DATA ANALYSIS

Week 4: Correlation

logistics: midterm 1

- Feb 23rd 2024 (Friday)
- content: Week 1 through Week 5
- two parts
  - in-person conceptual portion (quiz-like)
    - closed book (+ help sheet)
  - take-home computational portion (problem set-like: due Monday)
    - open book but NOT open person
- practice questions will be made available before Week 5

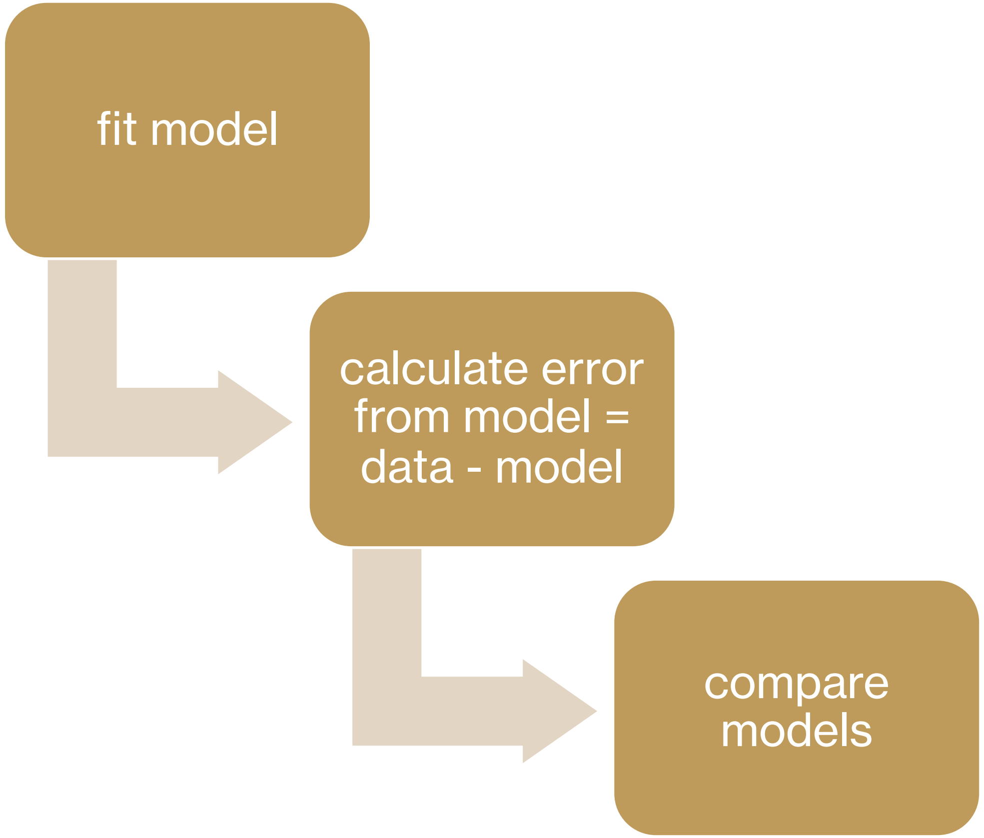
today's agenda

correlation

regression

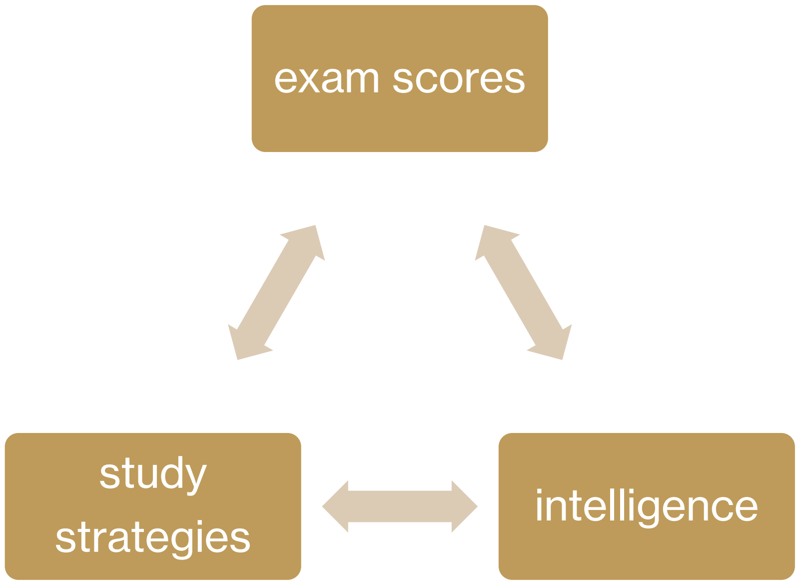
data = model + error

- simple but extremely powerful idea
- the types of “models” we have considered so far have been very simple
    - mean / median / mode
    - simply describe the data or variable based on its own characteristics
- often, we are interested in the relationships between variables



modeling relationships

- we often want to determine the relationship between two or more variables
- the statistical approach typically then becomes:
  - data (variable 1) = model (variables 2, 3, etc.) + error
- research question: how well can a set of variables (IVs) explain the variation in a key variable (DV)?

