***COURSERA: STATS W/ R SPECIALIZATION***

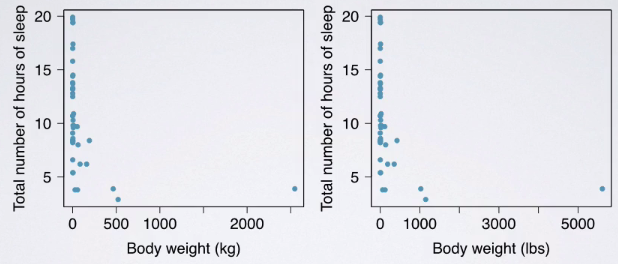
***COURSE 3 - Linear Regression and Modeling***

**WEEK 1 - Linear Regression**

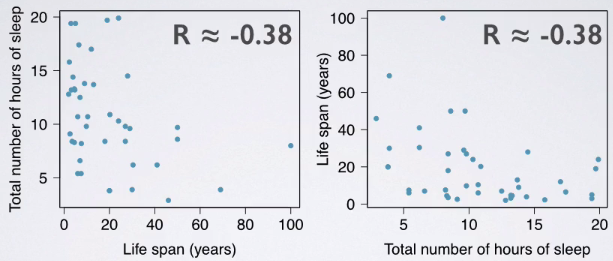
***5.1.1 Relationship between two numerical variables***

**Correlation**

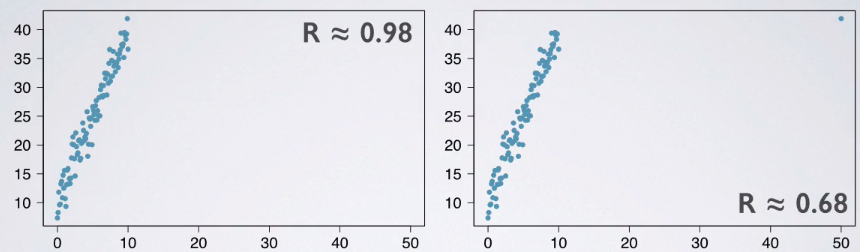
* **Correlation** = a measure of the strength of the linear relationship/association between 2 numerical variables.
* *Only* measure the *linear* association w/ correlation = **R**
* **Magnitude** = absolute value of the **correlation coefficient** **R** = Measures strength of linear association between 2 numerical variables.
* **Sign** of the correlation coefficient = direction of association.
* Correlation coefficient = always between -1 (perfect negative linear association) + 1 (perfect positive linear association), + correlation coefficient = 0 means no linear relationship.
* Correlation coefficient is **unitless** + is NOT affected by changes in center or scale of either variable, such as unit conversions.



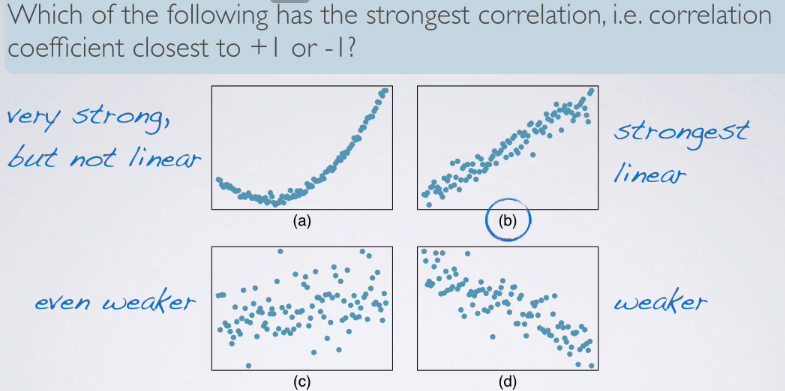
* Shape of relationship of the 2 variables looks very similar + for both plots correlation coefficient R = -0.34, no matter if x in terms of kg or lbs
* Probably wouldn't want to claim a linear relationship between these variables b/c the relationship between these 2 variables do not appear to be linear, but to demonstrate the fact that changing units does not affect the correlation coefficient, these plots still serve a purpose.
* **Correlation of X w/ Y = same as of Y / X**.
* Even if you swap the axis, correlation coefficient R should stay the same



* Correlation coefficient R = **sensitive to outliers**.



