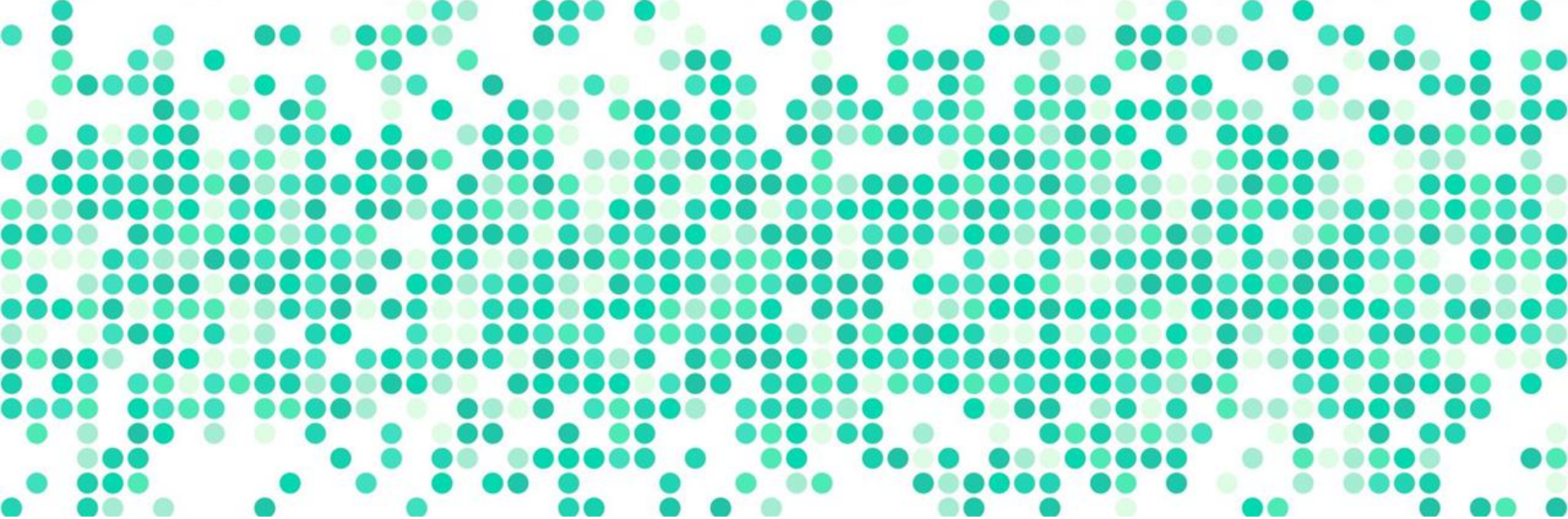


DATA ANALYSIS

Week 5: More correlation and regression

what's coming up

5	T: February 18, 2025	W5: More Correlation & Regression
5	Th: February 20, 2025	W6 continued…
5	Su: February 23, 2025	**Week 5 Quiz due**
5	M: February 24, 2025	**PS3 due**
6	T: February 25, 2025	W6: Loose Ends / Exam 1 review
6	W: February 26, 2025	**LA: Midterm Review (5-7.30 pm, Kanbar 101)**
6	Th: February 27, 2025	**Exam (Midterm) 1**
7	T: March 4, 2025	W7: Sampling and Hypothesis Testing
7	Th: March 6, 2025	W7 continued…
7	F: March 7, 2025	**PS3 revision due**
7	F: March 7, 2025	**Week 7 Quiz due**
8	T: March 11, 2025	**Spring Break!**
8	Th: March 13, 2025	**Spring Break!**
9	T: March 18, 2025	**Spring Break!**
9	Th: March 20, 2025	**Spring Break!**

logistics: midterm 1

Feb 27 2025

content: Week 1 through Week 5

two parts

- in-person conceptual portion (quiz-like + short answer)
  - closed book (+ help sheet)
- take-home computational portion (problem set-like: due Monday)
  - open book but NOT open person

