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

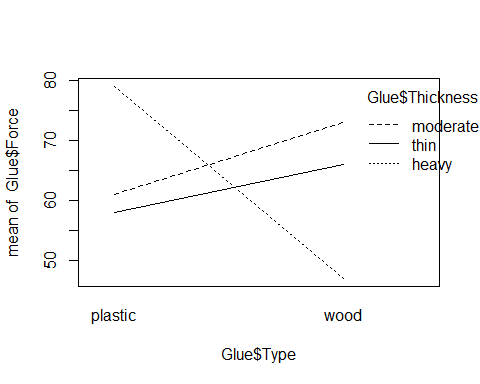
*See below for example on the flipping axises*

**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

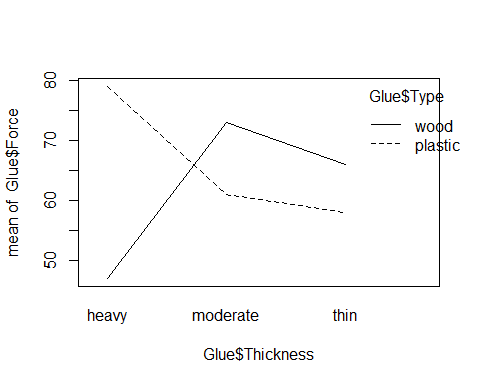
* The heavy amount of thickness is very different *There is some kind of interaction betcause heavy has a different interaction thatn the other lines*

**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*

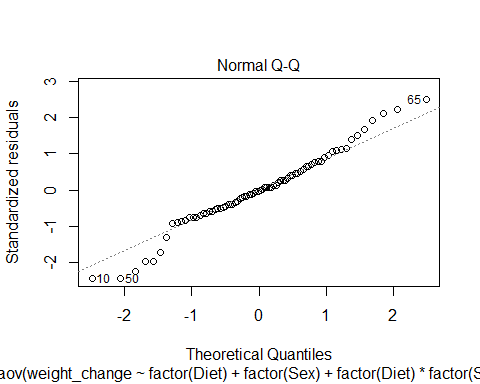
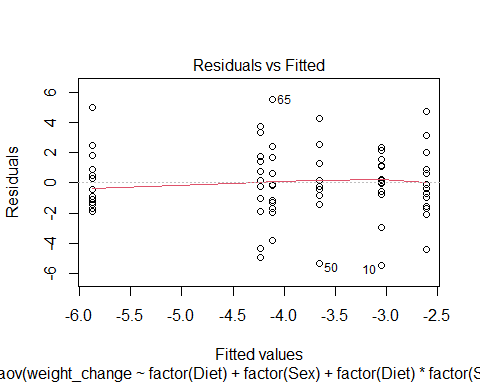
**Look at Resiudal analyss** - Nothing really stands out - Check for constant variance, they are spread similarly; one group is a little more compact, so may be an issue there, but not too bad - Normaility by QQPlot, some tail issues, but not too bad

**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

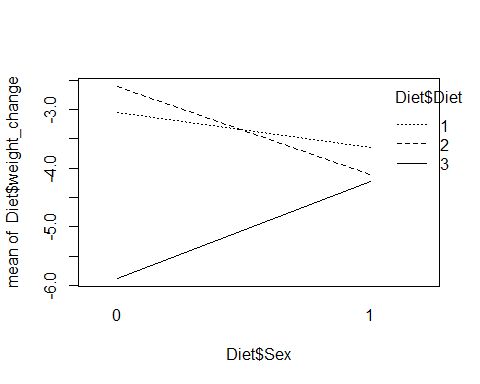
plot(mod1, 1:2)



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

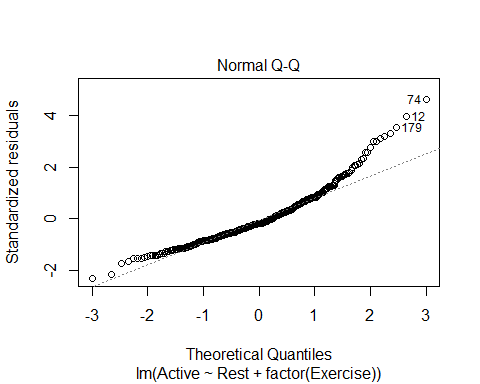
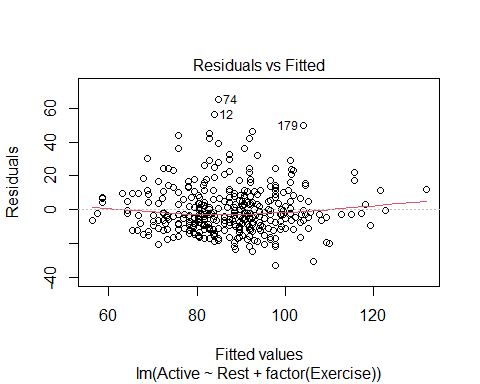
modp2 = lm(Active ~ Rest + factor(Exercise), data=Pulse)  
  
summary(modp2)

