Statistical Computing Final

Osman Batuhan Şahin

29 05 2022

Including Libraries

```
library(dplyr)
library(stringr)
library(corrplot)
library(ggplot2)
library(caret)
library(gridExtra)
library(imputeTS)
library(MASS)
library(RVAideMemoire)
library(car)


options(warn=-1)
```

1) Data Description

I found my dataset on on Kaggle. This dataset classifies patients according to their labels using biomechanical features. Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (each one is a column):

pelvic incidence pelvic tilt lumbar lordosis angle sacral slope pelvic radius grade of spondylolisthesis

column3Cweka.csv is the file with three class labels: Normal, Disk Hernia, Spondylolisthesis.

column2Cweka.csv is the file with two class labels: Normal, Abnormal.

```
data2c = read.csv("C:/Users/batuh/Desktop/r/finalsc/column_2C_weka.csv")
data3c = read.csv("C:/Users/batuh/Desktop/r/finalsc/column_3C_weka.csv")

head(data2c)
```

```
##   pelvic_incidence pelvic_tilt.numeric lumbar_lordosis_angle sacral_slope
## 1         63.02782           22.552586                39.60912     40.47523
## 2         39.05695           10.060991                25.01538     28.99596
## 3         68.83202           22.218482                50.09219     46.61354
## 4         69.29701           24.652878                44.31124     44.64413
## 5         49.71286            9.652075                28.31741     40.06078
## 6         40.25020           13.921907                25.12495     26.32829
##   pelvic_radius degree_spondylolisthesis    class
## 1      98.67292               -0.254400 Abnormal
## 2     114.40543                4.564259 Abnormal
## 3     105.98514               -3.530317 Abnormal
## 4     101.86850               11.211523 Abnormal
## 5     108.16872                7.918501 Abnormal
## 6     130.32787                2.230652 Abnormal
```

```
head(data3c)
```

```
##   pelvic_incidence pelvic_tilt lumbar_lordosis_angle sacral_slope pelvic_radius
## 1         63.02782   22.552586                39.60912     40.47523     98.67292
## 2         39.05695   10.060991                25.01538     28.99596    114.40543
## 3         68.83202   22.218482                50.09219     46.61354    105.98514
## 4         69.29701   24.652878                44.31124     44.64413    101.86850
## 5         49.71286    9.652075                28.31741     40.06078    108.16872
## 6         40.25020   13.921907                25.12495     26.32829    130.32787
##   degree_spondylolisthesis  class
## 1               -0.254400 Hernia
## 2                4.564259 Hernia
## 3               -3.530317 Hernia
## 4               11.211523 Hernia
## 5                7.918501 Hernia
## 6                2.230652 Hernia
```

2) EDA

```
dim(data2c)
```

```
## [1] 310   7
```

```
str(data2c)
```

```
## 'data.frame':    310 obs. of  7 variables:
##  $ pelvic_incidence        : num  63 39.1 68.8 69.3 49.7 ...
##  $ pelvic_tilt.numeric     : num  22.55 10.06 22.22 24.65 9.65 ...
##  $ lumbar_lordosis_angle   : num  39.6 25 50.1 44.3 28.3 ...
##  $ sacral_slope            : num  40.5 29 46.6 44.6 40.1 ...
##  $ pelvic_radius           : num  98.7 114.4 106 101.9 108.2 ...
##  $ degree_spondylolisthesis: num  -0.254 4.564 -3.53 11.212 7.919 ...
##  $ class                   : chr  "Abnormal" "Abnormal" "Abnormal" "Abnormal" ...
```

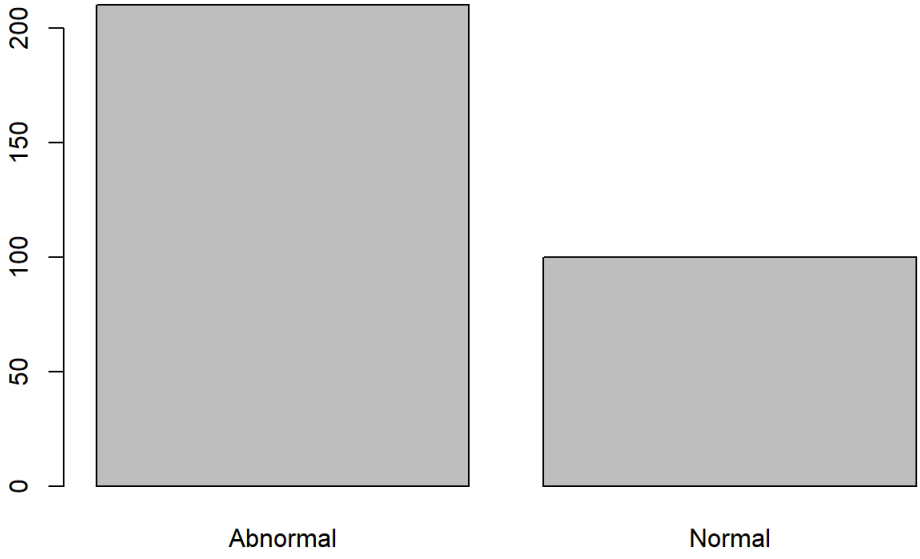
```
data2c=na_mean(data2c)
data3c=na_mean(data3c)
```

Dataset contains 310 row and 7 column. All columns are numeric except class which is a character, i will put it as a factor.

