Contents

[STOR 455 Class 34 R One Way ANOVA 1](#_Toc89722552)

[STOR 455 Class 35 One Way ANOVA (Again) 12](#_Toc89722553)

[We assumed in teh null that these all came from the same population. 17](#_Toc89722554)

[STOR 455 Class 36 Two Way ANOVA 24](#_Toc89722555)

[STOR 455 - Class 37 – ANCOVA and Homogeneity of Variances 40](#_Toc89722556)

[STOR 455 Class 38 Homogeneity of Variances and Contrasts 64](#_Toc89722557)

## STOR 455 Class 34 R One Way ANOVA

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

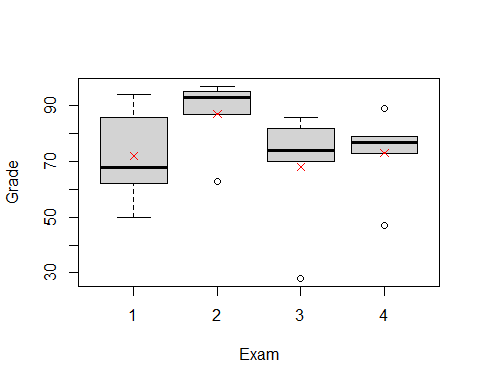
tapply(Exams4$Grade, Exams4$Exam, sd)

## 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)

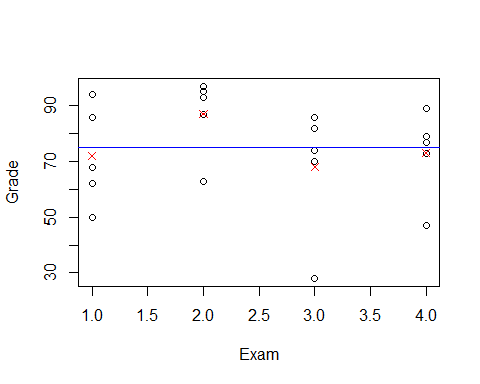
The SSE should be small = when we look at inviduval groups = support Ha

*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

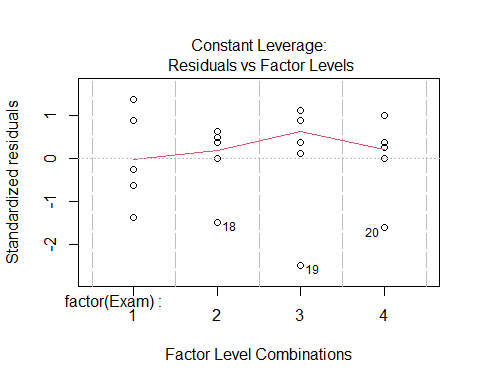
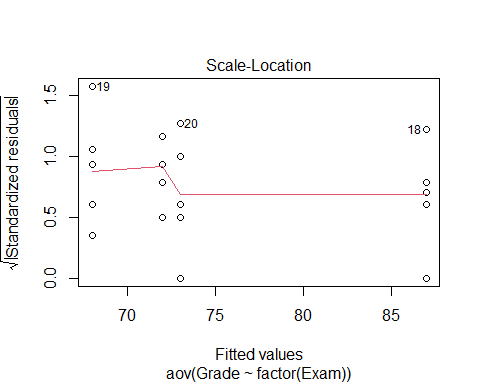
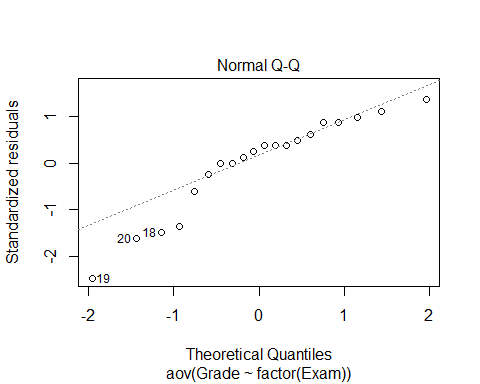
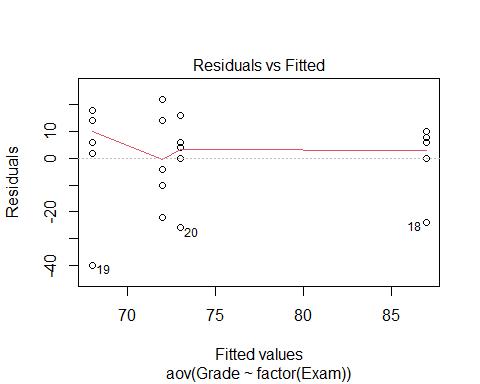
Total - how far are away from total mean, squared adn sum We cna see there are a lot of numbers in teh SSE term, which means that there isn’t a very sig relationship here.

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

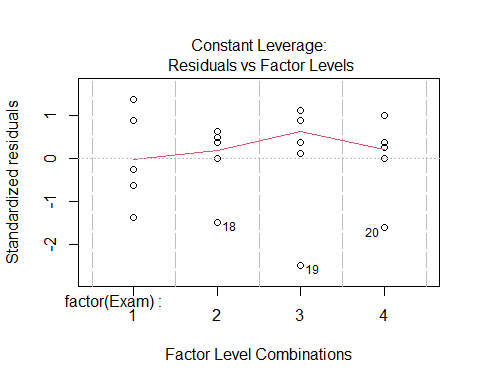
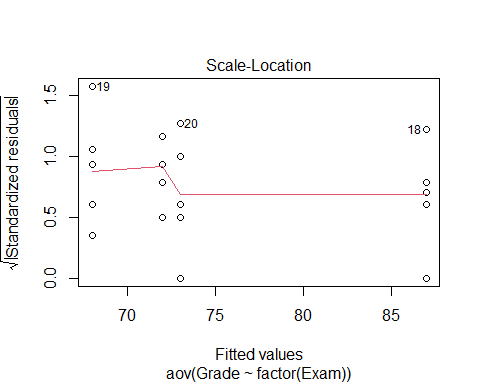
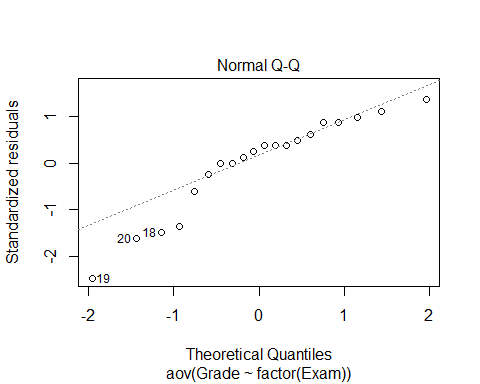
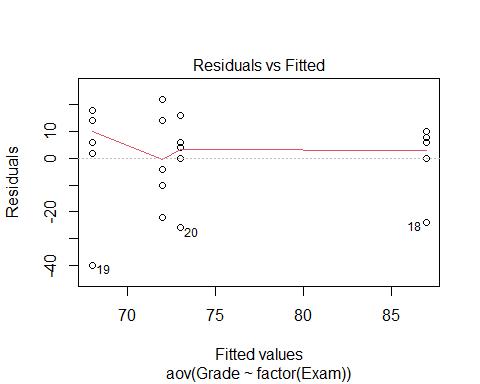
plot(amodG)



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

## STOR 455 Class 35 One Way ANOVA (Again)

Homework 8 comments - you have to pick 6 dif models of car and use a random sample selection of 50 cars - choose these cars that are from jap or us companies; they are not part of the sample, but you have the car names - you’ll have to deduce which are jap vs us cars; do based on what they car type name is; think about the company - choose 3 jap cars that are a compact, midsize and suv - you have to find this yourself - Goal: take samples of 50 and find the compact, midsize, and suv for - use rbind like HW 4 to put all in the same dataframe (left join) - make a new variable - a bit of a setup - reference hw 2 and 4

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

Last time talked about anova for a difference in k mean - if we split it up into groups and classify the mean for the group that it is in, does that explain a significant amount more of the variability? Aka: Should we classify things by their group mean or the overall mean?

*Hypo test* Ho: mu1 = mu2=…muk Ha: some mui != muj

We looked to see if there is sig dif in the average grade among the four exams? Or are teh results by chance?

*Partiioning Variability* Data = Model + Error Y = mui + E total varaition in the response, Y = variation explained by MODEL + Unexplained variation in Residuals

Question: Does the model explain a sig amount of the total variability?

We want to know if we should

the error term is what variabiltiy there is grom each group mean

*Partitioning Variability ANOVA for Group Means* Y = muk + E SSTotal = SSGroups + SSE

The SSTotal = the SSdifferences from the overal gran mean SSE = sum of the difference from each indiivdual group mean SSGroup = the variation explained by the groups that we have imposed on this model

If we have sig evi of the diff bt the means, then the model will have SSE = small compared to the SSGroup; b/c the SSGroup tells us if a lot of the variability has been explained by using the group means

*Below is the model we made last class*

Ho: all exam means are equal in population Ha: at least one exam mean is different

The pvalue is big, so there is no evidece for sig dif

**Below:** IF we sum these, then we get the total SS; SSTotal; if we have a good model, then we would want a lot of variability explained by our groups factor(Exam) and we see there is a lot in the residuals, which is below factor(Exam) - this is why we have a large p value

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

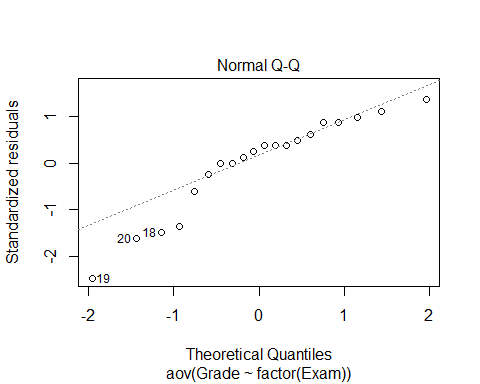
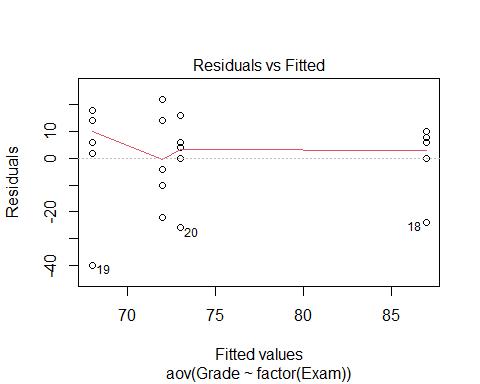
*Checking conditions for ANOVA* *Mostly have to focus on constant variance and normality* -Zero mean: Always holds for sample residuals - Constant Variance = plot residuals vs fits and/or compare std dev.’s of group (check if some groups is is more than twice another) *can be hard to check since we ar looking at groups; and sometimes its just teh mean values; we want to see a similar spread between the groups; can also think that as long as the largest standard deviation is not double the smallest standard deviation, then we ar egood enough for constant variance.* (Think more about the data, because if you have a really small dataset, you might have to throw that idea out the window) - normality: histogram/normal QQplot of residuals - Independence: Pay attention to how the data was collected

**Below: Checking Residuals** This amount of devation is not a problem

**Below: We check to see wht the std are of the functions**

This will split it by categories.

plot(amodG, 1:2)



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

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

*Question* Is there sig dif in the average grade among the five students? Is it a sig dif?

