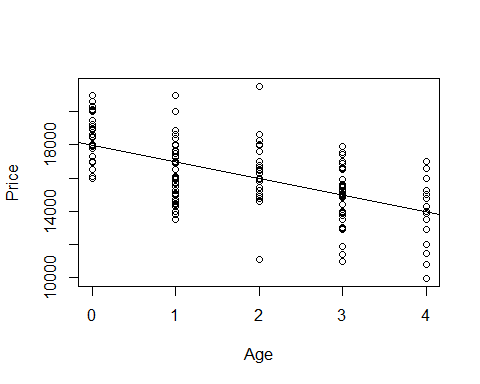
STOR 455 Class 11 R Assessing Multiple Linear Regression Models

library(readr)  
library(Stat2Data)  
  
data("Houses")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")  
  
# Or you can download the R script from Sakai  
# Save the script in the same folder as the notebook  
# comment out the above sourced script from github  
# Run code below  
#source("anova455.R)

# install if needed  
# dplyr package used for sample\_n() function  
  
library(dplyr)  
  
UsedCars <- read\_csv("UsedCars.csv")  
  
# selects same random sample each time  
set.seed(09132021)  
  
# random sample of 200 CamrySE  
CamrySE = sample\_n(subset(UsedCars, Model=="CamrySE"), 200)  
# Gets a random sample of a subset of the full dataset, where the model is camrySE, the second arguement is just how many it's taking

CamrySE$Age = 2017 - CamrySE$Year  
CamrySE\_Model = lm(Price~Age, data=CamrySE) # model  
plot(Price~Age, data=CamrySE) # Plot model   
abline(CamrySE\_Model)

 Look above; from here we have to think about the linear conditions - There is one situation were we need to look at a different way of the conditoins - When we have predictors wiht lots of values for the response for it ;we dont care about teh normalitiy to the residuals overall, we care about the normiality of the residuals for each response - In other words, for all the cars that are zero years old, are those residuals normally distirbuted for one, two, three and four are they all normally distributed? - Not a lot of years ar ehere, but we can try to see if we haev that normal distributio or are these some probailemns betwen the years

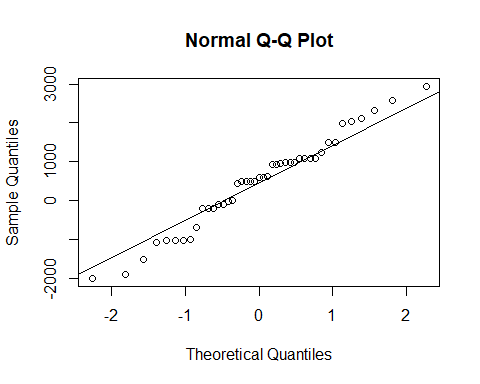
*How to lok at athat* - Can split it up by the ages of teh car

**Below:** - Looking at the camry’s by age

# Splits residuals by Age  
  
Resid\_by\_age = split(CamrySE\_Model$residuals, CamrySE$Age) # SUPER USEFUL FUNCTION  
Resid\_by\_age