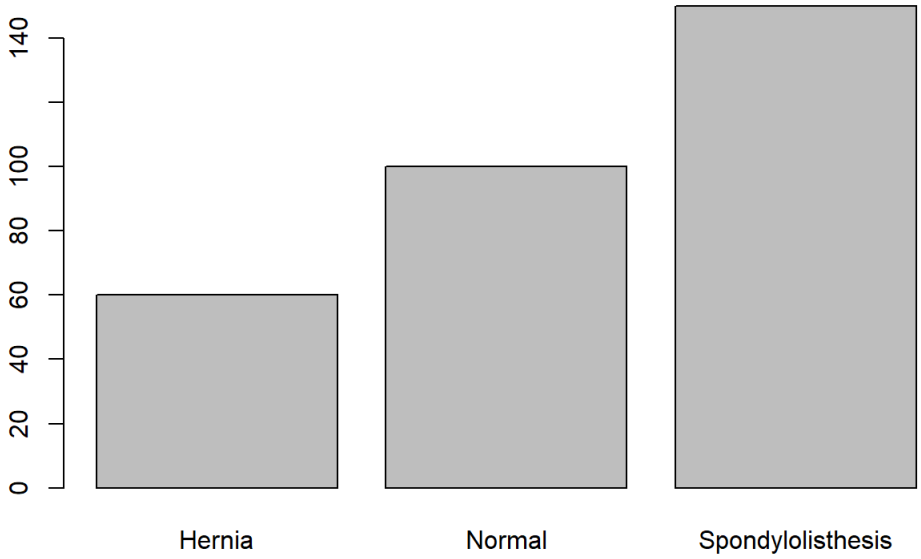
```
data2c$class<-as.factor(data2c$class)
data3c$class<-as.factor(data3c$class)
```

3) Data Visualization

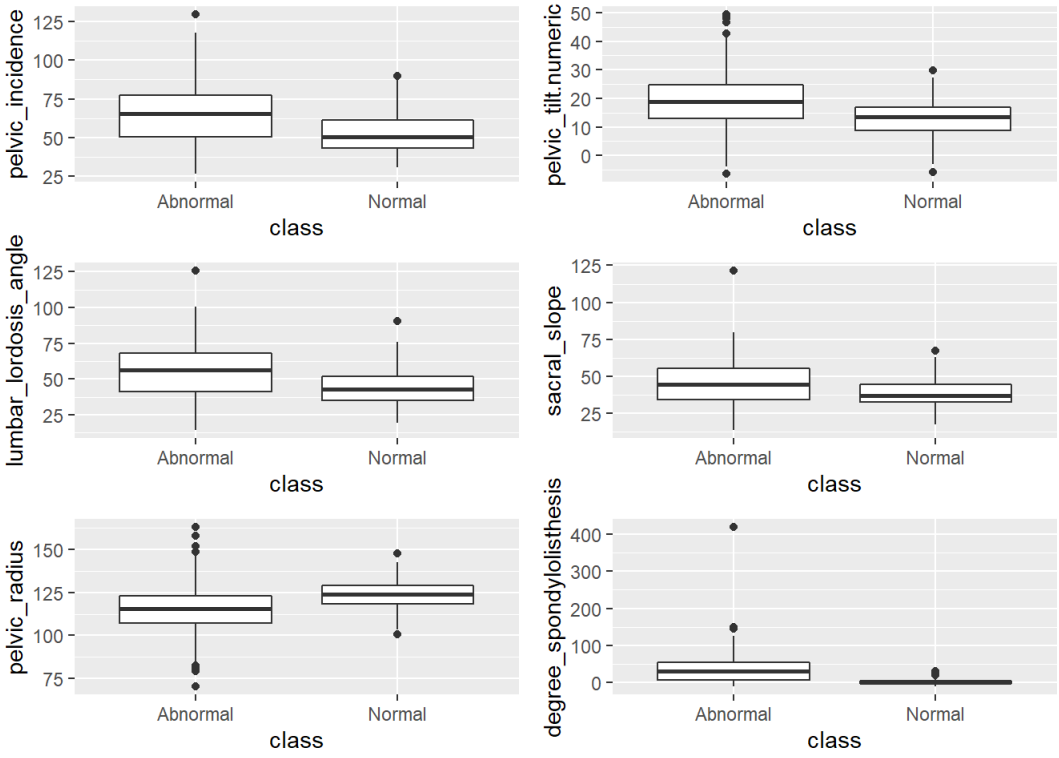
```
barplot(table(data2c$class))
```




```
barplot(table(data3c$class))
```




```
grid.arrange(ggplot(data2c, aes(x=as.factor(class), y=pelvic_incidence)) + geom_boxplot() + xlab("class"),
ggplot(data2c, aes(x=as.factor(class), y=pelvic_tilt.numeric)) + geom_boxplot() + xlab("class"),
ggplot(data2c, aes(x=as.factor(class), y=lumbar_lordosis_angle)) + geom_boxplot() + xlab("class"),
ggplot(data2c, aes(x=as.factor(class), y=sacral_slope)) + geom_boxplot() + xlab("class"),
ggplot(data2c, aes(x=as.factor(class), y=pelvic_radius)) + geom_boxplot() + xlab("class"),
ggplot(data2c, aes(x=as.factor(class), y=degree_spondylolisthesis)) + geom_boxplot() + xlab("class"))
```



Barplot shows frequency of each class. Boxplots shows that degree spondylolisthesis is the most important of the variables to explain the normal and abnormal of the patients.

