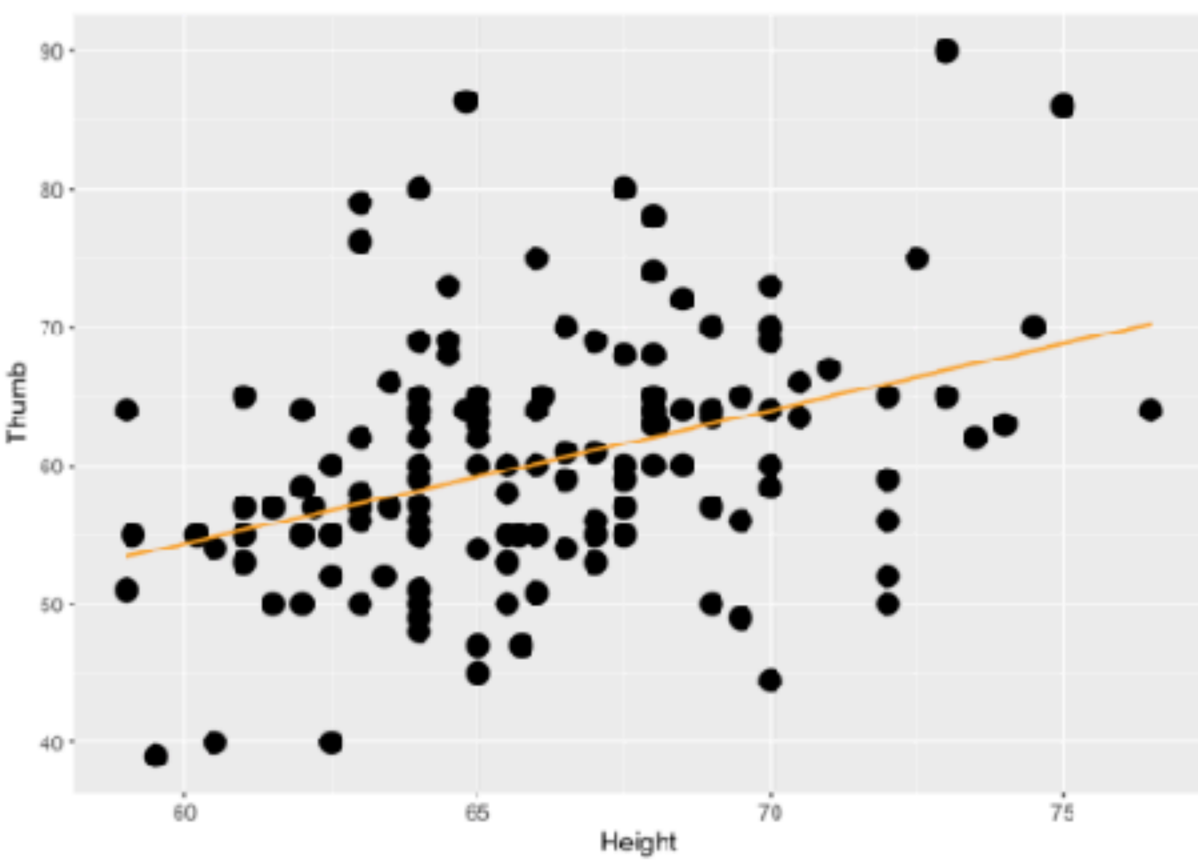
- Both slope of regression line and Pearson's r tell you something about the strength of a linear relationship between two variables.
- But they give you different kinds of information:
  - Pearson's r gives you information that is independent of the units used to measure both variables. Tells you how close the relationship is to a perfect linear relationship.
  - The slope of regression line tells you estimated change in value of outcome variable (Y) for each unit of change in predictor variable (X). Useful for making precise predictions.
  - Slope and Pearson's r are equal when sd of Y and X are equal.