## $`0`  
## 8 13 22 26 29 34   
## 1486.908341 -1022.091659 -204.091659 1095.908341 982.908341 -13.091659   
## 35 39 46 47 49 55   
## -1019.091659 1487.908341 2942.908341 591.908341 942.908341 2330.908341   
## 56 58 59 60 61 69   
## -1914.091659 937.908341 2043.908341 589.908341 -2008.091659 1983.908341   
## 71 75 80 85 88 89   
## -107.091659 987.908341 617.908341 2119.908341 1087.908341 -8.091659   
## 96 105 115 117 119 121   
## -204.091659 1095.908341 -1018.091659 430.908341 -1012.091659 495.908341   
## 123 128 129 137 143 149   
## 987.908341 495.908341 -1518.091659 -113.091659 1095.908341 924.908341   
## 150 159 165 171 175 193   
## -204.091659 2583.908341 -1077.091659 -704.091659 1237.908341 495.908341   
## 198   
## 495.908341   
##   
## $`1`  
## 2 5 16 18 24 25   
## -2036.580152 -1997.580152 -3506.580152 -1077.580152 -2722.580152 -1999.580152   
## 27 31 33 38 45 53   
## -891.580152 -26.580152 -1284.580152 1383.419848 257.419848 -942.580152   
## 57 62 63 65 67 68   
## -1100.580152 -3003.580152 -1514.580152 1868.419848 -2391.580152 -1342.580152   
## 77 78 79 81 83 84   
## -3175.580152 495.419848 -1493.580152 1384.419848 -992.580152 229.419848   
## 90 92 93 94 95 103   
## -3006.580152 593.419848 333.419848 -942.580152 -2609.580152 998.419848   
## 106 107 113 122 124 127   
## 1896.419848 3.419848 498.419848 293.419848 -1342.580152 673.419848   
## 130 135 138 139 140 141   
## -3210.580152 1003.419848 -2500.580152 -1996.580152 -96.580152 1007.419848   
## 144 148 152 155 157 161   
## -1091.580152 1612.419848 -492.580152 -2310.580152 1383.419848 -3006.580152   
## 164 167 170 172 174 180   
## -3506.580152 -2992.580152 508.419848 -1742.580152 -1891.580152 3003.419848   
## 185 190 191 196 199   
## -492.580152 2998.419848 599.419848 -1342.580152 4008.419848   
##   
## $`2`  
## 4 6 10 14 28 32   
## 2012.93135 2304.93135 2673.93135 530.93135 21.93135 -700.06865   
## 36 44 50 51 52 64   
## 997.93135 -976.06865 1643.93135 -976.06865 5548.93135 1003.93135   
## 70 74 76 97 99 101   
## 678.93135 477.93135 -67.06865 1014.93135 18.93135 1022.93135   
## 104 108 112 131 132 136   
## 323.93135 -189.06865 -549.06865 768.93135 11.93135 -978.06865   
## 142 145 146 156 160 162   
## -1094.06865 273.93135 2004.93135 1614.93135 1005.93135 -4898.06865   
## 166 169 188 189 192   
## -981.06865 -1377.06865 1018.93135 -1177.06865 2018.93135   
##   
## $`3`  
## 1 3 7 9 12 17   
## 289.442861 -1018.557139 -967.557139 -2060.557139 2448.442861 -960.557139   
## 19 20 21 23 37 43   
## -150.557139 -535.557139 30.442861 -521.557139 1939.442861 2019.442861   
## 48 54 66 72 73 86   
## 1534.442861 -1980.557139 260.442861 -1963.557139 37.442861 638.442861   
## 87 98 100 109 110 111   
## 1918.442861 539.442861 -1063.557139 499.442861 524.442861 1536.442861   
## 114 118 120 125 126 133   
## 651.442861 1530.442861 2939.442861 971.442861 -1970.557139 619.442861   
## 134 151 153 158 168 173   
## -3060.557139 -960.557139 -1461.557139 -3565.557139 -965.557139 4.442861   
## 176 177 178 179 181 182   
## 129.442861 1656.442861 -3975.557139 -1238.557139 -69.557139 2039.442861   
## 184 186 187 194 197 200   
## 2563.442861 2034.442861 1925.442861 36.442861 939.442861 1590.442861   
##   
## $`4`  
## 11 15 30 40 41 42   
## -450.04563 349.95437 50.95437 -3152.04563 1049.95437 2049.95437   
## 82 91 102 116 147 154   
## -1950.04563 -3988.04563 2649.95437 -2458.04563 3045.95437 849.95437   
## 163 183 195   
## -1057.04563 -70.04563 1304.95437

# Different lists, of the string 0; all residuals for 2017 models   
# Then 1, 2, and 3, etc old cars   
# Year year each year gets its own string   
  
# If want to look at normaility, then look at the qqnorm plot of it   
# Need to do for each thing individually   
  
# This is for the zero year old cars   
qqnorm(Resid\_by\_age$'0')  
qqline(Resid\_by\_age$'0')



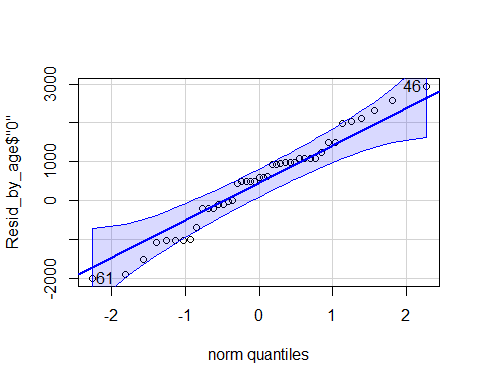
# install if needed  
# car package used for qqPlot()  
  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

