THE AMERICAN UNIVERSITY IN CAIRO | SCHOOL OF SCIENCES AND ENGINEERING

Mfäct | DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

Project Report

Multivariate Analysis (MACT4233-02)

Dr Ali Hadi

Spring 2023

Laila El Saadawi

900201723

Masa Tantawy

900201312

Table of Contents:

Introduction

The period of pregnancy, medically known as the ‘gestation’ period, begins at conception and ends with the delivery of the baby. During this time, mothers are instructed to keep their bodies at the optimal possible state of health. This is to ensure the safe delivery of the baby, which goes through stages of fetal development and growth that are greatly dependent on the mother’s bodily state. Medical professionals advise mothers to completely avoid alcoholic beverages and any form of smoking, as well as to limit their caffeine intake to the bare minimum. In the womb, such harmful behaviors could cause tissue damage in the unborn baby’s lungs or brain, or even lead to the termination of the pregnancy altogether in the form of a miscarriage or premature labor. After delivery, the mother and child(ren) are weighed and examined to ensure that the process has gone smoothly. Babies classified as having a low birth weight are those born weighing less than 5 pounds, 8 ounces (2,500 g). Issues with the baby’s weight and general health could arise as a result of the mother being too old or too young, or even as a consequence of a previous, turbulent pregnancy experience.

This project’s dataset (babies.csv) provides information on a variety of features that pertain to the pregnancy process, as well as measurements of the mother and child after the delivery’s completion. Identifying outliers is an integral part of this process, as any data points that fall too out of range (especially with features pertaining to the weights of mothers or children or the gestation period) indicate issues with the pregnancy that require immediate medical attention. We will be applying multivariate techniques, namely Hotelling's T^2 test and the BACON algorithm, to identify and handle outliers and establish statistical qualities.

The aim of this project is to study two sets of key traits included in the dataset: smokers and first-time mothers. Through the application of multivariate analysis, this project should clarify any differences between mothers who smoke and those who don’t, as well as between mothers experiencing first pregnancies and those experiencing successive ones. This should help in answering questions such as, but not limited to, how much of a role smoking plays in determining the gestation period or baby weight at birth, and whether or not first-time mothers endure more issues, in terms of weight at delivery or length of the gestation term, than mothers with added experience. In addition, through the identification of outliers, conclusions can be drawn regarding the degree of robustness of the data and consequently whether further analysis is reasonable.

Data Description

Data Source: Kaggle

https://www.kaggle.com/datasets/debjeetdas/babies-birth-weight

The dataset provides data on pregnancies between 1960 and 1967 among women in the San Francisco East Bay area.

Number of observations and variables: 1236 observations and 8 variables

Variables

	Name	Description	Type	Units of Measurement
1.	case	ID number	Quantitative	(Unitless)
2.	bwt	The baby's birth weight	Quantitative	Ounces (oz)
3.	gestation	The length of gestation/ pregnancy of the mother	Quantitative	Days
4.	parity	Whether or not this pregnancy is not the mother's first	Binary	0: First pregnancy 1: Not first Pregnancy
5.	age	Mother's age at birth	Quantitative	Years
6.	height	Mother's height	Quantitative	Inches (in)
7.	weight	Mother's weight	Quantitative	Pounds (lbs)
8.	smoke	Whether or not the mother smokes	Binary	0: Does not Smoke 1: Smokes

Data Preparation

To be able to work and analyze the data, it needs to be prepared first. The column *case* which represents an ID number for each observation was meaningless hence it was eliminated from the dataset.

i. Missing Values

Next, after checking for missing values, it appeared that the dataset contained 83 missing values in total. 62 observations out of 1,236 contained missing values. The columns with NA values were: *gestation, age, height, weight,* and *smoke*. To deal with missing values, missing value imputations were performed using 2 approaches; for continuous numeric columns, the missing values were replaced by the median, and for the binary column smoke, kNN using the 3 nearest neighbors then applying a threshold at 0.5 to classify the observations was used as median imputation is not reasonable. Even though the mean imputation technique for numeric columns is more popular than the median, replacing any missing values with the column's mean is non-robust as it is affected by outliers.