practice exams now available

Apply

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

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

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

$$a = M_y - bM_x$$

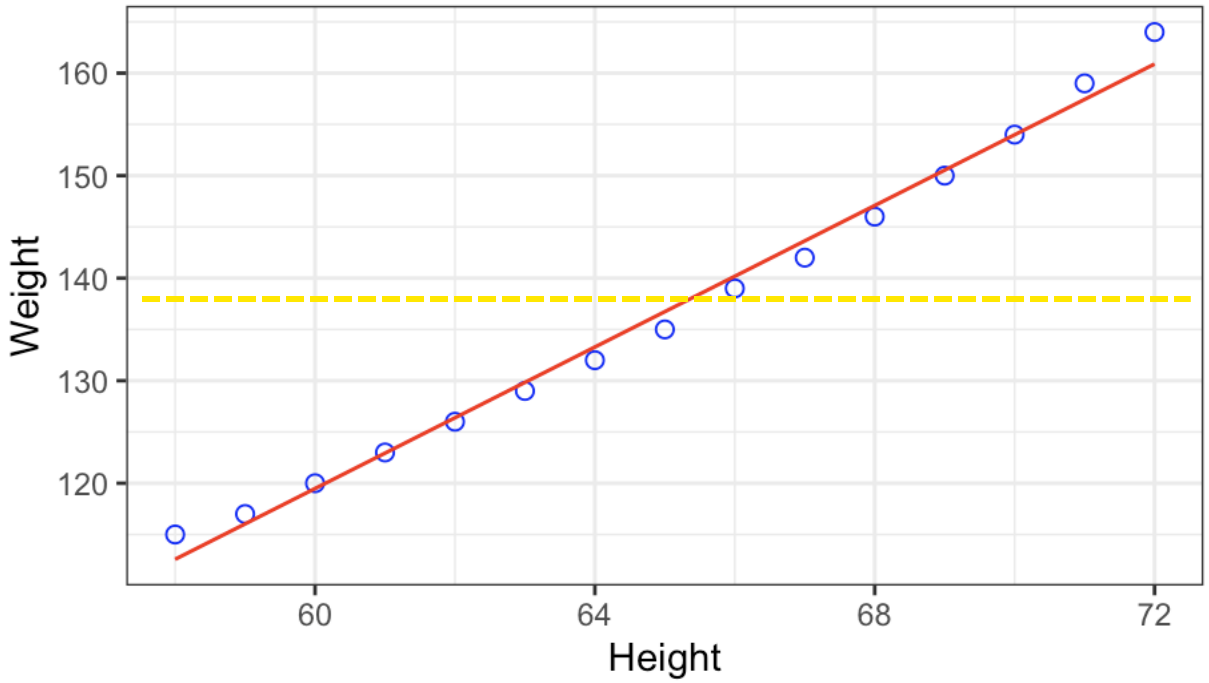
lingering question

- I'm still having trouble differentiating between samples and populations when calculating z-scores

populations	samples
variance (σ^2) = $\frac{\sum(X-\mu)^2}{N} = \frac{SS}{N}$	sample variance (s^2) = $\frac{\sum(X-M)^2}{n-1} = \frac{SS}{n-1}$
standard deviation (σ) = $\sqrt{\frac{\sum(X-\mu)^2}{N}} = \sqrt{\frac{SS}{N}}$	sample standard deviation (s) = $\sqrt{\frac{\sum(X-M)^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$
z-scores = $\frac{X-\mu}{\sigma}$	z-scores = $\frac{X-M}{s}$
correlation $\rho = \frac{\sum z_x z_y}{N}$	correlation r= $\frac{\sum z_x z_y}{N-1}$