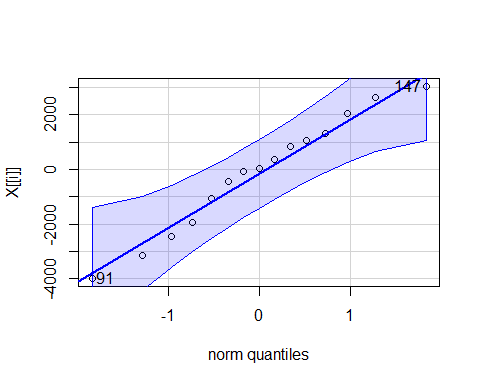
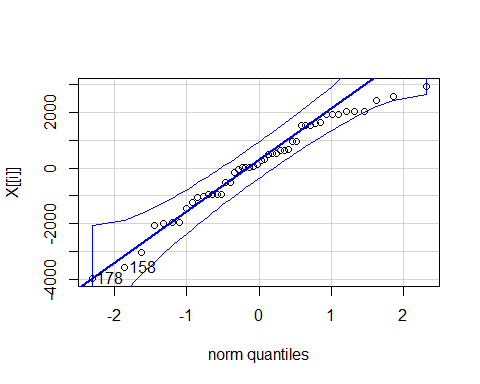
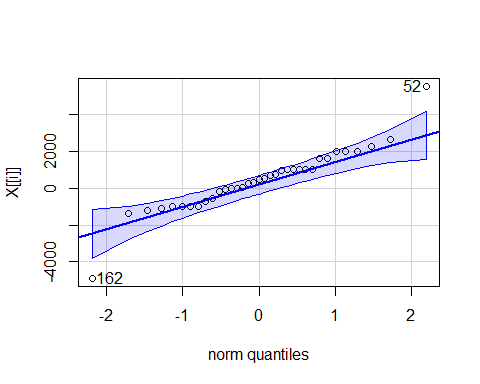
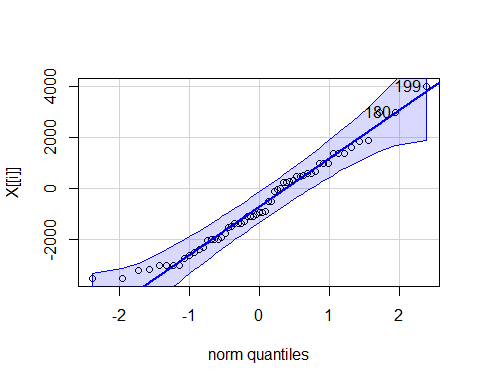
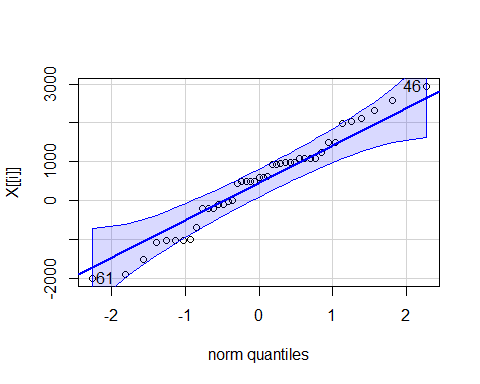
qqPlot(Resid\_by\_age$'0') # THIS IS A SUPER USEFUL FUNCTION



## 61 46   
## 17 9

# Shows teh QQNorm and QQLine plot if norma distributioned   
# Gives a band of what we would expect to see if this data of this size was normally distributed   
# The varibaility we would expect from a thing of this size   
# It looks pretty good in this interval   
# Want to use this for all the ages of the cars

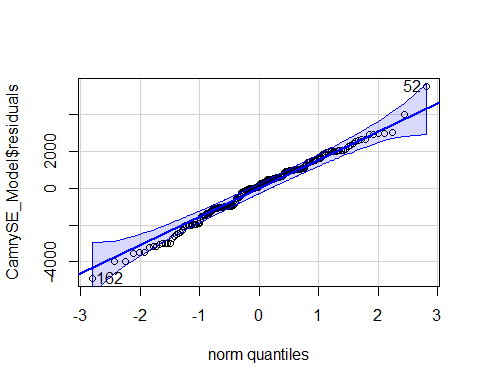
sapply(Resid\_by\_age, qqPlot)



