AMOD5240H - Term Project

Suraj Suresh Sajjan

December 11, 2018

Introduction to Analysis of Covarience(ANCOVA)

General Purpose and Description

ANCOVA belongs to the branch of Statistics that deals with comparision of means. The main purpose of this branch, as the name suggests, is to check if the means across various groups are equal. ANCOVA is an extension of ANOVA. It checks if the population mean of the dependent variable are the same across the levels of the independent variable, by adjusting to the defferences seen on the covariates. In other words, it is used to check if the adjusted group means vary considerable with repect to each other.

A one-way ANCOVA consists of three essential components, an independent variable or a factor, a dependent variable and a covariate. The factor divides the individuals into two or more levels or categories. The dependent variable and the covariate help in differentiating the individuals on quantitative dimensions.

The control and explaination of the variation in dependent variable can be done by either experimental control with the use of research design, or by statistical control, with the use of analysis of covarience. ANCOVA is primarily used as a procedure for statistical control of an irrelevant variable called covariate. It is is a combination of regression analysis and analysis of variance(ANOVA) done to contol the effects of covariate. The researcher can better estimate the effects of primary dependent variable by this way. The ANCOVA F-test as well as the p-value can be used to judge if the population means in the dependent variable, after being adjusted for differences on the covariates, differ across the categories.

Fundamental equations

Let us consider an example where we study the effects of viagra dosage on a person's libido. If we think about things other than Viagra that might influence a person's libido, then the first obvious answer will be the libido of the person's partner. There could be other factors also such as fatigue and medication. In this study, we consider the partner's libido as the covariate. If we enter the covariate first in the regression model, and then enter the variables that represent experimental manipulations, we can observe what effect the independent variable has after the effects of the covariate have been accounted for. This way, we partial out the effect of the covariate.

The effect of an experiment is assessed by comparing the variability of data the experiment can explain to the variabilty of data that it cannot explain. If we are able to exlain some of this "uneplained variance(SS_R)" with the help of the covariates, we shall be able to reduce the error variance. Effectively, this allows us to measure the effects due to independent variable(SS_M) more accurately.

The equation for the participant's libido is given as: $\text{libido}_i = b_o + b_3\text{partner's libido}_i + b_2\text{high}_i + b_1\text{low}_i + \epsilon_i$

Summary table containing formulae

For Covariate:
Sum of Squares = SS_{Cov}
Degrees of Freedo = 1
Mean Square = MS_{Cov}
F-Ratio =

$$\frac{MS\ Cov}{MS'\ w}$$

For Between groups:
Sum of Squares = SS'_B
Degrees of Freedo = K-1
Mean Square = MS'_B =

$$\frac{SS'\ B}{K-1}$$

F-Ratio =

$$\frac{MS'\ B}{MS'\ w}$$

For Within the group:
Sum of Squares = SS'_W
Degrees of Freedo = N-K-1
Mean Square = MS'_w =

$$\frac{SS'\ W}{N-K-1}$$

For Total:
Sum of Squares = SS'_T
Degree of Freedom = N-1

Assumptions in ANCOVA

1. Independence of covariate and the treatment effect.

Let us consider three equations.

Equation1 Total variance in libido = variance explained by viagra + unexplained variance.

Equation2 Total variance in libido = variance explined by viagra + unexplained variance, where unexplained variance is divided into two partitions (variance due to covariate + rest of unexplained variance).

Equation3 Total variance in libido = variance explined by viagra + unexplained variance (here the variance explained by covariate partially overlaps the variance explained by viagra and partially the unexplained variance)

We would want the scenerio seen in equation 2 and not equation 3, as in equation 3, the variance due to independent variable is being overlapped due to dependency of covarite on it which is undesired. Hence we

need to make sure that the indepedent variable and the covariate are independent of each other.

Example where the independent variable is independent of the covariate:

Let us consider the mtcars dataset. Let horse power be the covariate, auto-manual be the independent variable and miles per gallon be the dependent variable.

```
data(mtcars)
anova(lm(hp ~ am, data = mtcars))
```

```
## Analysis of Variance Table
##
## Response: hp
##           Df Sum Sq Mean Sq F value Pr(>F)
## am         1   8619  8619.5   1.886 0.1798
## Residuals 30 137107  4570.2
```