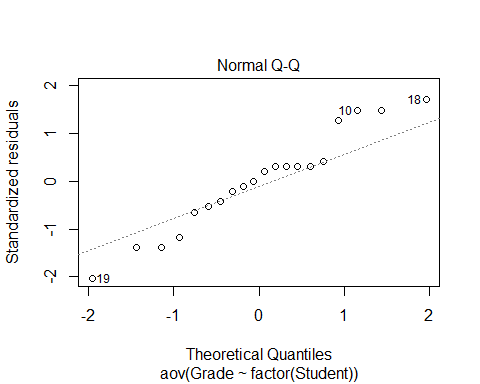
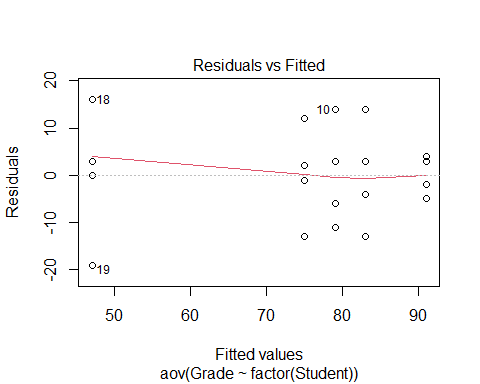
We can build the same model, but just change the group by student level instead of grade. we have evidence to say there is a sig dif.

This plot tells us where each persono’s scores are plotted in relation to each other

modS=aov(Grade~factor(Student),data=Exams4)  
summary(modS)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(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

plot(modS, 1:2)



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

## Barb Betsy Bill Bob Bud   
## 10.30 4.24 10.98 11.40 14.45

*How many comparisions?* Which one of the indivates caused the sig difference?

There could be a lot of comparisions to do.

*Problem of Multiplicity* When we are doing many pairwaise comparisions we are more likely to make a Tpye I Effor (finding a false difference)

**Possible fixes:** - Do only a few pre planned comparisons – If we know what we want to focus on, then we can do that - Adjust the sig level used for each test

* THe more testes we do, the more like we are th make a type 1 error = null is true, but we find evidence to reject it

*Example:* If we did tests at a .05 sig level, then there is a 5% chance that we will make a type 1 error; so there is a 95% chance that we won’t make a type 1 error.

We have to find out how we can minimize the chance of making a tpye 1 error.

# The probability of not making a type 1 error if you run 10 tests  
(1 - 0.05)^10

## [1] 0.5987369

# Chance that we make at least 1 type 1 error when we do 10 tests  
1 - (1 - 0.05)^10

## [1] 0.4012631

# The probability of not making a type 1 error if you run 100 tests   
(1 - 0.05)^100

## [1] 0.005920529

# Chance that we make at least 1 type 1 error when we do 100 tests  
1 - (1 - 0.05)^100

## [1] 0.9940795

*Pairwise comparision AFTER ANOVA* This is where you do every single test on the data

Compute a C.I for mui - muj

Pairwaise t-test for difference in means Ho: mui = muj Ha: mui != muj

*Modifications* - estimate any std with sqrt(MSE) = Se - Use teh error d.f for any t-values

We assumed the null came from a normal population.

Rather than using these ind groups to see the std, you can use the pooled value of all of the groups; we can estim the population std using the std of the residuals in teh model

*Pairwise Inference After ANOVA* C.I for mui - muj = (ybari - ybarj) +/- t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) #Note: use the error d.f for the t-star in both the C.I and Test # If 0 is in the C.I then that would give me evidence for a stat sig dif. # The hypothesis testing tells you how unlikely it would be to get that result

Test: Ho: mui = muj Ho: mui != muj

t = (ybari-ybarj)/(sqrt(MSE)\*sqrt((1/ni)+(1/nj)))

**T-Star tells you how many standard errors i need to go in each direction depending on the sampel size; gives you an approximation for what you think the population difference is**

* Bigger sample, fewer in each direction = trend towards a normal distribution
* MSE = standard devation of the residuals and it is weighted based on the size of the groups
* If zero is in the C.I, then we have evidence

**T-Test Stat** **Is what the actual difference is divided by that measure of the standard deviation scales in this way. Tells me how many standard deviations away from the null this data is**

*Fisher’s LSD* The least sig difference: adding or subing to those differences to see how much variability there is

LSD = t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) # concludes mui differs from uj if abs(ybari-ybarj)>LSD

*If looking at the t-distribution* The thing is centered at zero and we want to know how far in each direction we need to go to get the middle some percent of the data (whatever level you’re doing the sig test at); 95% = 0.05 sig level

We want to figure out what t-star is, because it will tell me how many i need to go in each direction. We want the value of t-star is 97.5% of the data under that curve to the left of it. **Look below, this is where we do the code for this**

# qt, give it an area under the curve, under the t-dist; and it will tell me the t value that is the bound for the left of it.   
  
#We also need df, and it's coming back from the anova model we made. We want the residuals df. We have 5 groups, because 5 students. This just gives us an idea of how much we trust the variability of the sample and how well it can predict the population   
  
# Big df = big sample, so the sample = more like the population   
# Small sample sizes = variability might not predict population as well   
  
qt(1-0.05/2, 15) #Tells you if you go OUTPUT number of standard deviations in each direction of this distribution, then I will capture the middle 95% of the data.

## [1] 2.13145

t\_LSD = qt(1 - 0.05/2, modS$df.residual)  
  
#WE want to make a prediction for the difference in population of the two groups, so we need to go 2.xxx standard deviations in each direction; we will calc the std by the sqrt(MSE)\*sqrt((1/ni)+(1/nj)) ((This is the last half of the LSD model))  
  
MSE = summary(modS)[[1]][,3][2] # This looks at the anova table output from the summary, that's what teh 1 is; the 3 gives me the third column and all rows; the 2 gives the second element of the 3rd column is what we are pulling out  
  
LSD = t\_LSD \* sqrt(MSE)\*sqrt(1/4 + 1/4) # it is 4 because each student took 4 exams   
LSD

## [1] 16.27921

*Interpreting the results above:* If we do each test basically at the 5% sig level, if any of your group means are more than this amount different, we have sig evidence to say there is a difference there;

we are predicting the mean s bt any two groups are this much or less by chance

*ANOVA for Grades vs Students* - Every compariasion we make; everyone’s score ocmpared with bud’s are more than 16.33 apart; so we have evideicent to say barb and bettey are different and others as well.

* THe multiplicity is an issue here; We are doing these tests at the 0.05 level, there are chances of making a type 1 error. This is a small example, but as it gets bigger it could be an error

*The Problem of Multiplicity* - When doing many pairwise comparisions there is an increased change to make a type 1 error (find a false difference) - Fisher’s LSD may be too lenient

*Possible fizes* - Do only a few preplaed comparsions - Use a smaller alpha for each test

OR

**Bonferroni Adjustment** When doing m tests with a *overall* error rate of alpha-star, use alpha = alpha-star/m for each test

* *This adjustment, basically lowering the levels of each one of the individual tests. So the chance of making a tpye 1 error for all the tests together can be defined and then from there you can figure out how twhat a sig level you need for each individual test*
* Bonferroni is going to let us do everything at the 0.05 level, and not a bigger aggregate of this **That’s really useful**
* IT does this by reducing the smaller alpha for each test. Reduce it by dividing the sig level by whatever number of comparisions you want to do.
* If I want the overall sig to be 0.05 I need to divide that by 10 (because there are 10 comparisons), then each test will be done at the 0.005 level.

#Bonferroni Method  
# No type 1 errors  
(1 - 0.05/10)^10

## [1] 0.9511101

# At least 1 type 1 error  
1 - (1 - 0.05/10)^10

## [1] 0.04888987

All that really changes, are the number of std that you have to go to trap that middle value gets bigger

We have to go much wider to track Bonferroni effect; and by going wider, we ar emaking it less likely that we make a type 1 error, but we are also demanding that we have stronger evidence of a difference to see that sig difference

*ANOVA for Grades vs Students*

*How to calculate the Bonferroni t-value* See below:

We get a value of 25, so using this method, we have to be 25 difference in term of average score to show some sig dif bt groups

The results are the same because of our data, but for other examples there may be a different conclusion

**Things to not do** - DOnt p-value hack; don’t choose what fits your narrative

t\_bf = qt(1 - 0.05/10/2, modS$df.residual) # What changed is the division by 10 ebcause of the total number of comparisions we wnat to do   
  
BSD = t\_bf \* sqrt(MSE)\*sqrt(1/4 + 1/4)  
BSD

## [1] 25.09753

*Tukey’s HSD: Honestly Sig Dif* *Replace t-star with value of q-star from teh studentized range distribution with R*

HSD = (q-star/sqrt(2))*sqrt(MSE)*sqrt((1/ni)+(1/nj))

q-star depends on slpha, # groups = K, and error df

tukey’s test is a bit more in the middle when it comes up with getting the dif values

* trying to get some middle group where we reduce the chance of a type 1 error, but not reducing it as much as the last case where we are inflating the case of making a type 2 error.
* the only change you see here is how many standard deivations you need to go in each difrection; it’s somewhere no quite in the middle; but its not as big as the last value
* Tukey’s is more the middle ground for what we would expect to see by chance

HSD = qtukey(1-0.05, 5, modS$df.residual) \* sqrt(MSE)\*sqrt(1/4 + 1/4)

*How can this be automated better by R?* *Automating the LSD and Bonferroni in R*

* Som eof the pvalues have changed

**LSD Method:**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'none') # What this is doing is doing each individual pairs of hypothesis test. It's not calcuing the conf int. It's teh same idea; where if the pvalue is below 0.05, then the dif bt these must be bigger than that 16.3 or something in the previous calculations.

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.05357 - - -   
## Bill 0.60812 0.13699 - -   
## Bob 0.31148 0.31148 0.60812 -   
## Bud 0.00229 3.8e-05 0.00079 0.00028  
##   
## P value adjustment method: none

# We see where the differences are in this table   
# We see what we saw with the confidence intervals, where one student has a sig difference, but others dont.   
# The table isn't filled out because where the dashes are, we already have those valeus; we have already compared bob and bill, but bill and bob would give us the same number

**Bonferroni Method in R**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'bonf')

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.53567 - - -   
## Bill 1.00000 1.00000 - -   
## Bob 1.00000 1.00000 1.00000 -   
## Bud 0.02293 0.00038 0.00789 0.00277  
##   
## P value adjustment method: bonferroni

# Similar, but the bonferronia adjustment   
# A little differently done by R;   
# When we made the C.I, we adj the sig level of test based on the number of groups we had; we lowered teh sig level, which made me needs a lower pvalue to find a difference  
# R's difference: it's calc the pvalues in the same way for the LSD method, then multiplying that by the number of comparisions; so rather than compariting the orignal pvalue to the smaller sig level, it's comparing an inflated pvalue to that same sig level   
# Mathmatically its teh same thing, and you can se it here   
  
# The 1st for the pvalues show it is mult the last case by 10 and if its 1, that's the upper limit because you cant have a prob higher than that   
  
# THis compares each of the pvalues to the same 0.05 level since they are all being multiplyed by 10  
  
# Bud is the big difference dude here again

*Tukey Function* Does the middle ground of making type one and type 2 errors;

diff = diff in mean scores for each pair - lower and upper = 95% conf int using the Tukey’s distribution - Trying to see if 0 in this distribution or not, if 0 is in it, we are predicting that there could be no difference in teh population, if 0 is not in it, then we’re - if there is no 0 invloved, then there may be a sig dif bt the two

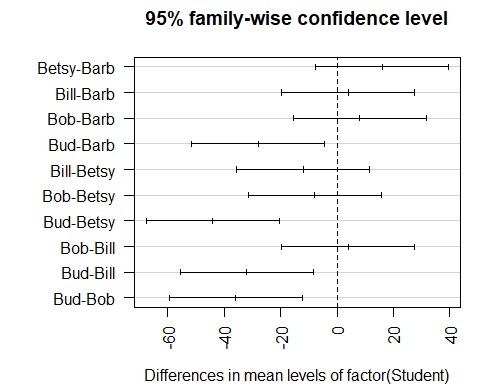
* we could look at the pvalue, which is being inflated; we compare then to the 0.05 level
* the results are the same

TukeyHSD(modS)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Grade ~ factor(Student), data = Exams4)  
##   
## $`factor(Student)`  
## diff lwr upr p adj  
## Betsy-Barb 16 -7.584413 39.584413 0.2720310  
## Bill-Barb 4 -19.584413 27.584413 0.9835150  
## Bob-Barb 8 -15.584413 31.584413 0.8295529  
## Bud-Barb -28 -51.584413 -4.415587 0.0166293  
## Bill-Betsy -12 -35.584413 11.584413 0.5360462  
## Bob-Betsy -8 -31.584413 15.584413 0.8295529  
## Bud-Betsy -44 -67.584413 -20.415587 0.0003116  
## Bob-Bill 4 -19.584413 27.584413 0.9835150  
## Bud-Bill -32 -55.584413 -8.415587 0.0060225  
## Bud-Bob -36 -59.584413 -12.415587 0.0021941

*Added Code* - Visual representation of this - We can plot all the hist on the same axis, and if any cross the vertial 0 line, then there is not sig dif there; - if they dont cross through it, then thats where we see those differences.

origpar = par()  
par(mar=c(4,7,3,1))  
hsd = TukeyHSD(modS)  
plot(hsd, las=2)



par(origpar)

## STOR 455 Class 36 Two Way ANOVA

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

*Diet Information* - Investigate the data to determine the impact of the type of diet and Sex on mean weight loss in separate models. - Evaluate the residuals for the appropriateness of using the ANOVA model in each model - Use pairwise methods to investigate differences between means if the overall ANOVA model is significant.

