November 20, 2020

1. Please open the file in the Oncourse resources folder entitled “Cholesterol.csv”

You will find the following data in this file. A researcher wanted to determine if individuals in three states had different cholesterol levels. Please test this hypothesis via ANCOVA. Be sure to tell me why ANCOVA is the best choice for this analysis. Be sure to test all assumptions to ensure that the data are appropriate for the ANCOVA

Variable Description Cholesterol: Cholesterol score for each subject Age: Age of each subject State: State of residence for each subject

#install.packages("gplots")  
library(gplots)

## Warning: package 'gplots' was built under R version 4.0.3

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(ggplot2)  
library(foreign)  
library(car)

## Loading required package: carData

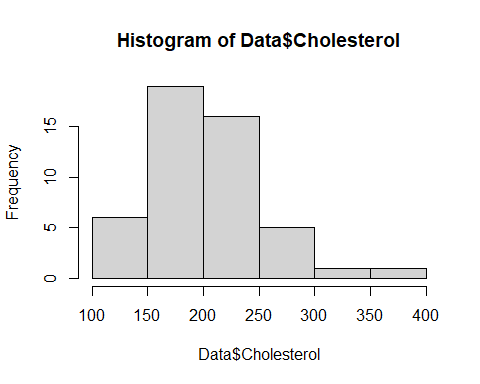
##reading and exploring data file   
Data=read.csv("cholesterol.csv" ,header = TRUE)  
summary(Data)

## Cholesterol Age State   
## Min. :110.0 Min. :18.00 Length:48   
## 1st Qu.:177.0 1st Qu.:35.25 Class :character   
## Median :198.5 Median :46.50 Mode :character   
## Mean :206.6 Mean :47.69   
## 3rd Qu.:242.0 3rd Qu.:58.50   
## Max. :356.0 Max. :78.00

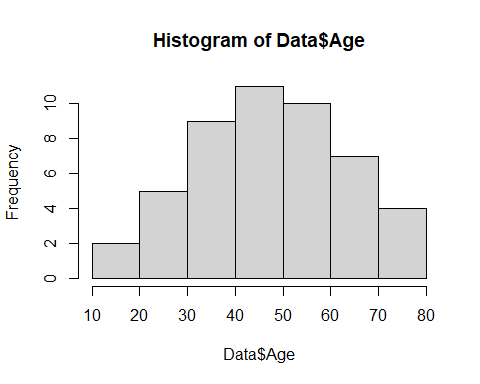
colnames(Data)

## [1] "Cholesterol" "Age" "State"

# The data contains a Cholesterol(dependent variable-y axis), a group with three groups (independent variable-State), and a covariate Age  
  
# Initial Plots too look into data  
  
hist(Data$Cholesterol)



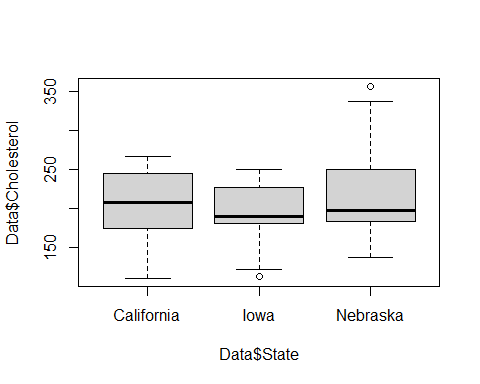
hist(Data$Age)



# checking normality by state with cholesterol  
by(Data$Cholesterol, Data$State, shapiro.test)

## Data$State: California  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.94602, p-value = 0.366  
##   
## ------------------------------------------------------------   
## Data$State: Iowa  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.91508, p-value = 0.2797  
##   
## ------------------------------------------------------------   
## Data$State: Nebraska  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.91796, p-value = 0.1038

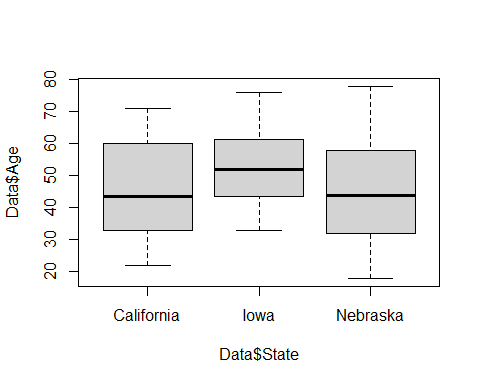
boxplot(Data$Cholesterol ~ Data$State)