special cases

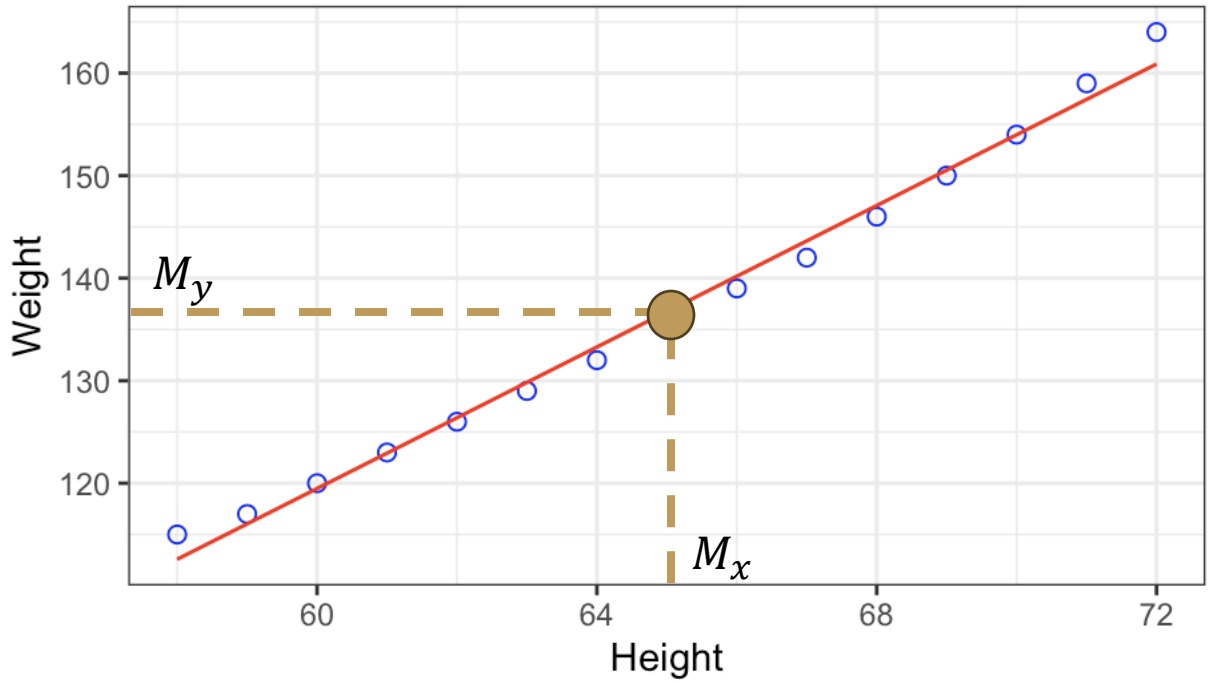
- no relationship between X and Y
  - r = 0, b = 0
  - Y = bX + a = a = $M_y - bM_x = M_y$
  - Y = M_y for all values of X
  - mean of Y is still our best model if there is no relationship between X and Y
- what is b when X and Y are standardized?
  - $b = r$ when $s_x = s_y = 1$



line of best fit & means

- $a = M_y - bM_x$
- $b = r\,\frac{s_y}{s_x}$
- rearranging the intercept equation:
  - $M_y = a + bM_x$
- the line of best fit passes through means of X and Y

