**INDIAN STATISTICAL INSTITUTE**

**Assignment on R-Programming.**

**ISEC- 73rd Batch**

**Specialisation: Data Processing**

**Name: FELIX MBOGOLELA**

**Country: Tanzania**

**Nature of the work: Assignment**

**Qn1:**

Read the data in R program, print the number of records, the number of variables, and the names of variables. Find out which variables are continuous, which are simply categorical (nominal), which are ordinal (maintaining the increasing or decreasing order of categories). Declare all the non-continuous variables as factor variables. Print the summary statistics of all the meaningful variables.

**Ans:** The number of records of a dataset is **189**, the number of variables is **11** and the names of the variables are:

"ID" "LOW" "AGE" "LWT" "RACE" "SMOKE" "PTL" "HT" "UI" "FTV" "BWT"

**Continuous variables** are: Identification code (ID), Age of mother (AGE), Weight of mother at last menstrual period (LWT) and birth weight (BWT).

**Categorical variables are**: Race (RACE), Smoking status during pregnancy (SMOKE), history of hypertension (HT), presence of uterine irritability (UI)

**Ordinal variables are**: History of premature labour (PTL), Low birth weight (LOW) and The number of physician visits during the first trimester (FTV).

Declaring all non-continuous variables as factor variables

1. bwdata$LOW=factor(bwdata$LOW)
2. bwdata$RACE = factor(bwdata$RACE)
3. bwdata$SMOKE=factor(bwdata$SMOKE)
4. bwdata$PTL=factor(bwdata$PTL)
5. bwdata$HT=factor(bwdata$HT)
6. bwdata$UI=factor(bwdata$UI)
7. bwdata$FTV=factor(bwdata$FTV)

Summary statistics for all meaningful information

Low birth weight (LOW): BWT <= 2500g =**59**

BWT > 2500 g =**130**

Age of mother (AGE):

Min. :14.00

1st Qu.:19.00

Median :23.00

Mean :23.24

3rd Qu.:26.00

Max. :45.00

Weight of mother at last menstrual period (LWT):

Min: 80.0

1st Qu.:110.0

Median :121.0

Mean :129.8

3rd Qu.:140.0

Max. :250.0

Race:

1=white=96

2=black=26

3=other=67

smoking status during pregnancy (SMOKE):

0=No=115

1=Yes= 74

History of premature labor (PTL),

0=None=159

1=One=24

2=Two= 5

3=Three=1

history of hypertension (HT)

0=No=177

1=Yes=12

presence of uterine irritability (UI)

0=No=161

1=Yes=28

The number of physician visits during the first trimester (FTV)

0=None=100

1=one=47

2=two=30

3=three= 7

4=four=4

6=six=1

birth weight (BWT).

Min. : 709

1st Qu.:2414

Median :2977

Mean :2945

3rd Qu.:3475

Max. :4990

QN2:

Present the univariate frequency distribution and then the sample percentage distribution of the patients having babies with normal birth weights and low birth weights and draw the bar chart with proper title. Write comments on your results.

**Ans: (i)** univariate frequency distribution table is

|  |  |
| --- | --- |
| Univariate frequency distribution of the patients | |
| having babies with normal birth weights | having babies with low birth weights |
| **130** | **59** |

|  |  |
| --- | --- |
| Sample percentage distribution of the patients | |
| having babies with normal birth weights | having babies with low birth weights |
| **69%** | **31%** |

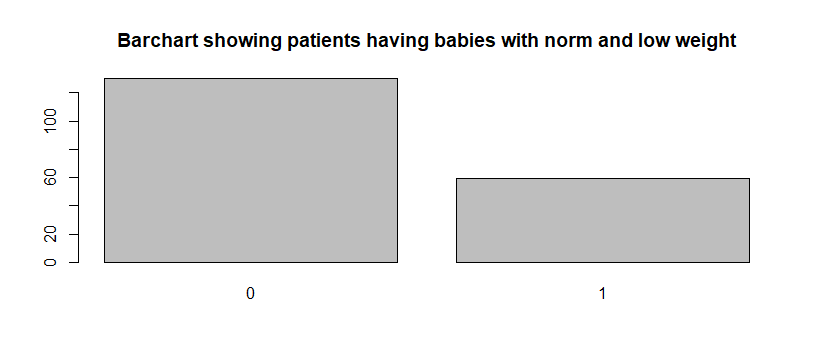


Figure 1 patients having babies with norm and low birth weight, o stands for normal weight above 2500g and 1 stands for low birth weight below 2500g

Qn3:

Present the bivariate sample percentage distribution of the patients having babies with normal birth weights and low birth weights by smoking habit and draw the suitable bar chart with proper title. Using this bar chart give the estimates for

Ans:

|  |  |  |
| --- | --- | --- |
|  | Norm birth weight | Low birth weight |
| Non smoking | 86=46% | 44=23% |
| smoking | 29=15% | 30=16% |
|  |  |  |