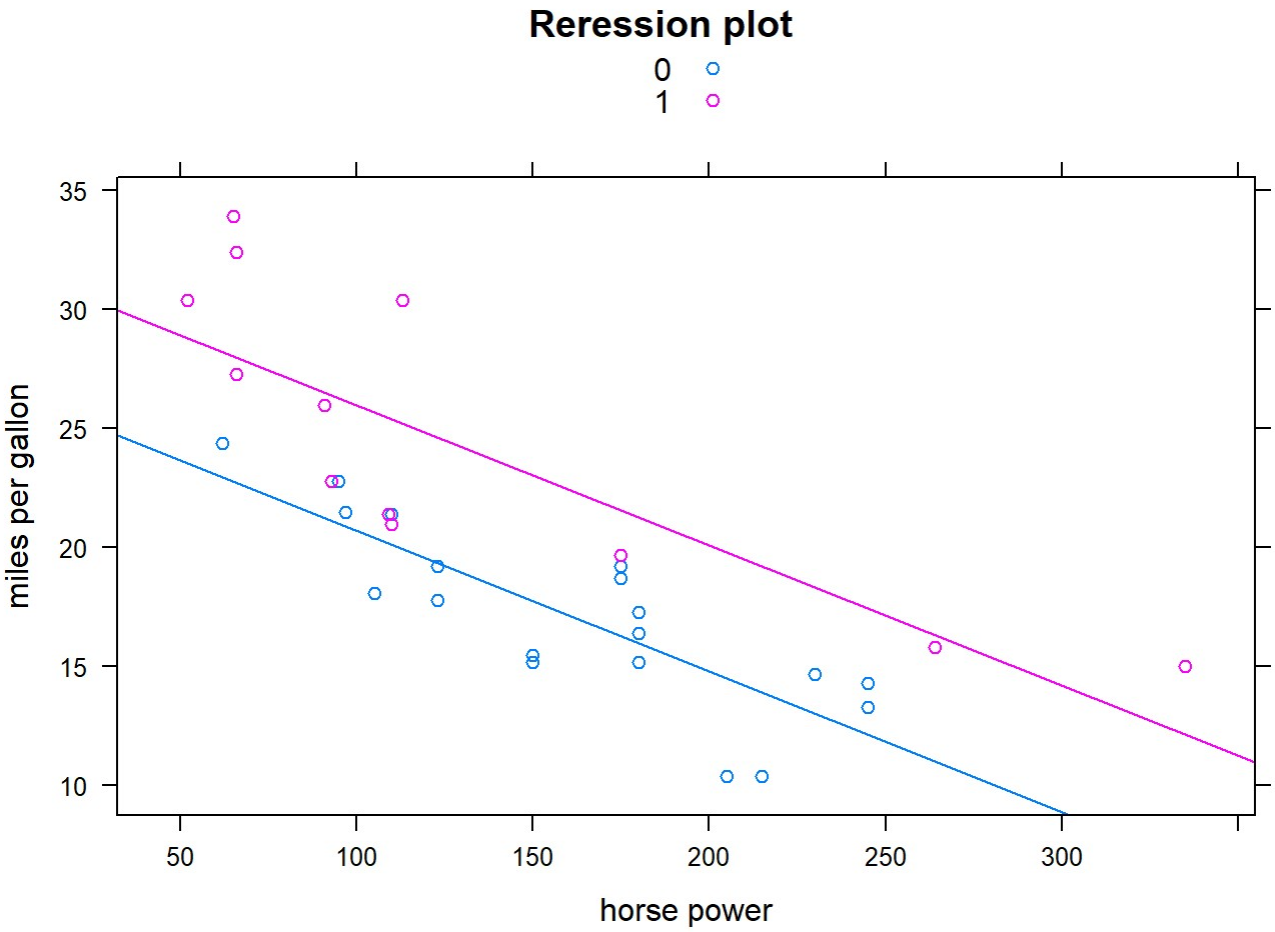
Null Hypothesis: The mean of horse power is same for both the transmission types.
Alternate Hypothesis: At least one of the mean is different from rest of the groups.
Here, we get a p-value of 0.1798 which is greater than the alpha of 0.05. Hence, we observe that the mean is same for both transmission groups. Hence we fail to reject null hypothesis and conclude that horse power can be used as a covariate.

2. Homegenity of regression slopes

When we conduct an ANCOVA, we observe the overall relationship between the dependent variable and the covariate. We fit a regresssion line to the entire dataset, irrespective of the group the individual belongs to. By fitting the overall model, we assume that for all the groups in the study, the relaionship is true. However, if the relationship is different for even one of the group, the homogeneity of regression slopes gets invalid, hence the regression model could be inaccurate. The best way to relaise this is by plotting a scatterplot. If we see that the regression lines of all the groups are almost parallel to each other, then the homogeneity of regression slopes is satisfied.

