Psyc 60: Intro to Statistics

Prof. Judith Fan

Spring 2022

Today

Lecture 19: Generalizing what you have learned

General announcements

What is correlation & how does it relate to causation?

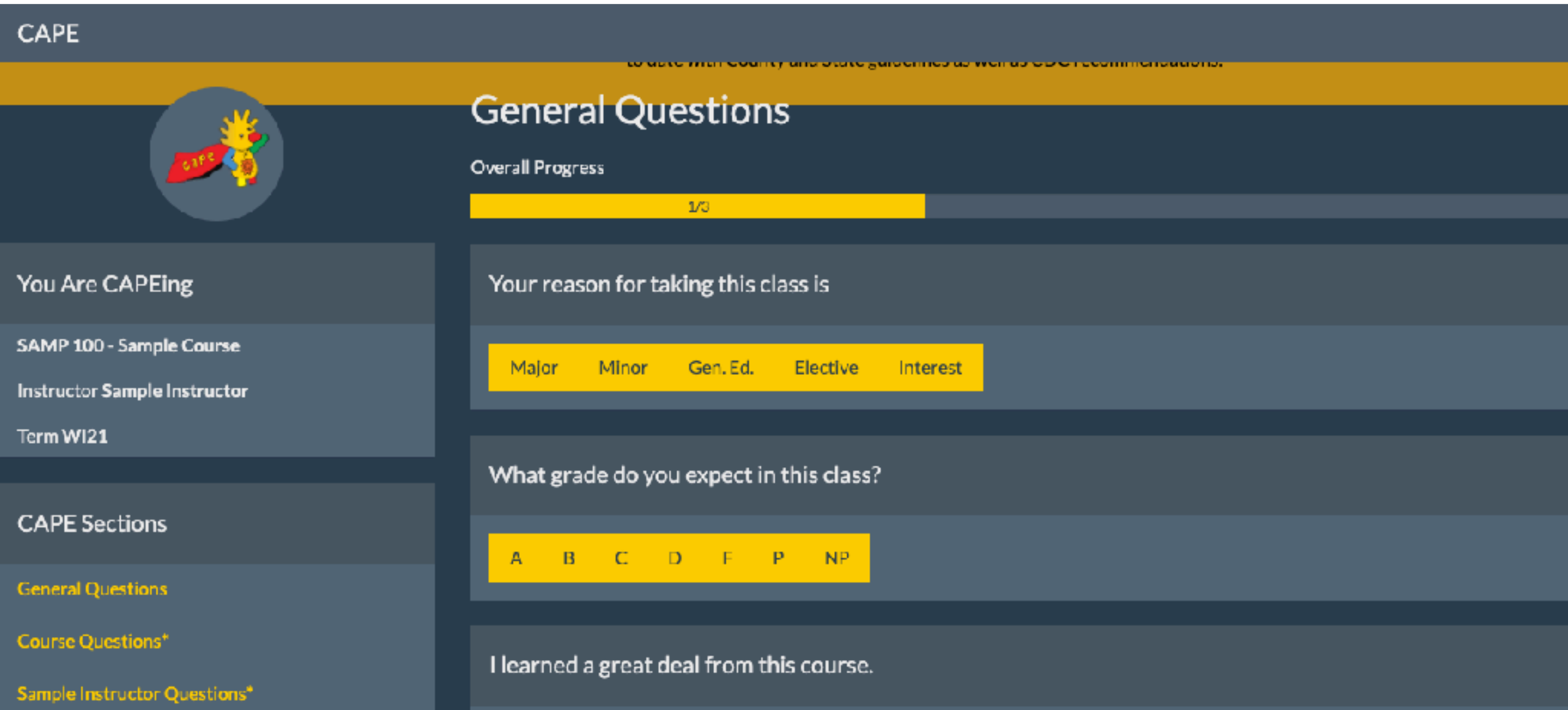
How to extend your model to multiple predictors?

All Remaining Due Dates

Assignment	Deadline
SONA credits	Wednesday, 6/1 at 4PM
CourseKata modules	Friday 6/3 at 11:59PM
Quiz 5	Friday 6/3 at 5PM
Milestone 5: Final project report	Friday 6/3 at 11:59PM
CAPEs	Saturday 6/4 at 8AM
Final project poster	Monday, 6/6 at 12PM (NOON)
Final project showcase	Tuesday, 6/7 at 8AM-10:50AM
Project peer reviews (during showcase)	Tuesday, 6/7 at 8AM-10:50AM
Project team evaluations (after showcase)	Tuesday, 6/7 at 11:59PM
Lab 5	Wednesday, 6/8 at 11:59PM

Please stay tuned for information about the project showcase over Slack & Canvas!

Every time someone fills out a CAPE, an actual unicorn does a little dance.



Today

Lecture 19: Generalizing what you have learned

General announcements

What is correlation & how does it relate to causation?

How to extend your model to multiple predictors?

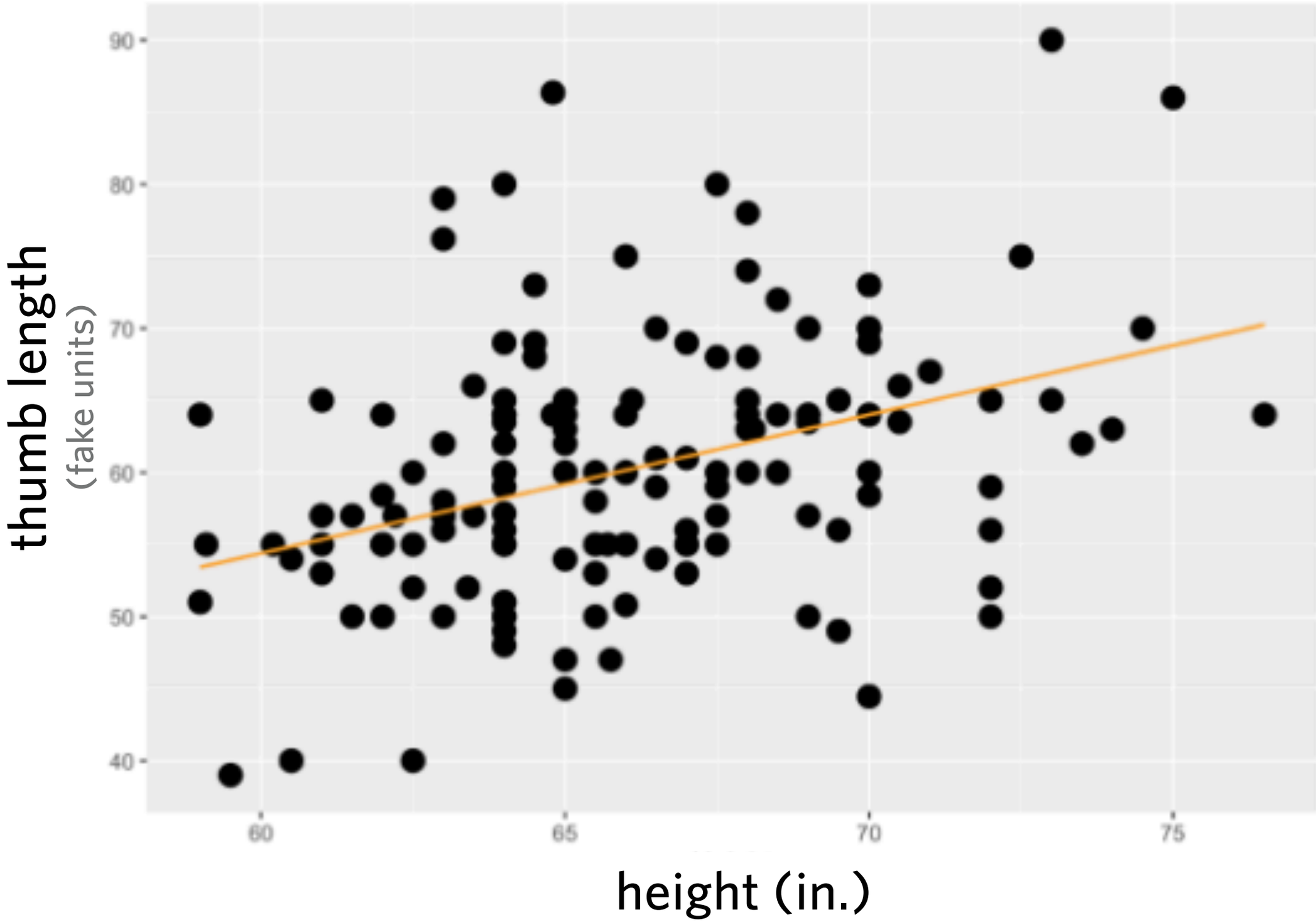
2

What is correlation & how does it relate to causation?

What does the word "correlation" mean to you?

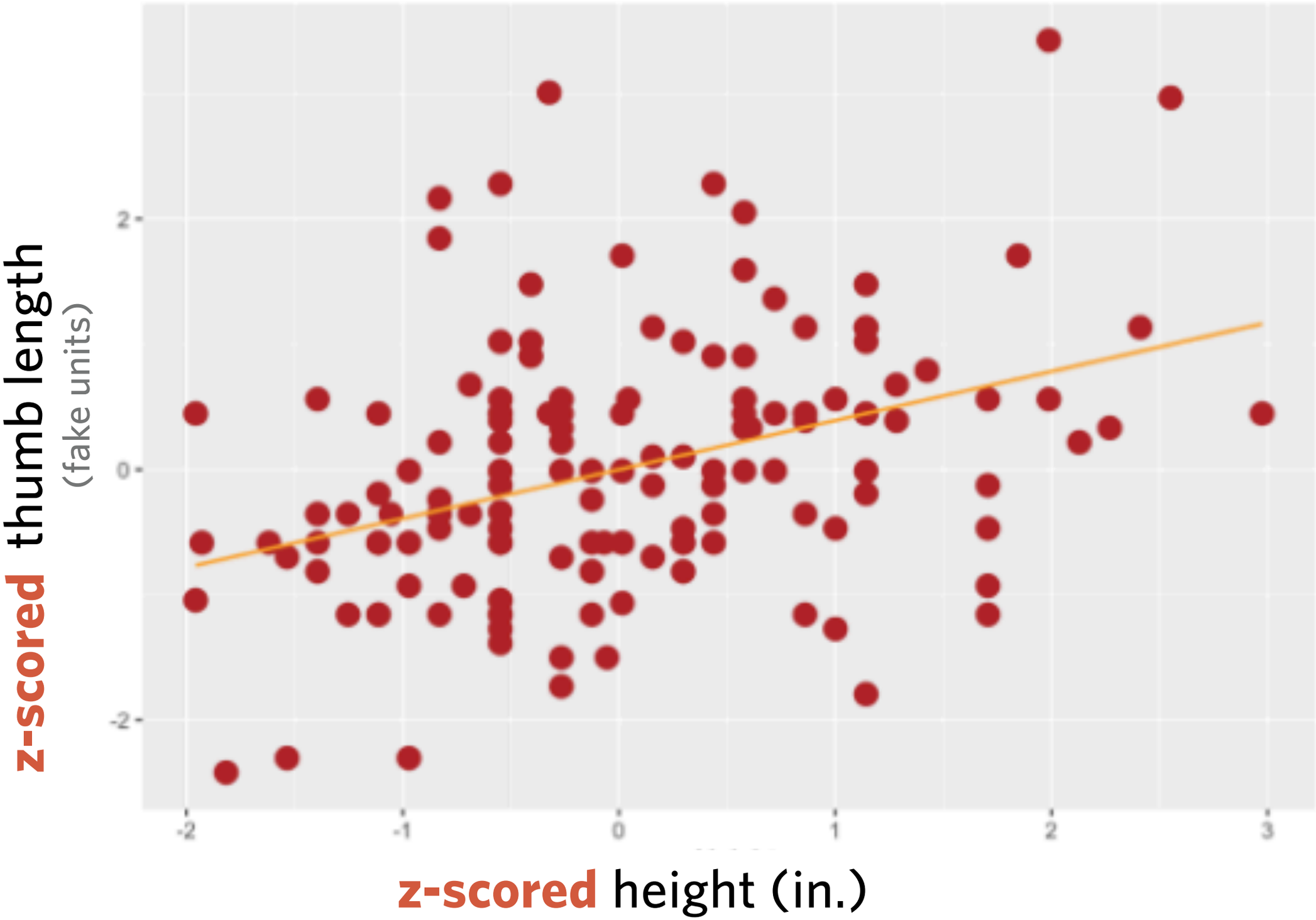
2 *What is correlation & how does it relate to causation?*

Pearson's correlation coefficient



2 *What is correlation & how does it relate to causation?*

Pearson's correlation coefficient

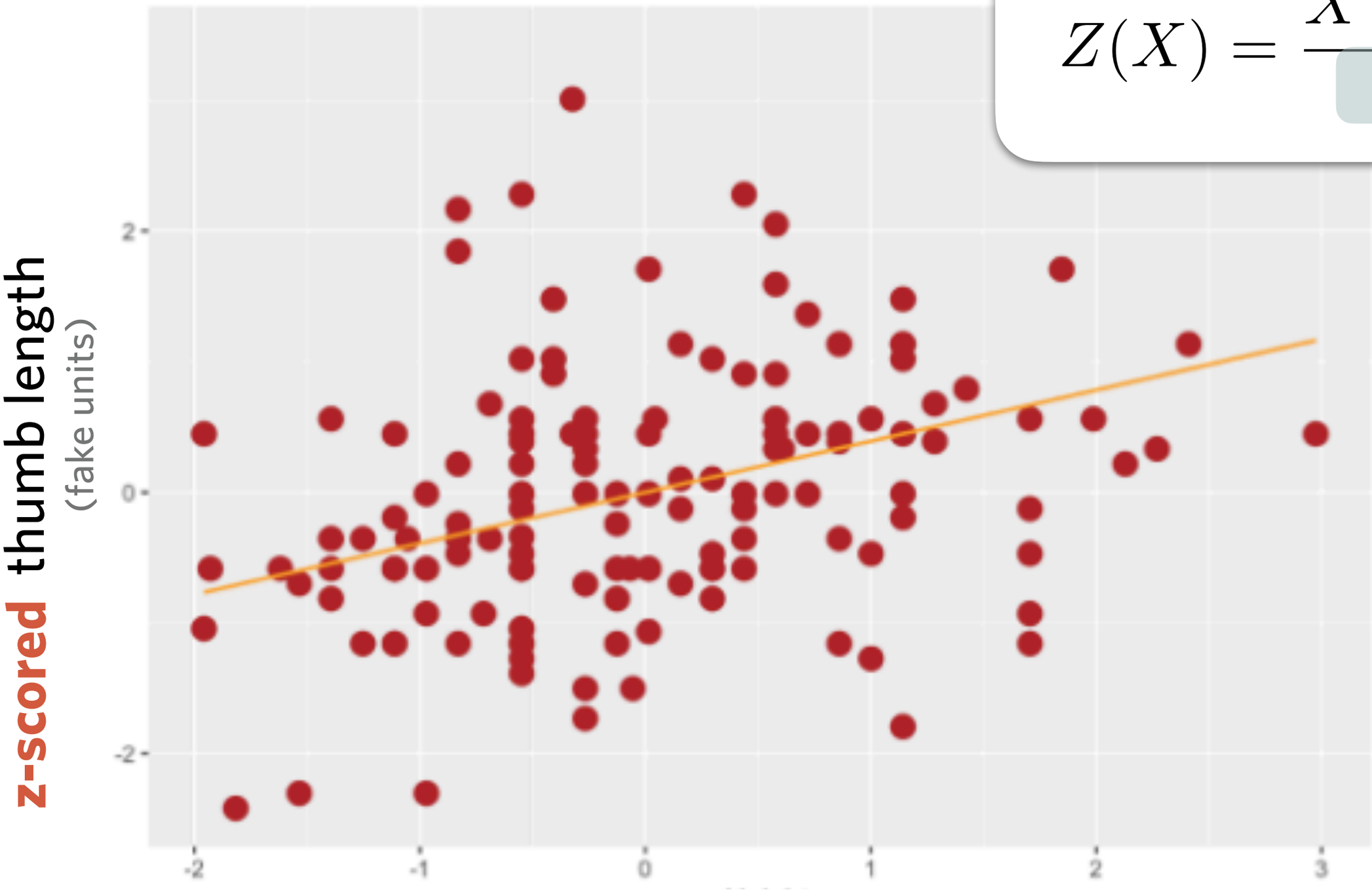


2 *What is correlation & how does it relate to causation?*

Pearson's correlation coefficient

How to "z-score" your data:

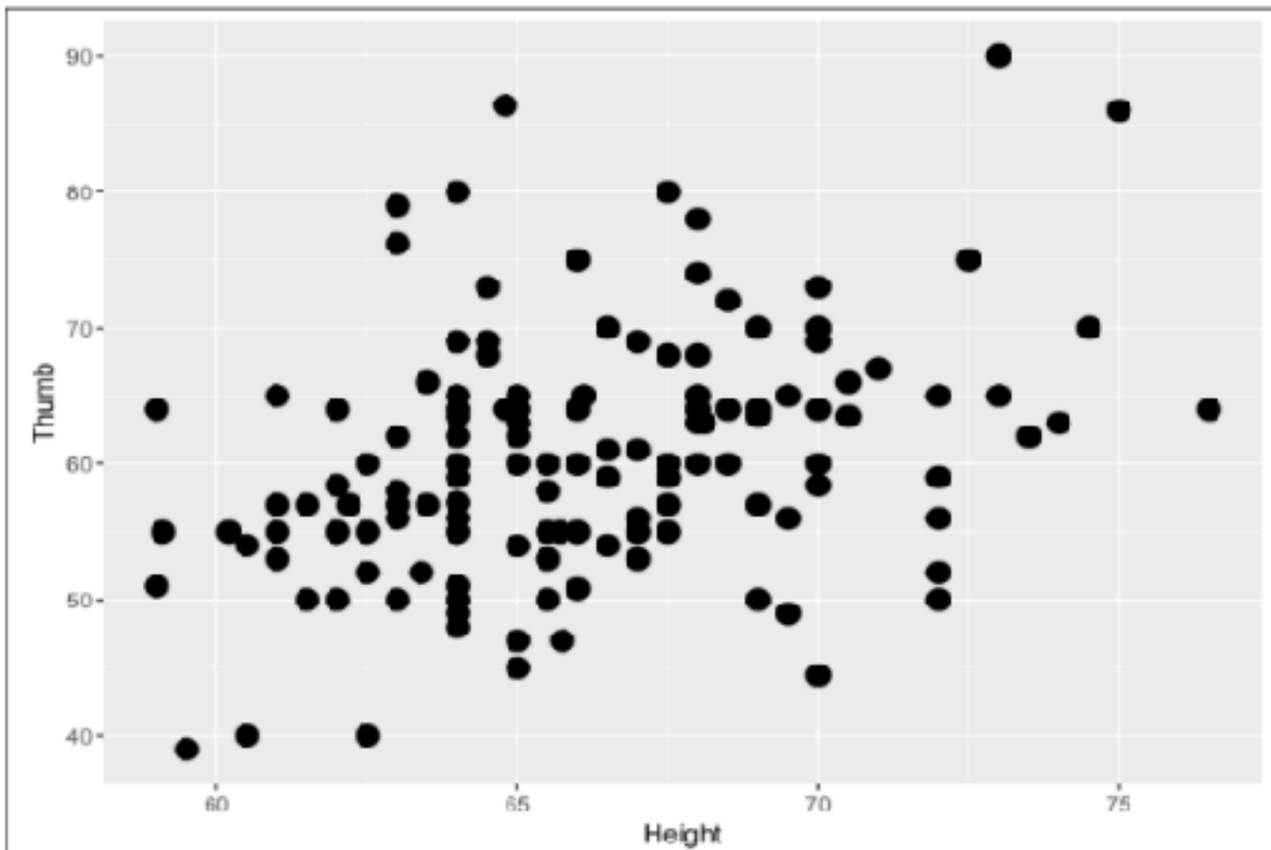
$$Z(X) = \frac{X - \mu}{\sigma}$$



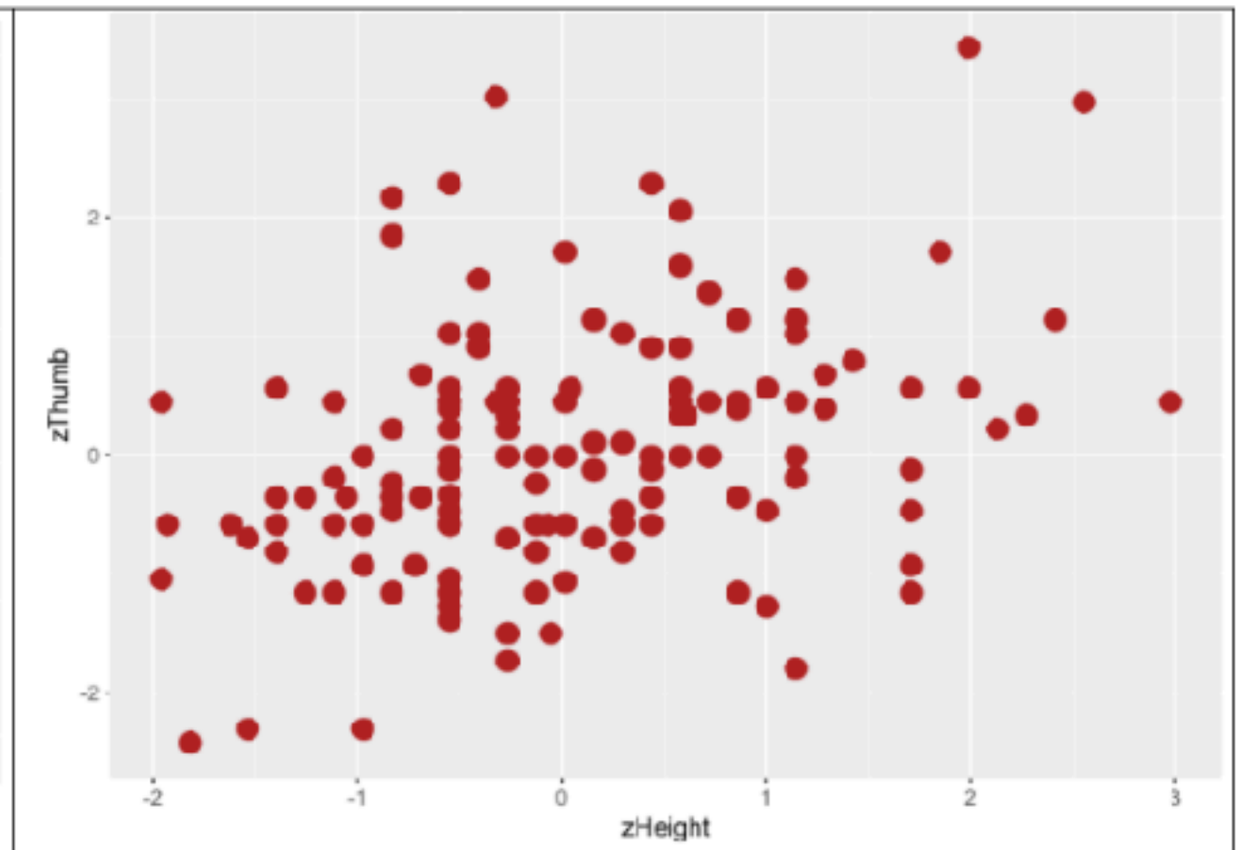
2 *What is correlation & how does it relate to causation?*

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:

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

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).

