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

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

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

$$a = M_y - bM_x$$

lingering question

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

populations

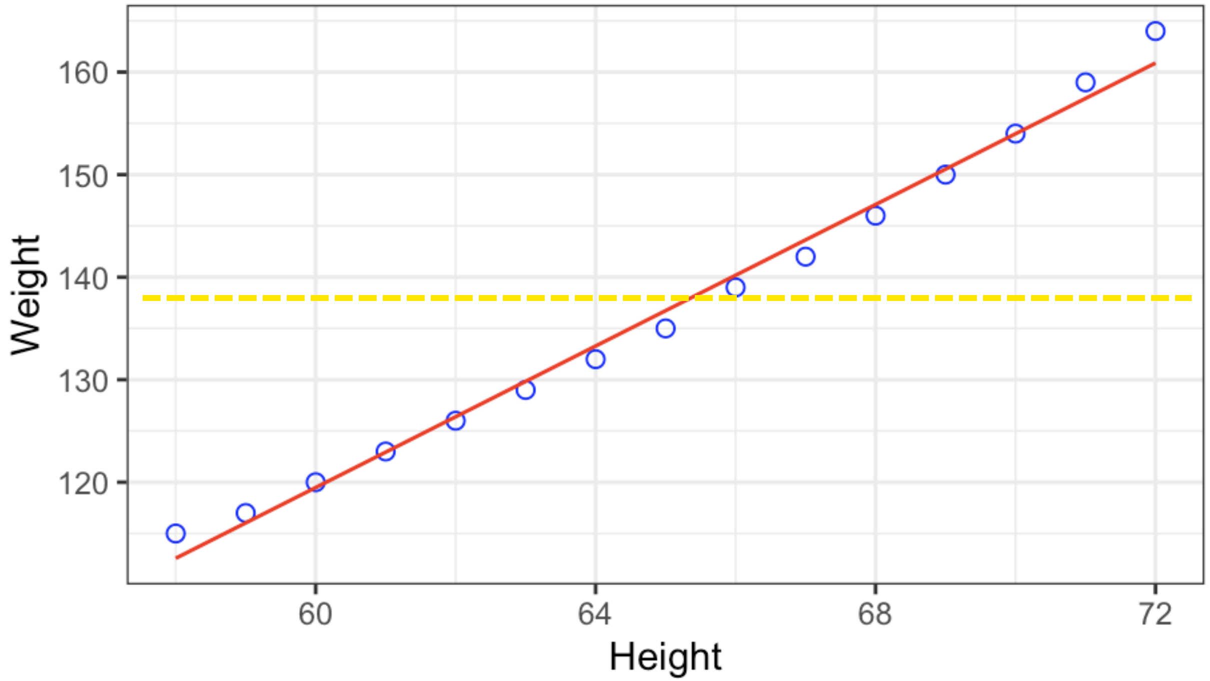
- variance (σ^2) $= \frac{\sum(X-\mu)^2}{N} = \frac{SS}{N}$
- standard deviation (σ)= $\sqrt{\frac{\sum(X-\mu)^2}{N}} = \sqrt{\frac{SS}{N}}$
- z-scores = $\frac{X-\mu}{\sigma}$
- correlation $\rho = \frac{\sum z_x z_y}{N}$