* Even w/ 1 outlier, b/c the correlation coefficient is sensitive, it will change greatly.
* So we can see that moving even one data



**Residuals**

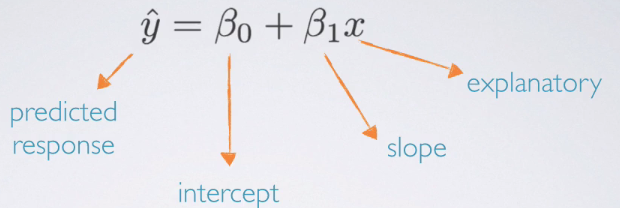
* **Residuals** = leftovers from the model fit = the difference between the observed + the predicted Y
* Observed data = the model fit + the residuals



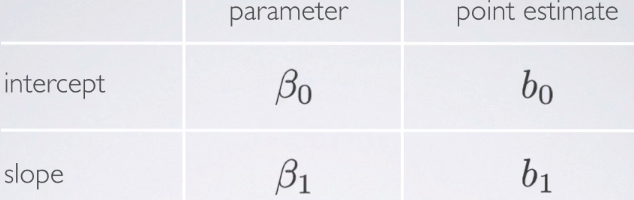
* Residuals tell us how much less/greater an actual DP is than what a model predicts. Similarly in D.C., the observed value is

**Least Squares Line**

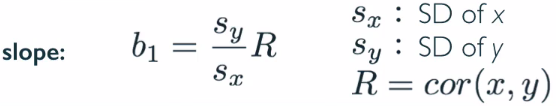
* Need the **least squares line** b/c can't simply add up all residuals b/c some are negative + some positive, depending on whether the model is over or underestimating certain DPs.
* Need to come up w/ a more clever approach:
* Minimize sum of *magnitudes* (absolute values) of their residuals
* Minimize sum of *square residuals* = the least squares
* Least squares = most commonly used approach + is also easier to compute by hand + w/ software.
* But most importantly, a residual twice as large as another = MORE than twice as bad (same idea as SD)



* If we have data from a sample, use that sample to estimate unknown population parameters.



* Greek = population, Latin = sample
* To minimize sum of squared residuals, could actually use calculus + calculate slope + intercept
* Can also use shortcut formula 🡪 **slope b1 = SD\_y / SD\_x (rise / run) times correlation coefficient R**



* Ex: SD of % living in poverty = 3.1% + of % of high school graduates = 3.73% in our data set. Given that the correlation between these variables = -.75, what is the slope of the regression line for predicting % living in poverty from % of high school graduates?

> sd.pov <- 3.1

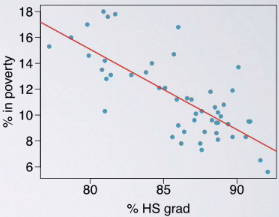
> sd.hs <- 3.73

> R <- -.75

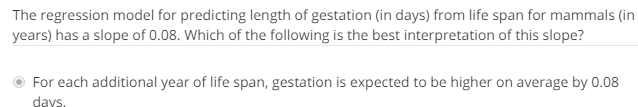
> (slope <- (sd.pov/sd.hs)\*R)

[1] -0.6233244

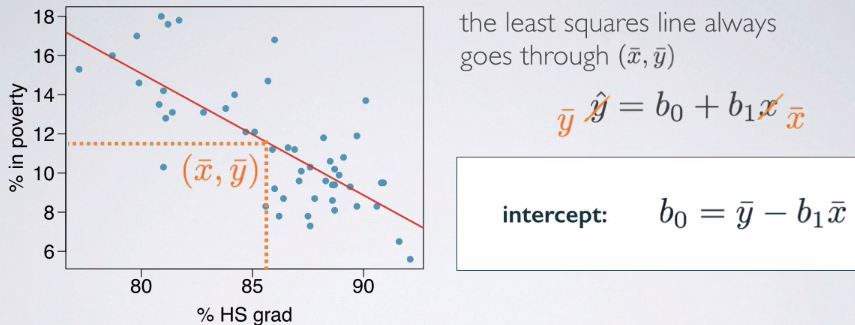
* Note: sign of the correlation coefficient R.
* Conceptually speaking, this is true, b/c we clearly seeing a negative relationship between the 2 variables so it makes sense the slope is negative.



