Home Exam On Data Analysis 1

Md Shamsuzzaman

10.16.2017

1 Conceptual

1.1 Answer of question 1

```
girlsBet7to14 <- subset(juul, age>=7 & age<=14)
girlsBet7to14
summary(girlsBet7to14)
```

1.2 Answer of question 2

Let?s assume two vectors are following below. Here a and b vectors contains same values, and on the other hand c and d vectors contain different values. Now, we can check the value either same or not. If they are same then the R analysis shows that as true while false for unequal vectors.

```
a <- c(100, 200, 300, NA, NA, 600, 700)
b <- c(100, 200, 300, NA, NA, 600, 700)
c <- c(100, 200, 300, NA, NA, 600, 700)
d <- c(150, 250, 350, NA, NA, 650, 750)
```

Analysis from R,

```
> all(is.na(a) == is.na(b)) && all((a == b)[!is.na(a)])
[1] TRUE

> all(is.na(c) == is.na(d)) && all((c == d)[!is.na(c)])
[1] FALSE
```

1.3 Answer of question 3

The histogram and true histogram

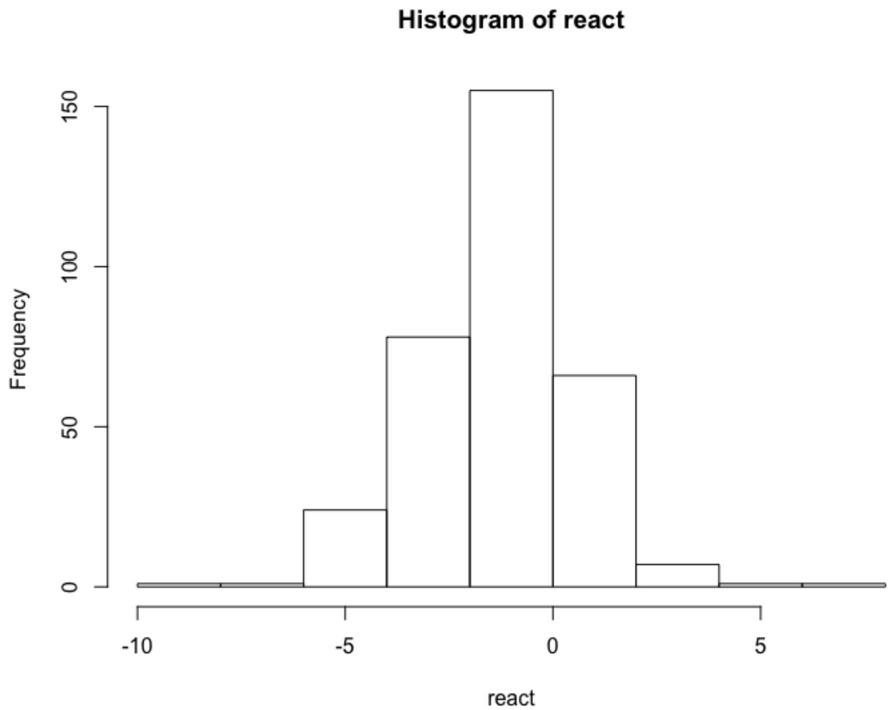

Fig 1. Histogram of react data

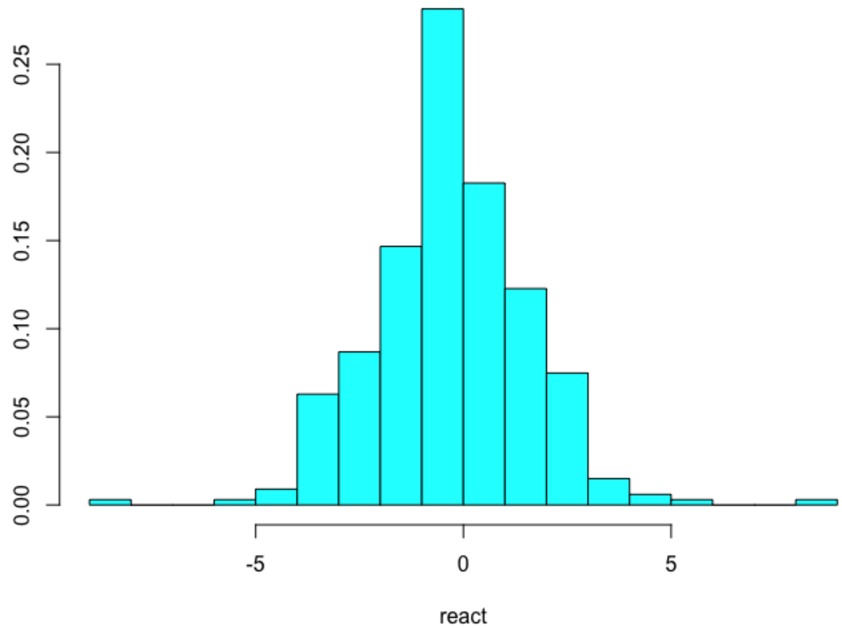

Fig 2. Truehist of react data

The react data are highly discretized so that this histogram is biased. From the react data, we can that the data contains integer values. Moreover, the data on the boundary are counted within the column to the left which shifts the histogram to left. A better histogram can be found by using truehist function from the MASS package which shows to specify a better set of breaks. The total area under the truehist is one.

2 Appropriate Analysis of the data

2.1 Answer of question 1

The group of frogs? data from nearby, 2 km, and 20 km away are independent (not related with each other). So, that is why I am going to use one-way analysis of variance (ANOVA) because through this test it can be determined whether there are any significant differences between the means of these three group of frogs. If the residuals of the data are not normally districted without constant variance then Kurskal-Wallis test is applicable instead of ANOVA. On the other hand, TukeyHSD, single-step multiple comparison procedure, can be used in conjunction with an ANOVA (post-hoc analysis) to find means that are significantly different from each other. The ANOVA statistic prevents up from having to do multiple t-tests puts all the data into one number.

```
Assumptions:
a) The populations of the frogs from where 20 frogs obtained are normally or approximately
b) The frogs samples are independent
c) The variances of the frogs populations are equal

Hypothesis
Null hypothesis: Population means of the frogs are equal
Alternate hypothesis: At least one mean is different
```

2.2 Answer of question 2

Paired t test can be used to analysis the data of the two measurements, with normal variance different, on this photocopy adverting experimental field. A paired t-test is being applied to compare the population means for, two samples of photocopies after and before, where one sample can be paired with observations in the other sample.

```
Paired T-Test Assumptions
1. The data are continuous (not discrete)
2. The data follow a normal probability distribution
3. The sample of pairs is a simple random sample from its population

Hypothesis
Null hypothesis: The mean difference of paired observations is equal
Alternate hypothesis: The mean difference of paired observations is not equal
```

If the mean differences are not normal then Wilcoxon signed rank test is applicable.

2.3 Answer of question 3

We can use two-way ANOVA for the analysis of this data set having normal distribution for residuals and constant variance. The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

```
Assumptions
a) The samples are independent
b) The variances of the populations must be equal
c) The groups have the same sample size
```

```
Hypothesis
Null hypothesis: The mean difference of all groups is same
Alternate hypothesis: The mean difference of all groups is not same
```

If the residuals are not normally distributed and the variances are not constant then we can use Friedman two-way ANOVA.

2.4 Answer of question 4

Here we can use the Two-sample t test to determine whether the means of two independent groups differ and to calculate a range of values that is likely to include the difference between the population means. 2-Sample t calculates a confidence interval and does a hypothesis test of the difference between two population means when standard deviations are unknown and samples are drawn independently from each other.

```
Assumptions
a) The two samples are independent
b) The two samples follow normal distributions, and can be done with normality check

Hypothesis
Null hypothesis: The population means are same
Alternate hypothesis: The population means are not same

When the assumptions are not met, other methods are possible based on the two samples:
a) Two dependent samples and follow Normal distribution, suggest Paired T-test
b) Two independent samples and does not follow Normal distribution, suggest WMW test
c) Two dependent samples and does not follow Normal distribution, suggest Signed Rank test
```

3 Executive Summary

Vitcap data set is consisted of 24 rows and 3 columns. It contains data on vital capacity for workers in the cadmium industry. It is a subset of the vitcap2 data set. This data set has been constructed on group, age, and vital capacity. This study is revealing the analysis, on two groups of workers who are exposed to cadmium for more than 10 years as group one and others in group 3 who are not exposed, two sample t-test to determine a comparison. The testing hypothesis of the analysis, the mean of the groups are same as null hypothesis while the alternate hypothesis reveals that the means are not same. In the analysis of the data set to determine the normality, it has been found that some of the tests indicate the data as normally distributed (box plot, QQ plot) while the others (histogram, scattered plot) do not support the normality distribution assumption. So, finally we can check the normality through Shapiro test. Finally in Shapiro test provides the p values (0.257 & 0.4269) which are more than 0.05 indicates the data are normally distributed. At 90 % confidence interval, the comparison of variance test provide the p value 0.1806 which indicates that the assumption of equal variance has been met. On the other hand, at 90 % confidence interval, it has been found (p value 0.007882) that the mean value of the groups is different. According to the test base on the ranked, the analysis got p-value 0.01783 (less than 0.05) indicating the not equal median of the groups.