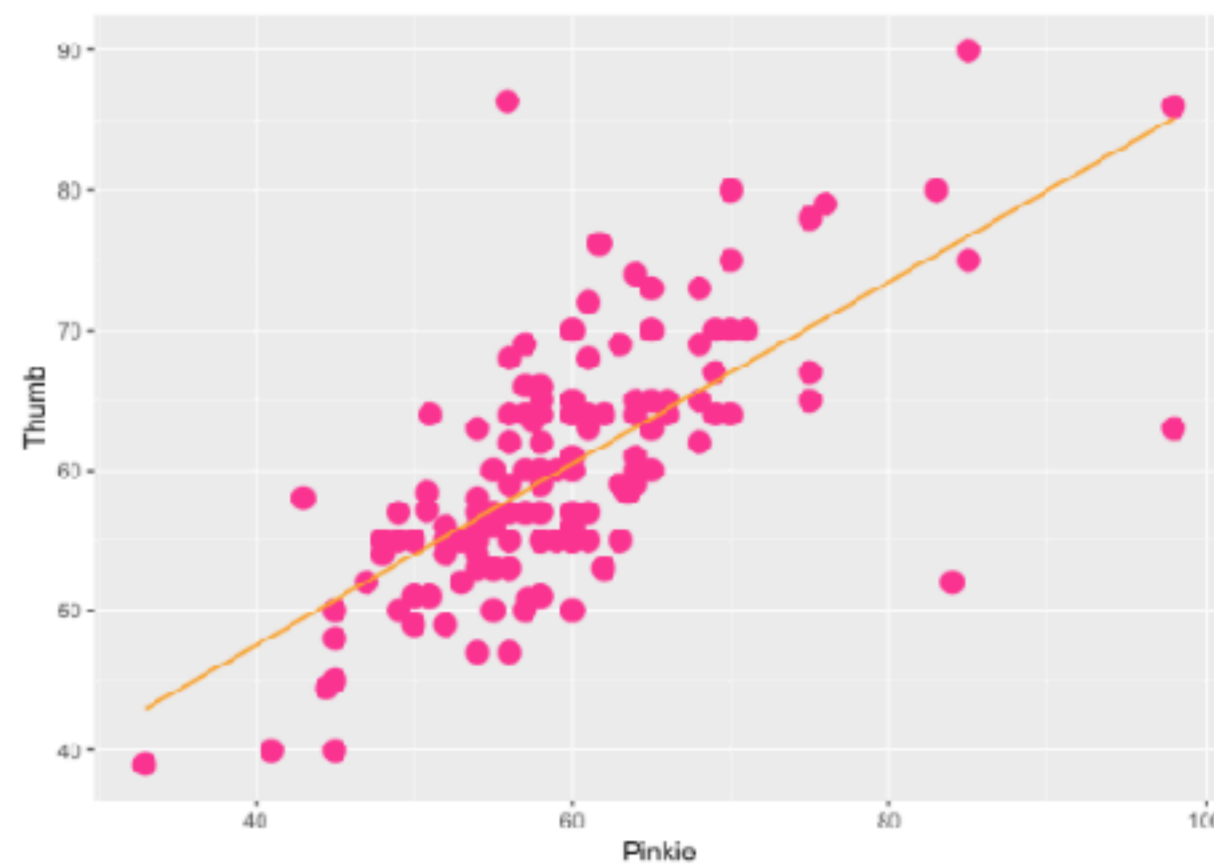
$$\beta_1 = r(Y, X) * \frac{s_Y}{s_X}$$

3

What is the (Pearson) correlation coefficient?