## 0 1 2 3 4  
## 61 17 59 30 39 8  
## 46 9 54 11 34 11

# Applies the qqPlot function to all the residual\_by\_age   
# This shows us that it's normal enough for each age, so it's good to go

qqPlot(CamrySE\_Model$residuals)



## [1] 52 162

# Narrow band because bigger dataset   
# Age 4 is wider ebcause there is less data   
# Less data = wider band, more data = smaller band (Same idea as confidence intervals)

**Three different types of tests** **T-test for Slope** - Ho: B1 =0 - Ha: B2 != 0

* t = b1/SEofb1
* Compare to t(subof n-2)
* Do we have evidence to say the relationship bt the predict and resposne has a non horizontal relationship/non zero slope?

**ANOVA for regression** - Ho: B0 = 0 - Ha: B1 != 0

* F = MSModel/MSE
* Compare to F1, n-2
* (tn-2)^2 = F1, n-2
* For simple linear models, did the same as the t test; tried to look at ho wmuch variability is being explained by odel compared to a null model (so a horizontal line)

**T-test for correlation** - Ho: p = 0 - Ha: p != 0

* t = ((r\*sqrt(n-2))/sqrt(1-r^2))
* Compare to tn-2
* Do we have evidence to say there is some kind of correlation, +/- for the two variables?

**Simple Linear Regression Model** - Y = B0 + B1X + Error - Where Error follows N(0,stdof error) and independent (normal and independent) *What if we have more than one potential predictor?* - We got teh same values if we did the above three different tests on a simple linear regression model, so its not super helpful when you’re doing a simple linear regression model

**Multiple Regression Model** - Y = B0 +B1X1 + B2X2 +….+BkXk + Error - where error is assumed to follow a normal and independent - THis is in many dimentions with more predictors

Data?  
We need n data cases, each with values for Y and all of the predictors X1,…,Xk.

**R - Correlation Matrix**

head(Houses)

## Price Size Lot  
## 1 212000 4148 25264  
## 2 230000 2501 11891  
## 3 339000 4374 25351  
## 4 289000 2398 22215  
## 5 160000 2536 9234  
## 6 85000 2368 13329

# Just a small dataset of 20 houses   
  
cor(Houses)

## Price Size Lot  
## Price 1.0000000 0.6848219 0.7157072  
## Size 0.6848219 1.0000000 0.7668722  
## Lot 0.7157072 0.7668722 1.0000000

# We can do this because they are all quanatitive variables   
# all have fairly strong, postiive correlations of eachother   
# Are tehse big enough to take this claim tot he population?

*Look at a test between them*

**t-test for Correlation**

#cor.test(Houses) # This doesnt work because there's too many   
  
# Correlation looks at 2 vars once and looks at the relationship bt two vars   
  
cor.test(Houses$Size, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Price  
## t = 3.9871, df = 18, p-value = 0.0008643  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3476582 0.8651583  
## sample estimates:  
## cor   
## 0.6848219

# Ho: No correlation between teh two in teh population   
# Ha: There is a correlation between teh two in the population   
# Want to see how likely it is that we would get the smapel line in the population   
# Output tells us the correlation bt the two, the t test of 3.9 (this tells us that it's pretty unlikely by chance); pvalue of 0.000008 - that is the probability that we would get a sample like this or one as extreme as this if there was no realtionship in teh popualtion; so this is pretty unlikely this would happen by chance   
  
# Can do the same thing above for the other realtionships as well   
cor.test(Houses$Lot, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Lot and Houses$Price  
## t = 4.3478, df = 18, p-value = 0.0003878  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3998134 0.8796343  
## sample estimates:  
## cor   
## 0.7157072

cor.test(Houses$Size,Houses$Lot)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Lot  
## t = 5.0694, df = 18, p-value = 7.991e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4909631 0.9029654  
## sample estimates:  
## cor   
## 0.7668722

# BOth have teh same conclusion and hypothesis that   
# Ho: No relationshio in the popualtion   
# Ha: Some realtionship in population   
# Both have small pvalues   
# Both have strong evidence to say there is a realtionship between these things in teh population

**t-test for Correlation** - Ho: p = 0 - Ha: p != 0

* t = ((r\*sqrt(n-2))/sqrt(1-r^2))
* Find p-value with t n-2

Use this to: 1. Identify potential good predictors of Y. 2. Look for relationships among predictors.

