STOR 455 Class 34 R Notebook

library(readr)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")  
  
Exams4

## # A tibble: 20 x 3  
## Student Exam Grade  
## <chr> <dbl> <dbl>  
## 1 Barb 1 62  
## 2 Barb 2 87  
## 3 Barb 3 74  
## 4 Barb 4 77  
## 5 Betsy 1 94  
## 6 Betsy 2 95  
## 7 Betsy 3 86  
## 8 Betsy 4 89  
## 9 Bill 1 68  
## 10 Bill 2 93  
## 11 Bill 3 82  
## 12 Bill 4 73  
## 13 Bob 1 86  
## 14 Bob 2 97  
## 15 Bob 3 70  
## 16 Bob 4 79  
## 17 Bud 1 50  
## 18 Bud 2 63  
## 19 Bud 3 28  
## 20 Bud 4 47

*Investigating relationships* 1. Y = Quantitative, X = Categorical - “dummy” variable and regression 2. Y = Binary Categorical, X = Quantitative - logistic Regression 3. Difference in two means - 2 sample t test 4. Difference in more than 2 means - ANOVA for Means

Goal TOday: See if there is evidence Many difference groups from same population and see if they have evidence that the means are difference btween the groups

*Samples from K Different Groups* Test: Ho: mu1 = mu2 = mu k Ha: some mui =/= muj

For each row is a different student

Is there a sig difference in average grade among the four exams?

Pull in the data

means = tapply(Exams4$Grade, Exams4$Exam, mean)  
  
tapply(Exams4$Grade, Exams4$Exam, mean)

## 1 2 3 4   
## 72 87 68 73

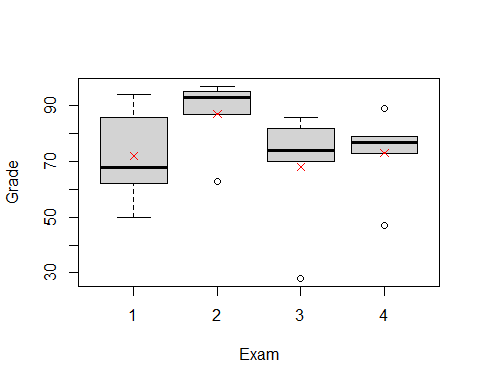
tapply(Exams4$Grade, Exams4$Exam, sd)

## 1 2 3 4   
## 17.88854 13.92839 23.23790 15.68439

tapply(Exams4$Grade, Exams4$Exam, length)

## 1 2 3 4   
## 5 5 5 5

boxplot(Grade ~ Exam, data = Exams4)  
points(means, col="red", pch=4)

 We want think about two completing model s to describe teh situtaion

maybe if i just looked at teh mean overall score, just added them all up = the scores in teh population

completing Ha: Instead, I should see each mu as based on teh exam score for the people who took that exam; just use the mean from each group individually

Note: The groups in teh mu in the group are going to be better overall, but we want to see if it’s a stat sig difference to see if we want to use either one.

*Predicting in ANOVA Model* - If the group means are the same (Ho); if H0 is true - yhat = ybar; for all groups > residual = y - ybar

* If the group means can be different (Ha); if Ha is true
  + yhat = ybar for the ith group = residual = y-ybar
* Do we do sig better with separate means?
  + Compare the sums of squared residuals
  + SSTotal = sum(y-ybar)^2 vs
  + SSE = sum(y-ybari)^2 (how different they are from the new model)

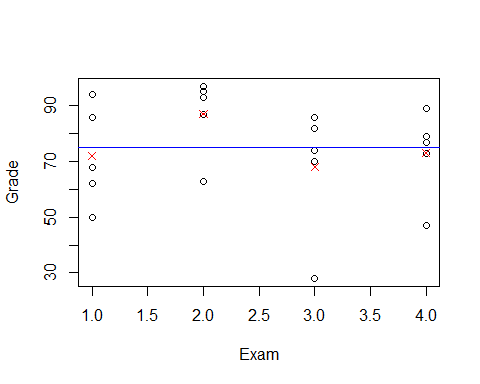
The SSE should be small = when we look at inviduval groups = support Ha

*Partitioning Variability* Data = Model + Error Y = mu i + E Total varation in response, Y = variation explained by MODEL, mui + Unexplained variation in RESIDUALS

Y=Mui +E (y-ybar) = (ybari - ybar) + (y-ybari) Sum(line 77)^2 each thing in parenthesis

SSTOTAL = SSGROUPS + SSE

plot(Grade ~ Exam, data = Exams4)  
points(means, col="red", pch=4)  
abline(h = mean(Exams4$Grade), col = "blue")



The above plots points for grades by exam; we are trying to see fi we can pair each group to the red xs and get a better line than comparing it to the horizontial line

If SSE = big, it tells us that this model may not be some much better than the null model; so we might not want to use the Ha.

