MA206, Lesson 19 - Two Quantitative Variables

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

We use a Scatter Plot. Typically, the explanatory variable is on the x-axis and response variable is on the y-axis. It is typically named "***Figure 1: Response Variable vs. Explanatory Variable***"

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

Direction, Form, and Strength

Define **Direction**.

The positive or negative relationship or association between two quantitative variables. As one increases, the other typically increases or decreases, respectively.

Define **Form**.

The general shape or pattern of the association. Specifically, we are interested in if the relationship is linear or otherwise.

Define **Strength**.

How closely the points follow the pattern (form).

What is the correlation coefficient?

The correlation coefficient measures the strength and direction of a linear association. It is denoted by r and has the property that $-1 \leq r \leq 1$. We calculate this in R using cor()

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

An outlier is an observation that doesn't follow the overall pattern. An influential observation is one that, if it is removed from the dataset, dramatically changes the calculations/results of our regression line slope.

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

$\hat{y} = \beta_0 + \beta_1 x_1$

where $\hat{y}$ is our expected y value, β_0 is the intercept, β_1 is the slope and x_1 is our (first) explanatory variable.

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

We use R^2, which for simple linear regression is equal to r^2, or $1 - \frac{SSR}{SST}$

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

Our null is that there is no linear association between the two variables; Our alternate is that there is a linear association (two-sided t-test)

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

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

t = $\frac{\beta_1 - 0}{SE(\beta_1)}$

Alternatively, with r calculated, you could calculate $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

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

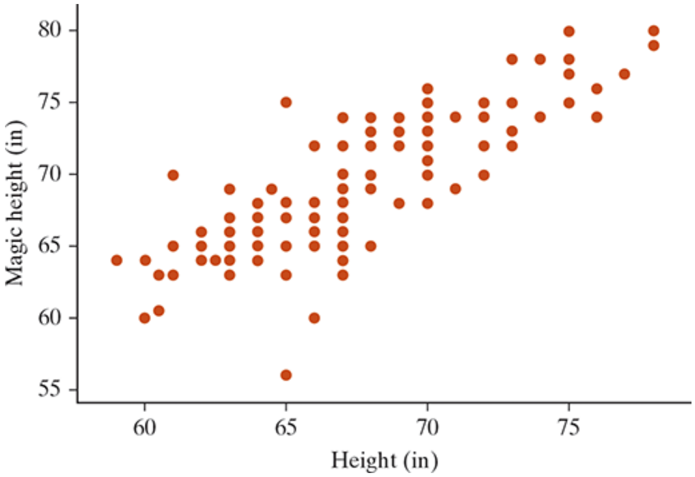
This will be a t-statistic using a t-distribution with *n-2* degrees of freedom. In general, this is a **NOT EQUAL TO** test, but greater than or less than tests are also *possible* in some applications.

$\beta_1 \neq 0$	$2 * (1 - pt(abs(t), n-2))$
$\beta_1 \geq 0$	$1 - pt(t, n-2)$
$\beta_1 \leq 0$	$pt(t, n-2)$

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

```
Model = dataset % > %
     lm(Y ~ X, data = ., na.action = na.exclude)
summary(Model)
```

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.

The direction is positive, the form is linear, and the association is strong.

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

For every additional inch of height, with all else being equal, the magical height is expected to increase by 0.9471 inches.

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

$R^2 = 0.842^2 = 0.709$. This indicates that 70.9% of the variation in magical height can be explained by the linear association with the *height* variable.

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?

$\widehat{time} = 198.33 - 7.76 \times (Sleep)$

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

0.012 (from our table)

c) If we were testing against the alternate hypothesis $\beta_1 < 0$, what is the p-value?
$\frac{0.012}{2} = 0.006$

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

With a p-value of 0.012, we can conclude that we have statistically significant evidence against the null that there is no association between sleep and the time to complete the maze.
Our best fit model estimates that for every additional hour of sleep, the amount of time to complete the maze will decrease by 7.76 seconds.

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?

Is there an association between the calculated Body Fat Percentage and true, water weighed Body Fat Percentage?

