DATA ANALYSIS

Week 4: Correlations and regression

lunch with Psychology faculty!

Lunch with Psychology Faculty

The Psychology Department is hosting lunches with faculty and students this semester.

All lunches will be in **Thorne Dining**! Please meet us at the check-in station at the times mentioned for the specific dates.

The lunches are on the following dates/times:

- Wednesday, February 21 2024 (**12 pm**): Prof. Erika Nyhus and Prof. Hannah Reese
- Tuesday, March 5 2024 (**12 pm**): Prof. Kacie Armstrong, Prof. Suzanne Lovett, and Prof. Thomas Small
- Friday, April 12 2024 (**1.10 pm**): Prof. Abhilasha Kumar and Prof. Samuel Putnam

We look forward to seeing you!

https://forms.gle/BGDXsRWFB6xNuoyE7

logistics: class survey (February)

- https://forms.gle/hw6kQzznP73Rrifh6
- link also on Canvas (under class surveys)
- due Feb 21 (**Wed morning**, so we can talk about it in class on Wed)
- 1 extra credit point that counts towards your final points/grade
  - submit on Canvas (it's an "assignment" on Canvas)
- I value your feedback
- anonymous survey! please be honest and reflective
- you will get a code at the end of the survey (on the thank you screen)
  - copy-paste this code on Canvas to get credit

what’s coming up

4	F: February 16, 2024	W4 continued…
5	M: February 19, 2024	**Problem Set 3 due**
5	W: February 21, 2024	W5: Loose Ends / Exam 1 review
5	F: February 23, 2024	**Exam (Midterm) 1**
6	W: February 28, 2024	W6: Probability & Sampling
6	F: March 1, 2024	sampling
7	M: March 4, 2024	**Problem Set Opt-out Deadline 2**
7	W: March 6, 2024	W7: Hypothesis Testing
7	F: March 8, 2024	W7 continued…
8	M: March 11, 2024	**Problem Set 4 due**
8	W: March 13, 2024	**Spring Break!**
8	F: March 15, 2024	**Spring Break!**
9	W: March 20, 2024	**Spring Break!**
9	F: March 22, 2024	**Spring Break!**

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

logistics: review for midterm 1

- practice midterm is available on Canvas (Modules > Midterm 1)
- conceptual portion (40% of total midterm)
  - 40 multiple-choice/true-false questions
  - try to practice in a timed/closed-book manner
- computational portion (60% of total midterm)
  - short answer questions
  - sheets-based questions
  - answers will be posted on Tuesday
  - actual exam: you will submit a **downloaded** PDF + **downloaded** Sheets file on Canvas

some bonus content

- guesssing correlations and tracking your performance!
- why is a correlation restricted to -1 and 1?