Below is an example for mtcars dataset:

```
data(mtcars)
xyplot(mpg~hp, data=mtcars, groups=am, type=c("p","r"), auto.key=TRUE, xlab="horse po
wer", ylab="miles per gallon", main = "Reression plot")
```



In the above plot, we observe that the slopes of the regression lines for both automatic(represented by 0) and manual(represented by 1) transmissions are almost parallel. Hence, the homogeneity of regression slopes is satisfied.

Null and alternate hypothesis in ANCOVA

The null and alternate hypothesis are very much similar to those of an ANOVA.

NUll hypothesis: There is no difference in means of the groups, even after being adjusted to the effects of covariate.
Alternate hypothesis: At least one of the mean in the groups is different from rest of the groups, after being adjusted to the effects of covariate.

R Implementation: ANCOVA using R

Packages for ANCOVA in R

In this test, we will be using car package(for Levene's test, type III sum of squares), effects(for adjusting means), multcomp(for post hoc tests), pastecs(for descriptive statistics), compute.es(for effect sizes),lattice(for plotting data).

Summary of the data

The dataset contains 3 variables.

- A Categorical variable called dose, where the dosage is categorised into group 1,2 and 3 which are further renamed as placebo, low dose and high dose respectively.
- Libido variable which is a continuous variable which contains libido values of the participant.
- Partner's libido variable, which is a continuous variable containing libido values of the partner. This is treated as the covariate.

General procedure for ANCOVA

1. Read the data file.

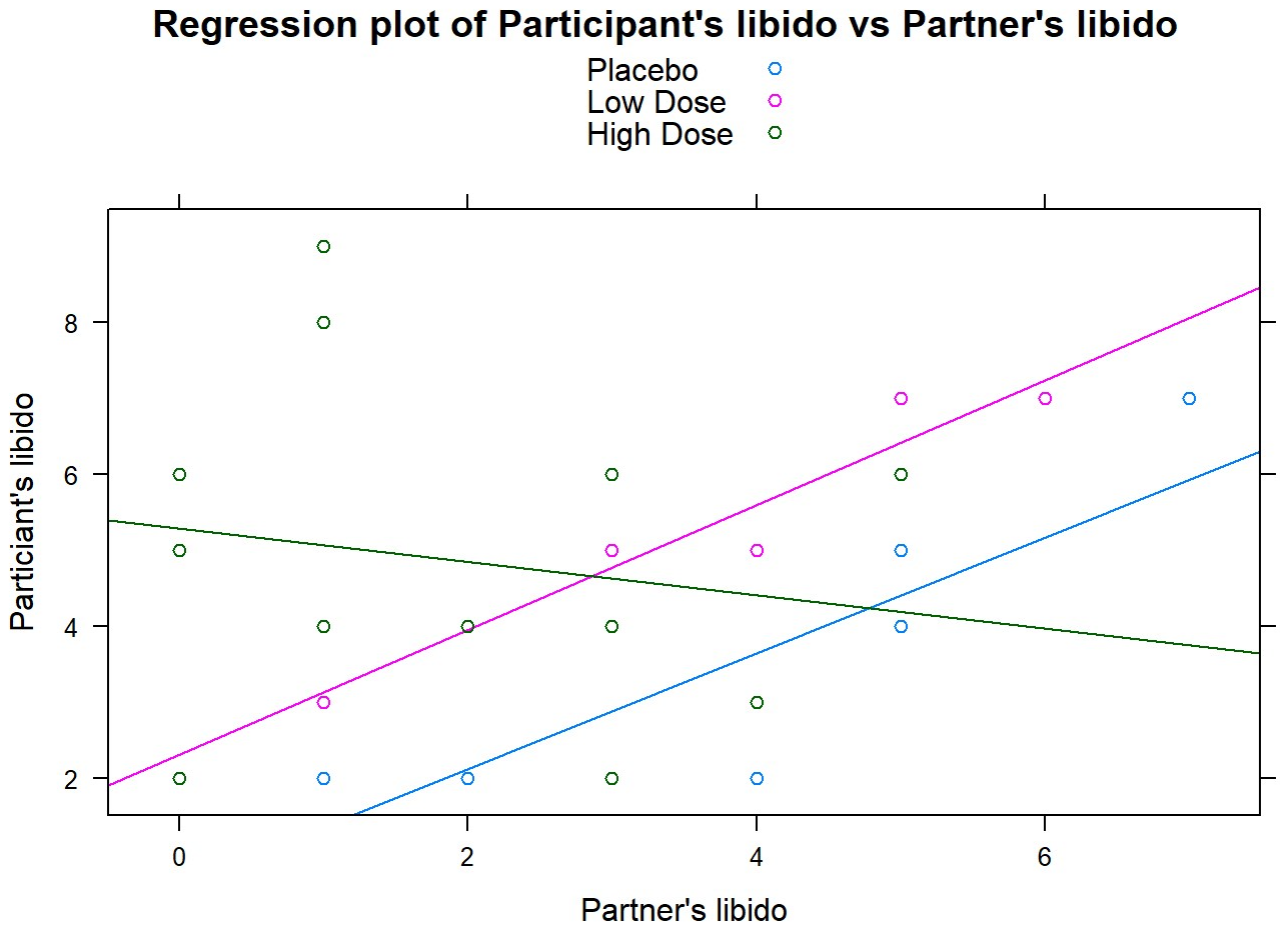
```
#Loading the libraries
library(car); library(compute.es); library(effects); library(ggplot2);
library(multcomp); library(pastecs); library(WRS); library(lattice)

#Reading the data file
viagraData<-read.delim("D:/AMOD/AMOD-5240H-A-Statistical aspects of modeling/FInal te
rm project/ViagraCovariate.dat", header = TRUE)

#Converting the dose into categorical data and renaming them
viagraData$dose<-factor(viagraData$dose, levels = c(1:3), labels = c("Placebo", "Low
Dose", "High Dose"))
```

