STOR 455 - Class Coding categorial variables

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

## # A tibble: 6 x 7  
## Active Rest Smoke Sex Exercise Hgt Wgt  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 97 78 0 1 1 63 119  
## 2 82 68 1 0 3 70 225  
## 3 88 62 0 0 3 72 175  
## 4 106 74 0 0 3 72 170  
## 5 78 63 0 1 3 67 125  
## 6 109 65 0 0 3 74 188

**Nested F-test** 𝐴𝑐𝑡𝑖𝑣𝑒 =𝛽\_0+𝛽1𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+ 3𝑅𝑒𝑠𝑡𝑆𝑒𝑥 +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.

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

**More than Two Categories** Example: (Active pulse)

* Exercise: – 1 = Slight – 2 = Moderate – 3 = Lots

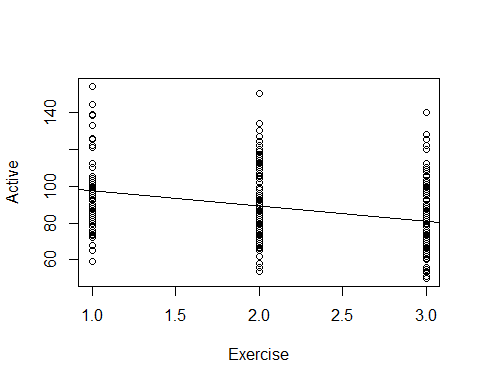
-Try a model to predict Y=Active pulse rates using X=Exercise. How should the coefficients be interpreted?

\_Predicting Active with Exercise\_\_

modelEX = lm(Active ~ Exercise, data=Pulse) # Predict active heart rate by exercise rate   
summary(modelEX)

##   
## Call:  
## lm(formula = Active ~ Exercise, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.613 -12.879 -1.613 9.121 60.754   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 105.979 2.878 36.829 < 2e-16 \*\*\*  
## Exercise -8.367 1.224 -6.834 3.37e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 373 degrees of freedom  
## Multiple R-squared: 0.1113, Adjusted R-squared: 0.1089   
## F-statistic: 46.71 on 1 and 373 DF, p-value: 3.372e-11

plot(Active ~ Exercise, data=Pulse)  
abline(modelEX)



# We are saying there is some change between exervise levels   
# Does exercising a moderate and a lot amount have the same impact as exercising a small and mdoerate amount?   
# We are summing that you exercising changes is constant regardless of group

* 105 = exercise rate of 0, but we don thave one
* exercise 1 = whatever the intercept is plus the slope because we are just going over on eunit and intercept is negative; r=poreict would be 105-8.37 = 89 prediction show the same distance bteween groups We dont want to make that asusmption here

Tkae more care with things that are not binary; we need to foroce varibales to be binary

**Active Pulse vs. Exercise Categories**

tapply(Pulse$Active, Pulse$Exercise, mean)

## 1 2 3   
## 96.24242 90.41290 80.29221

# Slpit groups by exercise levels   
# WE want to know what the average is for each thing   
#Is the “slope” from 1 to 2 the same as from 2 to 3?  
#Note: Using Exercise as a quantitative predictor forces the “slopes” to be the same.  
  
# The oringial model is telling me that there is no change between the mean heart rates based on exercise level; this is telling me that there is a change.  
# WE dont know if it's a significant change or not yet.   
  
#It's ordnal, the exercise levels

\_-Dummy Indicators for Multiple Categories\_\_ For a categorical predictor with k levels, we use k-1 dummy indicators. - X1 = 1 if group #1, 0 if otherwise - Xk-1 = 1 if graph is k-1, 0 if otherwise

*Below: R Trick: (To create indicator variables)* What happens to Group #k?

*Predicting Active Using Slight and Moderate Exercise Indicators* Call: lm(formula = Active ~ Slight + Moderate, data = Pulse)

Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) 80.292 1.392 57.670 < 2e-16 Slight 15.950 2.542 6.275 9.74e-10  
Moderate 10.121 1.966 5.148 4.27e-07

Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096 F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

Pulse$Moderate=(Pulse$Exercise==2)\*1 # Be careful! this is 2!  
# This says that if it is 2, it will be true  
Pulse$Slight=(Pulse$Exercise==1)\*1  
# this says if it is 1, then it will be true  
  
# Multiplying it by 1 will treat the trues and falses as 1 and 0   
# We only need to do this for all but 1, because if all false, then it's whwatever is left over  
  
modelEX2 = lm(Active ~ Slight + Moderate, data = Pulse)  
summary(modelEX2)