today's agenda

- more on correlations
- assessing model fit

recap: correlation and regression

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

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

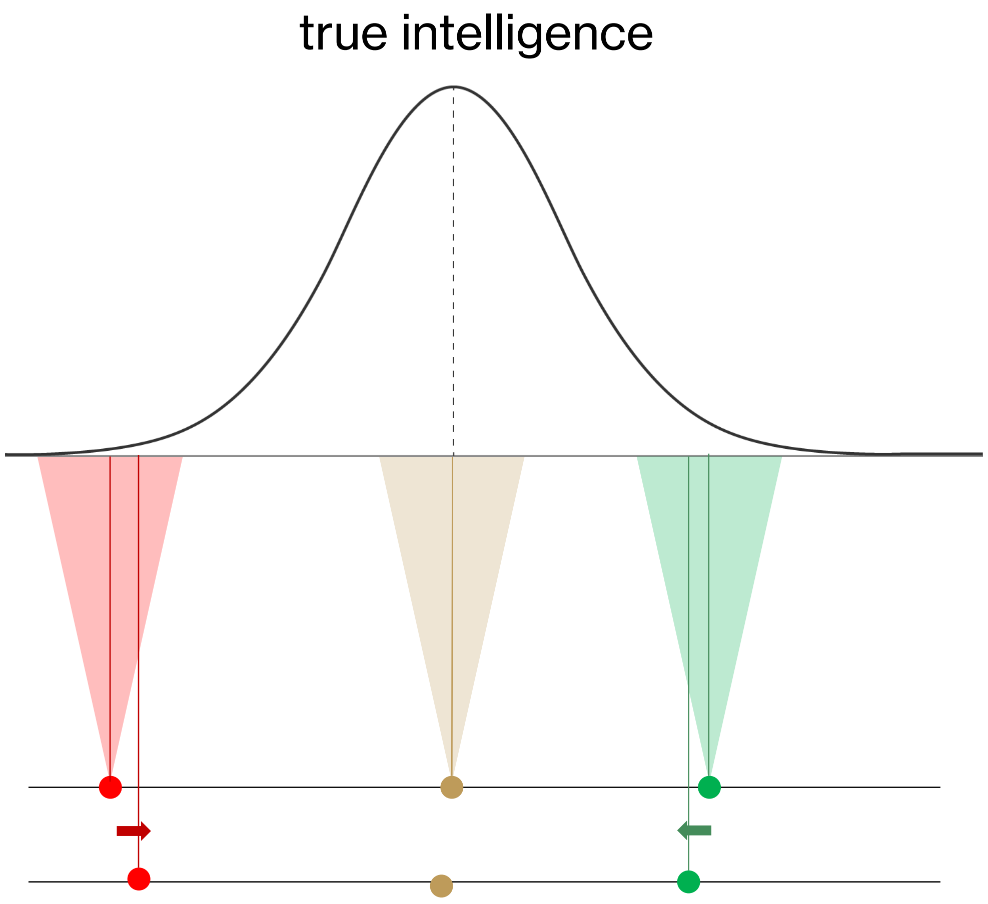
- linear regression uses r to fit a straight line to the data

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

$$a = M_y - bM_x$$

regression toward the mean

- if two variables are imperfectly correlated, extreme scores on one variable are associated with less extreme scores on the other variable, on average
- consider two measurements of intelligence, one before and one after a treatment
    - data = model + error
- the first measurement likely has some error with respect to the true value, due to several factors
- the second measurement will try to again estimate the true value
- since values closer to the mean are more likely, the second measurement is likely to be closer to the mean than the first extreme value



https://www.youtube.com/watch?v=2cB3m5aS3mw

regression toward the mean

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

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

$$a = M_y - bM_x$$

$$\hat{Y} = M_y - bM_x + b\,X = M_y + b(X - M_x)$$

$$\hat{Y} - M_y = b(X - M_x)$$

$$\hat{Y} - M_y = r\,\frac{s_y}{s_x}(X - M_x)$$

$$\frac{\hat{Y} - M_y}{s_y} = r\,\frac{(X - M_x)}{s_x}$$

$$\widehat{z_y} = r\,z_x$$

If $r \neq \mp 1$, $\widehat{z_y}$ (predicted value of Y) is less [extreme] than the value of X(z_x)

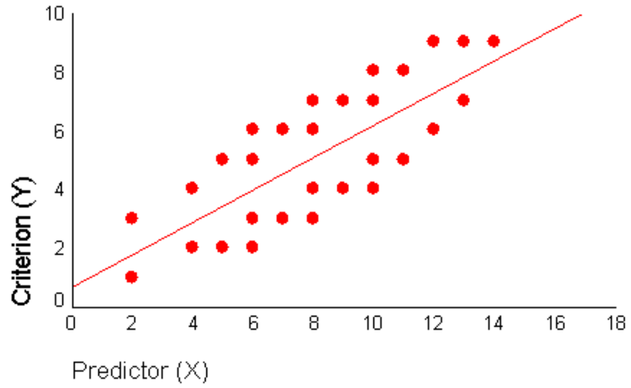
<u>Bonus</u>: If you know the z-score of X and the correlation, you can find the <u>predicted</u> z-score for Y!

