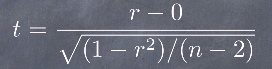
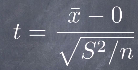
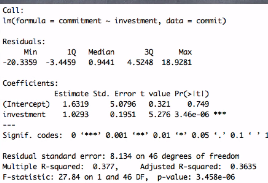
* A **linear model** which includes things like correlation, simple regression, and multiple regression, can use an inferential test for significance.
* It turns out these inferential tests use two distributions we are already familiar w/, namely the t distribution and the F distribution.
* Taken together these two distributions allow us to completely investigate a linear model.
* Our question asks about commitment in a relationship among college aged participants.
* We want to know about the relationship between commitment + some other quantitative variable, in this case, investment
* The 1st thing we want to know is if there is a significant relationship between these 2 quantitative variables.
* To do that we'll use a simple correlation.
* When using a Pearson correlation, make sure the relationship is in fact linear by examining a scatterplot.
* Once we know that the correlation is acceptable we can go ahead and ask for the actual Pearson correlation coefficient, which by itself, is free of context and scale + sometimes difficult to interpret.
* This is why we typically square the value to make *it more meaningful by giving us the* **coefficient of determination** *or proportion of variance accounted for.*
* But what about inference? Is there a statistical test we can use that tells us if this Pearson correlation coefficient is significantly different from zero?
* There is:
* William Gossett worked for Guinness brewery + worked on the concepts of small sample data + how a normal distribution didn't quite work well w/ a smallish n.
* Gossett's fix was to develop something call the t-distribution w/ Karl Pearson.
* The t-distribution is a perfect fit for the correlation coefficient b/c it stands to reason that we won't actually know the population parameters of a sample we’re using to calculate the Pearson Correlation, and that we should also take into account the size of the sample (n).
* It makes sense that this measure of relationship should also use the t-distribution + the idea of degrees of freedom.
* So, here's the formal t-test for a Pearson correlation:



* And the t-test that we are all used to when examining means:



* Notice the similarities.
* We are comparing some relational value to 0 (or, “no relationship.”)
* Also notice the denominator. W/ the t-test this was (variance / N)
* W/ the t-test for the Pearson correlation coefficient, it’s the *left over error in the relationship* denoted by 1 - minus R2
* So for any Pearson correlation coefficients we can calculate a t-statistic, which then allows us to examine the significance of that relationship.
* To put this into perspective here is simulated data based on a study looking at relationship commitment among University undergraduates in a monogamous relationships
* The Pearson correlation coefficient looking at the relationship between commitments and investments in the relationship was r = 0.6140
* And the corresponding t-test statistic for that Pearson correlation coefficient was t = 5.2764
* Using this t-test statistic, we can find the corresponding p-value + determine whether or not this slightly positive linear relationship is statistically significant.
* But remember, a correlation coefficient is only PART of the relational story.
* To fully explain the relationship between 2 quantitative variables, we can turn around and use **simple regression**, which also uses an underlying linear model, to help describe a relationship.
* Unlike the Pearson correlation coefficient, simple Regression allows us to **predict** a particular outcome.
* Here's the linear model predicting commitment score from investment score for the same data,



* The simple linear regression shows us this r-squared value, which is literally the Pearson correlation coefficient we found earlier squared.
* Also notice that T value for the investments coefficients matches the t-value found earlier
* This matches b/c there's only 1 variable in our regression model (*Investment is the only thing predicting commitment*), so we are effectively looking at the relationship between 2 variables, just like the Pearson correlation.
* Using the coefficient estimate for investment, we can continue to tell our relational story.
* See that for every 1 unit increase in investments, the commitment scale value for subjects changes by the amount is captured in the coefficient estimate = 1.0293.
* This **coefficient estimate** = the slope value of our independent variable.
* We can also find a confidence interval around any coefficient estimate (**confint()** in R)



* See the corresponding CI’s of slope for every coefficient in our regression model.
* We can say that we are 95% confidence that the true underlying population slope of investments predicting commitment is between these 2 values.
* Notice just like in a t-test, when these confidence interval values *don't capture zero*, we have a *significant difference in the test*.
* There's also something else that helps to tell our story: the **Standardized Beta coefficients,** whichallow us to talk about the coefficients *completely free from scale* (**lmBeta()** function in R)
* Remember that the slope value was on the scale of x and y 🡪 for a single unit change in x, then y changed by some amount, and *Standardized betas free us from scale.*
* Standardized scores for all coefficients in a model
* It's like calculating a z-score for the coefficients of every variable in the model, including the outcome variable.
* *It’s value is what the outcome variable moves by in Standard Deviation Units as the predictor changes by one whole Standard Deviation*