2. Explore the data file: graph the data, compute mean and variance, use Leven's test to check for homogeneity of variance.

```
#Exploring the data
xyplot(libido~partnerLibido, data=viagraData, groups=dose, type=c("p","r"), auto.key=
TRUE, xlab="Partner's libido", ylab="Particiant's libido", main = "Regression plot of
Participant's libido vs Partner's libido")
```



In the above plot, slope of the lines of placebo and low dose are almost equal. This shows the homogeneity of regression slopes.
Whereas, for high dose, we observe that there does not seem to be any relation between the partner's and participant's libido. This could lead us to doubt the presence of homogeneity of regression lines.

Heterogeneity of regression slopes does not necessarily mean a bad thing. If the homogeniety of regression slopes is violated, the it would account for an interesting hypothesis by itself. In such a case, the variation can be modeled explicitly by the use of "Multilevel linear models". The topic is beyond the scope of this chapter.

```
#Compute mean and variance for partner's as well as participant's libido.
libido.means<- tapply(viagraData$libido,viagraData$dose,FUN=mean)
cat("The mean values for libido in each dose group:", libido.means, sep="\n")
```

```
## The mean values for libido in each dose group:
## 3.222222
## 4.875
## 4.846154
```

```
libido.variance<- tapply(viagraData$libido,viagraData$dose,FUN=var)
cat("The variance values for libido in each dose group:", libido.variance, sep="\n")
```

```
## The variance values for libido in each dose group:
## 3.194444
## 2.125
## 4.474359
```

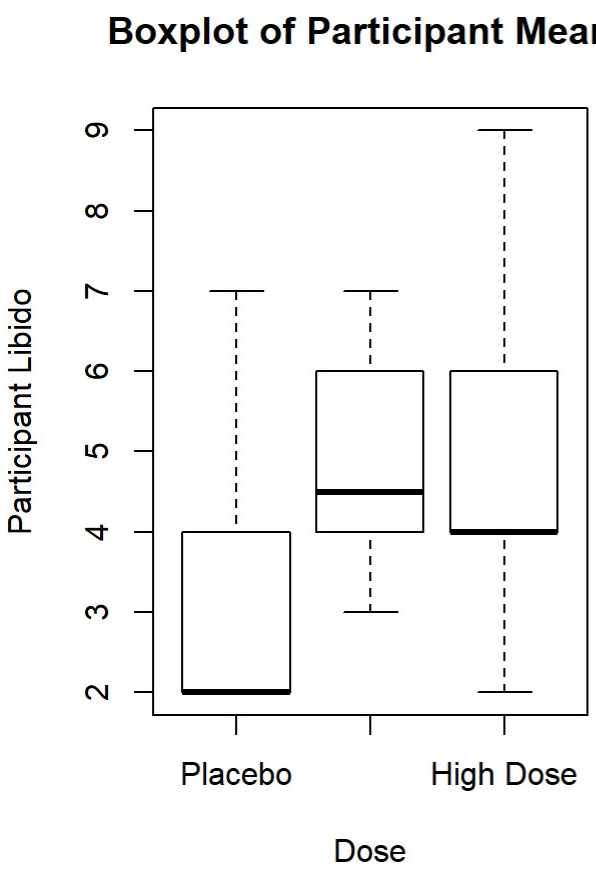
```
partner.libido.means<- tapply(viagraData$partnerLibido,viagraData$dose,FUN=mean)
cat("The mean values for partner's libido in each dose group:", partner.libido.means,
sep="\n")
```