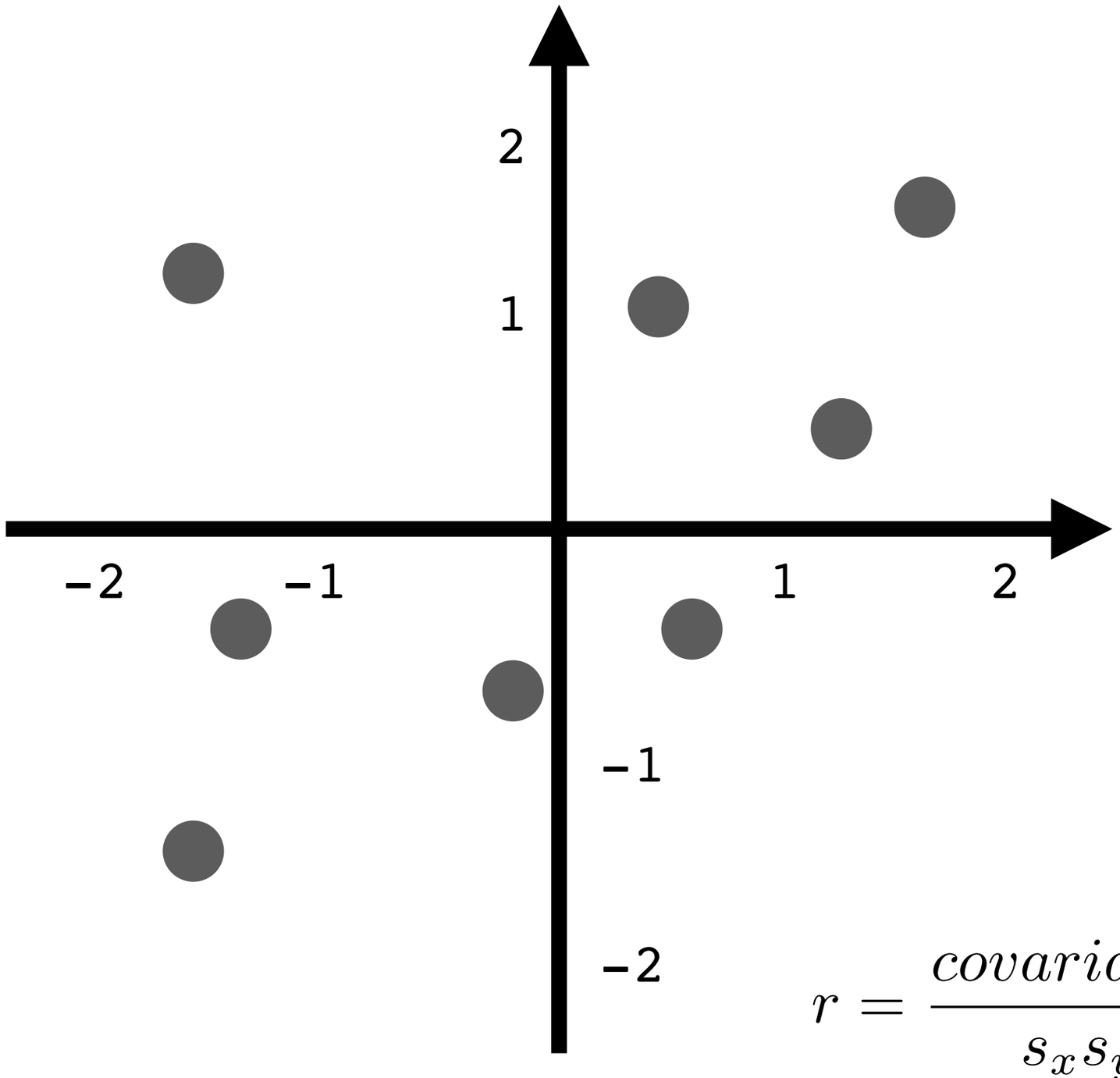
$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

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

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

2 *What is correlation & how does it relate to causation?*

Calculating Pearson's correlation coefficient

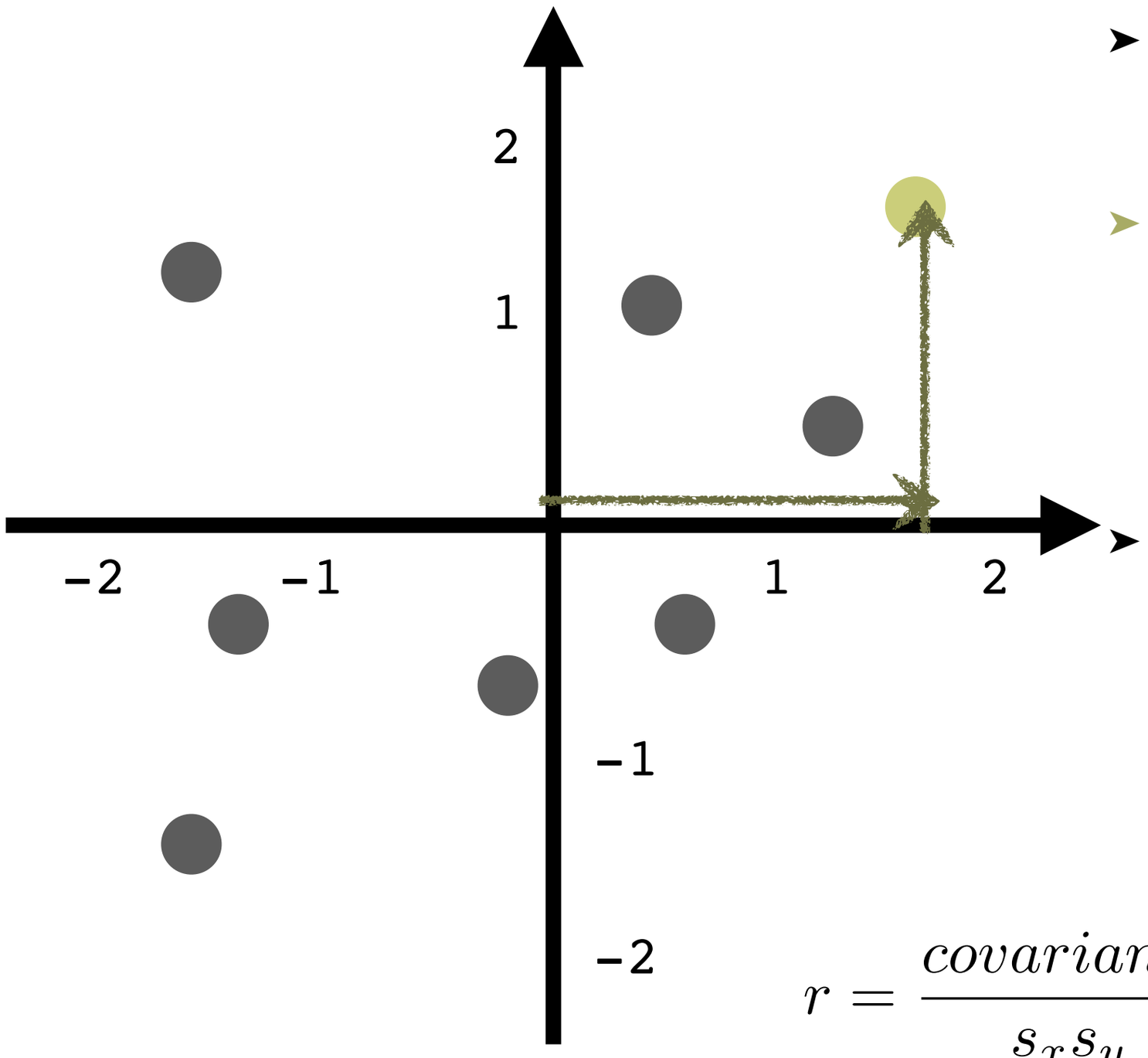


- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.

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

2 *What is correlation & how does it relate to causation?*

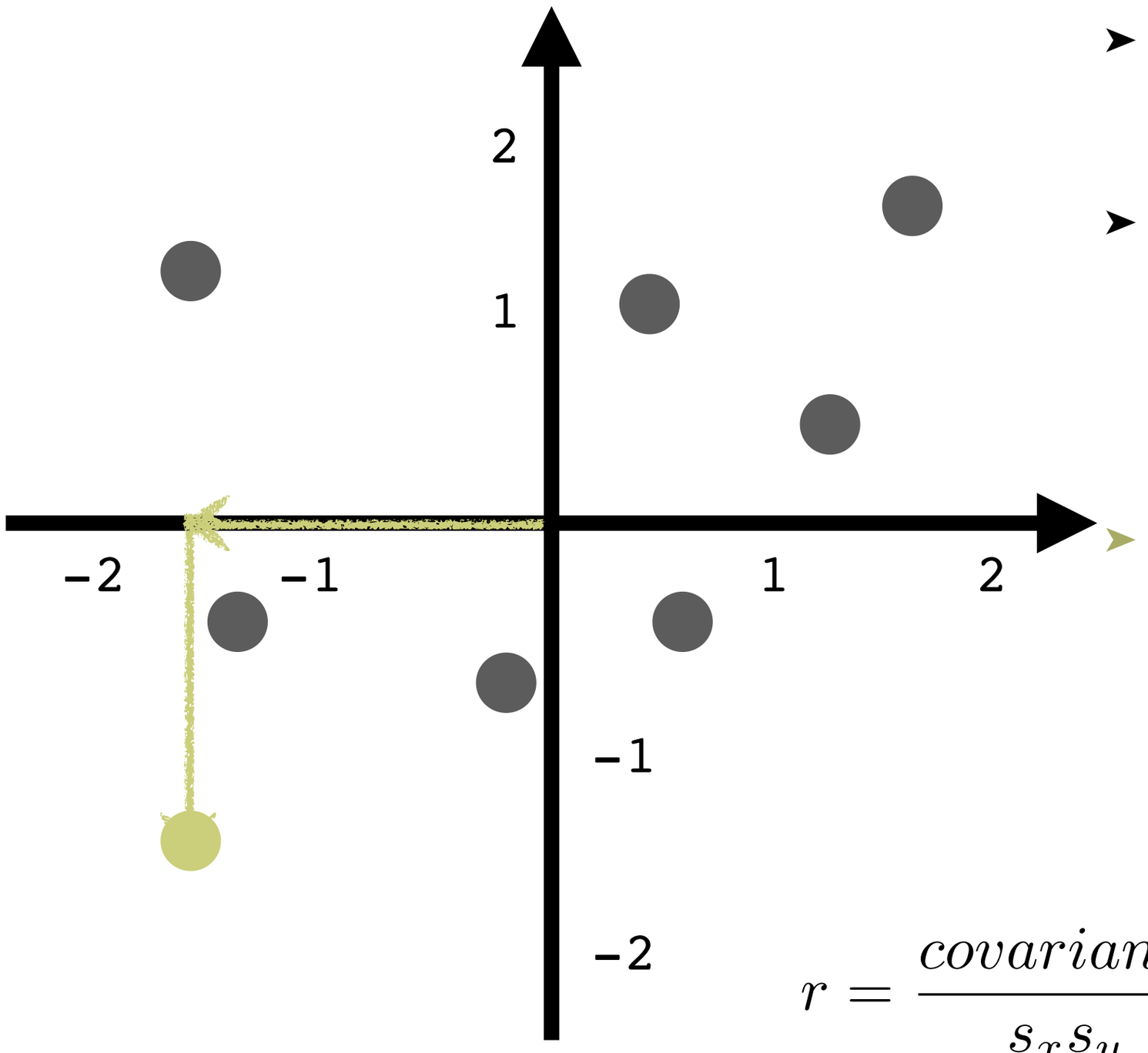
Calculating Pearson's correlation coefficient



- 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).

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

Calculating Pearson's correlation coefficient

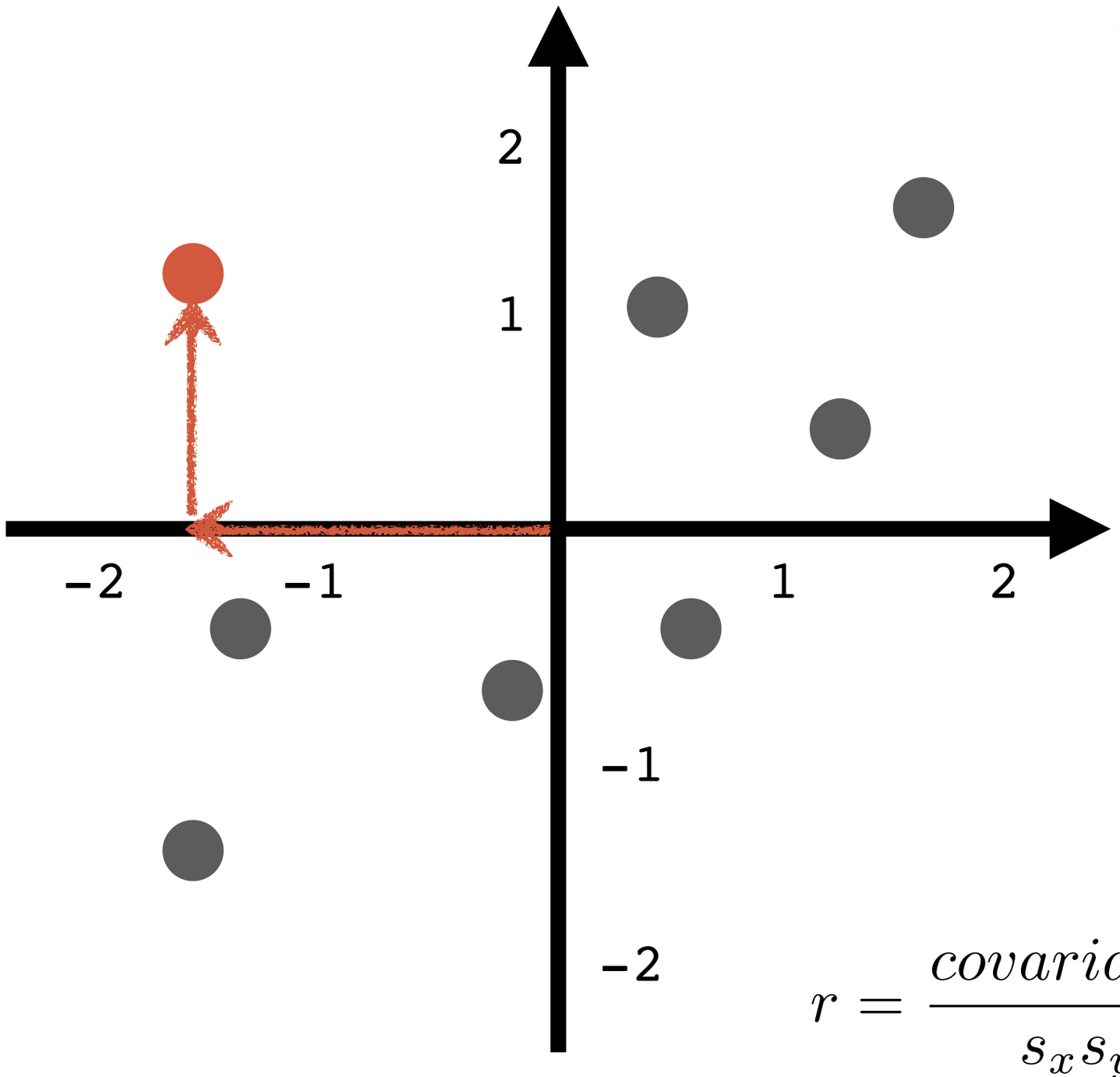


- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- A data point that is in the top right of this z-scored scatter plot will increase Pearson's r.
  - Its x-value > mean(x) and y-value > mean(y).
- A data point in the bottom left will also increase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

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

2 *What is correlation & how does it relate to causation?*

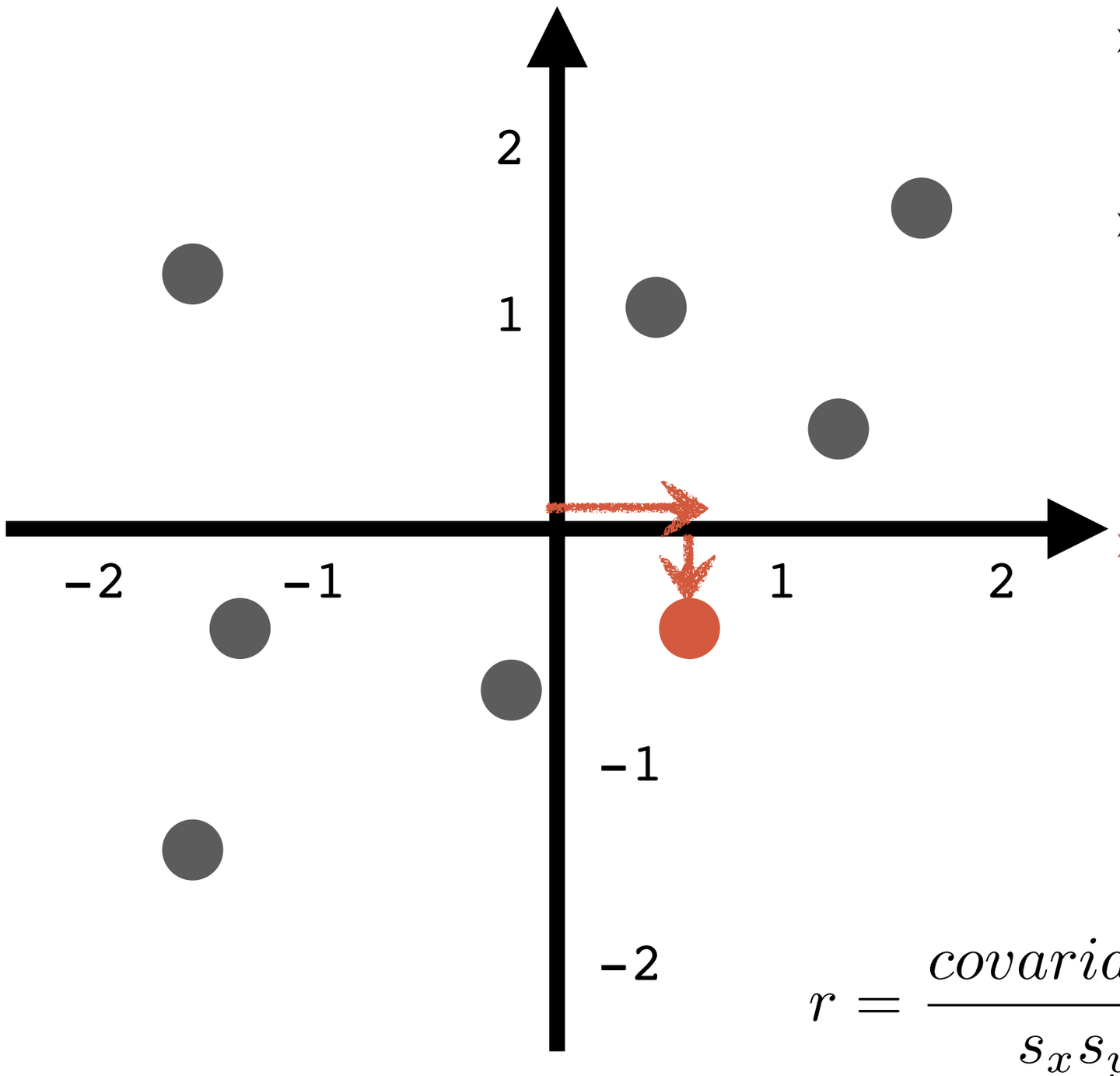
Calculating Pearson's correlation coefficient