**Goal:** - Want to see the impact of each diet and how the diet and sex effects teh mean weight loss - First model: ANOVA MOdel by WEight loss by Sex - SEcond: Anova by WEight loss by Diet type

* Want to look at residual plot to check the conditions and if we do see there are sig differences by sex or diet, then we want to follow up and do a pairwise comparision like we did last time to see where those differences are.

head(Diet)

## # A tibble: 6 x 7  
## Person Sex Age Height Preweight Diet weight6weeks  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 25 NA 41 171 60 2 60   
## 2 26 NA 32 174 103 2 103   
## 3 1 0 22 159 58 1 54.2  
## 4 2 0 46 192 60 1 54   
## 5 3 0 55 170 64 1 63.3  
## 6 4 0 33 171 64 1 61.1

There is no weight change variable so we will need to make a variable called weight change and subtract teh differences in weight begining and afterwards.

* Then look back, make an anova model for weight change by diet and then follow through the process of checking the conditions and see wehre teh sig difference with teh pairwise analysis if tehy exists
* Repeat for Sex
* Then we will talk about how to do one test that does the two things above twice

**THE DIET INFORMATION IS NOT ALREADY WRITTEN IN THE RMARKDOWN FILE, YOU WILL NEED TO WRITE THIS FROM THE SLIDE** *Diet Data Conclusions* **Question1:** Are there dif in weight change depending on the diet you are on and if there are, which diet(s) seem to be sig different than each other?

*There are changes* - **Best diet:** The 3rd one – They lost about on average 2kg more and there are sig dif – Practically this doesnt mean its better. Maybe it’s more miserable to be on than the others – Maybe it’s not worth it, but from a stats standpoint, there is a difference there

**Diet Code** - First: Make a new var of the dataframe = weight change - Second: Make a model and factor the diet because we want to see the 3 separate diets - Third: Look at teh summary of that

Ho: Mean weight change for each diet is equal Ha: At least one mean weight change for a diet is differnt

Pvalue: small, tells us

Conclusion: We have sig eviduce that the diets are not equal in the population, so we have evidence to support the alternative

*We should also check that it fits the conditions that we need, so we look at the plots of residuals by fitted and QQNorm* - NOrmality: Seems pretty normal - Residuals by fitted - constant variance; the spread of each of the diets are the same; nothing here stands out; what would stand out is if one diet was tightly packed adn the others were spread out. –This looks different than normal because it’s clustered by diet –We could have also gotten the sd of teh diets and seen if one was bigger than the others; –**Rule of thumb: If the biggest one is more than double the smallest, there may be some issues**

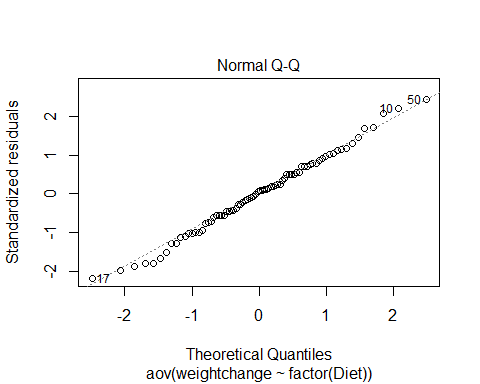
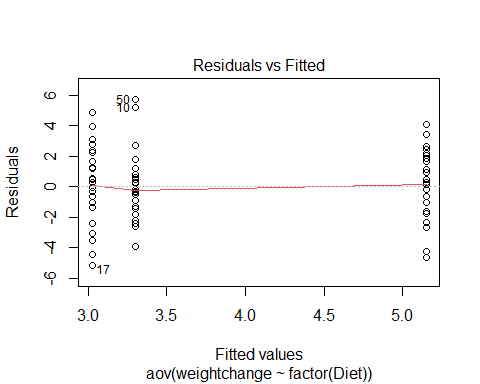
**To see where the differences are, look at the Tukey analysis** *Tukey’s* Gives us middle ground *Bonferroni* Increases chance of type 2 error, so we would need more evidence, which is why we don’t do that one. You can in the future if you ever need it though.

*Tukey’s output* - Inflated p-value: There are two places were the pvlaue is below a point 0.05 level and that’s wehre we are comparing diet 3:1 and 3:2; – Sig dif exist bt the mean weight change bt 3:1 and 3:2 –If you didn’t factor diet, Tukey would probably be mad at you and not run **Troubleshooting**

Diet$weightchange = Diet$Preweight - Diet$weight6weeks  
  
weightchange\_diet = aov(weightchange~factor(Diet), data = Diet)  
summary(weightchange\_diet)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 71.1 35.55 6.197 0.00323 \*\*  
## Residuals 75 430.2 5.74   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(weightchange\_diet, 1:2)



TukeyHSD(weightchange\_diet)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weightchange ~ factor(Diet), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 -0.2740741 -1.8806155 1.332467 0.9124737  
## 3-1 1.8481481 0.2416067 3.454690 0.0201413  
## 3-2 2.1222222 0.5636481 3.680796 0.0047819

tapply(Diet$weightchange, Diet$Diet, mean)

## 1 2 3   
## 3.300000 3.025926 5.148148

**Additional Analysis for Sex** Not much is different; iot says two obs were delted, because there were 2 NA’s for sex that weren’t listed

Looking at teh summary: Sum of Squeares tells us little variability is explained by this model, but we have a high pvalue, so we don’t have evidence to say that Sex plays a role in the weight change

Ho: Sexes are all equal Ha: Sexes are not equal

**Looking at the plots: Does it fit the conditions?** - No need to do a Tukey follow up because we don’t see a difference between the groups, so there’s no reason to keep going with that.

*LAst two classes* We looked at the level combinations of the variables: ex: Student and grade; here there are more because sex and diet; many people of sex 1 on different diets, so there are more than 1 person in each categories.

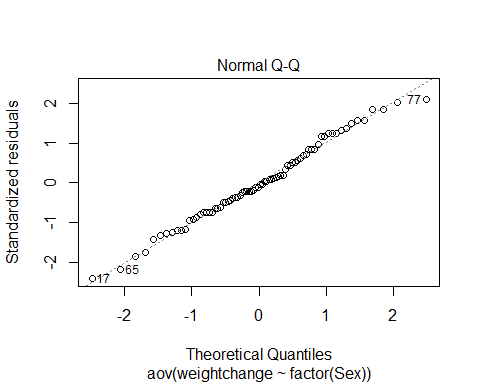
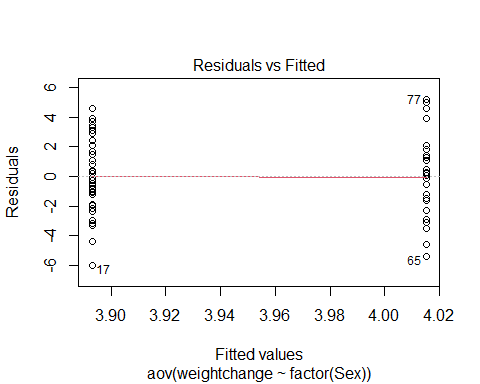
**HOw to tell if it is a good model?** - Look at teh summary to get an idea of how much variably is explained by this model compared to the total variablity(vs what is not being explained by the model is a better way to put it)

SSR, model 1: 71 w/ 430 left over in residual (unexplained) SSR, model 2: More left over unexplained, we are explaining very little of it by sex.

weightchange\_Sex <- aov(weightchange~factor(Sex), data = Diet)  
summary(weightchange\_Sex)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Sex) 1 0.3 0.278 0.044 0.835  
## Residuals 74 470.7 6.360   
## 2 observations deleted due to missingness

plot(weightchange\_Sex, 1:2)

 **Example: Exam Score** *We want to build a model with both Sex and the Diet* - We want to compare what is explained vs not explained - If there ar emore variables, then tehre will be less unexplained because they will explain some of the the differences that we see - When we build a bigger model, what you’re comparing (the variabilaity that is explained by that predictor too) is going to change

*We’ve shown (one-way):* *No significant differences between the exams* *Significant differences between the students*

**Question:** Can we use BOTH factors to help explain the variability in the exam scores?

**Repeat the models from last time** amodC is a different test now

**Simple Block Design**

**Define**: A simple block design has two factors with exactly one data value in each combination of the factors.

Assume: Factor A (Treatments) has K levels Factor B (Blocks) has J levels  n= K∙J data values

* Putting both model wants into one is a simpl block design; where we take 2 factors (Student and Exam) and we see which one is treatmetn or block and teh blocks (how we split them up) and treatment (apply different treatments)
* Treatmetns = levels (exams, student names, etc)
* We have one value for each person in our data

*thinking about what the model is doing* - Still thinking about the overall average of what is going one - we want to see how the different effects are on the model; what are the effects of it being this particualr student or this block that we are in (which exam or which student) - leaves us with what the differences are for each combination of the different levels of our treatment and block

**Two-way ANOVA: Main Effects Model** Y = mu + alphak + Bj+ E Y = GrandMean + effect for kth treatment + effect for jth block + Random Error

**Randomize Block - Calculations** 1. Find the mean for each treatment (row means), each block (column means), and grand mean. 2. Partition the SSTotal into three pieces: – SSTotal = SSA + SSB + SSE

SSTotal = sum(y-ybar)^2 = (n-1)s^2 SSA = row means SSB = column means SSE = SSTotal - SSA-SSB

**Randomized Block ANOVA Table** **Format:** Row(Header): [1,2],[1,3],[1,4],etc. Source: d.f, S.S., M.S, t.s., P-value Facotr A: K-1, SSA, SSA/(K-1), MSA/MSE Factor B: J-1, SSb, SSB/(J-1), MSB/MSE Error: (K-1)(J-1), SSE, SSE/(K-1)(J-1), [4,4], [4,5] Total: n-1, SSTotal

**Testing Two hypotheses** Ho: alpha1 = alpha 2=…= alphaK=0 Ha: Some alpha k != 0 Factor A: Difference in treatment means? Above, we are assuming that the factor a has no effect on the factor b

Ho: B1 = B 2=…= BJ=0 Ha: Some B j != 0 Facotr B: Difference in block means? THis second test is looking at the factor b and seeing if it has no effect on factor a; - we just swapped the two things we are testing, so we test both of the variables

* We get different results because the residuals left over will be smaller because we are combining to see which is being explained by each of these things at once

**Looking at the summary of amodA, B, and C** - WHen individually, there are no sig dif by exam, - When put together, there are sig difference

*Reason:* If you look at teh SSResid in teh Exam model, that is big, but if you look in the other it’s smaller because the var is being explained by Student - It’s not really testing the same thing; one is in a vaccum and one is in the model, but we are saying that if we are jsut looking at the unaccounted varibaility, then there are sig differences on the average scores by exam in the population.