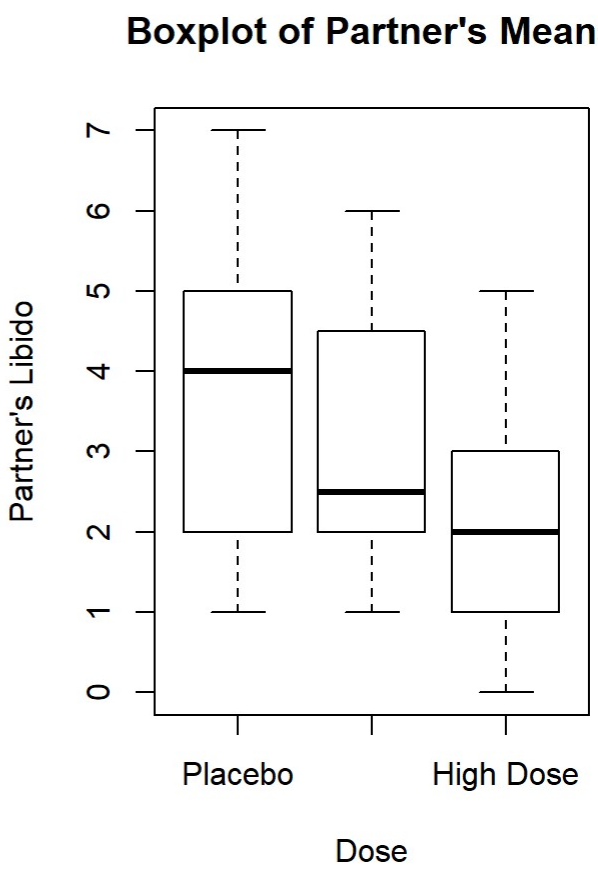
```
## The mean values for partner's libido in each dose group:
## 3.444444
## 3.125
## 2
```

```
partner.libido.variance<- tapply(viagraData$partnerLibido,viagraData$dose,FUN=var)
cat("The variance values for partner's libido in each dose group:", partner.libido.va
riance, sep="\n")
```

```
## The variance values for partner's libido in each dose group:
## 4.277778
## 2.982143
## 2.666667
```

```
#Plot the means for partner's as well as the prticipant's libido.
par(mfrow=c(1,2))
plot(x=viagraData$dose, y=viagraData$libido, xlab = "Dose", ylab = "Participant Libid
o", main = "Boxplot of Participant Means")
plot(x=viagraData$dose, y=viagraData$partnerLibido, xlab = "Dose", ylab = "Partner's
Libido", main = "Boxplot of Partner's Means")
```






```
#Levene's test
leveneTest(viagraData$libido, viagraData$dose, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.3256 0.7249
##       27
```

For the Levene's test:

Null Hypothesis: The Variance is constant across the groups.
Alternate hypothesis: The Variance of at least one of the group is different.

From the above p-value of 0.725, we observe that the value is much greater than alpha of 0.05, hence we fail to reject null hypothesis and state that the variance is similar across the groups.

3. Check if the independent variable and the covariate are independent of each other: Perform an ANOVA taking covariate as the outcome and the independent variable as the predictor. If the p value is greater than 0.05, we say that the independent variable and covariate are independent.

```
anova(lm(partnerLibido ~ dose, data = viagraData))
```

```
## Analysis of Variance Table
##
## Response: partnerLibido
##           Df Sum Sq Mean Sq F value Pr(>F)
## dose       2 12.769  6.3847  1.9793 0.1577
## Residuals 27 87.097  3.2258
```

Null Hypothesis: The mean of partner's libido is same across all 3 viagra groups.
Alternate Hypothesis: At least one of the mean is different from rest of the groups.
Here, we get a p-value of 0.1577 which is greater than the alpha of 0.05. Hence, we observe that the mean is same across all three viagra groups. Hence we fail to reject null hypothesis and conclude that partner's libido can be used as a covariate.