samples

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

special cases

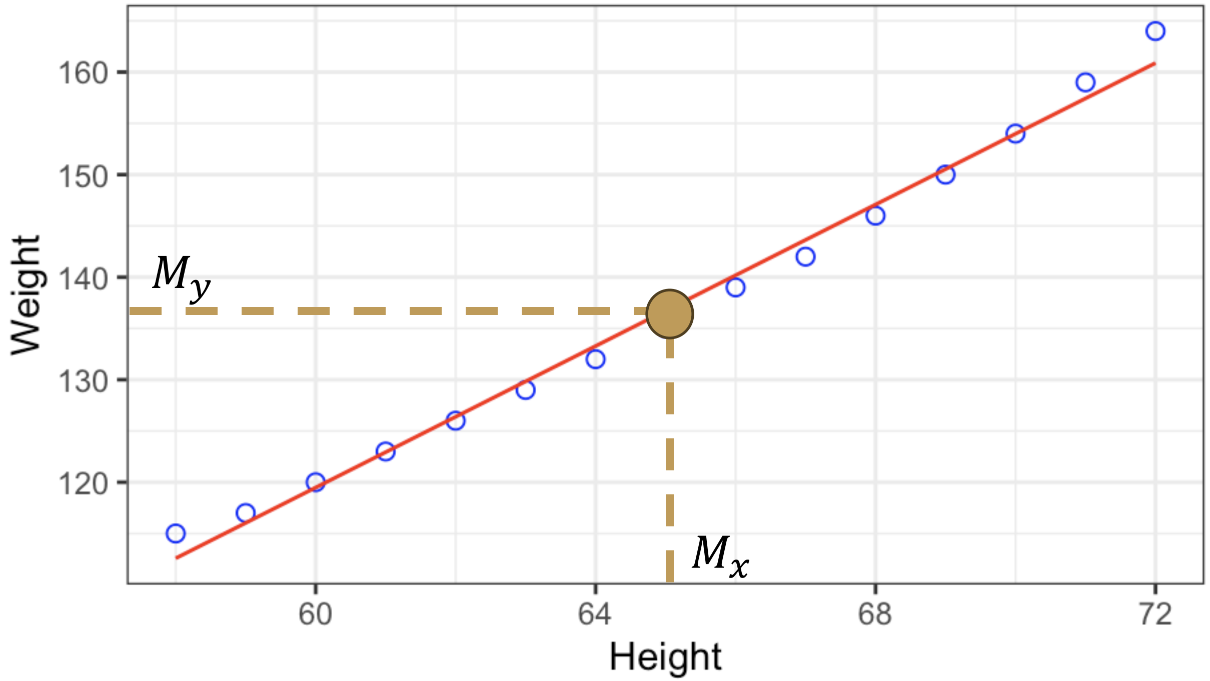
- no relationship between X and Y
  - *r* = 0, b = 0
  - Y = bX + a = a = $M_y - bM_x = M_y$
  - Y = M_y for all values of X
  - mean of Y is still our best model if there is no relationship between X and Y
- what is *b* when X and Y are standardized?
  - *b* = *r* when $s_x = s_y = 1$