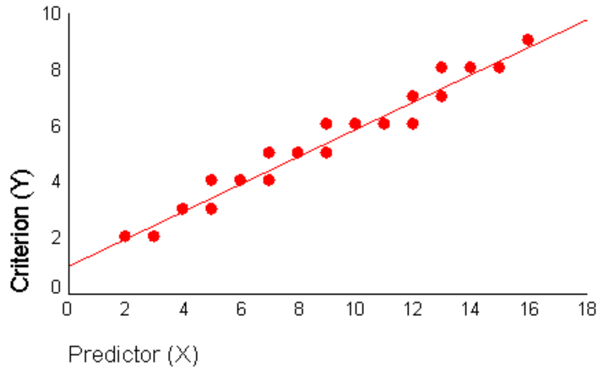
how good is the line of best fit?

- 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)!

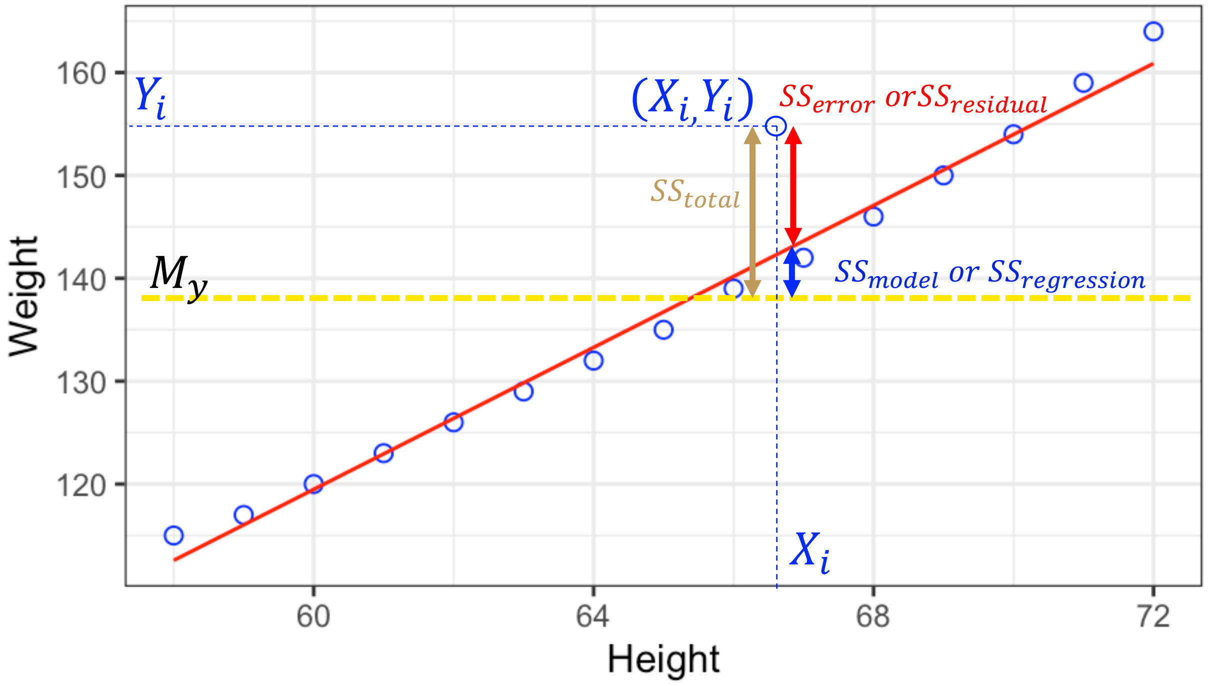
$$\hat{Y} = a + bX = \text{predictions}$$

