MA206, Lesson 19 - Two Quantitative Variables

How do we visually inspect association between two quantitative (numerical) variables?

What three aspects of association do we use when looking for correlation?

Define **Direction**.

Define **Form**.

Define **Strength**.

What is the correlation coefficient?

What is the difference between an outlier and an influential observation?

How do we express our least squares regression line to predict or explain our response variable?

How do we measure the goodness of fit for our regression line?

What is our null and alternate hypothesis for our regression line comparing two quantitative variables?

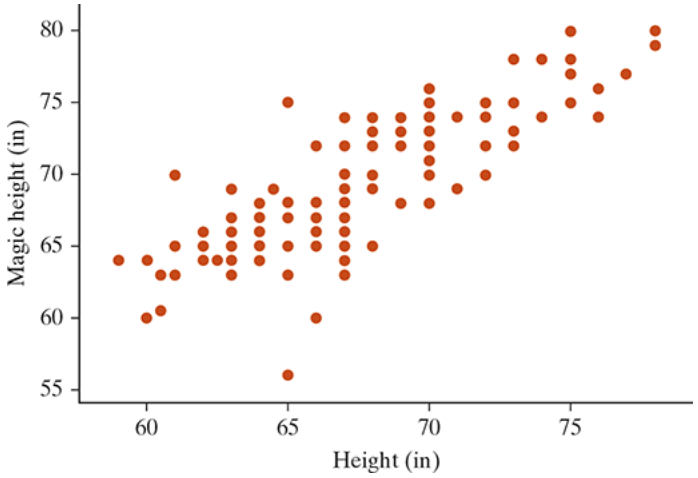
What is the standardized statistic for the population slope / association (β_1)?

t =

How do we calculate the p-value given this standardized statistic?

How do we use *RStudio* to calculate our regression line and its p-value?

1) The image below represents the results of a survey of college students, where data was gathered on their actual height (height) in inches and the height they would like to be if they could be any height (magical height). The results have a correlation of 0.842.



a) Describe the direction, form, and strength of the association between magical height and height.

b) The equation of the regression line is $\widehat{magic\ height} = 5.51 + 0.9471 \times (height)$. What does the slope mean in words?

c) What is the value of R^2, and what does this number mean in this context?

2) A regression table is shown below based on data used to test an association between the amount of sleep someone had the previous night (in hours) and the time needed to complete a paper and pencil maze (in seconds). Sleep is the explanatory variable and time needed to complete the maze is the response.

Term	Coefficient	SE	t-stat	p-value
Intercept	198.33	51.75	3.85	0.003
Sleep	-7.76	3.04	-2.55	0.012

a) What is the regression equation where time to complete the maze is predicted from the amount of sleep?

b) If we were testing against the alternative hypothesis $\beta_1 \neq 0$, what is the p-value?

c) If we were testing against the alternate hypothesis $\beta_1 < 0$, what is the p-value?

d) Interpret the results. Use a significance level of 0.05.

3) A 1997 study by Roger W. Johnson[1] measured the physical measurements of 184 randomly selected female students, aged 18 - 25, at Brigham Young University. The students measured also underwent an underwater weighing technique, found to be a very accurate measure of body fat percentage. Using the measurements data and the calculations prescribed in AR 600-9, we also calculated the Army Body Weight measurement and saved it as CIndex. For reference, for women, AR 600-9 estimates

$$Percent\ body\ fat = [163.205 \times Log_{10}(waist + hip–neck)]–[97.684 \times Log_{10}(height)]–78.387$$

We are interested in the accuracy of the Height and Weight results, as they directly relate to the careers of all members of the military. We will be comparing the variables Fat , the true body fat of each student, and CIndex , the Army estimate used by AR 600-9.

a) What is our research question?

b) Define and classify the two variables of interest as either categorical or quantitative.

c) Generate and label a scatterplot of the data.
Comment on the Form, Direction, and Strength of the association in the scatterplot.

d) What is the correlation coefficient value between the two variables of interest?

e) What is the expected R^2 for a simple linear regression line on our variables?

f) In words and symbols, write the null and alternate hypotheses

g) Calculate the regression equation with CIndex as the explanatory variable and Fat as the response.

h) Interpret the coefficient found in your model.

i) Generate a scatterplot with a least squares regression line over the points.

j) Report the standardized statistic and p-value for β_1. Interpret this p-value.

k) Record and interpret the R^2 of the model. How does this compare to part e above?

l) For females 21 and older, the acceptable body fat percentage allowed using AR 600-9 calculations is 32%. For simplicity, apply this standard to all 184 females 25 and under. Based on this dataset, how many would be flagged under 600-9 even if they are under body fat percentage? How many would pass 600-9 requirements even if they were over? How many are correctly identified as overweight?

[1] https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1971585