##   
## 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

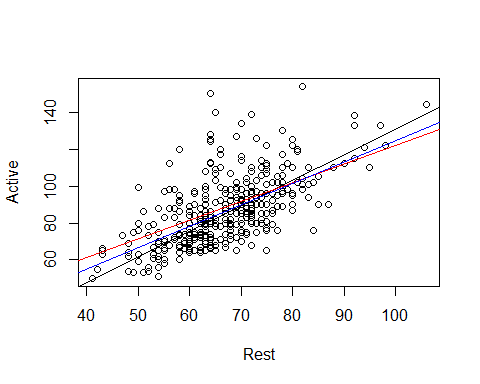
# How the active heart rate is adjected by the intercept   
# We are assuming that the resting heart rate follows teh same trend as the active, regardless of your exercise level

*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*

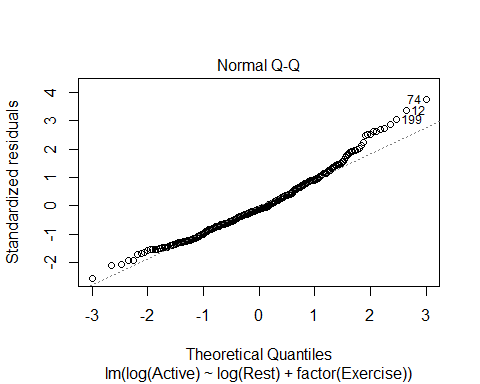
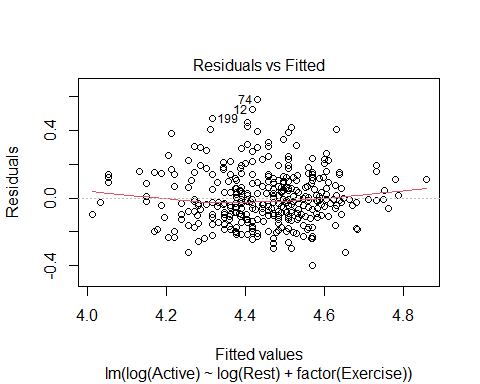
modp2log = lm(log(Active)~log(Rest) + factor(Exercise), data = Pulse)  
  
anova(modp2log)

## 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

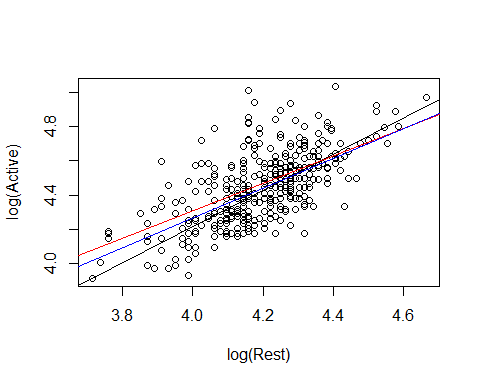
*Looking at Residuals* - There is some curve that is not ideal - Look at teh QQNorm, there are some big issues with that one tail on the untransformed thing - With the transformed data: the tail on QQNorm isn’t as bad and the little curve on the residual plot isn’t as bad.

*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)

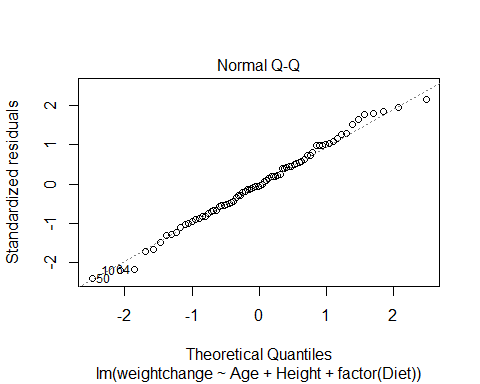
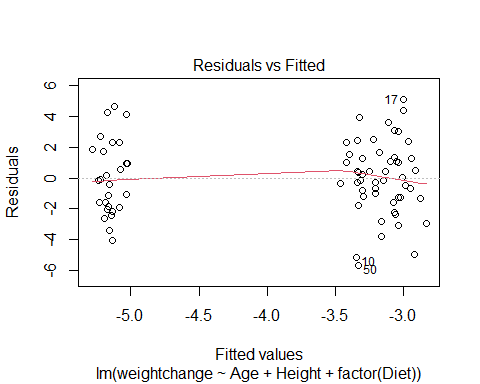
## Analysis of Variance Table  
##   
## Response: weightchange  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 71.09 35.547 6.1974 0.003229 \*\*  
## Residuals 75 430.18 5.736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**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.

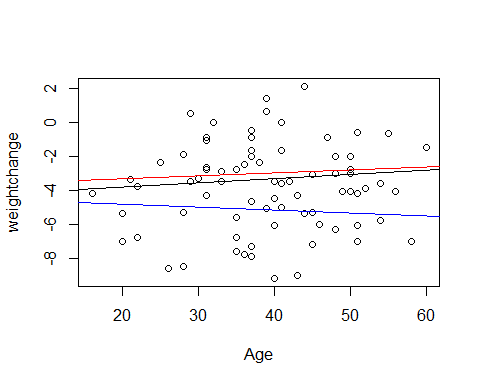
mod8 = lm(weightchange ~ Age + Height + factor(Diet), data=Diet)  
anova(mod8)

## 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')

