The Study of Effect of Smoking During Pregnancy on Baby Birth Weight

Chen, Mengting A14782590
Dai, Jixi A91084063
Xiang, Yixiao A92070053
Zhang, Haoran A92036029
Zhang, Su A91115828

Table of Content

0. List of Figures

1. Introduction

Background

Low birth weight is turning out to be one of the threats to infants' health, and because it will cause RDS, IVH, PDA, NEC, ROP[1], more and more doctors may pay attention to this issue. Smoking during pregnancy has been concerned as a factor that will result in low baby birth weight and fetal injury. This report summarizes a few measurable demonstration and examination related with the survey conducted by Child Health and Development Considers. The main purpose of this report is to examine the potential effects of smoking to baby weight from pregnant mothers. The information data set utilized in this investigation is the record "babies.txt", containing 1236 objects, each has 7 factors. We also use another record "babies23.txt", containing same observation in "babies.txt" but capturing more controlling factors.

Data

This study uses two data sets named baby.txt and baby23.txt. These two datasets have same observations and same birth weight data. The information in babies.txt is related to 7-feature vectors. Factors include birth weights, gestation period, mother's age, height, pre-pregnancy weight, whether the child is born first, and mother's smoking record. In babies23.txt, information is indicated with extra features and point by point in categorizations. The numbers "0", "1", "2", "3" and "9" represents smoker mothers, non-smoker mothers, mothers only smoked before pregnancy, mothers used to smoke and unknown smoking status mothers, respectively from "babies23.txt". For data collected by numbers "2", "3", and "9", we only used them to generate boxplots, while others exclude these data. Data collected for our study is an enl[1]arged portion of the mentioned CHDS data. The data is collected from all pregnancies recorded between 1960 and 1967 among women in Kaised Health Plan in Oakland, California. The research is comprised of 1236 male babies, with variables such as distinctive sexual orientation, twin status, and components which may affect weights of the newborn babies. Data accessible to us by means of information network incorporates mother's smoking status, which is a customary categorical perception, as well as baby's weight in ounces, a continuous, numerical variable that we considered to be related to mother's smoking status.

Purpose

The purpose of this report is to examine and analyze the effect of smoking during pregnancy on baby birth weight. We want to figure out if there is a difference between the birth weight of babies born to smokers and nonsmokers. We will further test the significance of the difference, if there is one, to the health and development of the baby.

Hypothesis:

We construct two hypothesis before our investigation and analysis.

[1] RDS: Respiratory distress syndrome; IVH:Bleeding in the brain; PDA: Patent ductus arteriosus; NEC: Necrotizing enterocolitis; ROP: Retinopathy of prematurity

1. There is no statistically significant difference between weight of babies born to mothers who smoked during pregnancy and who did not.
2. There is no association between smoking during pregnancy and low baby birth weight.

Theory

Methods that we apply in this study are listed in the following:

1. ***Mean and Median*****:** Mean represents the average of data and the center of certain data distribution graphically. The median is the observation right in the middle. If there are an even number of data collected, there will be two values in the middle, and the median is calculated as their average.
2. ***Variance and Standard Deviation*****:** Variance measures how far a set of data is spread out form their mean. Standard Deviation shows how far individual value may vary from the center of the distribution.
3. ***Histogram*****:** Histogram gives a clear picture of the data density. Higher bars represent where the data are recorded with higher frequency compared with others. Histogram is very straightforward for describing the shape of the data distribution. In this research, histograms are applied to form a contrast between pregnant mothers who are smokers and who are non-smokers.
4. ***Q-Q Plot*****:** A Q-Q plot is used to compare the shapes of distributions, providing a graphical observation of how features such as location, scale, and skewness vary in the two distributions.
5. ***Boxplot*****:** A boxplot summarizes a data set using five statistics including minimum, maximum, mean, 25th percentile, and 75th percentile but also plotting unusual observations such as outliers. The median of a boxplot splits the data into the top 50% and the bottom 50%. The total length of the box, shown vertically , is known as the interquartile range (IQR, for short), which, similar to the standard deviation, is a measure of variability in data. If there are more variabilities in a set of data, then the larger the standard deviation and IQR are. The lower bound and upper bound of the box are called the first quartile (25%) and the third quartile (75%), and are often labeled Q1 and Q3, respectively.
6. ***Normality*****:** In statistics, normality tests are applied to examine whether a set of data is a normal distribution or not and to how likely it is for randomly chosen variables.
7. ***Mann-Whitney U test:*** is a nonparametric test for hypothesis test that the randomly chosen values will be equally likely from one and other. This test, unlike the t-test, does not require the sample variables to be normal distributions.