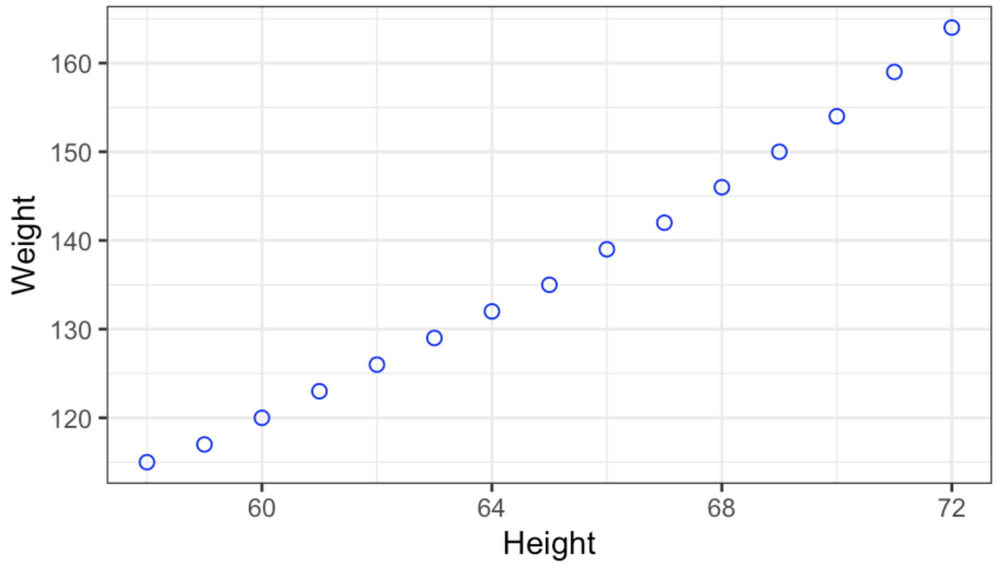


example

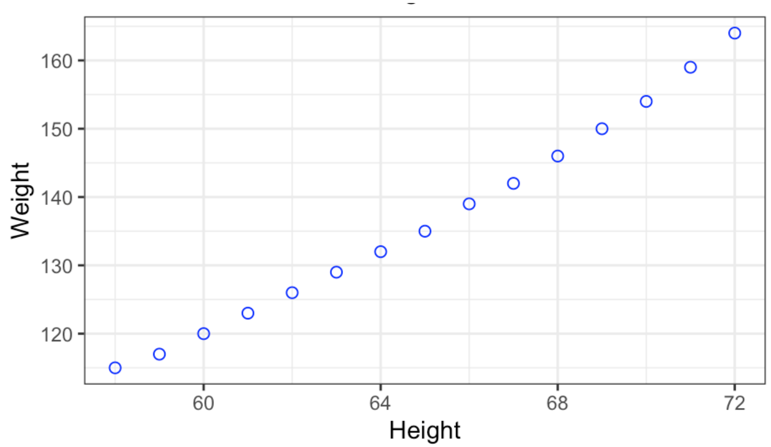
- a dataset of heights and weights for American women aged 30–39
- research question(s):
  - is there a relationship between height and weight?
  - how well can height explain the variation in weight?
- what causes weights to vary?
  - weight could vary independently of height
  - weight could vary with height
- we could represent the problem graphically
- we could formulate a preliminary model

  weight = b(height) + error

Woman	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164



covariance



- weight and height are on very different scales
- how can we bring them to the same scale? z-scores!
  - mean (z_{height}) = mean (z_{weight}) = 0
  - σ (z_{height}) = σ (z_{weight}) = 1
- once we have them on the same scale (their variances are the same), we can look at how weight and height *co-vary*
  - we multiply the z-scores together: $z_x z_y$
  - average them together to get an "average" estimate of covariance: $\frac{\sum z_x z_y}{N}$

Woman	z_height	z_weight	z_h*z_w	r
1	-1.62037037	-1.451485967	2.351676046	0.9954947681
2	-1.388888889	-1.317913639	1.830226406	
3	-1.157407407	-1.117555146	1.293318772	
4	-0.9259259259	-0.9171966539	0.8491590982	
5	-0.6944444444	-0.7168381616	0.497747384	
6	-0.462962963	-0.5164796692	0.2390836296	
7	-0.2314814815	-0.3161211768	0.07316783491	
8	0	-0.1157626845	0	
9	0.2314814815	0.151381972	0.03503811814	
10	0.462962963	0.3517404644	0.162824196	
11	0.6944444444	0.6188851209	0.4297322136	
12	0.9259259259	0.8860297774	0.8203041774	
13	1.157407407	1.153174434	1.334540088	
14	1.388888889	1.487105254	2.065187904	
15	1.62037037	1.821036075	2.950415653	

Pearson's *r* (correlation)

- measures the degree and direction of a **<u>linear</u>** relationship between two variables (X and Y)

$$r = \frac{degree\ to\ which\ two\ variables\ vary\ together\ (covary)}{degree\ to\ which\ two\ variables\ vary\ independently}$$

- degree
    - higher values of *r* imply that a strong relationship between X and Y
    - lower values of *r* imply that a weak relationship between X and Y
- direction
    - positive (+): as X increases, Y also increases
    - negative (-): as X increases, Y decreases

Pearson's *r* (correlation)

$$r = \frac{degree\ to\ which\ two\ variables\ vary\ together\ (covary)}{degree\ to\ which\ two\ variables\ vary\ independently}$$

but we calculated the relationship between height (X) and weight (Y) as follows:

$$r = \frac{\sum z_x z_y}{N}$$

$$r = \frac{\sum z_x z_y}{N} = \frac{1}{N}\sum\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N\ (\sigma_x\sigma_y)} = \frac{\sum(X-\mu_x)(Y-\mu_y)/_N}{\sigma_x\sigma_y} = \frac{covariance}{independent\ variance}$$

Pearson's *r* (correlation)

- more generally, you don't need to standardize or z-score the two variables to find the correlation

$$\rho(population) = \frac{\sum(X-\mu_x)(Y-\mu_y)}{(N)\sigma_x\sigma_y} = \frac{\sum z_x z_y}{N} \quad \text{OR } r(sample) = \frac{\sum(X-M_x)(Y-M_y)}{(N-1)s_x s_y} = \frac{\sum z_x z_y}{N-1}$$

- alternative formulas
  - SS = sum of squared errors
  - SP = sum of product of deviation scores

$$SP = \sum XY - \frac{\sum X \sum Y}{N}$$

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

activity 1

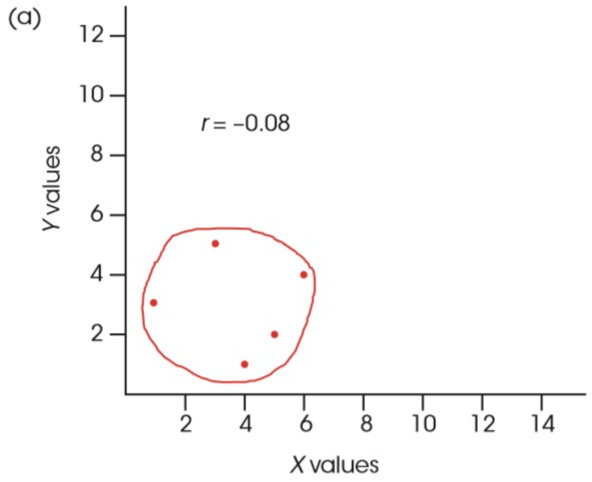
- science and history scores
- calculate the Pearson correlation

activity 2

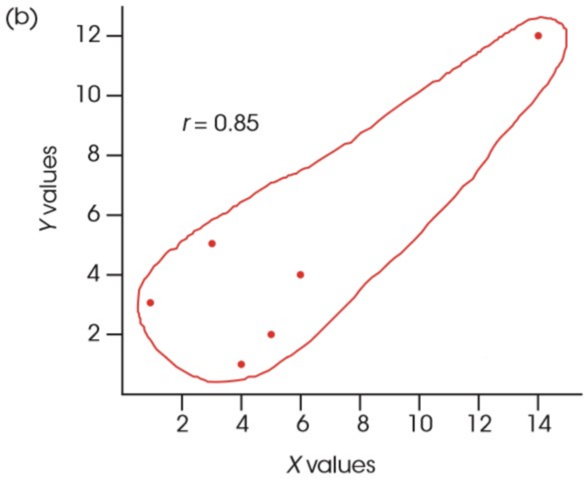
- try changing one of the history scores to an extreme value
- what happens to the correlation?

correlations and outliers

- outliers can have a dramatic effect on correlations
- always represent the problem graphically!