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





understanding goodness/errors



$$SS_{total} = SS_{model} + SS_{error}$$

$$SS_{total} = \sum (Y - M_y)^2$$

$$SS_{error} = \sum (Y - \hat{Y})^2$$

$$SS_{model} = \sum (\hat{Y} - M_y)^2$$

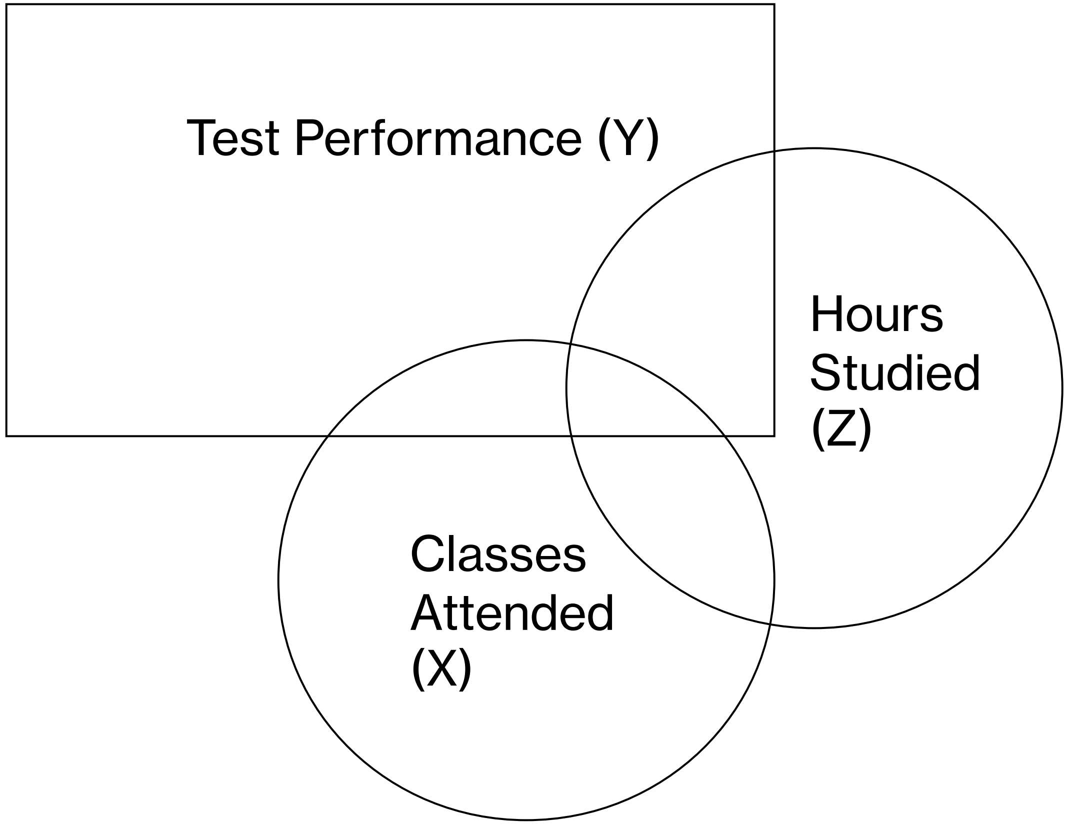
coefficient of determination (R^2)

- what proportion of the <u>total</u> error variance is explained by my model?
- $R^2 = \frac{SS_{model}}{SS_{total}} = r^2$ in the case of simple linear regression (i.e., Y = a + bX) (proof)
- R^2 denotes the **percentage of variance** explained in Y due to X
- when multiple variables are involved, R^2 reflects the variance explained by the full model