Limitations

There exist some limitations to this study that constrain the generalization of our conclusion.

1. Voluntary questionnaire should be polled with care due to research biases such as non-responses or deceptious responses.
2. The sample size in this study is much more smaller than the overall population. Furthermore, this study does not account for births that were preterm.
3. Methods applied in this study, including histogram, boxplot, qq-plot do not control the effect of confounders.

Confounders

Confounders are possible items that can mask the effect of smoking mothers on the babies' weight at birth. In our first dataset babies.txt. We include mother's weight, mom age in years, mom pre-pregnancy weight in pounds and gestation days as confounders, and then add more confounders in our second dataset babies23.txt. Additional confounders that we add include infant sex, mom race, mom height in inches, mom pre-pregnancy weight in pounds, dad race, dad age, dad height, and dad weight.

2. Investigations

2.1 Numerical

Basic summary

The total sample size for this study population is 1236 male babies and 1236 mothers. As shown in the *appendix(item3 in page 16)*, when we separate the data into two categories by smokers and non-smokers, we can see an overall smaller birth weight for the babies of smokers. Smokers' babies birth weight has a larger ***min*** value than that of nonsmokers' babies birth weight. However, nonsmokers' babies birth weight has a larger ***max*** value, ***25% percentile***, ***75% percentile***, and ***median*** than those of smokers' babies birth weight. We will break our data into the detailed analysis by the following factor: **mean**, **standard deviation**, and **low birth weight rate**.

Dataset babies23.txt breaks the variable "smoke" into more detailed categories, such as "until preg", "once did not now", and "unknown", but we define these factors as neither smoker nor nonsmoker. In this research, we only treat "never"(means moms that never smoked before) as nonsmokers and "yes now"(means moms that smoke until now) as smokers. Mostly, babies23 and babies generate the same results or different results with a minor difference.

Mean

The **mean** value of babies birth weight for smokers is smaller than the **mean** value of babies birth weight for nonsmokers.

Standard deviation

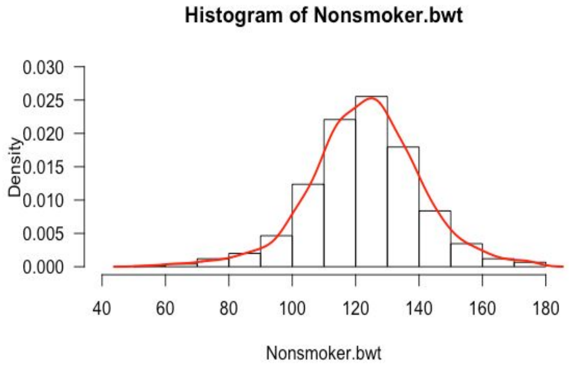
From the dataset babies23.txt, the **standard deviation** of smokers' babies birth weight is 18.09895, and the **standard deviation** of nonsmokers' babies birth weight is 17.10966. From the dataset babies.txt, the **standard deviation** of nonsmokers' babies birth weight is slightly larger but generates the same result as babies23.txt. From the data above, we can conclude the result that the **standard deviation** of smokers' babies birth weight is larger than the **standard deviation** of nonsmokers' babies birth weight.