4. Do the ANCOVA: Run the main analysis of covariance.

In our data set, N = 30, K = 3

Order matters!

```
covariateFirst <- aov(libido ~ partnerLibido + dose, data = viagraData)
doseFirst <- aov(libido ~ dose + partnerLibido, data = viagraData)
summary(covariateFirst)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## partnerLibido  1   6.73   6.734   2.215 0.1487
## dose           2  25.19  12.593   4.142 0.0274 *
## Residuals     26  79.05   3.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(doseFirst)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## dose           2  16.84   8.422   2.770 0.0812 .
## partnerLibido  1  15.08  15.076   4.959 0.0348 *
## Residuals     26  79.05   3.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above models, we observe that by changing the order of the independent variable and the covariate, the ancova analysis varies.

When we take covariate first, we observe that this models implies that the covariate's (partner's libido) effect on the participant's libido is not significant but that of the dose is.
When we take dose first, we observe that this models implies that the dose's effect on the participant's libido is not significant but that of the the covariate's (partner's libido) is.

This is observed because R, by default uses sequential or type I sum of squares where the predictor is evaluated only after any other predictor before it is evaluated. In our first model, partner's libido is evaluated, it

is considered as the only term, In second model, dose is evaluated first and then partner's libido.

To overcome this issue, we use type III sum of squares. In this case, the effects are evaluated by considering the effects of all others in the model. We also introduce contrasts in this model, withought which the type III sum of squares will not be accurate.

Assigning the weights is done in the fllowing manner:

Contrast1 = Positive Chunk 1 vs Negative chunk 2, Where chunk 1 contains high dose and low dose, chunk 2 contains placebo. We assign 1 to high dose, 1 to low dose in chunk 1 and -2 to placebo, because we need sum of both chunks to be zero. This makes the weights (1,1,-2)

Contrast2 = Here the positive chunk above is further divided into Positive Chunk 1 vs Negative chunk 2. Here, chunk 1 has low dose and chunk 2 has high dose. 1 is assigned to chunk 1 and -1 is assigned to chunk 2. 0 is assigned to placebo. This makes the weights (1,-1,0)

Now, we assign Contrasts to set the contrast for dose and run the ancova model as shown below:

```
contrasts(viagraData$dose)<-cbind(c(-2,1,1), c(0,-1,1))

viagraModel<-aov(libido ~ partnerLibido + dose, data = viagraData)
Anova (viagraModel, type= "III")
```

```
## Anova Table (Type III tests)
##
## Response: libido
##               Sum Sq Df F value    Pr(>F)
## (Intercept)   76.069  1 25.0205 3.342e-05 ***
## partnerLibido 15.076  1  4.9587   0.03483 *
## dose          25.185  2  4.1419   0.02745 *
## Residuals     79.047 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the main ANCOVA model

If we plot an anova of effect of viagra on libido, we get the below model:

```
anova(lm(libido ~ dose, data = viagraData))
```

```
## Analysis of Variance Table
##
## Response: libido
##           Df Sum Sq Mean Sq F value Pr(>F)
## dose       2 16.844  8.4219  2.4159 0.1083
## Residuals 27 94.123  3.4860
```

For the above model,
Null Hypothesis: The dose of viagra has no effet on the person's libido. i.e., the mean is equal across the groups.

Alternate Hypothesis: The mean of at least one of the dose groups is different from that of the others.

Looking at the p-value of 0.1083, which is greater than 0.05, we would have concluded that the viagra has no effect on participant's libido.

But after taking into consideration the covariate as partner's libido in the viagraModel,

Null Hypothesis: The dose of viagra has no effect on the person's libido after considering the effects of covariate, the mean is equal across the groups.
Alternate Hypothesis: The mean of at least one of the dose groups is different from that of the others.

When the effect of partner's libido is removed or controlled for, the p value of dose is 0.027. Hence, we reject the null hypothesis. We observe that viagra has a significant effect on a participant's libido.