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

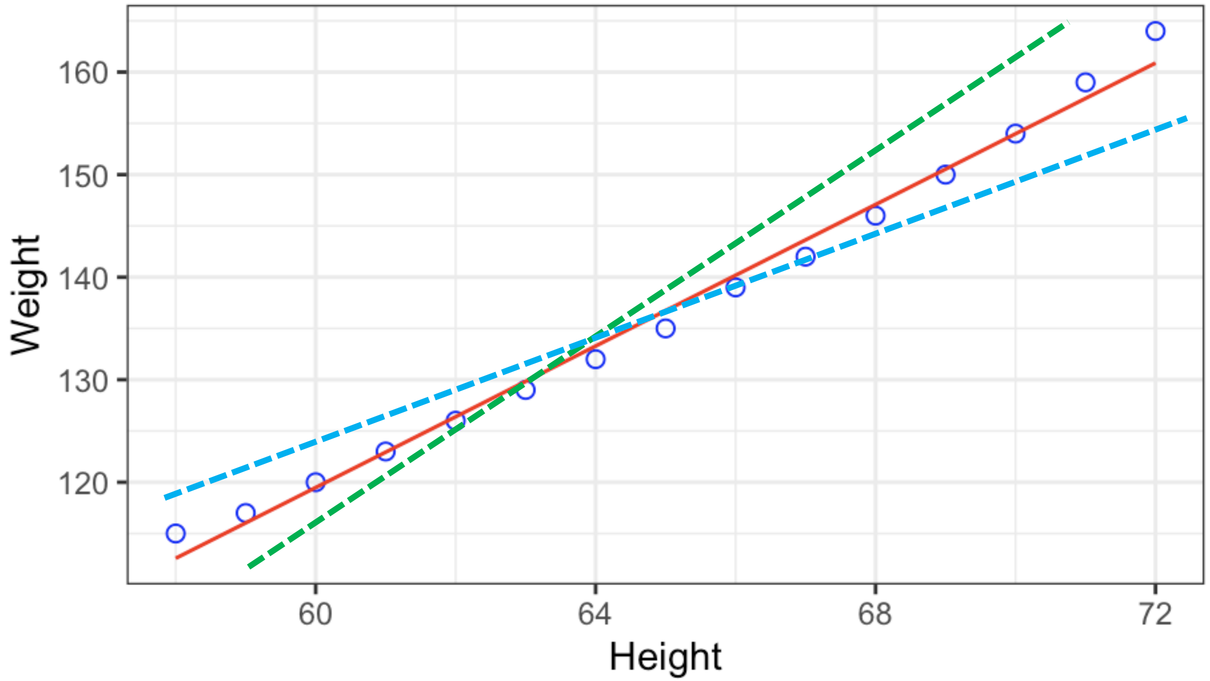
Y = bX + a + error



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 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

W5 Activity 1b

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

W5 Activity 1c

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

W5 Activity 2

- On Canvas: complete via link

today's agenda

- more on correlations
- assessing model fit

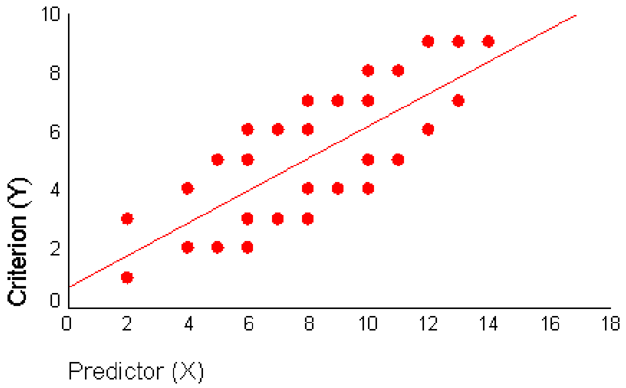
how good is the line of best fit?

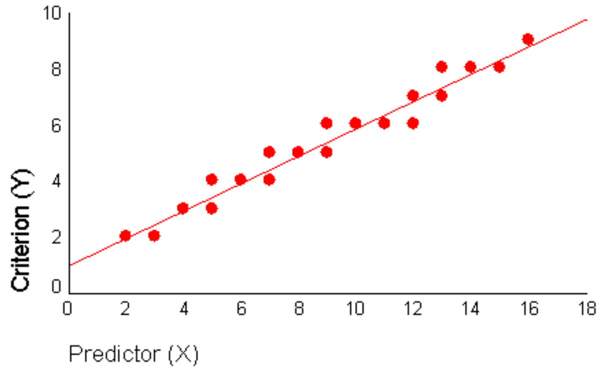
- even the line of “best” fit may ultimately not fit the data very well due to the inherent variability in the data
- how we assess model fit?
- data = model + error
- data = $a + bX$ + error
- our favorite friend: sum of squared errors (SS)!

