STOR 455 - Class 19 – Testing a subset of predictors

library(readr)  
library(leaps)  
  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

\*Is there a sig dif bt a model with these extra predictors compared to something smaller? **Comparing Two Regression Lines (with a multiple regression)** - dhould reg be considered the same? - The interaction terms - can interact slope and intercept depending on values

**Multiple regression model** - We had anova, but is there someone inbetween? - That’s th enested test! - Instead of comparing to anull, we compare to a subset of the model - and that subset is the base point - ANOVA looks at how much more is explained to a horizontal line - NOw a nested test is comparing the model to a different model

**Nested Models** - Definition: If all of the predictors in Model A are also in a bigger Model B, we say that Model A is nested in Model B. - Example: 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+ 𝜀 is nested in - 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+\_3 𝑅𝑒𝑠𝑡∗𝑆𝑒𝑥+𝜀 - Test for Nested Models: - Do we really need the extra terms in Model B? - i.e. How much do they “add” to Model A?

**Nested F-test** - Want to see how much variability is explained by adding these new values Basic idea: 1. Find how much “extra” variability is explained when the “new” terms being tested are added. 2. Divide by the number of new terms to get a mean square for the new part of the model. 3. Divide this mean square by the MSE for the “full” model to get an F-statistic. 4. Compare to an F-distribution to find a p-value.

**Nested F-test** Test: Ho: Bi=0 for a “subset” of predictors Ha: Bi != 0 for some predictors in the subset - F = ((SSModelFull - SSModelReduced)/# Predictors)/MSEFull - F = ((Explained by Full model - Explained by reduced model)/predictors tested in Ho)/ based on full model - Compared to a f distribution

**Nested F-test** 𝐴𝑐𝑡𝑖𝑣𝑒 =𝛽\_0+𝛽1𝑅𝑒𝑠𝑡+B\_2 𝑆𝑒𝑥+ B3𝑅𝑒𝑠𝑡𝑆𝑒𝑥 +𝜀 H0: β2=β3=0 Ha: Some βi≠0 -Compare mean square for the “extra” variability to the mean square error for the full model.

**Nested F-test Code Example**

modelPint=lm(Active~Rest+Sex+Rest\*Sex, data=Pulse) # Total model;   
# Predict active heart rate by resting rate, sex and the interaction bt rest and sex   
# including the interaction term makes sure that we don't assume that rest and sex have the same slope and intercept   
summary(modelPint)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex + Rest \* Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.251 -2.893 6.784 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.43987 7.47902 1.262 0.208   
## Rest 1.14319 0.11264 10.149 <2e-16 \*\*\*  
## Sex -0.28717 10.22830 -0.028 0.978   
## Rest:Sex 0.03907 0.15130 0.258 0.796   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.17 on 371 degrees of freedom  
## Multiple R-squared: 0.4056, Adjusted R-squared: 0.4008   
## F-statistic: 84.37 on 3 and 371 DF, p-value: < 2.2e-16

# If we want to test to see if adding sex to see if the slope adn itnercept are different, we want to comapre to one without sex and the interaction   
modelP\_Reduced = lm(Active~Rest, data=Pulse)  
  
# This compares the two models to tell us if the interaction and sex term are significant in our model   
anova(modelP\_Reduced, modelPint)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + Sex + Rest \* Sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74538 2 512.14 1.2746 0.2808

# How much extra variability is expalined? Its the difference int eh sum of squares; if that's a big differene, a higher SSqures is better? Yes

# This tells us, individually, if the predictors are significant in our model  
anova455(modelPint)

## ANOVA Table  
## Model: Active ~ Rest + Sex + Rest \* Sex   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 3 50854 16951.5 84.373 < 2.2e-16 \*\*\*  
## Error 371 74538 200.9   
## Total 374 125392   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova455(modelP\_Reduced)

## ANOVA Table  
## Model: Active ~ Rest   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 1 50342 50342 250.2 < 2.2e-16 \*\*\*  
## Error 373 75050 201   
## Total 374 125392   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The resting term has a sig relationship   
# The itneraciton model, its at 50854 and the other model is 50342  
# Subbing these gives us the additional variability exampled;  
  
# That's SSDif that is below

SS\_diff = anova455(modelPint)[1,2] - anova455(modelP\_Reduced)[1,2]  
SS\_diff # The additional variability explained

