DATA ANALYSIS

Week 5: More correlation and regression

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: midterm 1

- Practice assessments [5]: Before each exam, practice exams will be made available to you to help with your preparation. Submitting these practice exams and getting at least 50% on them will count towards class participation. Practice exams for midterms are worth 1.5 point each and the practice exam for the final will be worth 2 points.

- lingering question: I noticed that the bold section that is designed to link to the solution template for the practice midterm did not work for me. I don't know if this is only a problem with for me, but just wanted to say something in case other people were having the same issue. Thanks!
- answers to practice midterm 1 (conceptual + computational)
  - will be made available on **Tue noon** (before our review class)
- submissions count towards class participation credit (1.5 points)
  - need to come in before then (Tue noon)

today's agenda

- assessing model fit
- assumptions + more correlations

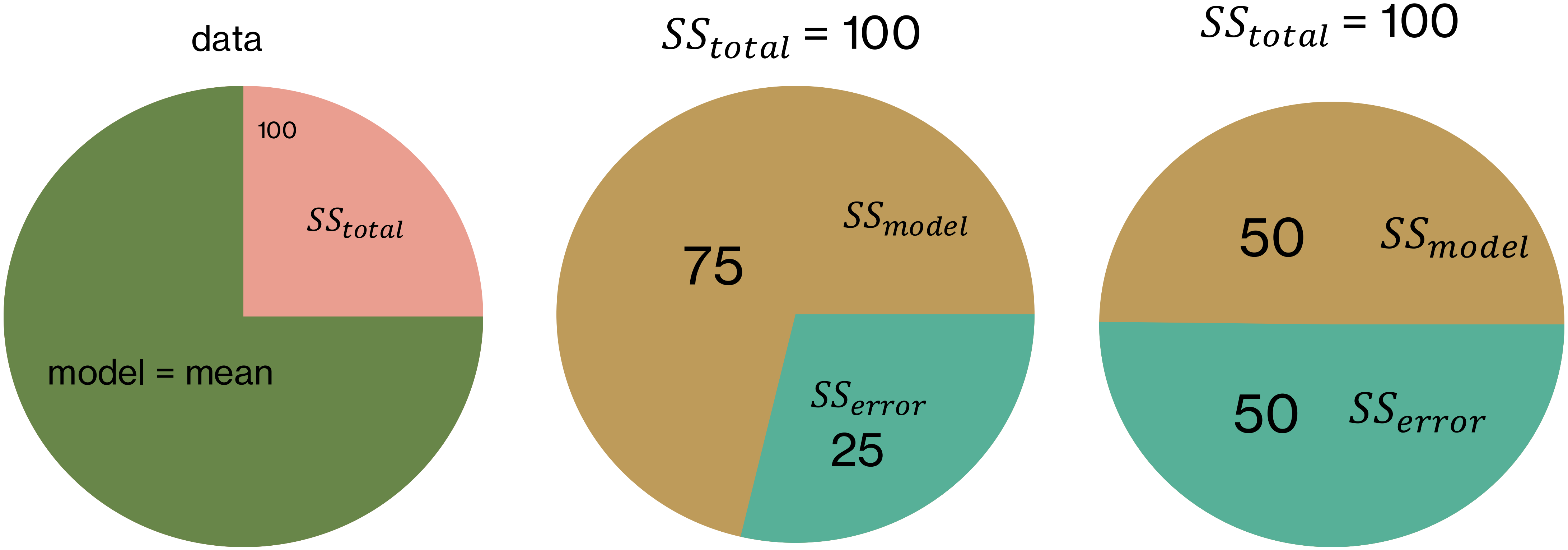
lingering question

- can you re-explain how we assessed the model fit relative to the mean? I found the conceptual part quite tricky. Thank you!