4 Introduction

This data set has been constructed on group, age, and vital capacity. This study is on two groups of workers who are exposed to cadmium for more than 10 years as group one and others in group 3 who are not exposed. The normality test, outliers and other presentations have been formulated by boxplot, histogram, scattered plot, and QQ plot. Outliers have been removed and the rank has been tested by wilcoxon sign rank test.

5 Data collection and summary

This data set has been collected from the study on the workers who work in a Cadmium industry. Some of them are expose to Cadmium and others are not exposed.

6 Analysis

At the beginning, we can analyze and understand the variables through the graphical analysis. Here we can check the data through boxplot, scattered plot, histogram, and QQ plot.

6.1 Answer of question No. a)

The representative graphical presentations are following below.

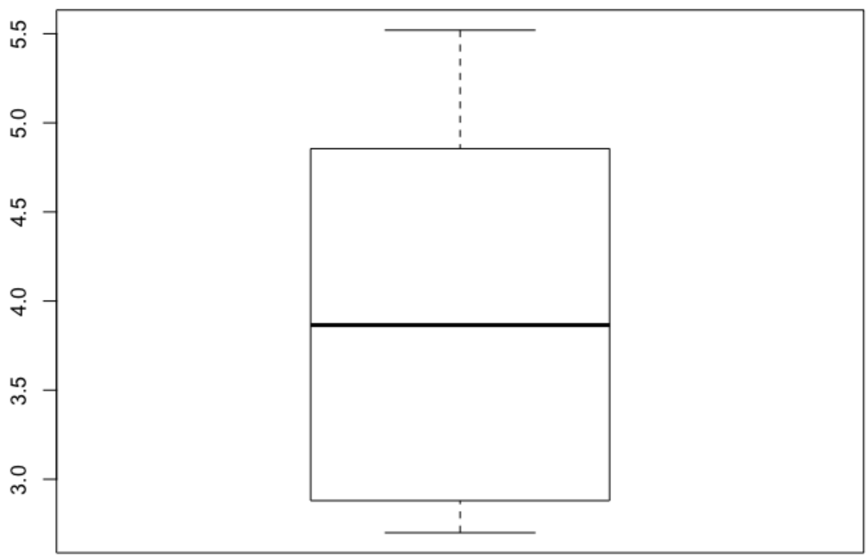

Fig 1. Boxplot of group 1 data

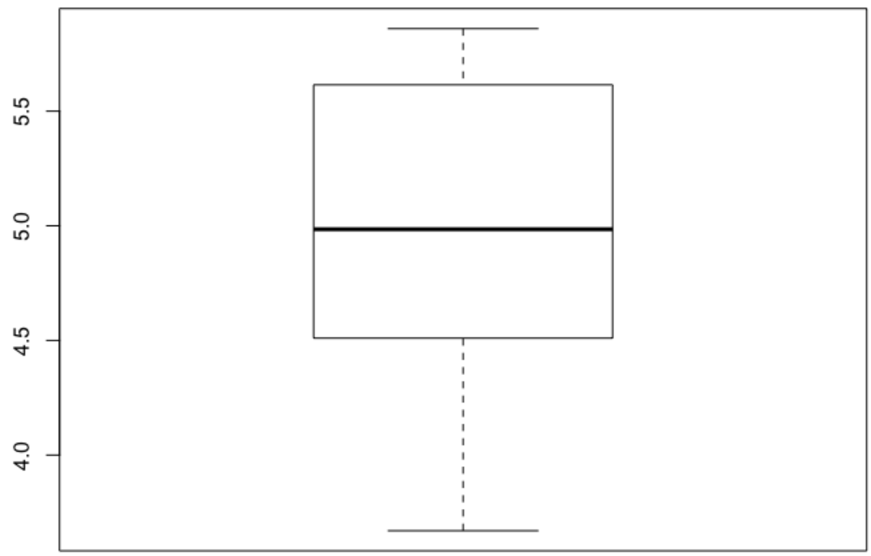

Fig 2. Boxplot of group 3 data

From the boxplot (figure 1 & 2), it looks that the both groups have normal distribution. However, we can double check their normality through histogram.

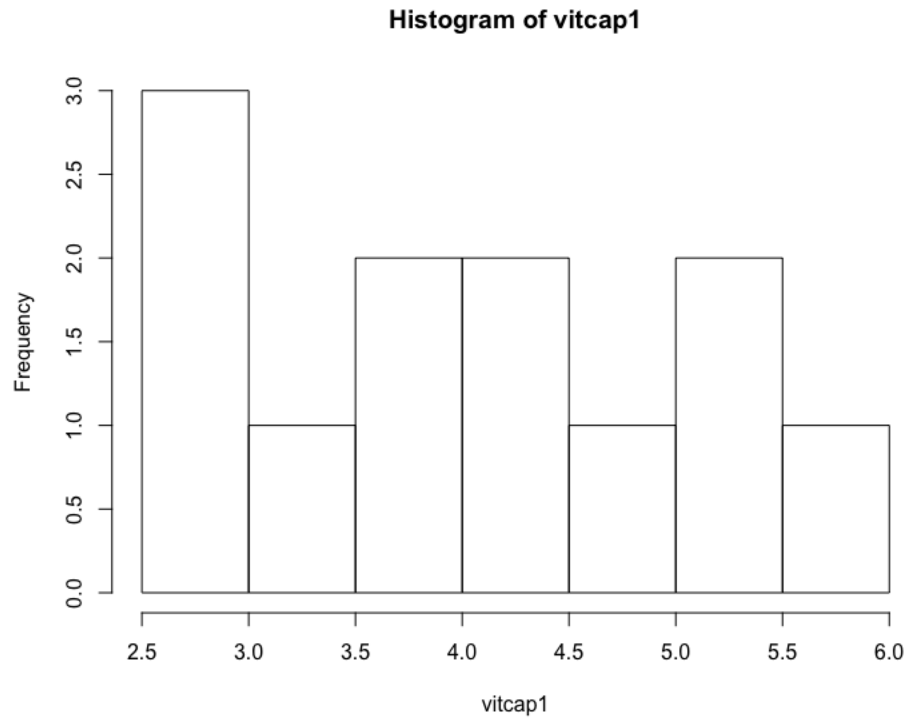

Fig 3. Histogram of group 1

From the histogram (figure 3), the data set looks like skewed to right which does not support normality assumption.

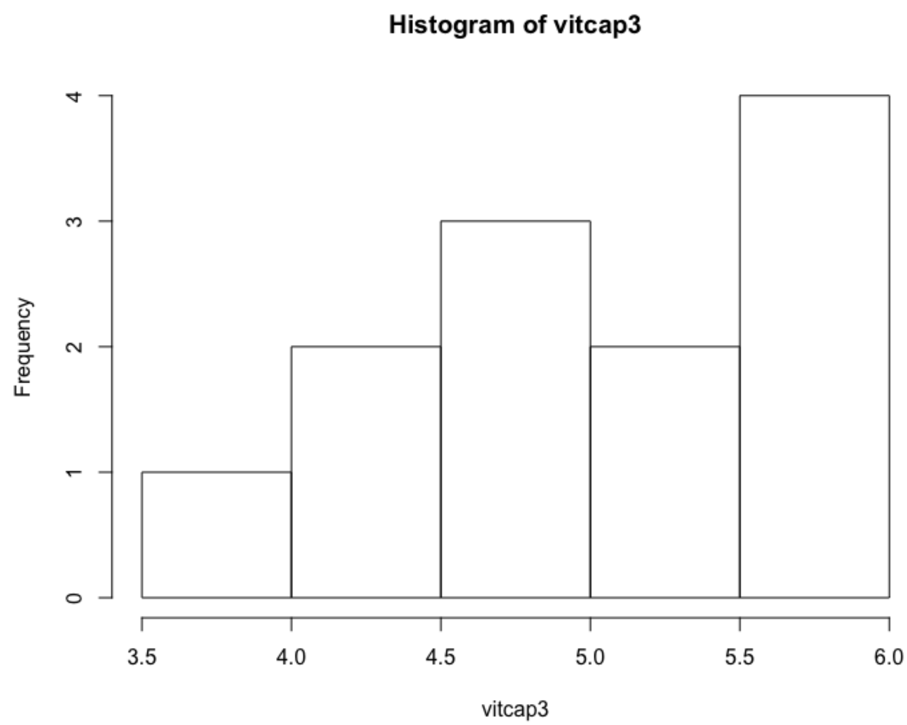

Fig 4. Histogram of group 3

Histogram from the figure 4, the data are skewed to left which also does not support the normality assumption.