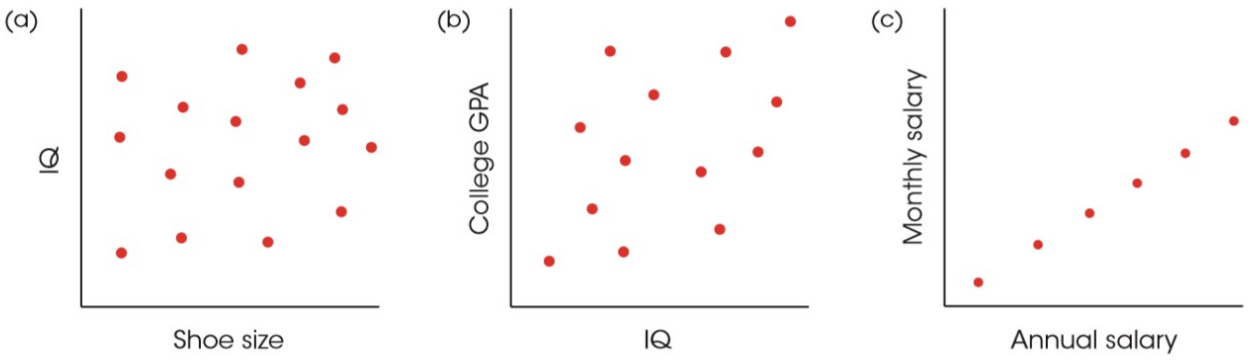
Original Data		
Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2

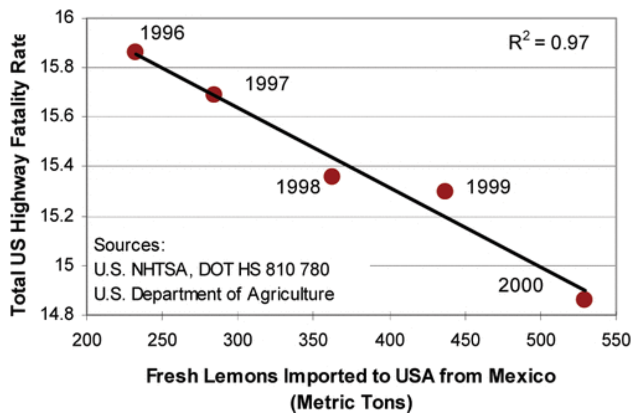


Data with Outlier Included		
Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2
F	14	12

correlation ≠ causation!

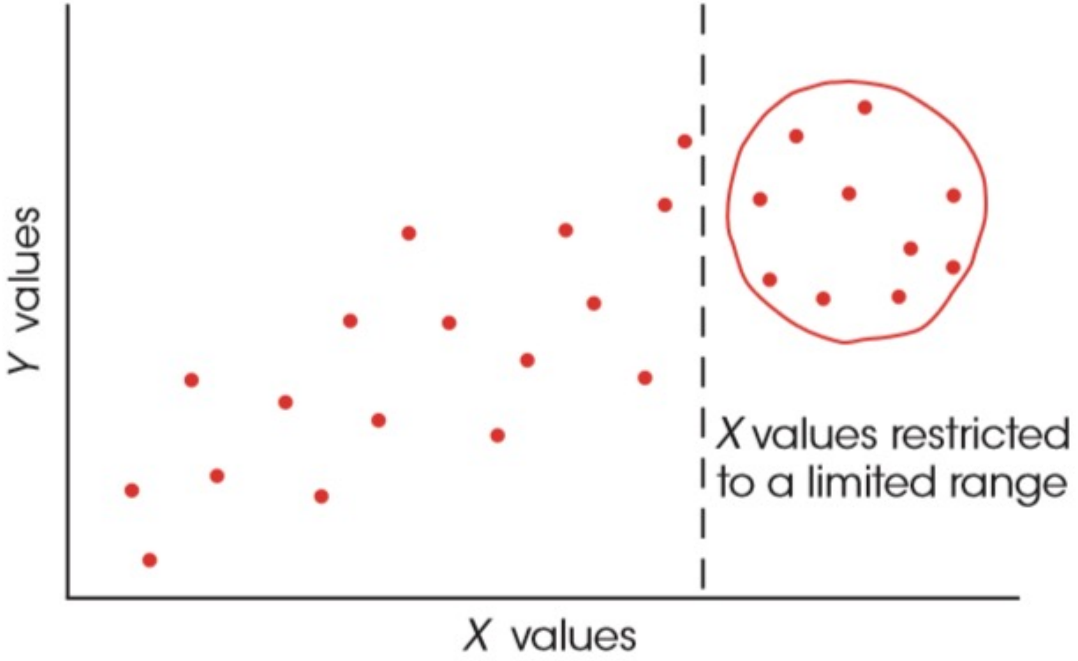
- for X to cause a change in Y:
  - X and Y must covary
  - X must <u>precede</u> Y
  - there should be no competing explanation or third variable





