Homework 2

Melissa Franco

9/18/2020

heartbchol <- read.csv("heartbpchol.csv")  
bupa <- read.csv("bupa.csv")

## Exercise 1: Analysis of Variance

Use the significance level of .05

The heartbpchol.csv data set contains continuous cholesterol (Cholesterol) and blood pressure status (BP\_Status) (category: High/ Normal/ Optimal) for alive patients. For the heartbpchol.xlsx data set, consider a one-way ANOVA model to identify differences between group cholesterol means. The normality assumption is reasonable, so you can proceed without testing normality.

## A)

Perform a one-way ANOVA for Cholesterol with BP\_Status as the categorical predictor. Comment on statistical significance of BP\_Status, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

We will begin by viewing the data set and checking to see if the data is balanced.

str(heartbchol)

## 'data.frame': 541 obs. of 2 variables:  
## $ Cholesterol: int 221 188 292 319 205 247 202 150 228 280 ...  
## $ BP\_Status : chr "Optimal" "High" "High" "Normal" ...

names(heartbchol)

## [1] "Cholesterol" "BP\_Status"

heartbchol$BP\_Status <- as.factor(heartbchol$BP\_Status)  
table(heartbchol$BP\_Status)

##   
## High Normal Optimal   
## 229 245 67

We note that the groups are unbalanced and we will perform a one-way ANOVA test.

Our ANOVA hypothesis is as follows:

Cholesterol is our response variable and BP Status is our categorical variable.

Ho: The means of each group are equal (i.e. m1 = m2 = m3)/ No Blood Pressure status effect on Cholesterol

Ha: At least one of the groups has a significantly different mean than the others/Blood Pressure status effects Cholesterol

We will first test our ANOVA hypothesis.

aov.heart <- aov(Cholesterol ~ BP\_Status, data=heartbchol)   
summary(aov.heart)

## Df Sum Sq Mean Sq F value Pr(>F)   
## BP\_Status 2 25211 12605 6.671 0.00137 \*\*  
## Residuals 538 1016631 1890   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our ANOVA test produces the following information: We obtain a P-Value of 0.00137, which is significantly less than our test statistic (0.05), therefore we will reject our null hypothesis and accept that at least one of the groups has a significantly different mean than the others/Blood Pressure status effect on Cholesterol.

Let us now take a look at the variation of Cholesterol that can be explained by our model.

R^2

lm.res <- lm(Cholesterol ~ BP\_Status, data=heartbchol)  
anova(lm.res)

## Analysis of Variance Table  
##   
## Response: Cholesterol  
## Df Sum Sq Mean Sq F value Pr(>F)   
## BP\_Status 2 25211 12605.4 6.6708 0.001375 \*\*  
## Residuals 538 1016631 1889.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm.res)$r.squared

## [1] 0.02419833

BP Status can explain 2.4% of the behavior of Cholesterol.

We will next check our model assumption of homoscedasticity through the Levene test.

The hypothesis for the Levene test is as follows:

Ho: Variances are equal

Ha: At least one group has a different variance

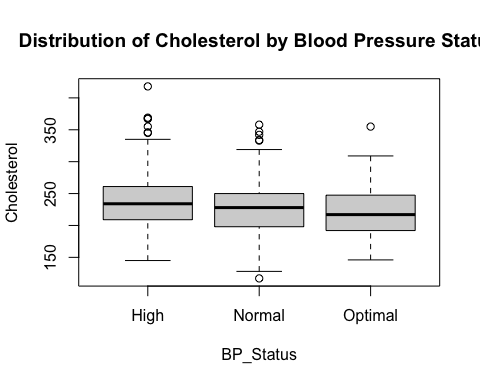
LeveneTest(aov.heart)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.1825 0.8332  
## 538

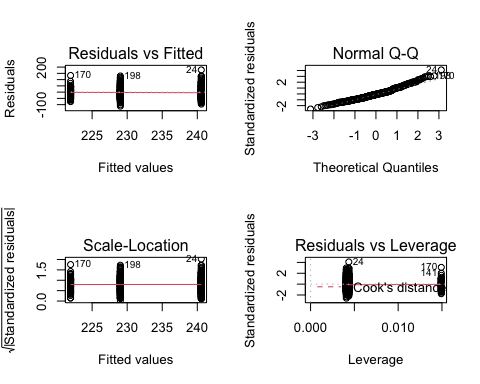
By the Levene test we receive a P-Value of 0.8332 which is significantly greater than our test statistic (0.05). Therefore, we will accept our null hypothesis that the groups have equal variance.