*Conclusion* Evidence of a difference in means between exams. and Evidence of a difference in means between students.

amodA = aov(Grade~factor(Exam), data=Exams4)  
summary(amodA) #This SSR Residuals is bigger because there is alot that is not being explained by the predictor

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

amodB = aov(Grade~Student, data=Exams4)  
summary(amodB) #This SSR Res id smaller bc when we look at mean by student, a lot of the variability is explained; when we predict Bub's score with the other average that he did, we know more where it is coming from

## 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

#amodC = combines amodA and amodB  
amodC = aov(Grade~factor(Exam)+Student, data=Exams4)  
summary(amodC)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Exam) 3 1030 343.3 5.722 0.0114 \*   
## Student 4 4480 1120.0 18.667 4.35e-05 \*\*\*  
## Residuals 12 720 60.0   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Following Up with Tukey** - This output is a series of the mult comparision with every poss comb of factor A and then every combo of factor B; - there are sig dif for each factor

*Looking at output by exam* Exam 2 and 1 have a low p value and 3 and 2 have a low p value; those are where we are saying *THERE IS SIG DIF IN THE POPULATION*; if its practical is another thing

*Looking at output by student* : simialr results,that bud has a lower average, bud and bob, bill, betsy and bard are all sig and none of the other values are because we don’t have a lot of added explained variability by having the exams

by adding the exam it didn;’t really show wehre to see diferences, but it worked the other way

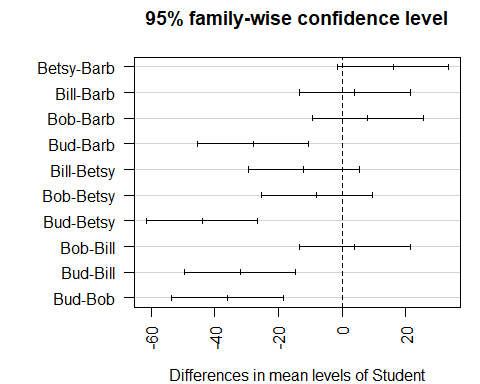
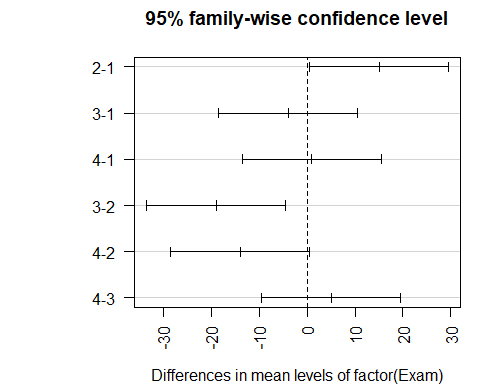
TukeyHSD(amodC)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Grade ~ factor(Exam) + Student, data = Exams4)  
##   
## $`factor(Exam)`  
## diff lwr upr p adj  
## 2-1 15 0.4554143 29.5445857 0.0425678  
## 3-1 -4 -18.5445857 10.5445857 0.8455825  
## 4-1 1 -13.5445857 15.5445857 0.9968212  
## 3-2 -19 -33.5445857 -4.4554143 0.0102092  
## 4-2 -14 -28.5445857 0.5445857 0.0605331  
## 4-3 5 -9.5445857 19.5445857 0.7409734  
##   
## $Student  
## diff lwr upr p adj  
## Betsy-Barb 16 -1.458285 33.458285 0.0782719  
## Bill-Barb 4 -13.458285 21.458285 0.9451946  
## Bob-Barb 8 -9.458285 25.458285 0.6039841  
## Bud-Barb -28 -45.458285 -10.541715 0.0019421  
## Bill-Betsy -12 -29.458285 5.458285 0.2467758  
## Bob-Betsy -8 -25.458285 9.458285 0.6039841  
## Bud-Betsy -44 -61.458285 -26.541715 0.0000293  
## Bob-Bill 4 -13.458285 21.458285 0.9451946  
## Bud-Bill -32 -49.458285 -14.541715 0.0006169  
## Bud-Bob -36 -53.458285 -18.541715 0.0002093

**Same visual from last time** - When they cross zero: means we predict there could be no difference betweent eh things - When they don’t cross zero: Visual to see wehre the differnces are that there could be a difference between tehse things

**We have been doing two way because we haev two predictors** *What about the conditions?*

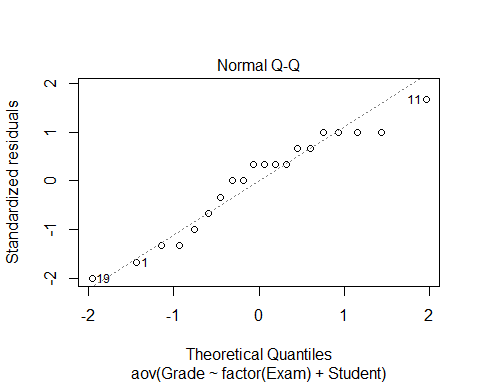
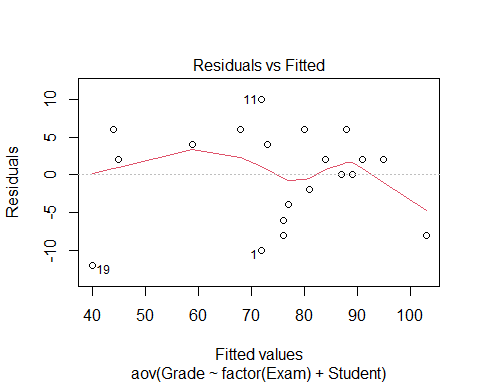
par(mar=c(4,7,3,1))  
hsd=TukeyHSD(amodC)  
plot(hsd,las=2)



par(mar=c(5,4,2,2))

*Looking at the conditions* - Constant variance: A little harder, but look at the fitted residuals, there are more columsn of data we haev to worry about here, overlal it looks good, but we can’t see much - NormalQQPlot of residuals : This is a small dataset, so nothing stands out as really bad.

plot(amodC, 1:2)

 **Looking back at the diet data** - GO in with a two way anova model Ho: Mean weights by diets are equal and themean weight changes are equal Ha: At least one is difference

Ho: Mean weight change by sex are equal Ho: Mean weight change by weight are not equal by sex

*Check COnditions* - None of them look sus

*Tukey tells us wehre changes are* - Tukey tells us that diets 3-1 and 3-2 are the different ones, - the dif bt sex, there ar eonly two here so we cant compare too mcuh - this test is giving us the same thing as before because its the same test

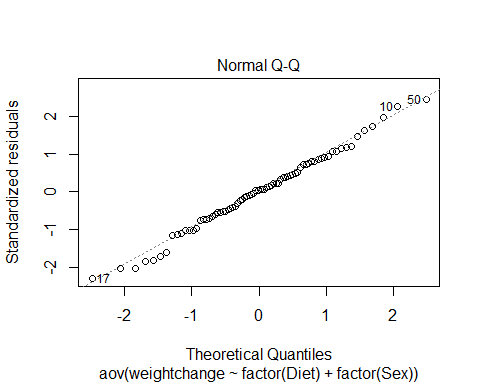
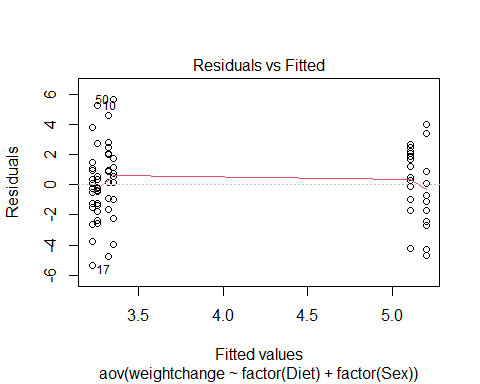
**Still need to think about: THe idea of interaction now** - We are saying that the diets there are sig differences there from weight changes and there are no sig difference by sex for weight change, but could there be interactions. I nother owords, woudl some diets work for some sexes thant he others and are there difference there?

If we look at the dif of weight change by sex and diet, could we explain more variabiltiy that’s left unexplained here.

weightchange\_both <- aov(weightchange~factor(Diet)+factor(Sex), data = Diet)  
summary(weightchange\_both)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 60.5 30.264 5.312 0.00705 \*\*  
## factor(Sex) 1 0.2 0.169 0.030 0.86387   
## Residuals 72 410.2 5.698   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

plot(weightchange\_both, 1:2)



TukeyHSD(weightchange\_both)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weightchange ~ factor(Diet) + factor(Sex), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 -0.032000 -1.6644380 1.600438 0.9987875  
## 3-1 1.848148 0.2455971 3.450699 0.0198145  
## 3-2 1.880148 0.2946561 3.465640 0.0160629  
##   
## $`factor(Sex)`  
## diff lwr upr p adj  
## 1-0 0.09502529 -1.006192 1.196243 0.8639058

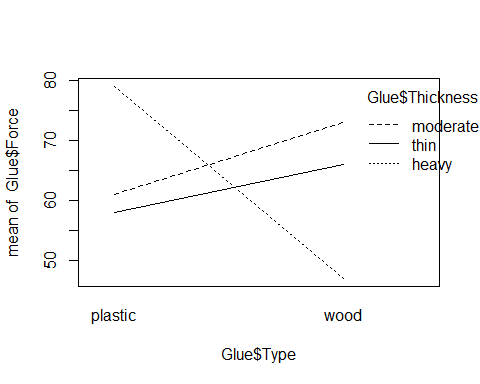
Glue\_model = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(Glue\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

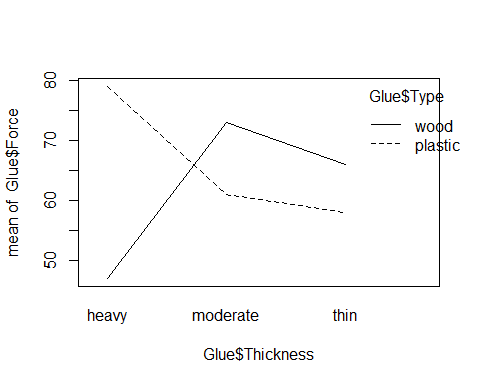
TukeyHSD(Glue\_model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## $Thickness  
## diff lwr upr p adj  
## moderate-heavy 4 -13.62589 21.62589 0.7743278  
## thin-heavy -1 -18.62589 16.62589 0.9834788  
## thin-moderate -5 -22.62589 12.62589 0.6768095  
##   
## $Type  
## diff lwr upr p adj  
## wood-plastic -4 -15.47703 7.477033 0.4264949  
##   
## $`Thickness:Type`  
## diff lwr upr p adj  
## moderate:plastic-heavy:plastic -18 -50.332428 14.3324278 0.3480798  
## thin:plastic-heavy:plastic -21 -53.332428 11.3324278 0.2331433  
## heavy:wood-heavy:plastic -32 -64.332428 0.3324278 0.0522177  
## moderate:wood-heavy:plastic -6 -38.332428 26.3324278 0.9689918  
## thin:wood-heavy:plastic -13 -45.332428 19.3324278 0.6259379  
## thin:plastic-moderate:plastic -3 -35.332428 29.3324278 0.9985907  
## heavy:wood-moderate:plastic -14 -46.332428 18.3324278 0.5637326  
## moderate:wood-moderate:plastic 12 -20.332428 44.3324278 0.6890028  
## thin:wood-moderate:plastic 5 -27.332428 37.3324278 0.9855637  
## heavy:wood-thin:plastic -11 -43.332428 21.3324278 0.7509689  
## moderate:wood-thin:plastic 15 -17.332428 47.3324278 0.5039318  
## thin:wood-thin:plastic 8 -24.332428 40.3324278 0.9074973  
## moderate:wood-heavy:wood 26 -6.332428 58.3324278 0.1170797  
## thin:wood-heavy:wood 19 -13.332428 51.3324278 0.3052653  
## thin:wood-moderate:wood -7 -39.332428 25.3324278 0.9431857

interaction.plot(Glue$Type, Glue$Thickness, Glue$Force)



interaction.plot(Glue$Thickness, Glue$Type, Glue$Force)



## STOR 455 - Class 37 – ANCOVA and Homogeneity of Variances

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

**What’s an Interaction Effect?** **DEFINE:** An interaction effect occurs when a significant difference is present at a specific combination of factors.