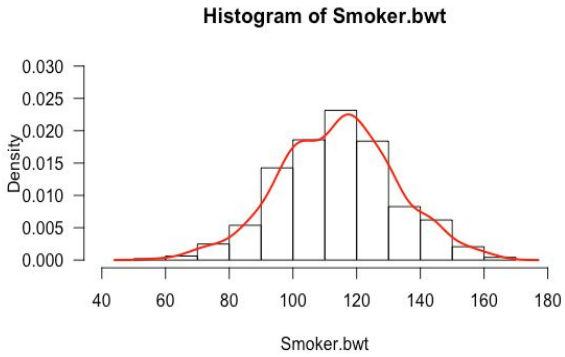
Low birth weight rate

Low birth weight is a term used to describe babies who are born weighing less than 88 ounces. **Low birth weight rate of smokers' babies** is equal to the number of smokers' babies with low birth weight divided by the number of smokers' babies. With same reasoning, **low birth weight rate of nonsmokers' babies** is equal to the number of nonsmokers' babies with low birth weight divided by the number of nonsmokers' babies. From the dataset babies23.txt, **low birth weight rate of smokers' babies** is 0.08264463, since we have 40 low birth weight babies out of 484 babies born to smokers. From the dataset babies23.txt, **low birth weight rate of nonsmokers' babies** is 0.03125, since we have 17 low birth weight babies out of 544 babies born to nonsmokers. From the data above, we can observe that smokers' babies may have a relatively larger chance of getting low birth weight, but we cannot be decide if smoking is the factor that influences the chance.

2.2 Histogram

To figure out the difference in weight of babies whose mothers smoked during pregnancy and the ones did not, we now adopt graphical methods. We first graph the histograms of baby weight using dataset from babies.txt. We separate our samples into a group of 484 mothers who are smokers and another group of 742 mother who are nonsmokers. Figure 2.2.1 below displays the distribution of weights of babies whose mother smoked during pregnancy and Figure 2.2.2 displays the distribution of weights of babies whose mother did not smoke. Because the numbers of samples in each group are different, we use density of baby weight to compare the difference. The y-axes represent density of the distribution and the x-axes represent weight of babies for each group. The baby birth weight is counted in ounces.





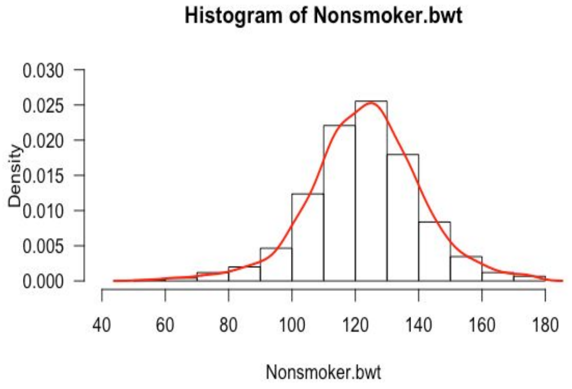
(babies.txt) Figure 2.2.1

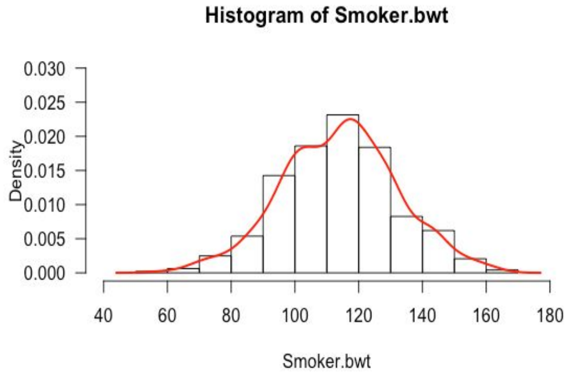
Figure 2.2.2

From the histograms, we see that the distribution of Nonsmoker baby birth weight (Figure 2.2.1) on the left is unimodal and slightly skewed to the left. Range of baby weight is approximately from 50 to 180 ounces. The median is between 120 and 130 ounces. The distribution of Smoker baby birth weight (Figure 2.2.2) on the right is approximately unimodal and symmetric. Range is approximately from 50 to 170 ounces. The median is between 110 and 120 ounces. From the density curve, we can see that two distributions are different in shape, the median of birth weight of smoker is less than that of nonsmoker. Baby birth weights that are less than 5.5 pounds (88 ounces) are considered low birth weight. The low birth weight frequency inferred from histogram of nonsmoker (figure 2.2.2) is approximately 3 percent while the low birth weight frequency of smoker is approximately 8 percent. In conclusion, percentage of low birth weight of smokers tends to be higher than that of nonsmokers.

However, histograms only give us a basic sense of difference in the distribution. To make our testing more reliable, we will check normality for both distributions by Q-Q plot and Shapiro-Wilk test of normality later. As we cannot infer the exact frequency of low-baby-birth weight from the histogram, the comparison of frequency is covered in the section of numerical analysis.

To ensure the accuracy and consistency, we now switch to the dataset babies23.txt. Both dataset use same observations but babies23.txt includes more control variables, including race of mom, weight of dad and age of dad. From the graph below, we can see that dataset babies23.txt and baby.txt display the same distributions for birth weight of babies whose mom smoked during pregnancy and whose mom did not. Same information can be inferred from figure 2.2.3 and 2.2.4.