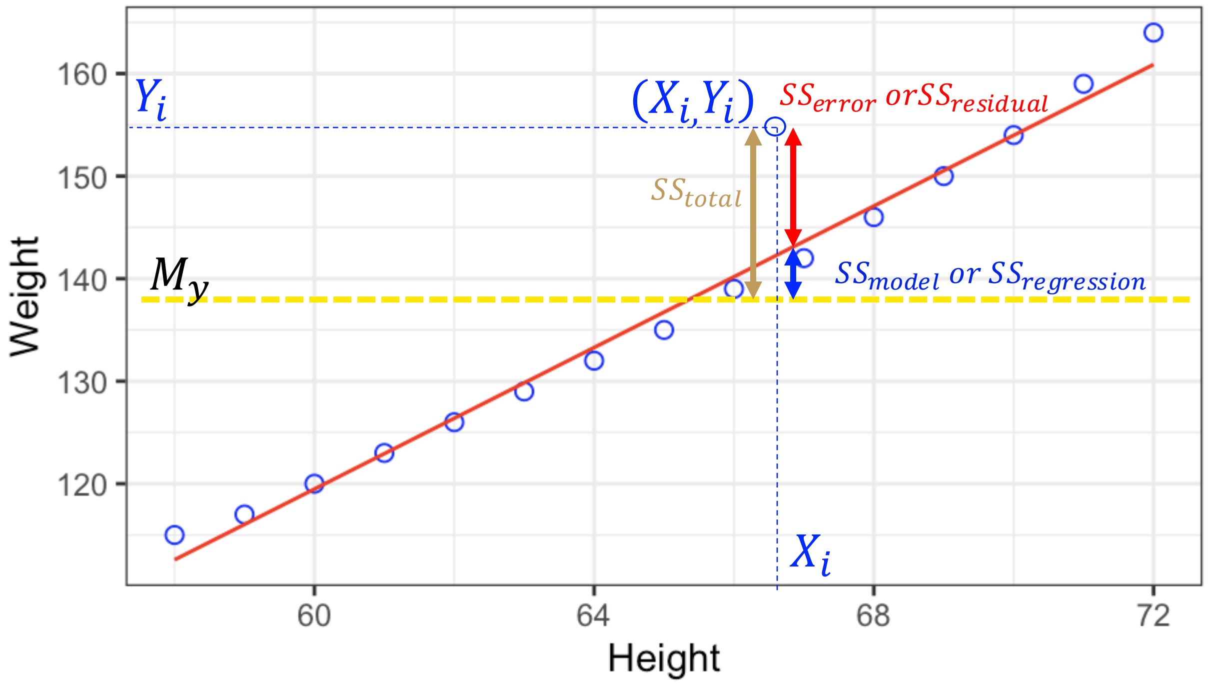
understanding model fit

- goal: explain variation in a variable Y (e.g., weight)
- our first approach is to "summarize" weights using the mean of all weights = M_y
  - the mean is our first, naive model, i.e., our "prediction" for each weight is simply the mean
  - $\widehat{Y_{mean}} = predicted\ weight\ based\ on\ mean = M_y$
  - the mean will not perfectly fit each point and will generate some error: $SS_{total} = \sum(Y - M_y)^2$
- our second approach is to reduce the error generated by the mean (SS_{total})
  - we build a more complex model, e.g., use height (X) to explain weight (Y)
  - $\widehat{Y_{line}} = predicted\ weightbased\ on\ height = a + bX$
  - the line will also generate some error for each data point, $SS_{error} = \sum(Y - \hat{Y})^2$
- we will then examine the improvement in our predictions by using a better model ($a + bX$) vs. the mean (M_y)
  - $SS_{model} = \sum(\hat{Y} - M_y)^2$
- we want SS_{model} to be high and SS_{error} to be low: $SS_{total} = SS_{model} + SS_{error}$

understanding model fit



understanding model fit



SS_{total} denotes the total error left over after the mean has been fit to Y

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

SS_{error} denotes the error left over after the line $\hat{Y} = a + bX$ has been fit

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

SS_{model} denotes the difference, i.e., the error that our line is able to explain vs. what was left over from the mean!

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

model fit is assessed relative to the mean, i.e., how much better did we do compared to the mean model?