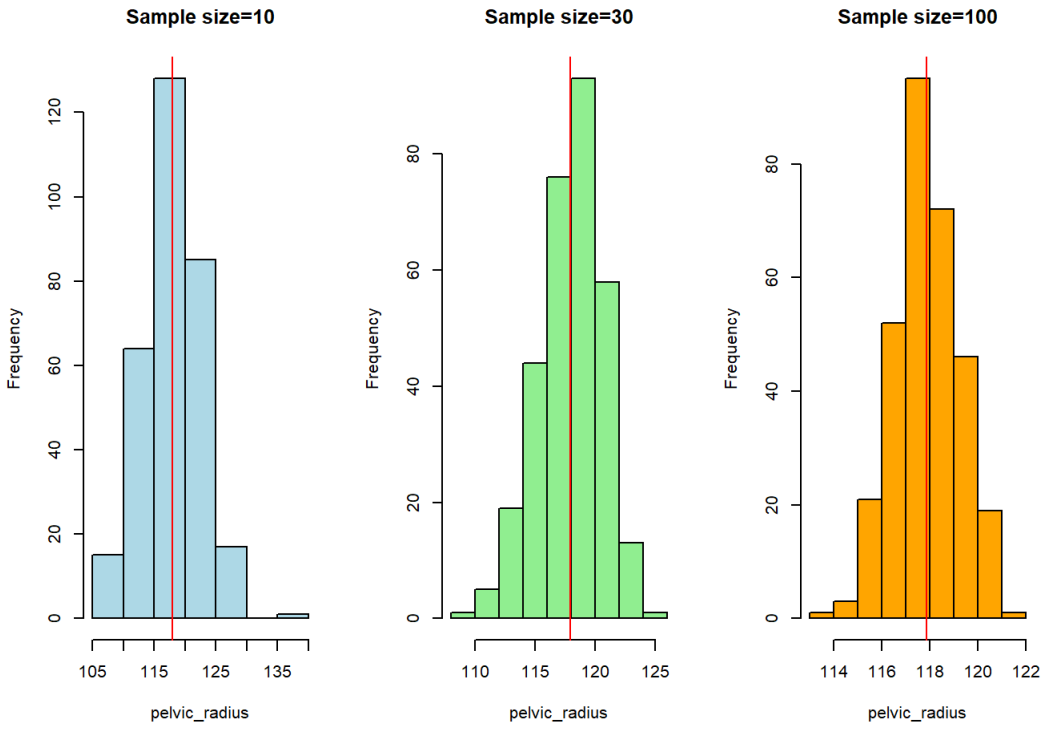
4) Central Limit Theorem

I chose pelvic_radius column. We will take sample size=10, 30 & 100 samples=310 Calculate the arithmetic mean and plot the mean of sample 310 times

```
s10 <- c()
s30 <- c()
s100 <- c()
n =310
for ( i in 1:n){
s10[i] = mean(sample(data2c$pelvic_radius,10, replace = TRUE))
s30[i] = mean(sample(data2c$pelvic_radius,30, replace = TRUE))
s100[i] = mean(sample(data2c$pelvic_radius,100, replace = TRUE))
}
par(mfrow=c(1,3))
hist(s10, col ="lightblue",main="Sample size=10",xlab ="pelvic_radius")
abline(v = mean(s10), col = "red")

hist(s30, col ="lightgreen", main="Sample size=30",xlab ="pelvic_radius")
abline(v = mean(s30), col = "red")

hist(s100, col ="orange",main="Sample size=100",xlab ="pelvic_radius")
abline(v = mean(s100), col = "red")
```



Sampling distribution approaches normal distribution as the sample sizes increase. Therefore, we can consider the sampling distributions as normal.

5) Confidence Intervals

```
model <- lm(pelvic_radius ~ 1, data2c)
```

```
confint(model, level=0.95)
```

```
##                2.5 %  97.5 %
## (Intercept) 116.4324 119.409
```

We are 95% confident that main of pelvic_radius between 116.4324 and 119.409.

```
confint(model, level=0.99)
```

```
##                0.5 %   99.5 %
## (Intercept) 115.9603 119.8811
```

We are 99% confident that main of pelvic_radius between 115.9603 and 119.8811. Confidence interval range grows when level grows.

6) Transformation

Shapiro-Wilk normality test to all columns.

```
df = data2c[-c(7) ]
apply(df,2,shapiro.test)
```

```
## $pelvic_incidence
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.97112, p-value = 7.132e-06
##
##
## $pelvic_tilt.numeric
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.96639, p-value = 1.321e-06
##
##
## $lumbar_lordosis_angle
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.97181, p-value = 9.221e-06
##
##
## $sacral_slope
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.96398, p-value = 5.887e-07
##
##
## $pelvic_radius
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.98872, p-value = 0.01661
##
##
## $degree_spondylolisthesis
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.69698, p-value < 2.2e-16
```

We can say all columns are not normally distributed with 0.95 confidence level.