other variables in the mix

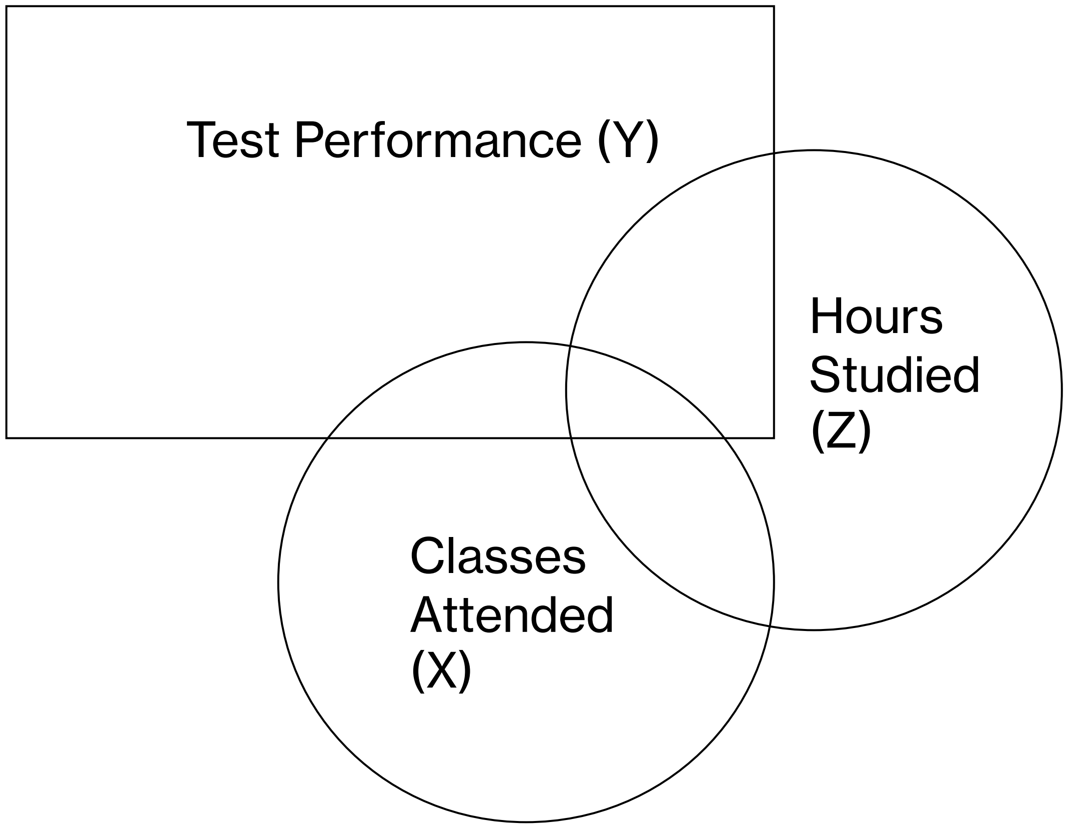
- sometimes, more than one variable (X and Z) may impact the key variable of interest (Y)
- in such cases, it is difficult to isolate the impact of one variable (X) on another (Y), without taking into account the variance shared by the variables (X and Z)
  - three relationships r_{xy}, r_{xz}, r_{yz}
- *partial* correlation of X and Y

$$r_{XY.Z} = \frac{r_{XY} - (r_{XZ} r_{YZ})}{\sqrt{(1 - r_{XZ}{}^2)(1 - r_{YZ}{}^2)}}$$



multiple regression

- multiple linear regression refers to finding a model that best predicts a variable of interest (Y) using more than one variable (X_1, X_2, etc.)
- data = model + error
    - *linear*: $Y = bX + a + \text{error}$
    - *multiple*: $Y = b_1 X_1 + b_2 X_2 + a + \text{error}$
- for two variables, we are fitting a *plane* to the data instead of a line
- more to come! we will discuss a family of models within the framework of “general linear models”



standard error of estimate / r

- 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 (regression model) = "average" SS_{error}

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

- standard error for correlation

$$r^2 = \text{ explained variance}$$

$$\text{unexplained variance} = 1 - \text{explained variance} = 1 - r^2$$

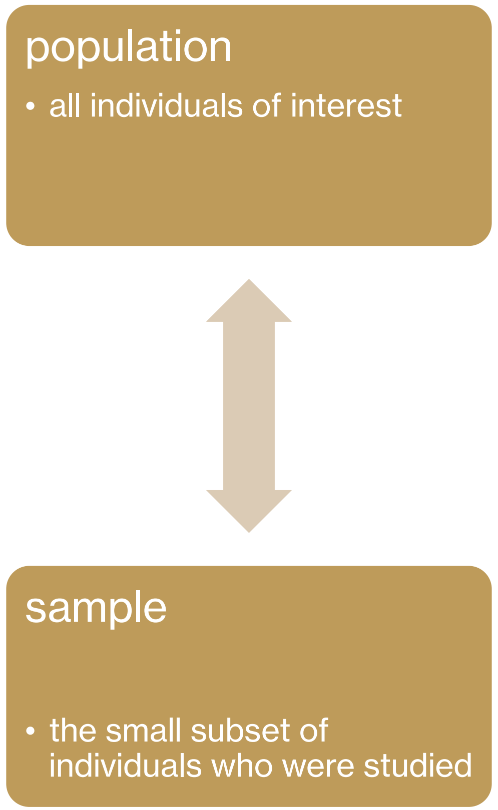
$$SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}}$$