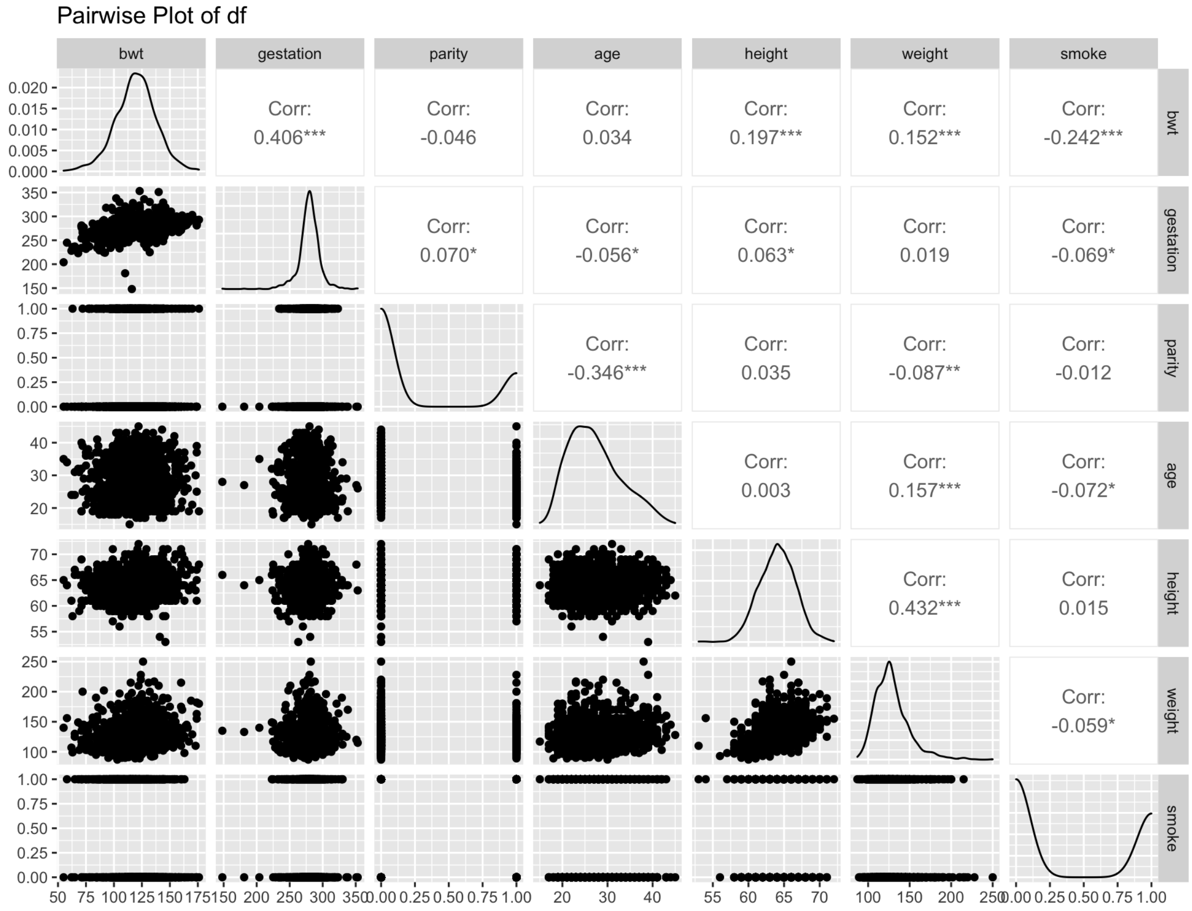
In fact, the missing values were initially replaced by the column means but when outliers were detected in the next stage, we decided that median imputation (a more robust measure) is a more plausible substitute that can improve analysis and reduce bias by reducing the effect of extreme outliers on the given data.

ii. Column Rank

To analyze the dependency relationship between the different features in our dataset, the column rank was calculated. The column rank was 7, which is equal to the number of variables in the dataset after dropping the column *case* as aforementioned. This indicates that the dataset is a full column rank matrix; all variables are linearly independent from each other.

iii. Variables

For the final step of preparing the data, the summary of the columns in addition to their relationship with each other were examined. For the relationship between the variables, the pairwise scatter plot and the correlation coefficient was used. This is shown below.



From the plots, it appears that the *height* and *weight* of the mother were the two variables with the strongest correlation (r = 0.432). The points on the corresponding plot reflected this positive correlation and were clustered in a generally increasing slope with only a few outliers.