##   
## Call:  
## lm(formula = Active ~ Slight + Moderate, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.242 -12.413 -1.292 8.647 59.708   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80.292 1.392 57.670 < 2e-16 \*\*\*  
## Slight 15.950 2.542 6.275 9.74e-10 \*\*\*  
## Moderate 10.121 1.966 5.148 4.27e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 372 degrees of freedom  
## Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096   
## F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

#small pvalue; so we do have some evi that at least one of the coef are not zero   
# other predictors look good   
# The rsquared, only 11% is explained, so it's not that its not explaining, buyt alone it's probably not best by itself   
  
# Look at thow th emodel is set up, we dont see exercise a lot   
# The intercept = those who exercise a lot   
# For those who exercise a lot, we predict their active heart rate is 80.29  
# IF you look a th em eanthe mean value = the same active heart rate   
# People who exercise a slight aount; then slight would be 1 and moderate would be zero ; then we would get intercept of 96 for slight   
# Doing it this way, we dont have to assume that the change is consistent among the levels of our categorical variables   
  
# WE dont need an extra variable, and if we include it, then we will probably get NA values

**Handling Categorical Predictors in R** - If a predictor in lm( ) has “text” values, R will automatically create indicators for all but one category. - Using factor( )around a quantitative predictor in lm( )creates the indicators. - If you let R decide, then R will decide which one to elave out and you might not know which one it stalking about - R Treats categorical varibales this way - If the categories were Slight, mdoerate and high, then R would factor it right - IF we want to use a numeric value as a category, then use factors

modelEX3=lm(Active~factor(Exercise),data=Pulse)  
summary(modelEX3)

##   
## Call:  
## lm(formula = Active ~ factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.242 -12.413 -1.292 8.647 59.708   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 96.242 2.127 45.253 < 2e-16 \*\*\*  
## factor(Exercise)2 -5.830 2.539 -2.296 0.0223 \*   
## factor(Exercise)3 -15.950 2.542 -6.275 9.74e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 372 degrees of freedom  
## Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096   
## F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

# Looks a little differen than before, because we have a different reference category   
# IT chose to leave out the people who exercise a slight amount   
# Intercept = slight amount average   
# Intercept + Eecise 2 = moderate maount   
# 96-15 = high amount   
  
# No reason we can't include more, so look below for more inclusions

**Multiple Categories in Regression** - With indicator variables for categories we can include quantitative and categorical predictors in the same model

modelEX4=lm(Active~Rest+factor(Exercise),data=Pulse)  
summary(modelEX4)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.653 -9.206 -2.629 7.231 65.073   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.25869 6.70517 1.381 0.168   
## Rest 1.15698 0.08611 13.436 <2e-16 \*\*\*  
## factor(Exercise)2 1.62128 2.15805 0.751 0.453   
## factor(Exercise)3 -0.51883 2.38266 -0.218 0.828   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.19 on 371 degrees of freedom  
## Multiple R-squared: 0.4043, Adjusted R-squared: 0.3995   
## F-statistic: 83.92 on 3 and 371 DF, p-value: < 2.2e-16

# We looked at the lines f coef table for exericse; maybe not useful due to pvalue   
# Need to do to nested test value because if one is small pvalue adn the other is big we dont want to use one level fothe categorical varible we want one or all   
# Unless we look at if exercise a lot has effect on heart rate; we just want to know if you exercise a lot or you dont; then just look at one category   
# IN general we wnast ot keep all of the categories   
  
# Could do a nested test to see if exercise is s auseful predictor in the odel   
mod = lm(Active~Rest, data=Pulse)  
anova(mod, modelEX4)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + factor(Exercise)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74699 2 350.9 0.8714 0.4192

# Careful here not comparing two predictors to one; its compare with 3 to 1   
# Cator excerise will give 3 var because it has 3 levels   
# Test will do is do a test is the coef of exercise factor 2 = to 0 and the coef of exercise factor 3 = 0 vs the alternative that at least one of them is nonzero?  
# We get a big pvalue; we dont have evidence that adding the exercise terms are improving the model   
# They are not a sig improvement   
  
# We then run into the same issue with binary cate variables that there is some relation between teh resting and active heart rate, but does that change for those who exercise a slight moderate and a lot?   
# Maybe the resting is not so different, but the active heart rates might be differnt?   
  