# California: W = 0.94602, p-value = 0.366>0.05  
# Iowa:W = 0.91508, p-value = 0.2797>0.05  
# Nebraska: W = 0.91796, p-value = 0.1038>0.05  
# no normality violation  
  
# checking normality by age and state  
by(Data$Age, Data$State, shapiro.test)

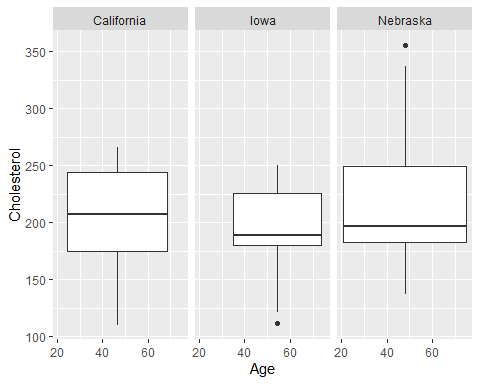
## Data$State: California  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.94486, p-value = 0.3502  
##   
## ------------------------------------------------------------   
## Data$State: Iowa  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.97287, p-value = 0.9139  
##   
## ------------------------------------------------------------   
## Data$State: Nebraska  
##   
## Shapiro-Wilk normality test  
##   
## data: dd[x, ]  
## W = 0.96795, p-value = 0.7348

boxplot(Data$Age~Data$State)

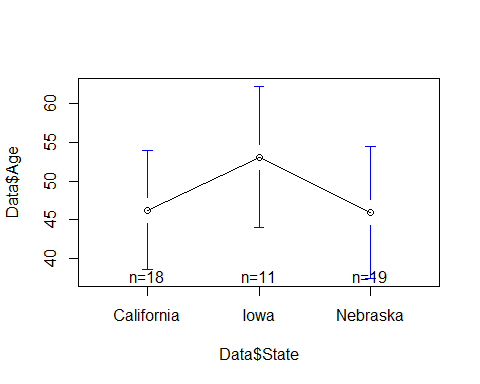


# California: W = 0.94486, p-value = 0.3502>0.05  
# Iowa:W = W = 0.97287, p-value = 0.9139>0.05  
# Nebraska:W = 0.96795, p-value = 0.7348>0.05  
# no normality violation  
  
# draw boxplot by group  
boxplot <- ggplot(Data, aes(Age, Cholesterol))  
boxplot + geom\_boxplot() + facet\_wrap(~State) + labs(x="Age", y ="Cholesterol")

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



## Assumptions of ANCOVA I: Are the IV(state) and CV(age) correlated?  
# Qualitative assessment  
plotmeans(Data$Age ~ Data$State)



# Quantitative assessment  
x=aov(Age~State,data=Data)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)  
## State 2 417 208.7 0.821 0.446  
## Residuals 45 11433 254.1

# F(2, 45) = 0.82, p = 0.44>0.05 . This means that for these data the variances are very similar.  
  
#Levene’s test  
Data$State<-as.factor(Data$State)  
leveneTest(Data$Age, Data$State, center = median)

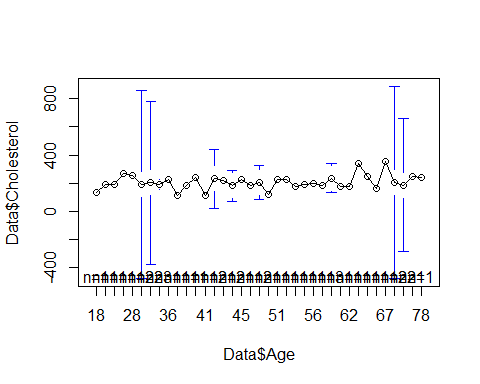
## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.5343 0.5897  
## 45

# F(2, 45) = 0.53, p = 0.58>0.05 . This means that for these data the variances are very similar.  
  
# All show that there is no statistical relationship between state and age  
  
#Assumption of ANCOVA II: Are DV(Cholesterol) and CV(age) correlated?  
  