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

understanding model fit

women weight ~ women height + error

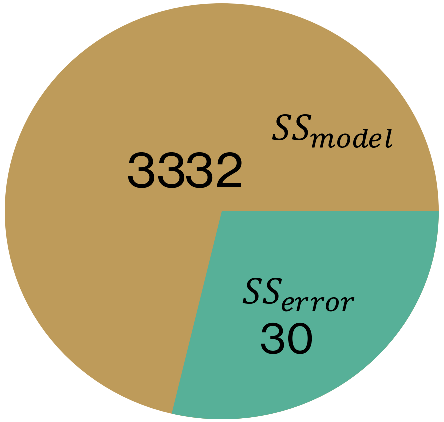
r	r^2
0.9954947678	0.9910098327
b	a
3.45	-87.51666667

SStotal
3362.933333

SSerror
30.23333333

SSmodel	SStotal-SSerror
3332.7	3332.7

SS_{total} = 3363



SS_{total} denotes the total error left over after the mean has been fit to Y

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

SS_{error} denotes the error left over after the line $\hat{Y} = a + bX$ has been fit

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

SS_{model} denotes the difference, i.e., the error that our line is able to explain vs. what was left over from the mean!

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

model fit is assessed relative to the mean, i.e., how much better did we do compared to the mean model?

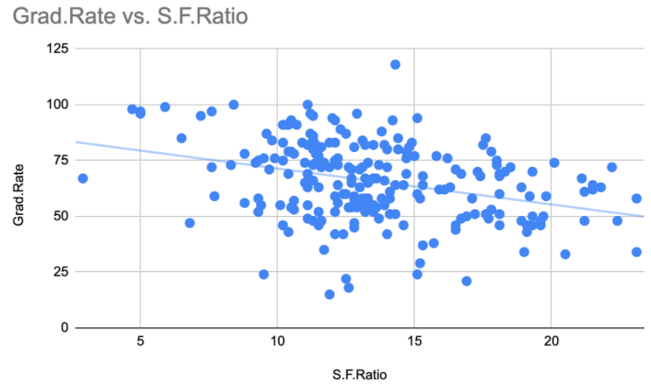
$$SS_{total} = SS_{model} + SS_{error}$$

W5 Activity 4a

- to what extent can student to faculty ratio explain graduation rates across colleges?
- Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.
- use the **data5** sheet to answer this question

College	S.F.Ratio	Grad.Rate
Abilene Christian University	18.1	60
Adelphi University	12.2	56
Adrian College	12.9	54
Agnes Scott College	7.7	59
Alaska Pacific University	11.9	15
Albertson College	9.4	55
Albertus Magnus College	11.5	63
Albion College	13.7	73
Albright College	11.3	80
Alderson-Broaddus College	11.5	52
Alfred University	11.3	73
Allegheny College	9.9	76
Allentown Coll. of St. Francis de Sales	13.3	74
Alma College	15.3	68
Alverno College	11.1	55
American International College	14.7	69
Amherst College	8.4	100
Anderson University	12.1	59
Andrews University	11.5	46

W5 Activity 4a debrief

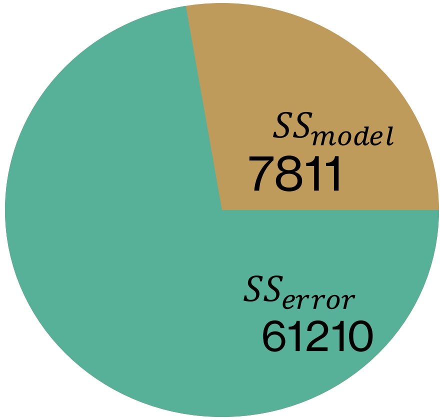


grad rate ~ student-faculty-ratio + error

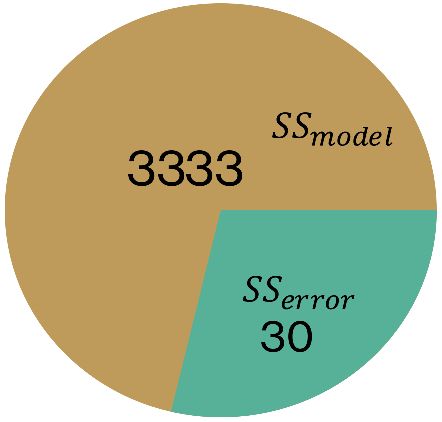
My	**r**
65.68888889	-0.3364151313
Mx	**b**
13.56088889	-1.609199866
Sy	**a**
17.55377226	87.51106948
sx	
3.669745893	