# THis model is assuming there is a same splot and same realtionship bt active and rest for all exercise levels   
# We are just changing the intercept

**Multiple Categories in Regression with Interactions** - With indicator variables for categories we can include quantitative, categorical, and interaction predictors in the same model

modelEX4int=lm(Active~Rest+factor(Exercise)+Rest\*factor(Exercise),data=Pulse)  
# Adds the interaction term   
# THis will add ac ouple of terms in here, but ti will tell you if the itneraction ebtween things is sig or not   
  
summary(modelEX4int)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise) + Rest \* factor(Exercise),   
## data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.420 -9.609 -2.467 7.008 64.374   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.5850 13.1326 -0.578 0.5639   
## Rest 1.3810 0.1731 7.977 1.91e-14 \*\*\*  
## factor(Exercise)2 28.6009 16.4065 1.743 0.0821 .   
## factor(Exercise)3 16.5284 15.7150 1.052 0.2936   
## Rest:factor(Exercise)2 -0.3715 0.2240 -1.659 0.0980 .   
## Rest:factor(Exercise)3 -0.2273 0.2216 -1.026 0.3056   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.18 on 369 degrees of freedom  
## Multiple R-squared: 0.4087, Adjusted R-squared: 0.4007   
## F-statistic: 51.01 on 5 and 369 DF, p-value: < 2.2e-16

anova(lm(Active~Rest, data=Pulse), modelEX4int)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + factor(Exercise) + Rest \* factor(Exercise)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 369 74146 4 903.87 1.1246 0.3446

# A line can show the differnce btw active and resting heart rate   
# Exercise elvel 1 which is the level not including this model, we have an intercept of -7.58 and a slope of 1.38 - that st he realtionship   
# For exercise level two , there would be 2 adj to the mopdel; the intercept is going to be the value for our intercept +28, because the intercept is going to change a bit and our resting relationship is going to be the 1.38 slope - .37  
# These are our adjustments   
# Looks like a drastic change   
# factor exercise 3, people who exercise a lto - the intercept will change by this amount and the slope will change by the .22; these are pretty differen tlines   
# If we plotted them we would see there is a big difference   
# We could do some tests to see if they are sid dif.

**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 - They dont all treat them in teh same way

**All subsets**

library(leaps)  
all = regsubsets(Active~   
 Rest+  
 factor(Exercise)+  
 Rest\*factor(Exercise)+  
 Hgt\*factor(Exercise)+  
 Wgt\*factor(Exercise),   
 data = Pulse, nvmax = 11)  
  
ShowSubsets(all)

## Rest factor(Exercise)2 factor(Exercise)3 Hgt Wgt  
## 1 ( 1 ) \*   
## 2 ( 1 ) \* \*   
## 3 ( 1 ) \* \* \*  
## 4 ( 1 ) \* \* \* \*  
## 5 ( 1 ) \* \* \* \*  
## 6 ( 1 ) \* \* \* \*  
## 7 ( 1 ) \* \* \* \* \*  
## 8 ( 1 ) \* \* \* \*  
## 9 ( 1 ) \* \* \* \*  
## 10 ( 1 ) \* \* \* \*  
## 11 ( 1 ) \* \* \* \* \*  
## Rest:factor(Exercise)2 Rest:factor(Exercise)3 factor(Exercise)2:Hgt  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 )   
## 6 ( 1 ) \*  
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*  
## 9 ( 1 ) \* \* \*  
## 10 ( 1 ) \* \* \*  
## 11 ( 1 ) \* \* \*  
## factor(Exercise)3:Hgt factor(Exercise)2:Wgt factor(Exercise)3:Wgt  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## 10 ( 1 ) \* \* \*  
## 11 ( 1 ) \* \* \*  
## Rsq adjRsq Cp  
## 1 ( 1 ) 40.15 39.99 14.57  
## 2 ( 1 ) 40.43 40.11 14.77  
## 3 ( 1 ) 41.70 41.23 8.57  
## 4 ( 1 ) 42.01 41.38 8.58  
## 5 ( 1 ) 42.84 42.07 5.20  
## 6 ( 1 ) 43.02 42.09 6.04  
## 7 ( 1 ) 43.29 42.20 6.35  
## 8 ( 1 ) 43.52 42.28 6.87  
## 9 ( 1 ) 43.58 42.19 8.45  
## 10 ( 1 ) 43.65 42.10 10.03  
## 11 ( 1 ) 43.65 41.94 12.00