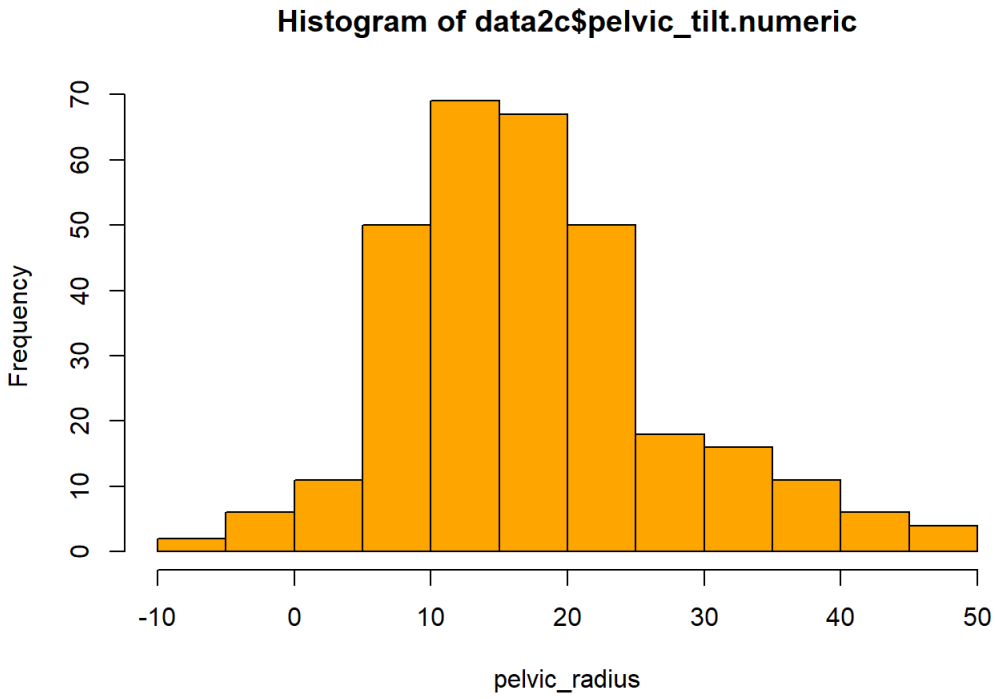
```
data2c$sqrt_pelvic_tilt.numeric = sqrt(data2c$pelvic_tilt.numeric)

shapiro.test(data2c$sqrt_pelvic_tilt.numeric)
```

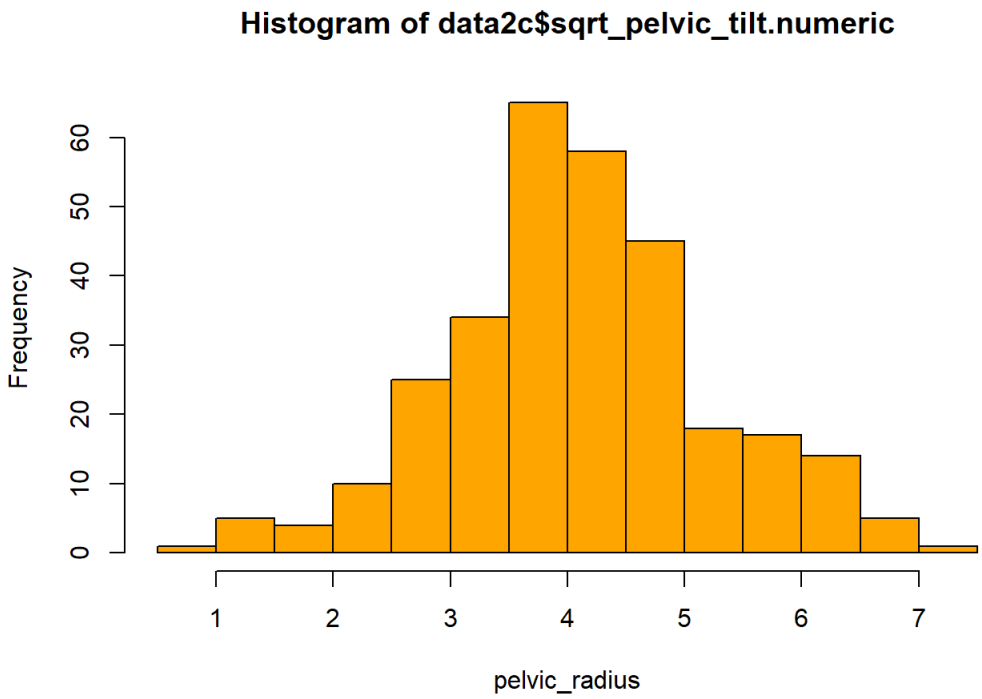
```
##
##  Shapiro-Wilk normality test
##
## data:  data2c$sqrt_pelvic_tilt.numeric
## W = 0.99174, p-value = 0.09009
```

p value increased for pelvic_tilt.numeric after sqrt transformation. Histograms before and after shows transformed data is more normally distributed.

```
hist(data2c$pelvic_tilt.numeric, col ="orange",xlab ="pelvic_radius")
```




```
hist(data2c$sqrt_pelvic_tilt.numeric, col ="orange",xlab ="pelvic_radius")
```



7)Single t-test

#a) Aim

Checking mean of sqrt transformed pelvic_tilt.numeric columns mean equal to 4

```
data2c=na_mean(data2c)
mean(data2c$sqrt_pelvic_tilt.numeric)
```

```
## [1] 4.097716
```

#b) Hypothesis

H0: μ = 4 H1: μ ≠ 4

α = 0.05

#c) Assumption Check

Is this a large sample? - Yes, because n > 30. Normality check - p-value = 0.09009(I did shapiro test at 6th step) The p-value of the test is 0.09009,

which is greater than alpha = 0.05. Thus, we can not reject the null hypothesis that our data is normally distributed.