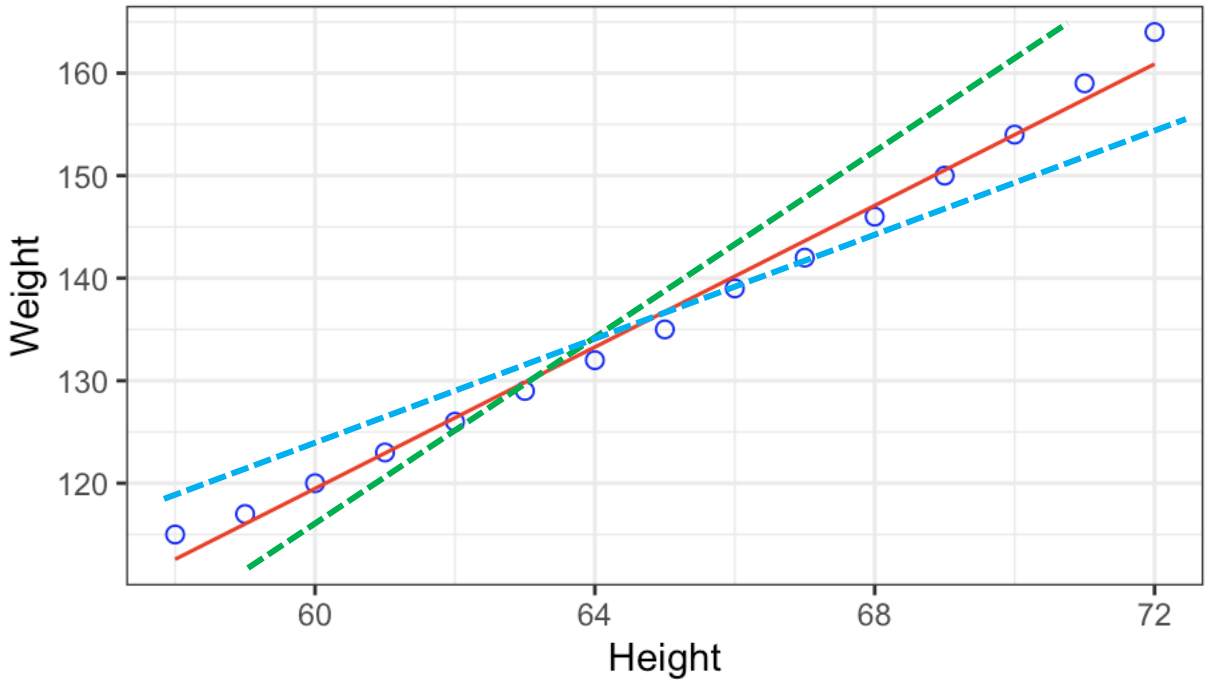
Y = a + bX + error



relationship between b and *r*

- $a = M_y - bM_x$
- $b = r\,\frac{s_y}{s_x}$
- the slope of the line (b) is simply the correlation adjusted to the original units of the data
- correlation and linear regression provide the same <u>conceptual</u> information about how two variables are related

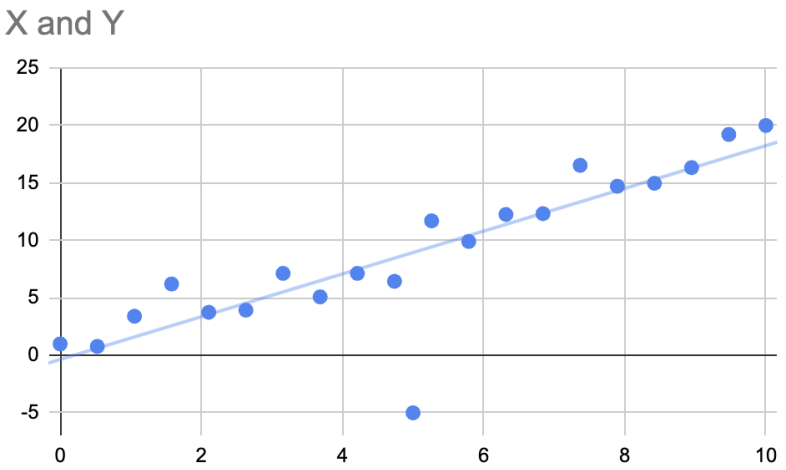
Y = bX + a + error



W5 Activity 1a

- calculate the correlation and slope for **data1**
- create a scatterplot with a trendline
- you can use the STDEV/CORREL formulas