SStotal
69022.22222
SSerror
61210.62252
SSmodel
7811.599703
SStotal-SSerror
7811.599703

SS_{total} = 69022



SS_{total} = 3363



women weight ~ women height + error

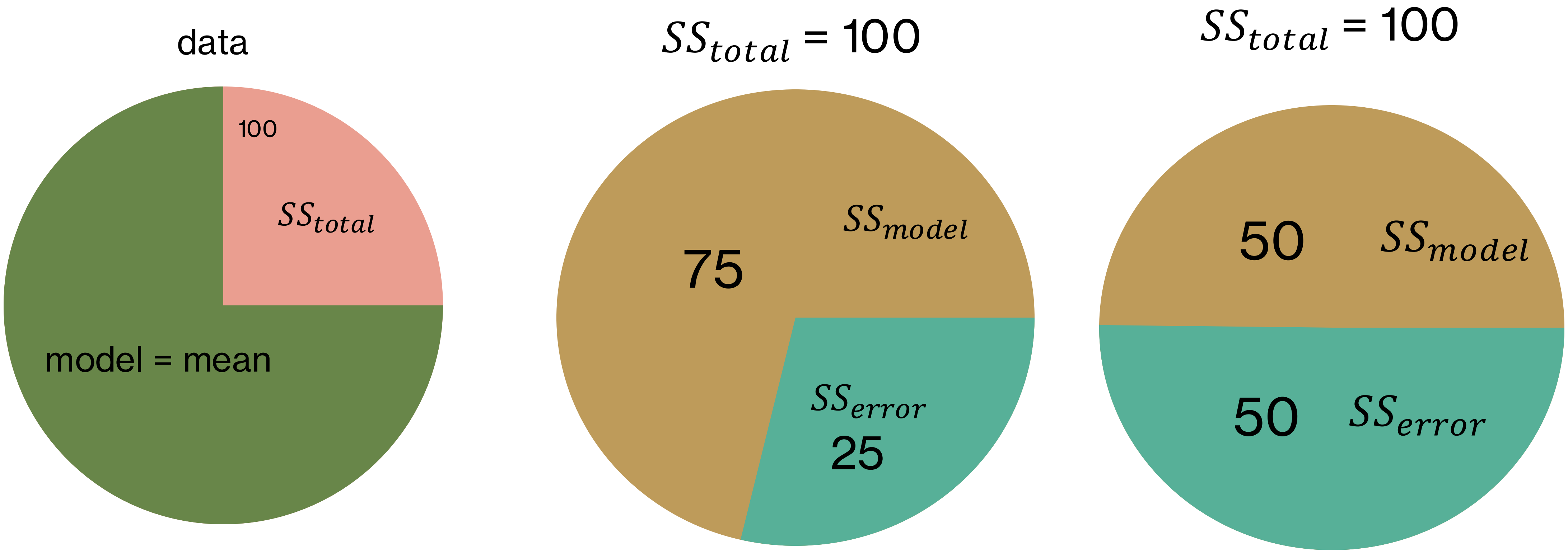
r	r^2
0.9954947678	0.9910098327
b	a
3.45	-87.51666667

SStotal	SSerror
3362.933333	30.23333333
SSmodel	SStotal-SSerror
3332.7	3332.7

coefficient of determination (R^2)

- what proportion of the 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 *100 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

coefficient of determination (R^2)



$$R^2 = \frac{SS_{model}}{SS_{total}} = .75$$

$$R^2 = \frac{SS_{model}}{SS_{total}} = .50$$

standard error of estimate: SE_{model} and SE_r

- how far away is an average data point from the line of best fit?
- similar concept to standard deviation, $s = \sqrt{\frac{SS}{n-1}}$ (how far is an average data point from the mean?)
- standard error of estimate (regression model) = "average" SS_{error}