*Example* Y=GPA Factor A = Year in School (FY, So, Jr, Se) Factor B = Major (Psych, Bio, Math)

FY is hard if alphaa1 < o (main effect) Bio is easy if B2 >0 (main effect) Jr in Math is hard if Y33 < 0 (interaction effect)

There may be times where one factor by itself are not sig; there may not be a sig diff by year or major; we may see for certain years for certain majors that is differnet

**Factorial Design** *In the past:* We hae only needed one data point that looked at each combo of the factor levels; for teh exam and student we needed one score for exam and student; here we can expand that if we only have C different values

*Future* Assume: Factor A has K levels, Factor B has J levels.

To estimate an interaction effect, we need more than one data value in each combination of factors.

Let nkj = sample size in (k,j)th cell

nkj = c

c = 1 = randomized block design c> 1 = blanaced factorial design

**Example: Glue Strength** Factor A: Thickness (thin, moderate, heavy) Factor B: Glue Type (plastic, wood) Factor B: Glue Type (plastic, wood)

Data: Plastic, Wood Thin: 52 64, 72 60 Moderate: 67 55, 78 68 Heavy: 86 72, 43 51

K = 3 j = 2 c = 2 n = 12

*Notes about the data* - They are in newtons - We could be using plastic or wood glue - How much glue we use is the other factor (Thin, Moderate, Heavy)

*Question of interest* Is there a dif in teh force reuired depending on the type of glue I use and/or a difference on how much glue I use? Is there an interaction? Is there some level of a combination that looks really different than the others and doesn’t follow the trend?

*Looking at the data* The thin and moderate force required, wood is higher until you get to heavy; a heavy amount of wood glue doesn’t seem to be as strong

*We’ll see:* there is some interaction going on between teh wood glue amount and the type of glue

# Want to try and predict the force requried to pull these things apart by the thickness and type of glue in the data and also that interaction  
Glue\_model = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(Glue\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This gives more information than we wanted because it gives the first line as a test by thickness   
# When we look at the itneraction, there is a sig level there; the SSQ for this is higher; Factor A and B are not sig on their own, but the interation is

**Two-way ANOVA Table (with interaction)** Header: Source, df, SS, MS, ts, p-value Row1: FactorA, K-1, SSA, SSA/(K-1), MSA/MSE, BLANK Row2: FactorB, J-1, SSB, SSB/(J-1), MSB/MSE, BLANK Row3: AxB, (K-1)(J-1), SSAB, SSAB/df, MSAB/MSE, BLANK Row4: Error, JK(n-1), SSE, SE/df, BLANK, BLANK Row5: total, n-1, SSY, BLANK———>

Ho: All alphak = 0 Ha: Some alphak != 0

Ho: All Bj = 0 Ha: Some Bj != 0

Ho: All ykj = 0 Ha: Some ykj != 0

*In the table above, we do not have evidence to say that there is impact from the glue type of thickness. We don’t have evidence to reject the null, but in the third case, we do have evidence that there is a sig interaction between these variables*

TukeyHSD(Glue\_model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## $Thickness  
## diff lwr upr p adj  
## moderate-heavy 4 -13.62589 21.62589 0.7743278  
## thin-heavy -1 -18.62589 16.62589 0.9834788  
## thin-moderate -5 -22.62589 12.62589 0.6768095  
##   
## $Type  
## diff lwr upr p adj  
## wood-plastic -4 -15.47703 7.477033 0.4264949  
##   
## $`Thickness:Type`  
## diff lwr upr p adj  
## moderate:plastic-heavy:plastic -18 -50.332428 14.3324278 0.3480798  
## thin:plastic-heavy:plastic -21 -53.332428 11.3324278 0.2331433  
## heavy:wood-heavy:plastic -32 -64.332428 0.3324278 0.0522177  
## moderate:wood-heavy:plastic -6 -38.332428 26.3324278 0.9689918  
## thin:wood-heavy:plastic -13 -45.332428 19.3324278 0.6259379  
## thin:plastic-moderate:plastic -3 -35.332428 29.3324278 0.9985907  
## heavy:wood-moderate:plastic -14 -46.332428 18.3324278 0.5637326  
## moderate:wood-moderate:plastic 12 -20.332428 44.3324278 0.6890028  
## thin:wood-moderate:plastic 5 -27.332428 37.3324278 0.9855637  
## heavy:wood-thin:plastic -11 -43.332428 21.3324278 0.7509689  
## moderate:wood-thin:plastic 15 -17.332428 47.3324278 0.5039318  
## thin:wood-thin:plastic 8 -24.332428 40.3324278 0.9074973  
## moderate:wood-heavy:wood 26 -6.332428 58.3324278 0.1170797  
## thin:wood-heavy:wood 19 -13.332428 51.3324278 0.3052653  
## thin:wood-moderate:wood -7 -39.332428 25.3324278 0.9431857

**Interpreting Interaction** IF the two-way ANOVA indicates a significant interaction, plot the cell means vs. one factor with separate lines/symbols for the second factor

Called a Cell means plot or INteraction plot

Cell Means: Plastic, Wood Thin: 58.0, 66.0 Moderate: 61.0, 73.0 Heavy: 49.0, 47.0

interaction.plot(FactorA,FactorB,Response)

*Class NOtes* - Will look at teh itneractions between teh groups and plot them by the force required along one axis if plastic or wood, then other; see the difference through this tabl instead of having two values in each cell, they’re just averaged - why don’t we make this interaction plot factor A and factor B reponse, so type thickness force, but if we flip those teh axis will flip

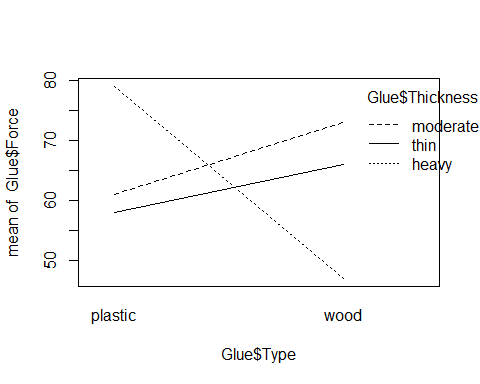
*See below for example on the flipping axises*

**Plot 1: interaction.plot(GlueThickness, Glue$Force)** - Looking at plastic and wood, then the force of the means for those - each one of the lines is the differnce amount of thickness, so wherever the poitns are that’s just the mean value - under 60, 62 or so and a bit higher. - we are looking at *are the trends as we go from plastic to wood similar for each level of thickness* – If they are similar, we expect lines t be similar in slope to eachother **The size of the data has a big impact here** - Small data is going to be skewed faster because it’s mean, so if there’s something out of wack, it’s going to show more easily since mean isn’t insoluated from outlies

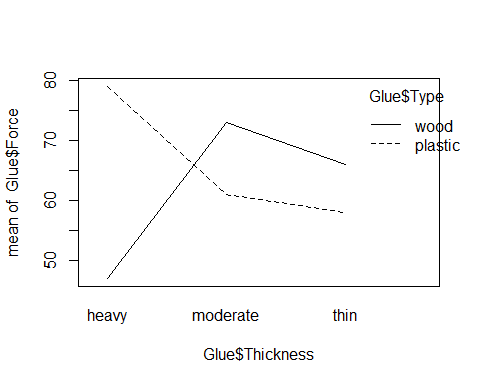
* The heavy amount of thickness is very different *There is some kind of interaction betcause heavy has a different interaction thatn the other lines*

**Plot 2: interaction.plot(GlueType, Glue$Force)** - Same relationship from a different perspective - Depending on what you’re looking at, it might look better or worse - Moderate to thin: – Those lines look similar in solpe for wood and plastic - - Heavy - Moderate the slope is different *Doesn’t show, but implied* - Going from Heavy to thin, it would be the very different relationship - Heavy to thin for one and thin to heavy to the other *Sometimes you have to assume there is no interaction to see if there is a relationship in the data*

interaction.plot(Glue$Type, Glue$Thickness, Glue$Force)



interaction.plot(Glue$Thickness, Glue$Type, Glue$Force)

 **Looking back at the Diet Data** *Make a 2-way anova model with interaction* - We made a thing by diet there were sig dif in the weight loss by which diet you were on - No sig dif by sex, - Made a two-way model to see if dif by diet and sex when in the same model, found similar results **One step further: 2-way with the interaction** *Question Interest: Are tehre any sig dif in weight loss by diet depending on which sex you are?* - Look at residuals and make these interaction plots to see where the sig difs are

*Diet Data* - Looking at the table, we have 3 different Hypo tests here

* Assuming there’s no effec ton the average weight change by diet, vs Ha: There is an effect; small p-value; – Type of diet shows sig there
* Sex has a high p-value, Ho: Assume there is no effect on average weight loss based on sex Ha: There is an effect on average weight loss based on sex

P-vale Conc: We dont have evidence to reject that

Interaction is below 0.05 level, so we are seeing a sig interaction between sex and diet, so tehre is some effect between what sex you are and which diet you are on

*Tukey* - It’s giving us 3 different outputs ’ 1. Dif by diet for all poss combos; 2. dif by sex for all combos; and 3. dif for interactions

The only real difference is in one situation; all other pvalues are huge, except for 3-0 and 1-0 there was a big difference.