- 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



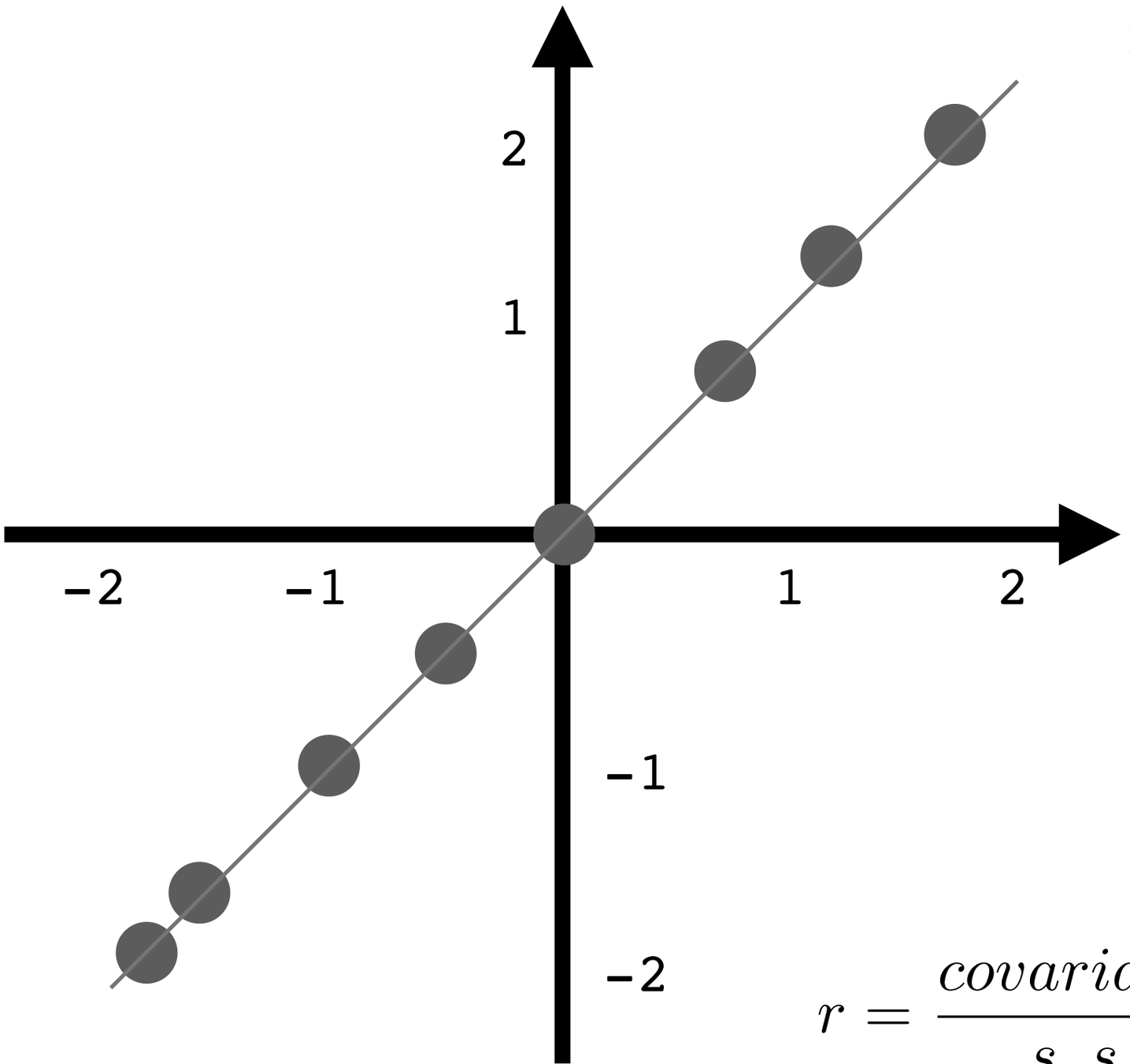
- Pearson's *r* is calculated by **adding** up a bunch of horizontal/vertical deviations from the mean.
- BUT a data point in the top left will *decrease* Pearson's *r.*
  - Its x-value < mean(x) but its y-value > mean(y), so product is negative.

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.

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

Calculating Pearson's correlation coefficient

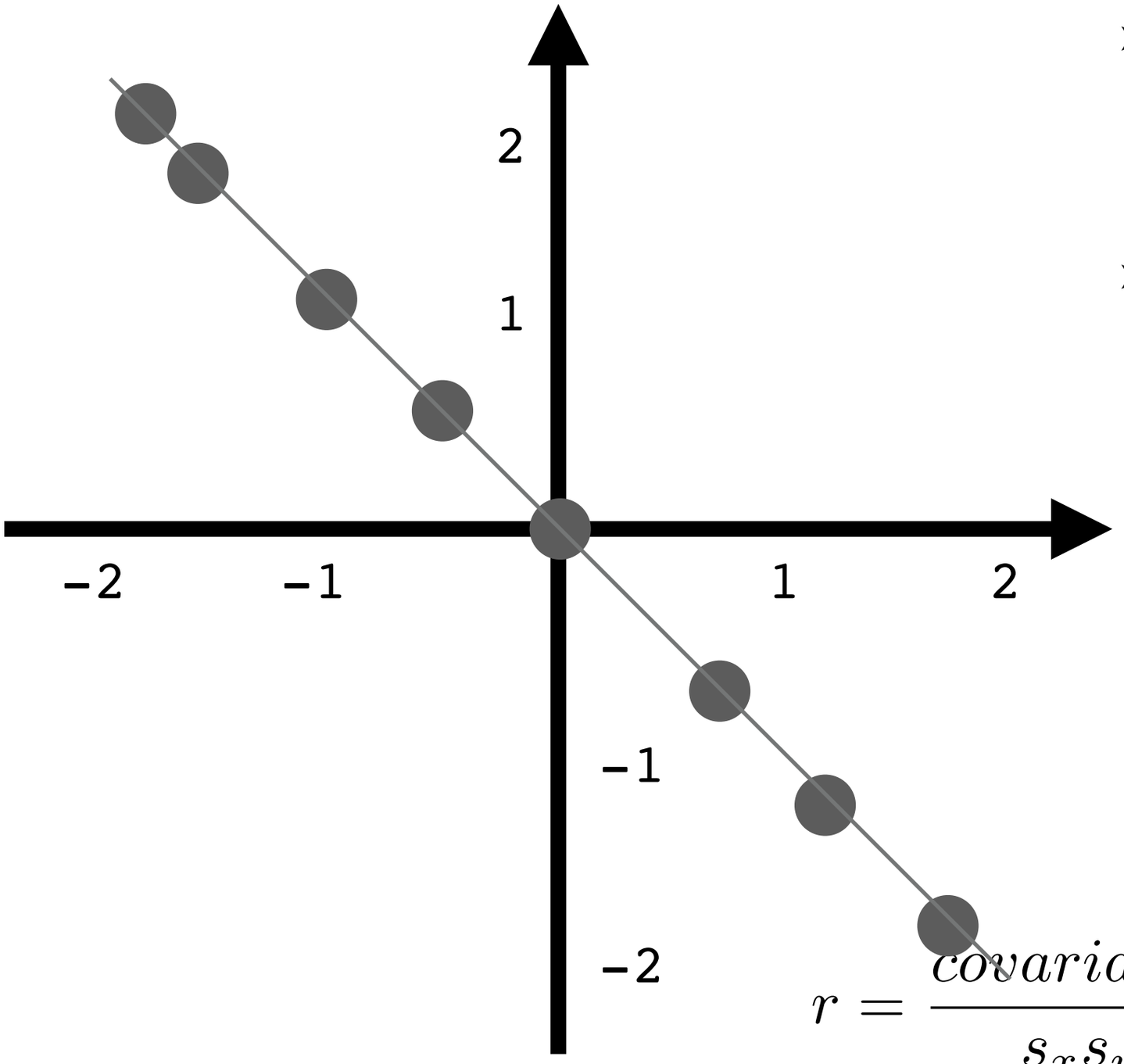


- Pearson's r is **maximized** ($r = +1$) when all data points fall perfectly onto a straight line with positive slope.

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

2 *What is correlation & how does it relate to causation?*

Calculating Pearson's correlation coefficient

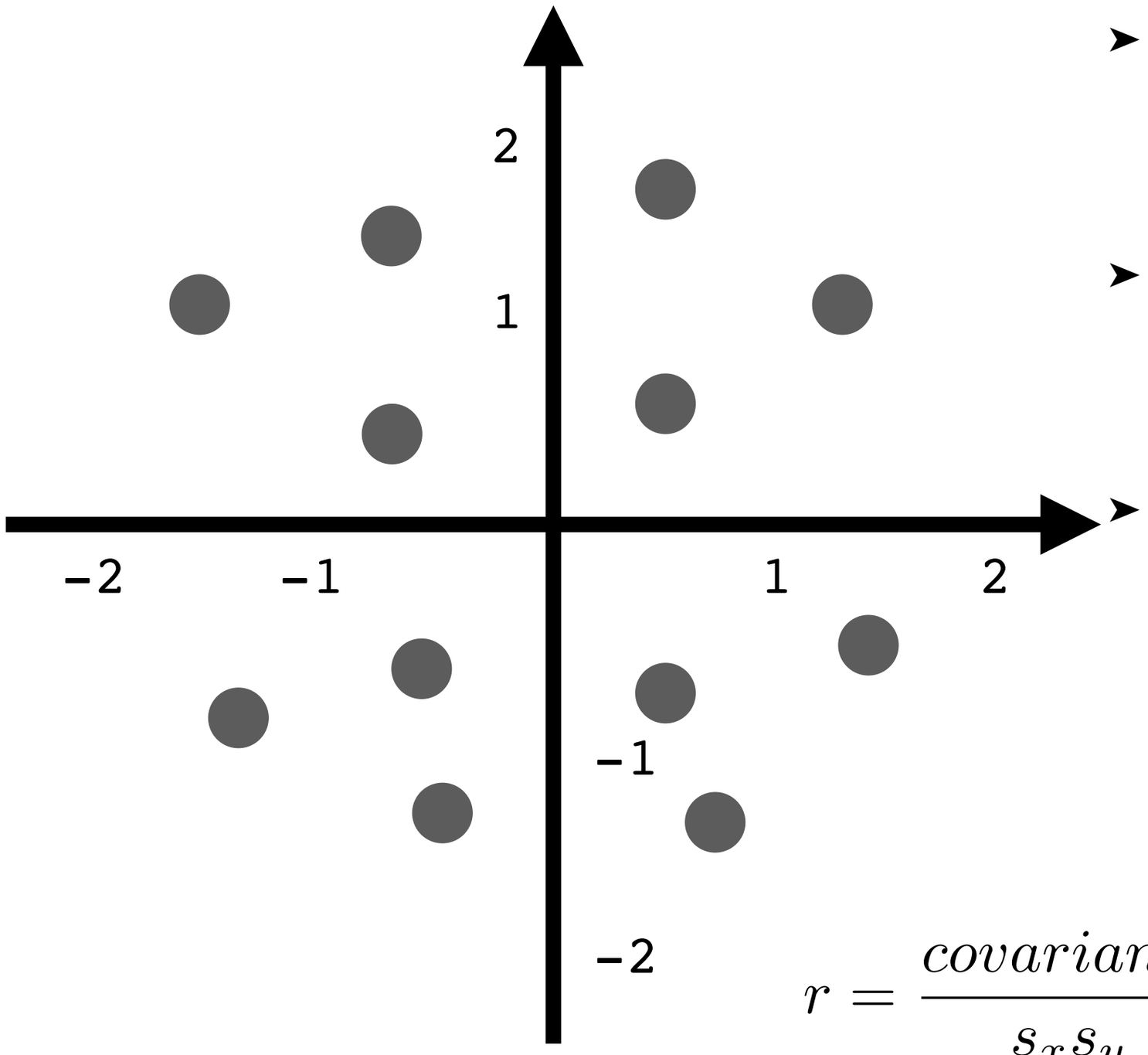


- Pearson's r is **maximized** (*r* = +1) when all data points fall perfectly onto a straight line with positive slope.
- Pearson's r is **minimized** (*r = -1)* when all data points fall perfectly onto a straight line with negative slope.

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

2 *What is correlation & how does it relate to causation?*

Calculating Pearson's correlation coefficient



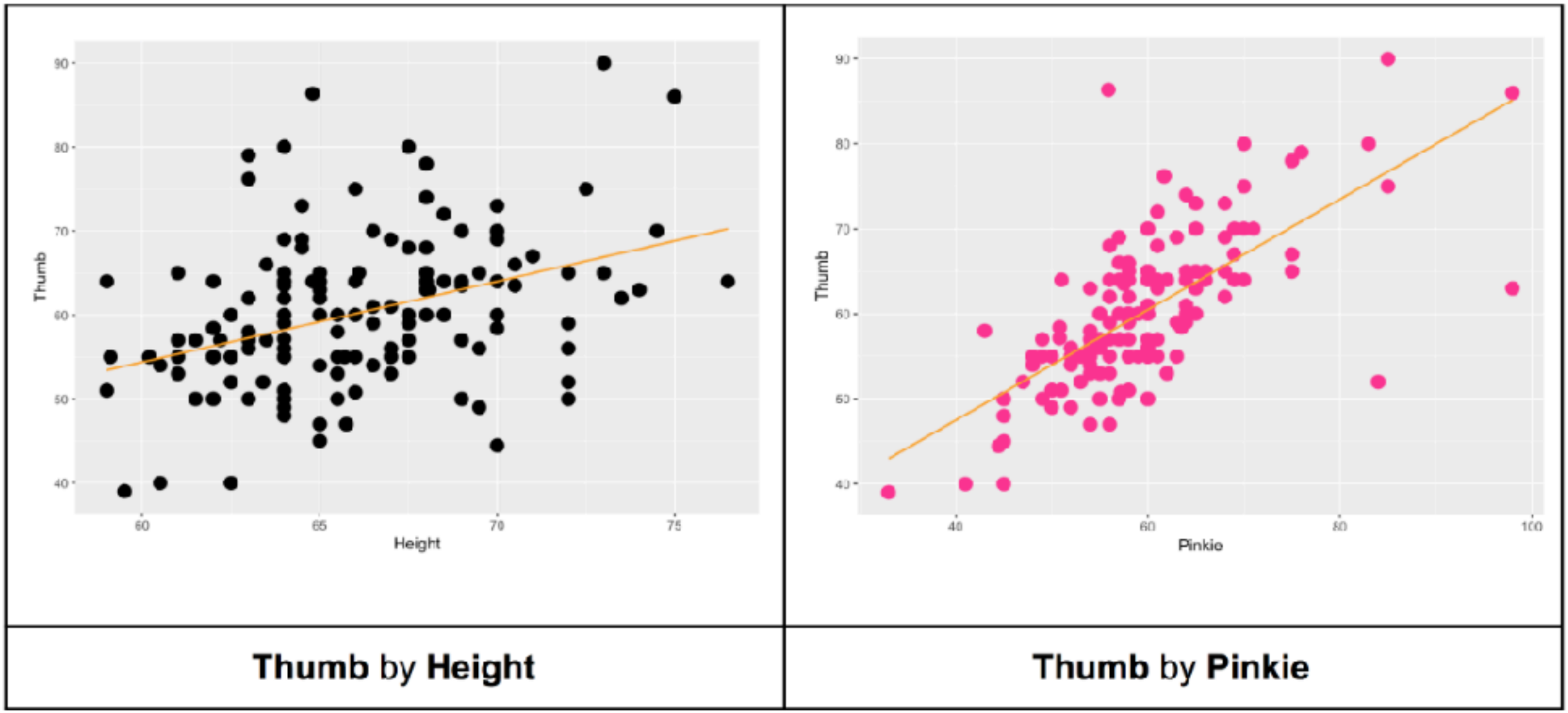
- Pearson's r is **maximized** (*r* = +1) when all data points fall perfectly onto a straight line with positive slope.
- Pearson's r is **minimized** (*r = -1)* when all data points fall perfectly onto a straight line with negative slope.

Pearson's r is **zero** (*r*=0) when data points form a ball-shaped cloud with no apparent tendency toward positive or negative slope.

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

2 *What is correlation & how does it relate to causation?*

Pearson's correlation coefficient



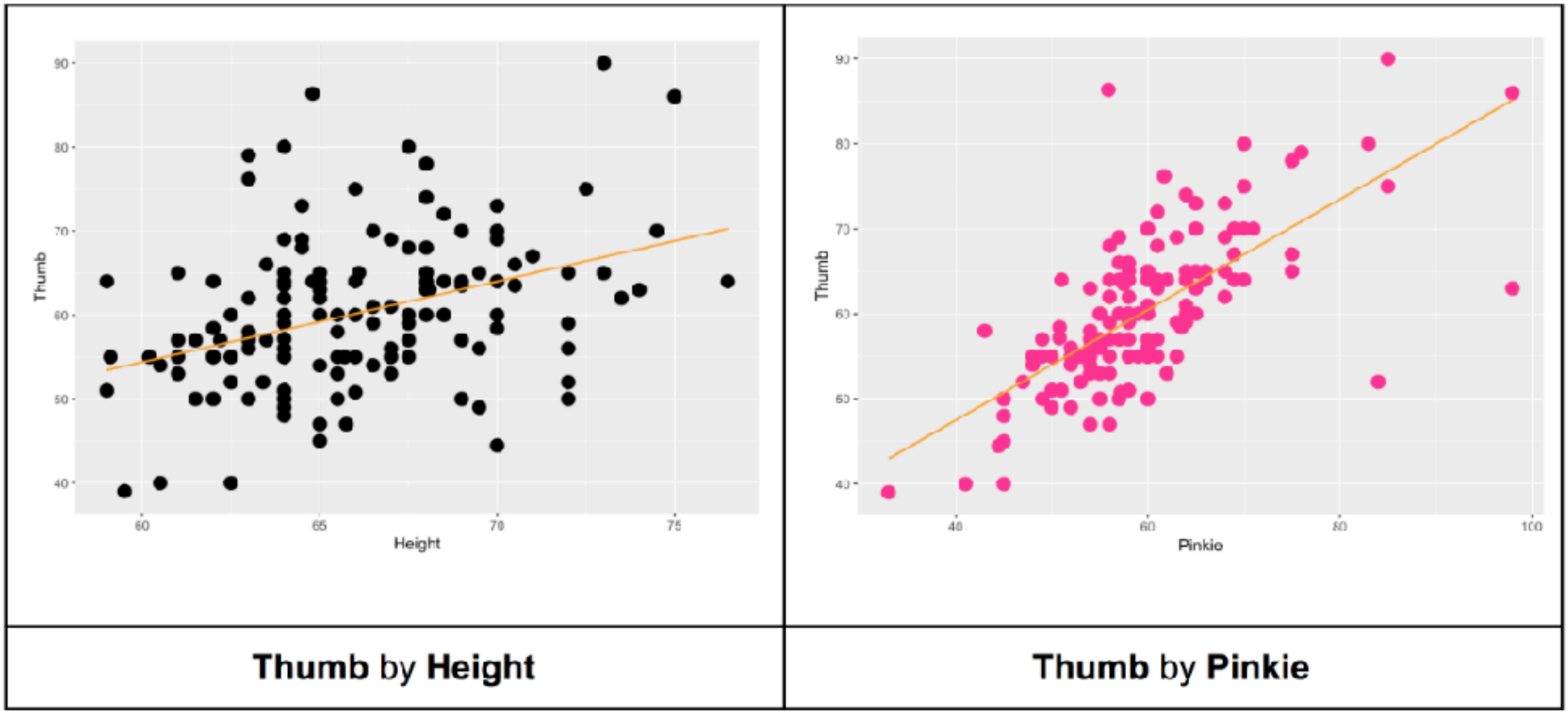
Thumb by Height	Thumb by Pinkie

Which pair of variables is more strongly correlated?