# Qualitative assessment  
plotmeans(Data$Cholesterol~Data$Age)

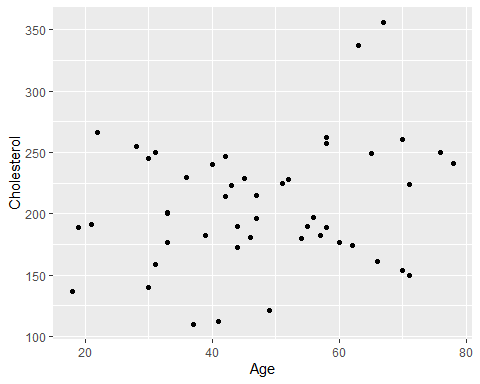
## Warning in qt((1 + p)/2, ns - 1): NaNs produced

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-  
## length arrow is of indeterminate angle and so skipped



graph <- ggplot(Data,aes(Age, Cholesterol))  
graph + geom\_point()



#Levene’s test  
agex<-as.factor(Data$Age)  
leveneTest(Data$Cholesterol, agex, center = median)

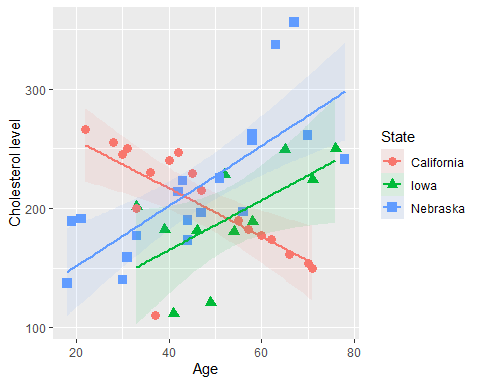
## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 36 1.4095 0.2785  
## 11

# F(36, 11) = 1.4, p = 0.27>0.05 . no normality violation.  
  
# Quantitative assessment  
x=lm(Cholesterol ~ Age, data=Data)  
summary(x)

##   
## Call:  
## lm(formula = Cholesterol ~ Age, data = Data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -89.773 -31.206 3.398 30.376 135.444   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 172.1531 23.0121 7.481 1.74e-09 \*\*\*  
## Age 0.7224 0.4583 1.576 0.122   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 49.89 on 46 degrees of freedom  
## Multiple R-squared: 0.05124, Adjusted R-squared: 0.03062   
## F-statistic: 2.485 on 1 and 46 DF, p-value: 0.1218

# F(1, 46) = 2.48, p = 1.74e-09 \*\*\*<0.05 . looks like there is strong relationship between age and cholesterol.  
  
# \*\*\*\* Assumption of ANCOVA III: Homogeneity of regression?  
# qualitative assesment  
graph <- ggplot(Data, aes(Age, Cholesterol, colour = State))  
graph + geom\_point(aes(shape = State), size = 3) +   
 geom\_smooth(method = "lm", aes(fill = State), alpha = 0.1) +  
 labs(x = "Age", y = "Cholesterol level")

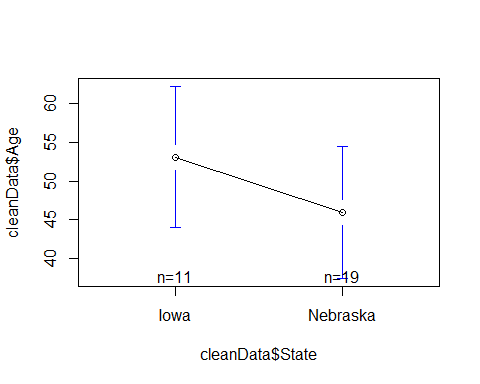
## `geom\_smooth()` using formula 'y ~ x'



## Clearly see the California is very different  
# there is negative slope between cholesterol and age in california state  
# positive relationship between cholesterol and age can be seen here in iowa and nebraska which looks normal  
  
  
  
# Quantitative assessment  
x=aov(Cholesterol~Age\*State,data=Data)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 6185 6185 4.595 0.0379 \*   
## State 2 6424 3212 2.386 0.1043   
## Age:State 2 51547 25773 19.147 1.23e-06 \*\*\*  
## Residuals 42 56536 1346   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# F(36, 2,7) = 34.6, 3.12,40.09, p = (0.0284 \*,0.2427, 0.0245 \*)<0.05   
# The interaction here is significant so homogeneity of regression is violated  
  
# Run ANCOVA (formula = (DV~CV+IV))  
  
# 1)give up: not chosen this here as its better to check further   
# 2) Interpret interaction :california changes relationship between age and cholesterol destroying positive relationships here.  
  