**Question: What is 3-0 and what is 1-0?** *Anwser: Zero is for women, and 3 = diet 3 and 1 = diet 1; there is sig dif in teh 3rd and 1st diet for women, but we are not seeing this sig dif for men*

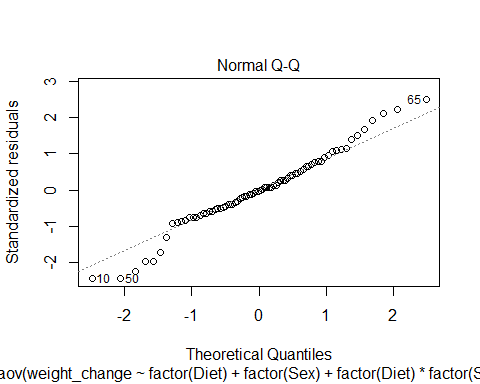
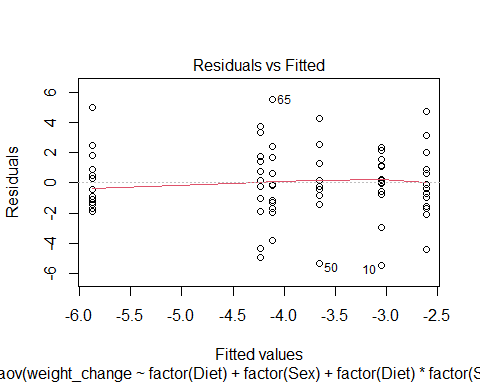
**Look at Resiudal analyss** - Nothing really stands out - Check for constant variance, they are spread similarly; one group is a little more compact, so may be an issue there, but not too bad - Normaility by QQPlot, some tail issues, but not too bad

**Interaction plot** - INteraction plot: sex on the horizontal axis: – For diet 3, it looks like we have a dif relationship than for diet 1 for each of teh sexes

Diet$weight\_change <- Diet$weight6weeks - Diet$Preweight  
  
mod1 <- aov(weight\_change ~   
 factor(Diet)+   
 factor(Sex)+  
 factor(Diet)\*factor(Sex),   
 data = Diet)  
  
summary(mod1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 60.5 30.264 5.629 0.00541 \*\*  
## factor(Sex) 1 0.2 0.169 0.031 0.85991   
## factor(Diet):factor(Sex) 2 33.9 16.952 3.153 0.04884 \*   
## Residuals 70 376.3 5.376   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

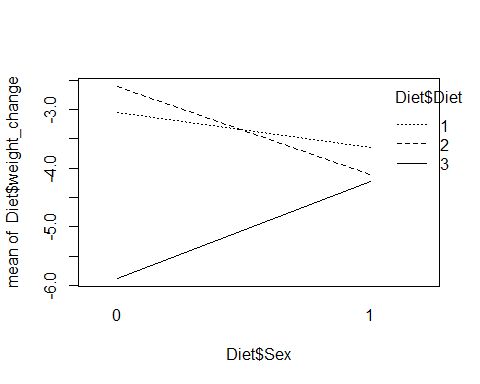
plot(mod1, 1:2)



TukeyHSD(mod1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weight\_change ~ factor(Diet) + factor(Sex) + factor(Diet) \* factor(Sex), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 0.032000 -1.554658 1.6186577 0.9987150  
## 3-1 -1.848148 -3.405757 -0.2905393 0.0160200  
## 3-2 -1.880148 -3.421176 -0.3391199 0.0128543  
##   
## $`factor(Sex)`  
## diff lwr upr p adj  
## 1-0 -0.09502529 -1.16524 0.9751891 0.8599508  
##   
## $`factor(Diet):factor(Sex)`  
## diff lwr upr p adj  
## 2:0-1:0 0.4428571 -2.1250148 3.0107291 0.9958151  
## 3:0-1:0 -2.8300000 -5.3547114 -0.3052886 0.0191170  
## 1:1-1:0 -0.6000000 -3.4129628 2.2129628 0.9887997  
## 2:1-1:0 -1.0590909 -3.7964516 1.6782698 0.8656520  
## 3:1-1:0 -1.1833333 -3.8560592 1.4893925 0.7855223  
## 3:0-2:0 -3.2728571 -5.7975685 -0.7481458 0.0040103  
## 1:1-2:0 -1.0428571 -3.8558199 1.7701056 0.8852416  
## 2:1-2:0 -1.5019481 -4.2393087 1.2354126 0.5963201  
## 3:1-2:0 -1.6261905 -4.2989163 1.0465354 0.4833188  
## 1:1-3:0 2.2300000 -0.5436187 5.0036187 0.1863470  
## 2:1-3:0 1.7709091 -0.9260048 4.4678230 0.3965102  
## 3:1-3:0 1.6466667 -0.9846191 4.2779524 0.4513580  
## 2:1-1:1 -0.4590909 -3.4275816 2.5093998 0.9975014  
## 3:1-1:1 -0.5833333 -3.4923292 2.3256625 0.9915569  
## 3:1-2:1 -0.1242424 -2.9601974 2.7117126 0.9999949

interaction.plot(Diet$Sex, Diet$Diet, Diet$weight\_change)



**ANOVA via Dummy Regression** *Recall: For a single categorical factor with K levels* 1. Create K-1 indicator (dummy) predictors 2. Run regression with the dummy predictors 3. Constant estimates the mean of the reference group 4. Coefficients estimate how each other group differs 5. ANOVA tables match (depending…)

*NOtes* - We are really just looking at regression - If we wanted tot hink of these as regression models, we could do that, we just have to combine some ideas we are looking at *ANOVA MODEL FOR DIFFERENC EIN MEANS: ONE-WAY MODEL* - That’s the same as a regression model that is useing K-1 predictor variables, and each of those coeff would be some measure of how different that group is from the overall mean

*IF look at glue data, see below:* - For a math standpoint, the below models are the same, the output is just different - There is no sig evidence by dif by just thickness - we did have it in terms of an interaction - If we did this for 1,2, 2w/interaction we can build these eitherway with a linear or aov model and we are building teh same math thing

**How is this useful?** - Because this is all just regression, we can see some slightly different relationships and think about this kind of analysis of difference in means after considering teh variability of other quanatitive variables as well in the model, not just categorical - So far we have looked at how much variability is explained by the type of glue vs some other things - There might be quanatitive var we want to compare with as well

mod1 = aov(Force ~ Thickness, data=Glue)  
summary(mod1)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.155 0.859  
## Residuals 9 1628 180.9

mod2 = lm(Force ~ Thickness, data=Glue)  
anova(mod2)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.00 0.1548 0.8588  
## Residuals 9 1628 180.89

summary(mod2)

##   
## Call:  
## lm(formula = Force ~ Thickness, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.00 -10.50 0.50 9.25 23.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.000 6.725 9.368 6.14e-06 \*\*\*  
## Thicknessmoderate 4.000 9.510 0.421 0.684   
## Thicknessthin -1.000 9.510 -0.105 0.919   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.45 on 9 degrees of freedom  
## Multiple R-squared: 0.03325, Adjusted R-squared: -0.1816   
## F-statistic: 0.1548 on 2 and 9 DF, p-value: 0.8588

mod3 = aov(Force ~ Thickness + Type, data=Glue)  
summary(mod3)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.142 0.870  
## Type 1 48 48.0 0.243 0.635  
## Residuals 8 1580 197.5

mod4 = lm(Force ~ Thickness + Type, data=Glue)  
anova(mod4)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.1418 0.8700  
## Type 1 48 48.0 0.2430 0.6353  
## Residuals 8 1580 197.5

summary(mod4)

##   
## Call:  
## lm(formula = Force ~ Thickness + Type, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.00 -10.50 0.00 8.25 21.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.000 8.114 8.011 4.32e-05 \*\*\*  
## Thicknessmoderate 4.000 9.937 0.403 0.698   
## Thicknessthin -1.000 9.937 -0.101 0.922   
## Typewood -4.000 8.114 -0.493 0.635   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.05 on 8 degrees of freedom  
## Multiple R-squared: 0.06176, Adjusted R-squared: -0.2901   
## F-statistic: 0.1755 on 3 and 8 DF, p-value: 0.91

mod5 = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(mod5)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod6 = lm(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
anova(mod6)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.4242 0.67247   
## Type 1 48 48 0.7273 0.42649   
## Thickness:Type 2 1184 592 8.9697 0.01574 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(mod6)

##   
## Call:  
## lm(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7 -6 0 6 7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 79.000 5.745 13.752 9.19e-06 \*\*\*  
## Thicknessmoderate -18.000 8.124 -2.216 0.06861 .   
## Thicknessthin -21.000 8.124 -2.585 0.04149 \*   
## Typewood -32.000 8.124 -3.939 0.00763 \*\*   
## Thicknessmoderate:Typewood 44.000 11.489 3.830 0.00866 \*\*   
## Thicknessthin:Typewood 40.000 11.489 3.482 0.01312 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.124 on 6 degrees of freedom  
## Multiple R-squared: 0.7648, Adjusted R-squared: 0.5689   
## F-statistic: 3.903 on 5 and 6 DF, p-value: 0.06386

*Go back to pulse data* - We looked at: How does exercise predict what your active heart rate was - Do we think the aveg heart rate is dif by exercise elvel

Ho: Do I see evidence of a difference? Ha: Do I not see evidence of a difference?

*Check the conditions for a linear model*

*ANOVA Table* - Tells us there is sig dif in teh average active heart rate based on what your exercise level is

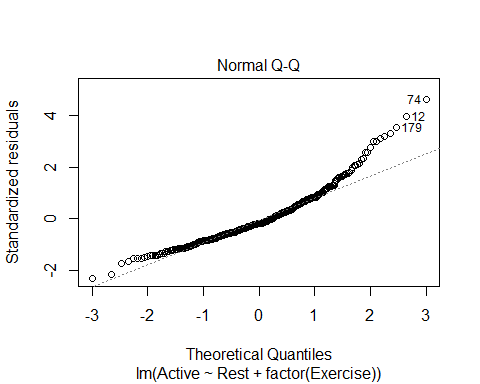
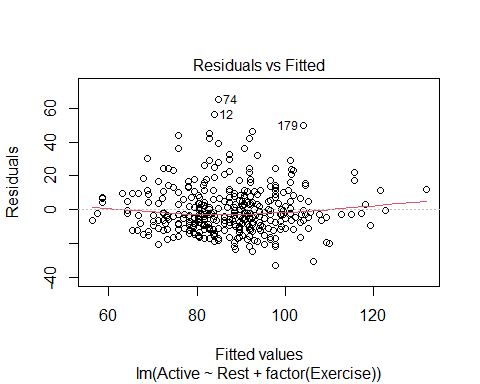
modp2 = lm(Active ~ Rest + factor(Exercise), data=Pulse)  
  
summary(modp2)

##   
## 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

anova(modp2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.0302 <2e-16 \*\*\*  
## factor(Exercise) 2 351 175 0.8714 0.4192   
## Residuals 371 74699 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(modp2, 1:2)



**Main Effects Two-way ANOVA via Dummy Regression** 1. Create indicator predictors for each factor 2. Run regression with the dummy predictors (leaving out one for each factor) 3. How to interpret the coefficients? 4. How to “recover” the two-way ANOVA table?

*NOtes* - Want to think about something a little differently *Do I still show these diferences after considering teh Varaiability explained by some over variable?* Is this sig AFTER looking at something else? - Look at the covar of the resting heart rate - We have that exercise level shows sig dif in resting heart rate on average; but does it show that after we consider the realtionship between your resting and active heart rate? *AKA, if we tell R that this is already a connection, will it still see a connection? I think that’s what this is saying*

modp2 <- lm(Active~Rest+factor(Exercise), data = Pulse)  
  
summary(modp2)

##   
## 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

# People who exercise a small amount, we can get rid of the last two lines and it tells us our realtions bt rest and active heart reate follows that intercept and lope   
  
anova(modp2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.0302 <2e-16 \*\*\*  
## factor(Exercise) 2 351 175 0.8714 0.4192   
## Residuals 371 74699 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

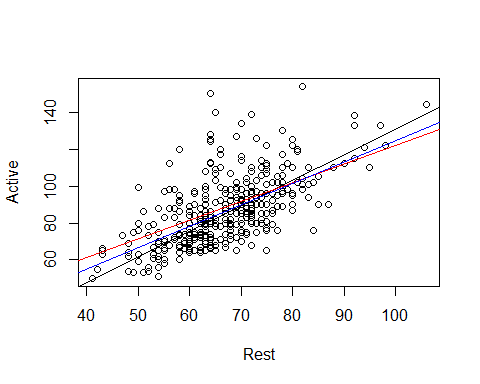
# How the active heart rate is adjected by the intercept   
# We are assuming that the resting heart rate follows teh same trend as the active, regardless of your exercise level

*Is it reasonable to make the assumption that resting heart rate follows the same trend as the active?* - What kind of relationships are there in the data if I subset by people who exercise a little, a moderate amount, and a lot?

**Two-way ANOVA w/Interaction via Dummy Regression** 1. Create indicator predictors for each factor 2. Run regression with the dummy predictors (leaving out one for each factor) 3. *To include interaction:* Use products of the (included) dummies. 4. How to interpret the coefficients? 5. How to “recover” the three ANOVA components?