2

What is correlation & how does it relate to causation?

Pearson's correlation coefficient



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

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

2 *What is correlation & how does it relate to causation?*

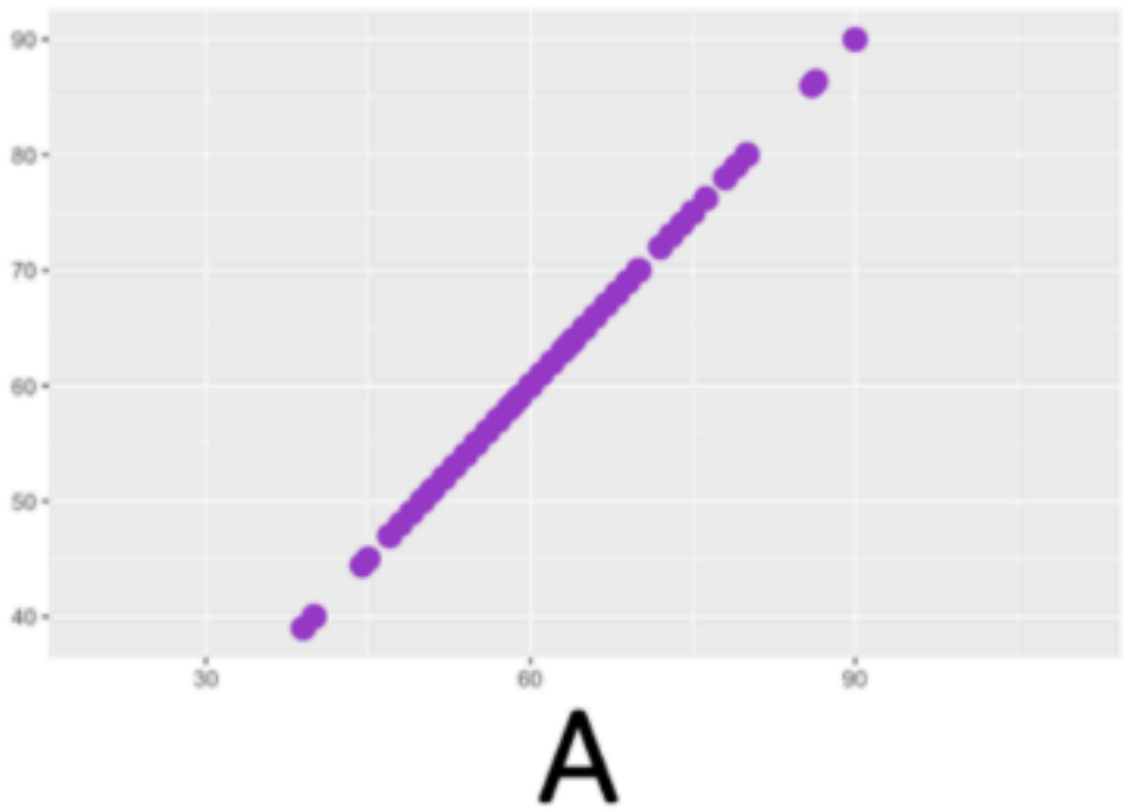
How is Pearson's correlation coefficient related to the slope of a linear regression model?

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

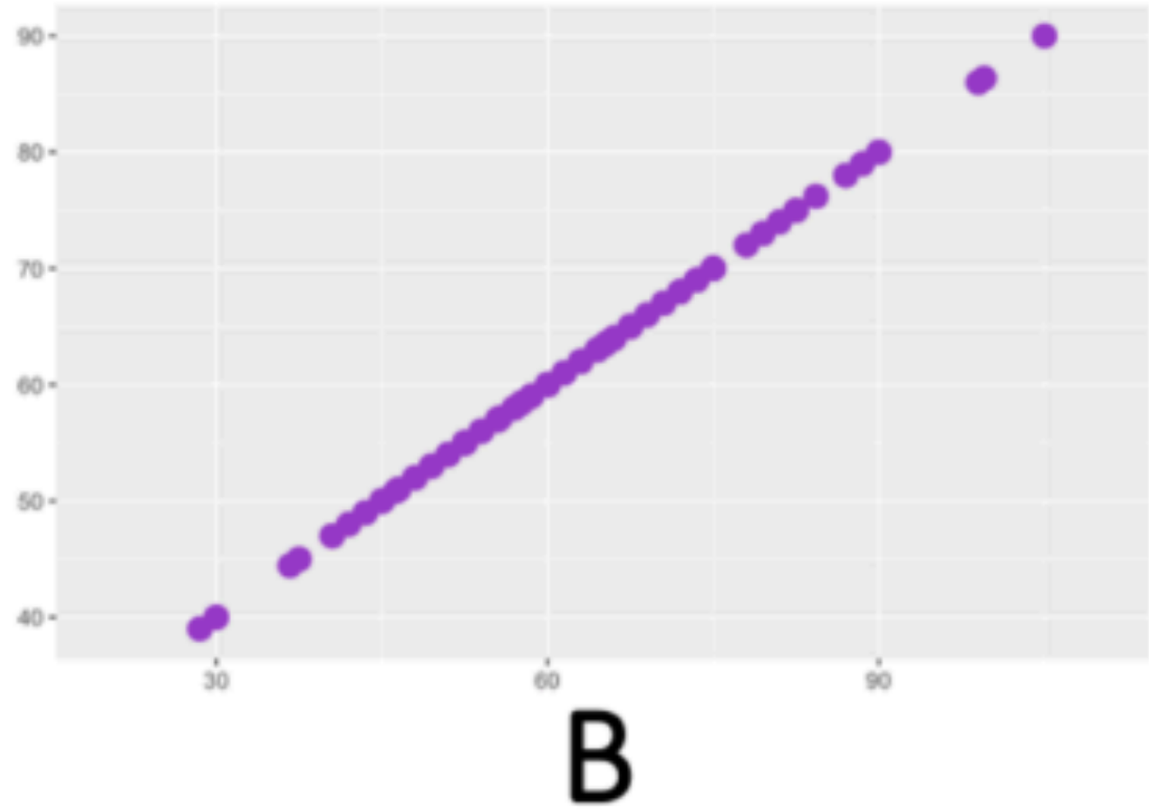
- Both slope of regression line and Pearson's r tell you something about the strength of a linear relationship between two variables.
- But they give you different kinds of information:
  - Pearson's r gives you information that is independent of the units used to measure both variables. Tells you how 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.

2 *What is correlation & how does it relate to causation?*

Pearson's correlation coefficient



A

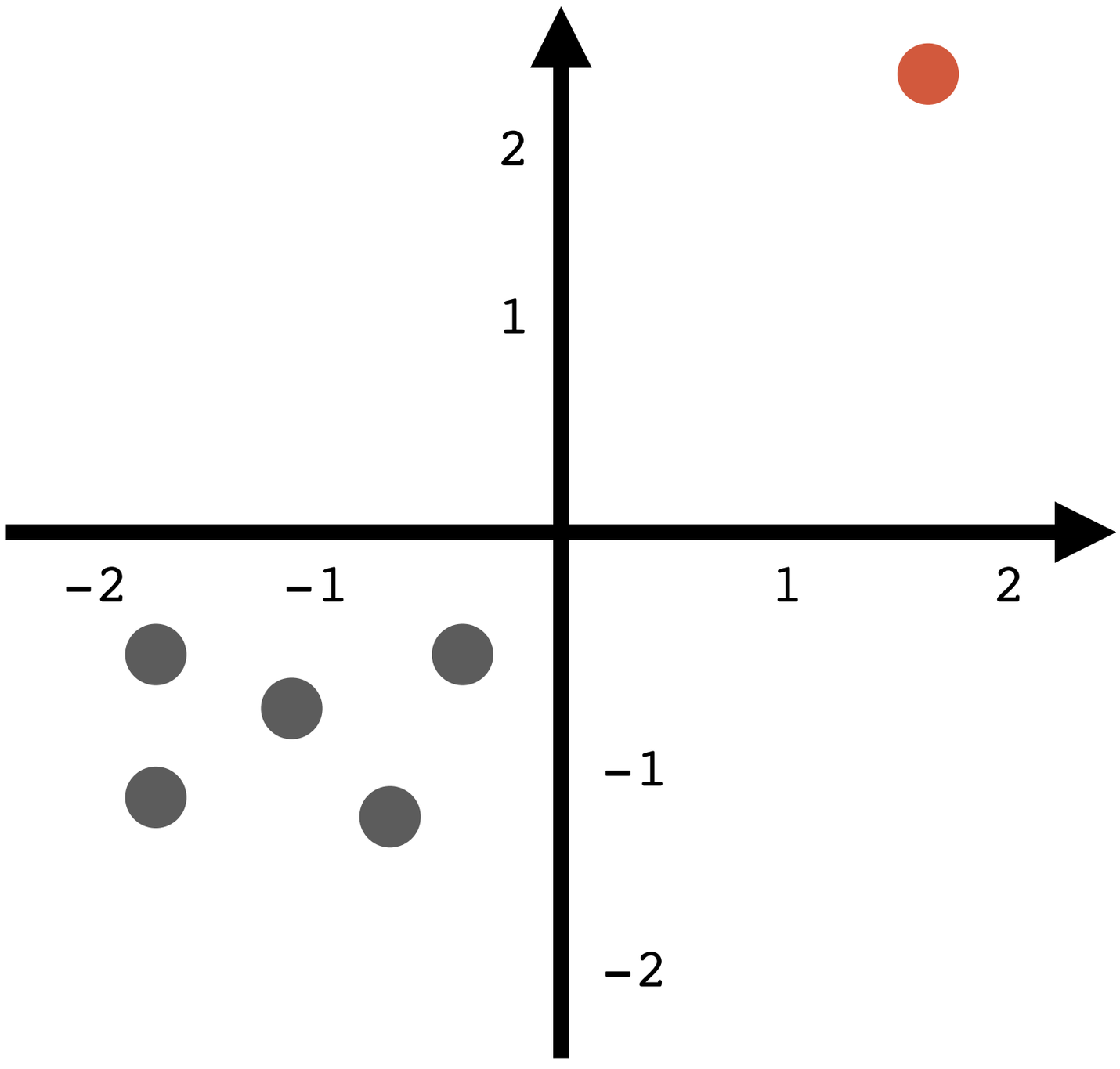


B

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

2 *What is correlation & how does it relate to causation?*

Remember to visualize your data!

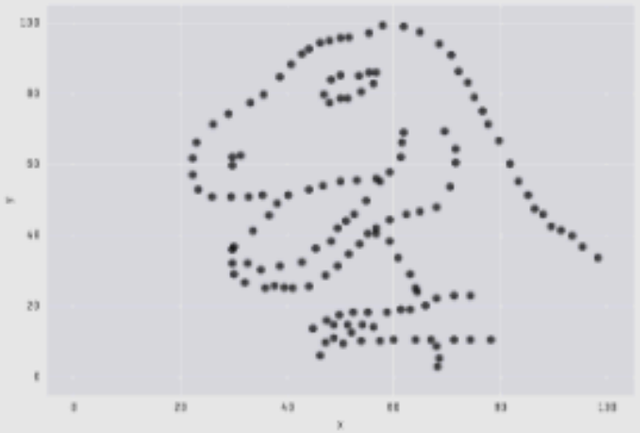


- Pearson's *r* is very sensitive to outliers.
- Also, you can calculate Pearson's r for any set of (x,y) coordinates .... but that doesn't mean that you are looking at a linear relationship!


100
80
60
40
20
0

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

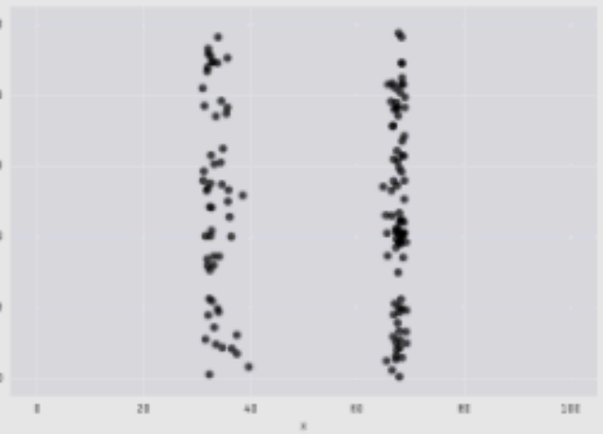
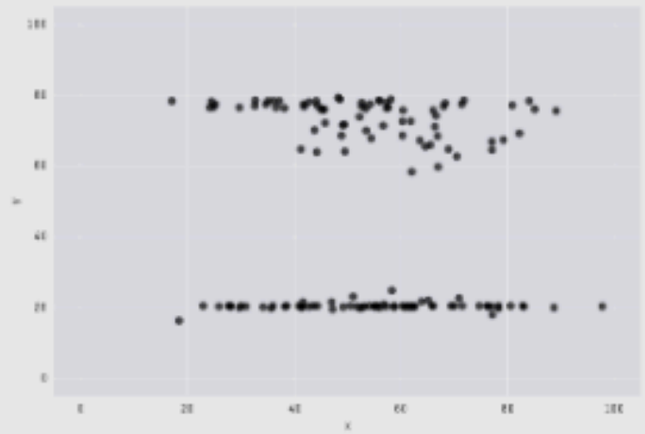
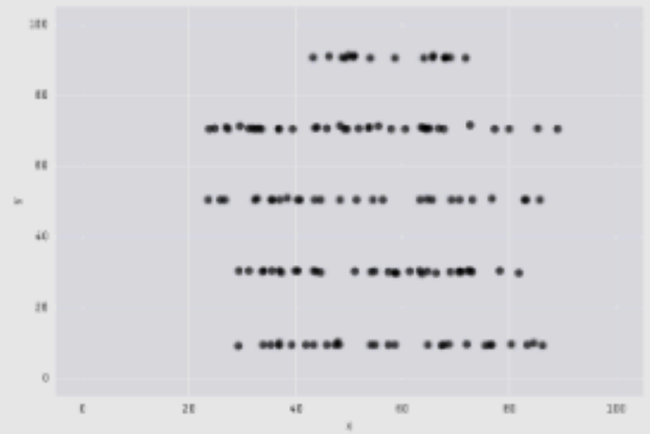
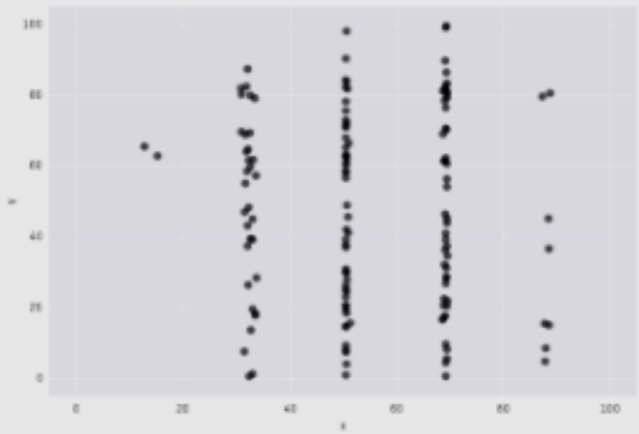
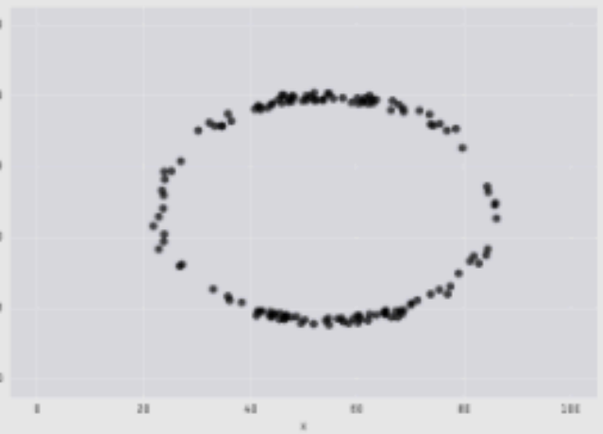
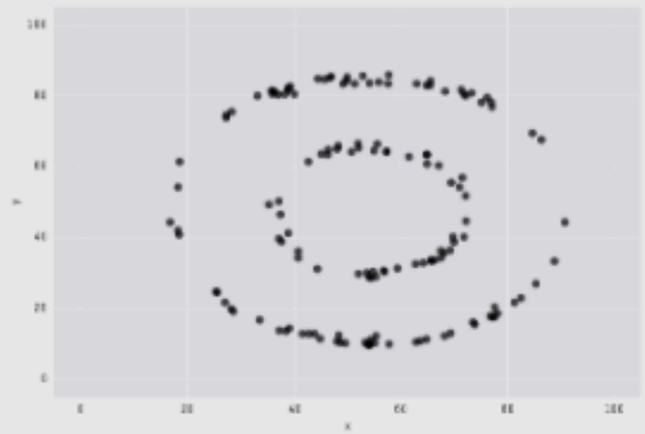
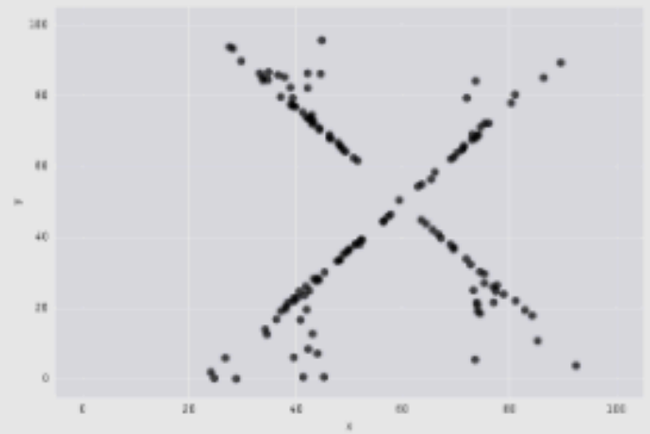
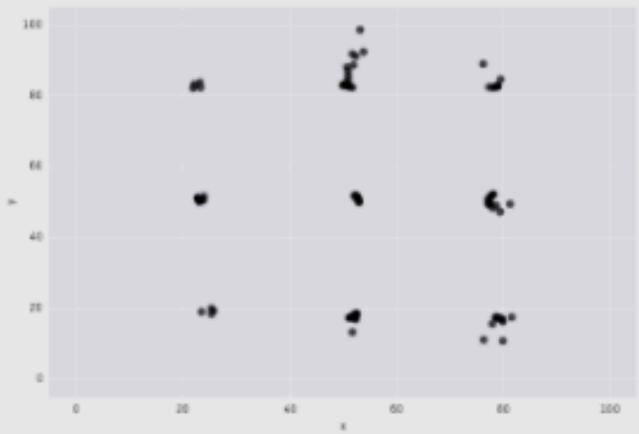
Summary statistics are identical in all 13 graphs.



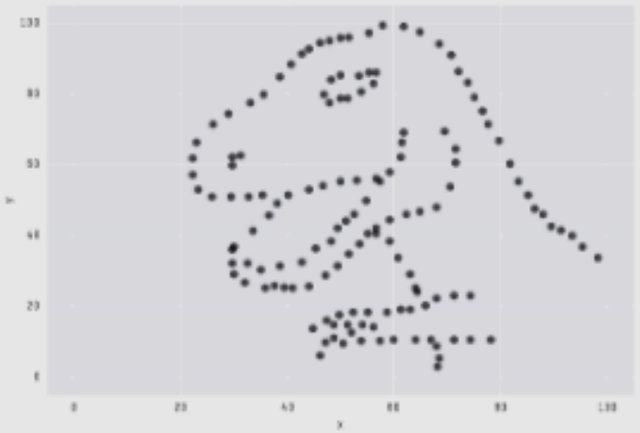
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06





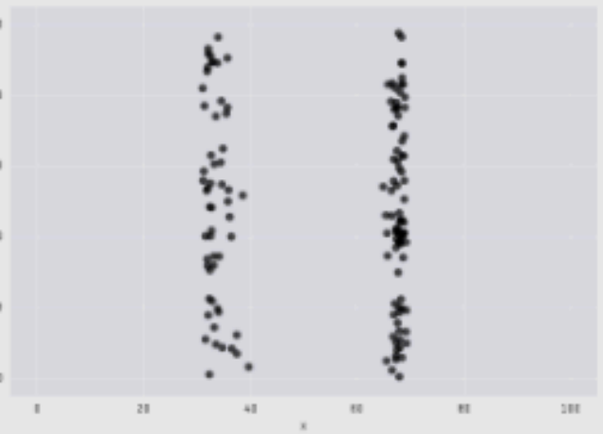
Summary statistics are identical in all 13 graphs.