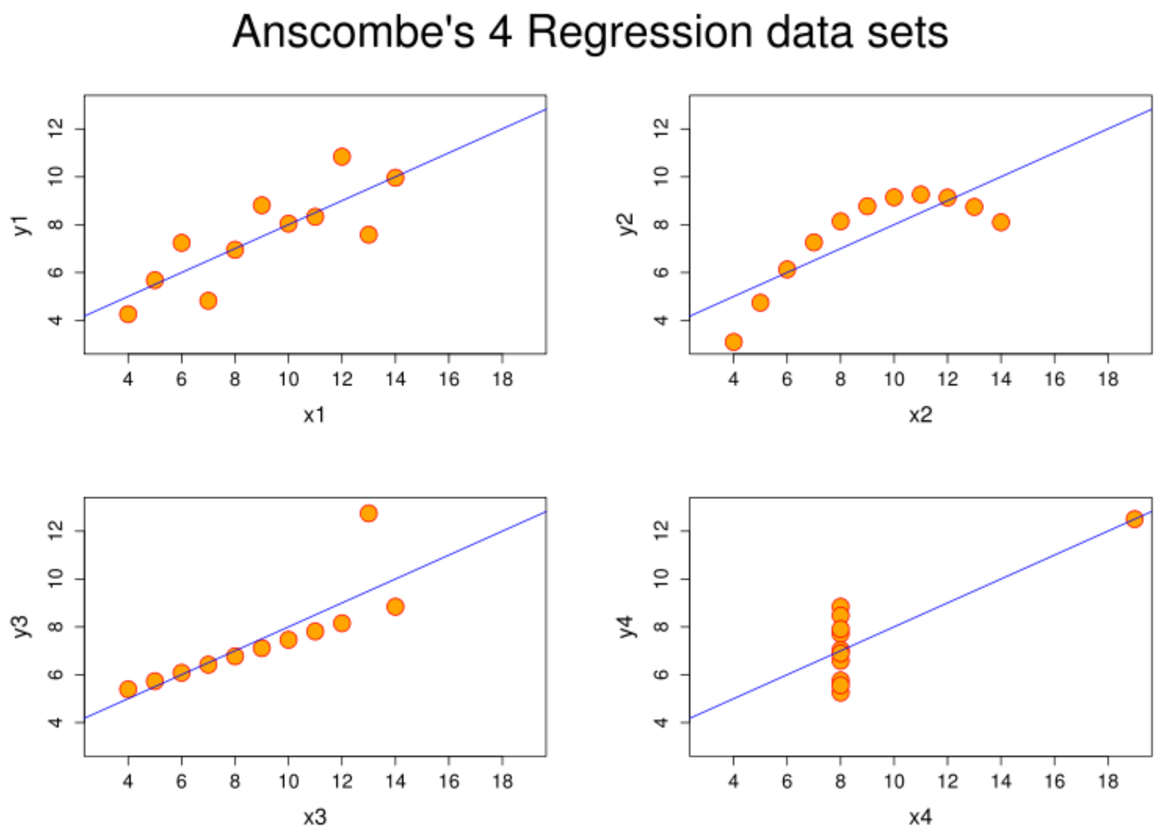
correlations and range restrictions

- correlations are greatly affected by the range of scores



Pearson's r and non-linearity

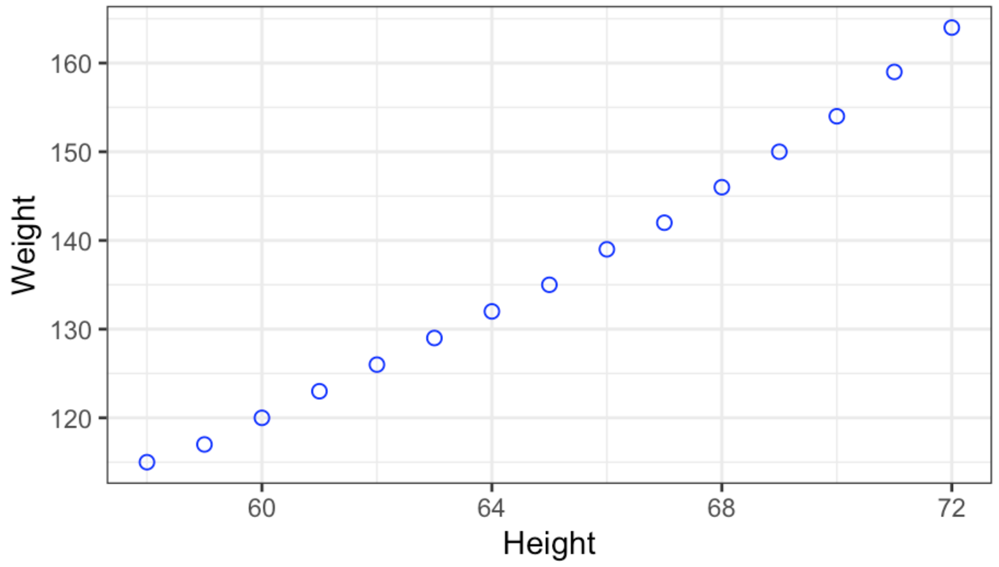
- Pearson's r measures the degree of *linear* relationship between two variables
- there can still be a consistent relationship, even if nonlinear but Pearson's *r* is not the appropriate model for these data
- more next time!



back to our example

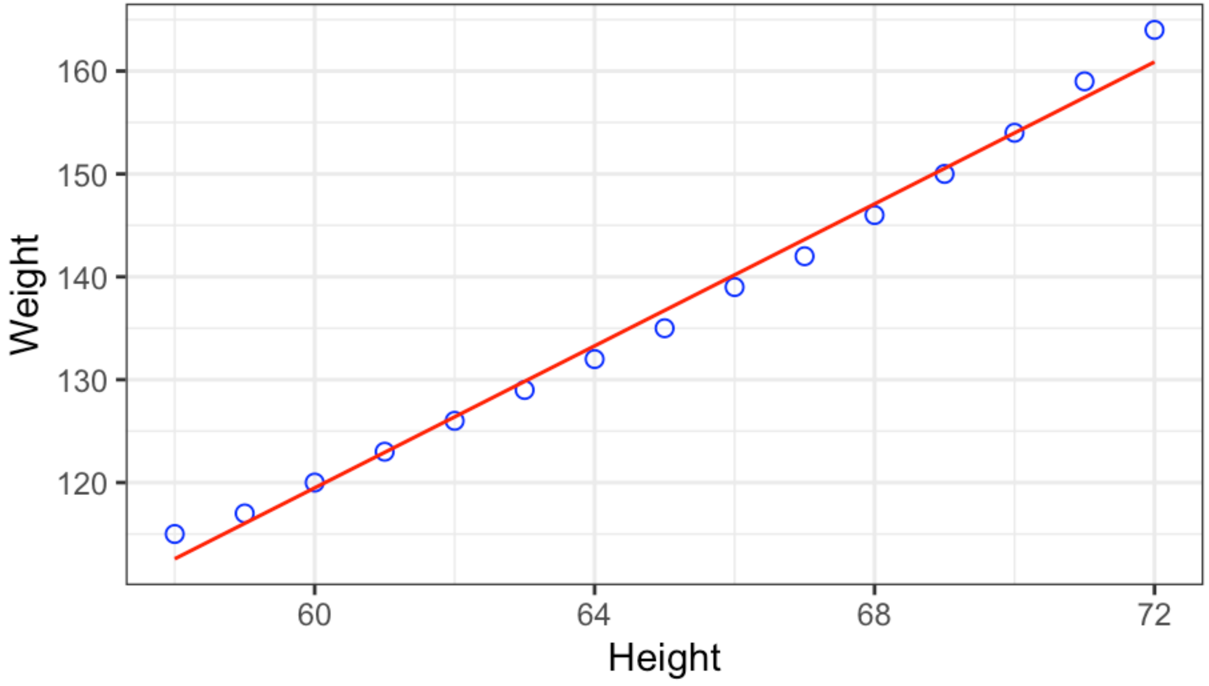
- we found that the *correlation* was $r \approx 0.9954$ for z-scored height and weight
- reviewing our modeling framework:
  - weight = b(height) + error
  - weight = 0.9954 (height) + error
  - a 1-unit increase in standardized height leads to a 0.9954-unit increase in standardized weight
- turns out, this is very close to the equation of a straight line!
  - Y = bX + a + error
  - Y? X? b? a?