* See that this Standardized beta matched the Pearson correlation value.
* We have just 2 variables in this simple regression, so the t-values for significance match, the r and R2 values match, and Standardized beta matches the Pearson correlation value!
* *WHY?* 🡪 Because just like the Pearson correlation value, this Standardized Beta is free from scale.
* There's also an F-test for the *overall model fit*, which comes back to **residuals**.
* For any model, there are 3 sources of error.
* Total error = **sums of squared deviations** (the *actual* value of y - the mean of y)^2



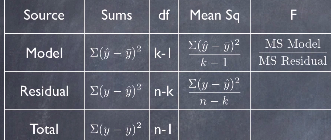
* if we don't know x, then best predictor of y is its mean.
* When we DO know there’s a relationship between x and y, we can define it w/ an intercept + slope + predict out outcome of y.
* So we effectively 'get better' at the prediction of y, which is captured w/ the **Model/Regression Error**, or the *predicted* value of y minus the mean of y



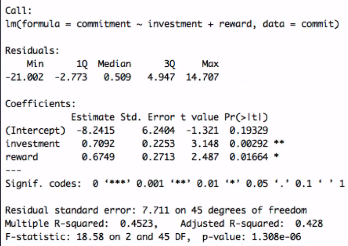
* Finally, there's **Residual Error** = how much is left completely unexplained by x (b/c we won't really ever get a perfect prediction).
* This found by calculating the **sums of squared residuals** = the actual value of y - the predicted value of y.



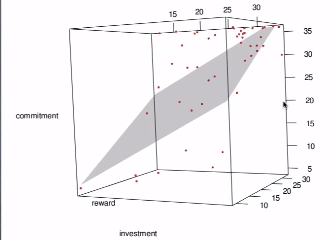
* We can put all 3 values into an ANOVA table + find an F-value evaluated w/ 2 sources of dF, which tells us if the *error saved by knowing the relationship between the IV + the outcome variable is more than the 'left over'/residual error.*



* If it IS, then we have a significant overall model.
* We are not limited to looking at just two quantitative variables.
* With **multiple regression**, we can examine the effects of multiple independent variables on a single quantitative outcome variables.
* With **simple regression** we use 1 variable to predict a single quantitative variable.
* In doing so, we expand the **Pearson correlation coefficient** by examining the **slope/impact** of the independent variable on the outcome variable.
* Can other variables also predict the commitment score amongst undergraduate students?
* Here's a new model predicting relationship commitment from 2 predictor variables of relationship investment + reward from relationship.



* Notice the slope for investment has changed from the simple regression model b/c now a new variable reward is in the model + it's pulling some of the effect of investment away from the commitment score.
* Investment is still significant, but now there's another player in the model.
* The effects of both variables are captured in the multiple regression model *simultaneously*
* If we ran 2 separate single regression models, the coefficient values we get from those model runs would not match the multiple regression b/c each variable is impacting the outcome alone
* The best way to see this is w/ a 3D graph.



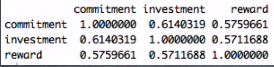
* Instead of a single line in 2D space, we have a plane of prediction in 3D space, effectively taking into account both independent variables simultaneously.
* Another way to think of this is w/ a Venn diagram
* For both independent variables using 2 simple regression models.



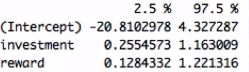
* Each area of overlap represents the proportion of variance accounted for in both models separately.
* But what happens when we run a multiple regression model?



* We effectively join both diagrams + the result allows us to see that the 2 independent variables in the multiple regression model are in fact related to one another.
* We can confirm this by looking at the **correlation matrix** of the variables in the model.



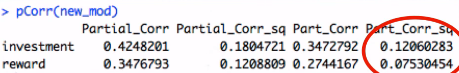
* Notice that there is in fact a relationship between investment + reward, + this relationship is the cause for the new multiple regression model having different slope values.
* Each coefficient of a multiple regression model is the impact of that particular variable on the outcome variable while *holding the other variables in the model* *constant.*
* Effectively these coefficients represent the *unique* impact of the variable that they represent.
* What's great about a multiple regression model is we can still find all of the things we found in a simple regression model:
* the t-test for the impact of the slope compared to 0
* the t-test significance of each of the slopes in the model
* a multiple R2 (which now represents the overall model variance accounted for)
* an F statistic that represents the overall model prediction.
* We can use the confint() function in R to again find the CI’s



* 95% confidence that the true underlying population slope of investments and of rewards predicting commitment is between these 2 values, respectively
* Can even use **lmBeta()** to get at the concept of a standardized beta for every coefficient in a model



* This is effectively freeing up every one from their scale.
* In addition, we can even talk about **unique proportion of variance** accounted for by each variable.
* Remember our Venn diagram.
* When we use **pCorr**, we can get the partial + semi-partial (Part) Correlation Coefficient squared value for every variable in the model.



