STOR 455 Class 9 Partitioning Variability - ANOVA

library(readr)  
library(Stat2Data)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
Domestic = subset(DistanceHome, Distance<250)

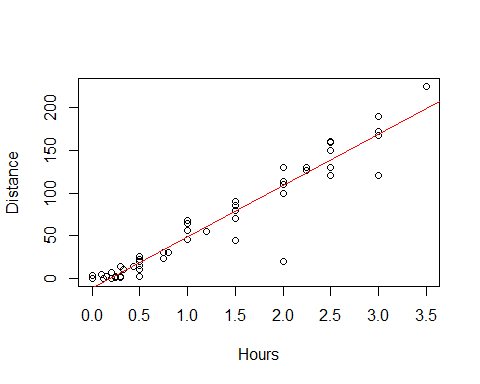
*NOtes* - Evidence to say if there is some kind of relationship exists? - How we center teh distributios is a little different

**T-test for Slope**  Ho: B1 = 0 (for Y = Bo + Error) Ha: B1 != 0 (for Y = Bo + B1X + Error)

t.s. = Bhat1/SEofBhat1 - Find p-value using a t-distribution and n-2 d.f. -p-value is small then Reject Ho

**How to find a P-value?** - Statistically significant evidence suggests ( p-value < 2.22\*10-16) that there is a relationship between the hours that a student spend travelling to campus and their distance from campus.

plot(Distance~Hours, data=Domestic) # Predict distance home, based onhow many hours it takes to get there   
moddist = lm(Distance~Hours, data=Domestic)  
abline(moddist, col="red") # Draw linear model on top of plot



summary(moddist) # Take summary of plot model

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

# In coef table,   
# Does a hypothesis test for us   
# COmpares two models to eachother   
# The linear model to the constant model, where the model is just the mean   
# If the model is just a horizontal line, the it's that the slope would be zero  
# slope zero = no cheged based on distance from home   
# Alternative: AS hours change, the ditsnce from home changes   
# How likely is it that we get a sampel slope like we did, what the chance we would get what we did   
# The test stat is the tstat/Se = how many std err are we from teh null; how unlikely is it that we get the thing that we did   
  
# Slope: 59.99  
# STd err: 2.48 = Tightly bound   
# tvalue = 24. = if our null is 0, then we are 24 std errs from the null hyp of 0; thats unlikely to happen by chance   
# Pvalue: How unlikely we are to get a sample that we did, if the null is true; and we have a really low pvalue for this, which means that it is not likely we would get the null hypothesis

**Finding Correlation in R** -For data in two variables:

cor(Domestic$Distance, Domestic$Hours)

## [1] 0.9581758

**Finding Correlation in R** For all variables in a dataframe:

data(Houses)  
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

# Can put in whole dataframes   
# This will give you a corr matrix; a table with a ll possible combination snad the correlations for those variables   
  
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

**Finding Correlation in R** Watch out – variables must be numeric! May need to choose numeric columns:

data(Cereal)  
head(Cereal)

## Cereal Calories Sugar Fiber  
## 1 Common Sense Oat Bran 100 6 3  
## 2 Product 19 100 3 1  
## 3 All Bran Xtra Fiber 50 0 14  
## 4 Just Right 140 9 2  
## 5 Original Oat Bran 70 5 10  
## 6 Heartwise 90 5 6

#cor(Cereal) <- this doesn't work because it has character vectors in it   
  
cor(Cereal[c(2:4)]) # Tells R Which columns we want to take from the thing

## Calories Sugar Fiber  
## Calories 1.0000000 0.5154008 -0.7150123  
## Sugar 0.5154008 1.0000000 -0.5025772  
## Fiber -0.7150123 -0.5025772 1.0000000

**Test for a Linear Relationship via Correlation** - Let p (rho) denote the population correlation

Ho: p = 0 Ha: p =!= 0

t = ((r\*sqrt(n-2))/sqrt(1-r^2)) - Compare to a t-distribution with n-2 d.f.

**Correlation t-test in R** *SEe below:*

cor.test(Domestic$Distance, Domestic$Hours) # If null is true (if 0 correlation) how likely is it that we have a sample with this strong correlation? That's what this will tell you

##   
## Pearson's product-moment correlation  
##   
## data: Domestic$Distance and Domestic$Hours  
## t = 24.144, df = 52, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9286787 0.9756277  
## sample estimates:  
## cor   
## 0.9581758

# OUtputs: Correlation bt the two things   
# What the cof int is for the population based here   
# t = 24.144 =how many stand dev from the null we are (Same value as before for test for slope)  
# Tells us really low pvalue, which means that its ar really low chance we could get this result by chance   
  
# All the results are going to be the same if it's simple linear regression   
# Gives us F test stat and p value

**ANOVA for Regression** Data = Model + ERror Total variation in response, Y = variation explained by MODEL + Unexplained variation in RESIDUALS

Key question: Does the MODEL explain a “significant” amount of the TOTAL variability?