The variables *gestation* and *bwt*, which represent the mother's pregnancy term and the baby's birth weight respectively, were also found to be moderately positively correlated (r = 0.406). This indicates that the longer the gestation period, the healthier the baby is by birth weight standards, a conclusion that supports the fact that many underweight babies are a result of incomplete pregnancy terms. The background research also revealed that the mother's weight played a role in the baby's birth weight, as stated in the introduction. However, the correlation between these two variables in the dataset (*bwt* and *weight*) was not a strong one (r = 0.152). As a result, this might indicate that the length of the gestation period plays a more prominent role in determining the baby's weight, as opposed to the mother's weight at the time of delivery.

In addition, the plots show a non-linear relationship between *age* and *bwt*, which is generally expected since research states that women who were either too young or too old were the main demographic at risk for complications with the baby's health. Therefore, this supports the outcome of the correlation matrix (r = 0.034) and the pairs plot.

With regards to smoking as a contributing factor in the mothers' and their babies' weights, there was a stronger negative correlation between smoking and baby birth weights (r = -0.242) than there was between smoking and mother's weights at delivery (r = -0.059 which is very close to 0). Being a smoker is thus assumed to be more detrimental to the baby's development (as demonstrated by inhibited growth) than it is a determinant of the mother's weight.

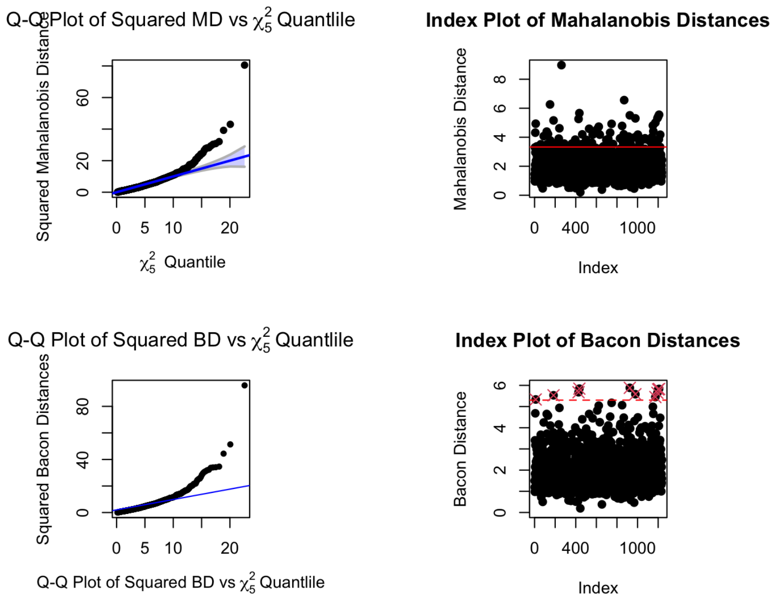
Interestingly, the variables *parity* and *age* were found to be negatively correlated (r = -0.346). For further clarification, *parity* is defined with '1' being a mother who has had no previous pregnancies. The negative correlation thus indicates that for higher parity values (i.e, values equal to 1 since *parity* is binary), ages were generally lower. In other words, the data demonstrated a trend of older mothers going through their first pregnancies as opposed to the expected alternative of older mothers having had previous pregnancy experiences.

Detection of Outliers

In order to identify outliers in our data, the blocked adaptive computationally efficient outlier nominators (BACON) algorithm proposed by Billor, Hadi, and Velleman (2000) was utilized. This algorithm is an effective method for the expeditious calculation of a set of multivariate outliers. This technique was applied only on numeric columns which were stored in a new dataframe. This new dataframe excludes the binary columns *smoke* and *parity* and is of dimensions n = 1236 and p = 5.

The initial basic subset has m = 20; the result of min (4*5, 1236*0.5). The algorithm computed the threshold value as equal to 5.303597 and adjusted the basic subset accordingly over 8 iterations (until a redundancy in the basic subset occurred). The final basic subset encompassed 1223 out of 1236 observations, meaning 13 outliers have been identified, which represent 1.05% of the total data. These points appear above the dotted horizontal line, representing the threshold value, in the index plot of the bacon distances (figure 2 in the appendix).

To validate these results, 4 plots were examined: q-q plot of the squared Mahalanobis distances against the chi square (5) quantile, index plot of Mahalanobis distances, q-q plot of the squared bacon distances against the chi square (5) quantile, index plot of bacon distances. These plots are shown below.