## [1] 512.1413

# that's where the 512 is coming from in teh table above, the difference in teh sum of squares  
  
MS\_diff = SS\_diff/(anova455(modelPint)[1,1] - anova455(modelP\_Reduced)[1,1])  
MS\_diff # Means squared difference

## [1] 256.0706

# Divide SS\_dif by the difference in predictors of the model   
# WE want to see what the differences are in teh df   
# 3 - 1 = 2   
  
F\_diff = MS\_diff/anova455(modelPint)[2,3]  
F\_diff # The F value difference

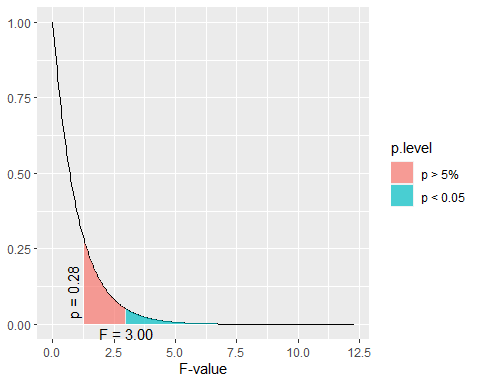
## [1] 1.274553

library(sjPlot)

## Warning: package 'sjPlot' was built under R version 4.1.2

## #refugeeswelcome

dist\_f(f = F\_diff,   
 deg.f1 = anova455(modelPint)[1,1] - anova455(modelP\_Reduced)[1,1],   
 deg.f2 = anova455(modelPint)[2,2],  
 )



# The area under the curve is a pvalue   
# Plots teh f distribution to see graphically how extreme it is   
# We need to tell it the difference of the predictors and the df of the error term (that's what the deg.f1 and f2 are)  
# This graph will vary depending onthe degrees of freedom   
# WE see that the 1.28 is around the p value of 0.28;   
# we would expect ot seet his variation about 28% of the time if there was no useful ness of adding things into the model   
# WE need an f test stat up to 3 to show sig results   
# This tells us that its not beneficial to add these terms to our model bcuase we don't see a stat sig dif bet the two models (using sex to predcit active heart rate)

**Example: State SAT Scores** Source: Statistical Sleuth, Case 12.1 pg. 339  
Response Variable:  
SAT =Average combined SAT Score Potential Predictors:  
Takers = % taking the exam Income = median family income ($100’s) Years = avg. years of study (SS, NS, HU) Public = % public school Expend = spend per student ($100’s) Rank = median class rank of takers

SATModel = lm(SAT~., data=StateSAT[,2:8])  
summary(SATModel)

##   
## Call:  
## lm(formula = SAT ~ ., data = StateSAT[, 2:8])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.046 -6.768 0.972 13.947 46.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -94.659109 211.509584 -0.448 0.656731   
## Takers -0.480080 0.693711 -0.692 0.492628   
## Income -0.008195 0.152358 -0.054 0.957353   
## Years 22.610082 6.314577 3.581 0.000866 \*\*\*  
## Public -0.464152 0.579104 -0.802 0.427249   
## Expend 2.212005 0.845972 2.615 0.012263 \*   
## Rank 8.476217 2.107807 4.021 0.000230 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.34 on 43 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618   
## F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# IF we think about polynomial regression we can make a good model with it   
# We have a few good predicotrs here

**R: Best Subsets for StateSAT**

all = regsubsets(SAT~., data=StateSAT[,2:8])  
ShowSubsets(all)

## Takers Income Years Public Expend Rank Rsq adjRsq Cp  
## 1 ( 1 ) \* 77.42 76.95 34.03  
## 2 ( 1 ) \* \* 84.71 84.05 10.22  
## 3 ( 1 ) \* \* \* 87.11 86.27 3.69  
## 4 ( 1 ) \* \* \* \* 87.71 86.61 3.58  
## 5 ( 1 ) \* \* \* \* \* 87.87 86.49 5.00  
## 6 ( 1 ) \* \* \* \* \* \* 87.87 86.18 7.00

# Tells us the best model is

SATModel1 = lm(SAT ~ Years + Expend + Rank, data = StateSAT)  
SATModel2 = lm(SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
summary(SATModel1)