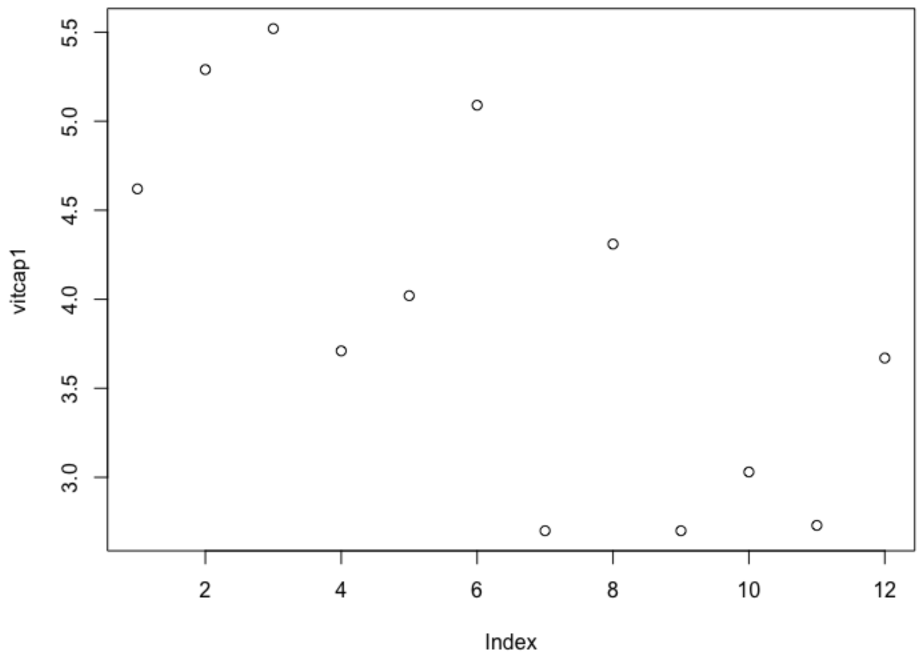

Fig 5. Scattered plot of group 1

Scattered plot (figure 5) of this data set shows that the data distribution is not normally distributed.

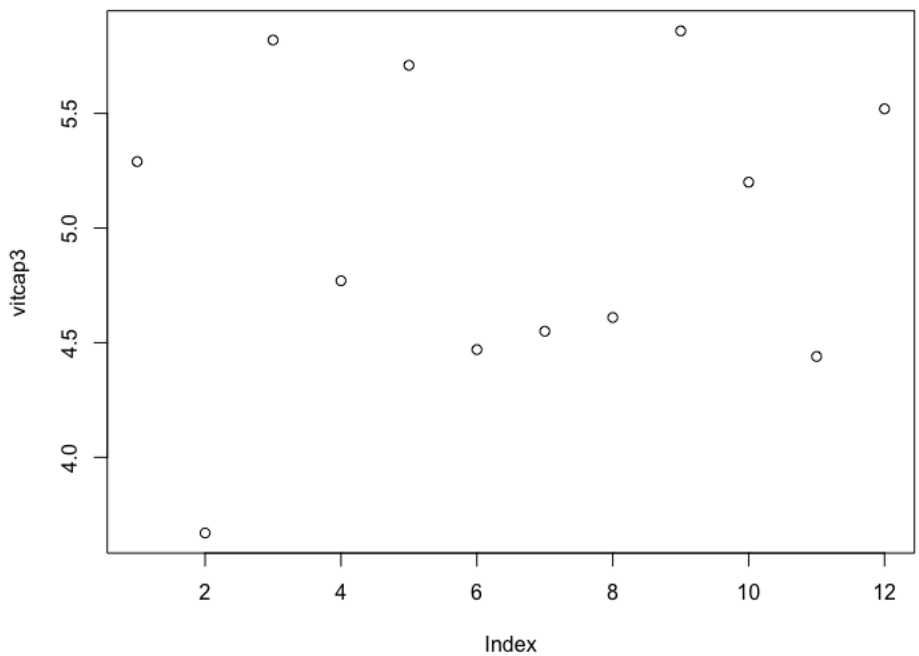

Fig 6. Scattered plot of group 2

Scattered plot (figure 6) of this data set shows that the data distribution is not normally distributed.

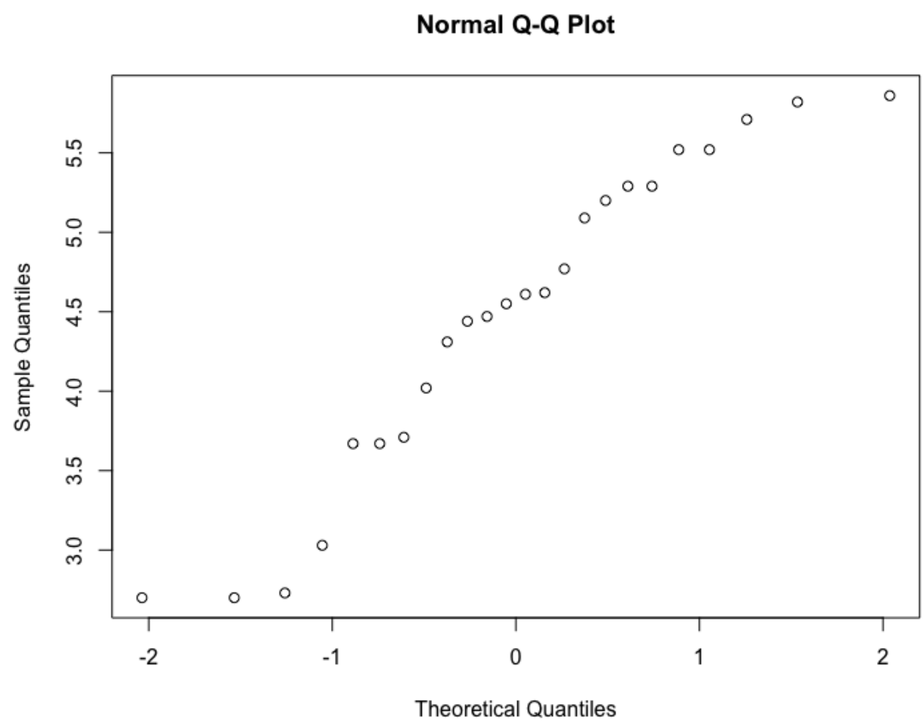

Fig 7. QQ plot

The QQ plot (figure 7) of the data set indicates that the data follow the normal distribution character as they are following the linear trend line. The analysis of the data, we can see that some of the analyses (box plot, QQ plot) indicate the normal distribution while the others (histogram, scattered plot) do not support the normality distribution assumption. So, finally we can check the normality through Shapiro test. In Shapiro test, we can have the hypothesis that indicates the normality.

```
Ho: Data are normally distributed
Ha: Data are not normally distributed

> shapiro.test(vitcap1)

Shapiro-Wilk normality test

data:  vitcap1
W = 0.91633, p-value = 0.257

> shapiro.test(vitcap3)
```

```
Shapiro-Wilk normality test

data:  vitcap3
W = 0.93421, p-value = 0.4269
```

From the p values of the shapiro tests, we have the p values (0.257 & 0.4269) more than 0.05. So, in this case, we fail to reject the null hypothesis (Ho) which means the data are normally distributed.

6.2 a)

Two-sample t-test

```
> Two Sample t-test

data:  vital.capacity by group
t = -2.9228, df = 22, p-value = 0.007882
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -2.04953499 -0.03713167
sample estimates:
mean in group 1 mean in group 3
       3.949167        4.992500
```

According to this t-test, the p-value is less than 0.01 (at 99 percent confidence interval), which means that we reject Ho. So, we can say the difference between the mean of group 1 and the mean of group 3 is not equal to 0. Two groups are different and that is between -2.04953499 and -0.03713167 at 90 percent interval.