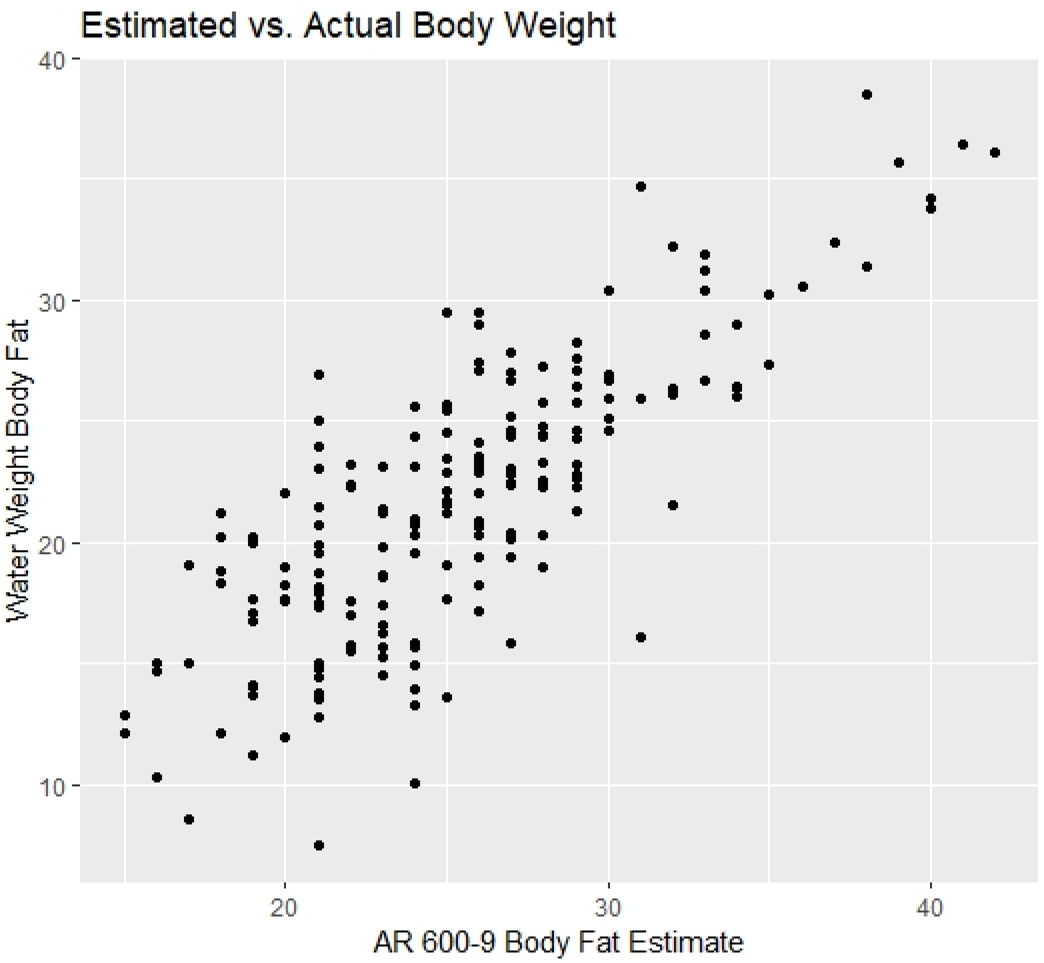
b) Define and classify the two variables of interest as either categorical or quantitative.

CIndex and Fat are both quantitative variables.

c) Generate and label a scatterplot of the data.

Comment on the Form, Direction, and Strength of the association in the scatterplot.

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



The form is linear, the Direction is positive, and the Strength appears fairly strong.

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

r = 0.7945106

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

$R^2 = 0.7945106^2 = 0.6312471$

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

$H_0 : \beta_1 = 0$. There is no linear association between the AR 600-9 Estimated Body Fat formula and the Water Weight Technique body fat calculation.

$H_a : \beta_1 \neq 0$. There is a linear association between the AR 600-9 Estimated Body Fat formula and the Water Weight Technique body fat calculation.