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

- standard error for correlation = "average" unexplained variance

$$r^2 = explained\ variance$$

$$unexplained\ variance = 1 - explained\ variance = 1 - r^2$$

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

W5 Activity 4b

- to what extent can student to faculty ratio explain graduation rates across colleges?
- calculate percentage of explained variance (R^2), SE_{model} and SE_r

College	S.F.Ratio	Grad.Rate
Abilene Christian University	18.1	60
Adelphi University	12.2	56
Adrian College	12.9	54
Agnes Scott College	7.7	59
Alaska Pacific University	11.9	15
Albertson College	9.4	55
Albertus Magnus College	11.5	63
Albion College	13.7	73
Albright College	11.3	80
Alderson-Broaddus College	11.5	52
Alfred University	11.3	73
Allegheny College	9.9	76
Allentown Coll. of St. Francis de Sales	13.3	74
Alma College	15.3	68
Alverno College	11.1	55
American International College	14.7	69
Amherst College	8.4	100
Anderson University	12.1	59
Andrews University	11.5	46

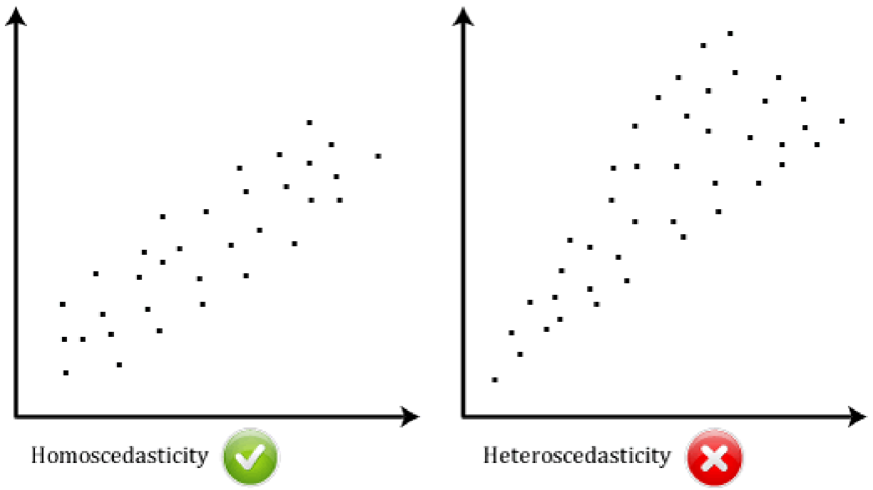
W5 Activity 4b debrief

- to what extent can student to faculty ratio explain graduation rates across colleges?
- calculate percentage of explained variance (R^2), SE_{model} and SE_r
- $R^2 = = \frac{SS_{model}}{SS_{total}} = r^2$ = 0.11
- $SE_{model} = \sqrt{\frac{SS_{error}}{n-2}}$ = 16.57
- $SE_r = \sqrt{\frac{1-r^2}{n-2}}$ = 0.06

College	S.F.Ratio	Grad.Rate
Abilene Christian University	18.1	60
Adelphi University	12.2	56
Adrian College	12.9	54
Agnes Scott College	7.7	59
Alaska Pacific University	11.9	15
Albertson College	9.4	55
Albertus Magnus College	11.5	63
Albion College	13.7	73
Albright College	11.3	80
Alderson-Broaddus College	11.5	52
Alfred University	11.3	73
Allegheny College	9.9	76
Allentown Coll. of St. Francis de Sales	13.3	74
Alma College	15.3	68
Alverno College	11.1	55
American International College	14.7	69
Amherst College	8.4	100
Anderson University	12.1	59
Andrews University	11.5	46