(babies23.txt) Figure 2.2.3

Figure 2.2.4

2.3 Check for normality

We check if birth weight of babies born to smokers and birth weight of babies born to non-smokers both follow normal distribution.

Figure 2.3.1 on the left is the normal qqplot for birth weight of babies born to smokers. On the right is the normal qqplot for birth weight of babies born to nonsmokers. We can see that there's no systematic departures from a straight line on the left but there's upward curve on the left tail and downward curve on the right tail in qqplot on the right.

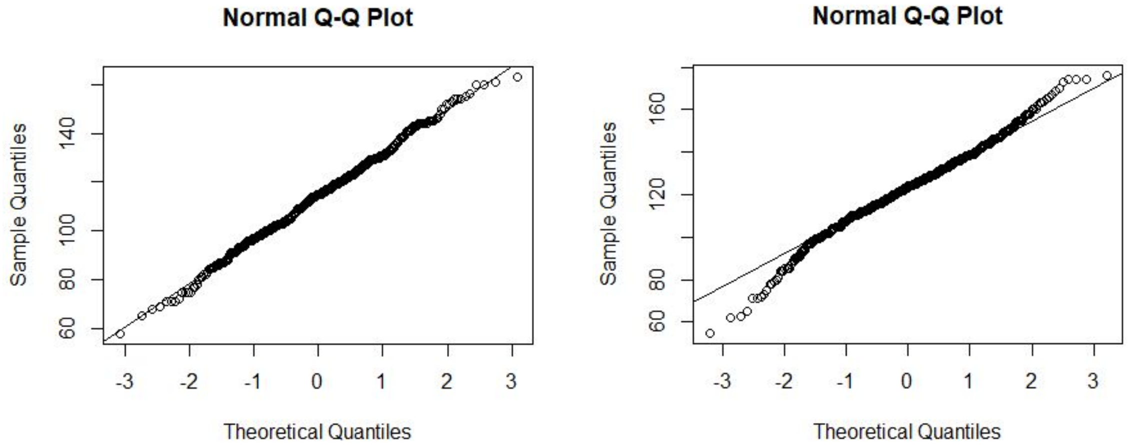


Figure 2.3.1

We check both of the distributions' normality first by simulation. First, we calculate that kurtosis of both birth weight data, and we get a kurtosis of 2.98 for birth weight of babies born to smokers and a kurtosis of 4.04 for birth weight of babies born to non smokers. Then to check normality for smokers' data, we will generate 484 pseudo-random observations from a normal distribution and calculate the kurtosis coefficient, and we will generate 742 pseudo-random observations from a normal distribution and calculate the kurtosis coefficient to check normality for nonsmokers' data.

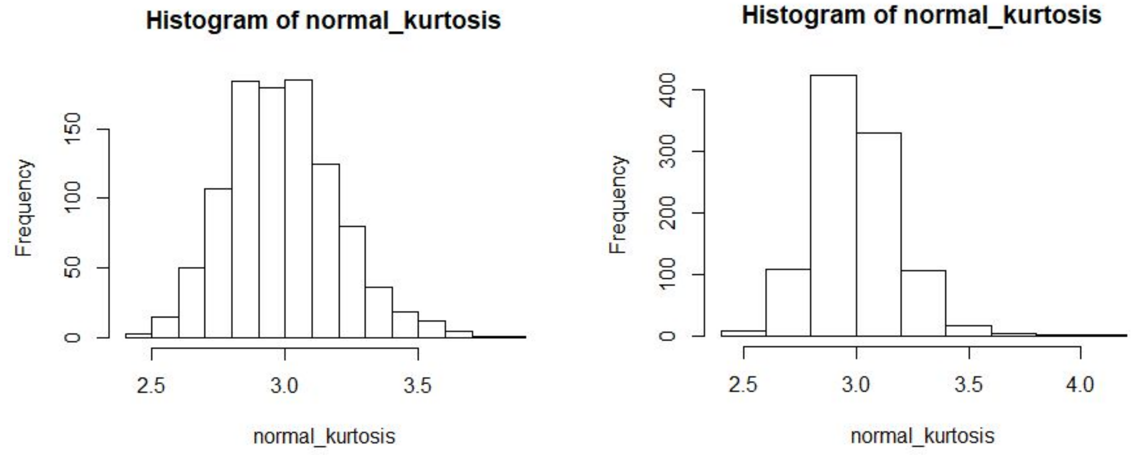


Figure 2.3.2

From the left part of figure 2.3.2, we can see that 2.98 is a typical kurtosis value of f the normal distribution with 484 samples, and from the right part of figure 2.3.2, we can see that 4.04 is on the right tail of the histogram. Therefore, distribution of birth weight of babies born to smokers does not departure from normality while distribution of birth weight of babies born to nonsmokers does not follow a normal distribution.

