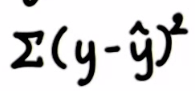
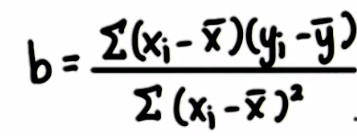
***Udacity Data Analyst Track***

**I. Into to Inferential Stats**

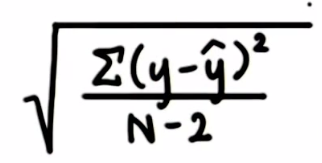
10. Linear Regression

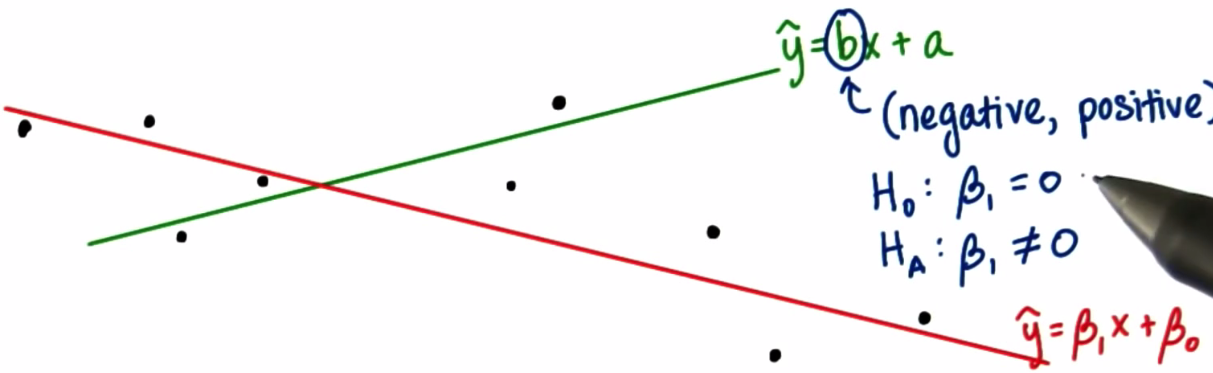
* Remember Pearson’s’ coefficient/correlation coefficient, r, measures strength of the relationship between 2 variables
* Looks at how x and y vary together (covariance) vs. how they vary w/in each other/independently
* Can describe linear data relationships via a straight line through the data **= regression line/line of best-fit**
* Describes the data + makes it easier to see relationships between x and y + then use it to make good predictions/estimates of y based on x
* *Expected value of DV based on IV(s)*
* Expected value of y from regression line = y(^), actual values = y 🡪 have 1 of these for each data point
* Difference between each y(^) and y = **residual** 🡪 basically error in prediction value
* **Slope** = the angle at which the regression line intercepts y-axis (amount y changes as x changes 1 unit)
* Regression line slope = how much y is *expected* to change w/ respect to x, not how much it will *actually* change
* **Regression coefficients** = the y-intercept + the slope of a variable
* To find line of best-fit, want to find the line that minimizes the sum of the residuals
* Must square the residuals, since some are negative (y(^) is less than y)



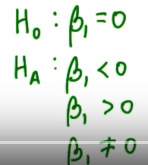
* This gives the sum of squared residuals = **sum of squares = SS**, which we want minimized via calculus + find out what our regression coefficients should be
* Via calculus, we get a slope of , which is equivalent to r multiplied by the ratio of the SD of y to the SD of x 🡪 **r\*(SD(y)/SD(x))**
* Then, to find the y-intercept, a, we need to know just 1 value of a data point *on the regression line* to solve for it (need it to be on the regression line so we know we’re calculating the actual y-intercept of the regression line 🡪 can’t just use any data point from the sample
* The y-intercept *should* go through the data point of the means of the variables (mu(x), mu(y)), since the best-fit line must go through the best-estimate of the x values and the best estimate of the y-values
* Then we can use these values of the slop, be, and value of y that a certain value of x should go through, to solve for the y-intercept