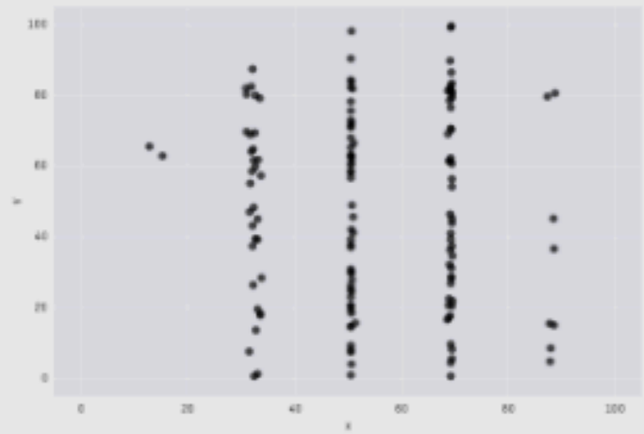


X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

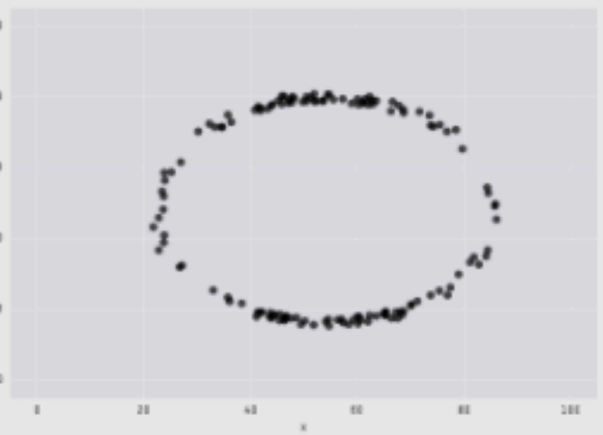
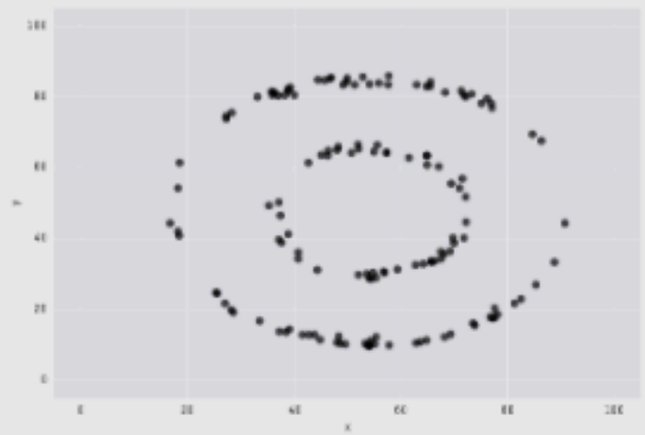
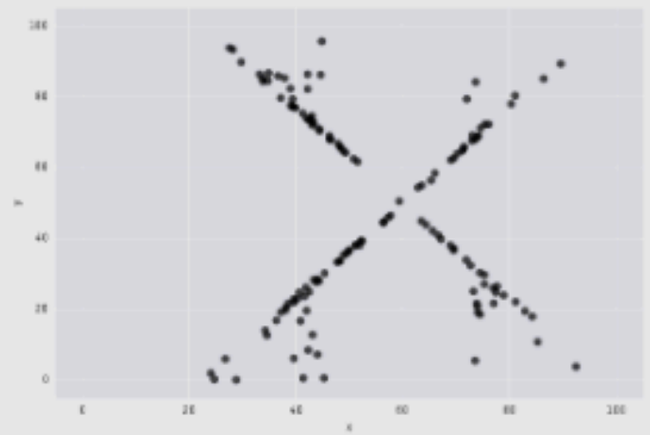
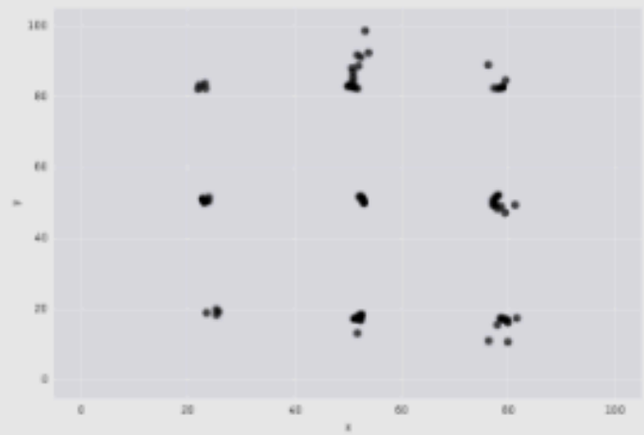






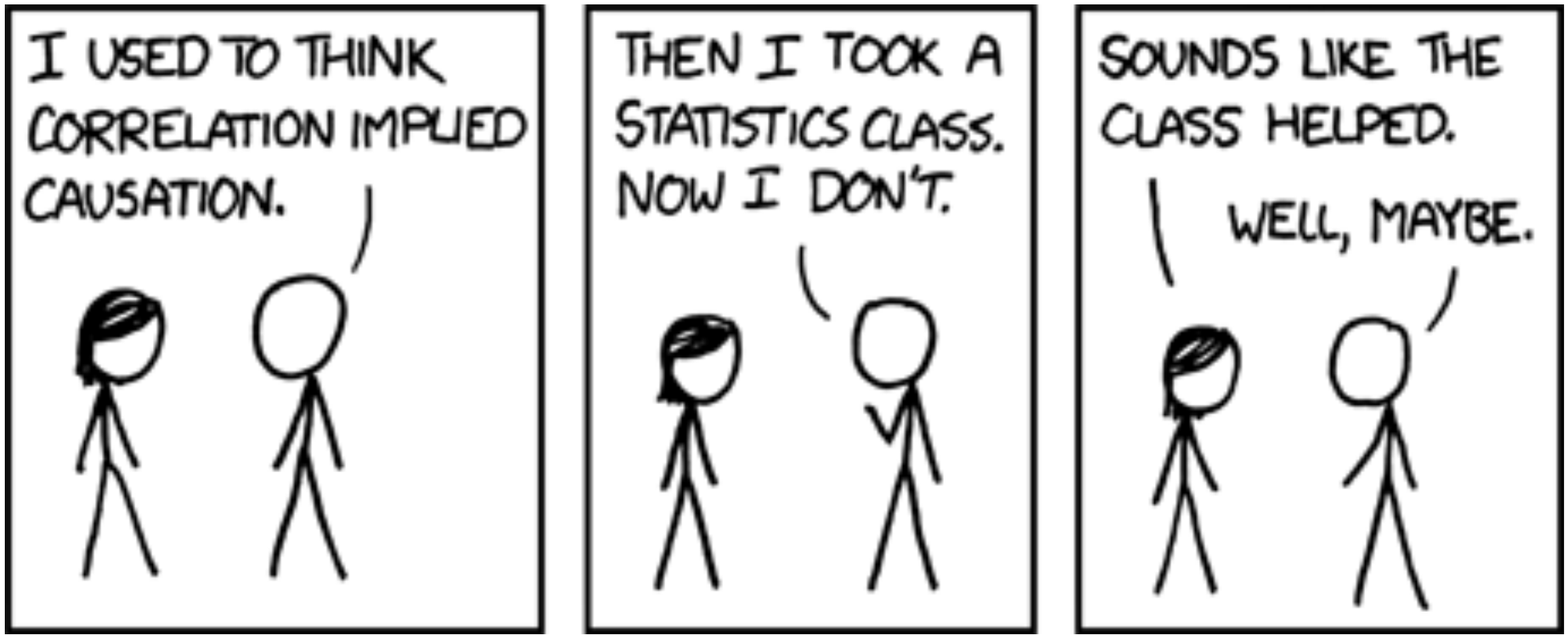



Don't be fooled: visualize your data!

2 *What is correlation & how does it relate to causation?*

Correlation and causation



https://xkcd.com/552/

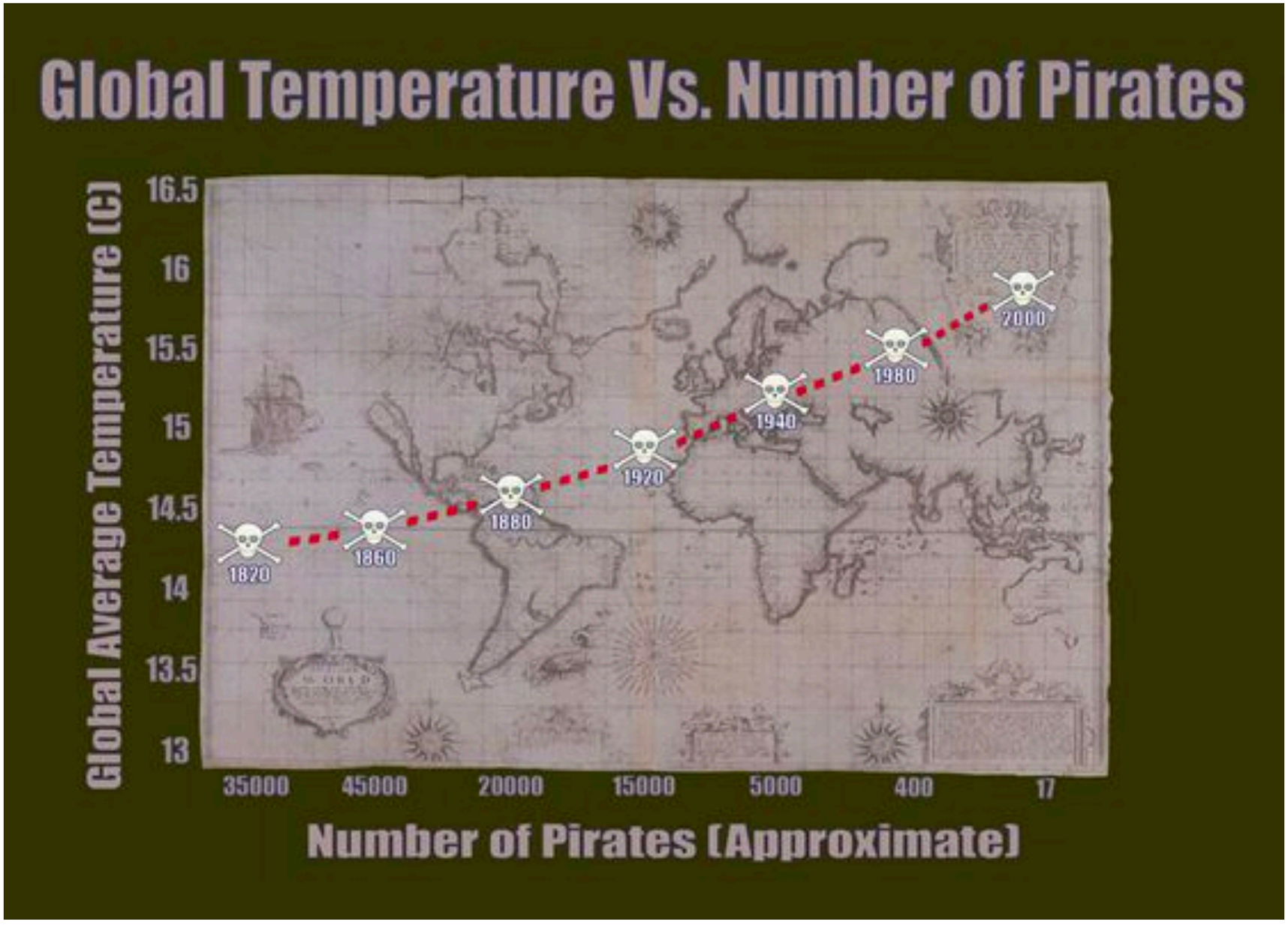
2 *What is correlation & how does it relate to causation?*

Correlation and causation

- So many of you might be familiar with the mantra: "correlation doesn't imply causation."
- It's a useful reminder because as human beings, we are extremely good at detecting patterns and finding meaning in the patterns we see.

2 *What is correlation & how does it relate to causation?*

Correlation and causation



https://www.forbes.com/sites/erikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming/

http://www.tylervigen.com/spurious-correlations

2 *What is correlation & how does it relate to causation?*

Correlation and causation

- So many of you might be familiar with the mantra: "correlation doesn't imply causation."
- It's a useful reminder because as human beings, we are extremely good at detecting patterns and finding meaning in the patterns we see.
- How should we think about the relationship between correlation and causation? Oftentimes observational data is the only kind available.

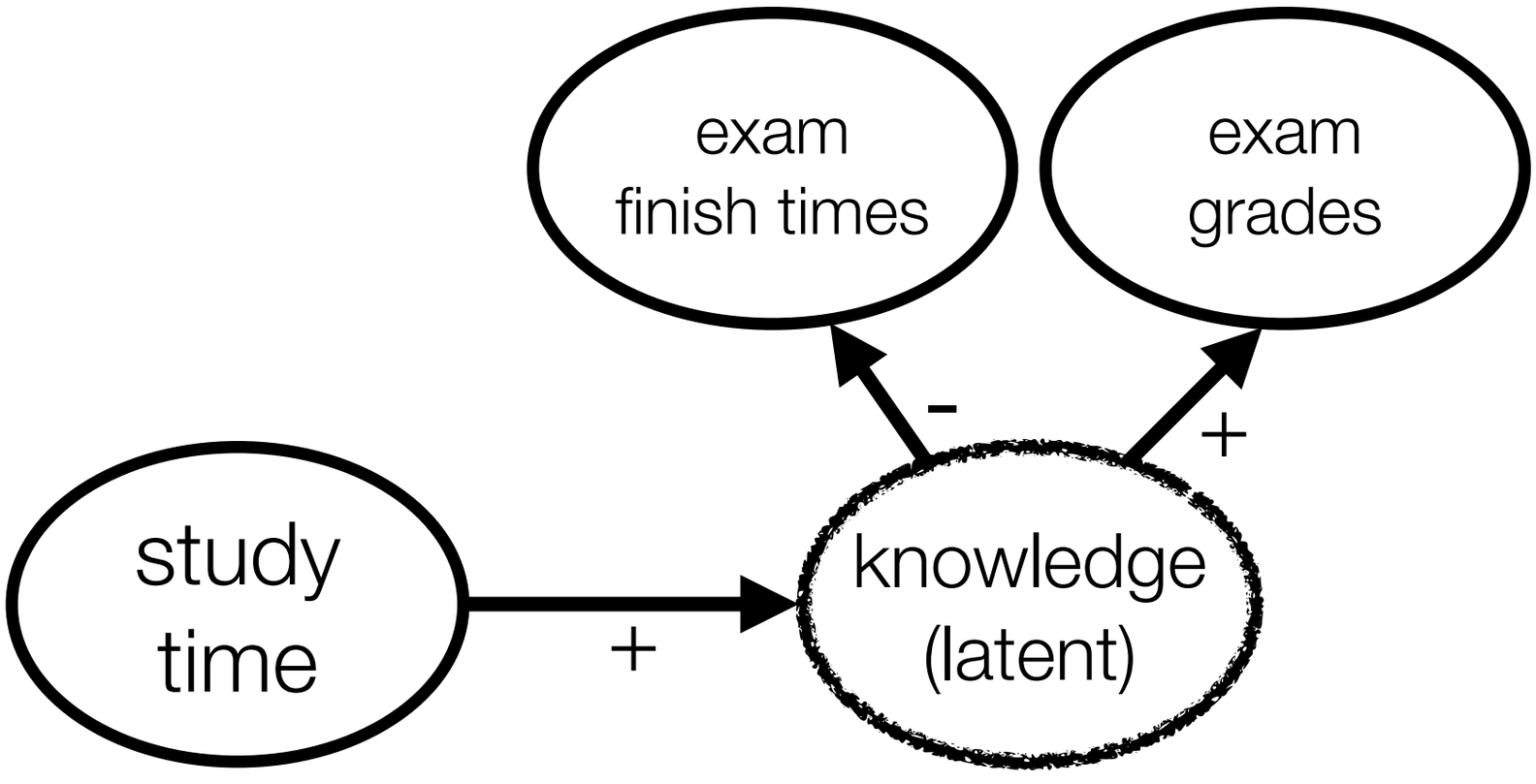
“

“Correlation does not imply causation, but it’s a pretty good hint.”

-Edward Tufte

Understanding causation using causal graphs

A causal graph describes the latent causal relations that give rise to the variables that we measure



arrows reflect
causal relations

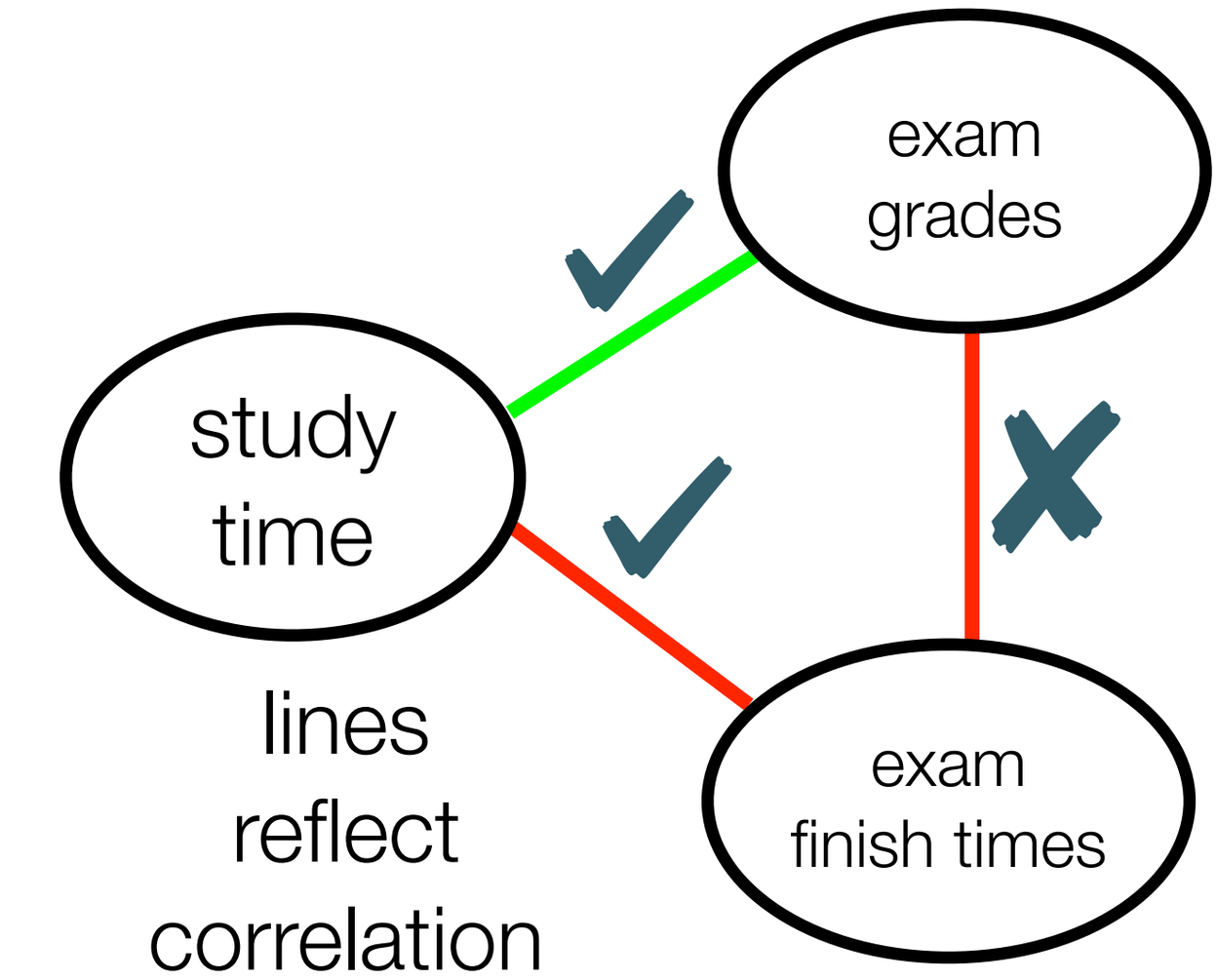
Causal relations mean that manipulating one variable will change another

Increasing study time will increase knowledge, which increases grades and reduces exam finishing time