Woman	z_height	z_weight	z_h*z_w	r
1	-1.62037037	-1.451485967	2.351676046	0.9954947681
2	-1.388888889	-1.317913639	1.830226406	
3	-1.157407407	-1.117555146	1.293318772	
4	-0.9259259259	-0.9171966539	0.8491590982	
5	-0.6944444444	-0.7168381616	0.497747384	
6	-0.462962963	-0.5164796692	0.2390836296	
7	-0.2314814815	-0.3161211768	0.07316783491	
8	0	-0.1157626845	0	
9	0.2314814815	0.151381972	0.03503811814	
10	0.462962963	0.3517404644	0.162824196	
11	0.6944444444	0.6188851209	0.4297322136	
12	0.9259259259	0.8860297774	0.8203041774	
13	1.157407407	1.153174434	1.334540088	
14	1.388888889	1.487105254	2.065187904	
15	1.62037037	1.821036075	2.950415653	



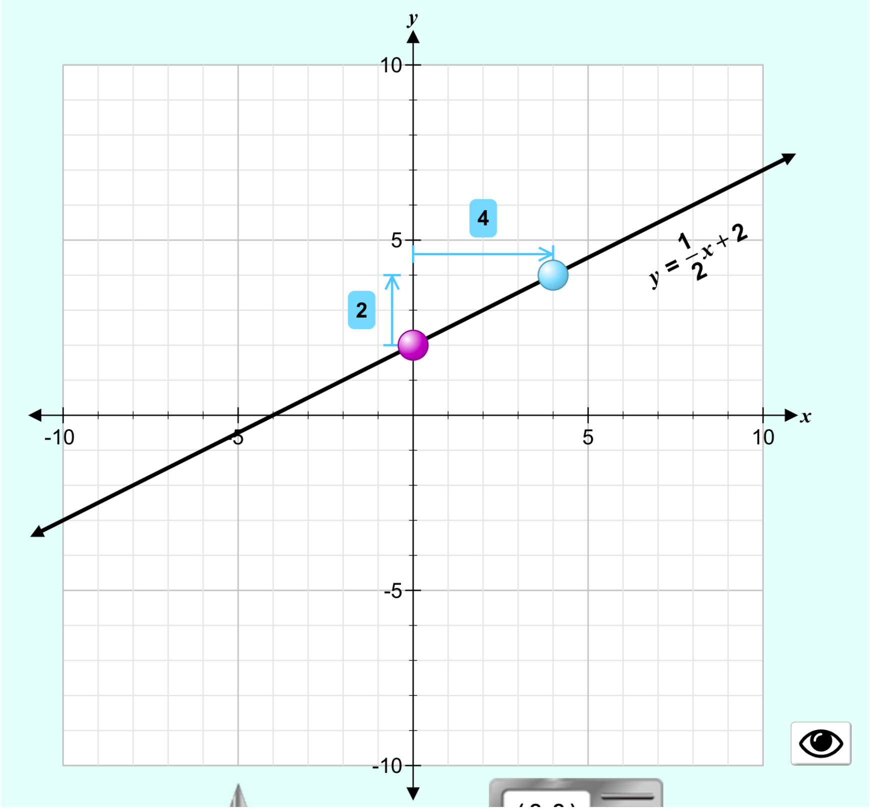
linear regression

- linear regression attempts to find the equation of a line that best fits the data, i.e., a line that could explain the variation in one variable using the other variable
- Y = bX + a + error
  - b: slope of the line
  - a: intercept
- extremely useful for prediction, i.e., given a score on X, we can predict a score on Y based on this line



activity: understanding lines

- Y = bX + a + error
- only two points are needed to define a line
- the slope (b) is the "rise" (y) over the "run" (x) for a given pair of points
- the intercept (a) is where the line cuts off the Y axis (i.e., when x = 0)
- example:
  - points = (0,2) and (4, 4)
  - b (slope) = $\frac{rise}{run} = \frac{4-2}{4-0} = \frac{2}{4} = \frac{1}{2}$
  - a (intercept) = 2
  - equation: Y = $\frac{1}{2}$X + 2