# 3) Removing "California" and repeat analyses  
  
k=which(Data$State!="California")  
cleanData=Data[k,]  
  
  
## Assumptions of ANCOVA I: Are the IV(state) and CV(age) correlated?  
  
# Qualitative assessment  
plotmeans(cleanData$Age~cleanData$State)



# Quantitative assessment  
  
leveneTest(cleanData$Age,cleanData$State, center = median)

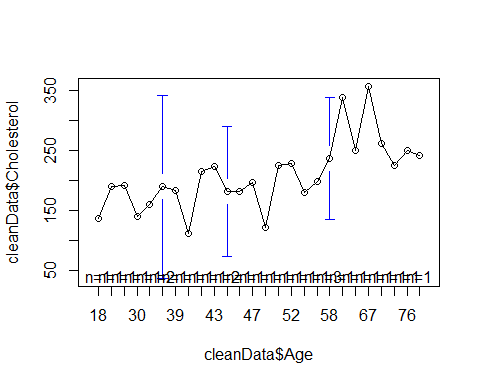
## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.9583 0.336  
## 28

#F(1,28) = 0.95, p = 0.33>0.05  
  
#anova  
  
x=aov(Age~State,data=cleanData)  
summary(x)

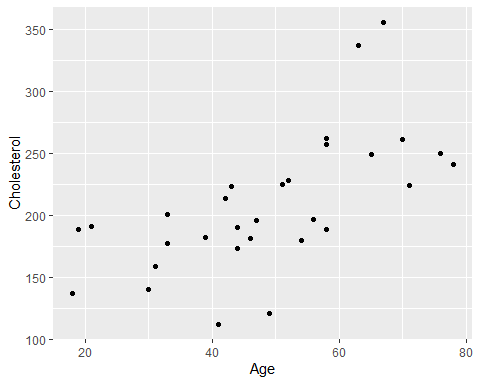
## Df Sum Sq Mean Sq F value Pr(>F)  
## State 1 356 355.5 1.346 0.256  
## Residuals 28 7394 264.1

#F(1,28) = 1.34, p = 0.25>0.05  
  
# looks like there is no relationships between state and age.  
  
#Assumption of ANCOVA II: Are DV(Cholesterol) and CV(age) correlated?  
  
# Qualitative assessment  
plotmeans(cleanData$Cholesterol~cleanData$Age)

## Warning in qt((1 + p)/2, ns - 1): NaNs produced



graph <- ggplot(cleanData,aes(Age, Cholesterol))  
graph + geom\_point()



# looks like there is significant relationship between age and cholesterol  
  
#Levene’s test  
agey<-as.factor(cleanData$Age)  
leveneTest(cleanData$Cholesterol, agey, center = median)

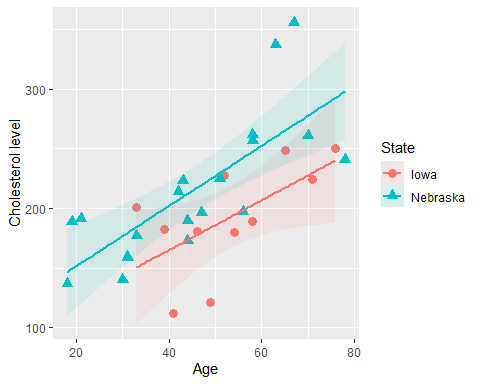
## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 25 0.0989 0.9999  
## 4

# F(25,4) = 0.0989, p = 0.9999>0.05 .There is no violation in assumption.  
  
# Quantitative assessment  
x=lm(Cholesterol ~ Age, data=cleanData)  
summary(x)

##   
## Call:  
## lm(formula = Cholesterol ~ Age, data = cleanData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -87.996 -26.570 -5.554 24.728 108.421   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.963 24.966 4.164 0.000270 \*\*\*  
## Age 2.143 0.488 4.392 0.000146 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 42.96 on 28 degrees of freedom  
## Multiple R-squared: 0.4079, Adjusted R-squared: 0.3868   
## F-statistic: 19.29 on 1 and 28 DF, p-value: 0.0001457