**Prediction Equation** - where the coefficients are chosen to minimize: SSE = sum(y-yhat)^2

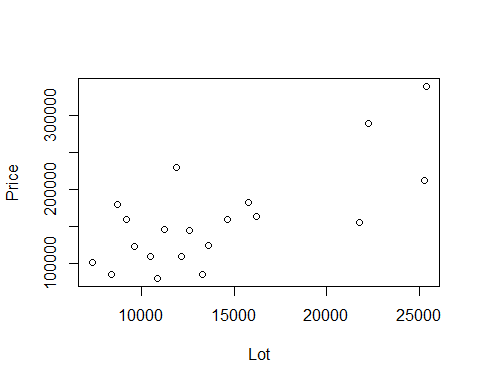
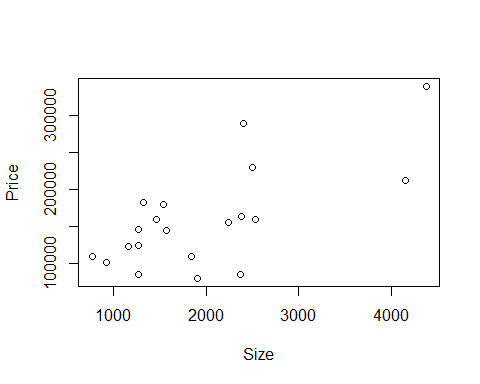
To fit a multiple regression in R: model=lm(y~pred1+pred2+pred3,data= ) - Other tests keep in mind the model that we are workign with, the correlation test always lookas only at 2, while the other tests look at more than 2 variables at once

**R Regression: Individual T-tests** - Look at the P value of the predictors (Size and lot in this example)

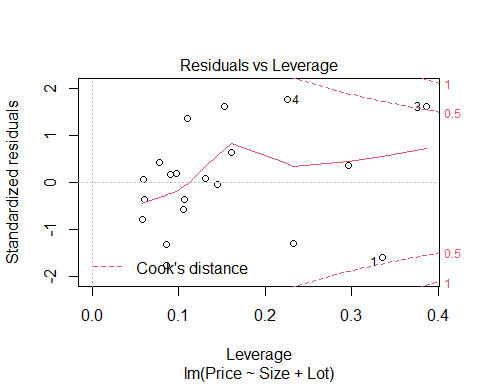
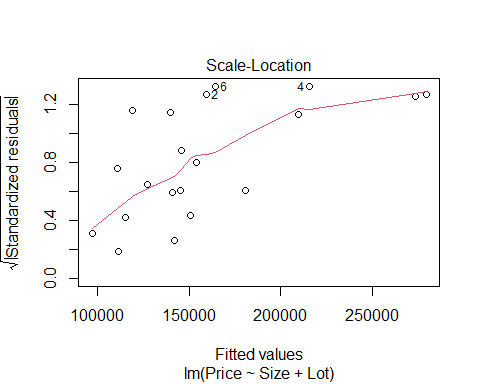
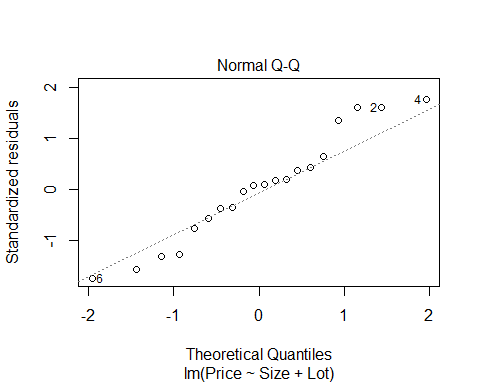
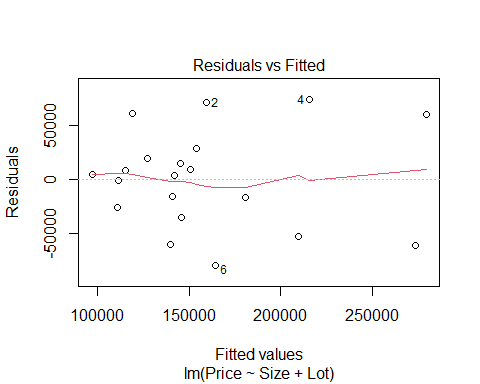
HouseModel=lm(Price~Size+Lot,data=Houses) # Multiple linear regression model   
summary(HouseModel)