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

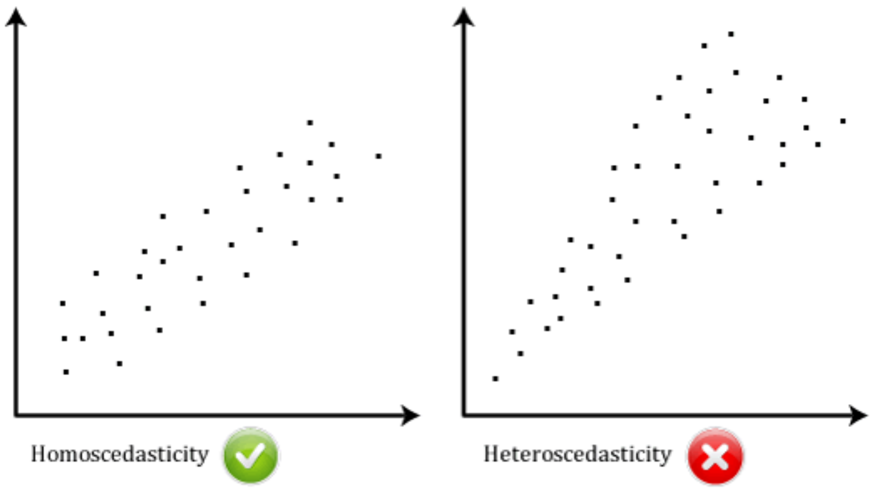
can we trust our models?

- 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!



Pearson's r assumptions

- interval/ratio scale: variables should be on interval / ratio scale: if the distance between the values is not equal, estimates of variability are difficult
- homoskedasticity: dispersion of Y remains relatively similar across the range of X
- no significant outliers
- variables should be approximately normally distributed

