STOR 455 - Class 17 – Comparing two regression lines

What do we do with categorical variables? **Example: Pulse Rates** - Response Variable:  
- Y =Active pulse - Predictors:  
- X1 = Resting pulse - X2 = Sex (0=M, 1=F) - Datat that looks at heart rate and how we can use sex as a categorical indicator varianble and how we should consider it if we add it to our model - If we just use sex –If we wanted to see the difference bt the binary groups, then we could do a two sample t test

**Categorical Predictor** Example:  
Response = Y = Active pulse Predictor = X = Sex  
- *Are active pulse rates related to sex? “Usual” procedure?* - Two-sample t-test (difference in means) – How different are the two means vs what I would expect to see from them? – A few ways you can do difference in means test – How unlikely would it be that they are this different by chance in our smaples

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

**(using pooled variances), Two-sample T-test for Means** Ho: mu1 = mu2 Ha: mu1 != mu2

where, (pooled standard deviation): 𝑠\_𝑝=√(((𝑛\_1−1) 𝑠\_1^2+〖(𝑛\_2−1)𝑠〗\_2^2)/(𝑛\_1+𝑛\_2−2))

𝑡=(𝑦̅\_1−𝑦̅\_2)/(𝑠\_𝑝 √(1/𝑛\_1 +1/𝑛\_2 )) Compare to t with 𝑛\_1+𝑛\_2−2 d.f.

**R - Two-sample T-test** - pvalue of 0.004853 is significant

* We want to see if active heart rates show any difference
* Will add var.equal=TRUE
* By default the dif in mean test assumes that the means are euqla bt the groups and wants to see how unlikely you get this result by chance
* ytou could also make the assumption that they come from teh same popuilation, so the means and spread should be equal as well
* We want withthe assumption that the mean active heart rates are equal by sex;
* Ho: Mean active heart rates = by sex
* Ha: No equal
* Evidence there is some difference by sex

t.test(Active~Sex, var.equal=TRUE, data=Pulse)

##   
## Two Sample t-test  
##   
## data: Active by Sex  
## t = -2.8329, df = 373, p-value = 0.004863  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -8.992040 -1.623594  
## sample estimates:  
## mean in group 0 mean in group 1   
## 84.60753 89.91534

**“Dummy” Predictors** - We can code a categorical predictor as (0,1) - How should this be interpreted in a regression? - Example: Y = Active pulse, where 0 = male and 1 = female

**For summary of modelP** Table, estimate, std, error, tvalue, P Intercept, mean for males, 1.330, 63.607, pvalue sex, difference for females, 1.874, 2.833, 0.00486 residual standard error….

**t.test(active~Sex, var.equal=TRUE, data = Pulse)** <data:Active> by Sex t = -2.8329, df = 373, pvalue = 0.004863 <- this is the t-test for significant difference

sample estimates: mean in group 0 mean in group 1 84.60753 (mean for males), 89.91534 (mean for female)

modelP=lm(Active~Sex, data=Pulse) #Active hr predicted by sex   
# In teh data,t eh intercept = mean for the meales ebc its where sex = 0   
# The mean for 0 = 84 = the male active heart rate   
# The mean for 1 = 89 = female active heart rate, if   
# If out linear model is just an intercept + sex\*slope, and sex is a 1 or a 0, then it ends up being intercept + 0 or intercept + slope   
# Getting together = female active heart rate   
summary(modelP)

##   
## Call:  
## lm(formula = Active ~ Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.915 -12.761 -1.915 9.392 69.392   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 84.608 1.330 63.607 < 2e-16 \*\*\*  
## Sex 5.308 1.874 2.833 0.00486 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.14 on 373 degrees of freedom  
## Multiple R-squared: 0.02106, Adjusted R-squared: 0.01844   
## F-statistic: 8.025 on 1 and 373 DF, p-value: 0.004863

# looking at t tests  
# same pvalue, dif sig digets   
# Can use a linear model like we did d fiference in means test in teh past

**Quantitative + Indicator Predictors** Example: Y = Active pulse rate X1 = Resting pulse rate X2 = Sex (0,1) (𝐴𝑐𝑡𝑖𝑣𝑒)̂=8.016+1.165𝑅𝑒𝑠𝑡+2.326𝑆𝑒𝑥