We further check their normality by using Shapiro-Wilk test of normality. And we get p-value of 0.5542, which is greater than 0.1, for birth weight of babies born to smokers so we can safely assume normality and p-value of 1.017e-05, which is smaller than 0.1, for birth weight of babies born to nonsmokers so we can't assume normality.

We switch to babies23.txt, a dataset with more variable. From normal qqplot, we get basically the same result from the previous dataset.

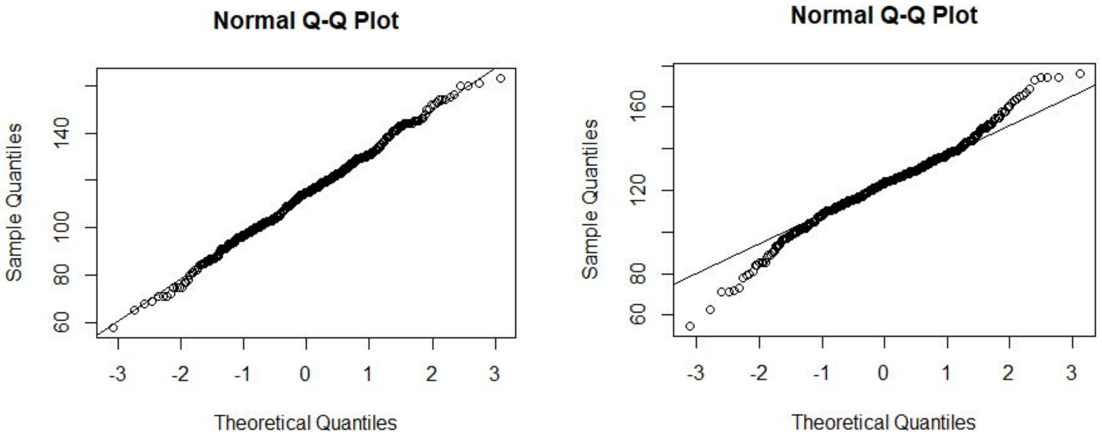


Figure 2.3.3

Still, we will perform Shapiro-Wilk test of normality. And we get p-value of 0.5542 for birth weight of babies born to smokers and 3.438e-06 for birth weight of babies born to nonsmokers. So we have the same result from the previous dataset: we can't assume normality.

2.4 Boxplot

We assign "0" and "1" to variables represents mothers who are non-smoker and smoker, respectively in data set *baby.txt*. Observations with unknown smoking status have been assigned to number "9". As our study focuses on the effect of smoking during pregnancy, we eliminate those observations with unknown smoking status by additional inspection and cleaning.

The boxplot on the left side of figure 2.4.1 is the boxplot of "nonsmoker". The line in the box roughly roughly splits the box equally, indicating that the distribution of the data set is symmetric. The median of birth weight of nonsmoker is roughly above 120 ounces. By the position of the box, it shows that 75% of birth weight of nonsmoker group are over 110 ounces. There exist some suspected outliers which are over 165 ounces. In another word, the whisker of babies' weight is from 85 ounces to 110 ounces.

Compared with the shape and distribution in boxplot of 'nonsmoker', the shape and distribution of "smoker" is similar to these of "nonsmoker". The range of 'smokers' babies weight is from 70 ounces to 160 ounces and there are less outliers. In this page, it looks like babies birth weight is alike between 'smoker' and 'nonsmoker'.