We have to factor exam below so that it looks at thte groups, otherwise it will assume some kind of relationship in the exams category

Beklow we see that the sum of squares have a lot

Below the bottom row is so much bigger, so we are not havcing a lot of variability explained by the model

amodG=aov(Grade~factor(Exam),data=Exams4)  
summary(amodG)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Exam) 3 1030 343.3 1.056 0.395  
## Residuals 16 5200 325.0

*Alternate Form: ANOVA MOdel For Means*

Y = mu + aplhai+ E Y = grand mean + effect for ith group + Random Error

mu-hat = y-bar alpha-hat = y-bari - y-bar

Hypothesis Testing: Ho: mu1 = mu2 = muk Ha: some Mui =/= muj

Above Hypo test is interchangable with Ho: alpha1 = alha2 - alphak =0 Ha: Some alphai =/= 0

We think about the effect of the difference. This is how we have phrased things for regression in teh past; so it’s like how we did regression; this is just a different way to do the same thing

We will do more pairwise comparisions in the future classes

*This about in a simualation persepctive* Going to StatKey website

If all of the exam scores are equal, then it wont matter where they are assigned. When you do a random; ANOVA for a difference in means; it randomly scatteres the values to see if the relation is stat sig. We are taking the mean of grades and splitting it up by exam

The ANOVA table in StatKey; looking at teh SSE - how far away each of those values are from their group means squared adn summed

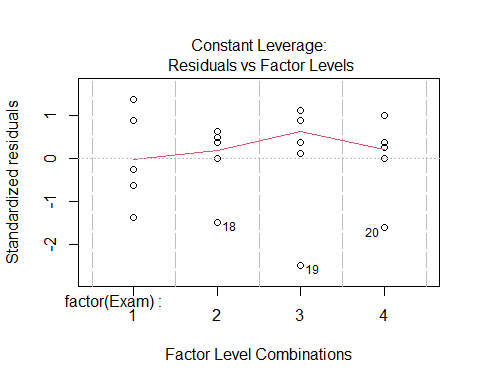
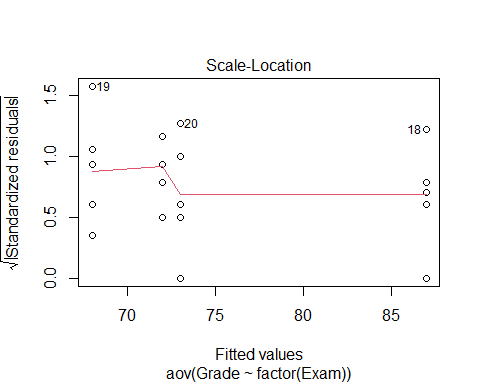
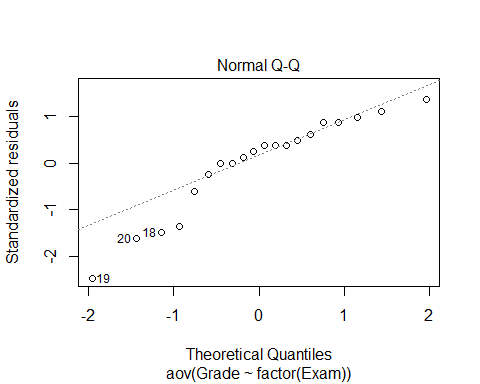
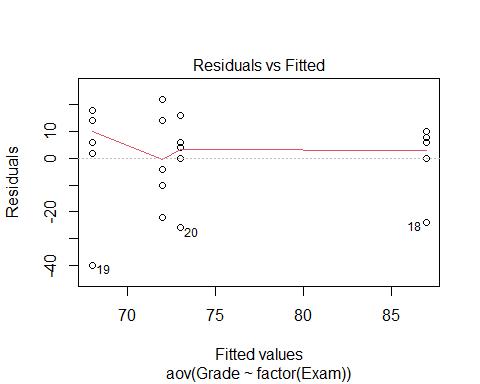
Total - how far are away from total mean, squared adn sum We cna see there are a lot of numbers in teh SSE term, which means that there isn’t a very sig relationship here.