*How do we interpret the coefficient of sex?* Ho: B2 = 0 Ha: B2 != 0

With pvalue of 0.116 (because B2 is Sex), There is not evidence to reject the null hypothesis and suggest that B2 != 0

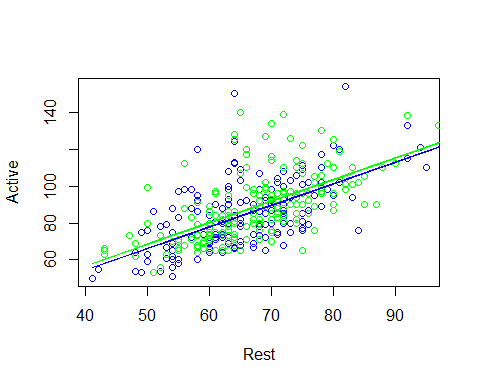
modelP2=lm(Active~Rest+Sex, data=Pulse) # model with quanti and categorical variables  
summary(modelP2)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.706 -9.396 -2.742 6.787 67.434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.01600 5.04661 1.588 0.113   
## Rest 1.16484 0.07511 15.508 <2e-16 \*\*\*  
## Sex 2.32642 1.47471 1.578 0.116   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.16 on 372 degrees of freedom  
## Multiple R-squared: 0.4055, Adjusted R-squared: 0.4023   
## F-statistic: 126.8 on 2 and 372 DF, p-value: < 2.2e-16

# Ho: Rest = 0   
# Ha: Rest != 0   
# Small pvalue = reject Ho  
  
# Ho: Sex = 0   
# Ha: Sex != 0   
# Higher pvalue = fail to reject Ho

**Same slope, different intercepts**

plot(Active~Rest, col="blue", data=subset(Pulse,Sex==0))  
  
points(Active~Rest, col="green", data=subset(Pulse,Sex==1))  
# plots the points on the graph   
  
# Below shows where we got the things from   
B\_Int = summary(modelP2)$coef[1,1]  
B\_Rest = summary(modelP2)$coef[2,1]  
B\_Sex = summary(modelP2)$coef[3,1]  
  
# Plots the line of males and females separately   
lines(  
 B\_Int + B\_Rest \* Rest ~ Rest,   
 col = "blue",   
 data = Pulse  
 )  
  
lines(  
 (B\_Int + B\_Sex) + B\_Rest \* Rest ~ Rest,   
 col = "green",   
 data = Pulse  
 )



# THis shows that it is forcing us to assume that there is the same slope per sex

**Fit Models to Subsets** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.440+1.1432 𝑅𝑒𝑠𝑡 (Males)

Males=subset(Pulse, Sex==0)  
modelPM=lm(Active~Rest, data=Males)  
summary(modelPM)

##   
## Call:  
## lm(formula = Active ~ Rest, data = Males)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.468 -9.426 -2.462 8.109 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.4399 7.4324 1.27 0.206   
## Rest 1.1432 0.1119 10.21 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.09 on 184 degrees of freedom  
## Multiple R-squared: 0.3618, Adjusted R-squared: 0.3583   
## F-statistic: 104.3 on 1 and 184 DF, p-value: < 2.2e-16

# Intercept 9.43 and a slope of 1.1432 for the male model

**Fit Models to Subsets** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.153+1.1823 𝑅𝑒𝑠𝑡 (Females)

Females=subset(Pulse, Sex==1)  
modelPF=lm(Active~Rest, data=Females)  
summary(modelPF)

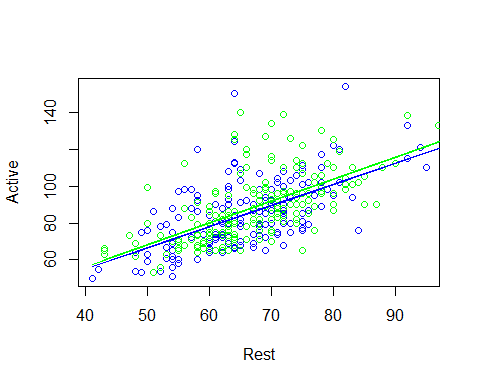
##   
## Call:  
## lm(formula = Active ~ Rest, data = Females)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.088 -3.177 6.010 54.000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.1527 7.0198 1.304 0.194   
## Rest 1.1823 0.1016 11.633 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.26 on 187 degrees of freedom  
## Multiple R-squared: 0.4198, Adjusted R-squared: 0.4167   
## F-statistic: 135.3 on 1 and 187 DF, p-value: < 2.2e-16

# Intercept of 9.1527 witha slope of 1.1823 for the female model

We see some difference between the output. Is this a signifigant difference or would I just expect to see this by chance?