It is clear from the 2 q-q plots that the points do not lie on the straight line for larger quantiles/ distances, indicating the presence of outliers. The index plot of the Mahalanobis distances shows that many points are outliers; however, this is incorrect as it uses the column means which is a non-robust measure that is affected by outliers. Hence, the index plot of the BACON distances identifies the true outliers in the dataset, marked by an X in the plot, as it uses a robust measure through an iterative process.

Hotelling's T^2 test

Now that outliers in the dataset were identified, it is ready for analysis. This project's aim is to determine whether the multivariate means for 2 groups are equal or not; to do this, the Hotelling's T^2 test will be used. This process will be implemented twice, once for the *parity* (split into one group for the first pregnancy and another for its complement) and another for *smoke* representing the 2 groups of women who smoke versus those who do not smoke.

Before applying the test, the assumptions of theHotelling's T^2 test must be checked. These are: the data follows a multivariate normal distribution and that the variances of the two subgroups are almost equal. This was done for the first 2 groups, first versus successive pregnancies, as well as the second 2 groups, smokers versus non-smokers.

First: First vs Successive Pregnancy

To validate the first assumption, the histograms of each variable for the 2 groups were examined, and it appeared that the 2 variables *age* and *weight* needed transformation as they did not follow a normal distribution. This is visible in figures 3.1 and 3.2, for first and successive pregnancies respectively, in the appendix. Thus, log transformation was applied to these 2 columns in both groups to make them approximately normal as the histograms illustrated higher normality after transformation. These histograms are shown in figures 4.1 and 4.2 in the appendix. As for the assumption of equality of variances, the mean relative difference between the covariances matrices of the 2 groups after transformation was 0.1861828 , which is relatively small as it is close to zero. Now the equality of the means can be tested. The null hypothesis (H_0) is that both means are equal ($\mu_{\text{first}} = \mu_{\text{successive}}$) while the alternative hypothesis (H_1) is that the two means are unequal ($\mu_{\text{first}} \neq \mu_{\text{successive}}$).

First, the non-robust Hotelling's T^2 test was used for hypothesis testing of the equality of the means of the 2 groups. This test concluded that the null hypothesis should be strongly rejected since the F value is much greater than the tabulated values based on 5 and 1230 degrees of freedom and a significance level of 0.05, or equivalently the p-value is equal to 0 which is less than the significance level. This means that the population means are unequal.

Next, the robust Hotelling's T^2 test was used for the same hypothesis; this measure should yield a more accurate conclusion since the dataset contains outliers. Similar to the non-robust test, this test concluded that the true difference between the means is not equal to zero as shown by the f-test and the p-value.

Second: Smokers vs Non-Smokers

For the second two groups (smokers v.s. non-smokers), the same methodology demonstrated above was applied in order to validate the two assumptions. To verify the normality assumption, histograms for each variable in the dataset were constructed for each subgroup (as seen in figures 5.1 and 5.2). Similar to the findings of the previous section, *age* and *weight* were found to be deviant from the normal distribution therefore log transformation was applied to these variables. Figures 6.1 and 6.2 in the appendix demonstrate the histograms after applying the log transformation to approximate *age* and *weight* as normal distributions.

The mean relative difference between the two subgroup's covariance matrices was calculated to test the second assumption of variance equality between the two groups, and it was found to be 0.16638. As this is close to zero, the second assumption is met and Hotelling's test can be performed. The null hypothesis (H_0) and alternative hypothesis (H_1) are that both means are equal ($\mu_{smokers} = \mu_{non\text{-}smokers}$) versus unequal ($\mu_{smokers} \neq \mu_{non\text{-}smokers}$) respectively.