**Partitioning Variability - SLM** Y = Bo + B1X + E SSTotal = SSModel +SSE

**NOtes** - ANOVA = analaysis of variances - we cant explain all variability, but we can try -*Total variations in reponse:* Is how far away each of our points are from the mean value - *Variation explained by model* total variability = full distance from teh mean down to the point; by fitting teh line to it, most of the variability is explained by the variation ; - *Resisudals* Distance that is left over from teh points - WE want a sig amount of variability in the model to be explained by the regression line vs a horizontal line - We really only care about teh sum of squares - The f test stat = how likely this variability is explained by model compared to what is left over in the error term if the null model were true that there is no relationship between tehse things **Can look at this by using ANOVA**

anova(moddist)

## Analysis of Variance Table  
##   
## Response: Distance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Hours 1 194417 194417 582.93 < 2.2e-16 \*\*\*  
## Residuals 52 17343 334   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# We have teh SSModel and SSError terms; does not give SSTotal, but if we add teh two, we could get that   
# What it is is that the SS in the HOurs row, is for the model, so if we had the line going back; how far away for each point the line is for that value away from teh mean of that value; how much variability is explained by the model and all squared and summed together   
# Residual SS; looks at the similar thing, but how far each point is from the line squared and summing them totgether   
  
# Think about that proportionally; how different are these two? Scale them by df.  
# MSE = comparing models for the vriabiilty and i'm taking the SSTotal/df. Hours df = 1, and residual df = 52 (which is 54-2)  
# F test stat = MSModel/MSError; that value is big it says that a lot of variability is explained by the model; when that value is small, it's saying not so much varibility is being explained by the model   
  
# Big and small are relative, depends on sample size   
# F test stat says its unlikely we would get this result by chance

**ANOVA Test for Regression** - Basic idea: Find two estimators of 2 - Model: SSModel/1 = MSModel ERror = variance of error = SSE/n-2 = MSE - We want to compare the model and the error

t.s. = MSModel/MSE - COmpare to F-dist with I and n-2 d.f

* ANOVA Test for regression Ho: B1 = 0 (ybar) Ha: B1 != 0 (yhat)
* Same null and alternative as slope test; asusming there is no relationship bt predictor and response
* so the slope of the model is 0 and the mean = the model that we use
* Alternative = some nonzero slope better descrbes this model and how likely we would get this kind of model if the null were true

Source, d.f, Sum of Squares, Mean Swuare, F, Pvalue Model, 1, SSModel, SSModel/1, MSModel/MSE, F(1,n-2) REsidual, n-2, SSE, SSE/n-2, See above got F and PValue TOtal, n-1, SSTOtal - in R use 1-pf(Fstat,1,n-2)

**What is r2?** r2 = proportion of total variability in the response (Y) that is “explained” by the model. 𝑟^2=𝑆𝑆𝑀𝑜𝑑𝑒𝑙/𝑆𝑆𝑇𝑜𝑡𝑎𝑙=1−𝑆𝑆𝐸/𝑆𝑆𝑇𝑜𝑡𝑎𝑙 - The amount of variability explaiend by the model out of the total variability - If we look at a plot ; the total variability (how far away each point is away from teh mean line), then our error term; varibaility explained by model = mean down tot the line - high values of rsquares = most of the variability in teh response is explained by the predictors a - low values = oppsitie

**Visualizing r2 for a SLM** Basic Idea: How much “better” does the least squares line do than a “prediction” that doesn’t use X at all? - Using NO predictor: 𝑦 ̂=𝑦 ̄

* Least Squares Line: 𝑦 ̂=𝛽 ̂\_𝑜+𝛽 ̂\_1 𝑥
* the cor coof being squared
* if corr coef is always bt -1 and 1; r^2 has to be bt 0 and 1
* big difference; r-sqqare we can look at multipel predictors at once

**Why is it called r2?** - Def: The correlation, r, measures the strength of linear association between two quantitative variables. -1< r <1 to 0 < r2 < 1 - 0 = Explains no variability - 1 = Explains all variability **Simple Linear Regression - R**

summary(moddist)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

**Which test is best?** -**T-test for slope:** Ho: B1=0 Ha: B1!=0

𝑡=𝑏\_1/(𝑆𝐸\_(𝑏\_1 ) )

Compare to t n-2 - How well related is this predictor with this response after taking into account the whole rest of the model, all other predicotrs - Not vacume - look at all in one room

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

𝐹=𝑀𝑆𝑀𝑜𝑑𝑒𝑙/𝑀𝑆𝐸

Compare to F1,n-2 - Bigger picutre - Have 1 predictor, but once we have more, it will let us see if there is any relation bt any of these predictors and repsonse here - one test to see if there is any relation anywhere - useful for more tests

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

𝑡=(𝑟√(𝑛−2))/√(1−𝑟^2 )

Compare to t n-2 (tn-2)2 = F1,n-2

* Focuse on a predictor ans response
* how are they related in a vaccume ingnoring everyhting else

**We have 3 different tests for when we get to multiple regression**