stdev_X	r	b
3.034858893	0.849782375	1.859624145
stdev_Y		
6.641343762		



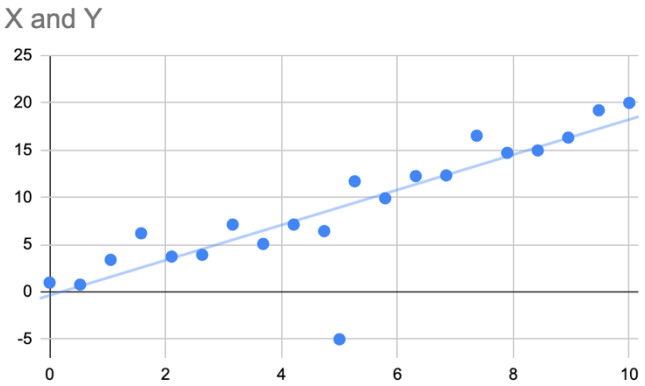
X	Y
0	0.993
0.526	0.776
1.053	3.401
1.579	6.204
2.105	3.742
2.632	3.929
3.158	7.128
3.684	5.082
4.211	7.129
4.737	6.437
5.263	11.696
5.789	9.904
6.316	12.252
6.842	12.322
7.368	16.524
7.895	14.706
8.421	14.961
8.947	16.323
9.474	19.209
10	19.99
5	-5

W5 Activity 1b

- as *r* increases, does *b* always increase?
- recalculate correlation and slope for **data2**

X	Y
0	0.993
0.526	0.776
1.053	3.401
1.579	6.204
2.105	3.742
2.632	3.929
3.158	7.128
3.684	5.082
4.211	7.129
4.737	6.437
5.263	6.696
5.789	4.904
6.316	7.252
6.842	7.322
7.368	11.524
7.895	9.706
8.421	9.961
8.947	11.323
9.474	14.209
10	14.99
5	10

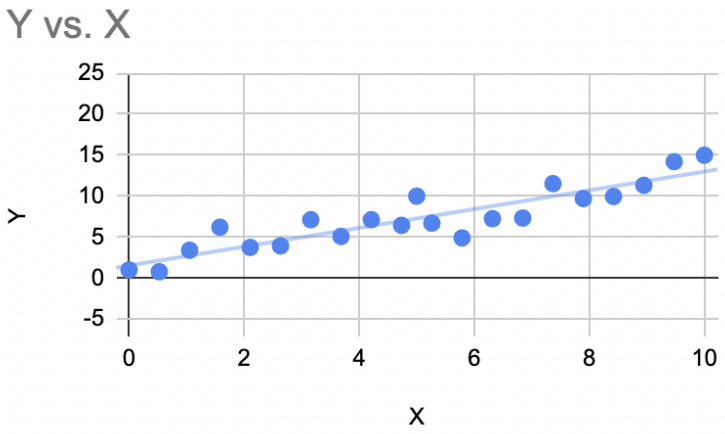
W5 Activity 1b



X	Y
0	0.993
0.526	0.776
1.053	3.401
1.579	6.204
2.105	3.742
2.632	3.929
3.158	7.128
3.684	5.082
4.211	7.129
4.737	6.437
5.263	6.696
5.789	4.904
6.316	7.252
6.842	7.322
7.368	11.524
7.895	9.706
8.421	9.961
8.947	11.323
9.474	14.209
10	14.99
5	10