The adjusted means can be calcualted as:

```
adjustedMeans<-effect("dose", viagraModel, se=TRUE)
summary(adjustedMeans)
```

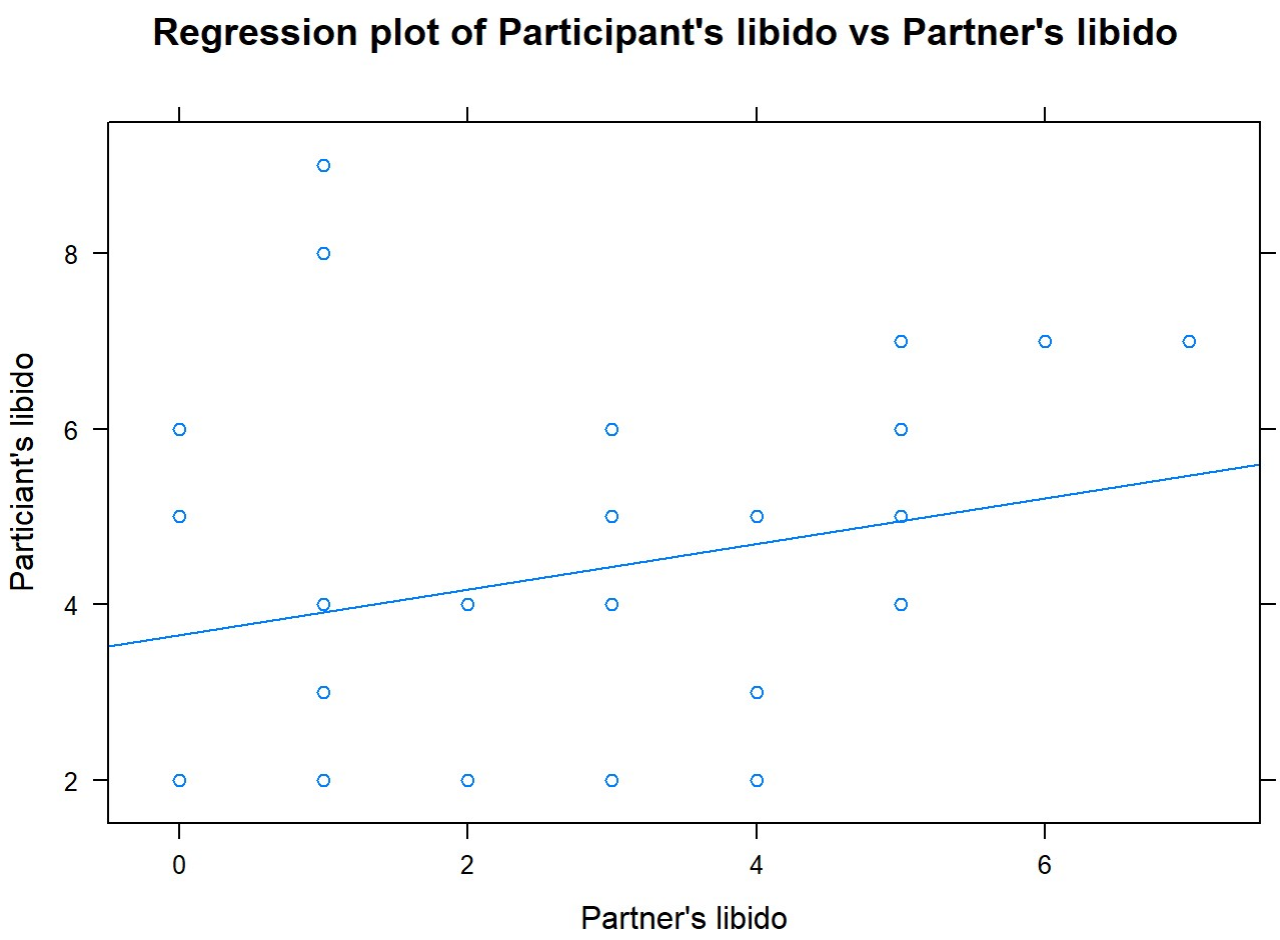
```
##
##  dose effect
## dose
##   Placebo  Low Dose High Dose
##  2.926370  4.712050  5.151251
##
##  Lower 95 Percent Confidence Limits
## dose
##   Placebo  Low Dose High Dose
##  1.700854  3.435984  4.118076
##
##  Upper 95 Percent Confidence Limits
## dose
##   Placebo  Low Dose High Dose
##  4.151886  5.988117  6.184427
```

From the adjusted means, unlike the means we saw earlier, we can clearly see that the mean values go on increasing as the dose increase.