Next we will graphically represent the relationships between the studied groups with regard to our response variable, Cholesterol.

boxplot(Cholesterol ~ BP\_Status, data = heartbchol, main = "Distribution of Cholesterol by Blood Pressure Status")



par(mfrow=c(2,2))   
plot(aov.heart)



Qualitatively by viewing the above box plot and QQ plot we see that the data is symmetrical, with several outliers, and does represent traits of normality with some right tail behavior.

## B)

Comment on any significantly different cholesterol means as determined by the post-hoc test comparing all pairwise differences. Specifically explain what that tells us about differences in Cholesterol levels across blood pressure status groups, like which group has the highest or lowest mean values of Cholesterol.

Our ANOVA test indicated that at least one of the groups has a significantly different mean than the others. To determine which group/groups have a different mean we will run a post-hoc procedure called Tukey’s honestly significant difference (HSD).

TukeyHSD(aov.heart)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Cholesterol ~ BP\_Status, data = heartbchol)  
##   
## $BP\_Status  
## diff lwr upr p adj  
## Normal-High -11.543481 -20.93394 -2.153023 0.0111929  
## Optimal-High -18.646679 -32.83690 -4.456460 0.0059898  
## Optimal-Normal -7.103198 -21.18815 6.981749 0.4624869

The output presents the pair-wise group mean differences, followed by their confidence levels, and then their corresponding p-values.

We will use our p-value for the following hypothesis:

Ho: There is no significant difference between the group means

Ha: There is a significant difference between the group means

Only the top two rows (Normal - High, Optimal - High) produce a p-value that is significantly smaller than our test statistic (0.05). Therefore, we reject our null hypothesis and accept the alternative hypothesis that there is a significant difference between the group means.

For Normal - High, it can be determined that High is greater than Normal. For Optimal - High, it can be determined that High is also greater than Optimal.

Since the bottom row, Optimal - Normal, produced a significantly large p-value, we accept our null hypothesis that there is no significant difference between the group means.

For Optimal - Normal, it can be determined that they have equal means.

Conclusively, high blood pressure has a significant effect on cholesterol and specifically: High > Normal = Optimal

## Exercise 2: Analysis of Variance

For this problem use the bupa.csv data set. The mean corpuscular volume and alkaline phosphatase are blood tests thought to be sensitive to liver disorder related to excessive alcohol consumption. We assume that normality and independence assumptions are valid.

## A)

Perform a one-way ANOVA for mcv as a function of drinkgroup. Comment on significance of the drinkgroup, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

We will begin by viewing our data and we assume that the normality and independence assumptions are valid. We will note that the data set is unbalanced and will move on to performing a one-way ANOVA test to determine if mean corpuscular volume (mcv) is effected by drinkgroup.

str(bupa)

## 'data.frame': 345 obs. of 3 variables:  
## $ mcv : int 85 85 86 91 87 98 88 88 92 90 ...  
## $ alkphos : int 92 64 54 78 70 55 62 67 54 60 ...  
## $ drinkgroup: int 1 1 1 1 1 1 1 1 1 1 ...

names(bupa)

## [1] "mcv" "alkphos" "drinkgroup"

bupa$drinkgroup <- as.factor(bupa$drinkgroup)  
table(bupa$drinkgroup)

##   
## 1 2 3 4 5   
## 117 52 88 67 21

Our ANOVA hypothesis is as follows:

Ho: The means of each group are equal (i.e. m1 = m2 = m3 = m4 = m5)/ No Drinkgroup effect on mcv

Ha: At least one of the groups has a significantly different mean than the others/Drinkgroup effect on mcv

Our response variable is mcv and drinkgroup is our categorical predictor.

We will first test our ANOVA hypothesis.

aov.mcv1<- aov(mcv ~ drinkgroup, data=bupa)   
summary(aov.mcv1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## drinkgroup 4 733 183.29 10.26 7.43e-08 \*\*\*  
## Residuals 340 6073 17.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our ANOVA test produces the following information:

We obtain a P-Value of 7.43e-08, which is significantly less than our test statistic (0.05), therefore we will reject our null hypothesis and accept that at least one of the groups has a significantly different mean than the others/drinkgroup effect on mcv.

Let us now take a look at the variation of mcv that can be explained by our model.

R^2

lm.res <- lm(mcv ~ drinkgroup, data=bupa)  
anova(lm.res)

## Analysis of Variance Table  
##   
## Response: mcv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## drinkgroup 4 733.2 183.294 10.262 7.429e-08 \*\*\*  
## Residuals 340 6073.1 17.862   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm.res)$r.squared

## [1] 0.1077214

Drinkgroup can explain 10.77% of the behavior of mcv.

We will check our model assumptions through the Levene test.