6.3 b)

F test to compare two varriances

```
> data:  vital.capacity by group
F = 2.3105, num df = 11, denom df = 11, p-value = 0.1806
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
  0.4343334 12.2911384
sample estimates:
ratio of variances
          2.310509
```

```
Null Hypothesis (Ho): m1-m2  is equal zero
Alternate hypothesis (Ha): m1-m3 is not equal zero
```

From the var.test, we can see that the p value (0.1860) is larger than 0.01, and it reveals that the null hypothesis is accepted. So, we can say, the variances are same at 99% confidence interval.

6.4 c)

According to the tests, this result might not give us the actual result because the assumption of normality is not met according to the histogram. So, at this point, we run a non-parametric test to see whether these two groups different on the basis of their ranks.

6.5 d)

We can run Wilcoxon test as a nonparametric test. Here, we have the hypothesis following below.

```
Null Hypothesis (Ho): m1-m2 is equal zero
Alternate hypothesis (Ha): m1-m3 is not equal zero

Verbatim
Wilcoxon rank sum test with continuity correction

data:  vital.capacity by group
W = 30.5, p-value = 0.01783
alternative hypothesis: true location shift is not equal zero.
```

According to the p value (0.01783), less than 0.05, from the Wilcoxon test, we can reject Ho. It means two groups are different.

6.6 e)

The assumptions have been performed through the graphical presentation (boxplot, histogram, scattered plot, QQ plot) in the figures at the beginning of the analysis section.

6.7 f)

```
shapiro.test(sorted_vitcap[-c (1,24)])

Shapiro-Wilk normality test

data:  sorted_vitcap[-c(1, 24)]
W = 0.94249, p-value = 0.2228
```

After removing the outlier (1st and 24th), the Shpiro test shows that the p value (0.2228) is more than 0.05. We can see there is the normality by sorting the data to remove the outliers.

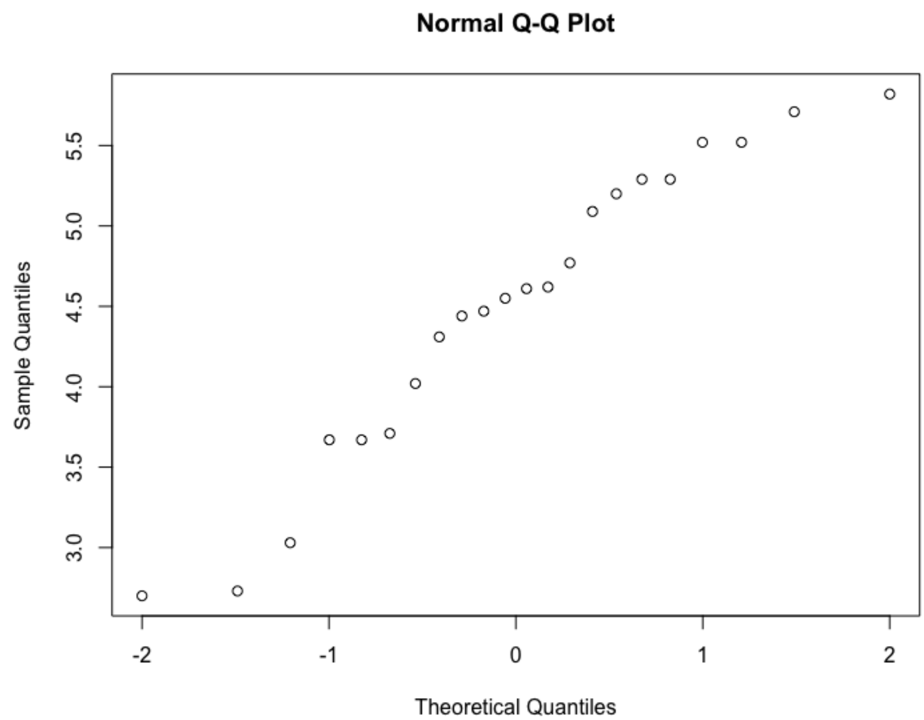

Fig 8. QQ plot of sorted data

7 Conclusion

Shapiro test provides the p values (0.257 & 0.4269) which are more than 0.05 indicates the data are normally distributed. At 90 % confidence interval, the comparison of variance test provide the p value 0.1806 which indicates that the assumption of equal variance has been met. On the other hand, at 90 % confidence interval, it has been found (p value 0.007882) that the mean value of the groups is different. According to the test base on the ranked, the analysis got p-value 0.01783 (less than 0.05) indicating the not equal median of the groups.

8 Executive Summary

These data have been collected by The National Institute of Diabetes and Digestive and Kidney Diseases from 768 female. This is the primary dataset from female adult patients who were living near Phoenix listed in 768 rows and 9 columns. This dada set has been used to analyse the relationship between diabetes and covariate. From the analysis, no relationship has not been found from the normality test. Several graphic visualising graphs have been used to see the relationship. Linear regression and fitted linear regression models revealed that the resulted p values are smaller than the predetermined value 0.05, and did not support to accept the null hypothesis. On the other hand the adjusted R squared value indicated the rejection of this model to use as a good model. Due to the higher p value for the relationship between log transformation of diabetes and covariate insulin, the null hypothesis is accepted. According to this model, the insulin information is not important and that is why we cannot use this model for determining diabetes level. Pearson correlation also does not support this model. So, in this case multiple regression model can be effective model for the analysis. It might be the case that it has not provided good model due to analysis only on the patients who are 33 years or more older.

9 Introduction

These data are to subject of the R analysis to find out more fitted linear regression model to carry out the relationship between diabetes and insulin. This analysis will be carried out on the patients who are 33 years old and more.

10 Data collection and summary

These data have been collected by The National Institute of Diabetes and Digestive and Kidney Diseases from 768 female. This is the primary dataset without any modification and fabrication from female adult patients who were living near Phoenix.

11 Analysis

11.1 a)

Setting all zero values of the five variable to NA

```
is.na(pima) <- !pima
> pima
pregnant glucose diastolic triceps insulin bmi diabetes age test
1 6 148 72 35 155.5482 33.6 0.627 50 1
2 1 85 66 29 155.5482 26.6 0.351 31 NA
```

```
3 8 183 64 NA 155.5482 23.3 0.672 32 1
4 1 89 66 23 94.0000 28.1 0.167 21 NA
5 NA 137 40 35 168.0000 43.1 2.288 33 1
6 5 116 74 NA 155.5482 25.6 0.201 30 NA
7 3 78 50 32 88.0000 31.0 0.248 26 1
8 10 115 NA NA 155.5482 35.3 0.134 29 NA
9 2 197 70 45 543.0000 30.5 0.158 53 1
10 8 125 96 NA 155.5482 NA 0.232 54 1
11 4 110 92 NA 155.5482 37.6 0.191 30 NA
12 10 168 74 NA 155.5482 38.0 0.537 34 1
13 10 139 80 NA 155.5482 27.1 1.441 57 NA
14 1 189 60 23 846.0000 30.1 0.398 59 1
15 5 166 72 19 175.0000 25.8 0.587 51 1
```

11.2 b)

Usage of mean value to replace the missing data

```
> pima$diastolic[which(is.na(pima$diastolic))]<-mean(pima$diastolic, na.rm = TRUE)
> pima$pregnant[which(is.na(pima$pregnant))]<-mean(pima$pregnant, na.rm = TRUE)
> pima$triceps[which(is.na(pima$triceps))]<-mean(pima$triceps, na.rm = TRUE)
> pima$glucose[which(is.na(pima$glucose))]<-mean(pima$glucose, na.rm = TRUE)
> pima
pregnant glucose diastolic triceps insulin bmi diabetes age test
1 6.000000 148.0000 72.00000 35.00000 155.5482 33.6 0.627 50 1
2 1.000000 85.0000 66.00000 29.00000 155.5482 26.6 0.351 31 NA
3 8.000000 183.0000 64.00000 29.15342 155.5482 23.3 0.672 32 1
4 1.000000 89.0000 66.00000 23.00000 94.0000 28.1 0.167 21 NA
5 4.494673 137.0000 40.00000 35.00000 168.0000 43.1 2.288 33 1
6 5.000000 116.0000 74.00000 29.15342 155.5482 25.6 0.201 30 NA
7 3.000000 78.0000 50.00000 32.00000 88.0000 31.0 0.248 26 1
8 10.000000 115.0000 72.40518 29.15342 155.5482 35.3 0.134 29 NA
9 2.000000 197.0000 70.00000 45.00000 543.0000 30.5 0.158 53 1
10 8.000000 125.0000 96.00000 29.15342 155.5482 NA 0.232 54 1
11 4.000000 110.0000 92.00000 29.15342 155.5482 37.6 0.191 30 NA
12 10.000000 168.0000 74.00000 29.15342 155.5482 38.0 0.537 34 1
13 10.000000 139.0000 80.00000 29.15342 155.5482 27.1 1.441 57 NA
14 1.000000 189.0000 60.00000 23.00000 846.0000 30.1 0.398 59 1
15 5.000000 166.0000 72.00000 19.00000 175.0000 25.8 0.587 51 1
```