2 *What is correlation & how does it relate to causation?*

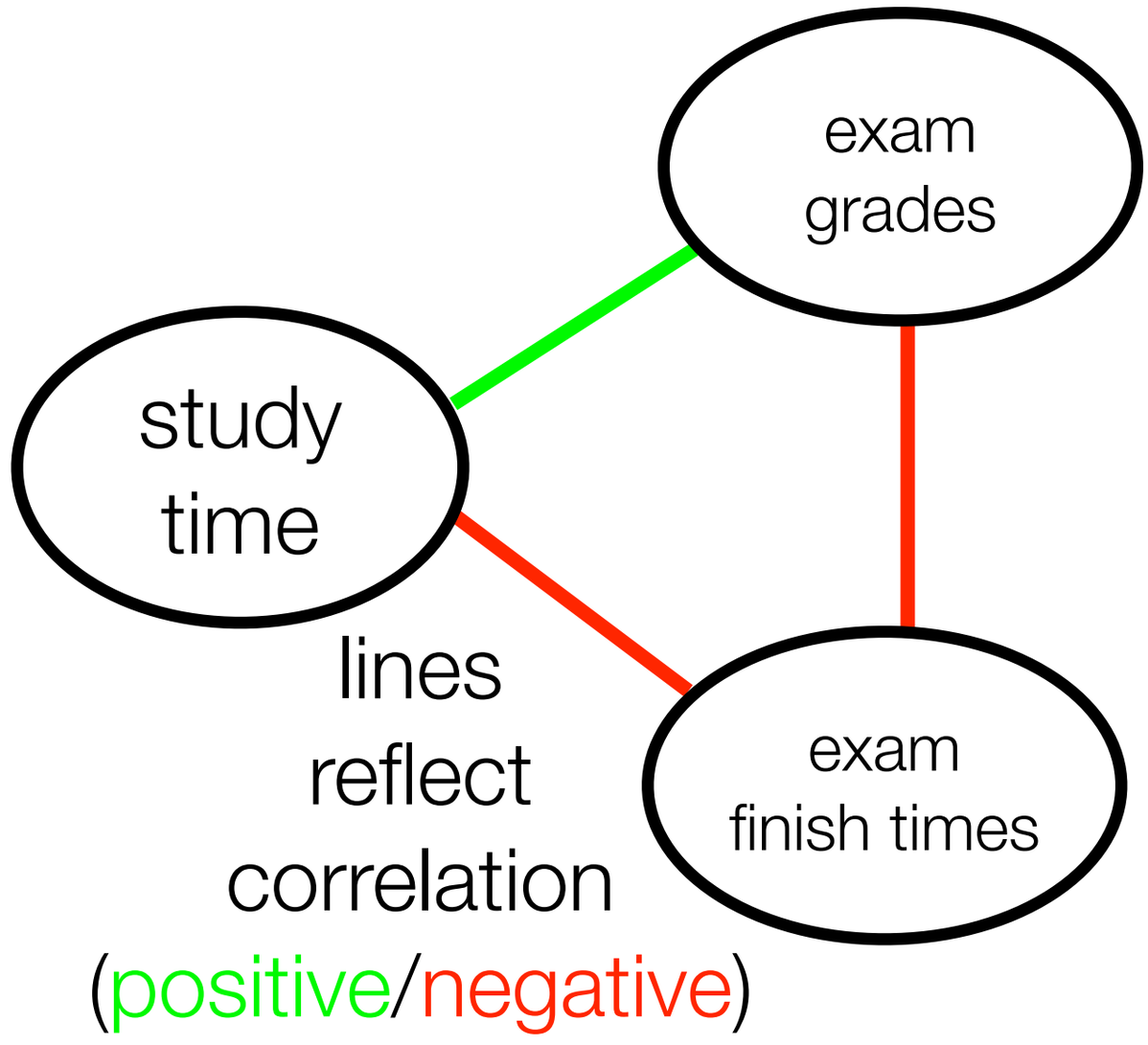
Correlation and causation

Correlations may reflect causal relations or the effects of common causes



Correlation and causation

- Correlations can sometimes imply the wrong causal relation
- Negative correlation between exam grades and exam finishing time
  - Might be interpreted to mean that finishing the exam faster will improve grades!
- So if we *only* measured exam grades & finish times in this study, we might fool ourselves!



Natural experiments

- Ideally, if we want to be able to draw stronger inferences about causal relationships between variables, we would run a **randomized controlled experiment**.
- But this isn't always possible! (e.g., randomly assigning smoking habits to individuals and tracking them over a long time)
- Instead, a still-powerful alternative is a **natural experiment.**
  - A natural experiment is an empirical study in which individuals (or clusters of individuals) are exposed to the experimental and control conditions that are determined by nature or by other factors outside the control of the investigators.

Example: Natural experiment exposing relationship between smoking and heart disease

- In Helena, Montana a smoking ban was in effect in all public spaces, including bars and restaurants, during the six-month period from June 2002 to December 2002.
- Helena is geographically isolated and served by only one hospital. The investigators observed that the rate of heart attacks dropped by 40% while the smoking ban was in effect.
- Opponents of the law prevailed in getting the enforcement of the law suspended after six months, after which the rate of heart attacks went back up.
- This study was an example of a natural experiment, called a case-crossover experiment, where the exposure is removed for a time and then returned.
- However, the inability to control variables in natural experiments can impede investigators from drawing firm conclusions. Critics argued that the particularly large percentage fluctuation in the rate of myocardial infarction was likely due to chance, given the small population size.

Today

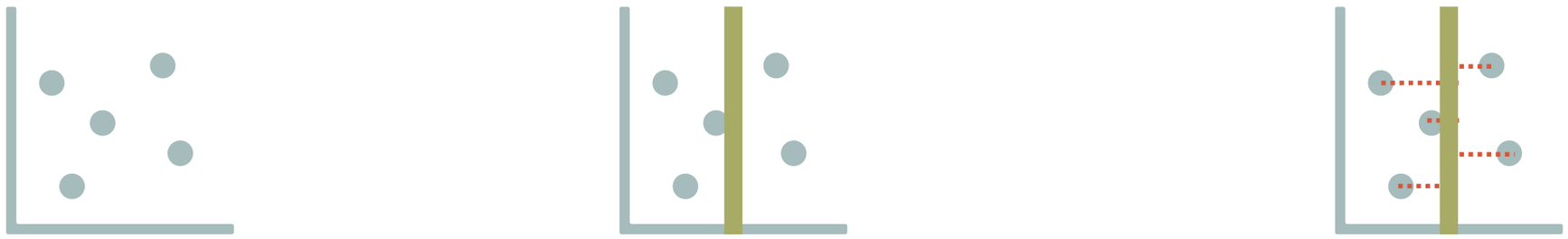
Lecture 19: Generalizing what you have learned

General announcements

What is correlation & how does it relate to causation?

How to extend your model to multiple predictors?

3 *How to extend your model to multiple predictors?*



data = model + error

what we actually observe

what we expect to observe

difference between expected and observed

3 *How to extend your model to multiple predictors?*

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*

3 *How to extend your model to multiple predictors?*

Example: making predictions using a linear model

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

observed value of outcome variable

intercept

slope

value of predictor variable

error

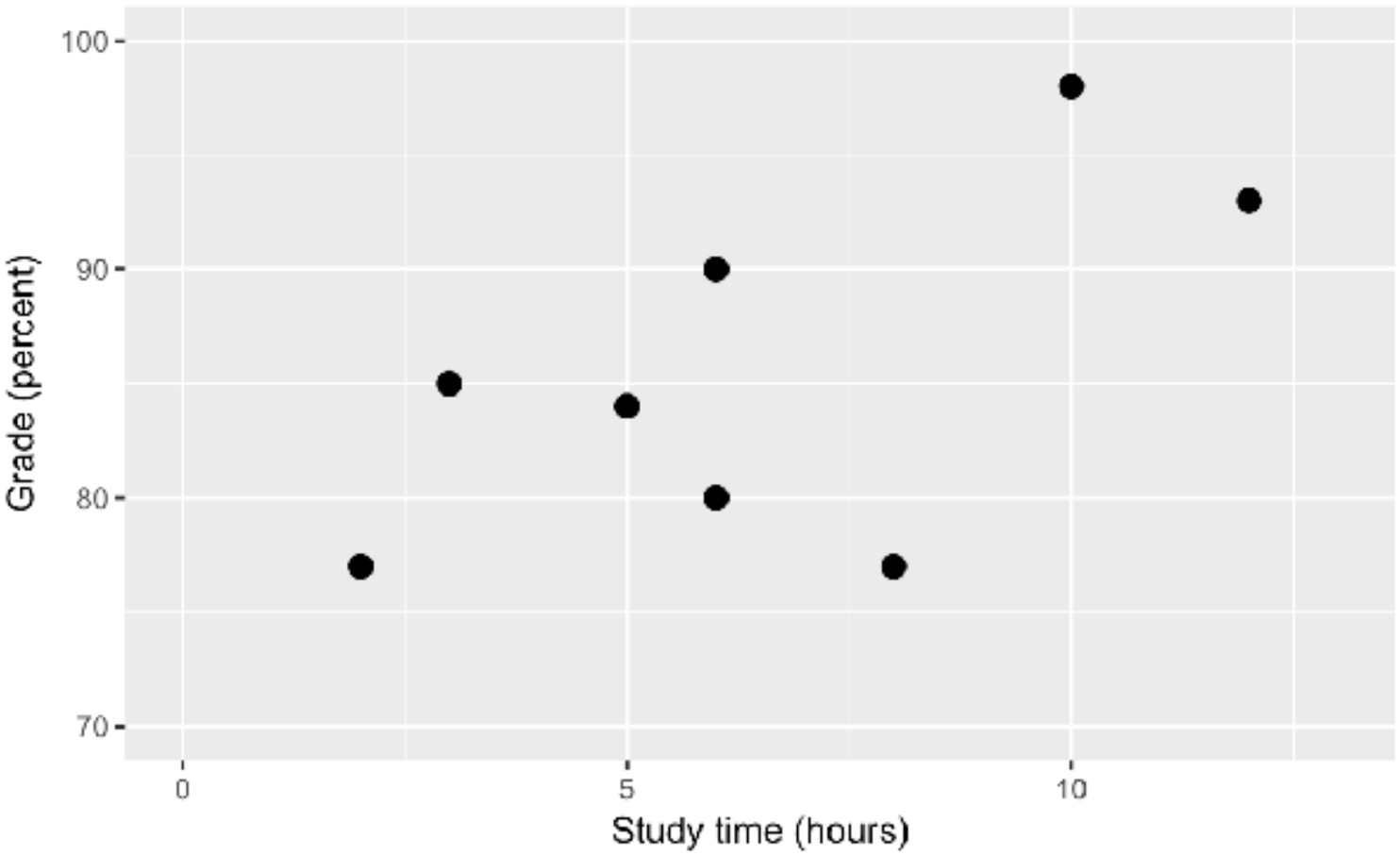
Suppose:

- **Your estimate of b0 (the y-intercept) = -3.**
- **Your estimate of b1 (the slope) = 2.5.**

What happens to Y if you increase X by 4?

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?

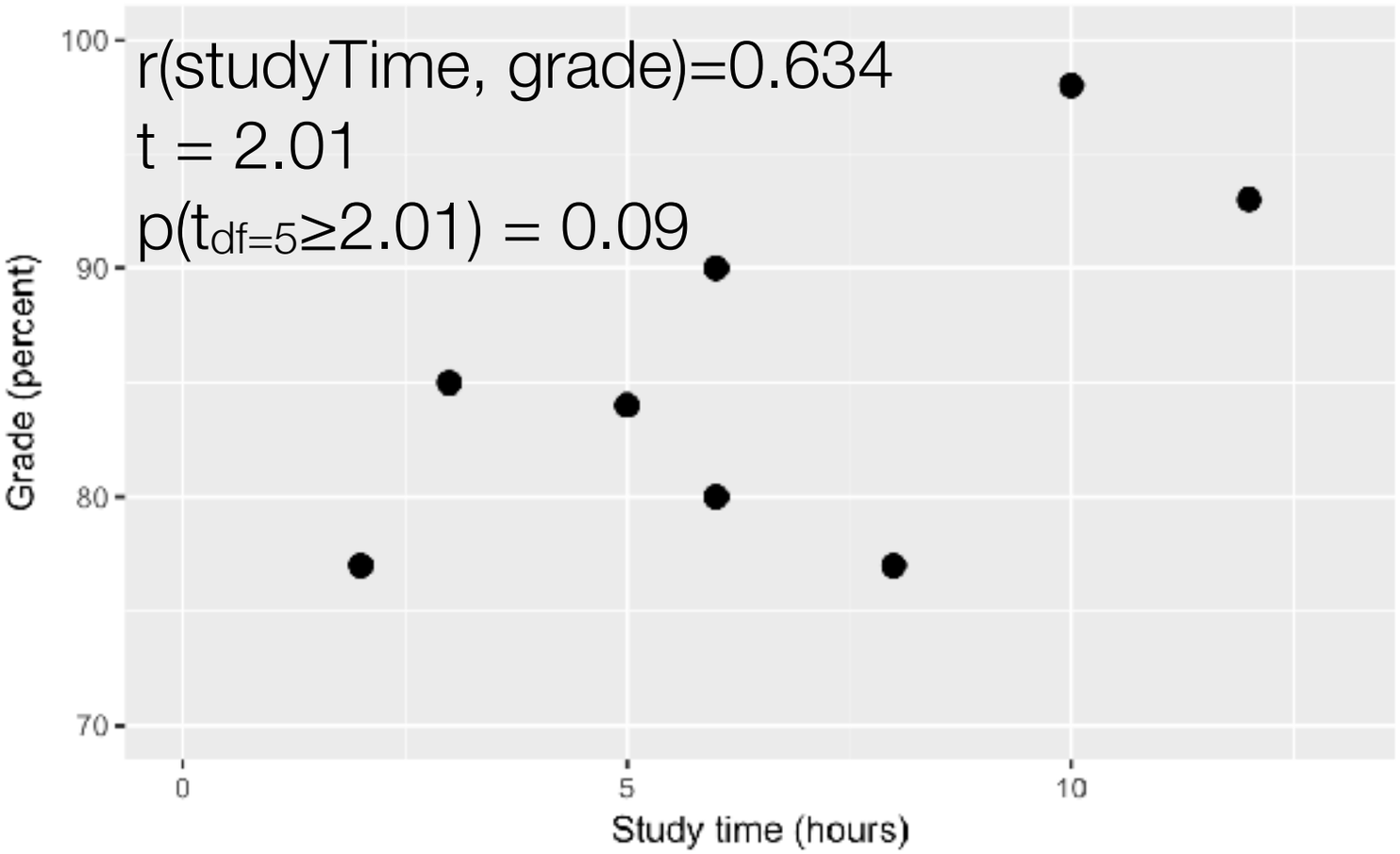


Two questions we might want to ask:

- Decide: Is there a relationship between study time and grade?
- Predict: Given a certain amount of study time, what grade would we predict?

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?



Two questions we might want to ask:

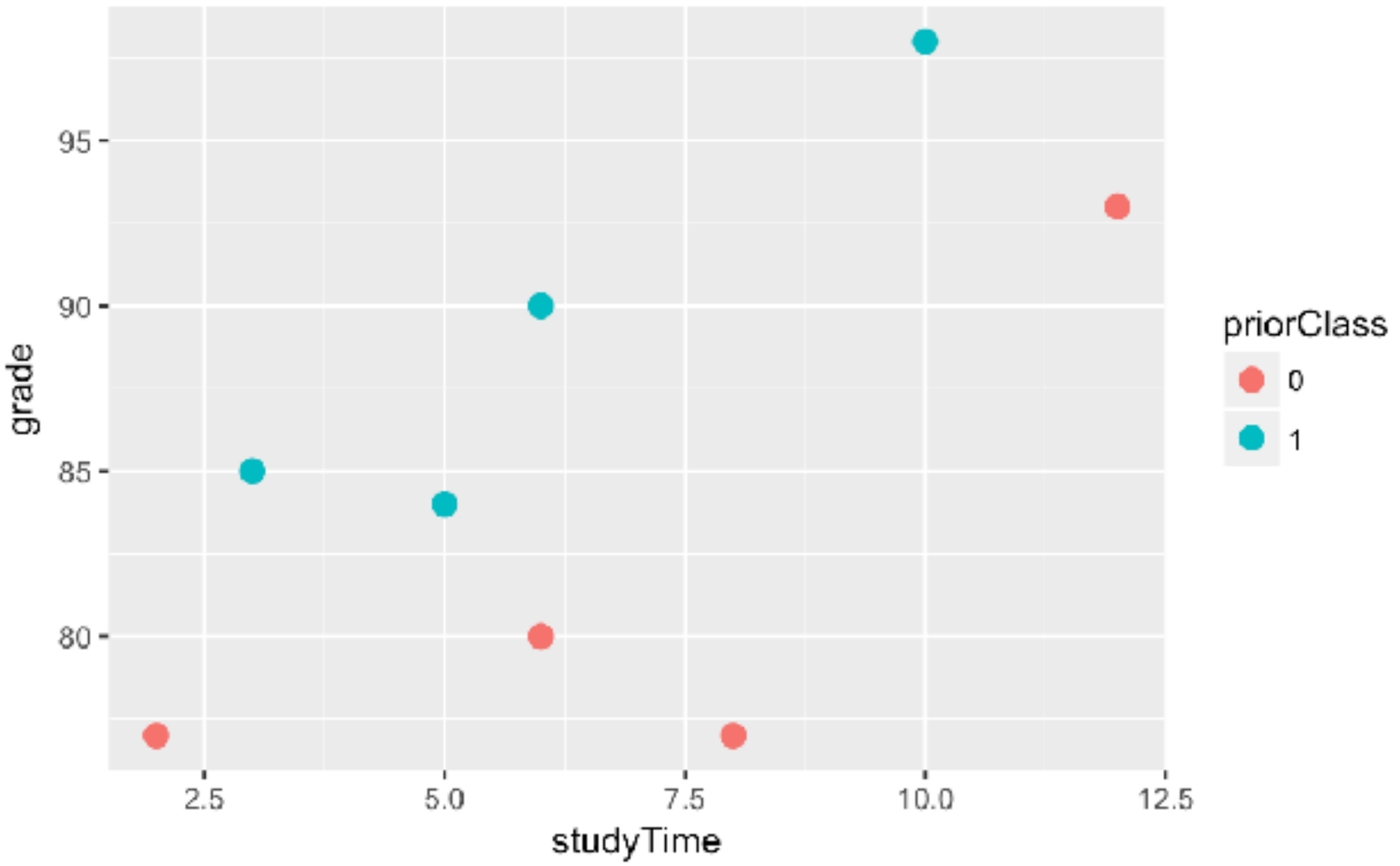
- Decide: Is there a relationship between study time and grade?
- Predict: Given a certain amount of study time, what grade would we predict?

```
lm(grade~studyTime,
        data=df)
Coefficients:
(Intercept)    studyTime
     76.156        1.438
```

$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_0$$

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?



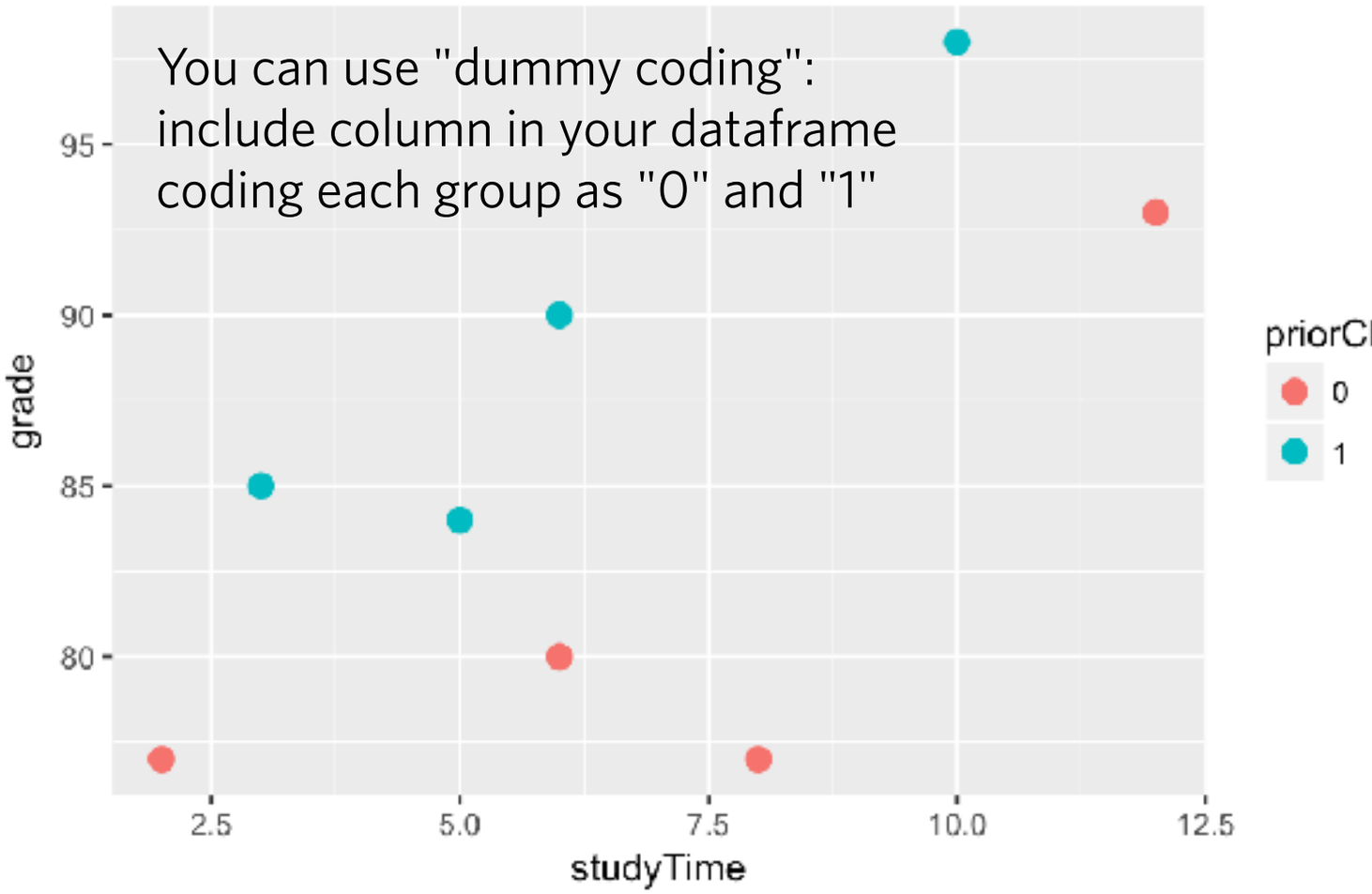
- Suppose that some students took a prior course on the topic
- Those students have substantially higher grades in the course, given the same amount of study time
- We can generate a more complex model that includes both of these factors

$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_2 * priorClass + \hat{\beta}_0$$

slope ***slope*** ***intercept***

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?