Pearson's correlation coefficient



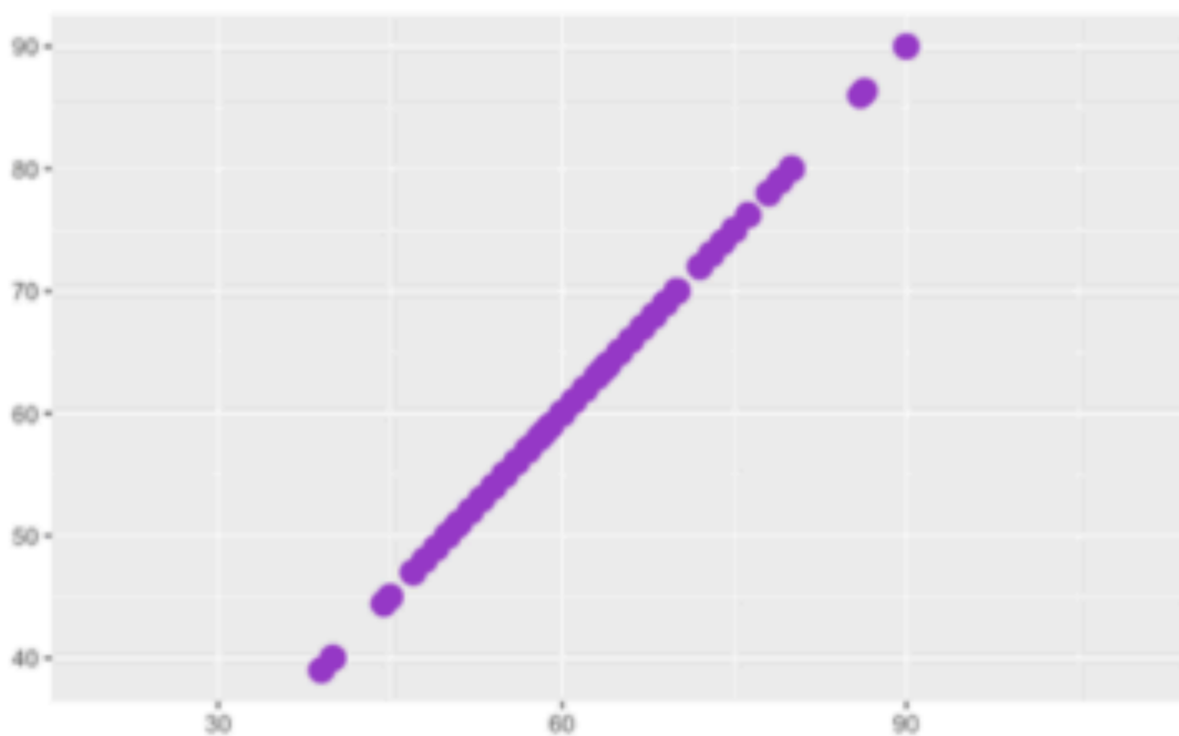


Thumb by **Height**	**Thumb** by **Pinkie**

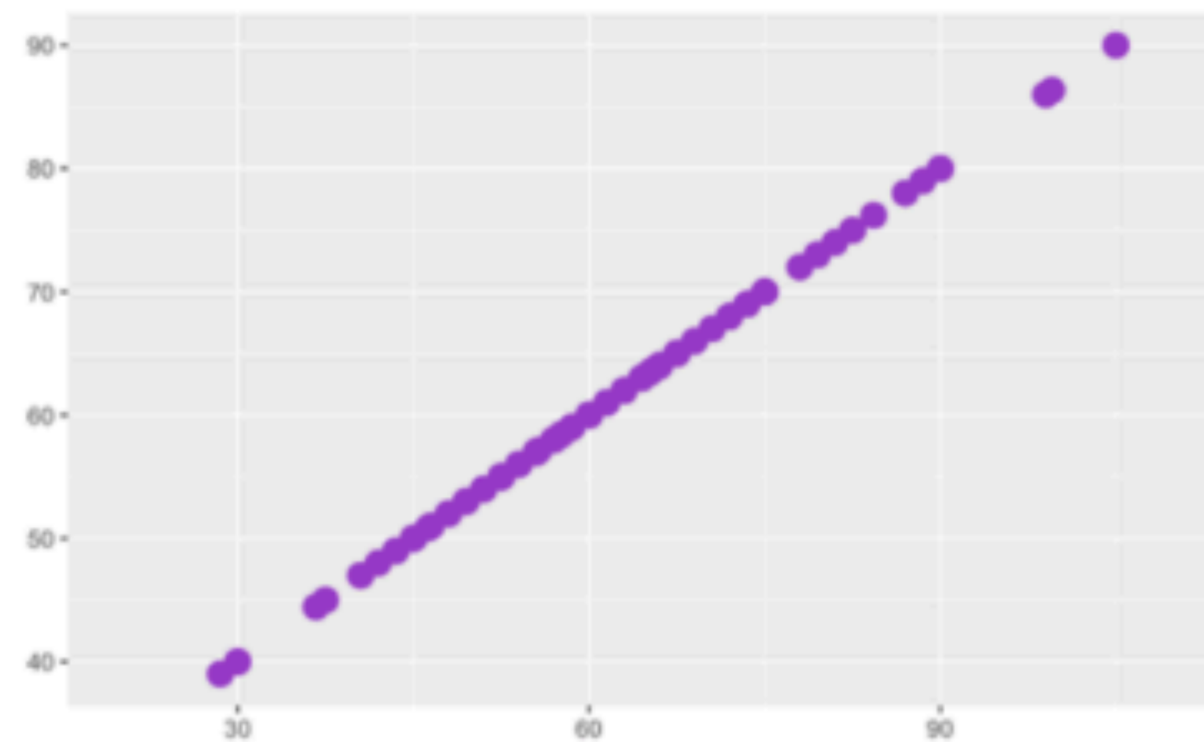
3

What is the (Pearson) correlation coefficient?