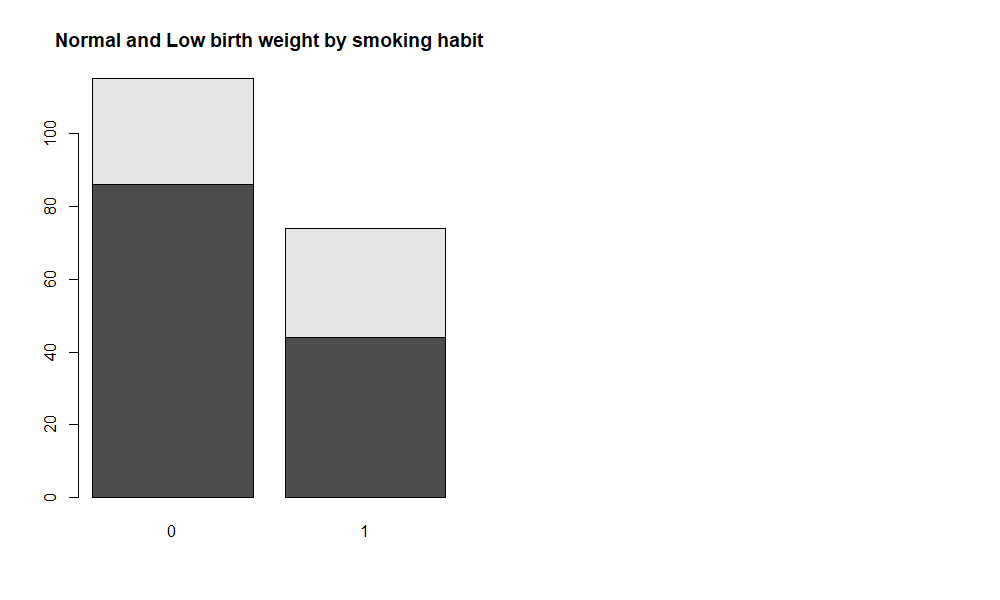


Figure 2 percentage distribution of patients having babies with norm and low birth weight, o stands for normal weight above 2500g and 1 stands for low birth weight below 2500g by smoking habit.

1. Percentage of low-weight babies for mothers who smoked during pregnancy were 16%
2. Percentage of low-weight babies for mothers who did not smoke during pregnancy were 23%
3. Test whether the occurrences of low weight babies is independent of the smoking habit of mothers.

Test: Ho: the two variables are independent, Vs H1: the two variables relate to each other.

Result: Since the p-value that came from the test is less than the predetermined significance level which is 0.05, then we reject the null hypothesis in favour of alternative hypothesis, that the smoking habit of pregnancy mothers causes low birth weight. The result of the test has shown below:

Pearson's Chi-squared test with Yates' continuity correction

data: bwdata$LOW and bwdata$SMOKE

X-squared = 4.2359, df = 1, p-value = 0.03958

**Qn4:**

Draw the boxplot of the age of the mothers in this data set with proper title. Use this boxplot to answer the following question

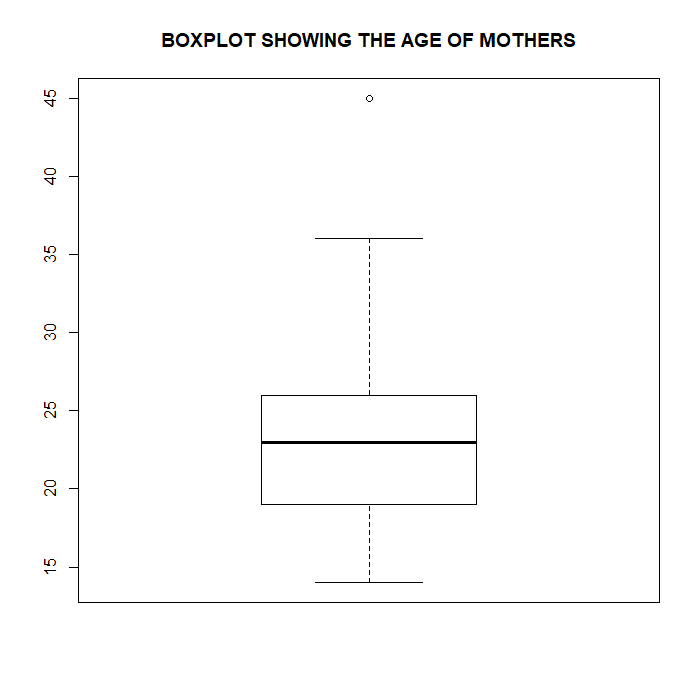
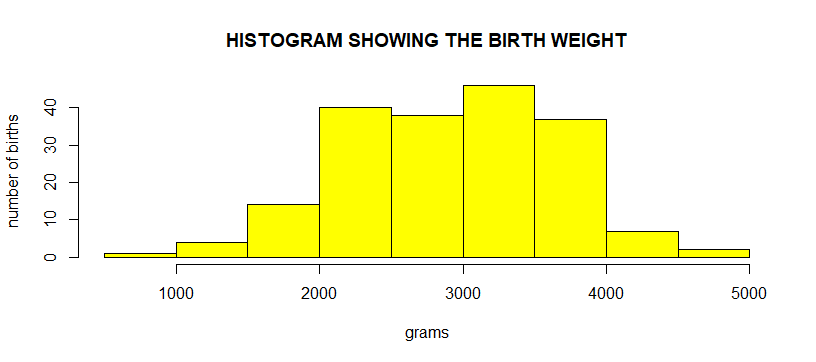
****