The hypothesis for the Levene test are as follows:

Ho: Variances are equal

Ha: At least one group has a different variance

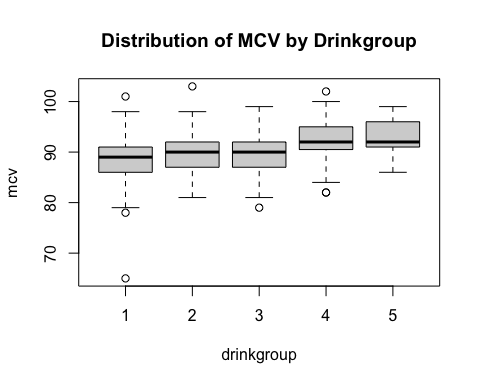
LeveneTest(aov.mcv1)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.3053 0.8744  
## 340

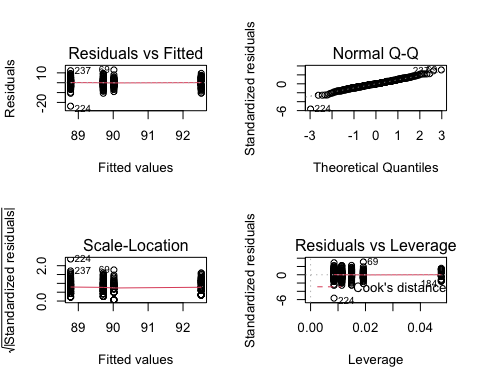
By the Levene test we receive a P-Value of 0.8744 which is significantly greater than our test statistic (0.05), thus, we will accept our null hypothesis that the groups have equal variance.

Next we will graphically represent the relationships between the studied groups with regard to our response variable, mcv.

boxplot(mcv ~ drinkgroup, data = bupa, main = "Distribution of MCV by Drinkgroup")



par(mfrow=c(2,2))   
plot(aov.mcv1)

 Qualitatively by viewing the above box plot and QQ plot we see that the data is symmetrical, with very few outliers on both the left and right tail, and does represent traits of normality.

## B)

Perform a one-way ANOVA for alkphos as a function of drinkgroup. Comment on statistical significance of the drinkgroup, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

aov.mcv2 <- aov(alkphos ~ drinkgroup, data=bupa)   
summary(aov.mcv2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## drinkgroup 4 4946 1236.4 3.792 0.00495 \*\*  
## Residuals 340 110858 326.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our ANOVA test produces the following information:

We obtain a P-Value of 0.00495, which is significantly less than our test statistic (0.05), therefore we will reject our null hypothesis and accept that at least one of the groups has a significantly different mean than the others/drinkgroup effect on alkphos.

Let us now take a look at the variation of alkphos that can be explained by our model.

R^2

lm.res1 <- lm(alkphos ~ drinkgroup, data=bupa)  
anova(lm.res1)

## Analysis of Variance Table  
##   
## Response: alkphos  
## Df Sum Sq Mean Sq F value Pr(>F)   
## drinkgroup 4 4946 1236.41 3.7921 0.00495 \*\*  
## Residuals 340 110858 326.05   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm.res1)$r.squared

## [1] 0.04270721

Drinkgroup can explain 4.27% of the behavior of alkphos.

We will check our model assumptions through the Levene test.

The hypothesis for the Levene test is as follows:

Ho: Variances are equal

Ha: At least one group has a different variance

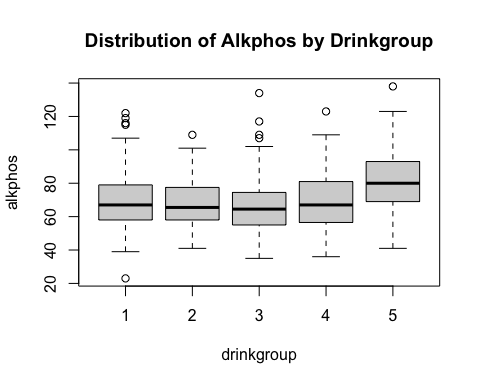
LeveneTest(aov.mcv2)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.8089 0.5201  
## 340

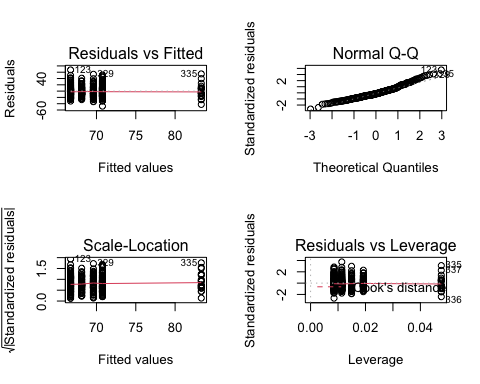
By the Levene test we receive a p-value of 0.5201 which is significantly greater than our test statistic (0.05), thus we will accept our null hypothesis that the groups have equal variance.