#d) Indicate “which test you choose” “for what reason”

I choose one sample t test to compare the mean of one sample.

#e) Result

```
res <- t.test(data2c$sqrt_pelvic_tilt.numeric, mu = 4)
res
```

```
##
##  One Sample t-test
##
## data:  data2c$sqrt_pelvic_tilt.numeric
## t = 1.5331, df = 309, p-value = 0.1263
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.972298 4.223134
## sample estimates:
## mean of x
##  4.097716
```

Since p value is not less than our significance level of 0.05, we can not reject the null hypothesis that mean is 4.

#f) Conclusion

With 95% confidence, mean of our sample could be equal to 4.

#g) What can be Type-1 and Type-2 error here?

If H0 is 4 and we reject it -> Type 1 error. If H0 is not 4 and we do not reject it -> Type 2 error.

8)Paired t-test

I create a new dataset for this chapter. I can not do this test with my data. It is about jumping height before and after training.

```
datajump <- data.frame(jumping = c(65 ,65, 58, 58, 72, 74, 71, 65, 52, 77,
                                   64, 75, 79, 60, 70, 68, 75, 70, 76, 69,
                                   64, 68, 68, 70, 72, 73, 71, 65, 60, 73,
                                   77, 80, 73, 71, 70, 67, 74, 63, 72, 75),
                       group = c(rep('before', 20), rep('after', 20)))

diff <- with(datajump, jumping[group == "after"] - jumping[group == "before"])
head(datajump)
```

```
##   jumping  group
## 1      65 before
## 2      65 before
## 3      58 before
## 4      58 before
## 5      72 before
## 6      74 before
```

```
tail(datajump)
```

```
##    jumping group
## 35      70 after
## 36      67 after
## 37      74 after
## 38      63 after
## 39      72 after
## 40      75 after
```

#a) Aim

Checking mean of before and after training jumping height is equal.

#b) Hypothesis

H0:m=0 H1:m≠0

m = difference of means.

α = 0.05

#c) Assumption Check

Are the two samples paired? - Yes Assumption 2: Is this a large sample? - No, because n < 30. Normality

```
shapiro.test(diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff
## W = 0.92307, p-value = 0.1135
```

The p-value of the test is 0.1135, which is greater than alpha = 0.05. Thus, we can not reject the null hypothesis that our data is normally distributed.

#d) Result

```
t.test(jumping ~ group, data = datajump, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  jumping by group
## t = 1.588, df = 19, p-value = 0.1288
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6837307  4.9837307
## sample estimates:
## mean of the differences
##                    2.15
```

Since our p-value is not less than our significance level of 0.05 we can not reject the null hypothesis that the two groups have statistically significant means.

#e) Conclusion With 95% confidence, mean of jumping before training and after training could be same.

9)Fisher’s exact test for count data

#a) Aim Checking women and men variables are independent or not. I have 210 abnormal and 100 normal data in my dataframe. I randomly part them into women and men.

```
dat <- data.frame(
  "Abnormal" = c(100, 110),
  "Normal" = c(30, 70),
  row.names = c("Women", "Men"),
  stringsAsFactors = FALSE
)
colnames(dat) <- c("Abnormal", "Normal")

dat
```

```
##       Abnormal Normal
## Women      100     30
## Men        110     70
```

#b) Hypothesis and level of significance: H0: The two categorical variables are independent. H1: The two categorical variables are dependent.

α = 0.05

#c) Result

```
fisher.test(dat)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dat
## p-value = 0.0045
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.245127 3.654743
## sample estimates:
## odds ratio
##    2.11614
```

Since our p-value is less than our significance level of 0.05 we can reject the null hypothesis that the two groups are independent.

#d) Conclusion With 95% confidence, men and women are not independent on each other from being normal or abnormal.

#e) Odds Ratio We can understand women are more likely to be abnormal from odds ratio.

10)ANOVA and Tukey Test

#a) Aim Checking pelvic_incidence column means of Hernia, Normal and Spondylolisthesis classes. I will use second dataset that has 3 factors on class column.

#b) Hypo H0: μ1 = μ2 = μ3 H1: All means are not equal.

α = 0.01

#c) Assumption Check The observations are obtained independently and randomly from the population defined by the factor levels. - Yes. The data of each factor level are normally distributed. - Yes, shapiro test below shows p values. They are greater than confidence level 0.01.

```
data3c$sqrt_pelvic_incidence = sqrt(data3c$pelvic_incidence)
```

sqrt transformation for normality.

```
byf.shapiro(sqrt_pelvic_incidence~class,data=data3c)
```

```
##
##  Shapiro-Wilk normality tests
##
## data:  sqrt_pelvic_incidence by class
##
##                      W p-value
## Hernia           0.9893 0.87730
## Normal           0.9775 0.08486 .
## Spondylolisthesis 0.9784 0.01830 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These normal populations have a common variance. Yes, levene test below show p value is greater than confidence level.

```
leveneTest(sqrt_pelvic_incidence ~ class, data = data3c)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   0.657 0.5192
##       307
```

#d) Result of ANOVA

```
res.aov <- aov(sqrt_pelvic_incidence ~ class, data = data3c)
summary(res.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## class         2  149.2   74.58   101.5 <2e-16 ***
## Residuals   307  225.5    0.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value is less than the significance level 0.01, we can reject null hypothesis that means are equal.