# F(1,28) = 19.29 , p = 0.0001457<0.05 . looks like there is significant relationship between age and cholesterol  
  
# \*\*\*\* Assumption of ANCOVA III: Homogeneity of regression?  
# qualitative assesment  
graph <- ggplot(cleanData, aes(Age, Cholesterol, colour = State))  
graph + geom\_point(aes(shape = State), size = 3) +   
 geom\_smooth(method = "lm", aes(fill = State), alpha = 0.1) +  
 labs(x = "Age", y = "Cholesterol level")

## `geom\_smooth()` using formula 'y ~ x'



# positive relationship between cholesterol and age can be seen here in iowa and nebraska which looks normal  
  
# Quantitative assessment  
x=aov(Cholesterol~Age\*State,data=cleanData)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 35606 35606 23.285 5.33e-05 \*\*\*  
## State 1 11661 11661 7.626 0.0104 \*   
## Age:State 1 262 262 0.171 0.6822   
## Residuals 26 39757 1529   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# F(1,1) = 19.29 , p = (5.33e-05 \*\*\*,0.0104 \*)<0.05 .   
#looks like there is significant relationship between age and cholesterol in two states here.  
# but for age and state p=0.68>0.05 so no correlation there.  
  
  
leveneTest(cleanData$Cholesterol, cleanData$State, center = median)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.3894 0.5377  
## 28

# F(1,28) = 0.38 , p = 0.53>0.05 . looks like there is significant relationship between cholesterol and state here.  
  
  
# Post hocs  
Data$State = as.factor(Data$State)  
  
TukeyHSD(x, which = 'State')

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: Age

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: Age,  
## State

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Cholesterol ~ Age \* State, data = cleanData)  
##   
## $State  
## diff lwr upr p adj  
## Nebraska-Iowa 39.96299 9.509856 70.41612 0.0121032

## looking into three levelsi.e.x=aov(Cholesterol~Age+State,data=Data);  
x=aov(Cholesterol ~ Age + State, data=Data)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Age 1 6185 6185 2.518 0.120  
## State 2 6424 3212 1.308 0.281  
## Residuals 44 108083 2456

# F(1,2) =(2.5,1.3) , p = (0.12,0.28)>0.05 .looks like there is no significant relationship between cholesterol and age in states here.  
  
# Post hocs  
cleanData$State = as.factor(cleanData$State)  
  
TukeyHSD(x, which = 'State')

## Warning in replications(paste("~", xx), data = mf): non-factors ignored: Age

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Cholesterol ~ Age + State, data = Data)  
##   
## $State  
## diff lwr upr p adj  
## Iowa-California -16.67429 -62.68049 29.33191 0.6561983  
## Nebraska-California 13.13716 -26.40294 52.67726 0.7014070  
## Nebraska-Iowa 29.81145 -15.73321 75.35611 0.2616160

# looking into all 3 levels in two states. Running ANCOVA (formula = (DV~CV+IV))  
x=aov(Cholesterol ~ Age + State, data=cleanData)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 35606 35606 24.022 3.97e-05 \*\*\*  
## State 1 11661 11661 7.867 0.00922 \*\*   
## Residuals 27 40019 1482   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# To get type III Sum of Squares use Anova from car package  
x=aov(Cholesterol ~ Age \* State, data=cleanData)  
summary(x)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 35606 35606 23.285 5.33e-05 \*\*\*  
## State 1 11661 11661 7.626 0.0104 \*   
## Age:State 1 262 262 0.171 0.6822   
## Residuals 26 39757 1529   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova(x,type="III")

## Anova Table (Type III tests)  
##   
## Response: Cholesterol  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 4101 1 2.6819 0.1135   
## Age 7943 1 5.1948 0.0311 \*  
## State 184 1 0.1203 0.7315   
## Age:State 262 1 0.1714 0.6822   
## Residuals 39757 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# F(1,2) =(2.5,1.3) , p = 0.0311 \*<0.05 and hence there strong relationship between age and cholesterol in two states here  
## Can also use multcomp  
library(multcomp)

## Warning: package 'multcomp' was built under R version 4.0.3

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.0.3

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.0.3

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

postHocs=glht(x, linfct = mcp(State="Tukey"))