Next we will graphically represent the relationships between the studied groups with regard to our response variable, alkphos.

boxplot(alkphos ~ drinkgroup, data = bupa, main = "Distribution of Alkphos by Drinkgroup")



par(mfrow=c(2,2))   
plot(aov.mcv2)

 Qualitatively by viewing the above box plot and QQ plot we see that the data is symmetrical, with a few outliers, and does represent traits of normality.

## C)

Perform post-hoc tests for models in a) and b). Comment on any similarities or differences you observe from their results.

Post- hoc test for model (a):

TukeyHSD(aov.mcv1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = mcv ~ drinkgroup, data = bupa)  
##   
## $drinkgroup  
## diff lwr upr p adj  
## 2-1 1.241452991 -0.690316778 3.173223 0.3973587  
## 3-1 0.938131313 -0.697363908 2.573627 0.5157202  
## 4-1 3.744610282 1.968846495 5.520374 0.0000002  
## 5-1 3.746031746 0.999127120 6.492936 0.0020039  
## 3-2 -0.303321678 -2.330666517 1.724023 0.9940309  
## 4-2 2.503157290 0.361050967 4.645264 0.0127884  
## 5-2 2.504578755 -0.492214115 5.501372 0.1499436  
## 4-3 2.806478969 0.927189295 4.685769 0.0005031  
## 5-3 2.807900433 -0.007037875 5.622839 0.0509321  
## 5-4 0.001421464 -2.897262735 2.900106 1.0000000

The output presents the pair-wise group mean differences, followed by their confidence levels, and then their corresponding p-values.

We will use our p-value for the following hypothesis:

Ho: There is no significant difference between the group means

Ha: There is a significant difference between the group means

Only four rows produce a p-value that is significantly smaller than our test statistic (0.05), therefore we accept the alternative hypothesis that there is a significant difference between the group means.

For 4-1, it can determined that group 4 is greater than group 1.

For 5-1, it can determined that group 5 is greater than group 1.

For 4-2, it can determined that group 4 is greater than group 2.

For 4-3, it can determined that group 4 is greater than group 3.

Alternatively the remaining rows (2-1, 3-1, 3-2, 5-2, 5-3, 5-4) each produced a p-value greater than our test statistic, thus we accept our null hypothesis that there is no significant difference between the group means.

Post-hoc test for model (B):

TukeyHSD(aov.mcv2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = alkphos ~ drinkgroup, data = bupa)  
##   
## $drinkgroup  
## diff lwr upr p adj  
## 2-1 -2.645299 -10.8987258 5.608127 0.9045115  
## 3-1 -4.056138 -11.0437411 2.931464 0.5035761  
## 4-1 -1.148743 -8.7356393 6.438152 0.9937509  
## 5-1 12.572650 0.8365845 24.308715 0.0288892  
## 3-2 -1.410839 -10.0726073 7.250929 0.9917361  
## 4-2 1.496556 -7.6555274 10.648639 0.9916117  
## 5-2 15.217949 2.4142438 28.021654 0.0107154  
## 4-3 2.907395 -5.1218122 10.936602 0.8583931  
## 5-3 16.628788 4.6020510 28.655525 0.0016498  
## 5-4 13.721393 1.3368544 26.105932 0.0214375

The output presents the pair-wise group mean differences, followed by their confidence levels, and then their corresponding p-values.

We will use our p-value for the following hypothesis:

Ho: There is no significant difference between the group means

Ha: There is a significant difference between the group means

Only four rows produce a p-value that is significantly smaller than our test statistic (0.05). Therefore we accept the alternative hypothesis that there is a significant difference between the group means.

For 5-1, it can determined that group 5 is greater than group 1.

For 5-2, it can determined that group 5 is greater than group 2.

For 5-3, it can determined that group 5 is greater than group 3.

For 5-4, it can determined that group 5 is greater than group 4.

Alternatively the remaining rows, (2-1, 3-1, 4-1, 3-2, 4-2, 4-3) produced a p-value greater than our test statistic, we accept our null hypothesis that there is no significant difference between the group means.

For groups 1,2,3, and 4 it can be determined that they have equal means.

Conclusively, high blood pressure has a significant effect on cholesterol and specifically: 5 > 1 = 2 = 3 = 4

## Instructor Note

For Exercise 3 and 4, we will NOT run the equal variance test (Levene’s Test) and will assume the homoscedasticity. As we have just a small number of samples for each cell for the implementation of Levene’s Test (or boxplot), we will not consider this for the class and just go ahead with the standard ANOVA that I covered in the class. No Welch’s ANOVA for Exercise 3 and 4.