#e) Conclusion of ANOVA With 99% confidence, pelvic_incidence column means of Hernia, Normal and Spondylolisthesis classes are not equal.

#f) Result of TUKEY

```
TukeyHSD(res.aov, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = sqrt_pelvic_incidence ~ class, data = data3c)
##
## $class
##                             diff        lwr       upr     p adj
## Normal-Hernia            0.2809539 -0.1298500 0.6917579 0.1121007
## Spondylolisthesis-Hernia 1.5498154  1.1655435 1.9340873 0.0000000
## Spondylolisthesis-Normal 1.2688615  0.9440925 1.5936305 0.0000000
```

As the p-value is less than the significance level 0.01, we can reject null hypothesis that means are equal. f) Conclusion of TUKEY With 99%

confidence, pelvic_incidence column mean of Spondylolisthesis is different than means of Hernia and Normal.

11) Multiple Linear Regression

#a) Aim

I want to build a model for degree_spondylolisthesis based on best columns.

#b) Regression Equation

pelvic_incidence = b0 + b1 * x + b2 * y

#c) Hypothesis and level of significance

H0: b1 = b2 = 0 H1: At least one of the coefficients ≠ 0

α = 0.05

#d) Find the Best Model

```
summary(lm1 <- lm(degree_spondylolisthesis  ~ pelvic_incidence+pelvic_tilt.numeric+lumbar_lordosis_angle+sacral_slope
                  +pelvic_radius, data = data2c))
```

```
##
## Call:
## lm(formula = degree_spondylolisthesis ~ pelvic_incidence + pelvic_tilt.numeric +
##     lumbar_lordosis_angle + sacral_slope + pelvic_radius, data = data2c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.587 -13.785  -1.591  10.470 303.311
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -109.9176    18.1273  -6.064 3.93e-09 ***
## pelvic_incidence         1.3301     0.1675   7.941 3.87e-14 ***
## pelvic_tilt.numeric     -0.1704     0.2151  -0.792   0.4289
## lumbar_lordosis_angle    0.2560     0.1266   2.022   0.0441 *
## sacral_slope                 NA         NA      NA       NA
## pelvic_radius            0.3854     0.1309   2.943   0.0035 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.4 on 305 degrees of freedom
## Multiple R-squared:  0.4358, Adjusted R-squared:  0.4284
## F-statistic: 58.89 on 4 and 305 DF,  p-value: < 2.2e-16
```

```
slm1 <- step(lm1)
```

```
## Start:  AIC=2079.66
## degree_spondylolisthesis ~ pelvic_incidence + pelvic_tilt.numeric +
##     lumbar_lordosis_angle + sacral_slope + pelvic_radius
##
##
## Step:  AIC=2079.66
## degree_spondylolisthesis ~ pelvic_incidence + pelvic_tilt.numeric +
##     lumbar_lordosis_angle + pelvic_radius
##
##                          Df Sum of Sq    RSS    AIC
## - pelvic_tilt.numeric     1       506 246459 2078.3
## <none>                                245953 2079.7
## - lumbar_lordosis_angle   1      3296 249248 2081.8
## - pelvic_radius           1      6984 252936 2086.3
## - pelvic_incidence        1     50857 296810 2135.9
##
## Step:  AIC=2078.3
## degree_spondylolisthesis ~ pelvic_incidence + lumbar_lordosis_angle +
##     pelvic_radius
##
##                          Df Sum of Sq    RSS    AIC
## <none>                                246459 2078.3
## - lumbar_lordosis_angle   1      3507 249965 2080.7
## - pelvic_radius           1      6479 252937 2084.3
## - pelvic_incidence        1     65173 311632 2149.0
```

```
summary(slm1)
```

```
##
## Call:
## lm(formula = degree_spondylolisthesis ~ pelvic_incidence + lumbar_lordosis_angle +
##     pelvic_radius, data = data2c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.062 -13.744  -1.327  10.466 309.685
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -105.7148    17.3230  -6.103 3.15e-09 ***
## pelvic_incidence        1.2571     0.1397   8.995  < 2e-16 ***
## lumbar_lordosis_angle   0.2633     0.1262   2.087  0.03775 *
## pelvic_radius           0.3586     0.1264   2.836  0.00487 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.38 on 306 degrees of freedom
## Multiple R-squared:  0.4346, Adjusted R-squared:  0.4291
## F-statistic:  78.4 on 3 and 306 DF,  p-value: < 2.2e-16
```