- as r increases, does b always increase?
- recalculate correlation and slope for **data2**
- the spread of Y relative to X got smaller, even though r increased
- $b = r \frac{s_y}{s_x}$

stdev_X	r	b
3.034858893	0.9040865356	1.145347621
stdev_Y		
3.844729764		



W5 Activity 1c

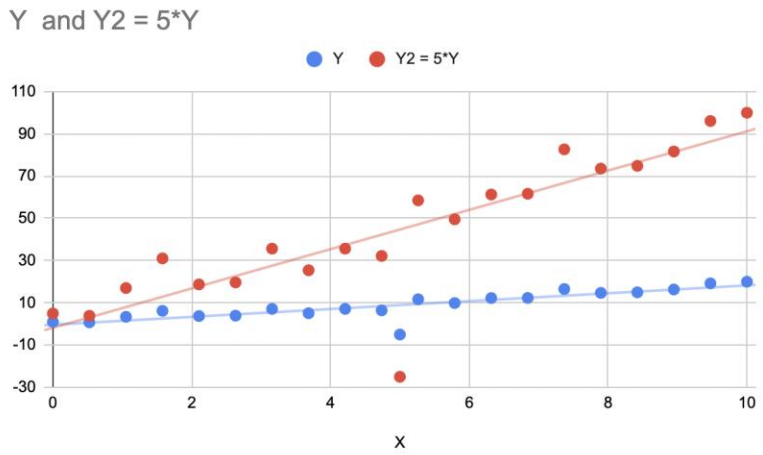
- if the spread of Y changes, do *r* and *b* both change?
- recalculate correlation and slope for **data3**

X	Y	Y2 = 5*Y
0	0.993	
0.526	0.776	
1.053	3.401	
1.579	6.204	
2.105	3.742	
2.632	3.929	
3.158	7.128	
3.684	5.082	
4.211	7.129	
4.737	6.437	
5.263	11.696	
5.789	9.904	
6.316	12.252	
6.842	12.322	
7.368	16.524	
7.895	14.706	
8.421	14.961	
8.947	16.323	
9.474	19.209	
10	19.99	
5	-5	

W5 Activity 1c

- if the spread of Y changes, do r and b both change?
- recalculate correlation and slope for **data3**
- correlation remains the same, slope increases
- correlation is scale-invariant
- $b = r \frac{s_y}{s_x}$

r	b
0.849782375	9.298120727
s_x	s_y
3.034858893	33.20671881



X	Y2
0	4.965
0.526	3.88
1.053	17.005
1.579	31.02
2.105	18.71
2.632	19.645
3.158	35.64
3.684	25.41
4.211	35.645
4.737	32.185
5.263	58.48
5.789	49.52
6.316	61.26
6.842	61.61
7.368	82.62
7.895	73.53
8.421	74.805
8.947	81.615
9.474	96.045
10	99.95
5	-25

W5 Activity 2

- on Canvas/course website
- two questions
- complete **on your own**
- discuss with a peer
- re-attempt

W5 Activity 2a debrief

- The average number of hours freshmen spend on social media per day is 5 (SD = 3). The average GPA in the freshman class is 3.5 (SD = 0.2). The correlation between time spent on social media and GPA is r = -0.75.
- Shaan's GPA is 3.7. What is your best prediction for how many hours Shaan spends on social media per day?
- X = GPA = 3.7, Y = social media
- $M_x = 3.5, M_y = 5$
- $s_x = 0.2, s_y = 3$
- $b = r\,\frac{s_y}{s_x}$ = -0.75 $\frac{3}{0.2} = -11.25$
- $a = M_y - bM_x = 44.375$
- $\hat{Y} = a + bX$ = $44.375 - 11.25\ (3.7)$ = 2.75

W5 Activity 2a debrief

- The average number of hours freshmen spend on social media per day is 5 (SD = 3). The average GPA in the freshman class is 3.5 (SD = 0.2). The correlation between time spent on social media and GPA is *r* = -0.75.
- Shaan's GPA is 3.7. What is your best prediction for how many hours Shaan spends on social media per day?
- X = GPA = 3.7, Y = social media
- $M_x = 3.5, M_y = 5$
- $s_x = 0.2$, $s_y = 3$
- $z_x = \frac{(X - M_x)}{s_x} = \frac{3.7-3.5}{0.2} = \frac{0.2}{0.2} = 1$
- $\widehat{z_y} = r\ z_x$ = -0.75 * (1) = -0.75
- $\widehat{z_y} = \frac{\hat{Y} - M_y}{s_y}$
- $\hat{Y}$ = $\widehat{z_y}\ s_y + M_y$ = -0.75*s_y + M_y = -0.75(3) + 5 = 2.75