##   
## Call:  
## lm(formula = Price ~ Size + Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34121.649 29716.458 1.148 0.2668   
## Size 23.232 17.700 1.313 0.2068   
## Lot 5.657 3.075 1.839 0.0834 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47400 on 17 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.505   
## F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

# The summary output:   
# The test we looked at before for individual slopes are still there, just an extra line is there for teh extra predictor   
# The resuls are not what we would expect; where invidually when we compare size with teh price of the house and the lot with t price of the house, we had small pvalues saying we would have a really low chance of getting this sample if there was no realtion   
# Same tests here that the coeff of size is vs teh alternatieve it's not, we get pvalues athat are higher and possibily not sign at this rate   
# There is something goeing on here, which is called multicolinearity (See next class's notes)  
# The bottome line gives the anova tests, does something different   
  
plot(Price~Size+Lot,data=Houses)



# It's not giving us one plot we can look at and think about lineariry and constant variance to the full modle   
# Its giving us two different plots, because this plot would be given in 3 dimentions   
# We have aplot by size by price and lot size by price   
# Hard to see the realtionship here with this visual format   
# We know teh realtionship with each redcitor in teh response could be useful when looking at t transformations in multiple regression   
  
plot(HouseModel)



# We see that it's not all conditions are really met   
# Same idea as simple lienar regression because for simple linear regression, just compare residuals to the fitted values or looking at teh normaility of the residuals   
# Here, no matter how many predictors we have, we still haev how far off are we and what are the residuals, so we can still look at the residuals and fitted plots   
# Does teh red line look roughly horizontal or some defined curve? IT's a small dataset, so we expect some kind of variability, that's looks pretty good. there's not some clear curve over the whole model   
# Normaility, can judge the same as simple lienar regression   
# the residuals by fitted, we can still look at this roughly the same, but leverage is a little different now  
# For simple linear models it was how far are we fro teh mean predictors, for mutkpel, we could be far from some predictors but close to others  
#Cook's distance, look at it the same as simple linear models outside bounds, teh have influence

We are saying that a price in house is about 34121.649 + 23.23size + 5.657lot **R - Multiple Regression** (𝑃𝑟𝑖𝑐𝑒) ̂=34121.6+23.232∙𝑆𝑖𝑧𝑒+5.657∙𝐿𝑜𝑡 - We are summing each test has a slope of zero - the t test stat = how many SE we are from zero - about the same, but different dfs

**Multicolinearity** is when size is a good predictor of lot already, so when we look at things in the future; if lots of predictors are correlated with eachother, then they may be explaineng a similar amount of variability

**T-test for Slope** -Note: We now have multiple “slopes” to test - Ho: B1 = 0 - Ha: Bi != 0

* t.s. = Bi/SEofBi
* All given in R with a p-value
* Compare to t n-k-1
* **lose 1 d.f. for each coefficient**
* Reject Ho if The ith predictor is useful in this model

**Coefficient of Multiple Determination** - 𝑅^2=𝑆𝑆𝑀𝑜𝑑𝑒𝑙/𝑆𝑆𝑇𝑜𝑡𝑎𝑙 - Now interpreted as the % of variability in the response variable (Y) that is “explained” by a linear combination of these predictors. *NOTes* - Variability in response being explained by the predictior - for Simple linear regression this was just correlation squared - now ti’s different because we haev 2 predictorsl its not just looing at each predictor and the response its lookig at the variability in teh resposne based on all teh predictors in the model - may explain overlaping responses in the mdoel

* Look at teh adjustesd R squared as 0.55, it says 55% of the variability in the price of the house is explained by the size of the house and the lot size of the house; that leaves an extray 44% that is not explained by these things by the data we have; so there may be other variables that explain that varibaility and we just don thaev those variablies

**t-test for Correlation vs. t-test for Slope** - **t-test for correlation:** Assesses the linear association between two variables by themselves. – in a vaccum; compare two things, ignore world - **t-test for slope:** Assesses the linear association after accounting for the other predictors in the model. – accounts for other predicotrs