Figure 3: boxplot showing the distribution of mothers ages

1. The median age of the mothers is **23**
2. 75-th percentile of the distribution age is **26.**
3. the shape of the underlying distribution of the variable AGE appeared to be symmetric distributed since the mean and the median appeared to be located in the same place with small minor difference.
4. The IQR in the distribution of the age variable is 7
5. There is one outlier in this distribution which is 45

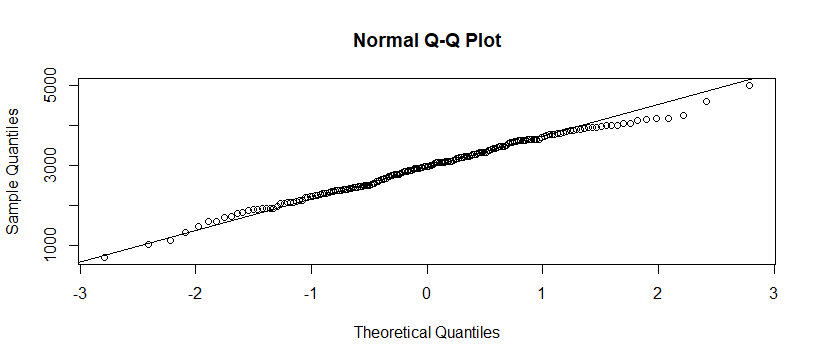
Qn5: Draw the histogram of birth weights (BWT) in this data set for all mothers and comment on the shape of the distribution of this data. Test whether this data can be considered to come from a Normal probability model. Answer graphically as well as based on some test.

Ans5:

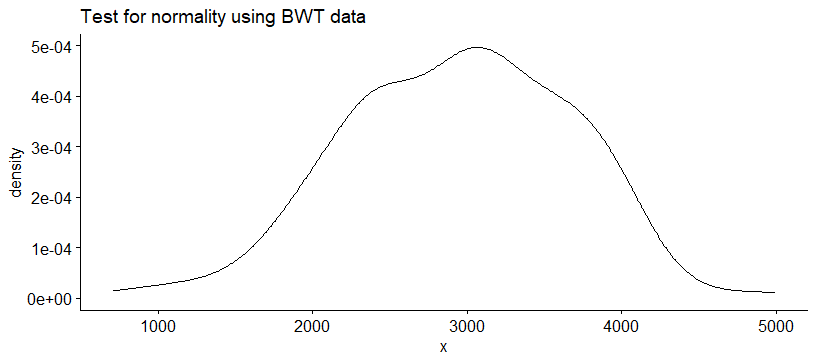


The histogram shows a bell-shaped picture which usually presents a normal distribution. Therefore, this data of birth weights is normally distributed.

1. Test whether this data can be considered to come from a Normal probability model
2. Graphically: Since all the points fall approximately along this reference line, then we can assume the normality, that all data come from a normal probability model as shown below;



1. Another test to check whether the BWT can be considered to come from a normal probability model has been tested using density plot which also proves the normality of the data as shown below: -



1. Normality based on **Shapiro-Wilk normality test**

**Null hypothesis**: The data is normally distributed. If p> 0.05, normality can be assumed.

The data has been tested by Shapiro-Wilk normality test

And the result shows that the p value is greater than 0.05, which also prove the

normality of these data. The p value is 0.4383.

Shapiro-Wilk normality test

data: bwdata$BWT

W = 0.99247, p-value = 0.4383

**Qn6:** Test whether the BWT data have come from a population having mean value greater than

2500 gm.

**Ans:** The BWT data was tested using t.test for one sample, that the data have come

from a population having mean greater than 2500.

Ho: population has mean greater than 2500 Vs H1: population has mean less than 2500

Since the p value is one which is greater than 0.05 then we can accept the hull hypothesis that

the data come from population having mean greater than 2500. The test is shown below

One Sample t-test

data: bwdata$BWT

t = 8.3852, df = 188, p-value = 1

alternative hypothesis: true mean is less than 2500

95 percent confidence interval:

-Inf 3032.312

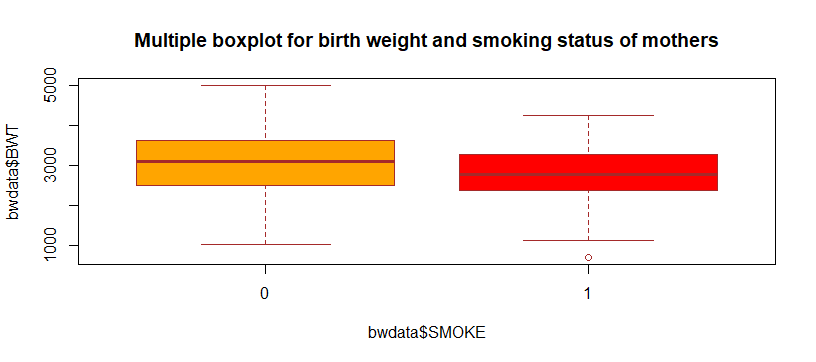
sample estimates:

mean of x

2944.656

**Qn7:** Draw the boxplot of the birth weights (BWT) in this data set separated by smoking status of mothers in a single boxplot window. Comment on your results.