Now we can check the normality through lille.test

```
 > lillie.test(log(diabetes))
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
data: log(diabetes)
D = 0.049932, p-value = 0.0001074
```

The p value (0.0001074) is less than 0.05 and it indicates that the data is not normally distributed. Before fitting a linear regression model for the log transformed diabetes versus covariate insulin for patients at age 33 or older, we ca perform the the data in graphs

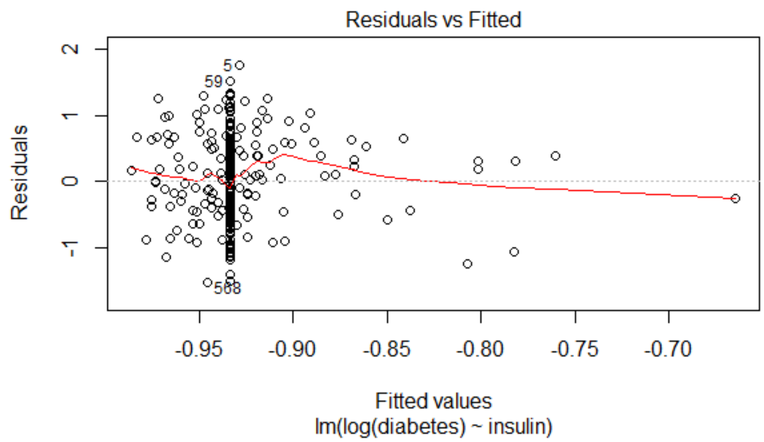

Fig 1. Residual plotting against fitted value

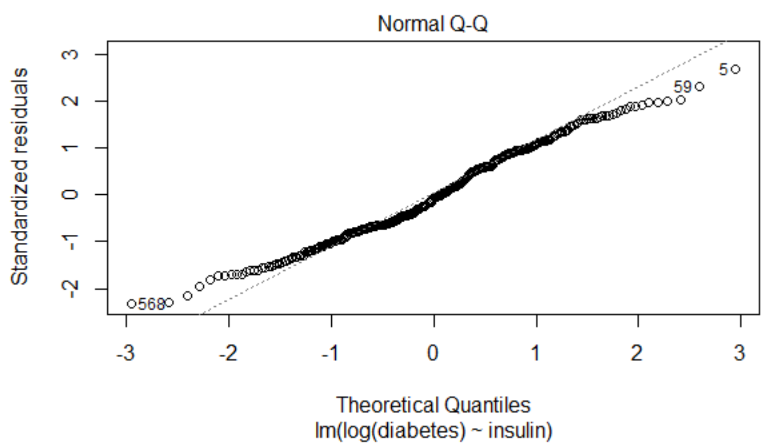

Fig 2. Residual plotting against leverage

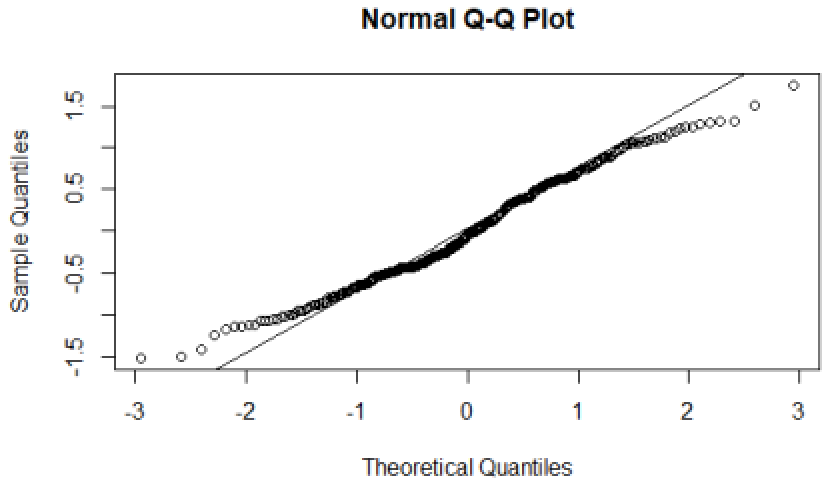

Fig 3. QQ plot

```
Call:
lm(formula = log(diabetes) ~ insulin, data = pima, subset = (subset = age >=
33))
Coefficients:
(Intercept) insulin
-0.9944202 0.0003897
```

So, now we can construct the fitted linear regression model for the log-transformed data for 33 years and more.

```
log(diabetes) = -0.9944202 + 0.0003897 insulin
```

```
> model.1=lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
> summary(model.1)
Call:
lm(formula = log(diabetes) ~ insulin, data = pima, subset = (subset = age >=
33))
Residuals:
Min 1Q Median 3Q Max
-1.51979 -0.47474 -0.03906 0.52784 1.75663
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9944202 0.0848103 -11.725 <2e-16 ***
insulin 0.0003897 0.0004590 0.849 0.396
---
```

```
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 0.6549 on 309 degrees of freedom
Multiple R-squared: 0.002328,Adjusted R-squared: -0.000901
F-statistic: 0.7209 on 1 and 309 DF, p-value: 0.3965
```

Correlation

```
> cor.test(log(diabetes), insulin)
Pearson?s product-moment correlation
data: log(diabetes) and insulin
t = 5.1633, df = 766, p-value = 3.095e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.1141309 0.2508837
sample estimates:
cor
0.1833944
```

From the Pearson correlation and other tests, we fail to reject the null hypothesis. So, we make the conclusion that age data is not important in this test.

12 Conclusion

Due to the higher p value for the relationship between log transformation of diabetes and covariate insulin, the null hypothesis is accepted. According to this model, the insulin information is not important and that is why we cannot use this model for determining diabetes level. Pearson correlation also does not support this model. So, in this case multiple regression model can be effective model for the analysis.

13 Appendix

13.1 R Code

```
attach(juul)
names(juul)
str(juul)
girlsBet7to14 <- subset(juul, age>=7 & age<=14)
girlsBet7to14
summary(girlsBet7to14)

## Applied
attach(vitcap)
```

```
print(vitcap)
vitcap1 <-vital.capacity[group=="1"]
vitcap3 <-vital.capacity[group=="3"]
plot(vitcap1)
plot(vitcap3)
boxplot(vitcap1)
boxplot(vitcap3)
hist(vitcap1)
hist(vitcap3)
qqnorm(vitcap$vital.capacity)
shapiro.test(vitcap1)
shapiro.test(vitcap3)
var.test(vital.capacity~group,conf.level=0.99)
t.test(vital.capacity~group,var.equal=T,conf.level=0.99)
wilcox.test(vital.capacity~group)
sorted_vitcap<-sort(vitcap$vital.capacity,decreasing=FALSE,na.last=NA)
shapiro.test(sorted_vitcap[-c(1,24)])
print(sorted_vitcap[-c(1,24)])
qqnorm(sorted_vitcap[-c(1,24)])
par(mfrow=c(1,2))
boxplot(vitcap1)
boxplot(vitcap3)
detach(vitcap)

Faraway
library(faraway)
attach(pima)
library(faraway)
attach(pima)
is.na(pima)<-!pima
pima
pima$insulin[which(is.na(pima$insulin))]<-mean(pima$insulin, na.rm = TRUE)
pima
pima$diastolic[which(is.na(pima$diastolic))]<-mean(pima$diastolic, na.rm = TRUE)
pima$pregnant[which(is.na(pima$pregnant))]<-mean(pima$pregnant, na.rm = TRUE)
pima$triceps[which(is.na(pima$triceps))]<-mean(pima$triceps, na.rm = TRUE)
pima$glucose[which(is.na(pima$glucose))]<-mean(pima$glucose, na.rm = TRUE)
pima$diabetes[which(is.na(pima$diabetes))]<-mean(pima$diabetes, na.rm = TRUE)
pima
lmpima<-lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
plot(lmpima)
```

```
abline(lmpima)
lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
model.1=lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
summary(model.1)
confint(model.1, level=.98)
pima_fit<-lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
plot(fitted(pima_fit),resid(pima_fit))
qqnorm(resid(pima_fit))
qqline(resid(pima_fit))
pima_fit
lillie.test(log(diabetes))
cor.test(log(diabetes), insulin)
```