*Reasonable to assume there is no interaction here. See the comments on the code below*

Exercise1 = subset(Pulse, Pulse$Exercise == 1)  
Exercise2 = subset(Pulse, Pulse$Exercise == 2)  
Exercise3 = subset(Pulse, Pulse$Exercise == 3)  
  
lme1 = lm(Active ~ Rest, data = Exercise1)  
lme2 = lm(Active ~ Rest, data = Exercise2)  
lme3 = lm(Active ~ Rest, data = Exercise3)  
  
plot(Active~Rest, data=Pulse)  
abline(lme1)  
abline(lme2, col='red')  
abline(lme3, col='blue')



# When we look at the three lines over the data, we are hoping that these will be similar in slope. IF they are much different in terms of slope, from a practical standpoint, there may be an interaction between these that we need to investigate.   
  
##Looking at the data, we are trying to see if they are similar enough with their slope, none look drastically different

*What if there are some issues? Lets try the above with a transformation*

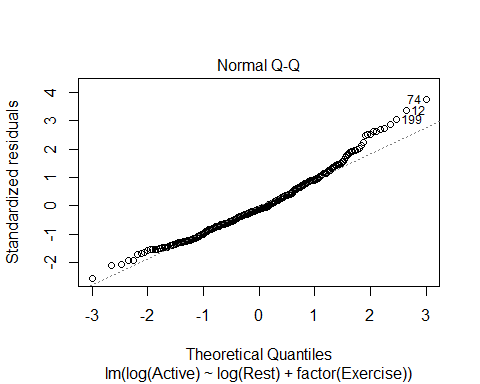
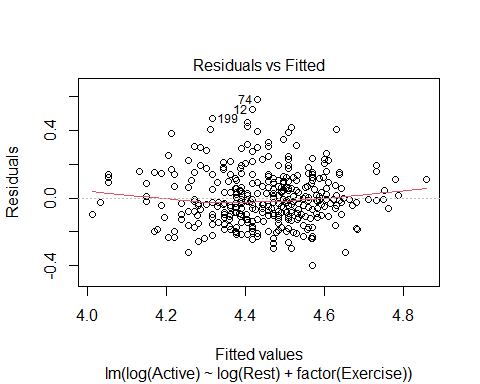
modp2log = lm(log(Active)~log(Rest) + factor(Exercise), data = Pulse)  
  
anova(modp2log)

## Analysis of Variance Table  
##   
## Response: log(Active)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(Rest) 1 6.7174 6.7174 277.991 <2e-16 \*\*\*  
## factor(Exercise) 2 0.0499 0.0250 1.033 0.3569   
## Residuals 371 8.9648 0.0242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

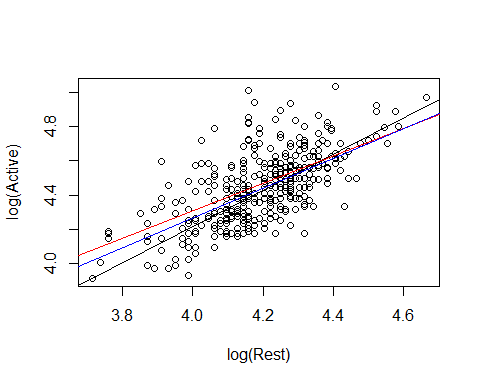
*Looking at Residuals* - There is some curve that is not ideal - Look at teh QQNorm, there are some big issues with that one tail on the untransformed thing - With the transformed data: the tail on QQNorm isn’t as bad and the little curve on the residual plot isn’t as bad.

*We also have to check and make sure that there is no interaction now that we have introduced the transformation* - Redo what you did before, but all in logs, See code below:

plot(modp2log, 1:2)



lme1.log = lm(log(Active) ~ log(Rest), data=Exercise1)  
lme2.log = lm(log(Active) ~ log(Rest), data=Exercise2)  
lme3.log = lm(log(Active) ~ log(Rest), data=Exercise3)  
  
plot(log(Active) ~ log(Rest), data=Pulse)  
abline(lme1.log)  
abline(lme2.log, col='red')  
abline(lme3.log, col='blue')



# Doesn't show anything idfferent than before   
# IT helped with teh conditions, but it doesn't effect the interaction to consider

**WHy do we do this?** - We want ot know for a hypo test: Is there a dif in teh active heart rates of people at each of the three exercise levels on average

Ho: Mean grousp are qual Ha: At least one mean group is difference

After accounting for the variability in their active heart rate; so basically doing a nested test to see if exercise variabile signifigant after considering active heart rate?

*look at ANOVA* FOcus on secon d line, We are thinking about after the var is explained in active heart rate by rest, do we see sig dif in the mean active heart rate by exercise level?

The pvalue is high, so we dont see a sig dif

anova(modp2log)

## Analysis of Variance Table  
##   
## Response: log(Active)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(Rest) 1 6.7174 6.7174 277.991 <2e-16 \*\*\*  
## factor(Exercise) 2 0.0499 0.0250 1.033 0.3569   
## Residuals 371 8.9648 0.0242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*How does this relate to the homework?* - Last questionon teh HW, are there dif in the mean price of the 6 models of car after taking into account another varibale?

* Account for the variability in mileage and then see if your car has sig difference in prices
* By chance you might have had cars of a certain model with high or low mileage
* if one car has only been out for a few years, then it might not have high mileage; so it might not be as useful.
* If you have cars that are different ages, then you might have a wide spread with some cars vs the other, then you can see after you take into account the variability ofX affects the price of car, are there any differences left by the model or is X really explaining that difference?

**Analysis of Covariance (ANCOVA)** **Basic idea:** If we can use dummy predictors to convert an ANOVA for means into a regression model, why not also include quantitative predictors?

**Analysis of Covariance (ANCOVA)** WeightLoss = β0 + β1Age + β2Height + β3Diet1 + β4Diet2 + ε Age and Height = Covariates Diet1 and Diet2 = Factor

* If want to see if sig dif by diet after taking into account other variables, we cna see that as well; see below
* We dont know how they chose the groups for these people, so we don’t know if its skewed. This will keep that in mind.
* If they didn’t, we could test to see if someone’s age explains how much weight they loss and if there is a sig diff bt the diet they were on after considering their age or height

Diet$weightchange = Diet$weight6weeks - Diet$Preweight  
  
mod7 = lm(weightchange ~ factor(Diet), data=Diet)  
anova(mod7)

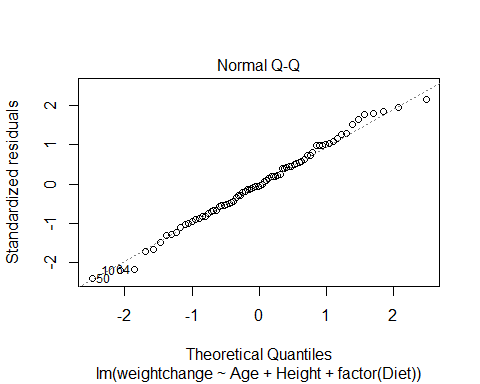
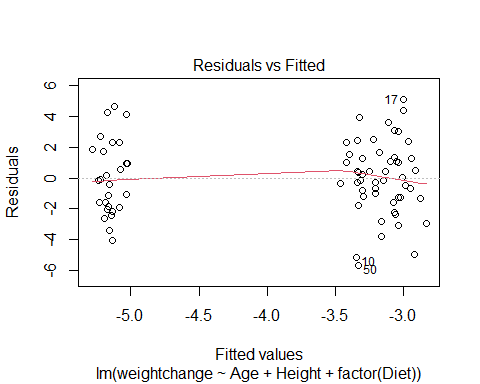
## Analysis of Variance Table  
##   
## Response: weightchange  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 71.09 35.547 6.1974 0.003229 \*\*  
## Residuals 75 430.18 5.736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**What does this tell us?** - After explaining the variabilit yin weight change by age and height, neither were sig, we still saw that there were sig dif bt the weight change by which diet you were on - There could be other variables that did not show this and owuld not show any difference diet after thinking about that.

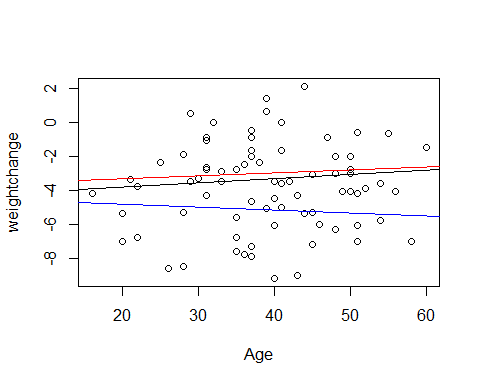
mod8 = lm(weightchange ~ Age + Height + factor(Diet), data=Diet)  
anova(mod8)

## Analysis of Variance Table  
##   
## Response: weightchange  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 1.69 1.689 0.2870 0.593795   
## Height 1 6.23 6.227 1.0580 0.307068   
## factor(Diet) 2 63.68 31.838 5.4092 0.006448 \*\*  
## Residuals 73 429.68 5.886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

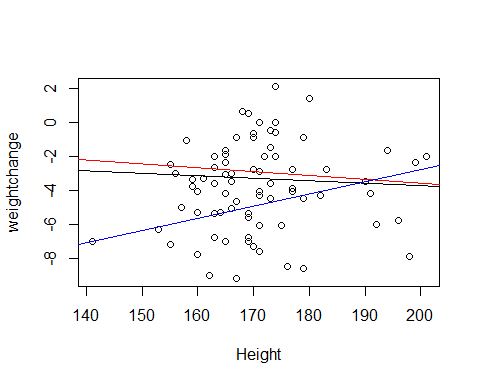
plot(mod8, 1:2)



Diet1 = subset(Diet, Diet==1)  
Diet2 = subset(Diet, Diet==2)  
Diet3 = subset(Diet, Diet==3)  
  
lme1.diet = lm(weightchange ~ Age, data=Diet1)  
lme2.diet = lm(weightchange ~ Age, data=Diet2)  
lme3.diet = lm(weightchange ~ Age, data=Diet3)  
  
plot(weightchange ~ Age, data=Diet)  
abline(lme1.diet)  
abline(lme2.diet, col='red')  
abline(lme3.diet, col='blue')



lme4.diet = lm(weightchange ~ Height, data=Diet1)  
lme5.diet = lm(weightchange ~ Height, data=Diet2)  
lme6.diet = lm(weightchange ~ Height, data=Diet3)  
  
plot(weightchange ~ Height, data=Diet)  
abline(lme4.diet)  
abline(lme5.diet, col='red')  
abline(lme6.diet, col='blue')



## STOR 455 Class 38 Homogeneity of Variances and Contrasts

library(readr)  
library(car)  
library(Stat2Data)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")

*Goal: Trying to see if there is a dif in means bt groups* - Look how we can test variability between means

**ANOVA for Difference in K Means** Data: Samples from K different groups Summary statistics: n1, ybar1, s1 for each group Combine all

Test: Ho: mu1 = mu2 = muk Ha: Some mui != muj

**Checking Conditions for ANOVA** E~N(0, Std of Errors) Check with residuals

* Zero Mean: Always holds for sample residuals
* COnstant Variance: Plot residuals vs. fits and/or compare std. dev.’s of groups (Check if some group si is more than twice another)
* Normaility: Histogram/normal plot of residuals
* Independence:Pay attention to data collection

**Example: Five Students** *Question of intereset: Is there a significant difference in average grade among the five students?* - Test tells us if there is sig, but it doesn’t always mean useful - Found there are sig dif by student, and the variability are suspect to assume there weas constant variance here (Betsy’s scores were all sim to each other while Bud was all close) - Laready groups so that the first student is grouped with their scores, etc, etc