**Plotting the lines** - Are these lines “significantly” different?

plot(Active~Rest, col="blue", data=subset(Pulse,Sex==0))  
  
points(Active~Rest, col="green", data=subset(Pulse,Sex==1))  
# The above code puts the dots on the graph   
  
# the below code puts the line of the models with male and female on the graph  
  
lines(  
 summary(modelPM)$coef[1,1] + summary(modelPM)$coef[2,1] \* Rest ~ Rest,   
 col = "blue",   
 data = Pulse  
 )  
  
lines(  
 summary(modelPF)$coef[1,1] + summary(modelPF)$coef[2,1] \* Rest ~ Rest,   
 col = "green",   
 data=Pulse  
 )



# Now we are working with 2 different models instead of one   
  
# The slopes are slightly different, and we are allowing for the different rate of change; we want to do this with the entire dataset.

Y = B0 + B1X1 + B2X2 + B3X1\*X2+Error - We want to create a line that by changing the value of one indivator variable, we can change what teh intercept of the prediction is, as well as what the slope of that model is - The interaction term does this - When sex = 1 what should slope be and when sex = 0 what should slope be? - When sex = 0, then the some things go away, but when sex = 1, the B2 Term will effect the intercept of the model and B3 will affect the slope of the model

**Comparing Two Regression Lines (with a multiple regression)** - Example:  
- Y=Active pulse - X1= Resting pulse - X2= Sex(0,1)

𝑌=𝛽\_𝑜+𝛽\_1 𝑋\_1+𝛽\_2 𝑋\_2+𝛽\_3 𝑋\_1 𝑋\_2+𝜀 Y = Intercept + Quantitative + Indicator + Interaction

**Quantitative + Indicator +Interaction** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.440+1.1432𝑅𝑒𝑠𝑡−0.287𝑆𝑒𝑥+0.039𝑅𝑒𝑠𝑡∗𝑆𝑒𝑥 - How does this relate to the two lines? - Substitute Sex=0 and Sex=1

# Interaction terms  
# ANOVA Assumptions   
# Ho: All Bi = 0   
# Ha: At least one Bi != 0   
modelPint=lm(Active~Rest+Sex+Rest\*Sex, data=Pulse)  
# BEcause rest is sig, it doesn't appear that the intercept change is useful for us to do   
# interaction bt rest and sex has a high pvalue, tells us we might not need the interaction term   
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

*How can we make a sig test for this?* - When we dont appear to have evidence for sig difference in slope or intercept? - Are there different lines to predict what we want? - If there are, then the last two coeff would be 0

**Tests to Compare Two Regression Lines** Y = Bo + B1X1 + B2X2 + B3X1X2 + Error

* **Different Slope** - T test – Ho: B3 = 0 – Ha: B3 != 0
* **Different Intercept** - T test – Ho: B2 = 0 – Ha: B2 != 0
* **Different lines** - See Multiple Regression Model section below – Ho: B2=B3=0 – Ha: B2 != 0 or B3 != 0

**Multiple regression model** - **Testing one term at a time:** – T-test – Ho: B1 = 0 – Ha: B1 != 0

* **Testing all terms at once** – ANOVA – Ho: B2=B3=0 – Ha: Some Bi != 0

*Is there anything in between?*

**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?

# Something in between   
# ANOVA = all coef are zero vs at least one is nonzero - we compare to null mode, how much do we explain vs a null model?   
# We can sub a different model to the null model   
# Caveat: The sub model has to have a nested subset of what we are working with in the bigger model   
  
modelP\_Reduced = lm(Active~Rest, data=Pulse)  
  
# NEsted F Test   
# Do we need all these things in the model?   
# IF we add things tothe model we may get a smaller mallow CP, but is it signifigant improvement? This tells you   
anova(modelP\_Reduced, modelPint) # Nested test

## 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

# First line = test compare model to null model, and its sig   
# THen build a model with rest and sex in it and comparing the model before to this model   
# To see if addingteh sex predictor increases teh varaibility signifigant;y (It is not in this explame)  
# Third line compares rest, sex, and interaction to just the model with rest and sex in it  
# Tells you if we are explaining an extra amount of the varibility by adding the interactiont erm (This tells you you are not explaining a good amount extra)   
# Only showed up this way ebcause fo the order you put it in ANOVA, if you change the order, you change the order it analyzes things and it might change what it says   
  