- Suppose that some students took a prior course on the topic
- Those students have substantially higher grades in the course, given the same amount of study time
- We can generate a more complex model that includes both of these factors

$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_2 * priorClass + \hat{\beta}_0$$

slope ***slope*** ***intercept***

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?

Model 0: Null model (intercept only)

$$grade = \hat{\beta}_0$$

Model 1: Simple linear regression (1 predictor)

$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_0$$

Model 2: Multiple linear regression (2 predictors)

$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_2 * priorClass + \hat{\beta}_0$$

3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?

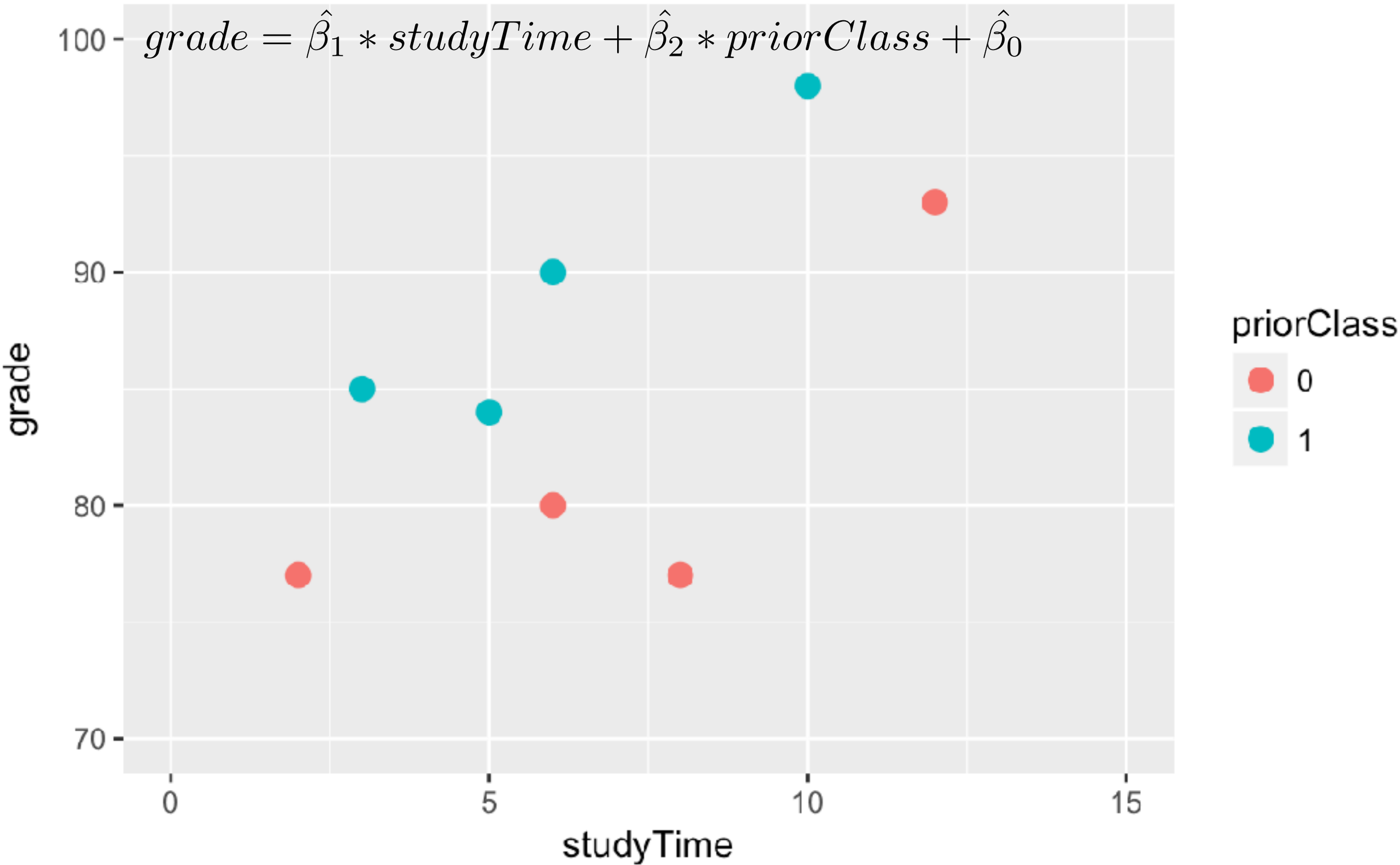
```
Call:
lm(formula = grade ~ studyTime + priorClass, data = df)

Residuals:
       1        2        3        4        5        6        7
8
 3.58333  0.75000 -3.58333 -0.08333  0.75000 -6.41667  2.08333
2.91667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.0833     3.7680  18.600 8.27e-06 ***
studyTime     1.6667     0.4553   3.661   0.0146 *
priorClass    9.1667     2.8793   3.184   0.0244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.021 on 5 degrees of freedom
Multiple R-squared:  0.8028,   Adjusted R-squared:  0.724
F-statistic: 10.18 on 2 and 5 DF,  p-value: 0.01726
```

3

How to extend your linear model to multiple predictors?

Example: Are exam grade and study time related?

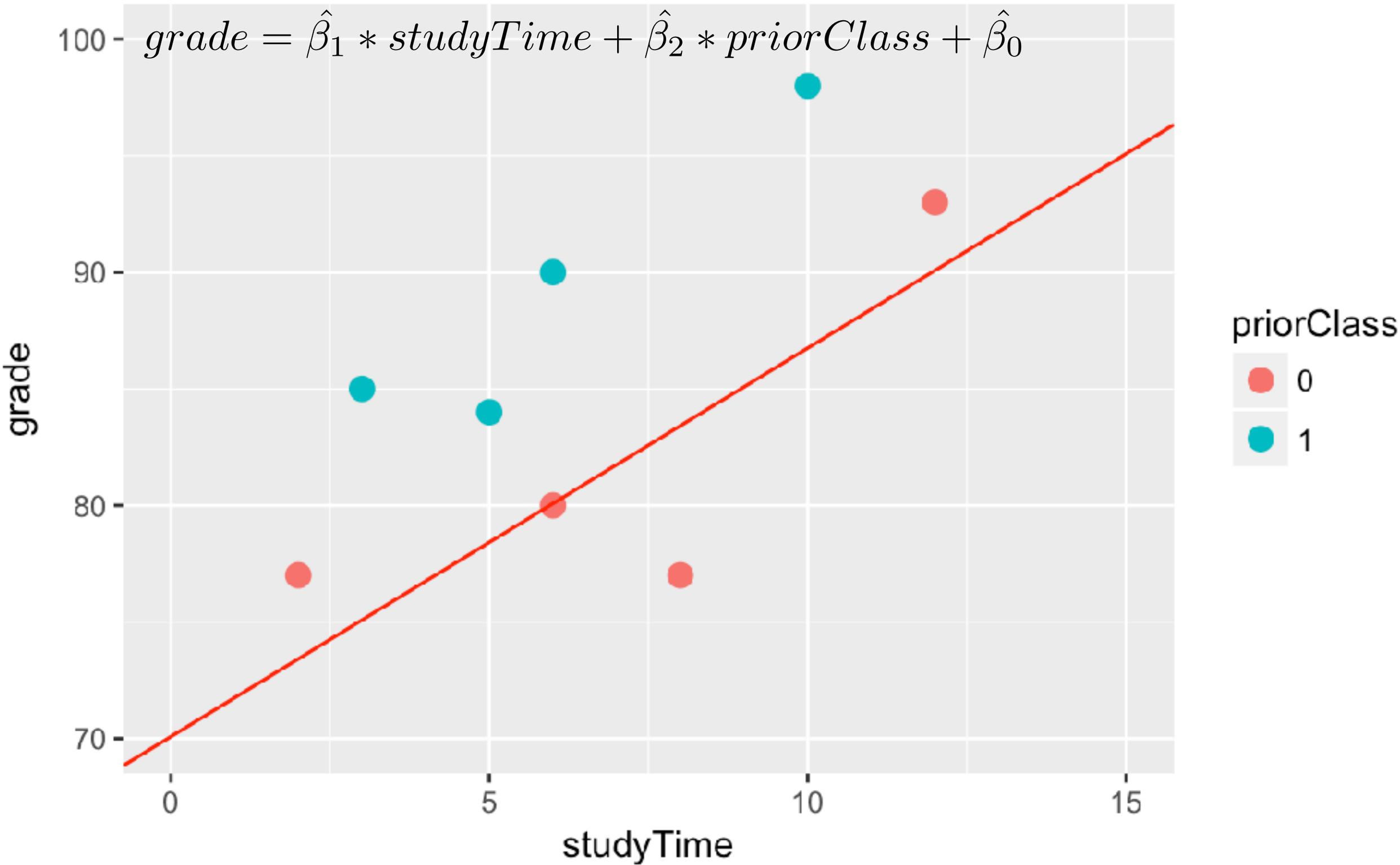
$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_2 * priorClass + \hat{\beta}_0$$



3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?

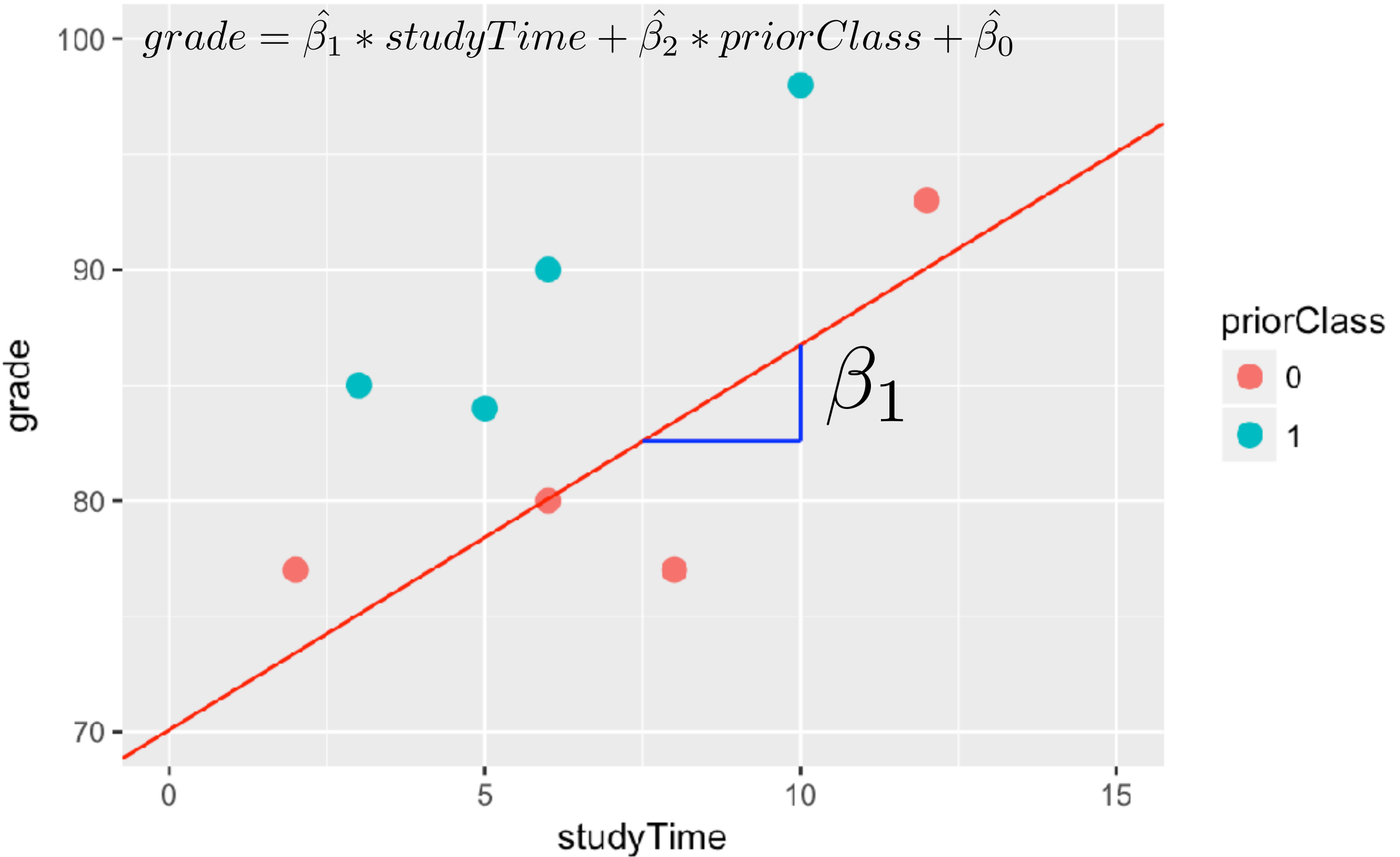
$$grade = \hat{\beta}_1 * studyTime + \hat{\beta}_2 * priorClass + \hat{\beta}_0$$



3

How to extend your linear model to multiple predictors?

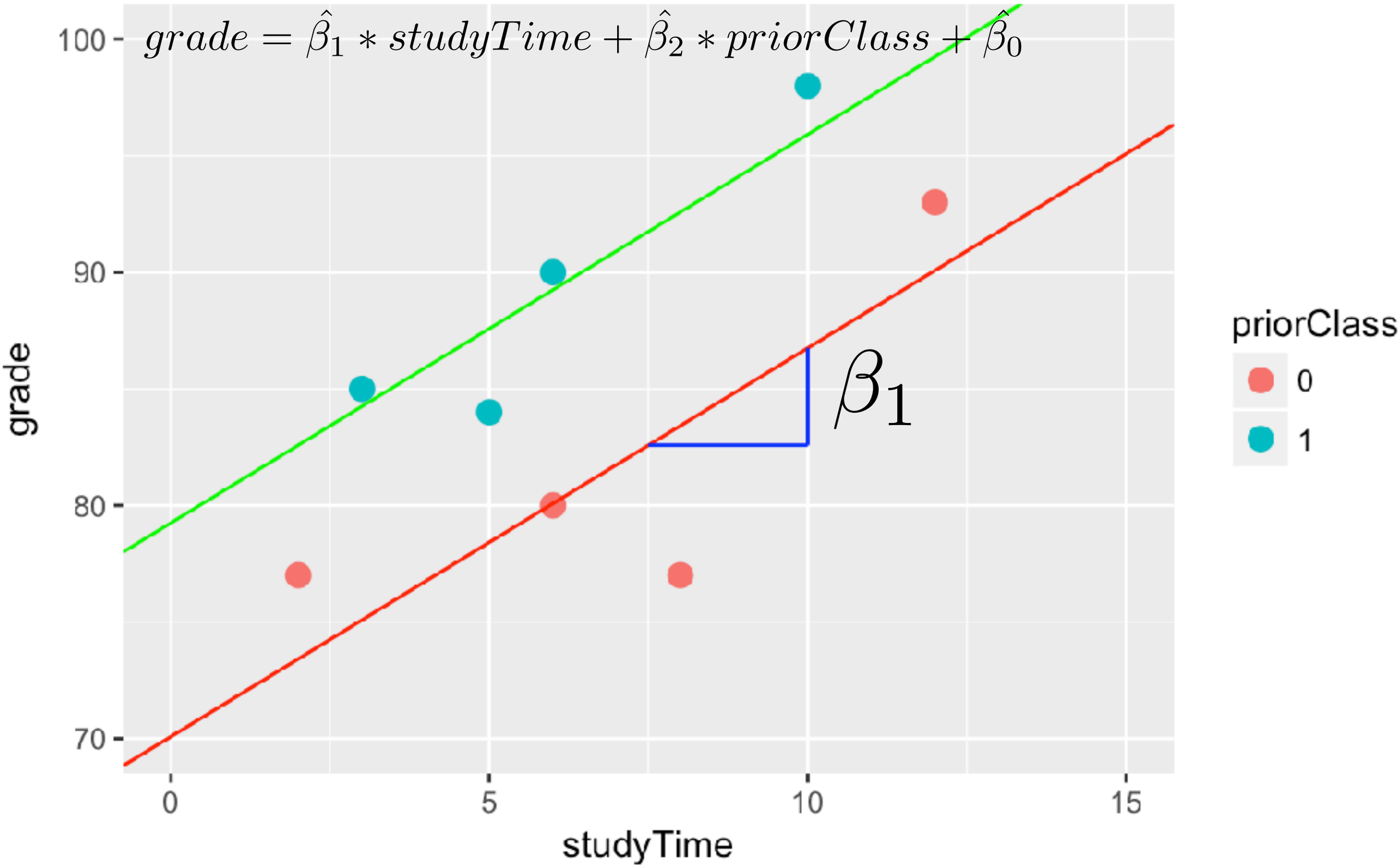
Example: Are exam grade and study time related?



3

How to extend your linear model to multiple predictors?

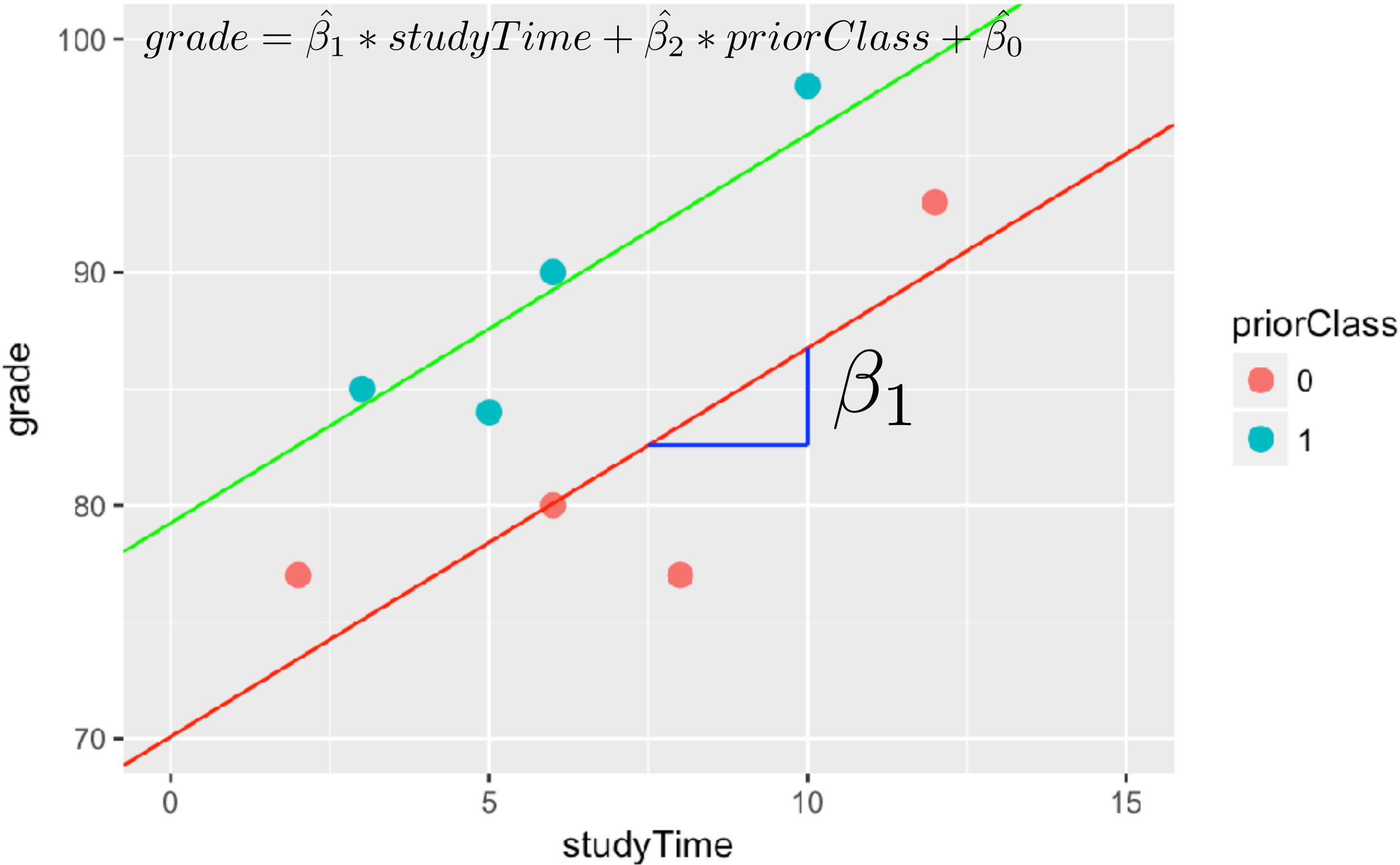
Example: Are exam grade and study time related?