## Warning in mcp2matrix(model, linfct = linfct): covariate interactions found --  
## default contrast might be inappropriate

summary(postHocs)

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Tukey Contrasts  
##   
##   
## Fit: aov(formula = Cholesterol ~ Age \* State, data = cleanData)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)  
## Nebraska - Iowa == 0 19.49 56.18 0.347 0.731  
## (Adjusted p values reported -- single-step method)

# F(1,2) =(2.5,1.3) ,p=0.00922 \*\*<0.05 and hence there strong relationship between age and cholesterol

Conclusion:Since Anova III gives F(1,2) =(2.5,1.3) , p = 0.0311 \*<0.05 and posthocs gives F(1,2) =(2.5,1.3) ,p=0.00922 \*\*<0.05 and hence there strong relationship between age and cholesterol. Overall it has been seen that california state showed negative slope so after excluding it it has been found that there is not strong correlation between cholesterol and states but there is relationship between age and cholesterol.Its obvious that cholesterol increases with increase in age which is normally acceptable. California showed weird results can be possibly due to food and life style causing obesity there which drives high cholesterol regardless of age.

1. An experiment was performed to examine if carriers of a genetic mutation have an increased probability of a schizophrenia diagnosis. Please look at the following data and run the appropriate statistical test to answer this question.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Schizophrenia Diagnosis** | | |
| **Genetic Mutation** |  | ***No*** | ***Yes*** |
| ***No*** | 25 | 65 |
| ***Yes*** | 60 | 50 |

# Hypothesis:  
#H0:Mutation=SCh\_Diag (NO effect of genetic mutation on Schizophrenia)  
#H1:Mutation!=SCh\_Diag (genetic mutation on Schizophrenia diagnosis)  
  
## creating data frame  
Mutation\_No<-c(25,65)  
Mutation\_YES<-c(60,50)  
  
# Chi-square test   
  
chisq.test(Mutation\_YES,p= Mutation\_No, rescale.p = TRUE)

##   
## Chi-squared test for given probabilities  
##   
## data: Mutation\_YES  
## X-squared = 39.287, df = 1, p-value = 3.659e-10

# X-squared = 39.287, df = 1, p-value = 3.659e-10<0.05  
## Checking expected values  
model=chisq.test(Mutation\_YES,p= Mutation\_No, rescale.p = TRUE)  
model$expected

## [1] 30.55556 79.44444

# here expected values are >5 and no issue  
# Test of independence:  
#install.packages("vcd")  
library(vcd)

## Warning: package 'vcd' was built under R version 4.0.3

## Loading required package: grid

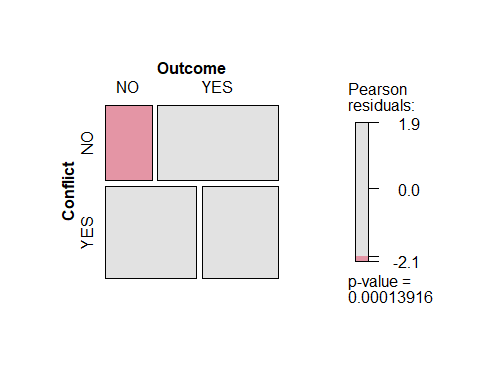
## Create a table manually based on the data set from Scientific American   
Table = matrix(c(25,65,60,50), 2, 2, byrow=TRUE)  
dimnames(Table) = list("Conflict"=c("NO", "YES"),"Outcome"=c("NO", "YES"))  
res = chisq.test(Table, correct=FALSE)  
res

##   
## Pearson's Chi-squared test  
##   
## data: Table  
## X-squared = 14.513, df = 1, p-value = 0.0001392

# X-squared = 14.513, df = 1, p-value = 0.0001392<0.05   
res$expected

## Outcome  
## Conflict NO YES  
## NO 38.25 51.75  
## YES 46.75 63.25

# Expected values are >10 so no issue again  
# All counts are >10 here   
mosaic(Table, shade=T, legend=TRUE)



Conclusion: Since p=0.0001392<0.05 (very small), it looks strongly statistically significant and hence there is increase in probability of a schizophrenia diagnosis by genetic mutation and hence null hypothesis here is rejected.