**Ans:**

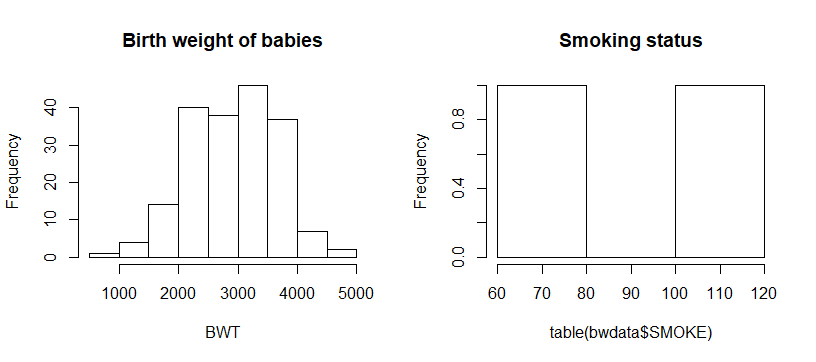
****

The boxplot shows that the weight of babies from non smoking mother tend to have a higher average birth weight than babies from smoking mother. The upper and lower whiskers in the two boxplots show that the birth weight of babies from smoking mothers has a higher variance than babies from nonsmoking mothers. Note 0 is equivalent to nonsmoking mothers and 1 is

equivalent to smoking mothers.

Qn8: Draw the histogram of the birth weights (BWT) in this data set separated by smoking status of mothers in a single plot window. Does the distribution of the variable BWT appear to have the same general shape for mothers who smoked and who did not smoke during pregnancy.

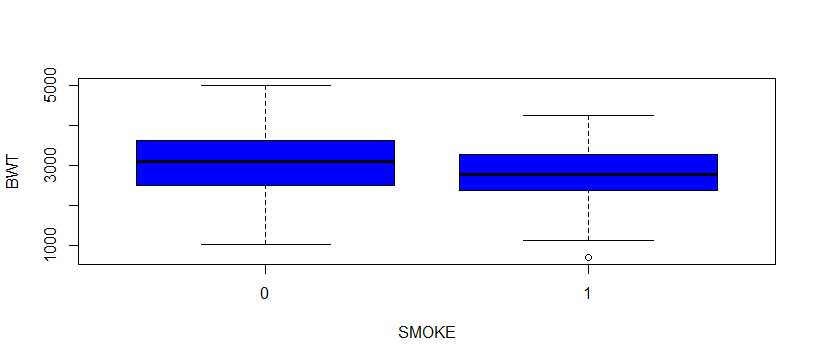
Ans:



The distribution of the variables appears to have the same general shape for both mothers who smoke and those who did not smoke during pregnancy.

QN9: Test whether the normal probability model can be considered as a plausible probability model for the distributions of BWT for mothers who smoked and who did not smoke. Answer graphically as well as by using testing of hypothesis.

Ans: (a) graphically



Since the box is symmetrical with the mean and the median at the centre, and also there is only one outlier hence the normal probability model can be considered as a plausible probability model for the distributions of BWT for mothers who smoked and who did not smoke.

Since the p-value is greater than 0.05 after testing by Chi squared test, then we can conclude that normal probability model can be considered as a plausible probability model for the distributions of BWT for mothers who smoked and who did not smoke, the output from the test has shown down below.

Pearson's Chi-squared test

data: bwdata$BWTS and (bwdata$SMOKE)

X-squared = 146.67, df = 132, p-value = 0.1808

QN10: Estimate the mean, variance, standard deviation for BWT for mothers who smoked and did not smoke during pregnancy. Also estimate the standard error (standard deviation / sqrt(n)) for estimating mean BWT for mothers who smoked and did not smoke during pregnancy. Then compute the 95% confidence interval for the mean birth weight to mother who smoked during pregnancy and comment on your results.

Ans: (a)

* Mean for BWT for smoking mothers was **2773.243**
* Mean for BWT for non-smoking mothers was **3054.957**
* Variance for BWT for smoking mothers was **435699.2**
* Variance for BWT for non-smoking mothers was **566119.3**
* Standard deviation for BWT for smoking mothers was **660.0752**
* Standard deviation for BWT for non-smoking mothers was **752.4090**
* Standard error for BWT for smoking mothers was **48.0134**
* Standard error for BWT for non-smoking mothers was **54.7297**

(b) The 95% confidence interval for the mean birth weight to mother who smoked during pregnancy are: **Lower CI was 2679.14 and Upper CI was 2867.35.**

We are 95% confident that the mean of BWT for smoking mothers in the population is lying between the 2679.137 and 2867.350.

Qn11. Estimate the 10th, 25th, 50th, 60th, 75th, 90th BWT percentiles by smoking status of the.

Ans: The percentiles of BWT by smoking status are as follows

10% 25% 50% 60% 75% 90%

78.10 84.25 94.50 98.60 104.75 110.90

Qn12: Test whether the sample data have come from population for which the mean birth

weight for non-smoker mothers is greater than for smoker mothers.

Ans: t test was done by assuming Ho: µo has come from the population in which the mean of

BWT for non-smoker mothers is greater than for smoker mother’s vs H1 that the mean came

from the population in which the mean of BWt for nonsmoker mothers is less.