interactive demo

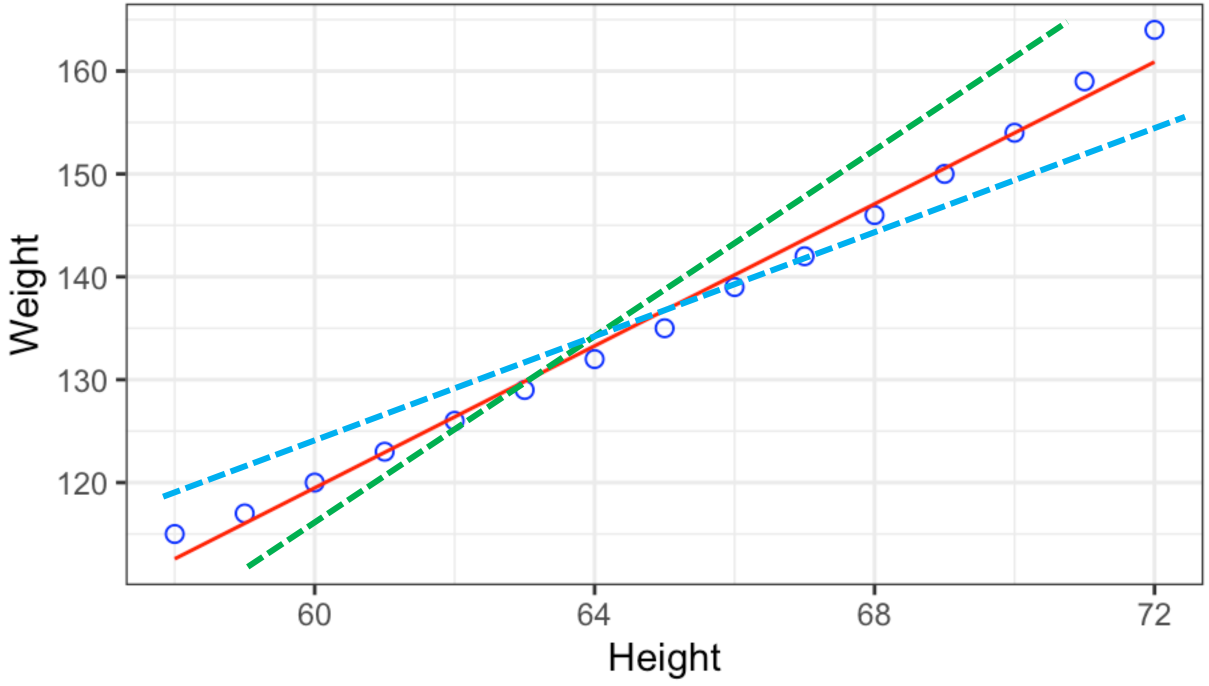
linear regression: finding a and b

- when fitting a line to multiple points, finding the value of the slope (b) is not straightforward, because several lines could potentially fit the full dataset
- how do we find the one that *best fits the data*?
- we could plug in ALL possible values of *b* and *a* and compute the error?

$$\text{error} = Y_i - (bX_i + a)$$

- find the combination of *b* and *a* that minimizes this error

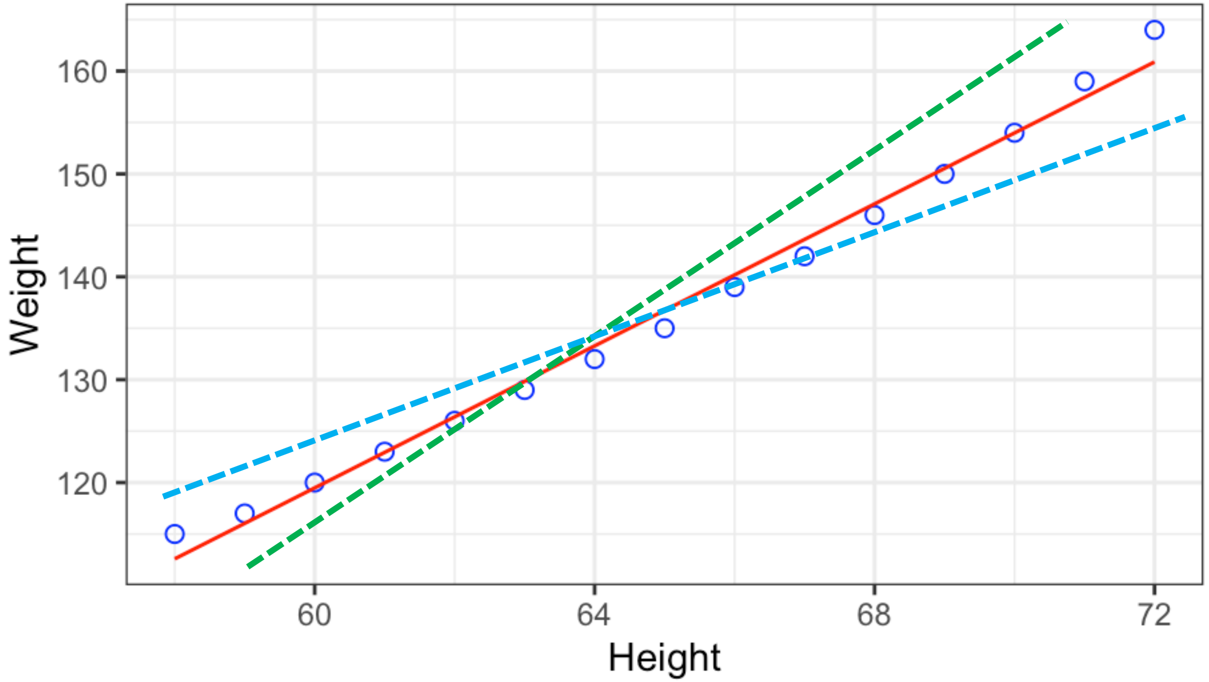
$$Y = bX + a + \text{error}$$



linear regression: finding a and b

- calculus provides a way to find the slope and intercept of the best-fitting line
- errors are first squared (to avoid canceling out!) and then summed, i.e., sum of squared errors (SS)
- $argmin(\sum_{i=1}^{n}(y_i - a - bx_i)^2)$
- partial derivatives are taken with respect to *a* and *b* (to find the minima) to yield
  - $a = M_y - bM_x$
  - $b = \frac{\sum(X-M_x)(Y-M_y)}{\sum(X-M_x)^2}$