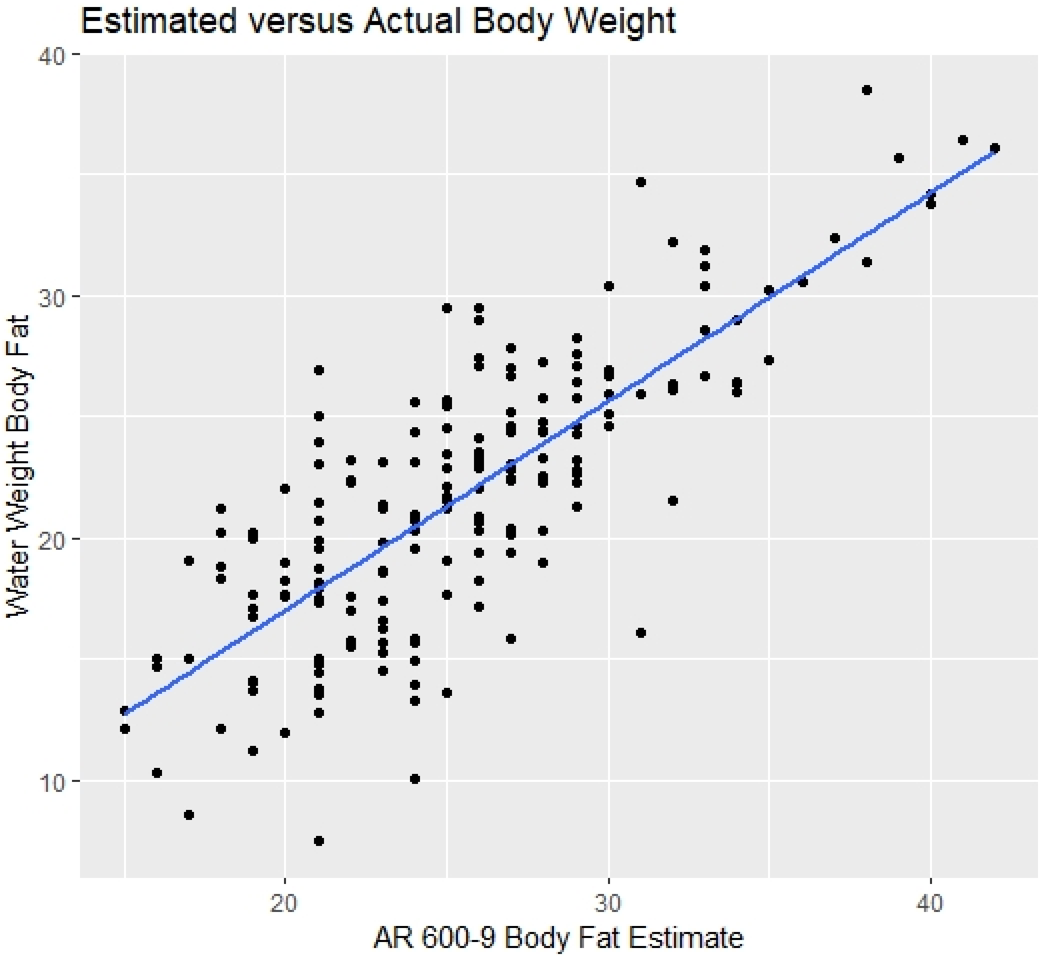
g) Calculate the regression equation with CIndex as the explanatory variable and Fat as the response.

$\widehat{Fat} = -0.23937 + 0.86234(CIndex)$

h) Interpret the coefficient found in your model.

For every point of 600-9 Body Fat Estimate increase, the actual water weight Fat percentage is expected to increase by 0.86234 percent.

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.

t = 17.65

p-value < $2e^{-16}$

The probability of observing an association at least as strong as our sample, assuming there truly is no association, is less than $2e^{-16}$. Therefore, we have very strong evidence against the null and conclude that there is a positive linear association between AR 600-9 Estimates and Water Weight body fat.

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

$R^2 = 0.6312$, which is equal to part *e* above.

We can account for 63.12% of the variability in water weight measurements if we know the AR 600-9 body fat estimates.

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?

Of the 184, there are 13 who are flagged under 600-9 but have a body fat percentage within tolerance.
There are 2 who are not flagged under 600-9, but ave a body fat percentage over 32%.
There are 7 correctly identified as having a body fat percentage over 32%.

To summarize, 13 of the 20 females that would be flagged (65%) actually have a body fat percentage within tolerance. This is a "false positive" flag.
2 of the 9 females that are overweight are not flagged (22%). This is a "false negative" flag.
Of the 175 females within tolerance, 7 of them (7.4%) would be erroneously flagged.