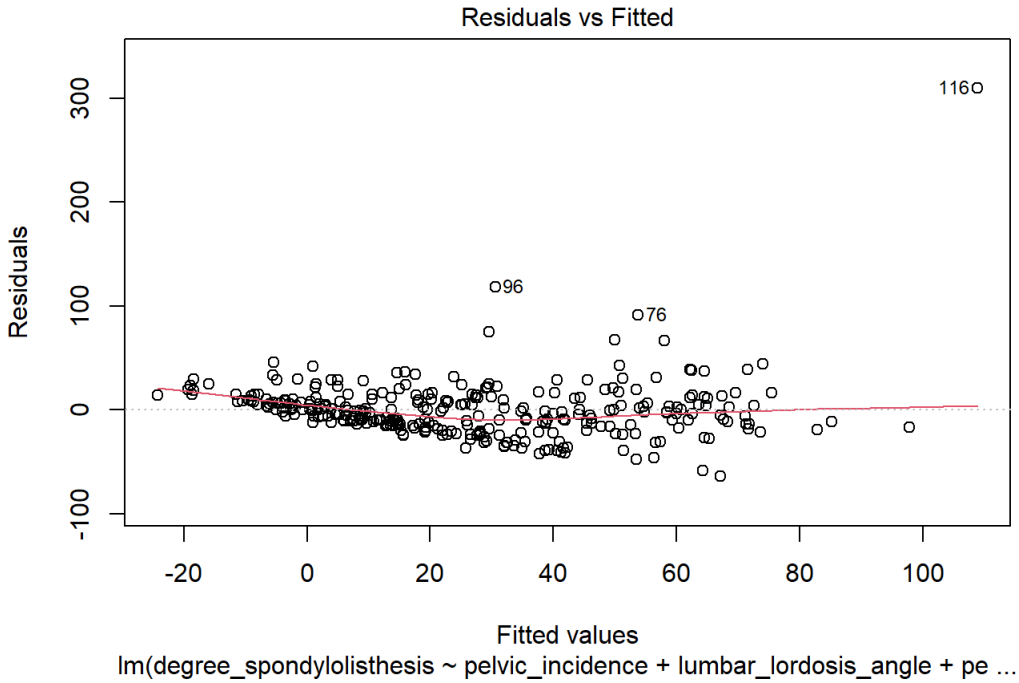
degree_spondylolisthesis = b0 + b1 * pelvic_incidence + b2 * lumbar_lordosis_angle + b3 * pelvic_radius is best model. P value shows that.

```
model <- lm(degree_spondylolisthesis ~ pelvic_incidence + lumbar_lordosis_angle+pelvic_radius, data = data2c)
```

#e) Assumption Check

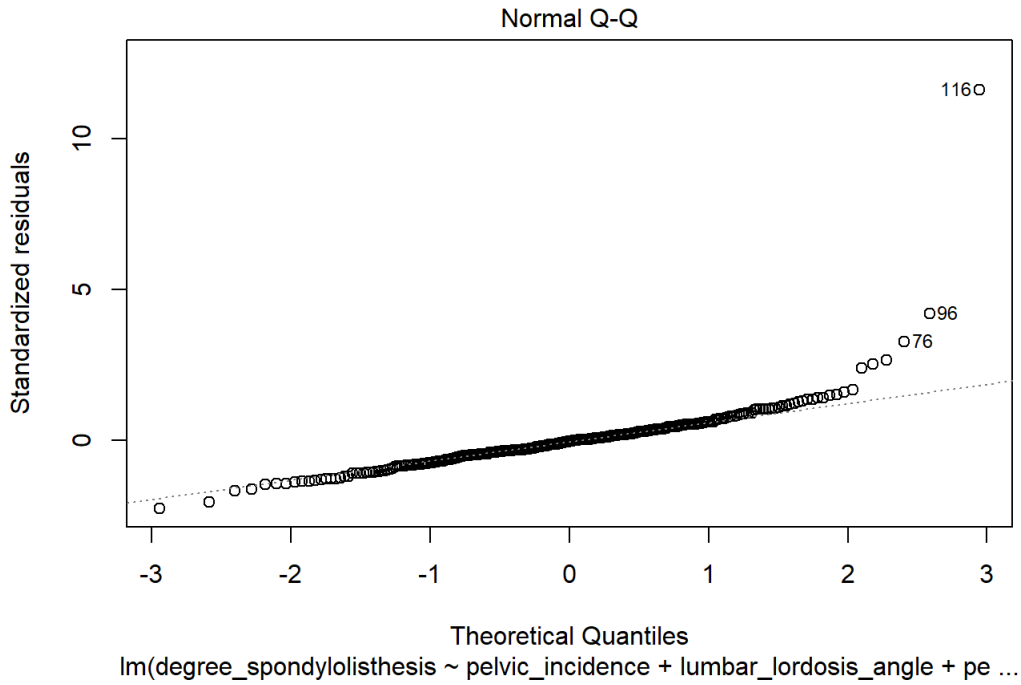
Linearity of the data. Plot below shows relationship between the predictors and the outcome is linear.

```
plot(model, 1)
```