**ANOVA for Grades vs. Students** *Can we assume equal variances?*

tapply(Exams4$Grade, Exams4$Student, median)

## Barb Betsy Bill Bob Bud   
## 75.5 91.5 77.5 82.5 48.5

medians = rep(tapply(Exams4$Grade, Exams4$Student, median), each=4)  
medians # Median scores for each student; some skew, which makes it different than means

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 75.5 75.5 75.5 75.5 91.5 91.5 91.5 91.5 77.5 77.5 77.5 77.5 82.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 82.5 82.5 82.5 48.5 48.5 48.5 48.5

# Want to keep track of not comparing this with an anova analysis   
# Want to see how far away each score is from the median value   
#Analysis of the variances of these absolute devations

* Need to take the abso score - media to make a new array with all the values
* It’s going to work nicely here, but it’s not going to look unless we make the data look like this data

**Levene’s Test for Equality of Variances**

Test: Ho: variance1 = variance2 =…=variancek H1: some variancei!=variancej

(y-ybar) = (ybark-ybar) + (y-ybark) where (y-ybark) conducts analysis of variance on teh collection of absolute deviations abs(y-ytildek); where ytildek = median of group k

LeveneData = abs(Exams4$Grade - medians)  
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student)  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

leveneTest(Exams4$Grade, factor(Exams4$Student))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Levene’s Test for Grades versus Students** Ho: variance(barb) = variance2 = variance3 = etc Ha: variancei != variancej; for at least one pair of students (i,j)

*Conclusion*: There is not statistically significant evidence to suggest that students’ grades have difference variances.

medians=rep(tapply(Exams4$Grade, Exams4$Student,median),each=4) # Each = 4 because we want barbs 4 times, betsy 4 times, etc.   
LeveneData=abs(Exams4$Grade-medians) #They are lined up perfectly right in this dataset; so all the grades will be subed by each student's median score.   
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student) # Do we see any sig dif in the subbed median by actual score?   
#Testing: Build model assume the var are equal in each group and trying to see if we have edv that there are some dif here; same as like anova model, but at var   
# most of the sum of squares are in resdu, which assum there are equal variance; little are in sumsquares of the model itself (grand variance)   
# Go in with teh same ideas, but with variances and assume that htey are equal, then the p value is big, we dont have evidence to support Ha; even though the rule of variance not being double another, but the test it didnt show sign  
# Comes down to sample size; we would have a lot of dif to have a sig dif   
# tehere is dif here, but its not big enough to reject Ho  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

library(car)   
leveneTest(Exams4$Grade, Exams4$Student) # It wants grade and student (what we are splitting grade up with); it shouldnt really matter if you facotr it

## Warning in leveneTest.default(Exams4$Grade, Exams4$Student): Exams4$Student  
## coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Cancer Survival with Ascorbate Supplement**  - Use this when you have issues with constant variance and how to use transofmraitons to help with that

**INformation about the data:** *In the 1970’s doctors wondered if giving terminal cancer patients a supplement of ascorbate would prolong their lives. They designed an experiment to compare cancer patients who received ascorbate to cancer patients who did not receive the supplement. The result of that experiment was that, in fact, ascorbate did seem to prolong the lives of these patients. But then a second question arose. Was the effect of the ascorbate different when different organs were affected by the cancer? The researchers took a second look at the data. This time they concentrated only on those patients who received the ascorbate and divided the data up by which organ was affected by the cancer. They had 5 different organs represented among the patients (all of whom only had one organ affected): Stomach, bronchus, colon, ovary, and breast.*

* Looking at treatmetn for cancer; is the outcome dife for the type of cancer the people have?

data("CancerSurvival")  
head(CancerSurvival)

## Survival Organ  
## 1 124 Stomach  
## 2 42 Stomach  
## 3 25 Stomach  
## 4 45 Stomach  
## 5 412 Stomach  
## 6 51 Stomach

**Cancer Survival with Ascorbate Supplement** Test:

Ho: mu1 = mu2 = …= muk Ha: Some mui != muj

* Want to see if the avg # days are equal depn on type of cancer or if no dif in treatment vs cancer type

Cancermod = aov(Survival~Organ, data=CancerSurvival)  
summary(Cancermod) # Sig evid to say surivial time changes depending on the organ that has cancer

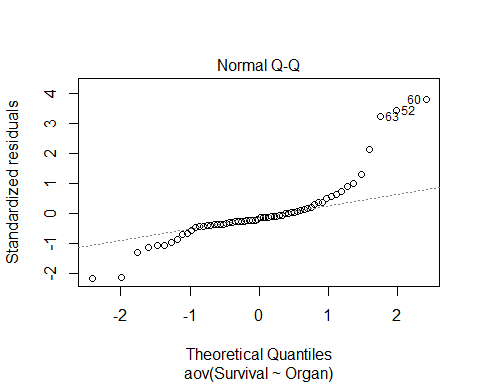
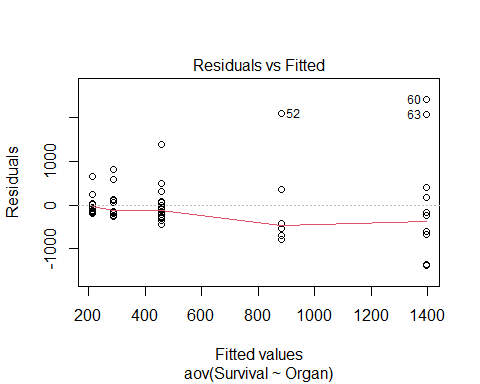
## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 11535761 2883940 6.433 0.000229 \*\*\*  
## Residuals 59 26448144 448274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# THIS IS ASSUME CONDITIONS ARE MET

**Cancer Survival with Ascorbate Supplement**  *Check conditions* - See there is a trend in variability increases - The organ that has the lowest survial rate = more compact - as survival rate for organ increase as - clear pattern in plot

* Normal resiudal s
* off because the tails are wonky

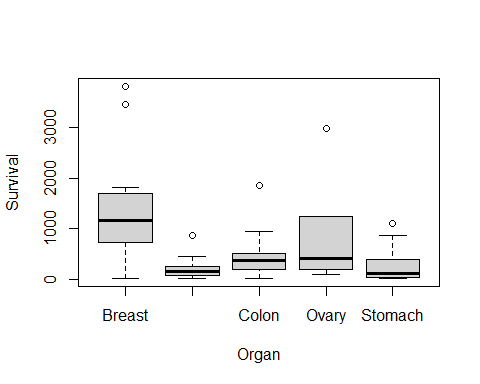
plot(Cancermod, 1:2)



**Cancer Survival with Ascorbate Supplement**  Ho: variance1 = variance2=…=variancek Ha: Some variancei != variancej

* Looking at visuals of survival by organ

boxplot(Survival~Organ, data=CancerSurvival)



tapply(CancerSurvival$Survival, CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1238.9667 209.8586 427.1686 1098.5788 346.3096

leveneTest(CancerSurvival$Survival, CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 4 4.4524 0.003271 \*\*  
## 59   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

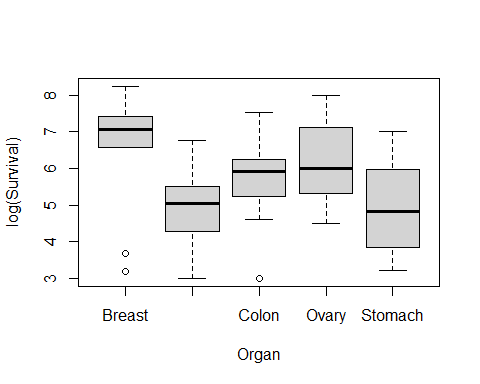
# Constant variance is a huge issue here   
  
#Levenetest gives a really small pvalue

* are these sim by chance or
* we have evid to say that the var is dif than the others
* we cant really use the results because the conditions aren’t good

**Would a transformation help?** - Yes, if you can make it look more normal then you will be good

LeveneTest Ho: variance1=varaince2=…varicnek Ha: some variancei != variancej

boxplot(log(Survival)~Organ, data=CancerSurvival)



tapply(log(CancerSurvival$Survival), CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1.6477550 0.9534041 0.9974766 1.2569313 1.2502073

leveneTest(log(CancerSurvival$Survival), CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.6685 0.6164  
## 59

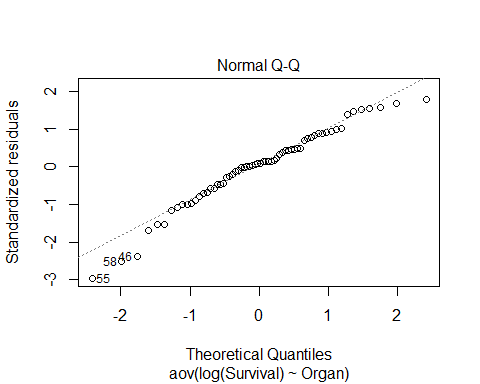
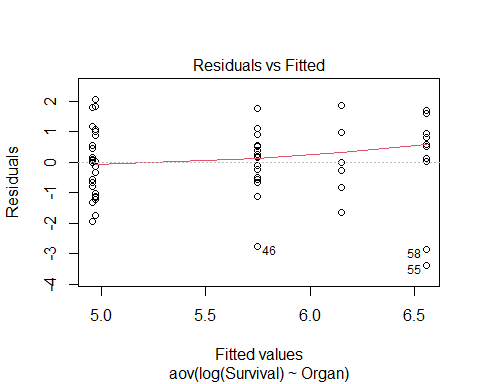
* above looks a lot better
* the sd are much closer and they are squished better
* the levenetest is also better with teh pvalue is high
* now we dont see there is dif
* we like higher pvalue in this

Cancermod\_log = aov(log(Survival)~Organ, data=CancerSurvival)  
summary(Cancermod\_log)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 24.49 6.122 4.286 0.00412 \*\*  
## Residuals 59 84.27 1.428   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Cehck conditions* - not as big of an issue - not perfect for QQNorm, but its beter

plot(Cancermod\_log, 1:2)



TukeyHSD(Cancermod\_log) # shows us that Bronchus-Breast and stomach-breast are sig

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log(Survival) ~ Organ, data = CancerSurvival)  
##   
## $Organ  
## diff lwr upr p adj  
## Bronchus-Breast -1.60543320 -2.906741 -0.3041254 0.0083352  
## Colon-Breast -0.80948110 -2.110789 0.4918267 0.4119156  
## Ovary-Breast -0.40798703 -2.114754 1.2987803 0.9615409  
## Stomach-Breast -1.59068365 -2.968399 -0.2129685 0.0158132  
## Colon-Bronchus 0.79595210 -0.357534 1.9494382 0.3072938  
## Ovary-Bronchus 1.19744617 -0.399483 2.7943753 0.2296079  
## Stomach-Bronchus 0.01474955 -1.224293 1.2537924 0.9999997  
## Ovary-Colon 0.40149407 -1.195435 1.9984232 0.9540004  
## Stomach-Colon -0.78120255 -2.020245 0.4578403 0.3981146  
## Stomach-Ovary -1.18269662 -2.842480 0.4770864 0.2763506

*Below shows the difference between small level what is “sig” and big sample level what is “sig”* - As sample size increases, if its sig doesn’t mean its pracitcal to us - we want ot make sure its true roughly, but we might have practical difference - we dont always use it because it can be sus

n=10 # 10 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.673 0.5742  
## 36

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 1.2604865 0.9596752 1.1035292 0.6933545

# D is almost double and levne pval = 0.2313  
# If we run a few times, then we eventually get something closer to 0.05  
# Gives idea of the dif that we would need to see for levene test to not like the dat aor say its sig

n=400 # 400 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.6395 0.5896  
## 1596

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 0.9938886 1.0174135 1.0623579 1.0265391