##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years 26.0952 5.3894 4.842 1.49e-05 \*\*\*  
## Expend 1.8609 0.6351 2.930 0.00526 \*\*   
## Rank 9.8258 0.5987 16.412 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.25 on 46 degrees of freedom  
## Multiple R-squared: 0.8711, Adjusted R-squared: 0.8627   
## F-statistic: 103.6 on 3 and 46 DF, p-value: < 2.2e-16

summary(SATModel2)

##   
## Call:  
## lm(formula = SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.931 -5.471 1.932 14.980 43.280   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -204.5982 117.6871 -1.738 0.088963 .   
## Years 21.8905 6.0372 3.626 0.000731 \*\*\*  
## Public -0.6638 0.4500 -1.475 0.147154   
## Expend 2.2416 0.6782 3.305 0.001868 \*\*   
## Rank 10.0032 0.6033 16.581 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.93 on 45 degrees of freedom  
## Multiple R-squared: 0.8771, Adjusted R-squared: 0.8661   
## F-statistic: 80.25 on 4 and 45 DF, p-value: < 2.2e-16

# Null: Added coefficients for the added predictors are equal to zero   
# Alternative: At least 1 is nonzero   
# tehre is only 1 added predictor; we are testing that public = 0 vs the alternative that it is nonzero

# Nested test on the things   
# This tells us the same pvaule resut  
# Doing a nested test for the difference with one term in our model is the same as doing those individual tests for slope   
##IMPORTANT ABOVE  
# The below anova is less useful with one term at a time, but it's pretty useful if judgeing multiple terms at a time   
anova(SATModel1, SATModel2)

## Analysis of Variance Table  
##   
## Model 1: SAT ~ Years + Expend + Rank  
## Model 2: SAT ~ Years + Public + Expend + Rank  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 46 31708   
## 2 45 30246 1 1462.5 2.1759 0.1472

**Model Selection with Categorical and Interaction Predictors** Use each of the four model selection methods discussed in class (AllSubsets, Backwards, Forwards, and Stepwise) and compare the processes and outcomes for the predictor pool: Rest, Exercise, Hgt, Wgt, Rest & Exercise, Hgt & Exercise, and Wgt & Exercise

* WE saw in teh past that the regsubsets method wasn’t the best becuase it included things that weren’t as useful
* it picked a chose levels of things when we wanted all of the levels or none of the levels; and it also liked to pick and choose certain interaction terms, some of which were not included in the model
* If you want to include an interaction term, you have to have both terms already in the model

# THis is setting things up  
Full=lm(Active~Rest+Hgt+Wgt+factor(Exercise)+Rest\*factor(Exercise)+ Hgt\*factor(Exercise) + Wgt\*factor(Exercise), data=Pulse)  
none=lm(Active~1,data=Pulse)  
MSE=(summary(Full)$sigma)^2

#Backwards selection  
back\_mod = step(Full,scale=MSE, trace=FALSE)  
back\_mod

##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise),   
## data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest Hgt   
## 84.97301 1.13968 -1.33728   
## Wgt factor(Exercise)2 factor(Exercise)3   
## 0.10212 -4.19657 -70.52397   
## Hgt:factor(Exercise)2 Hgt:factor(Exercise)3   
## 0.09612 1.02785

# Forward selection  
forward\_mod = step(none,scope=list(upper=Full), scale=MSE,direction="forward", trace=FALSE)  
forward\_mod

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# Stepwise selection  
step\_mod = step(none,scope=list(upper=Full),scale=MSE, trace=FALSE)  
step\_mod

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# Comaring the nested backwards selection model to the stepwise selection method  
# IF we look at the nested test values of these   
# Ho: At there is no difference between the models   
# Ha: At least one variable is non zero   
# Do we have sig evidence that at least one of these predictors coefficient is non zero?   
anova(back\_mod, step\_mod)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise)  
## Model 2: Active ~ Rest  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 367 71441   
## 2 373 75050 -6 -3608.9 3.0899 0.005788 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# There are 6 predictors different from the two   
# The factor exercise has 2 additional dummy variables and the interactio nhas 2 addiitonal variables   
# is the coeff for these 6 extra terms equal to zero or evidence that non zero   
# Small pvalue, evidence that at least 1 is non zero   
# mallow Cps may not fit for this model, we might have a lower mallow Cp for rest, it slooks like its a sig imporvement ot add these different criteria to it   
  
#It is an addiitonal tool to build a bigger model