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

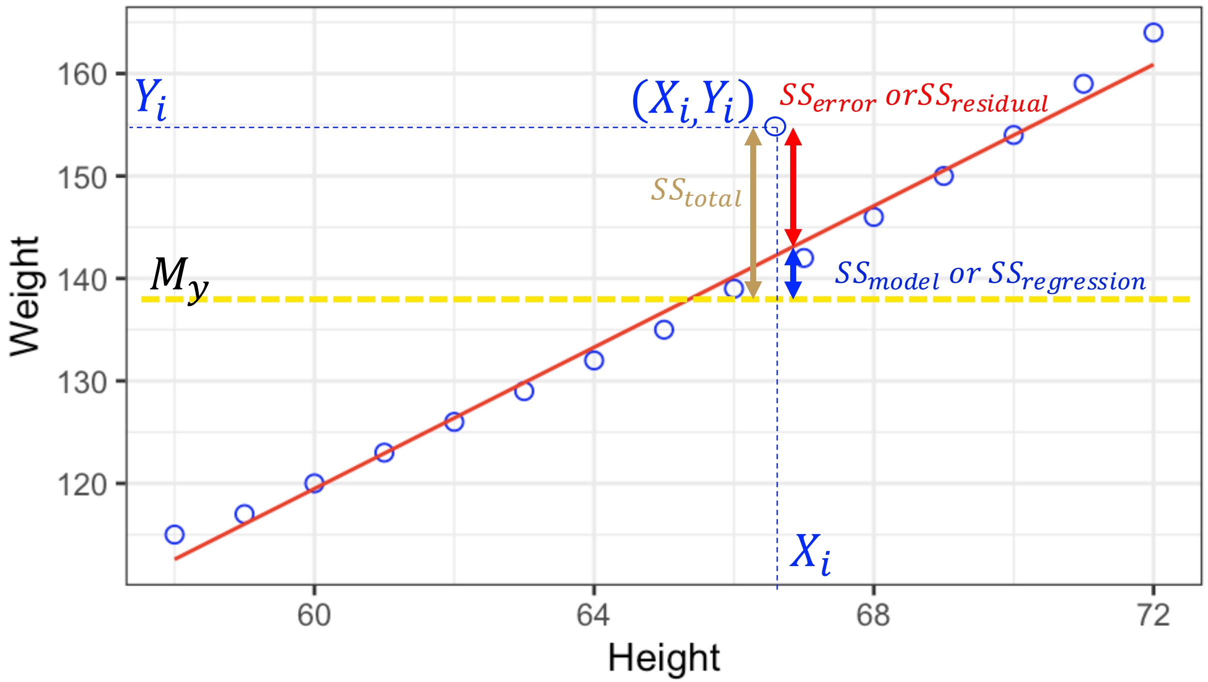
$$SS_{error} = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum(Y - \hat{Y})^2$$

- 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?

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

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 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

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}

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

- standard error for correlation

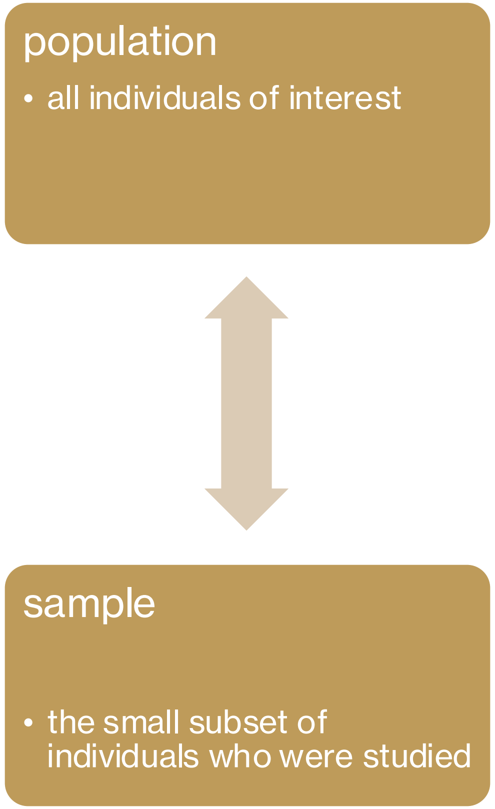
$$r^2 = \text{explained variance}$$

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

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

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?
- we will compare what we have observed (in the sample) vs. what is expected (in the population), by making some assumptions
- after midterm 1!



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