**Partitioning Variability** - Y = B0+B1X1+…+BkXk + Error - SSTotal = SSModel + SSE - SSModel = Total explained by the regression - SSE = Error after regression - SSTotal = Total variability in Y

* About the same thing, but just in different dimentions
* we have amodel that predicts that data dn teh error around that model that we haev
* can look at teh SS the same as before and this case, we jsut condense teh SSModel not just from 1 predicotr, but from multiple predictors we are looking at
* ANOVA wise, the idea of accounting for that avariability is about the same

**ANOVA test for Overall Fit** - Ho: B1 = B2 = …= Bk = 0 (weak model) - Ha: Some Bi != 0 (Effective model)

Source, d.f, Sum of Squares (SS), MeanAquare, t.s., P-value Model, k, SSModel, SSModelk, MSModel/MSE, Fk, n-k-1 Residual, n-k-1, SSE, SSE/(n-k-1), t.s. is the same, p value is the same Total, n-1, SSTotal

* Only difference than before is that there can be multiple predictors
* Still haev the other things, just accounts for other htings in the model
* still trying to figure out how good is the mdoel doing
* WE ARE ASSUMING NOTHING IN THE MODEL IS USEFUL TO US
* We are assuming that the coeffe of the predictors are zero **ANOVA Hypo Test**
* Ho: There is no point of using the model; all coeffs are zero
* Ha: Assume at least one is non-zero
* Use to see if the predictors are useful
* will talk more about this when it comes to errors later
* if we haev 10 preictors in the mode, we dont wat to test each of teh predicotrs individuals because tehres agood chance we will amek an error an dhow a type 1 error
* if we do any overal test, it will give us an way to see if anything is useful to us
* good to use for nestsed tests
* comapres the mdoel with all predictors to a null modle where teh coeff are both 0
* the same process can be used to compare nested models together
* think: Model with price by size and lot, compared to a model of just price by size; does adding this extra variabile explain a sig amount more of the variability?

**R - Regression ANOVA** - **Important note:** R shows a “sequential” sum of squares in the ANOVA table, i.e. how much new variability is explained as each predictor is added. Add the components to find the SSModel.

anova(HouseModel)

## Analysis of Variance Table  
##   
## Response: Price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Size 1 4.0447e+10 4.0447e+10 18.0018 0.0005485 \*\*\*  
## Lot 1 7.6013e+09 7.6013e+09 3.3831 0.0833990 .   
## Residuals 17 3.8196e+10 2.2468e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# ANOVA by default will looka at ta sequnetial SS  
# This iwll loo at the first line of the model that has price predicted by size and it says is teh slope of that mdoel nonzero? we have pvalue of 0.0005,   
# Next line says compares a model with asize and lot in it to a model of just size; if we add lot to our odel, does it explain more varibaility? In this case, no.   
# Use ANOVA455 if you DONT want it to look at it sequentially

**A “Local” ANOVA Function** - To find ANOVA for a multiple regression model that is NOT split sequentially for each predictor…

anova455(HouseModel)

## ANOVA Table  
## Model: Price ~ Size + Lot   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 2 4.8048e+10 2.4024e+10 10.693 0.000985 \*\*\*  
## Error 17 3.8196e+10 2.2468e+09   
## Total 19 8.6244e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# It condesnes the model and looks at it instead of line by linea nd adding it one at a time, its looking at teh overal model   
# It condesnes all teh SS together rather than the SS of each individual predictor   
  
# We see that its going in twith teh assumption that the coeffes for lot and size are 0 and ttrying to see if we have evidence to say at least one of them is non zero   
# Have as amll pvalue that at least one is nonzero, but it seems that it contradicts what we have before   
# Where we lookeda t teh summary, then both the pvalues were big, so it's a bit contradictory because of teh multiocolinearity that we talk about next class

**Example: Houses** 1. Test #1: Compute and test the correlation between Size and Lot in Houses

cor.test(HousesLot) t = 5.0694, df = 18, p-value = 7.991e-05

1. Test #2: Compute and test the coefficient of Size in a multiple regression model (along with Lot) to predict Price. (Estimate Std Error t value Pr(>|t|) Intercept) 34121.649 29716.458 1.148 0.2668 Size 23.232 17.700 1.313 0.2068 Lot 5.657 3.075 1.839 0.0834

F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985