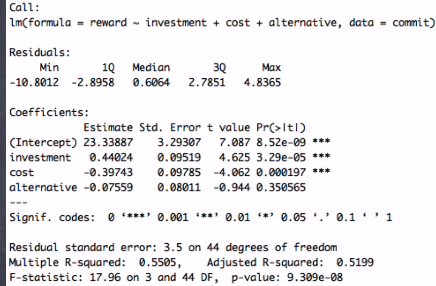
* This effectively models the concept of the unique sliver represented by this single coefficient in the model.
* We can evaluate the overall proportion of variance accounted for *uniquely* by each coefficient by examining which one has the biggest impact.
* When it comes to multiple regression, we've got to do a little bit of back ground checking to make sure that our model is okay.
* **Multicollinearity** 🡪 install **car** package + use the **Variance Inflation Factor** function, **vif()** on a model



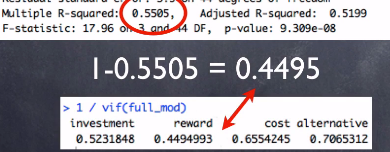
* See the variance inflation factor for each coefficient in the model.
* Take the reciprocal of the Variance Inflation Factor to get **tolerance**



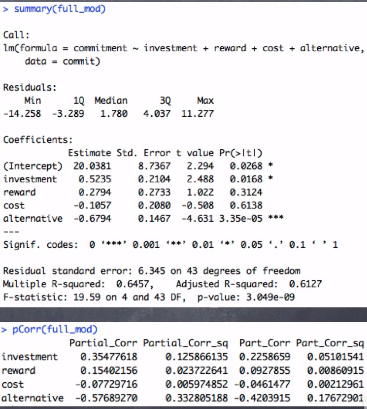
* Run another multiple regression model, this time predicting rewards from every other IV that was in our full multiple regression model (cost, investments, viable alternatives)

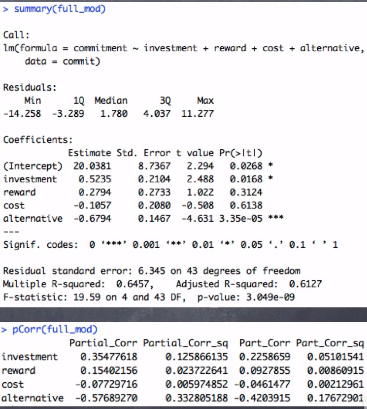


* *Multiple R2 value tells us how much rewards is accounted for in terms of its variance by the other independents.*
* If we look at the amount of *remaining* variance still left over/unexplained in the prediction of rewards, what do we get?

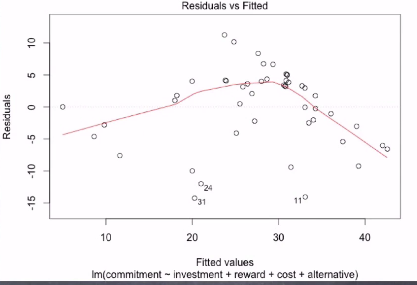


* We get **tolerance** = how much variance is left over in *this particular* independent variable once I know all of the other independents in the model”
* A really low amount of variance left over means that a particular IV is highly redundant w/ everything else in the model.
* Rules of Thumb
* Any IV w/ a Variance Inflation Factor VIF > five is kind of redundant.
* corresponds to a tolerance value < 0.2
* Effectively saying we’ve only got about 20% of the variance of this IV left over once I account for all of the other variables in the model.