# F test stat is calc similiarlly, the dif is when we think about sum of squares; its not how much teh SS in this model it's how much they are in this model taking away what is in the reduced model   
# How much variability is being explained by these extra things, not the whole model as a whole

anova(modelPint)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.5710 <2e-16 \*\*\*  
## Sex 1 499 499 2.4824 0.1160   
## Rest:Sex 1 13 13 0.0667 0.7964   
## Residuals 371 74538 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Nested F-test** - 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.

* Test: Ho: Bi = 0 for a “subset” of predictors Ha: Bi != 0 for some predictors in the subset
* 𝐹=(((𝑆𝑆𝑀𝑜𝑑𝑒𝑙\_𝐹𝑢𝑙𝑙−𝑆𝑆𝑀𝑜𝑑𝑒𝑙\_𝑅𝑒𝑑𝑢𝑐𝑒𝑑))⁄(# 𝑝𝑟𝑒𝑑𝑖𝑐𝑡𝑜𝑟𝑠))/𝑀𝑆𝐸𝐹𝑢𝑙𝑙
* F = ((Explained by full model - explained by reduced model)/#predictors tested in Ho)/MSEFullthat is based on the full model
* Compare to F-distribution

**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.

modelPint2 = lm(Active~Sex+Rest+Sex\*Rest, data = Pulse)  
anova(modelPint2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Sex 1 2641 2641 13.1454 0.0003285 \*\*\*  
## Rest 1 48200 48200 239.9081 < 2.2e-16 \*\*\*  
## Sex:Rest 1 13 13 0.0667 0.7963721   
## Residuals 371 74538 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# First one being done is sex a sig pred of active heart rate?   
# Yes it is, but it looks different than the other tests because they were doing soemthing different   
# This test is after considering teh variability explained in teh active heart rat eby a resting heart ratel is sex sig after that?   
# We have a big p value so its not,   
#If we dont take into account the varability in account by resting heart rate first, then sex alone is useful   
# ADding rest and then sex adn rest together has a pvalue of 2.2 to the -16, to its a sig amount explained   
# The last row is teh same because we are still comparing teh same test as we are before   
# A mdoel wtih all 3 vs a model withjust sex and rest

modelPint3 = lm(Active~Rest\*Sex, data = Pulse)  
summary(modelPint3)

##   
## Call:  
## lm(formula = Active ~ 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

# Even though its jsut an interaction term, R assumed we wanted each indidivual terms as well   
# YOu dont want an interaction without each of the terms in teh model before that   
# R knew that I should do that, so it gave it to you

ANOVA Tests - We dont want ot just put one function into it becase thatis depend on the order of the predictors of the model - ONly lets us test one predictor at a time - Adds one layer at a time

We want to compare a model with all 3 predictors with just 1 predictor (just restin gheart rate)

modelP\_Reduced = lm(Active~Rest, data = Pulse)  
summary(modelP\_Reduced)

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.675 -9.142 -2.725 7.062 66.309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.15295 5.05592 1.613 0.108   
## Rest 1.18029 0.07462 15.818 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.18 on 373 degrees of freedom  
## Multiple R-squared: 0.4015, Adjusted R-squared: 0.3999   
## F-statistic: 250.2 on 1 and 373 DF, p-value: < 2.2e-16

* Then we wnat to do the ANOVA on the reduced and others
* To see how much variability is being explaine dby adding those new terms to it

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

# This tells us the df is 2, whih tells us tehre are 2 variables different between tehse two models   
# We get an idea of teh variability explained bt these two models and see tha tthe SS (the 512.14) = the amount of varabiltiy explained by adding teh sex and interaction term to the model ; that is a small amount in the long run and that is why the f test stat is small and the p value is high   
# This lets you test both those terms together   
  
# BEnefit: We are jsut comparing two extra terms; if there were 10 terms different, then we could just do tests of each 10 terms individually, but tehn we'll run into more error issues   
# When we do it all at once, we get a big pitcure if there are any differences and if we need to we can investigate further to see wehre teh differences are

Nested tests tell you if you’re looking at noise or if you’re looking at something signifignat