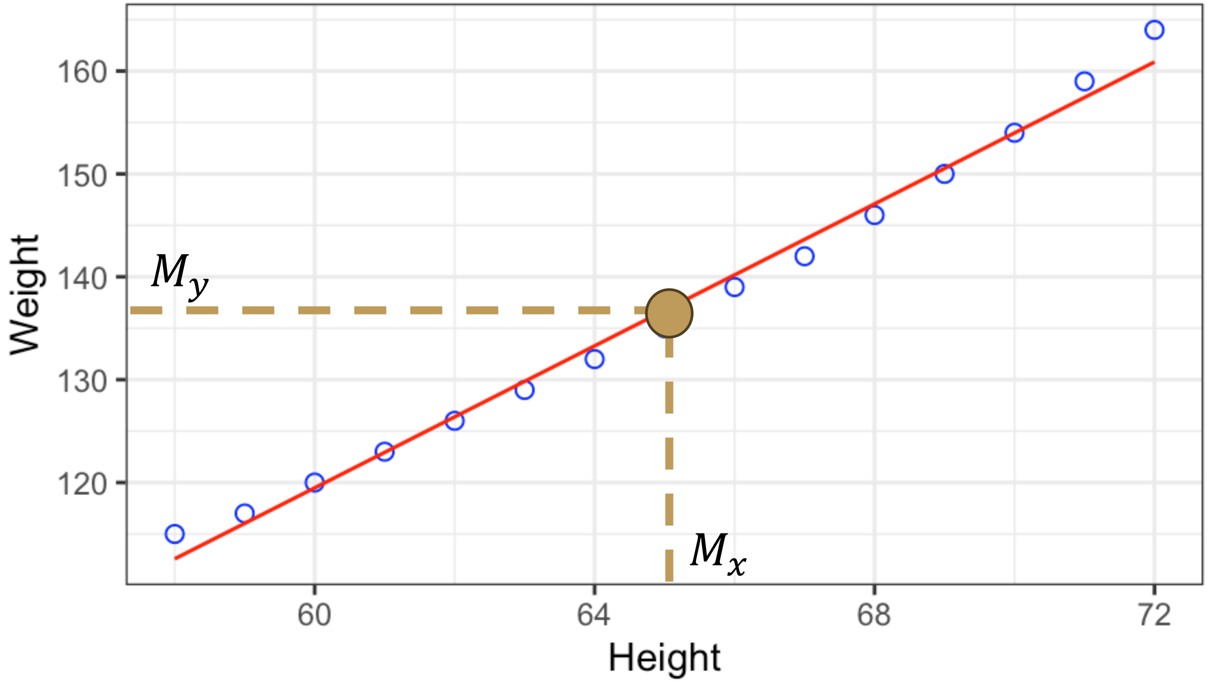
Y = bX + a + error



linear regression: finding a and b

- $a = M_y - bM_x$
- $b = \frac{\sum(X-M_x)(Y-M_y)}{\sum(X-M_x)^2}$
- rearranging the intercept equation:
  - $M_y = a + bM_x$
- the line of best fit passes through means of X and Y

Y = bX + a + error



linear regression and correlation

- but we already found the correlation between weight and height, $r \approx 0.9954$
- how are b and r related?

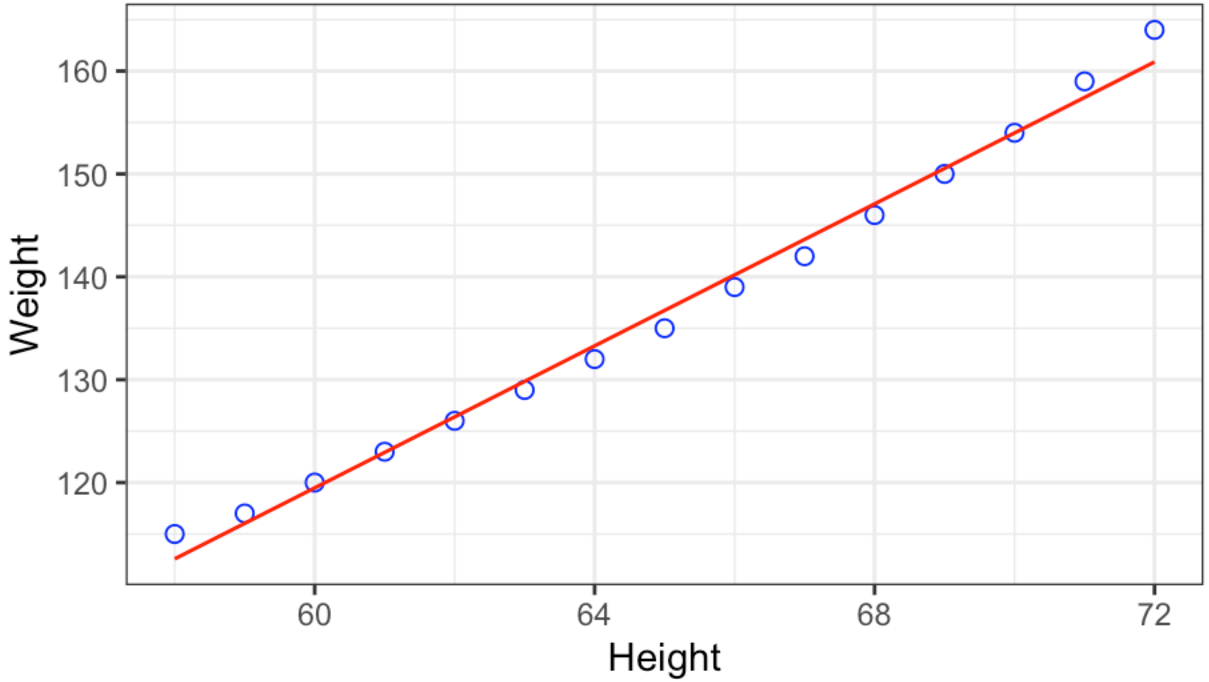
$$r = \frac{\sum(X - M_x)(Y - M_y)}{(N-1)s_x s_y}$$

$$b = \frac{\sum(X - M_x)(Y - M_y)}{\sum(X - M_x)^2} = \frac{\sum(X - M_x)(Y - M_y)}{(N-1){s_x}^2}$$

$$= \frac{r\, s_x s_y}{{s_x}^2} = r\,\frac{s_y}{s_x}$$

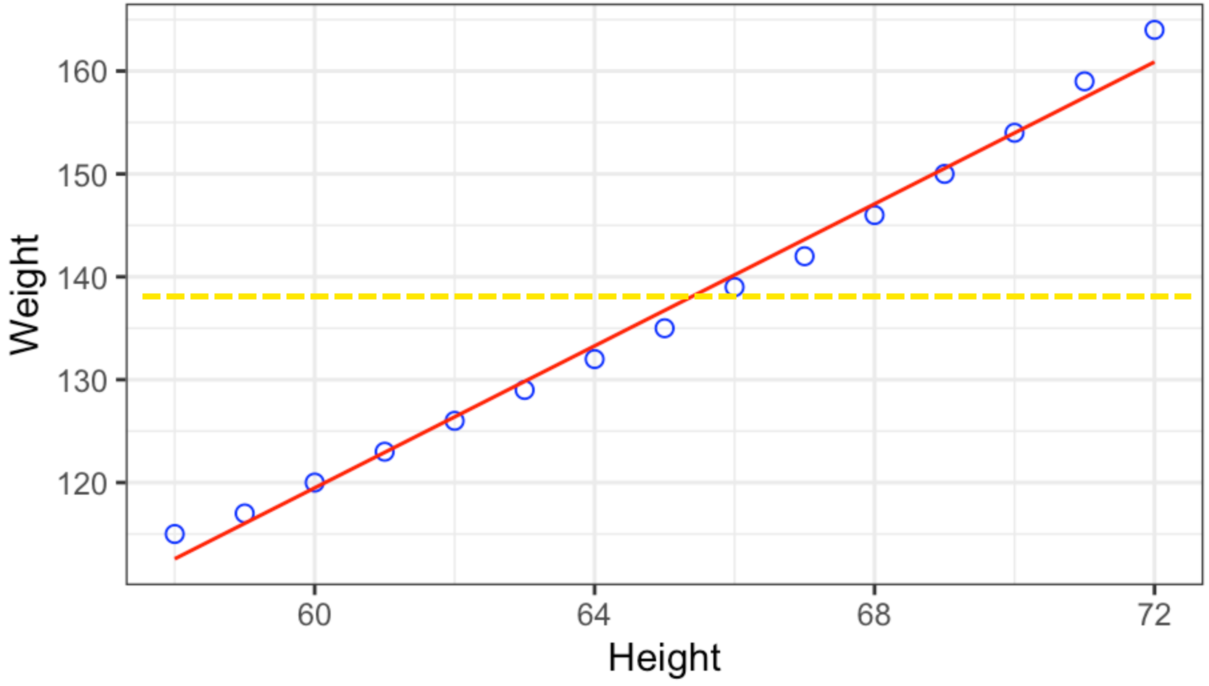
$$b = r\,\frac{s_y}{s_x}$$

Y = bX + a + error



special cases

- no relationship between X and Y
  - r = 0, b = 0
  - Y = bX + a = a = $M_y - bM_x$ = M_y
  - Y = mean value of Y for all values of X
- what is b when X and Y are standardized?
  - b = r when $s_x = s_y = 1$