# Scroll, over to see where the lowest mallow cp is   
  
ShowSubsets(all)[5,] # Best mallow Cp

## Rest factor(Exercise)2 factor(Exercise)3 Hgt Wgt  
## 5 ( 1 ) \* \* \* \*  
## Rest:factor(Exercise)2 Rest:factor(Exercise)3 factor(Exercise)2:Hgt  
## 5 ( 1 )   
## factor(Exercise)3:Hgt factor(Exercise)2:Wgt factor(Exercise)3:Wgt  
## 5 ( 1 ) \*   
## Rsq adjRsq Cp  
## 5 ( 1 ) 42.84 42.07 5.2

# This is not idea because it isnt taking all levels of the varibaile; it might include an interaction term, but i t might not include the indivudal values; which is bad

Full = lm(Active~Rest+Hgt+Wgt+Wgt\*factor(Exercise)+Rest\*factor(Exercise)+ Hgt\*factor(Exercise), data = Pulse)  
# Fullmodel with all predictors we want   
  
none = lm(Active~1, data = Pulse)  
# Model with non   
  
MSE = (summary(Full)$sigma)^2  
# Pull out MSE  
  
# Sets up the process

**Backwards Selection**

step(Full, sclae=MSE)

## Start: AIC=1988.5  
## Active ~ Rest + Hgt + Wgt + Wgt \* factor(Exercise) + Rest \* factor(Exercise) +   
## Hgt \* factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## - Wgt:factor(Exercise) 2 368.83 71026 1986.5  
## - Rest:factor(Exercise) 2 388.85 71046 1986.6  
## - Hgt:factor(Exercise) 2 600.51 71258 1987.7  
## <none> 70657 1988.5  
##   
## Step: AIC=1986.45  
## Active ~ Rest + Hgt + Wgt + factor(Exercise) + Rest:factor(Exercise) +   
## Hgt:factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## - Rest:factor(Exercise) 2 414.99 71441 1984.6  
## <none> 71026 1986.5  
## - Hgt:factor(Exercise) 2 1233.28 72259 1988.9  
## - Wgt 1 1606.43 72632 1992.8  
##   
## Step: AIC=1984.64  
## Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## <none> 71441 1984.6  
## - Hgt:factor(Exercise) 2 1270 72711 1987.2  
## - Wgt 1 1683 73123 1991.4  
## - Rest 1 34858 106298 2131.7

##   
## 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

# Not saying we could take out hgiehg, says we would haev to remove the interaction term as well   
# Removing weight is possible because th te interaction is gone   
# Takes itno account the restrictions for the model

**forward Method**

step(none, scope=list(upper = Full), sclae = MSE, direction = "forward")

## Start: AIC=2181.6  
## Active ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + Rest 1 50342 75050 1991.1  
## + factor(Exercise) 2 14342 111050 2140.1  
## + Hgt 1 3238 122154 2173.8  
## <none> 125392 2181.6  
## + Wgt 1 397 124995 2182.4  
##   
## Step: AIC=1991.12  
## Active ~ Rest  
##   
## Df Sum of Sq RSS AIC  
## <none> 75050 1991.1  
## + Hgt 1 350.00 74700 1991.4  
## + Wgt 1 148.18 74902 1992.4  
## + factor(Exercise) 2 350.90 74699 1993.4

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

# starts with none, puts in rest, adn doesnt igve option to add interactions  
# Can only add interaciton if the two thigns were in it   
# Tells you to use just rest

**Stepwise**

step(none, scope = list(upper=Full), scale=MSE)

## Start: AIC=271.2  
## Active ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + Rest 1 50342 75050 14.568  
## + factor(Exercise) 2 14342 111050 201.516  
## + Hgt 1 3238 122154 256.563  
## + Wgt 1 397 124995 271.162  
## <none> 125392 271.200  
##   
## Step: AIC=14.57  
## Active ~ Rest  
##   
## Df Sum of Sq RSS Cp  
## <none> 75050 14.568  
## + Hgt 1 350 74700 14.770  
## + Wgt 1 148 74902 15.806  
## + factor(Exercise) 2 351 74699 16.765  
## - Rest 1 50342 125392 271.200

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

# Tells you about the same thing, with only rest   
# Stepwise and forward are very different based on what they do   
  
# Backwards eleminiation goes backwards, least compuational, but you might have a bigger model thatn you need   
# Forward start with nothign and risk a too small method   
# Stepwise is noramlly between, but in this case it was like forward   
  
# We like thes other methods becuase they treat the intearciton terms differently.  
  
#I would say if there are a lot of interaction terms, then you should probably use backwards selection