13.2 Log File

```
> install.packages("ISwR", dependencies = FALSE)
> library("ISwR", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
> attach(juul)
> names(juul)
[1] "age"      "menarche" "sex"      "igf1"     "tanner"   "testvol"
> str(juul)
'data.frame': 1339 obs. of  6 variables:
 $ age     : num  NA NA NA NA NA 0.17 0.17 0.17 0.17 0.17 ...
 $ menarche: int  NA NA NA NA NA NA NA NA NA NA ...
 $ sex     : num  2 2 2 2 2 2 2 2 2 2 ...
 $ igf1    : num  90 88 164 166 131 101 97 106 111 79 ...
 $ tanner  : int  NA NA NA NA NA 1 1 1 1 1 ...
 $ testvol : int  NA NA NA NA NA NA NA NA NA NA ...
> girlsBet7to14 <- subset(juul, age>=7 & age<=14)
> girlsBet7to14
      age menarche sex igf1 tanner testvol
99   7.01       NA   2   NA      1       1
100  7.04       NA   2  149      1       1
101  7.07       NA   2   NA      1       1
102  7.07       NA   2  187      1       1
103  7.08       NA   2   NA      1       1
104  7.22       NA   2  103      1       1
105  7.24       NA   2   NA      1       1
106  7.24       NA   2  145      1       1
```

107	7.25	NA	2	NA	1	1
108	7.25	NA	2	117	1	1
109	7.26	NA	2	88	1	1
110	7.29	NA	2	NA	1	1
111	7.29	NA	2	186	1	1
112	7.30	NA	2	235	1	1
113	7.36	NA	2	NA	1	1
114	7.47	NA	2	NA	1	1
115	7.48	NA	2	300	1	1
116	7.49	NA	2	188	1	1
117	7.50	NA	2	NA	1	1
118	7.50	NA	2	110	1	1
119	7.50	NA	2	198	1	1
120	7.54	NA	2	134	1	1
121	7.54	NA	2	46	1	1
122	7.64	NA	2	NA	1	1
123	7.79	NA	2	NA	1	1
124	7.81	NA	2	NA	1	1
125	7.82	NA	2	NA	1	1
126	7.88	NA	2	221	1	1
127	7.90	NA	2	225	1	1
128	8.01	NA	2	NA	1	1
129	8.04	NA	2	NA	1	1
130	8.09	NA	2	166	1	1
131	8.10	NA	2	324	1	1
132	8.11	NA	2	NA	1	1
133	8.14	NA	2	146	1	1
134	8.19	NA	2	485	1	1
135	8.20	NA	2	152	1	1
136	8.25	NA	2	278	1	1
137	8.27	NA	2	315	1	2
138	8.30	NA	2	206	1	1
139	8.31	NA	2	624	1	1
140	8.33	NA	2	318	1	1
141	8.33	NA	2	187	1	1
142	8.37	NA	2	141	1	1
143	8.39	NA	2	NA	1	1
144	8.44	NA	2	152	1	1
145	8.44	NA	2	219	1	1
146	8.54	NA	2	169	1	1
147	8.55	NA	2	NA	1	3

148	8.62	NA	2	115	1	1
149	8.64	NA	2	223	1	1
150	8.64	NA	2	295	1	1
151	8.65	NA	2	NA	1	1
152	8.65	NA	2	117	1	1
153	8.68	NA	2	416	1	1
154	8.69	NA	2	NA	1	1
155	8.69	NA	2	149	1	2
156	8.72	NA	2	NA	1	1
157	8.80	NA	2	160	1	1
158	8.80	NA	2	99	1	1
159	8.83	NA	2	NA	1	1
160	8.83	NA	2	490	1	1
161	8.85	NA	2	NA	1	1
162	8.86	NA	2	NA	1	1
163	8.88	NA	2	NA	1	1
164	8.89	NA	2	101	1	1
165	8.90	NA	2	238	1	1
166	8.91	NA	2	283	1	1
167	8.96	1	2	NA	1	NA
168	8.96	NA	2	NA	1	1
169	8.96	NA	2	279	1	1
170	8.97	NA	2	NA	1	1
171	9.00	NA	2	NA	1	2
172	9.01	NA	2	171	1	1
173	9.05	NA	2	NA	1	1
174	9.07	NA	2	NA	1	2
175	9.09	NA	2	224	1	2
176	9.13	NA	2	174	1	1
177	9.14	NA	2	179	1	1
178	9.23	NA	2	104	1	1
179	9.25	NA	2	NA	1	1
180	9.32	NA	2	NA	1	1
181	9.33	NA	2	279	1	1
182	9.34	NA	2	NA	1	1
183	9.38	NA	2	NA	1	1
184	9.41	NA	2	222	1	1
185	9.42	NA	2	156	1	1
186	9.43	NA	2	288	1	1
187	9.45	NA	2	269	1	2
188	9.46	NA	2	262	1	2