We can accept the null hypothesis since the p-value is greater than 0.05, that the mean came from the population of mothers who were non-smokers during their pregnancy. Likewise, a non-parametric test was also done which again shows the p-values of 0.9965 which confirms the

claim. The results are given below from each test.

1. Welch Two Sample t-test

data: bwdata$BWT by bwdata$SMOKE

t = 2.7095, df = 170, p-value = 0.9963

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 453.6725

sample estimates:

mean in group 0 mean in group 1

3054.957 2773.243

(ii) Wilcoxon rank sum test with continuity correction

data: bwdata$BWT by bwdata$SMOKE

W = 5243.5, p-value = 0.9965

alternative hypothesis: true location shift is less than 0

Qn13: (a) Estimate the mean BWT for mothers of various races.

Ans: 1=White=3103.740

2=Black=2719.692

3=Other=2804.105

(b) Assess the normality assumption for BWT data for different races.

Ans:

(c) Test whether the variances for BWT data for mothers of different races are equal or not.

Ans: The variances for BWT data for mothers of different races are seems to be equal because

the test of equal variances shows that the p-values is greater than 0.05. the p value is 0.7209.

the result of the test is shown below.

Bartlett test of homogeneity of variances

data: bwdata$BWT by bwdata$RACE

Bartlett's K-squared = 0.65453, df = 2, p-value = 0.7209

(d) determine which test function is suitable to test whether the sample data have come from population for which the mean birth weight for mothers of different races are same or not.

Ans: **Bartlett test of homogeneity of variances**

(e) Perform the test accordingly.

Ans:

Bartlett test of homogeneity of variances

data: bwdata$BWT by bwdata$RACE

Bartlett's K-squared = 0.65453, df = 2, p-value = 0.7209

(f) If the test rejects the hypothesis of equal means, then determine which pairs have different means.

Ans: The test doesn’t reject the hypothesis since the p-value is 0.7209 which is greater than 0.05.

Qn14. Test whether the proportion of low weight births among mothers of various races are same or not.

Ans: Chi-square test and Fisher’s exact test were used to test the proportion of low weight births among mothers of various races are same or not. **Null hypothesis** was, all the proportional of low weight among mothers of different races are the same Vs **Alternative hypothesis** that the weight were different. The results showed that the p-values of **0.082** for chi-square test and **0.079** for Fisher’s exact test indicate that there is a chance by 8.2% or 7.9% that the Null hypothesis is correct. These fairly high values mean that we fail to reject the null hypothesis, and we conclude by saying races of pregnancy mothers has nothing to do with low births weight. All tests are shown below:

fisher. Test (table (bwdata$LOW, bwdata$RACE))

Fisher's Exact Test for Count Data

data: table (bwdata$LOW, bwdata$RACE)

p-value = 0.07889

alternative hypothesis: two. Sided

Pearson's Chi-squared test

data: table (bwdata$LOW, bwdata$RACE)

X-squared = 5.0048, df = 2, p-value = 0.08189

Qn15. Test whether the proportion of low weight births among mothers of two smoking status are same or not.

Ans: Ho: The same proportion of low birth mothers of 2 smoking status

H1: the proportion is not the same

Result: 2-sample test for equality of proportions with

continuity correction

data: table (bwdata$LOW, bwdata$SMOKE)

X-squared = 4.2359, df = 1, p-value = 0.03958

alternative hypothesis: two. Sided

95 percent confidence interval:

0.006401121 0.333624955

sample estimates:

prop 1 prop 2

0.661538 0.4915254

Since the p value is less than 0.05 then we can conclude that there is a difference in proportion of low weight birth among mothers of two smoking status.

The 95% confidence interval estimate of the difference between the low weight birth proportion of two smoking status is between 0.64% and 33.35%.

Qn16. Test whether the occurrences of low weight babies is independent of the races of mothers.

Ans: Chi-square test was used to test the independence of low weight babies against the mother

Races. Null hypothesis: there is independence or no association Vs Alternative hypothesis: there

is association. Results: since the p-value is 0.08189 which is greater than 0.05 at the significance

level we can accept null hypothesis that there is independence or no association between the low

weight babies and the races of their mothers. The test is shown below:

Pearson's Chi-squared test

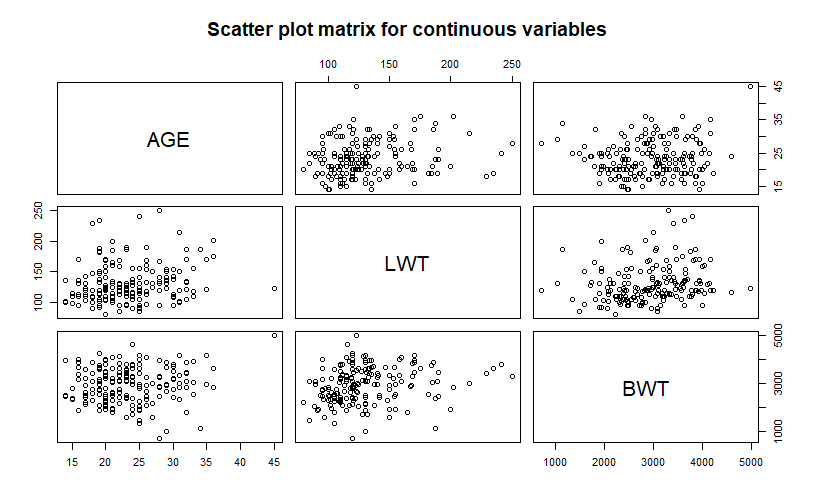
data: table (bwdata$LOW, bwdata$RACE)

X-squared = 5.0048, df = 2, p-value = 0.08189

Qn17. Draw the scatter plot matrix among all the continuous variables. Print the sample correlation coefficient matrix among them. Comment on your results.