Interpreting the coveriate

Construct a scatterplot of Participant's libido against partner's libido:

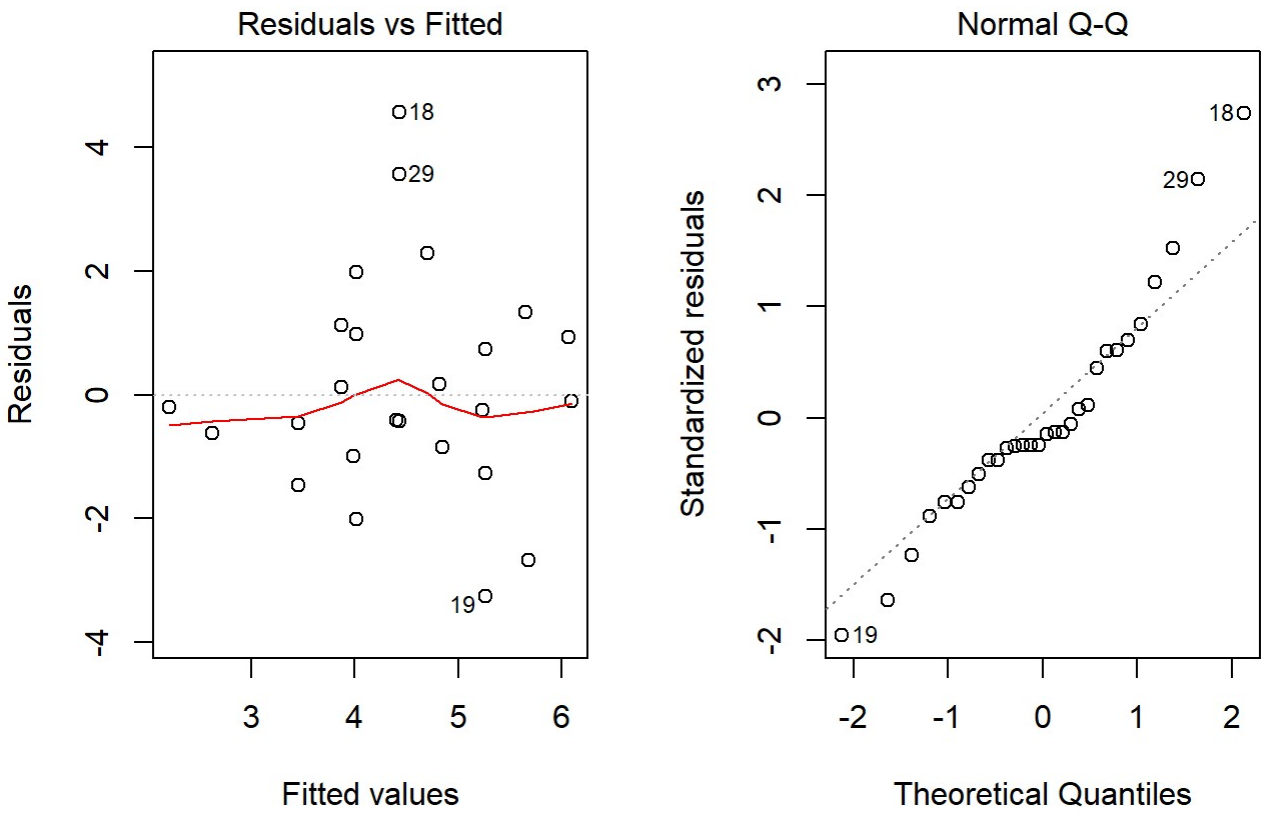
```
xyplot(libido~partnerLibido, data=viagraData, type=c("p","r"), auto.key=TRUE, xlab="P
artner's libido", ylab="Particiant's libido", main = "Regression plot of Participant'
s libido vs Partner's libido")
```



From the above scatterplot, we observe the effect of covariate on participant's libido is that, as the Partner's libido increases, so does that of the participant.

Plots in ANCOVA

```
par(mfrow=c(1,2))
plot(viagraModel, which = 1)
plot(viagraModel, which = 2)
```



From the normal QQ-plot, we observe that the residuals are almost normally distributed as they follow the line. From the Residuals vs fitted plot, we observe that the points have no particular pattern, the line is almost parallel to the x axis.

References

[1] file - https://github.com/lawrence009/dsur/blob/master/data/ViagraCovariate.dat (https://github.com/lawrence009/dsur/blob/master/data/ViagraCovariate.dat)
[2] text book - Discovering Statistics using R by Andy Field
[3] http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/ancova/understanding%20ancova.pdf (http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/ancova/understanding%20ancova.pdf)