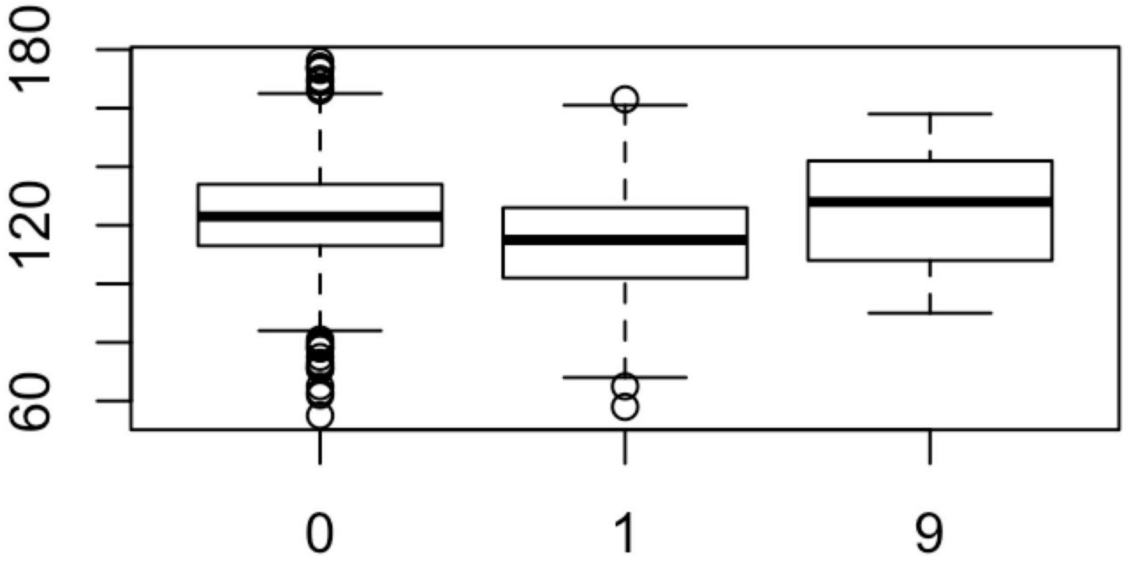


Figure 2.4.1

To make sure the accuracy of the result we get from the data set *baby.txt*, we graph another boxplot based on the dataset, *babies23.txt* and get Figure 2.4.2. The figure indicates that most of boxplots are ranged similar, which means most of their babies birth weight are ranged in a similar way. For an example, when we looked at boxplot of "mom smoked until pregnancy", its range, median and whisker all is like the boxplot of "mom never smoke". Furthermore, there is more extreme baby birth weight within mom never smoke. So we can conclude that mom's smoke situation does not directly affect baby birth weight, there may be other reasons that cause baby birth weight light.

In this database, there are three different types of 'smoker'. '1' represents that mom keeps smoking; '2' represents that mom only smoked before pregnancy; '3' represents that mom used to smoke. In the same case, '9' is unknown as irregular values, which we can calculus as inputs by mistakes. In figure 2.4.2, that mom who ever smoked before pregnancy or not, but no smoke currently, we can find out that those two figures are quite similar. Both median and the box distribution are ranged from 110 to 130 ounce. If we compare those figures with the first figure which represent nonsmoker mom, there are more outliers under nonsmoker figure. Also, all of the medians are among 125 ounces.