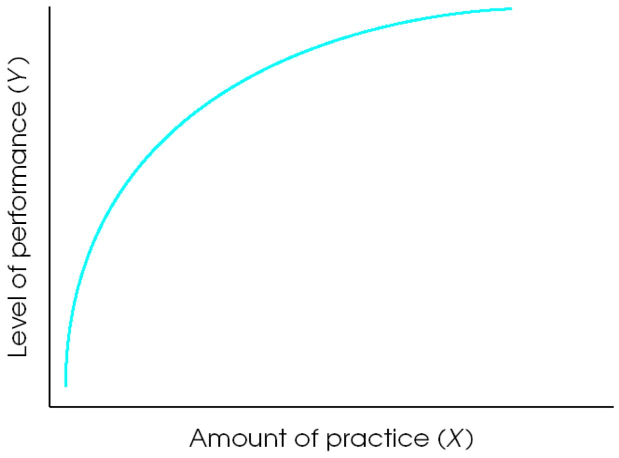


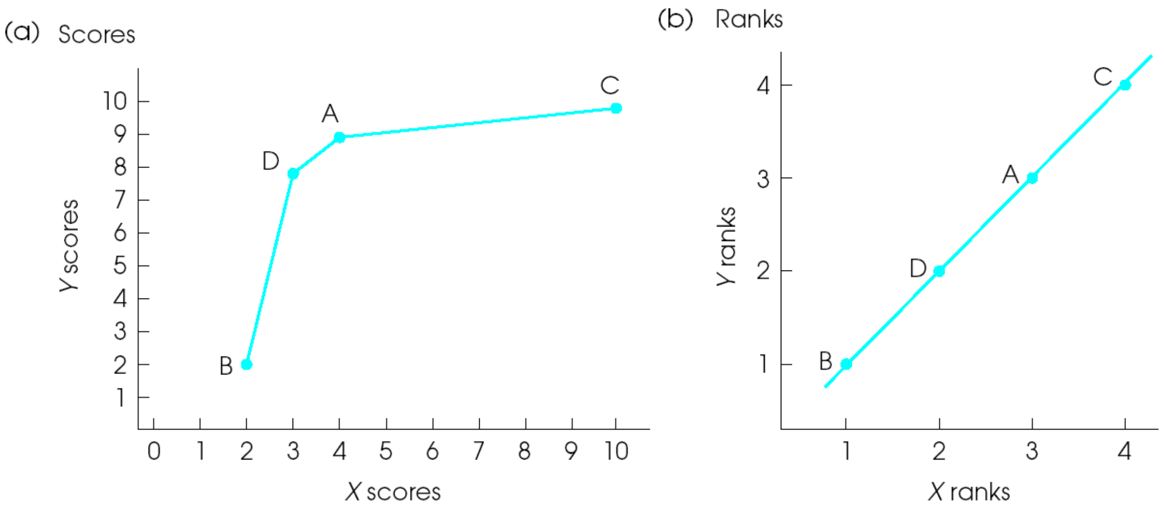
alternatives to Pearson's r

- when data are not interval/ratio, Pearson's r is not appropriate
- other alternatives exist
  - both variables ordinal: spearman's *rho*
  - one variable dichotomous (binomial): point biserial
  - both variables dichotomous: phi
- all alternatives are simply variations/extensions of Pearson's r
- remember, data = model + error
- when the data changes, the model also changes

spearman's *rho*

- typically used for ordinal scales, non-linear relationships, or when outliers may need to be included
- uses ranks / ordering of scores instead of the raw scores themselves
- Pearson's r may underestimate the relationship but ranks may reveal a strong relationship
- if r is higher than rho, that typically means there is more of a linear trend in the data



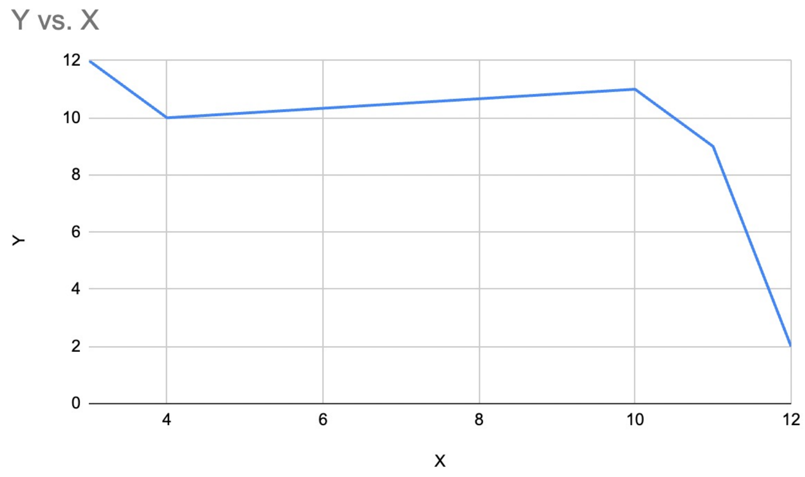


example

- a set of scores
- we first calculate Pearson's *r*

  =CORREL(X,Y)
- then we compute ranks
  - lowest numbers get lower ranks
- compute the pearson's *r* for ranks!

  =CORREL(rank_x, rank_y)



Person	X	Y
A	3	12
B	4	10
C	10	11
D	11	9
E	12	2

rank_x	rank_y
1	5
2	3
3	4
4	2
5	1

pearson
-0.6485442507

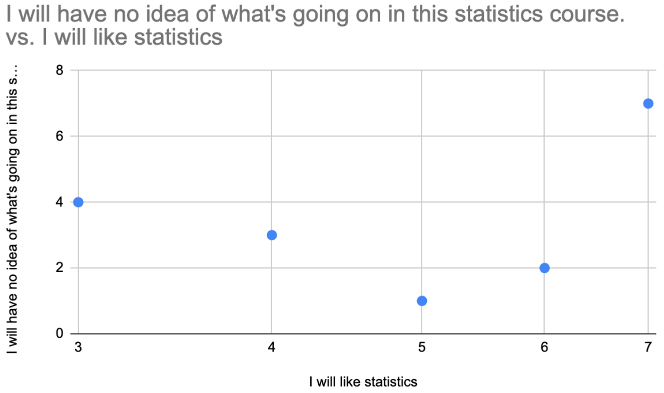
spearman
-0.9

activity: calculate spearman’s rho

- calculate the correlation between two items from the statistics survey from class
- sheet (fake data)