Pearson's correlation coefficient



A



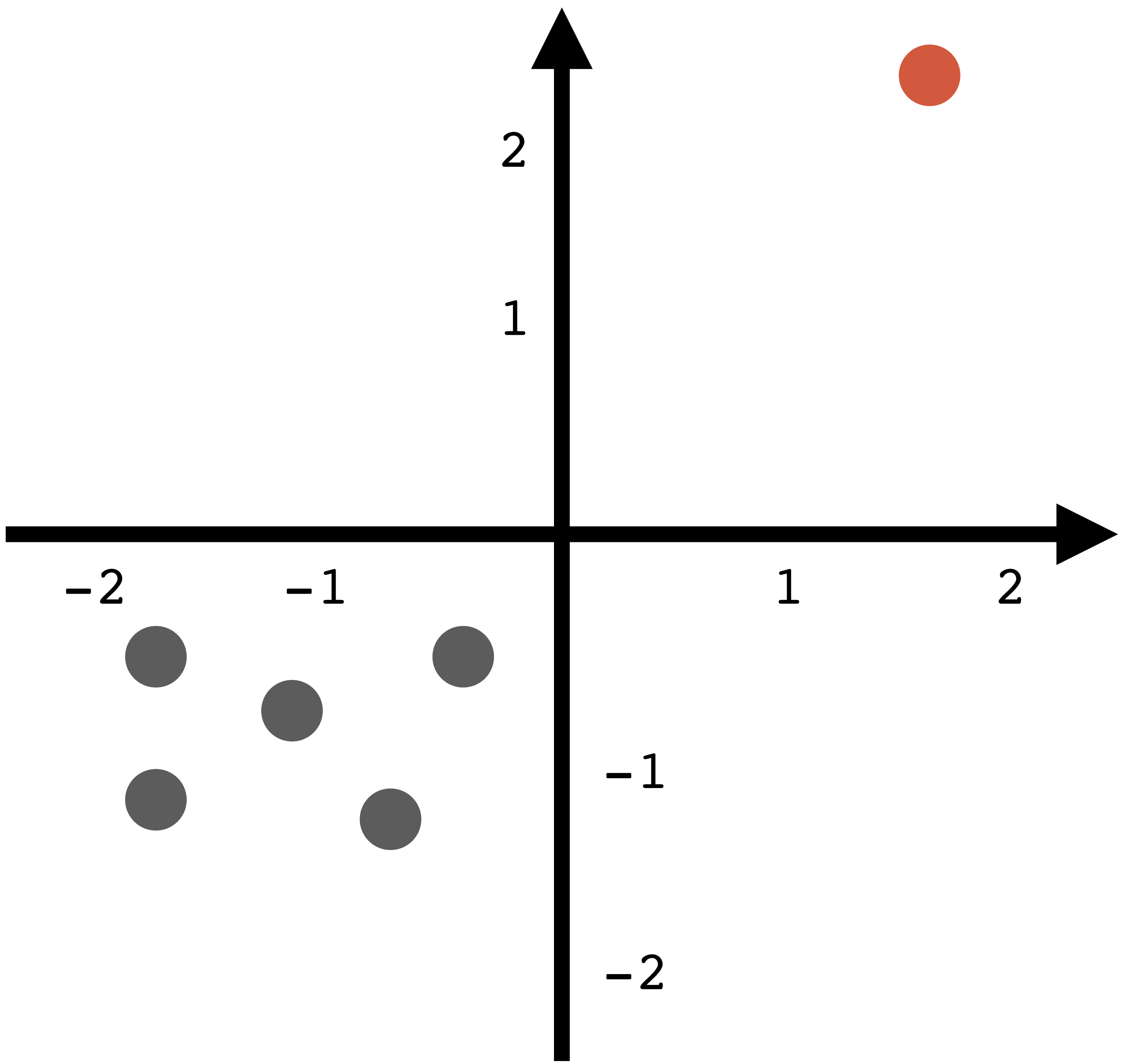
B

Which of these unstandardized scatterplots shows a stronger correlation between the two variables?