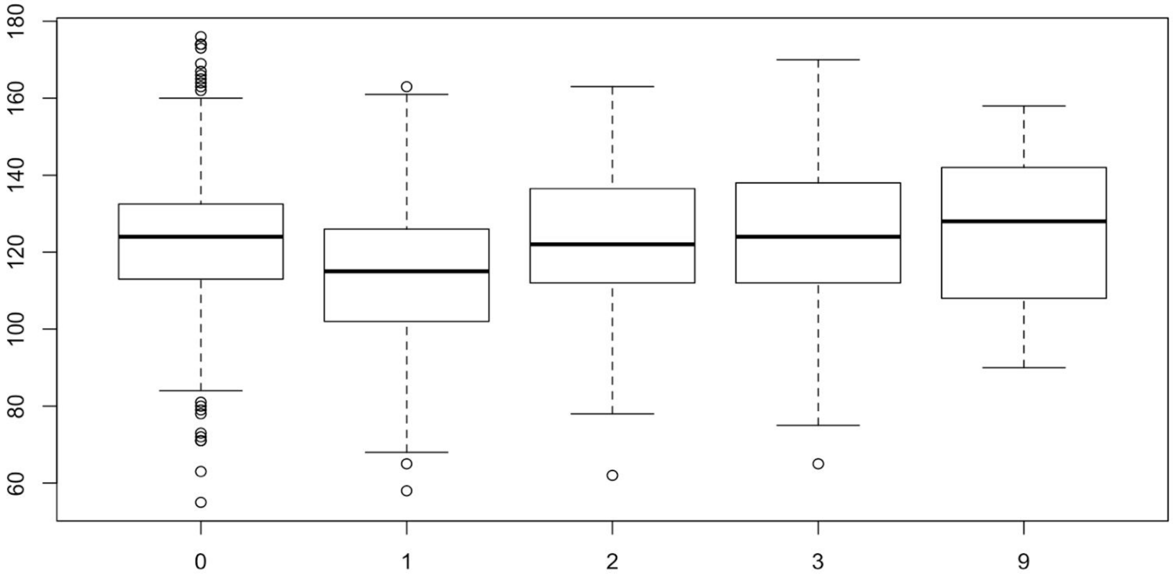


Figure 2.4.2

2.5 Check for significance

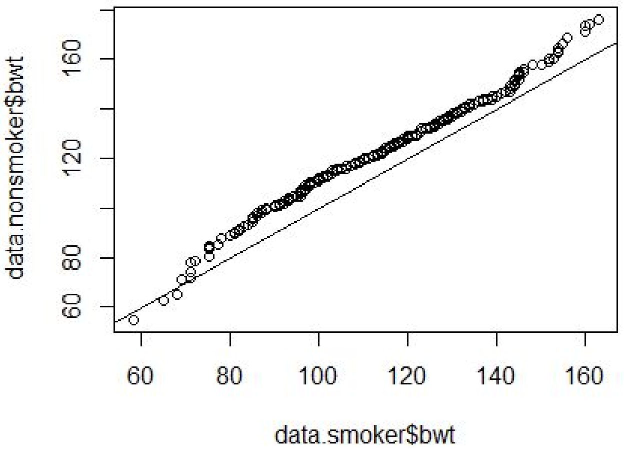


Figure 2.5.1

Figure 2.5.1 is the qq plot between birth weight of babies born to smokers and birth weight of babies born to non smokers, and the reference line is y = x, which has a slope of 1 and intercept of 0. We can see from the graph that the plotted points deviate from the reference line and although the points are roughly linear, the intercept is different from 0.

Again, we check if the qq plot changes when we change the dataset to babies23.txt. Below is the qq plot for birth weight data in babies23.txt. Still plotted points are roughly linear but the intercept is different from 0.

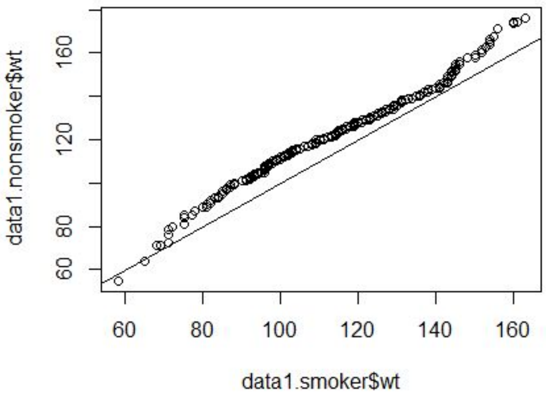


Figure 2.5.2

In this case, where normality of two distribution cannot be assumed, we will use non-parametric test to test if there is a significant difference between birth weight of babies born to smokers and birth weight of babies born to nonsmokers.

We use Mann-Whitney test to test if there is a statistically significant difference between two birth weight data. The null hypothesis is that the mean of birth weight of babies born to smokers is the same as the mean of birth weight of babies born to nonsmokers, and the alternative is the mean of birth weight of babies born to nonsmokers is greater than the mean of birth weight of babies born to smokers.

After we perform the one-sided test, we get a p-value of 2.2e-16, which is smaller than 0.05 significance level, so we reject the null hypothesis. Therefore, from hypothesis testing, we know that the average of birth weight of babies born to nonsmokers is statistically greater than the average of birth weight of babies born to smokers.

3. Discussion

The main purpose of this report is to examine the potential effects of smoking on baby birth weight. From above investigation, the result indicates that there is no significant difference between the birth weight of babies born to smokers and nonsmokers. However, there are certain limitations of this study include small sample size compared to the overall population and no control on effect of confounders. Future studies on the effect of smoking during pregnancy and other potential factors may bring more accurate information for people who concerned about fetus health.