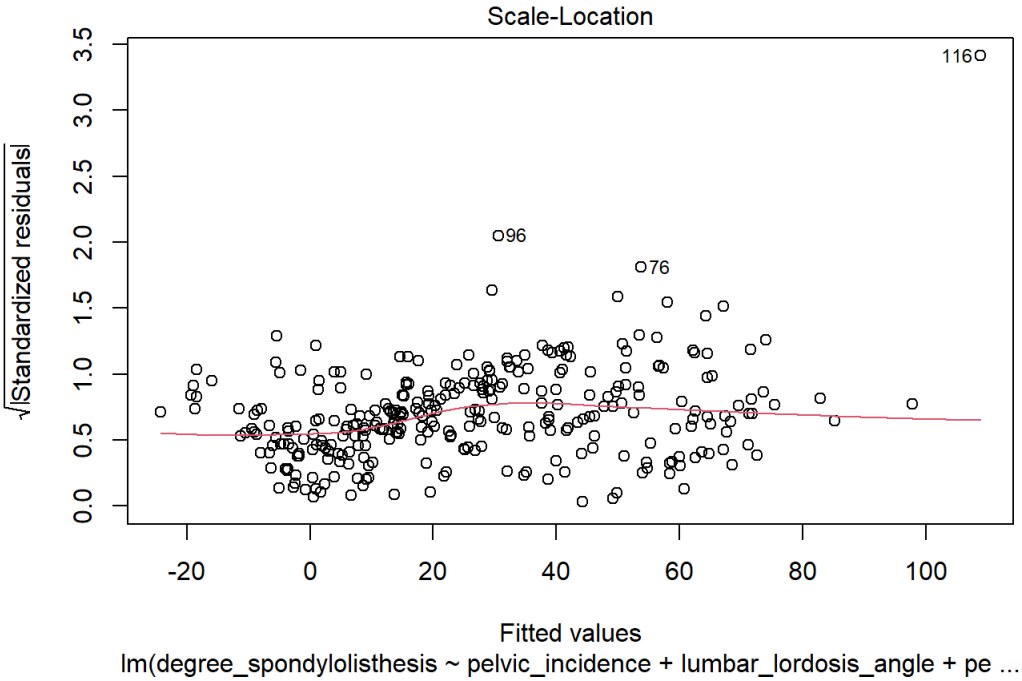
Normality of residuals. Plot below shows residual errors are normally distributed.

```
plot(model, 2)
```



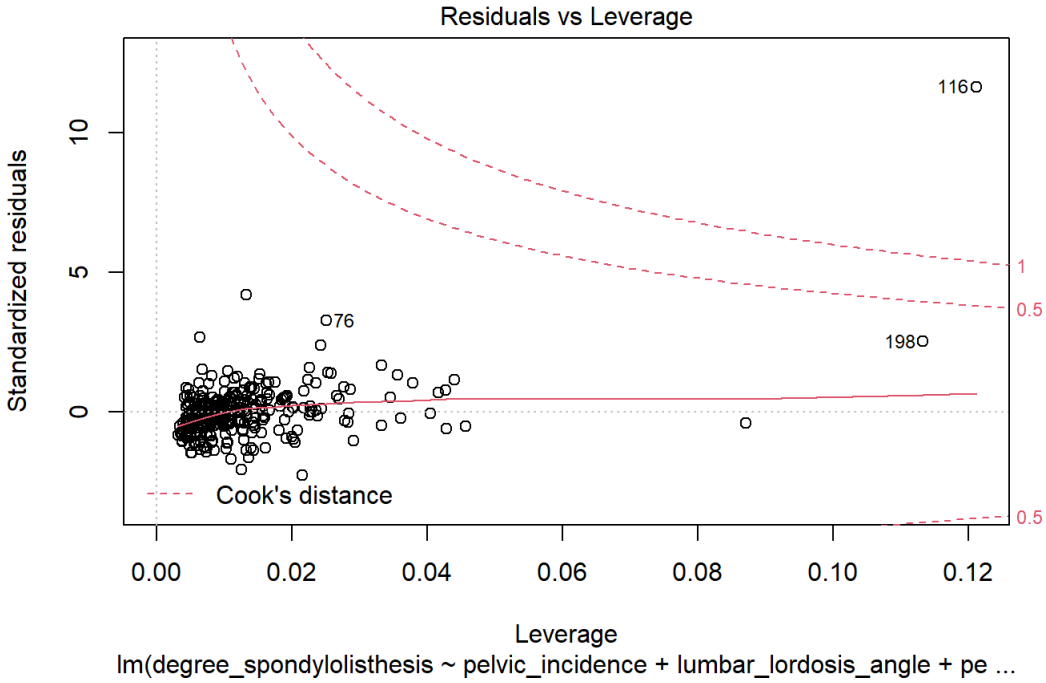
Homogeneity of variance. Plot below shows variance is homogen.

```
plot(model, 3)
```



Residuals vs Leverage. There are some outliers.

```
plot(model, 5)
```



#f) Result

```
summary(model)
```

```
##
## Call:
## lm(formula = degree_spondylolisthesis ~ pelvic_incidence + lumbar_lordosis_angle +
##     pelvic_radius, data = data2c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.062 -13.744  -1.327  10.466 309.685
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -105.7148    17.3230  -6.103 3.15e-09 ***
## pelvic_incidence       1.2571     0.1397   8.995  < 2e-16 ***
## lumbar_lordosis_angle  0.2633     0.1262   2.087  0.03775 *
## pelvic_radius          0.3586     0.1264   2.836  0.00487 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.38 on 306 degrees of freedom
## Multiple R-squared:  0.4346, Adjusted R-squared:  0.4291
## F-statistic:  78.4 on 3 and 306 DF,  p-value: < 2.2e-16
```

43% of the variance in the measure of degree_spondylolisthesis can be predicted by pelvic_incidence, lumbar_lordosis_angle and pelvic_radius. Our model equation can be written as follow: degree_spondylolisthesis = -105.7148 + 1.2571 * pelvic_incidence + 0.2633 * lumbar_lordosis_angle + 0.3586 * pelvic_radius.

#g) Conclusion:

With pelvic_incidence, lumbar_lordosis_angle and pelvic_radius columns, we can predict degree_spondylolisthesis with 43% accuracy.

#h) Prediction

```
predict(model,newdata = data.frame(pelvic_incidence =c(60),lumbar_lordosis_angle =c(40),pelvic_radius =c(100)))
```

```
##       1
## 16.10429
```