Student	I will like statistics	I will have no idea of what's going on in this statistics course.
1	6	2
2	5	1
3	3	4
4	7	7
5	4	3

activity: calculate spearman's rho



Student	I will like statistics	I will have no idea of what's going on in this statistics course.	rank_like	rank_idea	rho	r
1	6	2	4	2	0.1	0.3434014099
2	5	1	3	1		
3	3	4	1	4		
4	7	7	5	5		
5	4	3	2	3		

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different

score
7
8
2
7
4
2
4

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different

score	initial_ranks
7	6
8	7
2	2
7	5
4	4
2	1
4	3

spearman's *rho*: handling ties

- when two or more scores are the same, their ranks are the average of the ranks they would have gotten if the scores were different

score	initial_ranks	final_ranks
7	6	5.5
8	7	7
2	$\boxed{2}$	1.5
7	5	5.5
4	4	3.5
2	$\boxed{1}$	1.5
4	3	3.5

spearman's *rho*: other formula

$$r = \frac{\sum(X - \mu_x)(Y - \mu_y)}{(N)\sigma_x\sigma_y}$$

- given that ranks do away with the original scores, this formula can be simplified when there are no ties

$$r_s = 1 - \frac{6\sum D^2}{n\,(n^2 - 1)}$$

where D is difference between X and Y ranks for each data point

- proof

X	Y	rank_x	rank_y	D	D^2
3	12	1	5	-4	16
4	10	2	3	-1	1
10	11	3	4	-1	1
11	9	4	2	2	4
12	2	5	1	4	16

spearman's *rho*: other formula

- what is D if the ranks of X and Y are in the same order?
- what is r?

$$r_s = 1 - \frac{6\sum D^2}{n\,(n^2 - 1)}$$

X	Y	rank_x	rank_y	D	D^2
3	12	1	5	-4	16
4	10	2	3	-1	1
10	11	3	4	-1	1
11	9	4	2	2	4
12	2	5	1	4	16

point biserial and phi

- similar idea as Pearson's r but now our variables are not interval/ratio
- just converting the dichotomous variable to 0/1 numeric representations
  - point biserial : one variable dichotomous
  - phi : both variables dichotomous
- convert to numeric representations
- proceed as before

puzzle score	group
11	0
9	0
4	0
5	0
6	0
7	0
12	0
10	0
7	1
13	1
14	1
16	1
9	1
11	1
15	1
11	1
meanX	meanY
10	0.5

point biserial and phi

- similar idea as Pearson's r but now our variables are not interval/ratio
- just converting the dichotomous variable to 0/1 numeric representations
  - point biserial : one variable dichotomous
  - phi : both variables dichotomous
- convert to numeric representations
- proceed as before

puzzle score	group	sqx	sqy	z_x	z_y	z_x*z_y
11	0	1	0.25	0.2901905	-1	-0.2901905
9	0	1	0.25	-0.2901905	-1	0.2901905
4	0	36	0.25	-1.741143	-1	1.741143
5	0	25	0.25	-1.4509525	-1	1.4509525
6	0	16	0.25	-1.160762	-1	1.160762
7	0	9	0.25	-0.8705715001	-1	0.8705715001
12	0	4	0.25	0.5803810001	-1	-0.5803810001
10	0	0	0.25	0	-1	0
7	1	9	0.25	-0.8705715001	1	-0.8705715001
13	1	9	0.25	0.8705715001	1	0.8705715001
14	1	16	0.25	1.160762	1	1.160762
16	1	36	0.25	1.741143	1	1.741143
9	1	1	0.25	-0.2901905	1	-0.2901905
11	1	1	0.25	0.2901905	1	0.2901905
15	1	25	0.25	1.4509525	1	1.4509525
11	1	1	0.25	0.2901905	1	0.2901905
meanX	meanY	SSx	SSy			r
10	0.5	190	4			0.5803810001
		sd_x	sd_y			
		3.446012188	0.5			

next time

- **before** class
  - *complete*: Week 4 quiz
  - *submit*: PS3
  - *fill out*: class survey (February)
  - *practice*: midterm 1 review questions
- **during** class
  - reviewing concepts + preparing for midterm 1!