## Exercise 3:

The psychology department at a hypothetical university has been accused of underpaying female faculty members. The data represent salary (in thousands of dollars) for all 22 professors in the department. This problem is from Maxwell and Delaney (2004).

## A)

Fit a two-way ANOVA model including sex (F, M) and rank (Assistant, Associate) the interaction term. What do the Type 1 and Type 3 sums of squares tell us about significance of effects?

Is the interaction between sex and rank significant? Also comment on the variation explained by the model.

We will begin by viewing our data and checking if balanced or unbalanced. We see that our data is unbalanced.

psych <- read.csv("psych.csv")  
names(psych)

## [1] "sex" "rank" "salary"

psych$sex <- as.factor(psych$sex)  
psych$rank <- as.factor(psych$rank)  
str(psych)

## 'data.frame': 22 obs. of 3 variables:  
## $ sex : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 2 2 2 2 ...  
## $ rank : Factor w/ 2 levels "Assist","Assoc": 1 1 1 1 1 1 1 1 1 1 ...  
## $ salary: int 33 36 35 38 42 37 39 38 40 44 ...

table(psych$sex);table(psych$rank)

##   
## F M   
## 10 12

##   
## Assist Assoc   
## 10 12

We will begin by running a type I two-way ANOVA test for interaction between sex and rank:

aov.salary <- aov(salary ~ sex \* rank, data = psych)  
aov.salary1 <- aov(salary ~ rank \* sex, data = psych)  
summary(aov.salary)

## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 155.15 155.15 17.007 0.000637 \*\*\*  
## rank 1 169.82 169.82 18.616 0.000417 \*\*\*  
## sex:rank 1 0.63 0.63 0.069 0.795101   
## Residuals 18 164.21 9.12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov.salary1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## rank 1 252.22 252.22 27.647 5.33e-05 \*\*\*  
## sex 1 72.76 72.76 7.975 0.0112 \*   
## rank:sex 1 0.63 0.63 0.069 0.7951   
## Residuals 18 164.21 9.12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We interpret our type I two-way ANOVA test as follows:

Since we are wanting to determine if there is an interaction between sex and rank. Our test hypothesis is as follows:

Ho: No interaction

Ha: Interaction

Both of our outputs provided the interaction of the two variables (sex:rank) is associated with a p-value of 0.795 which is significantly larger than our significance level of 0.05. For this reason we accept our null hypothesis that the interaction is not significant which we interpret as suggesting that the effect (if any) of sex upon salary is the same for rank (Assistant and Associate).

We will next run a Type III two-way ANOVA test for interaction between sex and rank:

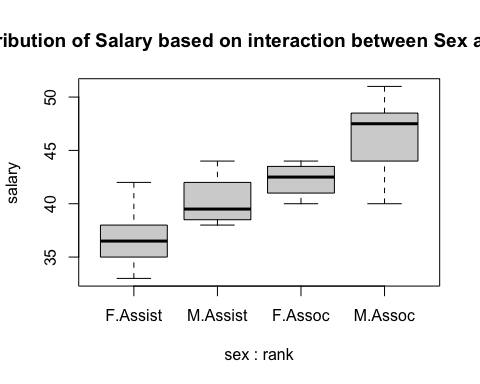
Anova(aov.salary, type = 3)

## Anova Table (Type III tests)  
##   
## Response: salary  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 8140.2 1 892.2994 < 2e-16 \*\*\*  
## sex 28.0 1 3.0711 0.09671 .   
## rank 70.4 1 7.7189 0.01240 \*   
## sex:rank 0.6 1 0.0695 0.79510   
## Residuals 164.2 18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The type III ANOVA test produces a p-value of 0.795 for interaction of the two variables (sex:rank). Thus our conclusion is the same: The interaction is not significant which we interpret as suggesting that the effect (if any) of sex upon salary is the same for rank (Assistant and Associate).

Let us examine the boxplots for the interaction:

boxplot(salary ~ sex:rank, data = psych, main = "Distribution of Salary based on interaction between Sex and Rank")

 The resulting boxplots reveal that both male and females who hold a rank of associate have more earned salary, while those in the the assistant rank tend to be lower, regardless of sex. However we can observed that females typically average a lower salary regardless of rank.

## B)

Refit the model without the interaction term. Comment on the significance of effects and variation explained.

Report and interpret the Type 1 and Type 3 tests of the main effects. Are the main effects of rank and sex significant?