* Mathematically speaking, remember SD = square root of the **variants** + is a measure of variability
* So the SD of y + x are always going to be positive, so we’re always dividing 2 positive #’s + multiplying by a negative or positive value depending on the direction of the relationship between the 2 variables.
* So, mathematically speaking, **sign of the slope always = sign of the correlation coefficient**.
* Calculating the slope by hand is clearly very simple and actually sometimes
* For each % point increase in HS graduate rate, we’d expect the % living in poverty to be lower, *on average*, by 0.62% points.



* Slope = relationship between explanatory + response variable = how we expect response variable to change as we increase explanatory variable by 1 unit.
* When interpreting these, also make sure if dealing w/ an *observational study, avoid causal language*
* We expect this to happen on average
* Next, we want to estimate the intercept (where regression line crosses y axis).
* For this, make use of the property that least squares line always goes through x\_, y\_ = always goes through mean of y + x.
* We know we can write the linear model: y^ = b0 + b1\*x 🡪 plug in x\_ + y\_ in b/c we know the line MUST go through this point
* Then rearrange things a bit to get the formula for the intercept.



* Intercept = average value of response variable – slope times average value of explanatory variable.
* Given the average % living in poverty = 11.35% + average % of HS grads = 86.01%, what is the intercept of the regression line for predicting % living in poverty from % of high school graduates?

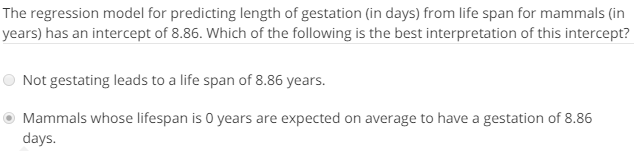
> pov\_bar <- 11.35

> hs\_bar <- 86.01

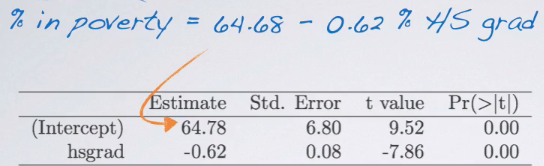
> (int <- pov\_bar - slope\*hs\_bar)

[1] 64.96213

* Understand that the regression line always goes through the *center of the data* + we interpret the intercept as where the regression line crosses the Y axis.
* In other words, it's the expected value of the response variable when the explanatory variable = 0
* So, in context, what we can say states w/ no HS graduates are expected, *on average*, to have 64.68% of residents living below the poverty line.
* Does this seem realistic, that there would be states in the U.S. w/ absolutely no HS graduates?
* Looking at the data we have, it actually seems very unlikely as all states in the US actually have HS graduation rate varying somewhere between 75% to ~95%, maybe.
* So mathematically speaking, this is a construct that is important for putting together a linear model
* However, in context, it is not a very useful number.



* So putting the info together from the previous 2 steps, we can write our linear model as a predicted % living in poverty (64.68%, the intercept) - minus 0.62 (slope) \* % that’re HS graduates.



* Estimate column = we can find our **parameter/coefficient estimates** for slope + intersect.
* So to recap, **intercept**: when x = 0, y is expected to = the intercept (may be a meaningless value in context of the data + in those cases it might only be serving to adjust the height of the line)
* The interpretation of the slope is slightly different = about the relationship between the 2 variables
* For each unit increase in X, Y is expected to be higher/lower on average by the value of the slope.
* Depending on type of study, be careful about interpreting the slope as causal vs. correlational

***5.1.2 Linear Regression w/ 1 Predictor***

**Prediction + Extrapolation**