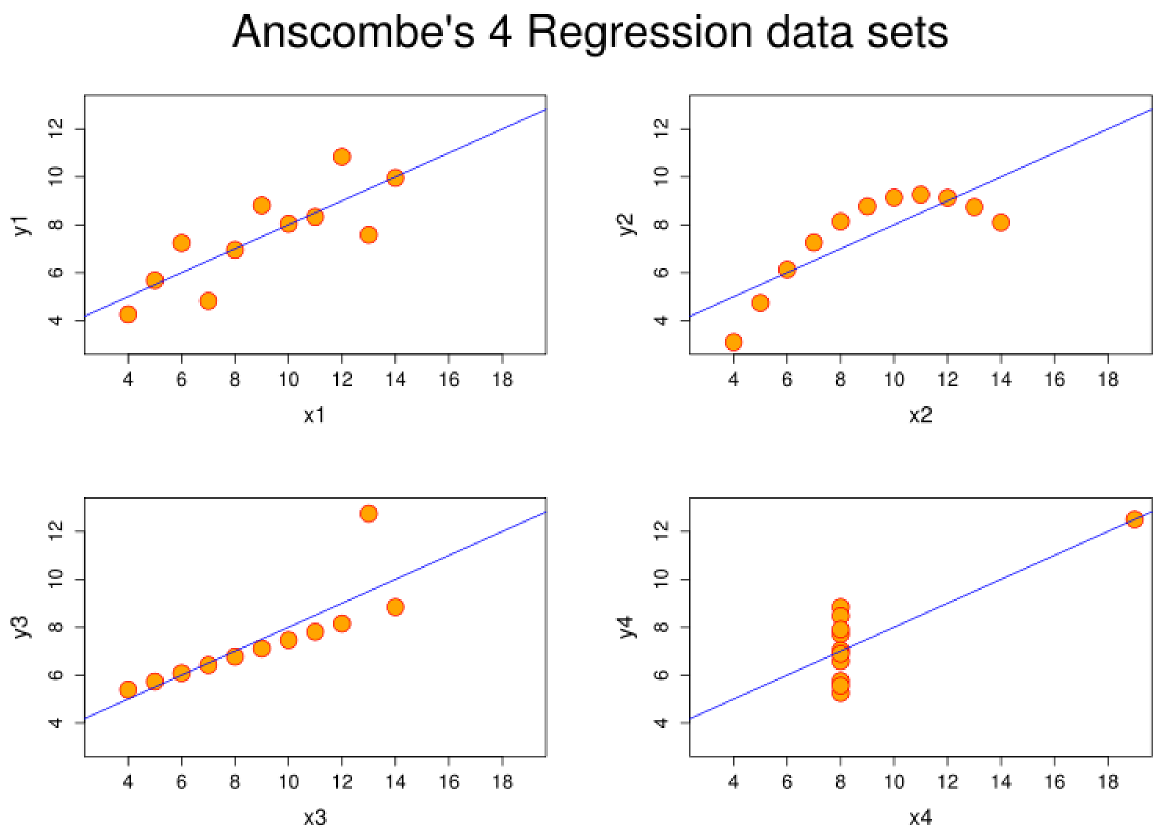
Pearson's r assumptions

- 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



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

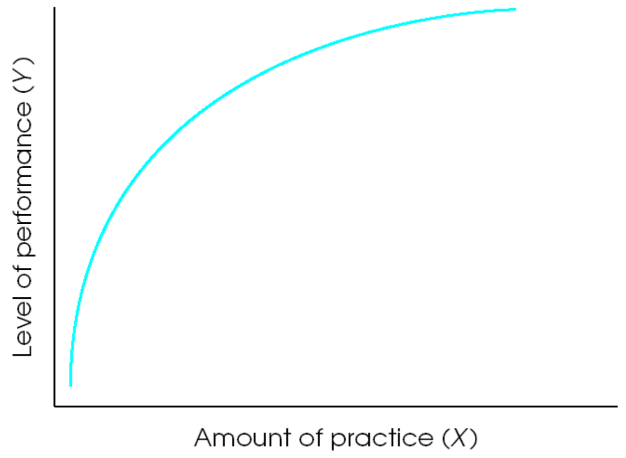


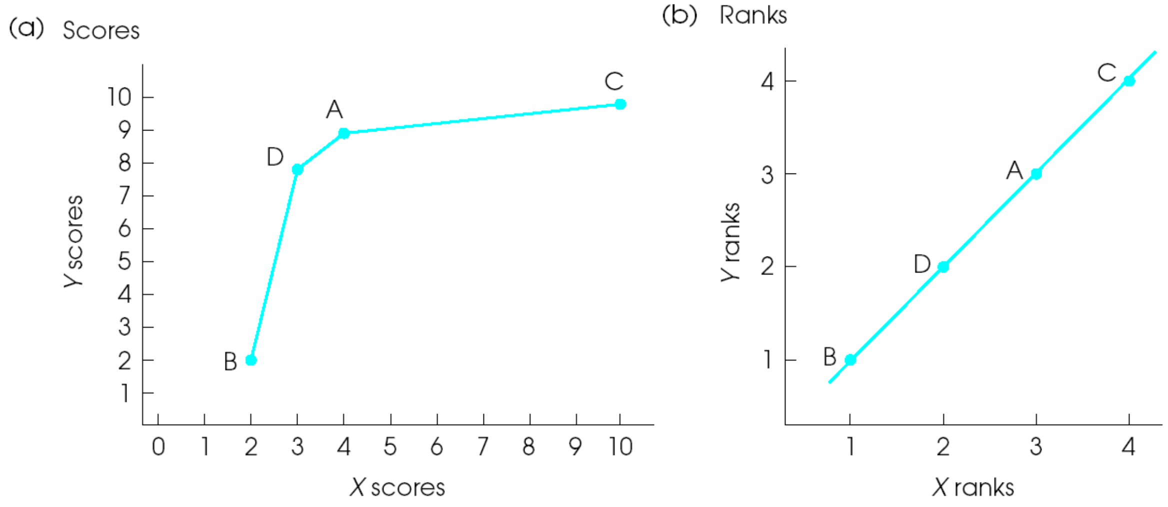
alternatives to Pearson's r