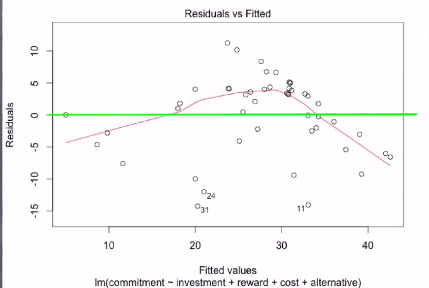




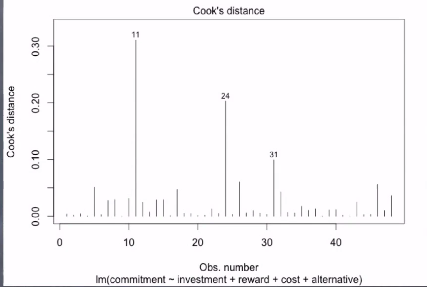
* In running our final multiple regression model, we've gotten a much fuller story to the prediction of commitment scores amongst undergraduate students.
* It's not the amount of investment into relationship that drives commitment, nor the number of rewards you get out of the relationship, nor the amount that the relationship is costing you that determines your commitment.
* The best predictor of commitment is, unfortunately, viable alternatives, which makes for a pretty interesting story.
* Once we're done running a regression model + we have a feel for the story that were telling, we're still not done examining the validity of our model.
* In order to do that, we need to look at some diagnostics for the model
* To determine the **validity** of a regression model, we have to look at 2 Basic graphs.
* Residual versus fitted plot
* This works with simple regression just as well w/ multiple regression
* We'll use it w/ multiple regression to get the most out of the model diagnostic procedure.

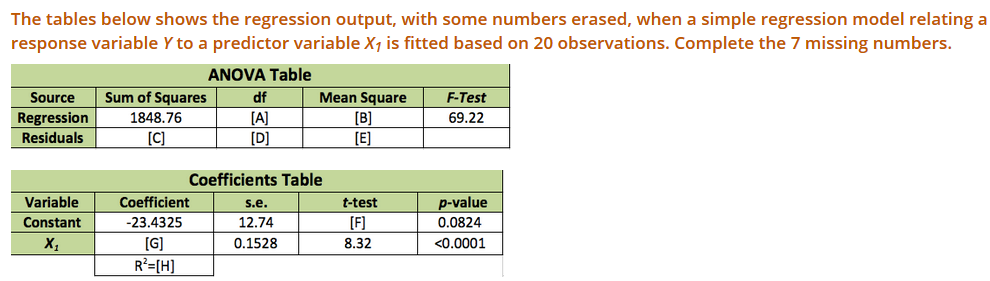


* This allows us to see whether or not there's a funneling effect to our residuals, down to a 0 value of residuals across all of the fitted values.
* Across the horizontal of this graph, there’s a 0 value for the residual score.
* It allows us to see the dispersion of residual values around that value of 0.



* A *good fitting* regression model will have a normal distribution of residuals w/ mean = zero, so we shouldn't see any odd outliers (very far away from 0 at any point in the graph)
* We should also see a random dispersion of the points on this graph + shouldn't see any funneling or any non-linear concepts.
* If we don't, we know we have **homoscedasticity** (our regression model fits the model effectively the same across all values of the fitted values of y = doesn’t get better at certain points vs. others)
* Can also see if we have **linearity**
* Both **homoscedasticity + linearity** are good things to have for regression
* Next diagnostic plot that's important to look = Cook's Distance Plot.



* In regression, **cook's distance** is a measure of influence of a single observation w/in the regression model.
* The higher the value = the more influence a particular observation plays in the overall model fit, as well as in particular concepts such as slopes
* As a cutoff of influence, we use the value **4 / dF** in the multiple regression model.
* So our cutoff of **Cook’s D** changes from model to model, depending upon the size of the data set used.
* We can ask for a Cook’s D plot for any regression model run, along w/ the cut off of 4/dF
* Doing so will allow us to see those observations that are highly influential.
* Using both of these methods we can determine if our regression model, be it simple or multiple regression, is a good model for our underlying data.
* We can make sure that there is linearity, homoscedasticity, and no influential outliers.
* 
* [A] = 1
* [B] 🡪 MS.reg = SS.reg / dF = 1848.76 / 1 = **1848.76**
* [E] 🡪 MS.res = MS.reg / F = 1848.76 /69.22 = **26.71**
* [D] 🡪 n – 2 = 20 – 2 = **18**
* [C] 🡪 SS.res = MS2 \* dF = 26.71 \* 18 = **480.78**
* [F] 🡪 t = Coeff / SE = -23.4325 / 12.74 = -**1.84**
* [G] 🡪 Coeff = t\*SE = 8.32 \* 0.1528 **= 1.27**
* [H] 🡪 R2 = SS.reg / SS.t = 1848.76 / (1848.76 + 480.78) 🡪 **0.7936**
* 