189	9.48	NA	2	NA	1	2
190	9.49	NA	2	NA	1	1
191	9.50	NA	2	NA	1	2
192	9.50	NA	2	NA	2	2
193	9.55	NA	2	264	1	1
194	9.56	NA	2	240	1	2
195	9.56	NA	2	126	1	1
196	9.56	NA	2	158	1	1
197	9.59	NA	2	258	1	1
198	9.59	NA	2	146	1	1
199	9.60	NA	2	NA	1	1
200	9.64	NA	2	203	1	1
201	9.68	NA	2	288	1	2
202	9.71	NA	2	NA	1	1
203	9.71	NA	2	NA	1	1
204	9.74	NA	2	151	1	2
205	9.74	NA	2	161	1	2
206	9.75	NA	2	179	1	1
207	9.76	NA	2	209	1	1
208	9.79	NA	2	NA	1	1
209	9.80	NA	2	292	1	1
210	9.82	NA	2	NA	1	2
211	9.83	NA	2	284	1	1
212	9.83	NA	2	295	1	1
213	9.89	NA	2	NA	1	1
214	9.92	NA	2	138	1	2
215	10.03	NA	2	NA	1	1
216	10.03	NA	2	224	2	2
217	10.04	NA	2	204	1	2
218	10.17	NA	2	245	1	1
219	10.18	NA	2	267	1	1
220	10.26	NA	2	195	1	1
221	10.26	NA	2	418	1	2
222	10.26	NA	2	223	1	1
223	10.27	NA	2	232	1	1
224	10.37	NA	2	138	1	2
225	10.40	NA	2	190	1	2
226	10.41	NA	2	NA	1	1
227	10.41	NA	2	234	1	2
228	10.42	NA	2	218	1	1
229	10.43	NA	2	272	1	1

```
230  10.43        NA   2  367      1       1
231  10.44        NA   2  239      1       1
232  10.46        NA   2  222      1       1
233  10.48        NA   2  163      1       2
234  10.49        NA   2   NA      1       1
235  10.50        NA   2  180      2       4
236  10.51        NA   2  347      1       1
237  10.52        NA   2  154      1       1
238  10.57        NA   2   NA      1       3
239  10.57        NA   2   NA      1       2
240  10.60        NA   2  312      1       2
241  10.61        NA   2  211      1       2
242  10.62        NA   2  231      1       1
243  10.65        NA   2  281      1       1
244  10.68        NA   2  465      2       8
245  10.70        NA   2  171      1       1
246  10.71        NA   2  388      1       1
247  10.73        NA   2   NA      1       1
248  10.74        NA   2   NA      1       2
249  10.74        NA   2  244      1       3
250  10.77        NA   2  201      1       2
251  10.80        NA   2  184      1       1
252  10.83        NA   2   NA      1       1
253  10.92        NA   2   NA      1       1
254  10.92        NA   2   NA      1       2
255  11.03        NA   2   NA      1       2
256  11.03        NA   2  225      1       1
257  11.07        NA   2   NA      1       1
258  11.09        NA   2  280      2       2
259  11.14        NA   2  179      1       2
260  11.16        NA   2   NA     NA       2
261  11.19        NA   2  246      1       1
262  11.22        NA   2  157      1       1
263  11.22        NA   2  280      2       5
264  11.23        NA   2  284      1       1
 [ reached getOption("max.print") -- omitted 452 rows ]
> summary(girlsBet7to14)
      age            menarche          sex          igf1            tanner
 Min.   : 7.000   Min.   :1.000   Min.   :2   Min.   : 46.0   Min.   :1.000
 1st Qu.: 8.902   1st Qu.:1.000   1st Qu.:2   1st Qu.:199.0   1st Qu.:1.000
 Median :10.550   Median :1.000   Median :2   Median :269.0   Median :1.000
```

```
 Mean   :10.532   Mean   :1.132   Mean   :2   Mean   :317.0   Mean   :1.681
 3rd Qu.:12.127   3rd Qu.:1.000   3rd Qu.:2   3rd Qu.:413.5   3rd Qu.:2.000
 Max.   :13.990   Max.   :2.000   Max.   :2   Max.   :915.0   Max.   :5.000
                  NA's   :278                 NA's   :204     NA's   :63
    testvol
 Min.   : 1.000
 1st Qu.: 1.000
 Median : 1.000
 Mean   : 3.079
 3rd Qu.: 3.000
 Max.   :25.000
 NA's   :351


## Applied
> attach(vitcap)
The following objects are masked from vitcap (pos = 3):

    age, group, vital.capacity

> print(vitcap)
   group age vital.capacity
1      1  39           4.62
2      1  40           5.29
3      1  41           5.52
4      1  41           3.71
5      1  45           4.02
6      1  49           5.09
7      1  52           2.70
8      1  47           4.31
9      1  61           2.70
10     1  65           3.03
11     1  58           2.73
12     1  59           3.67
13     3  27           5.29
14     3  25           3.67
15     3  24           5.82
16     3  32           4.77
17     3  23           5.71
18     3  25           4.47
19     3  32           4.55
```

```
20     3  18           4.61
21     3  19           5.86
22     3  26           5.20
23     3  33           4.44
24     3  27           5.52
> vitcap1 <-vital.capacity[group=="1"]
> vitcap3 <-vital.capacity[group=="3"]
> plot(vitcap1)
> plot(vitcap3)
> boxplot(vitcap1)
> boxplot(vitcap3)
> hist(vitcap1)
> hist(vitcap3)
> qqnorm(vitcap$vital.capacity)
> shapiro.test(vitcap1)

Shapiro-Wilk normality test

data:  vitcap1
W = 0.91633, p-value = 0.257

> shapiro.test(vitcap3)

Shapiro-Wilk normality test

data:  vitcap3
W = 0.93421, p-value = 0.4269

> var.test(vital.capacity~group,conf.level=0.99)

F test to compare two variances

data:  vital.capacity by group
F = 2.3105, num df = 11, denom df = 11, p-value = 0.1806
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
  0.4343334 12.2911384
sample estimates:
ratio of variances
          2.310509
```

```
> t.test(vital.capacity~group,var.equal=T,conf.level=0.99)

Two Sample t-test

data:  vital.capacity by group
t = -2.9228, df = 22, p-value = 0.007882
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -2.04953499 -0.03713167
sample estimates:
mean in group 1 mean in group 3
       3.949167        4.992500

> wilcox.test(vital.capacity~group)

Wilcoxon rank sum test with continuity correction

data:  vital.capacity by group
W = 30.5, p-value = 0.01783
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(4.62, 5.29, 5.52, 3.71, 4.02, 5.09,  :
  cannot compute exact p-value with ties
> sorted_vitcap<-sort(vitcap$vital.capacity,decreasing=FALSE,na.last=NA)
> shapiro.test(sorted_vitcap[-c(1,24)])

Shapiro-Wilk normality test

data:  sorted_vitcap[-c(1, 24)]
W = 0.94249, p-value = 0.2228

> print(sorted_vitcap[-c(1,24)])
 [1] 2.70 2.73 3.03 3.67 3.67 3.71 4.02 4.31 4.44 4.47 4.55 4.61 4.62 4.77 5.09 5.20 5.29
[18] 5.29 5.52 5.52 5.71 5.82
> qqnorm(sorted_vitcap[-c(1,24)])
> detach(vitcap)


Faraway
> library(faraway)
```

```
> attach(pima)
The following objects are masked from pima (pos = 3):

    age, bmi, diabetes, diastolic, glucose, insulin, pregnant, test, triceps

The following objects are masked from pima (pos = 4):

    age, bmi, diabetes, diastolic, glucose, insulin, pregnant, test, triceps

The following objects are masked from pima (pos = 5):

    age, bmi, diabetes, diastolic, glucose, insulin, pregnant, test, triceps

The following objects are masked from pima (pos = 6):

    age, bmi, diabetes, diastolic, glucose, insulin, pregnant, test, triceps

The following object is masked from package:faraway:

    diabetes

> is.na(pima)<-!pima
> pima
     pregnant  glucose diastolic  triceps  insulin      bmi diabetes age test
1    6.000000 148.0000  72.00000 35.00000 155.5482 33.60000    0.627  50    1
2    1.000000  85.0000  66.00000 29.00000 155.5482 26.60000    0.351  31   NA
3    8.000000 183.0000  64.00000 29.15342 155.5482 23.30000    0.672  32    1
4    1.000000  89.0000  66.00000 23.00000  94.0000 28.10000    0.167  21   NA
5    4.494673 137.0000  40.00000 35.00000 168.0000 43.10000    2.288  33    1
6    5.000000 116.0000  74.00000 29.15342 155.5482 25.60000    0.201  30   NA
7    3.000000  78.0000  50.00000 32.00000  88.0000 31.00000    0.248  26    1
8   10.000000 115.0000  72.40518 29.15342 155.5482 35.30000    0.134  29   NA
9    2.000000 197.0000  70.00000 45.00000 543.0000 30.50000    0.158  53    1
10   8.000000 125.0000  96.00000 29.15342 155.5482 32.45746    0.232  54    1
11   4.000000 110.0000  92.00000 29.15342 155.5482 37.60000    0.191  30   NA
12  10.000000 168.0000  74.00000 29.15342 155.5482 38.00000    0.537  34    1
13  10.000000 139.0000  80.00000 29.15342 155.5482 27.10000    1.441  57   NA
14   1.000000 189.0000  60.00000 23.00000 846.0000 30.10000    0.398  59    1
15   5.000000 166.0000  72.00000 19.00000 175.0000 25.80000    0.587  51    1
16   7.000000 100.0000  72.40518 29.15342 155.5482 30.00000    0.484  32    1
17   4.494673 118.0000  84.00000 47.00000 230.0000 45.80000    0.551  31    1
```

18	7.000000	107.0000	74.00000	29.15342	155.5482	29.60000	0.254	31	1
19	1.000000	103.0000	30.00000	38.00000	83.0000	43.30000	0.183	33	NA
20	1.000000	115.0000	70.00000	30.00000	96.0000	34.60000	0.529	32	1
21	3.000000	126.0000	88.00000	41.00000	235.0000	39.30000	0.704	27	NA
22	8.000000	99.0000	84.00000	29.15342	155.5482	35.40000	0.388	50	NA
23	7.000000	196.0000	90.00000	29.15342	155.5482	39.80000	0.451	41	1
24	9.000000	119.0000	80.00000	35.00000	155.5482	29.00000	0.263	29	1
25	11.000000	143.0000	94.00000	33.00000	146.0000	36.60000	0.254	51	1
26	10.000000	125.0000	70.00000	26.00000	115.0000	31.10000	0.205	41	1
27	7.000000	147.0000	76.00000	29.15342	155.5482	39.40000	0.257	43	1
28	1.000000	97.0000	66.00000	15.00000	140.0000	23.20000	0.487	22	NA
29	13.000000	145.0000	82.00000	19.00000	110.0000	22.20000	0.245	57	NA
30	5.000000	117.0000	92.00000	29.15342	155.5482	34.10000	0.337	38	NA
31	5.000000	109.0000	75.00000	26.00000	155.5482	36.00000	0.546	60	NA
32	3.000000	158.0000	76.00000	36.00000	245.0000	31.60000	0.851	28	1
33	3.000000	88.0000	58.00000	11.00000	54.0000	24.80000	0.267	22	NA
34	6.000000	92.0000	92.00000	29.15342	155.5482	19.90000	0.188	28	NA
35	10.000000	122.0000	78.00000	31.00000	155.5482	27.60000	0.512	45	NA
36	4.000000	103.0000	60.00000	33.00000	192.0000	24.00000	0.966	33	NA
37	11.000000	138.0000	76.00000	29.15342	155.5482	33.20000	0.420	35	NA
38	9.000000	102.0000	76.00000	37.00000	155.5482	32.90000	0.665	46	1
39	2.000000	90.0000	68.00000	42.00000	155.5482	38.20000	0.503	27	1
40	4.000000	111.0000	72.00000	47.00000	207.0000	37.10000	1.390	56	1
41	3.000000	180.0000	64.00000	25.00000	70.0000	34.00000	0.271	26	NA
42	7.000000	133.0000	84.00000	29.15342	155.5482	40.20000	0.696	37	NA
43	7.000000	106.0000	92.00000	18.00000	155.5482	22.70000	0.235	48	NA
44	9.000000	171.0000	110.00000	24.00000	240.0000	45.40000	0.721	54	1
45	7.000000	159.0000	64.00000	29.15342	155.5482	27.40000	0.294	40	NA
46	4.494673	180.0000	66.00000	39.00000	155.5482	42.00000	1.893	25	1
47	1.000000	146.0000	56.00000	29.15342	155.5482	29.70000	0.564	29	NA
48	2.000000	71.0000	70.00000	27.00000	155.5482	28.00000	0.586	22	NA
49	7.000000	103.0000	66.00000	32.00000	155.5482	39.10000	0.344	31	1
50	7.000000	105.0000	72.40518	29.15342	155.5482	32.45746	0.305	24	NA
51	1.000000	103.0000	80.00000	11.00000	82.0000	19.40000	0.491	22	NA
52	1.000000	101.0000	50.00000	15.00000	36.0000	24.20000	0.526	26	NA
53	5.000000	88.0000	66.00000	21.00000	23.0000	24.40000	0.342	30	NA
54	8.000000	176.0000	90.00000	34.00000	300.0000	33.70000	0.467	58	1
55	7.000000	150.0000	66.00000	42.00000	342.0000	34.70000	0.718	42	NA
56	1.000000	73.0000	50.00000	10.00000	155.5482	23.00000	0.248	21	NA
57	7.000000	187.0000	68.00000	39.00000	304.0000	37.70000	0.254	41	1
58	4.494673	100.0000	88.00000	60.00000	110.0000	46.80000	0.962	31	NA