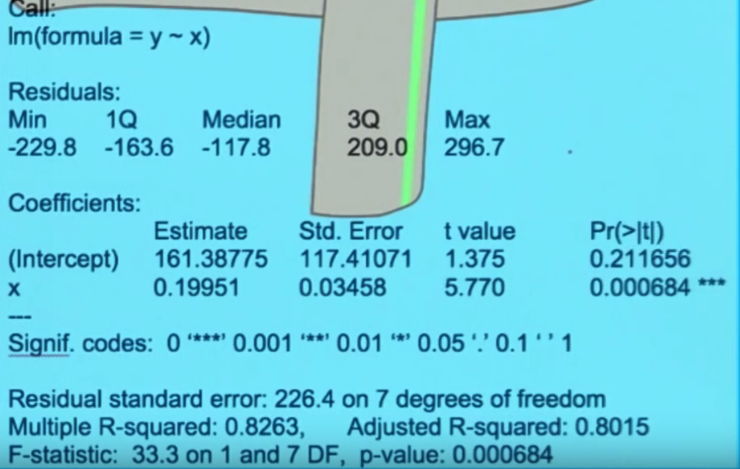
* Here a =160.03
* To calculate our error, take all residuals, square them, and find the average square while correcting for the fact that this is a sample, and then take the square root to get back to the original units
* **Standard Error of the Estimate =** 
* This can help us assess the accuracy of our predictions, but we can use a CI for predicted values to make more accurate predictions, as well as a CI for the true slope



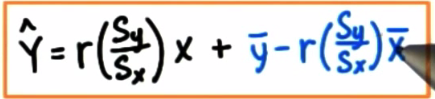
* Green = sample, red = population
* Sample regression line is positive sloping and population line is negatively sloping
* So sample slope CI has a negative lower bound + positive upper bound ( b/c 0 is w/in the range)
* So if we ran a 2-tailed hypothesis test to see if the slope was = 0, we’d *fail* to reject h(0): B1 = 0, meaning there’s no evidence of a linear relationship between x and y, based on this sample
* Hypothesis Testing for Slope tells us basically the same as hypothesis testing for r
* If the test for r is significant, then the test for slope will also be significant b/c both tests ask the same general question, “are the 2 variables linearly related?)
* B(1) = population slope + B(0) = population y-intercept
* b = slope of sample best-fit line + y = y-intercept of sample best-fit line

 🡪 another type of t-test

* Linear regression for the flight distance + cost data from Lesson 9

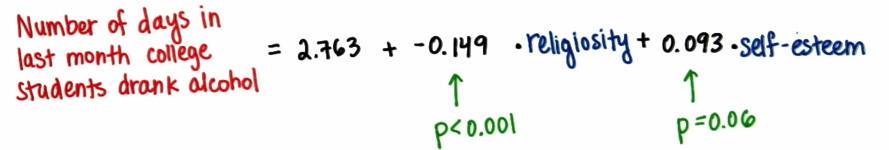


* See hypothesis test for the slope in the 2nd row 🡺 large t-value + p value small enough to reject null (\*\*\* = reject null @ level 0.001)
* See hypothesis test for the y-intercept in the 1st row 🡺 smaller t-value + p value larger @ 0.2 w/ no \*’s = not sure the true intercept is at 161.375
* In most cases the intercept estimate is not of interest to researches b/c this value could have no meaning real life b/c x = 0 may not be realistic (price of iPods in 1955, when they didn’t exist
* We know outliers can affect r, the correlation coefficient, and they can also affect the linear regression line (can pull it up or down)
* Full Linear Regression equation:



* Can also use multiple predictors to do **multiple regression** to explain more of the variation in outcome variable y
* We **regress** y on predictor1\*x1 and on predictor2\*x2 through predictor(n)\*x(n)



* Easier to do 1 big equation than a bunch of single regressions for each predictor + also we can calculate relationships between predictors + y independent of other predictors
* We no longer get simple slopes + get regression coefficients for each predictor variable, telling us the change in y in y units for a change of 1 x unit, holding the others constants
* i.e. see the mathematical influence of 1 variable while statistically controlling for the influence of the other variables
* **R = multiple correlation coefficient** (similar to Pearson’s coefficient, r, but includes more than 1 predictor variable) 🡪 tells us strength of the relationship between outcome variable + the combined set of predictors
* Usually more interested in **R2 =** proportion of variability in Y explained by the *set* of predictors
* We also remove variables + their coefficients if the regression coefficient is not sig
* 
* i.e. remove self-esteem variable from data