We will begin by performing a type I two way ANOVA test (sequential). Since order matters we will start with sex then rank and then the reverse.

aov.salary3 <- aov(salary ~ sex + rank, data = psych)  
summary(aov.salary3)

## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 155.2 155.15 17.88 0.000454 \*\*\*  
## rank 1 169.8 169.82 19.57 0.000291 \*\*\*  
## Residuals 19 164.8 8.68   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our results provide significantly small P-Values. Our test hypothesis for ANOVA is as follows:

Ho: Sex and Rank have no effect on Salary

Ha: Sex and Rank have effect on Salary

The first test produce p-values for both sex and rank that are significantly less than our test statistic (0.05) thus we we will reject our null hypothesis and accept our alternative hypothesis that sex and rank have an effect on salary.

New we will run another type I two-way ANOVA test with rank and sex.

aov.salary4 <- aov(salary ~ rank + sex, data = psych)  
summary(aov.salary4)

## Df Sum Sq Mean Sq F value Pr(>F)   
## rank 1 252.22 252.22 29.071 3.34e-05 \*\*\*  
## sex 1 72.76 72.76 8.386 0.00926 \*\*   
## Residuals 19 164.84 8.68   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

This test also produces p-values that are significantly less than our test statistic, thus the same conclusion holds. We accept that rank and sex have an effect on salary.

Next we will run a type III two-way ANOVA test:

aov.salary4 <- aov(salary ~ sex + rank, data = psych)  
Anova(aov.salary4, type=3)

## Anova Table (Type III tests)  
##   
## Response: salary  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 10227.6 1 1178.8469 < 2.2e-16 \*\*\*  
## sex 72.8 1 8.3862 0.0092618 \*\*   
## rank 169.8 1 19.5743 0.0002912 \*\*\*  
## Residuals 164.8 19   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our Type III test focuses on unique contribution and for this type of test order is inconsequential. Since our p-values are significantly less that our test statistic we will accept our alternate hypothesis that sex and rank have significant effect on salary.

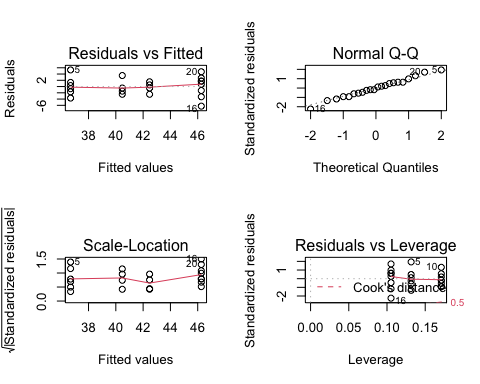
## C)

Obtain model diagnostics to validate your Normality assumptions.

par(mfrow=c(1,2))  
boxplot(salary ~ sex, data = psych, main = "Salary by Sex", xlab="Sex", ylab = "Salary (in thousands of dollars)")  
boxplot(salary ~ rank, data = psych, main = "Salary by Rank", xlab="Rank", ylab = "Salary (in thousands of dollars)")



par(mfrow=c(2,2))  
plot(aov.salary3)

 Qualitatively by viewing the above box plot and QQ plot we see that the data is symmetrical and does represent traits of normality.

## D)

Choose a final model based on your results from parts (a) and (b). Comment on any significant group differences through the post-hoc test. State the differences in salary across different main effect groups and interaction (if included) between them.

Hint: For interpretations of differences for the main effects, state quantitative interpretations of the significantly different groups (e.g. estimated differences between groups and what the difference tells us about salary). For interaction term, identify significant interactions, but no need to interpret it quantitatively.

Based on our tests we determined that there was no indication of interaction we will select our ANOVA test type I. Since order does matter with a ANOVA test type I we will use the test with sum of squares for sex being taken first followed by rank.

Let us now take a look at the variation of salary that can be explained by our model.

R^2

lm.res.salary <- lm(salary ~ sex + rank, data = psych)  
anova(lm.res.salary)

## Analysis of Variance Table  
##   
## Response: salary  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 155.15 155.152 17.883 0.0004544 \*\*\*  
## rank 1 169.82 169.825 19.574 0.0002912 \*\*\*  
## Residuals 19 164.84 8.676   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm.res.salary)$r.squared

## [1] 0.6634627

Sex and rank can explain 66.34% of the behavior of salary.

TukeyHSD(aov.salary3)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = salary ~ sex + rank, data = psych)  
##   
## $sex  
## diff lwr upr p adj  
## M-F 5.333333 2.693648 7.973019 0.0004544  
##   
## $rank  
## diff lwr upr p adj  
## Assoc-Assist 5.377778 2.738092 8.017463 0.0004193

The output of our Tukey HSD test presents the pair-wise group mean differences, followed by their confidence levels, and then their corresponding p-values.