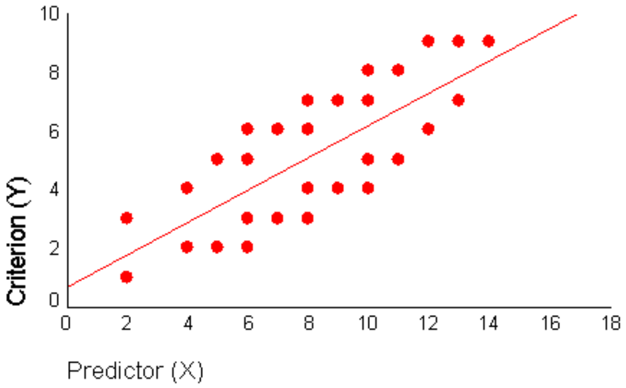
how good is the line of best fit?

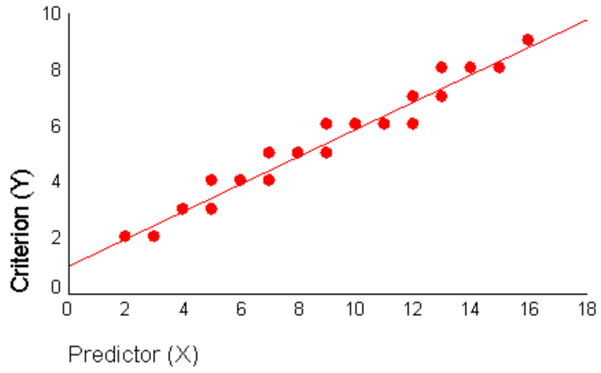
- even the line of “best” fit may ultimately not fit the data very well due to the inherent variability in the data
- how we assess model fit?
- data = model + error
- data = a + bX + error
- our favorite friend: sum of squared errors (SS)!

$$\hat{Y} = a + bX = \text{predictions}$$

$$SS_{error} = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum Y - \hat{Y}$$

- later: we compare SS_{model} to SS_{error} via the F-test!





standard error of estimate

- how far away is an average data point from the line of best fit?
- similar concept to standard deviation, $s = \sqrt{\frac{SS}{df}}$
- standard error of estimate = “average” SS

$$SE_{model} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{SS_{error}}{n-2}}$$

- why 2? number of estimated parameters (*a* and *b*)
- later in the course:
  - standard errors can be calculated for specific *a* and *b* associated with X by rescaling this error
  - how much better is this model than the mean model?

how good is a correlation?

$$r = \frac{degree\ to\ which\ two\ variables\ vary\ together\ (covary)}{degree\ to\ which\ two\ variables\ vary\ independently}$$

- coefficient of determination: r^2

$$r^2 + unexplained\ variance = 1$$

$$unexplained\ variance = SS_{error} = 1 - r^2$$

$$SE_r = s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

r^2 denotes the percentage of variance explained in Y due to X

putting it all together...

- Pearson's correlation (*r*) measures the linear relationship between two variables

$$\rho(population) = \frac{\sum(X-\mu_x)(Y-\mu_y)}{(N)\sigma_x\sigma_y} = \frac{\sum z_x z_y}{N} \quad \text{OR} \quad r(sample) = \frac{\sum(X-M_x)(Y-M_y)}{(N-1)s_x s_y} = \frac{\sum z_x z_y}{N-1}$$

- linear regression uses *r* to fit a straight line to the data

$$b = \frac{\sum(X-M_x)(Y-M_y)}{\sum(X-M_x)^2} = r\,\frac{s_x}{s_y}$$

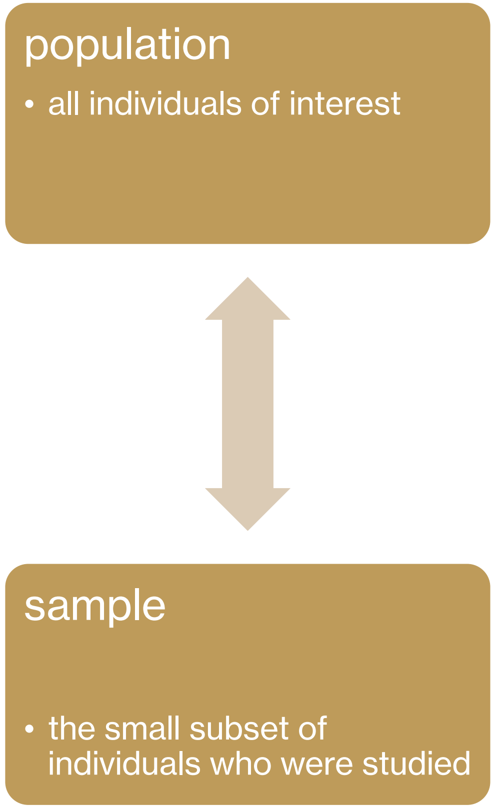
$$a = M_y - bM_x$$

conceptual differences

- technically, regression involves predicting a random variable (Y) using a fixed variable (X). In this situation, no sampling error is involved in X, and repeated replications will involve the same values for X (this allows for prediction)
  - example: X is an experimental manipulation
- correlation describes the situation in which both X and Y are random variables. In this case, the values for X and Y vary from one replication to another and thus sampling error is involved in both variables
  - example: X and Y both naturally vary

can we trust our models?

- our goal is to find the best model for our data and generalize to the population
- but how do we know that our sample is representative of the population? how do we know our models are good enough?
- after midterm 1!



next time

- **before** class
  - *work on*: PS 3 (Chapter 15/16 problems)
  - *watch*: Pearson correlation and Linear regression
- **during** class
  - more on correlation / regression!