3

How to extend your linear model to multiple predictors?

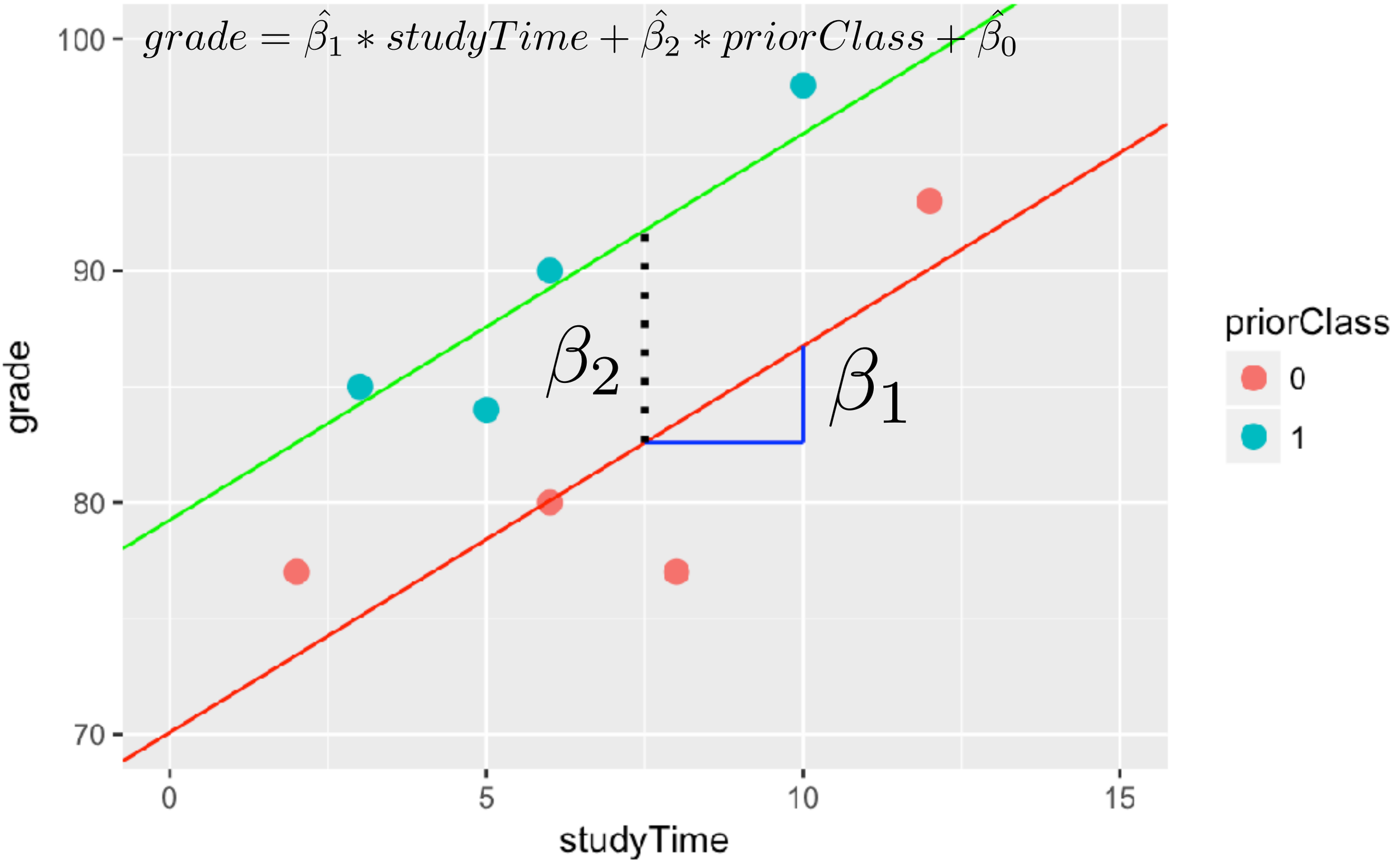
Example: Are exam grade and study time related?



3

How to extend your linear model to multiple predictors?

Example: Are exam grade and study time related?

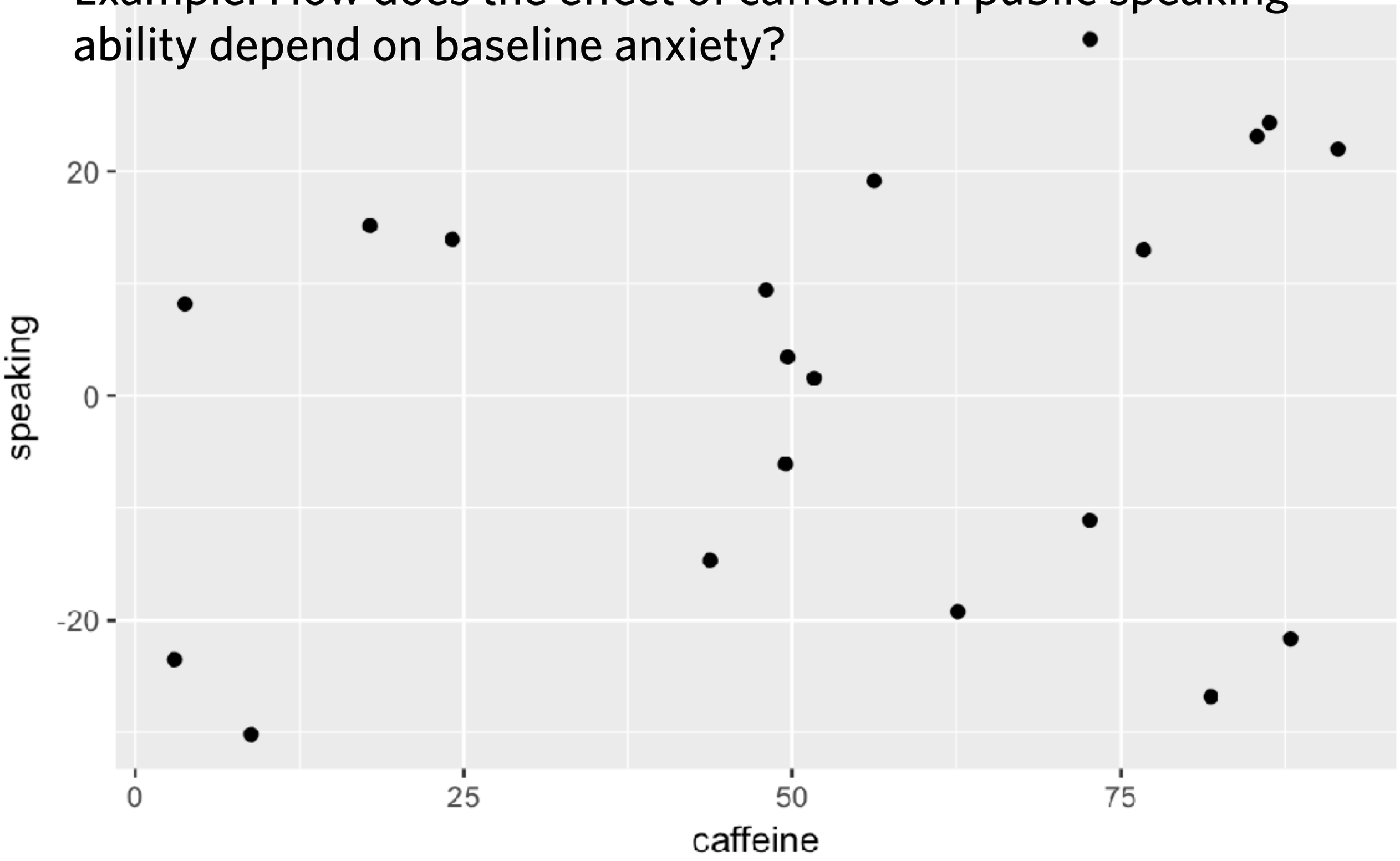


What is a statistical interaction?

- Sometimes the effect of one predictor variable *depends* on another variable.
- For example, "How does the effect of caffeine on public speaking ability depend on baseline anxiety?"
- When that is the case, we say that these two variables "interact with each other" or that "there is an interaction between these two variables."

3 *How to extend your linear model to multiple predictors?*

Example: How does the effect of caffeine on public speaking ability depend on baseline anxiety?



3 *How to extend your linear model to multiple predictors?*

Example: How does the effect of caffeine on public speaking ability depend on baseline anxiety?

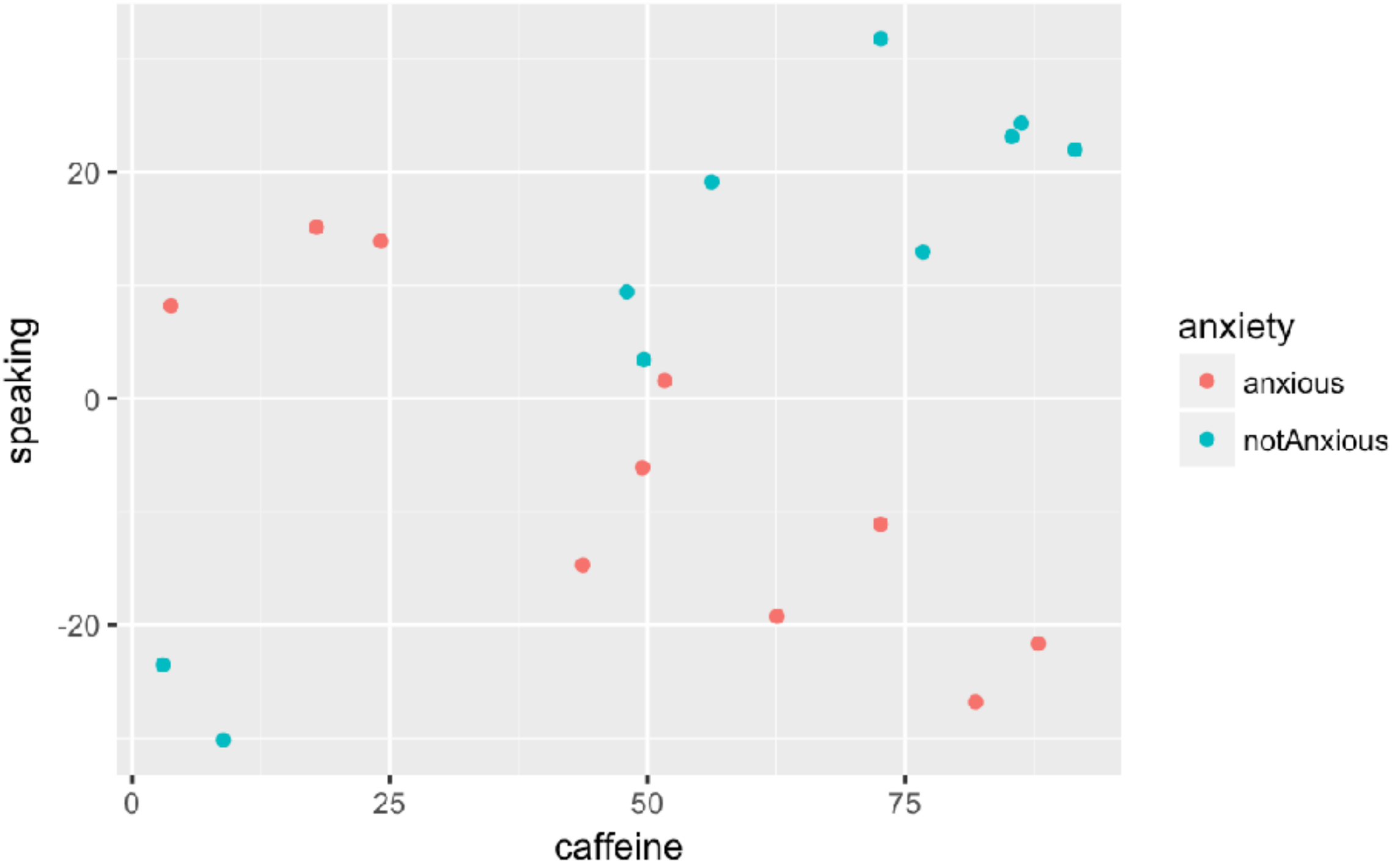
```
lmResultCaffeine = lm(speaking~caffeine,data=df)
summary(lmResultCaffeine)

Call:
lm(formula = speaking ~ caffeine, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-33.10 -16.02   5.01  16.45  26.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.413      9.165   -0.81     0.43
caffeine       0.168      0.151    1.11     0.28
```

What is a statistical interaction?

- Sometimes the effect of one predictor variable *depends* on another variable.
- For example, "How does the effect of caffeine on public speaking ability depend on baseline anxiety?"
  - By making non-anxious people better at speaking
  - By making anxious people worse at speaking
- When that is the case, we say that these two variables "interact with each other" or that "there is an interaction between these two variables."

3 *How to extend your linear model to multiple predictors?*

Example: How does the effect of caffeine on public speaking ability depend on baseline anxiety?



3 *How to extend your linear model to multiple predictors?*

Example: How does the effect of caffeine on public speaking ability depend on baseline anxiety?

```
lmResultCafAnx = lm(speaking ~ caffeine + anxiety,data=df)
summary(lmResultCafAnx)

Call:
lm(formula = speaking ~ caffeine + anxiety, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-32.97  -9.74    1.35   10.53   25.36

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -12.581      9.197   -1.37     0.19
caffeine             0.131      0.145    0.91     0.38
anxietynotAnxious   14.233      8.232    1.73     0.10
```

"main effects"

3 *How to extend your linear model to multiple predictors?*

Example: How does the effect of caffeine on public speaking ability depend on baseline anxiety?

- **An interaction occurs when the effect of one variable depends on the value of another variable**
- To include an interaction between variables in a model, we multiply them in the formula:

```
lm(formula = speaking ~ caffeine + anxiety +
     caffeine * anxiety, data = df)
```

3 *How to extend your linear model to multiple predictors?*

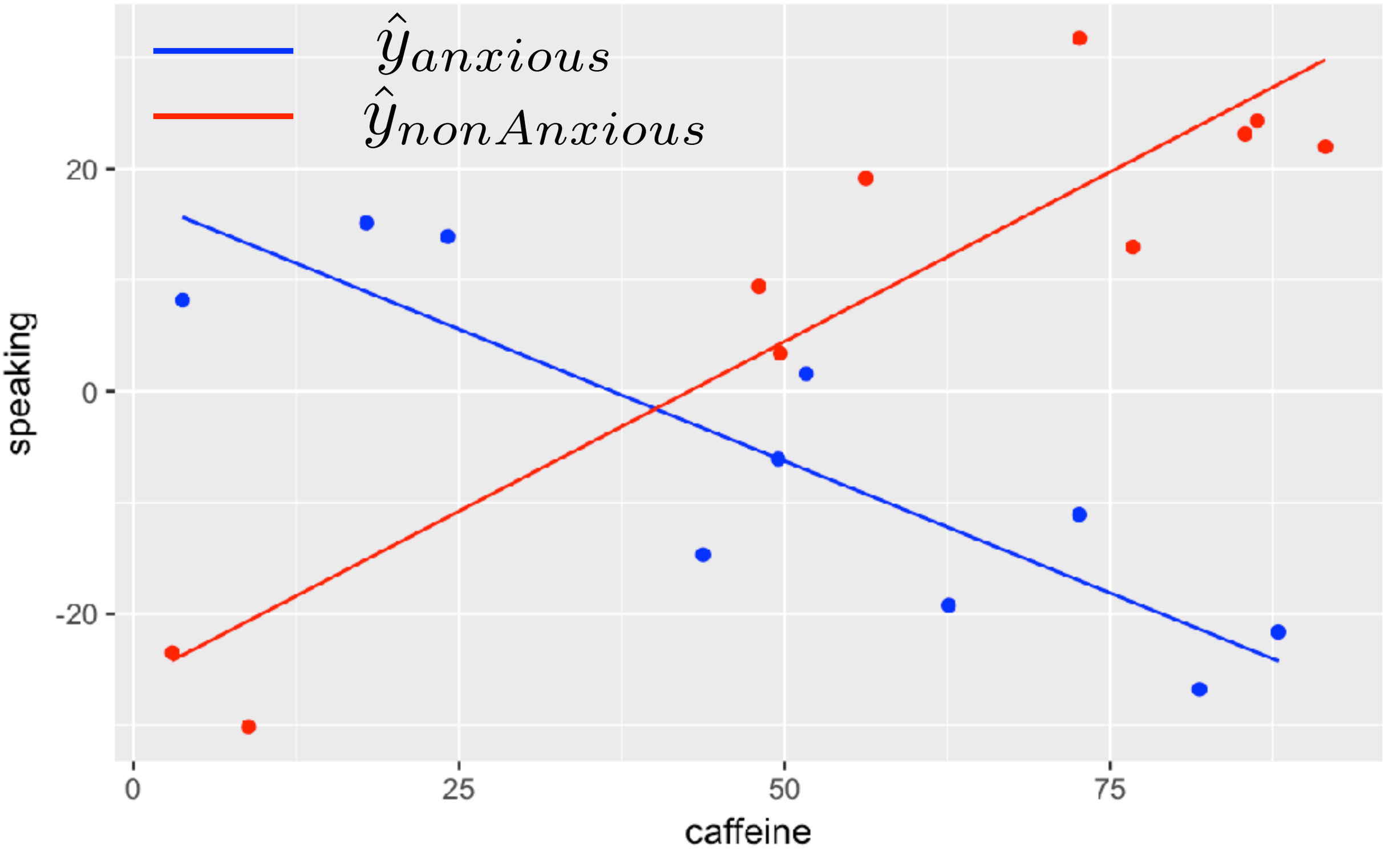
```
Call:
lm(formula = speaking ~ caffeine + anxiety + caffeine * anxiety,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-11.385  -7.103  -0.444   6.171  13.458

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                17.4308     5.4301    3.21  0.00546 **
caffeine                   -0.4742     0.0966   -4.91  0.00016 ***
anxietynotAnxious         -43.4487     7.7914   -5.58  4.2e-05 ***
caffeine:anxietynotAnxious  1.0839     0.1293    8.38  3.0e-07 ***
```

interaction term

how much does the effect of caffeine differ between anxious and non-anxious people?

3 *How to extend your linear model to multiple predictors?*

$$\hat{y} = \beta_1 * caffeine + \beta_2 * anxiety + \beta_3 * (caffeine * anxiety) + \beta_0$$



3 *How to extend your linear model to multiple predictors?*

Example: Are exam grade and study time related?

Model 0: Null model (intercept only)

$$publicSpeaking = \hat{\beta}_0$$

Model 1: Simple linear regression (1 predictor)

$$publicSpeaking = \hat{\beta}_0 + \hat{\beta}_1 * caffeine$$

Model 2: Multiple linear regression (2 predictors)

$$publicSpeaking = \hat{\beta}_0 + \hat{\beta}_1 * caffeine + \hat{\beta}_2 * anxiety$$

Model 3: Multiple linear regression (2 predictors w/ interaction term)

$$publicSpeaking = \hat{\beta}_0 + \hat{\beta}_1 * caffeine + \hat{\beta}_2 * anxiety + \hat{\beta}_3 * (caffeine * anxiety)$$

Today

Lecture 19: Generalizing what you have learned

General announcements

What is correlation & how does it relate to causation?

How to extend your model to multiple predictors?

Student Daily Feedback Survey

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

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

Feedback

We welcome student fe[illegible] d your TA a Slack message, or[illegible] 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