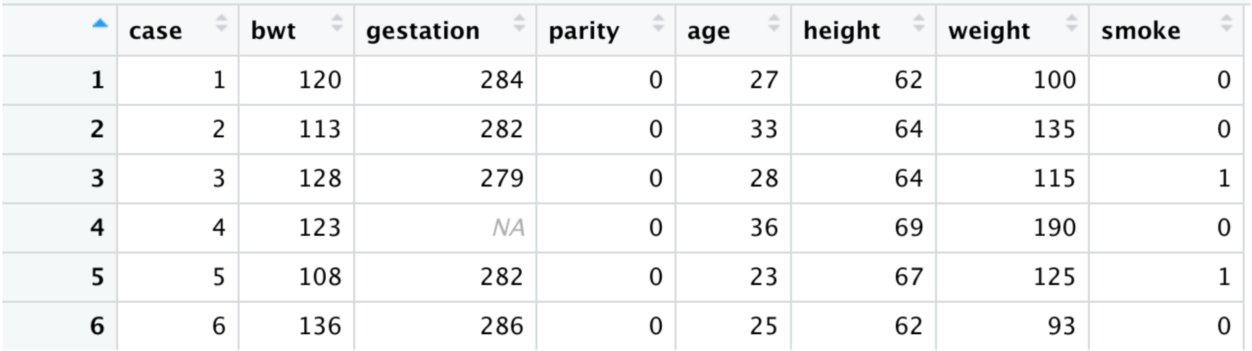
The non-robust Hotelling's T^2 test was used for hypothesis testing of the equality of the means of the 2 groups followed by the robust test. Both tests concluded that the null hypothesis should be strongly rejected since the F value is much greater than the tabulated value based on 5 and 1230 degrees of freedom and a significance level of 0.05, or equivalently the p-value is equal to 0 which is less than the significance level. This means that the population means are unequal.

Conclusion

Finally, the dataset used in this project describes the details of pregnancies of multiple women. The aim was to identify if the dataset contained any outliers, which was tested using BACON, as well as perform analysis on it to identify whether first vs successive pregnancies differed and whether smoking affected pregnancy. The equality of the means of each of the 2 groups was tested using Hotelling's T^2 test. It was found that 13 observations out of the 1236 in the dataset were outliers, representing 1.05% of the total data. Since the vast majority of the observations is robust, further analysis can be performed. The data was prepared for testing the equality of the means, such as through transformations and dividing the observations into groups. It was concluded that for the first 2 groups, women for whom this was their first pregnancy compared to those for whom this was a successive pregnancy, the null hypothesis should be rejected as the two population means are not equal (shown in figure 7). Similarly, for the second 2 groups, the population means for women who smoke compared to those who do not were unequal, as visible from figure 8, so the null hypothesis was also rejected. The population means showed that the mean birth weight for first pregnancies were higher than for non-first ones despite that the other variables did not show any magnified difference. For future research, it would be interesting to further investigate the reason for this and understand what exactly causes this difference. Unsurprisingly, it appeared that women who smoke on average have babies with lower birth weights than those who do not smoke which may be due to their shorter gestation periods on average. This proves that smoking is extremely harmful to babies and has further implications than on a child's health.

Appendix

Figure 1: Head of initial Dataset



	case	bwt	gestation	parity	age	height	weight	smoke
1	1	120	284	0	27	62	100	0
2	2	113	282	0	33	64	135	0
3	3	128	279	0	28	64	115	1
4	4	123	NA	0	36	69	190	0
5	5	108	282	0	23	67	125	1
6	6	136	286	0	25	62	93	0