Ans: scatter plot matrix showing all continuous variables

(a)



(b) sample correlation coefficients matrix

[AGE] [LWT] [BWT]

[AGE,] 1.00000000 0.1800732 0.08986639

[LWT] 0.18007315 1.0000000 0.18578871

[BWT] 0.08986639 0.1857887 1.00000000

From this sample correlation matrix, it is seen that, there is a weak positive correlation between the age of the mother and her weight at last menstrual period by 0.18. likewise, the birth

weight of her baby by 0.09. On the other hand, there is also a weak positive correlation

between the weight of mother at last menstrual period and the birth weight of her baby by 0.19

**Qn18.** Perform the multiple linear regression model for relating baby’s birth weight (BWT) to mothers age (AGE), mother’s weight at the last menstrual period (LWT), and also to the categorical variables : smoking status of mothers during pregnancy (SMOKE), History of premature labour (PTL), History of hypertension (HT), Presence of uterine irritability (UI), Number of physician visits during the first trimester (FTV).

Ans: Multiple linear regression performed as shown below:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2471.700 295.641 8.360 1.89e-14 \*\*\*

bwdata$AGE 3.713 9.623 0.386 0.700070

bwdata$LWT 4.372 1.763 2.480 0.014068 \*

bwdata$SMOKE1 -167.409 102.902 -1.627 0.105561

bwdata$PTL1 -436.902 153.914 -2.839 0.005067 \*\*

bwdata$PTL2 -54.476 304.867 -0.179 0.858390

bwdata$PTL3 1352.231 677.395 1.996 0.047460 \*

bwdata$HT1 -591.917 206.873 -2.861 0.004734 \*\*

bwdata$UI1 -527.699 141.682 -3.725 0.000264 \*\*\*

bwdata$FTV1 207.513 122.063 1.700 0.090897 .

bwdata$FTV2 33.644 141.308 0.238 0.812087

bwdata$FTV3 -312.468 260.211 -1.201 0.231442

bwdata$FTV4 276.736 344.088 0.804 0.422339

bwdata$FTV6 -198.285 697.767 -0.284 0.776616

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 656.8 on 175 degrees of freedom

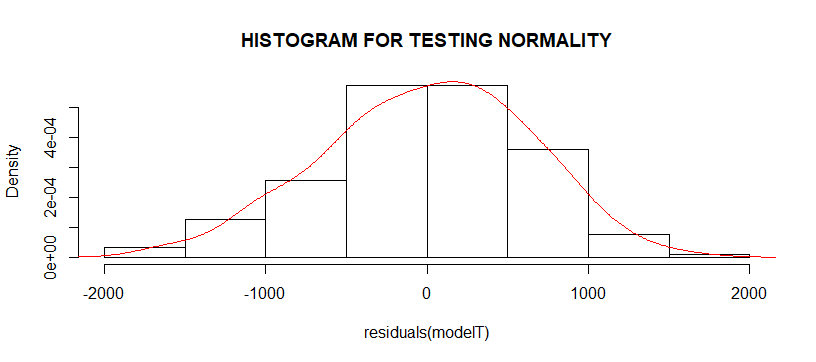
Multiple R-squared: 0.2444, Adjusted R-squared: 0.1882

F-statistic: 4.353 on 13 and 175 DF, p-value: 2.419e-06

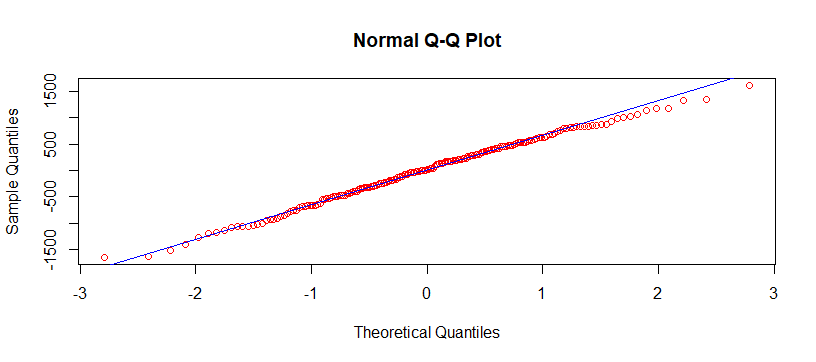
After the extraction of the model residuals normality was tested graphically as well as

hypothetically and here are the results:

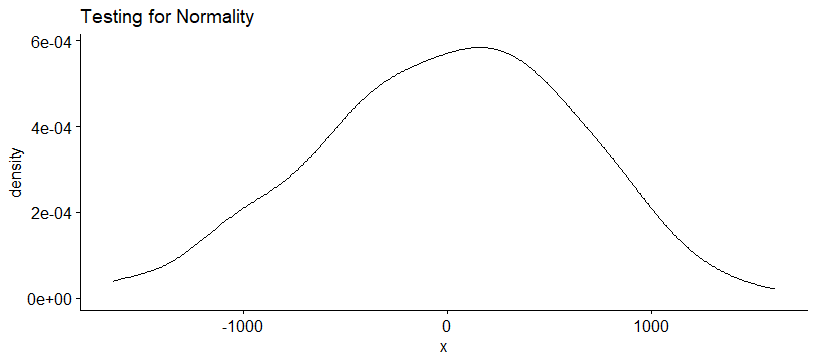
1. Graphically
2. By histogram aligned with density curve for checking normality



1. By q-q plot for checking normality



1. By density plot for normality checking.



1. One sample t test and shapiro. test were used to test the normality and each test

showed the positive result, thus residuals data are also normally distributed by

p-values which were equal to 1 and the other is 0.58. Here are the results.

Shapiro-Wilk normality test

data: residuals(modelT)

W = 0.99352, p-value = 0.5755

One Sample t-test

data: residuals(modelT)

t = 8.8551e-16, df = 188, p-value = 1

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

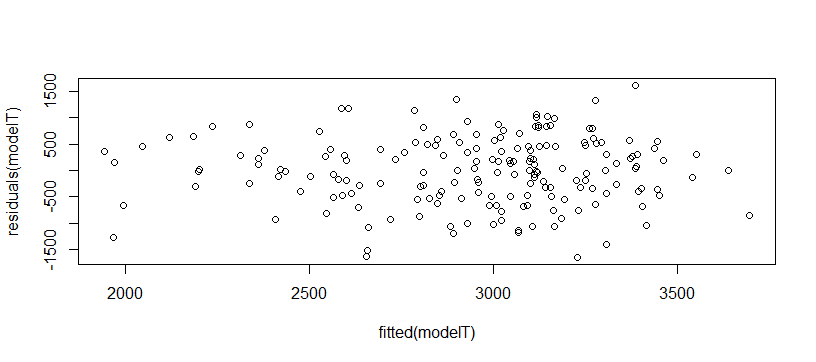
-90.93246 90.93246

sample estimates:

mean of x

4.081861e-14

1. plotting the model residuals Vs model fitted values.



From the data it can be seen that not all explanatory variables have significant effect on the bwt data.

R-squared =0.2444, it can be seen that 24.4% of R-squared implies that this is a good model since this small variance fits the data and there is close relationship between the independent variables and the outcome.

Qn19: From the above regression output, exclude the unimportant explanatory variables and perform the regression analysis again. Print the output of new regression analysis, and check now whether all explanatory variables have significant effect or not and also check that R^2-adjusted is improved (increased) and choose the model having largest R^2-adjusted value and fewest explanatory variables. Write down the final model.

**Ans:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2464.503 224.852 10.961 < 2e-16 \*\*\*

bwdata$LWT 4.672 1.680 2.781 0.00598 \*\*

bwdata$HT1 -663.585 208.893 -3.177 0.00175 \*\*

bwdata$UI1 -568.368 140.979 -4.032 8.09e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 678.5 on 185 degrees of freedom

Multiple R-squared: 0.1476, Adjusted R-squared: 0.1338

F-statistic: 10.68 on 3 and 185 DF, p-value: 1.65e-06

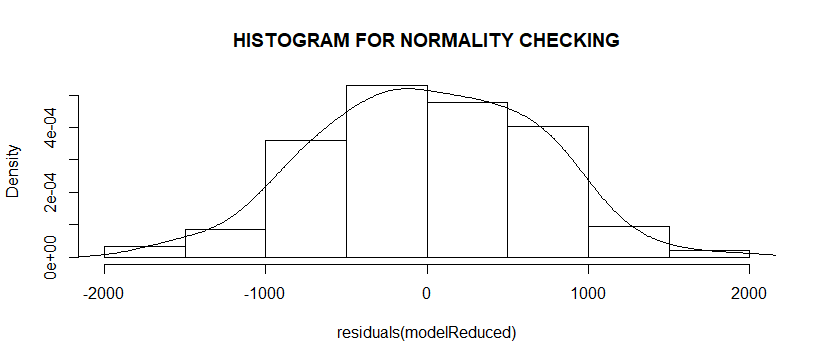
Since the R-square is now 0.1476 which is equivalent to 14.76 it shows that the model now is better than before by removing some explanatory variables.

The normality has been tested graphically as well as hypothetically and here down below are

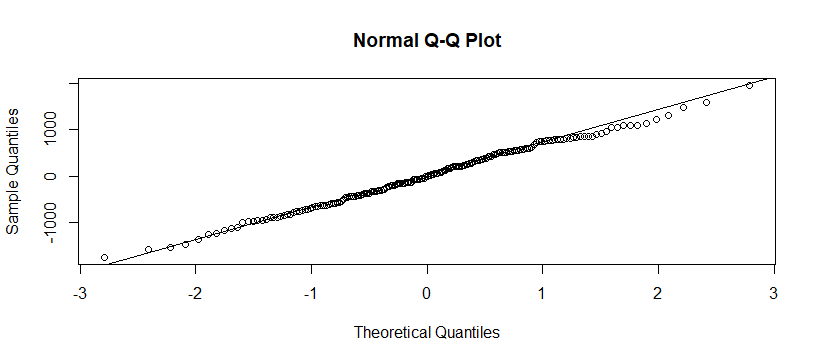
the results of those tests.