We will use our p-value for the following hypothesis:

Ho: There is no significant difference between the group means

Ha: There is a significant difference between the group means

Both sex and rank produce a p-value that is significantly smaller than our test statistic (0.05). Therefore we accept the alternative hypothesis that there is a significant difference between the group means.

For male and female, it can be determined that males average a higher salary than females.

For rank, associate and assistant, it can be determined that associates average a higher salary than assistants.

## Exercise 4:

Use the cars\_new.csv. See HW1 for detailed information of variables.

## A)

Start with a three-way main effects ANOVA and choose the best main effects ANOVA model for mpg\_highway as a function of cylinders, origin, and type for the cars in this set. Comment on which terms should be kept in a model for mpg\_highway and why based on Type 3 SS. For the model with just predictors you decide to keep, comment on the significant effects in the model and comment on how much variation in highway fuel efficiency the model describes.

We will begin by viewing our data set and note that the data is unbalanced.

cars <- read.csv("cars\_new.csv")  
  
cars$cylinders <- as.factor(cars$cylinders)  
cars$origin <- as.factor(cars$origin)  
cars$type <- as.factor(cars$type)  
str(cars)

## 'data.frame': 180 obs. of 4 variables:  
## $ type : Factor w/ 2 levels "Sedan","Sports": 1 1 1 1 1 2 1 1 1 1 ...  
## $ origin : Factor w/ 2 levels "Asia","USA": 1 1 1 1 1 1 2 2 2 2 ...  
## $ cylinders : Factor w/ 2 levels "4","6": 1 1 2 2 2 2 2 2 2 2 ...  
## $ mpg\_highway: int 31 29 28 24 24 24 30 29 30 28 ...

table(cars$cylinders);table(cars$origin)

##   
## 4 6   
## 86 94

##   
## Asia USA   
## 104 76

We will select the best model by running a backwards elimination model selection test.

First we will perform a type III ANOVA test with cylinders, origin, and type.

aov.cars <- aov(mpg\_highway ~ cylinders + origin + type, data = cars)  
Anova(aov.cars, type=3)

## Anova Table (Type III tests)  
##   
## Response: mpg\_highway  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 69548 1 6501.6715 < 2e-16 \*\*\*  
## cylinders 1453 1 135.8499 < 2e-16 \*\*\*  
## origin 1 1 0.0786 0.77948   
## type 108 1 10.1018 0.00175 \*\*   
## Residuals 1883 176   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The categorical predictor ‘Origin’ produces the largest p-value that is greater than criteria/cut off of 0.05. So we will fist remove origin.

Next we will run the type III ANOVA test again with just cylinders and type.

aov.cars1 <- aov(mpg\_highway ~ cylinders + type, data = cars)  
Anova(aov.cars1, type=3)

## Anova Table (Type III tests)  
##   
## Response: mpg\_highway  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 88449 1 8311.96 < 2.2e-16 \*\*\*  
## cylinders 1482 1 139.27 < 2.2e-16 \*\*\*  
## type 116 1 10.88 0.001175 \*\*   
## Residuals 1883 177   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The categorical predictor ‘Type’ has the largest p-value of 0.001175 that is less than our cirteria/cut off (0.05) so we keep both cylinders and type.

The best effects for our ANOVA model are cylinders and type.

## B)

Starting with main effects chosen in part (a), find your best ANOVA model by adding in any additional interaction terms that will significantly improve the model. For your final model,comment on the significant effects and variation explained by the model.

Since we determined that cylinders and type are our best affects for our ANOVA model, we will now check their interaction.

aov.cars2 <- aov(mpg\_highway ~ cylinders \* type, data = cars)  
Anova(aov.cars2, type=3)

## Anova Table (Type III tests)  
##   
## Response: mpg\_highway  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 85471 1 8358.838 < 2.2e-16 \*\*\*  
## cylinders 1558 1 152.397 < 2.2e-16 \*\*\*  
## type 198 1 19.392 1.844e-05 \*\*\*  
## cylinders:type 84 1 8.201 0.004696 \*\*   
## Residuals 1800 176   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Interaction (cylinders:type) has p-value 0.004696 which is significantly less than our test statistic (0.05) so we reject the null and accept that there is an interaction between cylinders and type.

For type III two-way ANOVA test we will want to keep cylinders and type and their interaction.

Let us now take a look at the variation of mpg-highway that can be explained by our model.