Figure 2: BACON distances plot

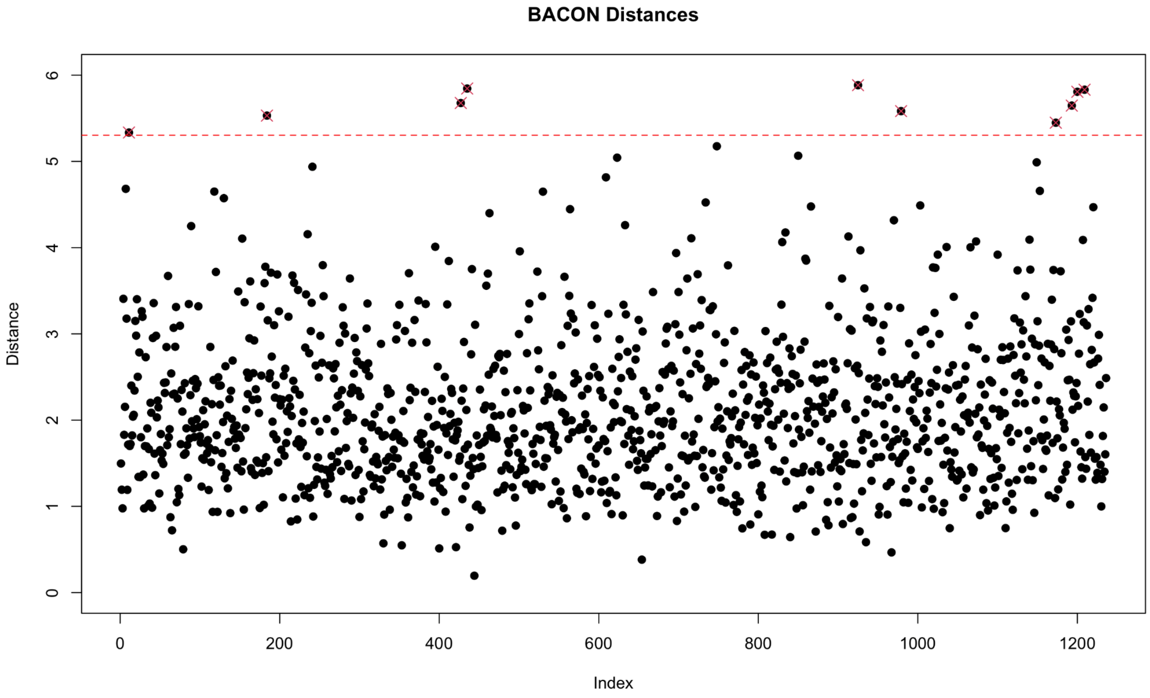


Figure 3.1: Histogram of First Pregnancies Before Transformation

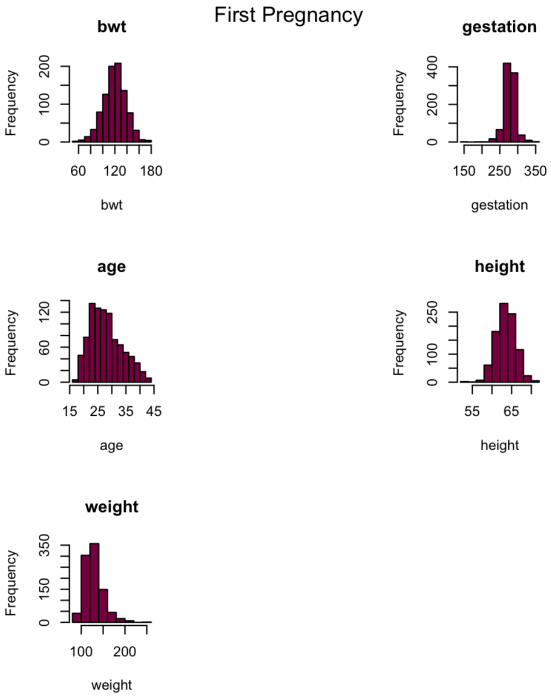


Figure 3.2: Histogram of Successive Pregnancies Before Transformation

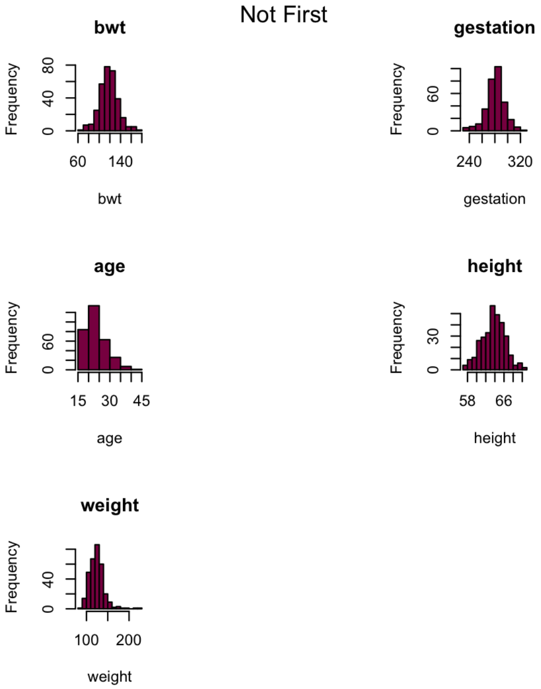


Figure 4.1: Histogram of First Pregnancies After Transformation

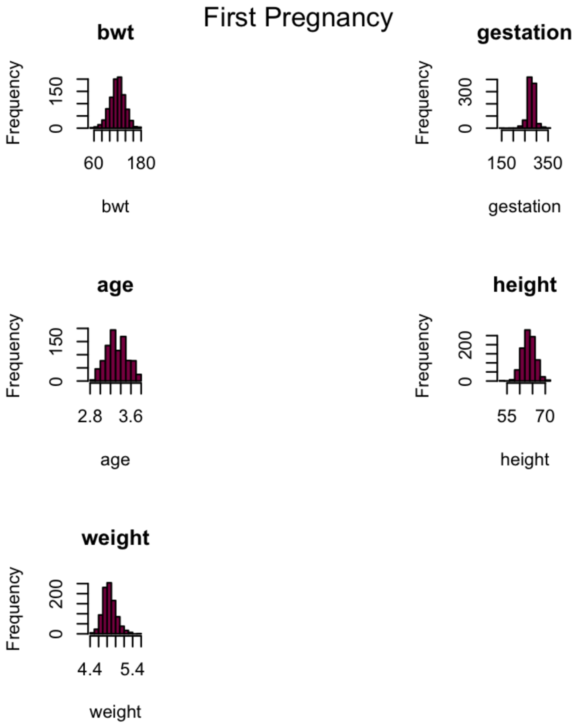


Figure 4.2: Histogram of Successive Pregnancies After Transformation

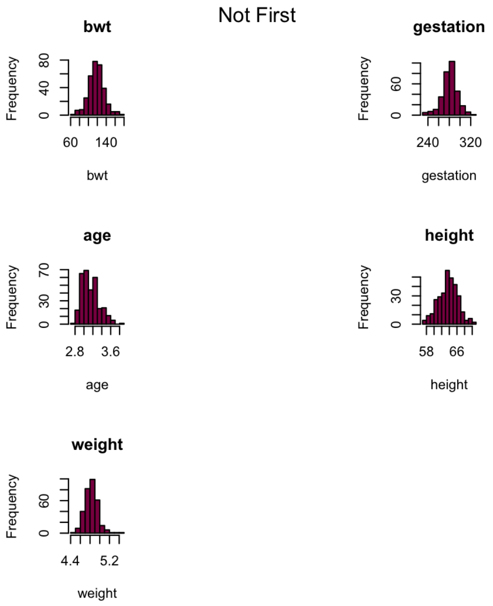