- 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

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





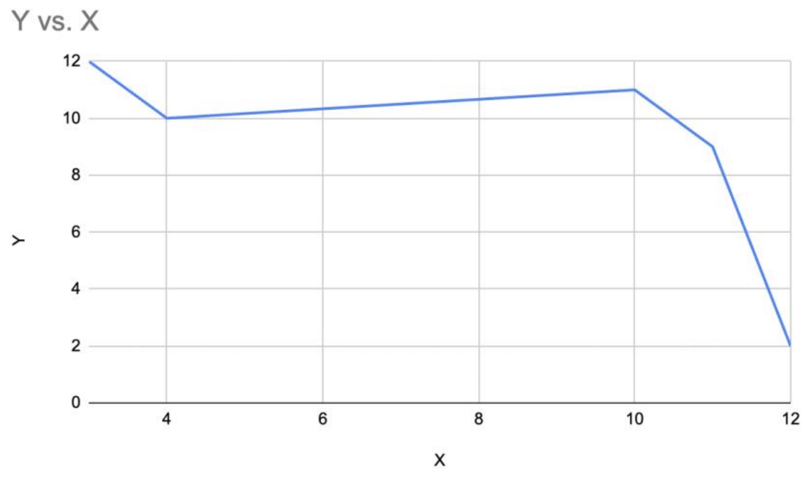
example

- 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	rank_x	rank_y
A	3	12	1	5
B	4	10	2	3
C	10	11	3	4
D	11	9	4	2
E	12	2	5	1