W5 Activity 2b

- The average number of hours freshmen spend on social media per day is 5 (SD = 3). The average GPA in the freshman class is 3.5 (SD = 0.2). The correlation between time spent on social media and GPA is r = 0.75.
- Selena spends 6 hours on social media. What is your prediction for Selena's GPA?
- X = social media = 6, Y = GPA *notice the flip of X and Y!!
- could we use the same line's equation for this?
- $\hat{Y} = a + bX$ = $44.375 - 11.25X$

W5 Activity 2b

- The average number of hours freshmen spend on social media per day is 5 (SD = 3). The average GPA in the freshman class is 3.5 (SD = 0.2). The correlation between time spent on social media and GPA is r = 0.75.
- Selena spends 6 hours on social media. What is your prediction for Selena's GPA?
- X = social media = 6, Y = GPA *notice the flip of X and Y!!
- the same line's equation cannot be used, prediction is NOT symmetric!
- $M_y = 3.5, M_x = 5$
- $s_y = 0.2, s_x = 3$
- $b = r \frac{s_y}{s_x}$ = -0.75 $\frac{0.2}{3} = -0.05$
- $a = M_y - bM_x = 3.75$
- $\hat{Y} = a + bX$ = $3.75 - 0.05\ (5)$ = 3.45

W5 Activity 2b

- The average number of hours freshmen spend on social media per day is 5 (SD = 3). The average GPA in the freshman class is 3.5 (SD = 0.2). The correlation between time spent on social media and GPA is $r = 0.75$.
- Selena spends 6 hours on social media. What is your prediction for Selena's GPA?
- X = social media = 6, Y = GPA *notice the flip of X and Y!!
- $M_y = 3.5, M_x = 5$
- $s_y = 0.2, s_x = 3$
- $z_x = \frac{(X - M_x)}{s_x} = \frac{6-5}{3} = \frac{1}{3} = 0.33$
- $\widehat{z_y} = r\, z_x$ = -0.75 * (0.33) = -0.25
- $\widehat{z_y} = \frac{\hat{Y} - M_y}{s_y}$ and so $\hat{Y} = \widehat{z_y}\, s_y + M_y$ = -0.25*s_y + M_y = -0.25(0.2) + 3.5 = 3.45

today's agenda

more on correlations

assessing model fit

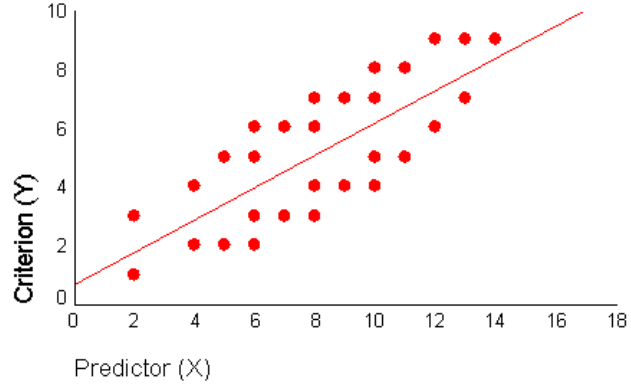
how good is the line of best fit?