MSE = Sum of squares/dof

F test = MSGroup/MSE

*Checking Conditions for ANOVA* E~N(0,stdE) <- check with residuals 1. Zero Mean: Always holds for sample residuals 2. Constant Variance: Plot residuals vs fits and/or compare std. dev’s og groups (Check if some other groups si is more than two another) 3. Normaility: histogram/normal plot of residuals 4. Indepdnence: Pay attention to data colelction

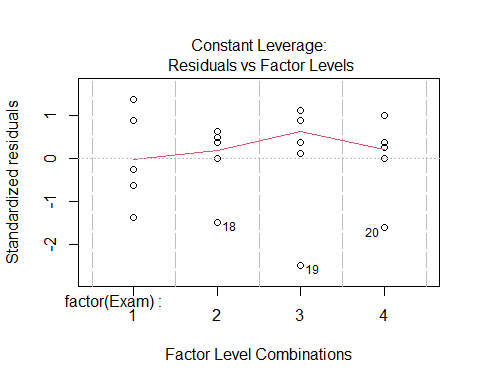
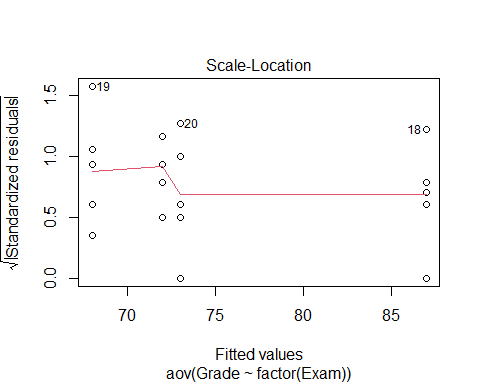
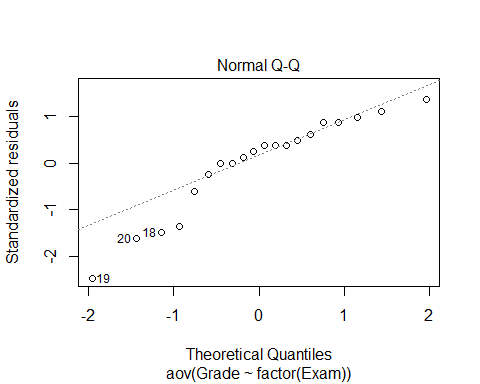
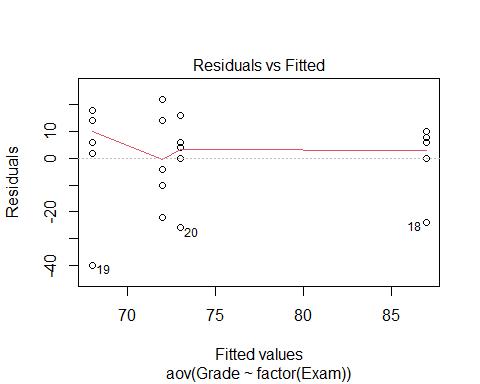
plot(amodG)



round(tapply(Exams4$Grade,Exams4$Exam,sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

plot(amodG)



round(tapply(Exams4$Grade, Exams4$Exam, sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

What about the other direction?

Is there a sig dif in average grade among the 5 students? construct a model by student

There is a stat sig relationship we ahve evidence to say at least 1 is stat sig

but where are those differences?

Bud is probably lower than the rest, but do the other 5 students have smaller things?

There are a lot of different ways to look at this

amodS=aov(Grade~Student,data=Exams4)  
summary(amodS)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Student 4 4480 1120.0 9.6 0.000468 \*\*\*  
## Residuals 15 1750 116.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

NOw that we know there is a difference, hwere is the difference?

N groups \* (n-1)/2 = how many comparisions we have to do

In this example we have 5 students, so 5\*4/2 = 10 combinations

If we do lots of test, we might run into an error; if we are saying then

If an outcome would happen only 5% if the null is true, we’re going to reject it and say teh alternative is suported instead. Well, that could happen, but we might do a type I error; we don’t know that we a re doing a type I error though,

Each time there is a change of making at ype 1 error when you test a combination; each time there is a 5% chance of making a type 1 error when you are working with an alpha level of 0.05.

There is a 95% chance no type 1 Error; so 0.95^20 = chance of making no type 1 errros in a test that has 20 combinations; We have a 0.358 chance of making no errors; we have a 0.6415 chance of making AT LEAST 1 TYPE 1 ERROR.

This is an issue; we need to look at how can we decrease the chance of mkaing type I errors.

*Ways to possibly fix this* a) do a few preplanned comparisions - we cna’t always just test teh difference because we are cheating and we cant really find wehre other differences are b) Adjust teh sig level used for each test - make the sig level higher; so you have a decreased chance of mkaing a type I error; with 1% sig level, so we have a 0.99 chance of NOT makign a TYPE 1 Erric

So we would still have a chance of getting a type 1 Error, but ti woudl be lower.

*Above* Shows one way to show no sig dif vt exames and sig dif bt students but: *Question*: Can we use both factors to help explain teh variability int eh exam scores?

*Looking forwards* - will look at teh ANOVA model for dif in means with 1 predi - how to dod comparisions if there are differenecnes in data - all have difference ways of makign errors and how to avoid it