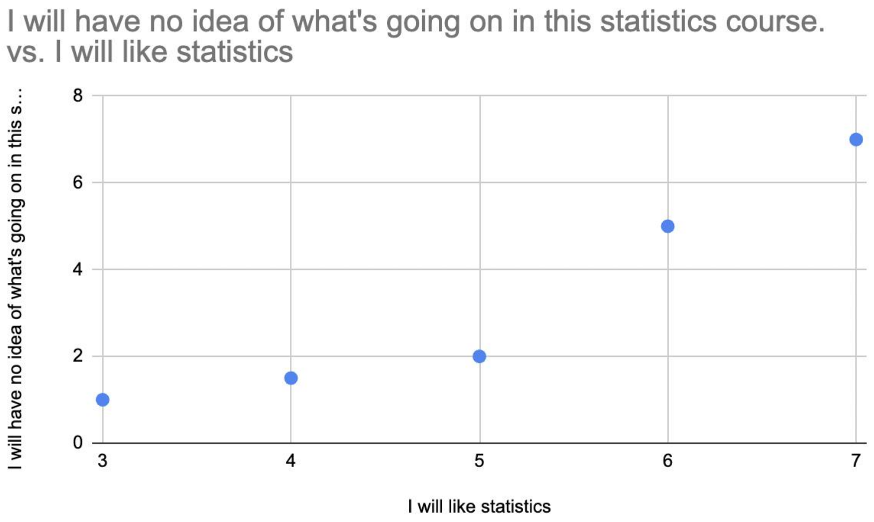
pearson	spearman
-0.6485442507	-0.9

W5 Activity 5

- calculate the correlation between two items from the statistics survey from class
- **data6**

Student	I will like statistics	I will have no idea of what's going on in this statistics course.
1	6	5
2	5	2
3	3	1
4	7	7
5	4	1.5

W5 Activity 5 debrief



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	5	4	4	1	0.9468131938
2	5	2	3	3		
3	3	1	1	1		
4	7	7	5	5		
5	4	1.5	2	2		

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	2	1.5
7	5	5.5
4	4	3.5
2	1	1.5
4	3	3.5

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
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

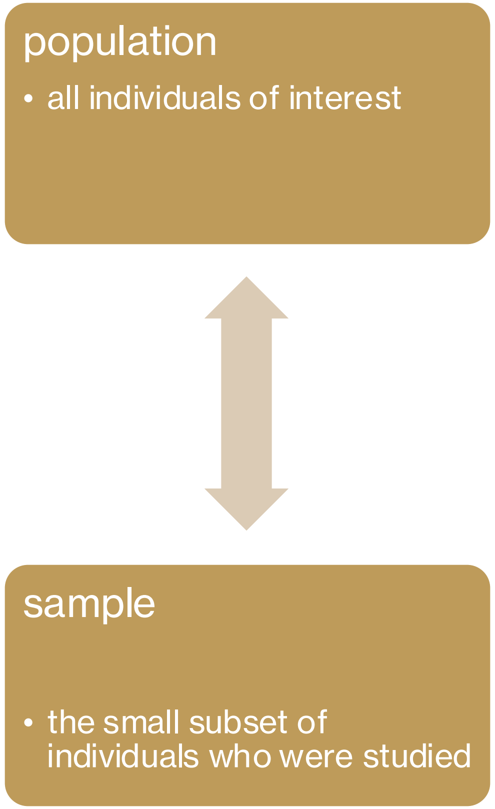
puzzle score	group	sqx	sqy	z_x	z_y	z_x*z_y
11	0	1	0.25	0.2901905	-1	-0.2901905
9	0	1	0.25	-0.2901905	-1	0.2901905
4	0	36	0.25	-1.741143	-1	1.741143
5	0	25	0.25	-1.4509525	-1	1.4509525
6	0	16	0.25	-1.160762	-1	1.160762
7	0	9	0.25	-0.8705715001	-1	0.8705715001
12	0	4	0.25	0.5803810001	-1	-0.5803810001
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			

W5 Activity 6

- Link will take you to canvas, 5 questions
- complete on your own
- discuss with a peer
- re-attempt the questions
- come back for a debrief

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

- midterm review

Before Tuesday

Try to complete these or at least skim through them by Tuesday. They will remain open until Wednesday night.

- Complete Practice Midterm 1 (Conceptual)
- Complete Practice Midterm 1 (Computational)

Thursday

- Submit Midterm 1 (Conceptual): IN CLASS

Here are the to-do's for this week:

- Submit Week 5 Quiz
- Submit Problem Set 3
- Complete Practice Midterm 1 (Conceptual)
- Complete Practice Midterm 1 (Computational)
- Submit any lingering questions here!
- Extra credit opportunities:
  - Submit Exra Credit Questions
  - Submit Optional Meme Submission

optional: spearman's *rho* D 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

optional: spearman's *rho* D 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