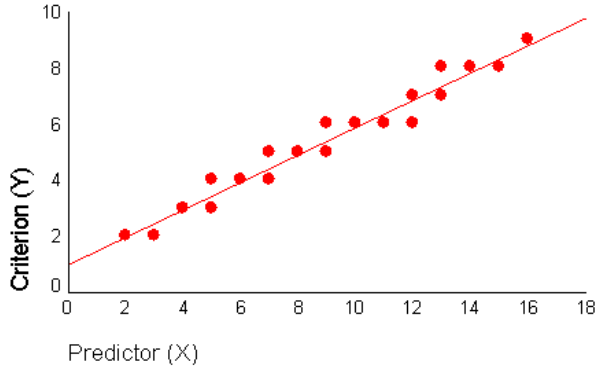
- even the line of “best” fit may ultimately not fit the data very well due to the inherent variability in the data
- how we assess model fit?
- data = model + error
- data = $a + bX$ + error
- our favorite friend: sum of squared errors (SS)!

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

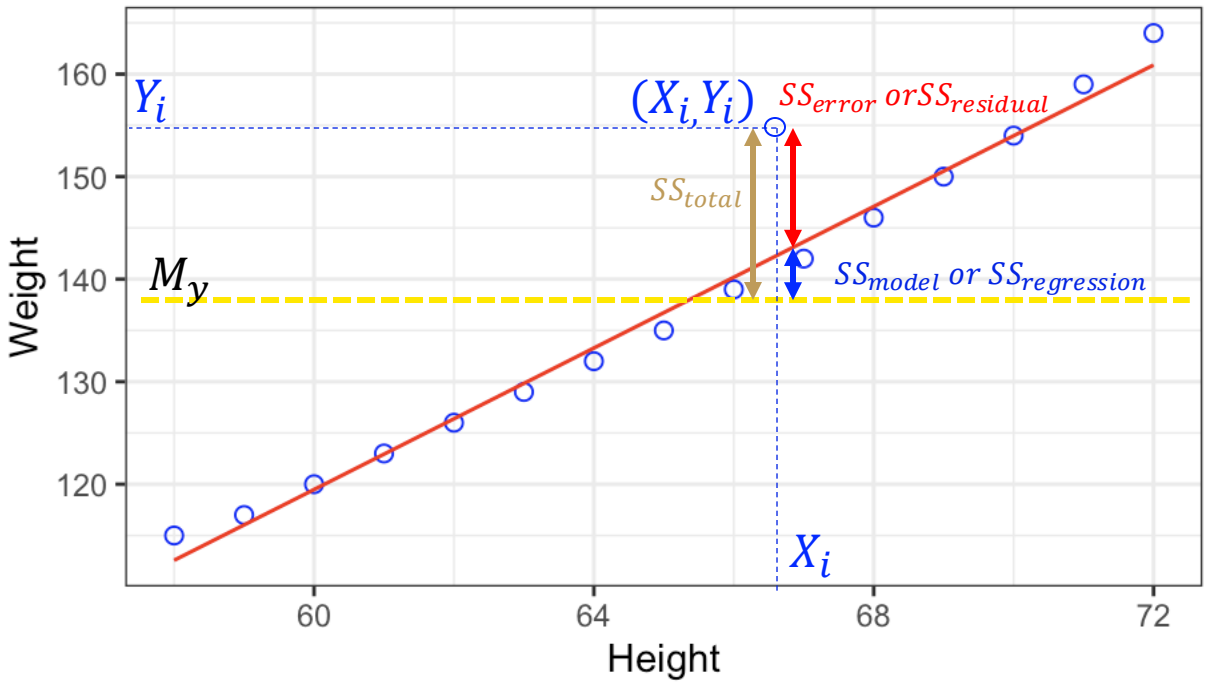
$$SS_{error} = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum(Y - \hat{Y})^2$$

- this represents the error left over after a line has been 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?

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

W5 Activity 3

- calculate all of these values for the women dataset in the data4 sheet

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?

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

next time

- **spearman & point biserial correlations**

Prep

Before Tuesday

- Start preparing for Midterm 1. Practice midterm is now available: see the Apply section. Submitting the practice midterm by next Monday counts towards class participation.

Before Thursday

- Watch: Spearman and Point Biserial Correlations.

After Thursday

- See Apply section.

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