R^2

lm.res.cars <- lm(mpg\_highway ~ cylinders \* type, data = cars)  
anova(lm.res.cars)

## Analysis of Variance Table  
##   
## Response: mpg\_highway  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cylinders 1 1470.79 1470.79 143.839 < 2.2e-16 \*\*\*  
## type 1 115.78 115.78 11.323 0.0009397 \*\*\*  
## cylinders:type 1 83.86 83.86 8.201 0.0046958 \*\*   
## Residuals 176 1799.64 10.23   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

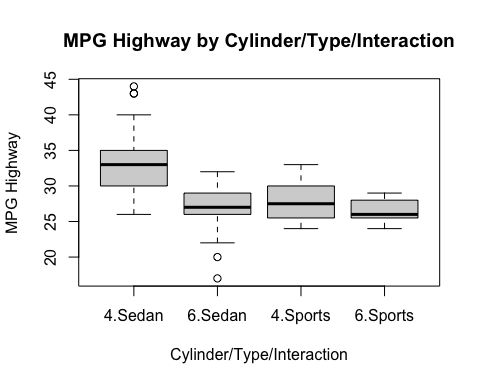
summary(lm.res.cars)$r.squared

## [1] 0.4813821

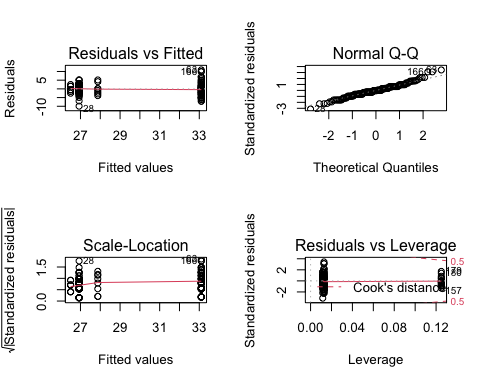
Cylinders, type, and their interaction can explain 48.138% of the behavior of mpg\_highway.

We will next obtain model diagnostics to validate our Normality assumptions.

boxplot(mpg\_highway ~ cylinders \* type, data = cars, main = "MPG Highway by Cylinder/Type/Interaction", xlab="Cylinder/Type/Interaction", ylab = "MPG Highway")



par(mfrow=c(2,2))  
plot(aov.cars2)

 Qualitatively by viewing the above box plot and QQ plot we see that the data is symmetrical and does represent traits of normality.

## C)

Comment on any significant group differences through the post-hoc test. What does this tell us about fuel efficiency differences across cylinders, origin, or type groups? See Hint in Exercise 3

We will run our Tukey HSD test:

TukeyHSD(aov.cars2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = mpg\_highway ~ cylinders \* type, data = cars)  
##   
## $cylinders  
## diff lwr upr p adj  
## 6-4 -5.722662 -6.664343 -4.780981 0  
##   
## $type  
## diff lwr upr p adj  
## Sports-Sedan -2.817931 -4.470787 -1.165075 0.0009407  
##   
## $`cylinders:type`  
## diff lwr upr p adj  
## 6:Sedan-4:Sedan -6.1723315 -7.469178 -4.875485 0.0000000  
## 4:Sports-4:Sedan -5.2275641 -8.306639 -2.148489 0.0001079  
## 6:Sports-4:Sedan -6.6025641 -9.681639 -3.523489 0.0000006  
## 4:Sports-6:Sedan 0.9447674 -2.120956 4.010491 0.8546517  
## 6:Sports-6:Sedan -0.4302326 -3.495956 2.635491 0.9834567  
## 6:Sports-4:Sports -1.3750000 -5.521993 2.771993 0.8253946

We will analyze each row with a test statistic of 0.05 and the following hypothesis:

Ho: There is no significant difference between the group means

Ha: There is a significant difference between the group means

Starting with cylinders, the p-valued produced is less than our test statistic so we will reject our null and accept the alternative hypothesis that there is a difference between 6 and 4 cylinders and conclusively, 4 > 6

Next with sports and sedans, the p-valued produced is less than our test statistic so we will reject our null and accept the alternative hypothesis that there is a difference between sports and sedans and conclusively, Sedans > Sports

For cylinders with type interaction there are only three rows that produce a p-value that is significantly smaller than our test statistic (0.05). Therefore we accept the alternative hypothesis that there is a significant difference between the group means.

4 cylinder sedan > 6 cylinder sedan

4 cylinder sedan > 4 cylinder sports

4 cylinder sedan > 6 cylinder sports

Our test indicates that for 4 cylinder sports and 6 cylinder sedans, 6 cylinder sports and 6 cylinder sedans, and 6 cylinder sports and 4 cylinder sport all produced a p-value greater than our test statistic thus we will accept our null hypothesis that there is not a significant difference between the group means.

These results allow us to determine that 4 cylinder vehicles average a higher highway mileage than 6 cylinder vehicles and sedans have a higher average highway mileage than sports cars.