Figure 5.1: Histogram of Smokers Before Transformation

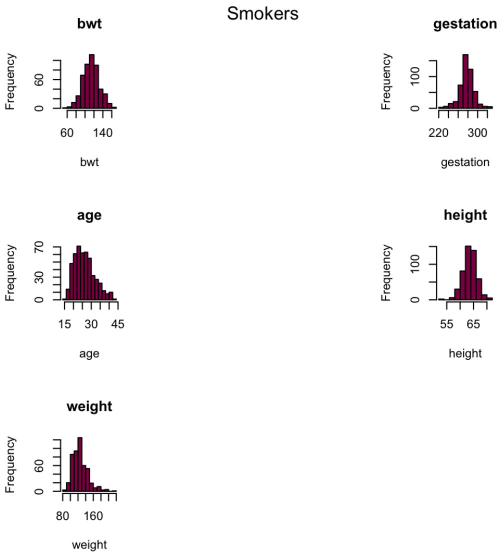


Figure 5.2: Histogram of Non-Smokers Before Transformation

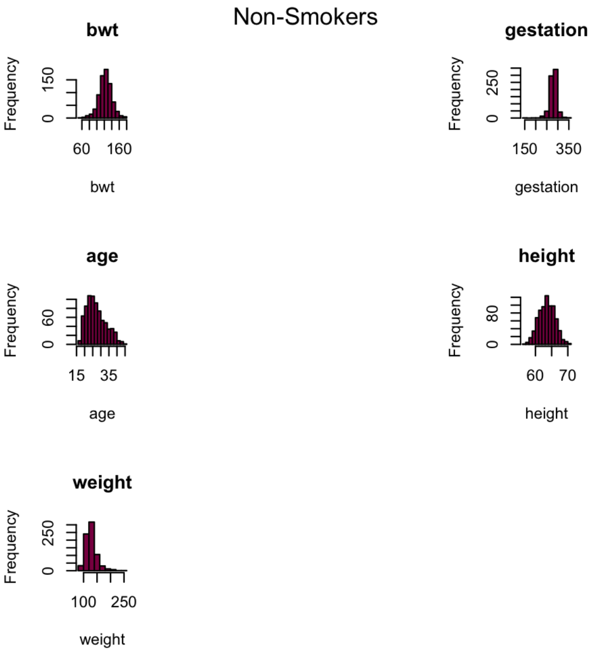


Figure 6.1: Histogram of Smokers After Transformation

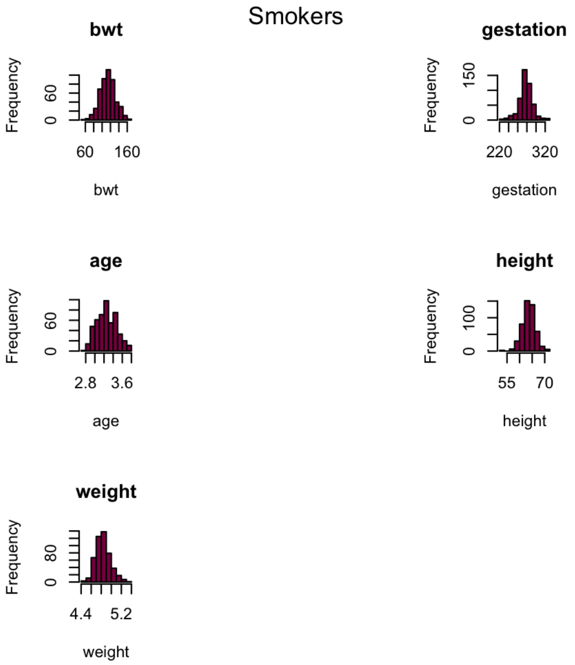


Figure 6.2:Histogram of Non-Smokers After Transformation

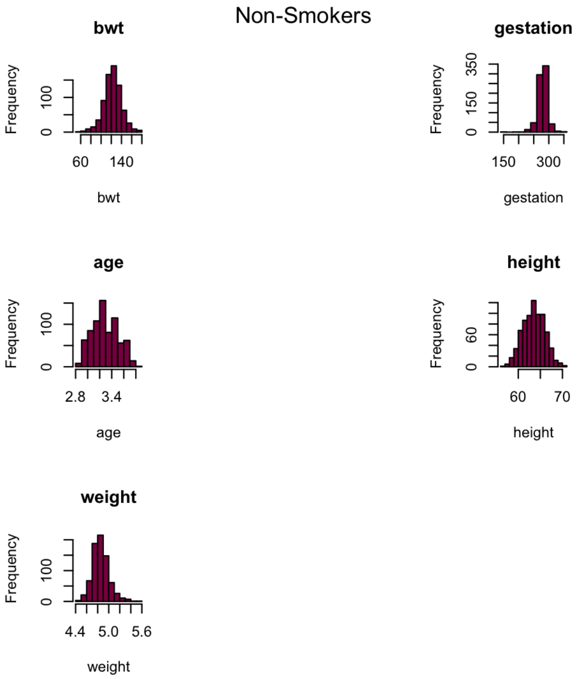
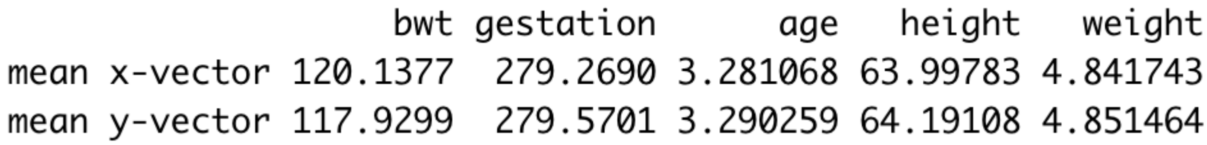
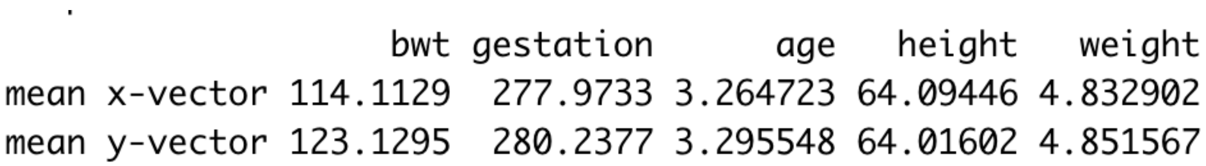


Figure 7: First versus Successive Pregnancies means using robust test



```
                     bwt gestation      age   height   weight
mean x-vector 120.1377  279.2690 3.281068 63.99783 4.841743
mean y-vector 117.9299  279.5701 3.290259 64.19108 4.851464
```

Figure 8: Smokers versus Non-smokers means using robust test



```
                     bwt gestation      age   height   weight
mean x-vector 114.1129  277.9733 3.264723 64.09446 4.832902
mean y-vector 123.1295  280.2377 3.295548 64.01602 4.851567
```