59	4.494673	146.0000	82.00000	29.15342	155.5482	40.50000	1.781	44	NA
60	4.494673	105.0000	64.00000	41.00000	142.0000	41.50000	0.173	22	NA
61	2.000000	84.0000	72.40518	29.15342	155.5482	32.45746	0.304	21	NA
62	8.000000	133.0000	72.00000	29.15342	155.5482	32.90000	0.270	39	1
63	5.000000	44.0000	62.00000	29.15342	155.5482	25.00000	0.587	36	NA
64	2.000000	141.0000	58.00000	34.00000	128.0000	25.40000	0.699	24	NA
65	7.000000	114.0000	66.00000	29.15342	155.5482	32.80000	0.258	42	1
66	5.000000	99.0000	74.00000	27.00000	155.5482	29.00000	0.203	32	NA
67	4.494673	109.0000	88.00000	30.00000	155.5482	32.50000	0.855	38	1
68	2.000000	109.0000	92.00000	29.15342	155.5482	42.70000	0.845	54	NA
69	1.000000	95.0000	66.00000	13.00000	38.0000	19.60000	0.334	25	NA
70	4.000000	146.0000	85.00000	27.00000	100.0000	28.90000	0.189	27	NA
71	2.000000	100.0000	66.00000	20.00000	90.0000	32.90000	0.867	28	1
72	5.000000	139.0000	64.00000	35.00000	140.0000	28.60000	0.411	26	NA
73	13.000000	126.0000	90.00000	29.15342	155.5482	43.40000	0.583	42	1
74	4.000000	129.0000	86.00000	20.00000	270.0000	35.10000	0.231	23	NA
75	1.000000	79.0000	75.00000	30.00000	155.5482	32.00000	0.396	22	NA
76	1.000000	121.6868	48.00000	20.00000	155.5482	24.70000	0.140	22	NA
77	7.000000	62.0000	78.00000	29.15342	155.5482	32.60000	0.391	41	NA
78	5.000000	95.0000	72.00000	33.00000	155.5482	37.70000	0.370	27	NA
79	4.494673	131.0000	72.40518	29.15342	155.5482	43.20000	0.270	26	1
80	2.000000	112.0000	66.00000	22.00000	155.5482	25.00000	0.307	24	NA
81	3.000000	113.0000	44.00000	13.00000	155.5482	22.40000	0.140	22	NA
82	2.000000	74.0000	72.40518	29.15342	155.5482	32.45746	0.102	22	NA
83	7.000000	83.0000	78.00000	26.00000	71.0000	29.30000	0.767	36	NA
84	4.494673	101.0000	65.00000	28.00000	155.5482	24.60000	0.237	22	NA
85	5.000000	137.0000	108.00000	29.15342	155.5482	48.80000	0.227	37	1
86	2.000000	110.0000	74.00000	29.00000	125.0000	32.40000	0.698	27	NA
87	13.000000	106.0000	72.00000	54.00000	155.5482	36.60000	0.178	45	NA
88	2.000000	100.0000	68.00000	25.00000	71.0000	38.50000	0.324	26	NA
89	15.000000	136.0000	70.00000	32.00000	110.0000	37.10000	0.153	43	1
90	1.000000	107.0000	68.00000	19.00000	155.5482	26.50000	0.165	24	NA
91	1.000000	80.0000	55.00000	29.15342	155.5482	19.10000	0.258	21	NA
92	4.000000	123.0000	80.00000	15.00000	176.0000	32.00000	0.443	34	NA
93	7.000000	81.0000	78.00000	40.00000	48.0000	46.70000	0.261	42	NA
94	4.000000	134.0000	72.00000	29.15342	155.5482	23.80000	0.277	60	1
95	2.000000	142.0000	82.00000	18.00000	64.0000	24.70000	0.761	21	NA
96	6.000000	144.0000	72.00000	27.00000	228.0000	33.90000	0.255	40	NA
97	2.000000	92.0000	62.00000	28.00000	155.5482	31.60000	0.130	24	NA
98	1.000000	71.0000	48.00000	18.00000	76.0000	20.40000	0.323	22	NA
99	6.000000	93.0000	50.00000	30.00000	64.0000	28.70000	0.356	23	NA

```
100  1.000000 122.0000  90.00000 51.00000 220.0000 49.70000    0.325  31    1
101  1.000000 163.0000  72.00000 29.15342 155.5482 39.00000    1.222  33    1
102  1.000000 151.0000  60.00000 29.15342 155.5482 26.10000    0.179  22   NA
103  4.494673 125.0000  96.00000 29.15342 155.5482 22.50000    0.262  21   NA
104  1.000000  81.0000  72.00000 18.00000  40.0000 26.60000    0.283  24   NA
105  2.000000  85.0000  65.00000 29.15342 155.5482 39.60000    0.930  27   NA
106  1.000000 126.0000  56.00000 29.00000 152.0000 28.70000    0.801  21   NA
107  1.000000  96.0000 122.00000 29.15342 155.5482 22.40000    0.207  27   NA
108  4.000000 144.0000  58.00000 28.00000 140.0000 29.50000    0.287  37   NA
109  3.000000  83.0000  58.00000 31.00000  18.0000 34.30000    0.336  25   NA
110  4.494673  95.0000  85.00000 25.00000  36.0000 37.40000    0.247  24    1
111  3.000000 171.0000  72.00000 33.00000 135.0000 33.30000    0.199  24    1
 [ reached getOption("max.print") -- omitted 657 rows ]
> pima$insulin[which(is.na(pima$insulin))]<-mean(pima$insulin, na.rm = TRUE)
> pima$diastolic[which(is.na(pima$diastolic))]<-mean(pima$diastolic, na.rm = TRUE)
> pima$pregnant[which(is.na(pima$pregnant))]<-mean(pima$pregnant, na.rm = TRUE)
> pima$triceps[which(is.na(pima$triceps))]<-mean(pima$triceps, na.rm = TRUE)
> pima$glucose[which(is.na(pima$glucose))]<-mean(pima$glucose, na.rm = TRUE)
> pima$diabetes[which(is.na(pima$diabetes))]<-mean(pima$diabetes, na.rm = TRUE)
> lmpima<-lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
> plot(lmpima)
Hit <Return> to see next plot: abline(lmpima)
Hit <Return> to see next plot: lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
Hit <Return> to see next plot: model.1=lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
Hit <Return> to see next plot: summary(model.1)
> confint(model.1, level=.98)
                     1 %         99 %
(Intercept) -1.192747037 -0.796093325
insulin     -0.000683601  0.001463019
> pima_fit<-lm(log(diabetes)~insulin,data=pima,(subset=age>=33))
> plot(fitted(pima_fit),resid(pima_fit))
> qqnorm(resid(pima_fit))
> qqline(resid(pima_fit))
> pima_fit

Call:
lm(formula = log(diabetes) ~ insulin, data = pima, subset = (subset = age >=
    33))

Coefficients:
(Intercept)      insulin
```

```
 -0.9944202    0.0003897

> lillie.test(log(diabetes))

Lilliefors (Kolmogorov-Smirnov) normality test

data:  log(diabetes)
D = 0.049932, p-value = 0.0001074

> cor.test(log(diabetes), insulin)

Pearson's product-moment correlation

data:  log(diabetes) and insulin
t = 1.9035, df = 766, p-value = 0.05735
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.002139533  0.138686197
sample estimates:
       cor
0.06861512
```