3

What is the (Pearson) correlation coefficient?

Remember to visualize your data!



- Pearson's *r* is very sensitive to outliers.
- Also, you can calculate Pearson's r for any set of (x,y) coordinates .... but that doesn't mean that you are looking at a linear relationship!

100
80
60
40
20
0
y

X Mean: 54.26

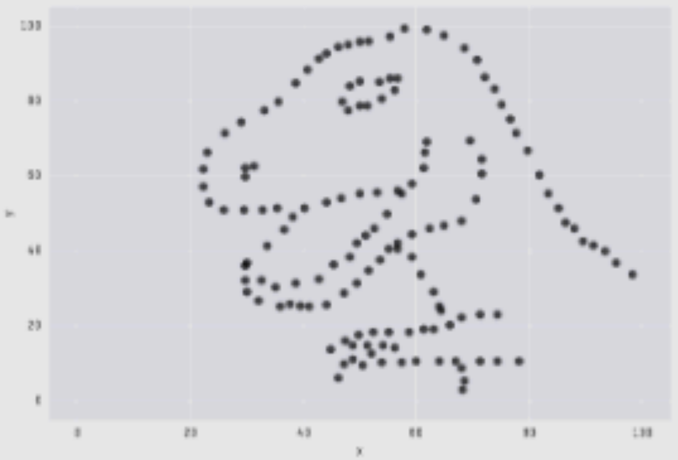
Y Mean: 47.83

X SD : 16.76

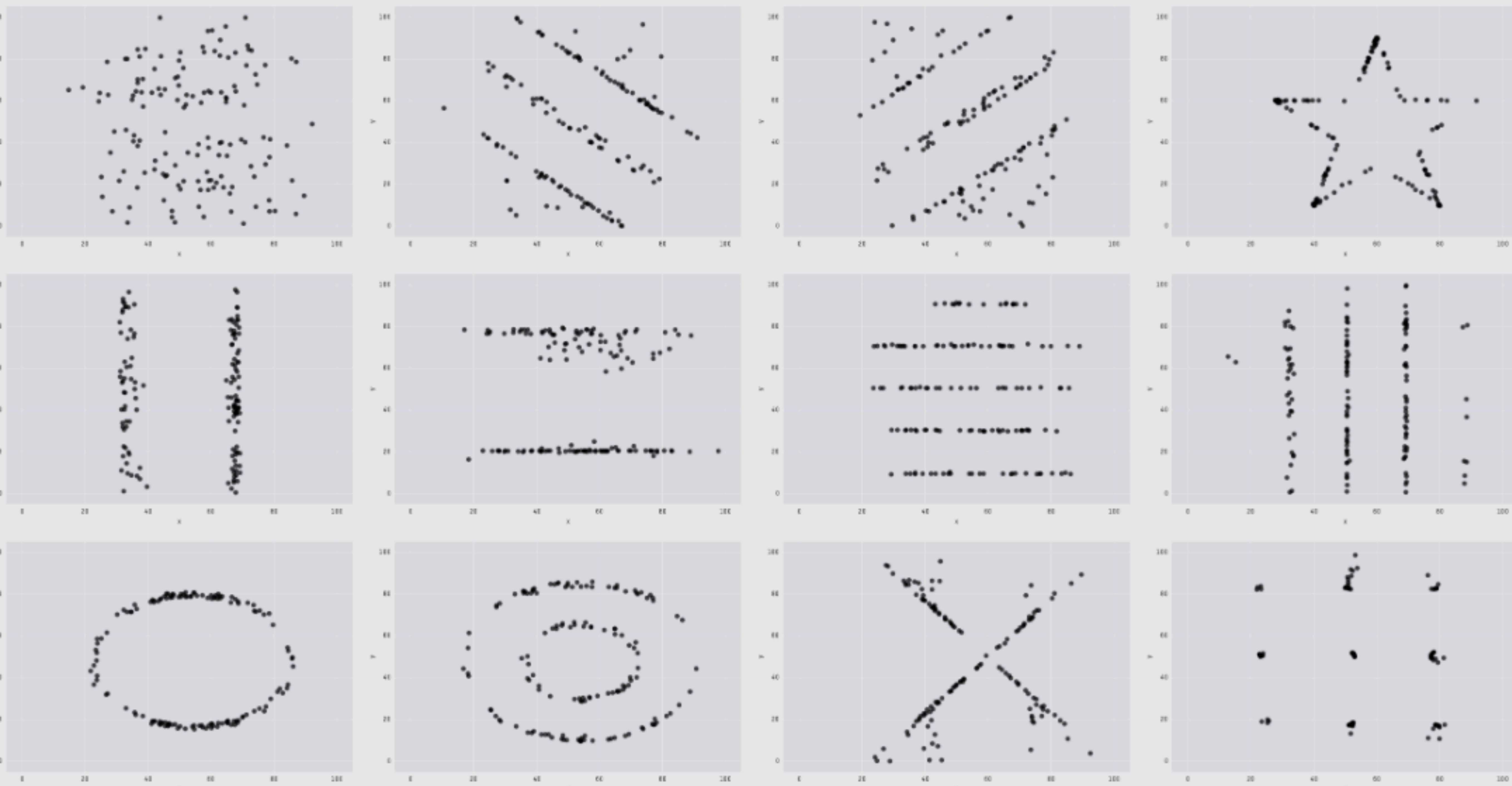
Y SD : 26.93

Corr. : -0.06

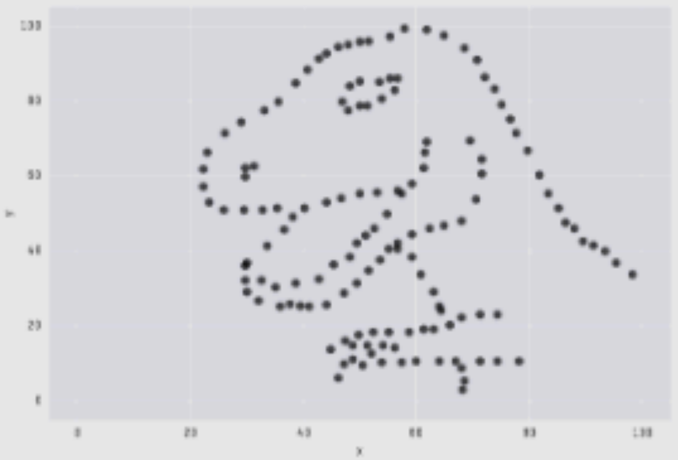
Summary statistics are identical in all 13 graphs.



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



Summary statistics are identical in all 13 graphs.



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

Don't be fooled: visualize your data!

Today

Mini-Review Session #3

1 → 2 → 3

Using an explanatory variable to model variation in an outcome variable

Quantifying effects using confidence intervals

What is the (Pearson) correlation coefficient?

Go to: https://psyc60.github.io/syllabus

Student Daily Feedback Survey

...se complete the linked daily feedback survey. The purpose of this ... better understand how things are going for you in this class, and ... reflect on what you have been learning.

Feedback

We welcome student fe... ...d your TA a Slack message, or ... ...nline form.

Before leaving class, please complete daily feedback survey!

Acknowledgements

Many thanks to Prof. Ji Son, Prof. James Stigler, everyone in the UCLA Teaching and Learning Lab, Prof. Russ Poldrack and Prof. Tobias Gerstenberg for generously sharing their instructional materials.

PSYC 60: How was class today?

Hi there!

I would love to know about your experience in today's class. Could you please take 2 minutes to answer the following few questions? It will be hugely useful for helping me know what is working well, what isn't, and how to keep improving this class.

Best,
Prof. Fan

jefan@ucsd.edu Switch account

Your email will be recorded when you submit this form

* Required

How are you finding the pace of this class so far? *

	1	2	3	4	5	6	7	
Much too slow	○	○	○	○	○	○	○	Much too fast

Do you feel like you are learning new things? *

	1	2	3	4	5	6	7	
Not learning anything new	○	○	○	○	○	○	○	Learning lots of new things