The final model is:

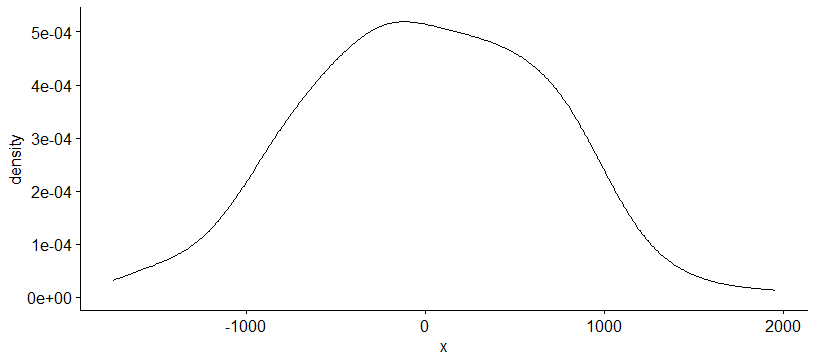
1. Graphically
2. By histogram aligned with density curve for checking normality



1. By q-q plot for checking normality

****

1. By density plot for normality checking.



1. One sample t test and shapiro. test were also used to test the normality and each

test showed the positive result, thus residuals data was also normally distributed by

p-values which were equal to 1 and the other is 0.8526. Here are the results.

Shapiro-Wilk normality test

data: residuals(modelReduced)

W = 0.99554, p-value = 0.8526

One Sample t-test

data: residuals(modelReduced)

t = -1.8736e-16, df = 188, p-value = 1

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-96.5774 96.5774

sample estimates:

mean of x

-9.172594e-15

Qn20: Consider the binary response variable of having low weight birth (LOW= 0 for no and = 1 for yes). Fit a logistic regression model for this variable of having or not having low weight baby based on the explanatory variables mother’s age, mother’s weight at last menstrual period and the smoking status of mother during pregnancy.

Ans: the output is:

Call:

glm(formula = bwdata$LOW ~ bwdata$AGE + bwdata$LWT + bwdata$SMOKE,

family = binomial(link = logit))

Deviance Residuals:

Min 1Q Median 3Q Max

-1.2829 -0.8650 -0.6938 1.2624 2.0103

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.368225 1.014262 1.349 0.1773

bwdata$AGE -0.038995 0.032726 -1.192 0.2334

bwdata$LWT -0.012139 0.006135 -1.979 0.0479 \*

bwdata$SMOKE1 0.670764 0.325878 2.058 0.0396 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom

Residual deviance: 222.88 on 185 degrees of freedom

AIC: 230.88

Number of Fisher Scoring iterations: 4

From this output it is seen that AGE has no effect on low birth weight (LOW) while smoking habit (SMOKE) has some effects. Each unit increase in smoking habit increases the log odds of getting low birth weight by 0.67 and the other explanatory variable

that is LWT is that each one-unit change will decrease (LOW) by an effect of -0.01. Note that,

the p-value indicates that it is some what significant in determining the low birth weight. Also,

the difference between null deviance and residual deviance tells us that the model is a good fit. The greater the difference the greater the model.

Qn21: In above logistic regression output, check whether all the explanatory variables have significant effect on having the low weight baby. If not, exclude from the model the unimportant variables and re-perform the logistic regression analysis.

**Ans: Output is seen below.**

Call:

glm(formula = bwdata$LOW ~ bwdata$LWT + bwdata$SMOKE, family = binomial(link = logit))

Deviance Residuals:

Min 1Q Median 3Q Max

-1.2487 -0.8459 -0.7374 1.2491 2.0808

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.62200 0.79592 0.781 0.4345

bwdata$LWT -0.01332 0.00609 -2.188 0.0287 \*

bwdata$SMOKE1 0.67667 0.32470 2.084 0.0372 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom

Residual deviance: 224.34 on 186 degrees of freedom

AIC: 230.34

Number of Fisher Scoring iterations: 4

The final model is

Each unit increase in smoking habit increases the log odds of getting low birth weight by 0.67,

each one-unit change in LWT will decrease (LOW) by an effect of -0.01. Note that, the p-value

indicates that it is somewhat significant in determining the low birth weight. Also, the difference between null deviance and residual deviance tells us that the model is a good fit. The greater

the difference the greater the model.

Qn22: From model of previous question, extract the estimates and compute the Odds Ratios as the exponential of the coefficients. Comment on the Odds Ratios.

Ans: odds\_ratious

(Intercept) bwdata$AGE bwdata$LWT bwdata$SMOKE1

3.9283727 0.9617559 0.9879348 1.9557304

Note that after output the AGE and the weight of mother at her last menstrual period have no

effect in increasing or decreasing the probability of low birth weight for a baby, but for smoking

variable if it will be kept fixed, the female smoking mother have a value of 1.9557 higher chance to affect the low birth weight of her baby