Objectives

As our investigation regarding effect of smoking during pregnancy on baby birth weight has certain limitations, we start to explore another possible factor that may lead to low baby birth weight. In the following study, we will examine the association between maternal age and low baby birth weight.

Reference

The following study refers to the published thesis *The Association between Maternal Age and Low Birth Weight Offspring, NHANES 2007-2008* by Dianna Johnson at Georgia State University. The study of Dianna Johnson aims to examine the association between young age of mothers and low baby birth weight. The null hypothesis in this study states that "there is no positive association between young motherhood and low birth weight offspring." (Dianna Johnson, 13). Studied variables in this study are ethnicity, education, annual family income, smoking status, young motherhood and low birth weight. Methods involved are univariate logistic regression analysis and multivariate logistic regression analysis. Conclusion of this study is that "the results of this study did not indicate young maternal age increases odd of low birth weight" (Dianna Johnson, 43).

Hypothesis

Based on the research result of Dianna Johnson's study, the following study constructs the same hypothesis: There is no positive association between maternal age and low birth weight.

Methods

We use same datasets baby.txt and baby23.txt to examine the effect of maternal age on probability of low baby birth weight. We are using univariate linear regression to estimate the association between paternal age and low birthweight. After we clean the data by excluding data that has age of 99, we divide data into six groups by age: 15-20 years old, 21-26 years old, 27-32 years old, 33-38 years old, 39- 44 years old and older than 45 years old. Then we generate an array of proportion of number of babies with low birth weight in all babies born to each age group. We will find association between maternal age and low birth weight by running linear regression on the array we generate and age groups, which are represented by 1,2,3,4,5,6 from young ages to old ages.

Results:

After we run the regression using the threshold that below 88 ounces are low birth weight, we get a negative coefficient of -0.07 between age and incidence of low birth weight. And p-value of 0.2829,which is greater than 0.05, shows that the relationship is not significant as we can't reject the null hypothesis that coefficient is equal to zero. Then we change the threshold to 80, and we get the following result: coefficient of -0.0005 and p-value of 0.90. More detailed summary can be found in *Appendix(page 16)*. So after doing the regression, we find that there's no significant relationship between maternal age and probability of low birth weight. The graph below showing the relationship between maternal age and incidence of low birth weight more clearly shows that a strict negative relationship does not stand in our dataset.

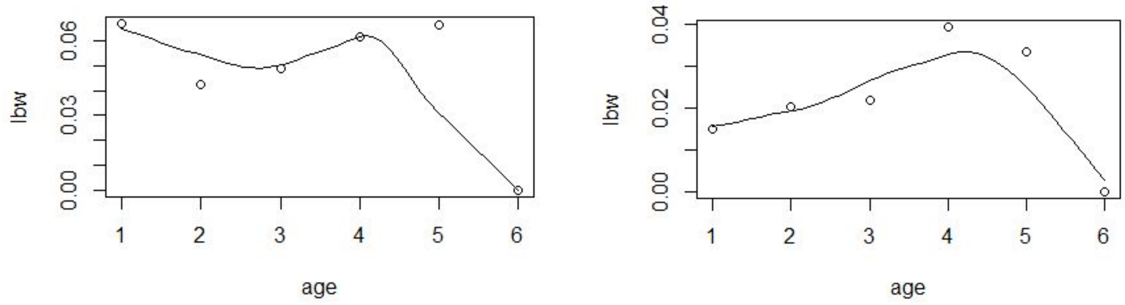


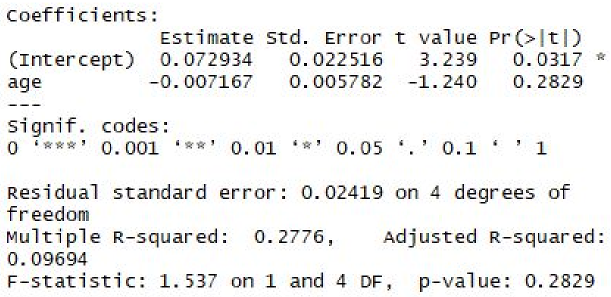
Figure 3.1

4. Conclusion

The results of this study does not indicate that smoking during pregnancy will increase the possibility of low birth weight. But there is a statistically significant difference between the average birth weight of babies born to nonsmokers and average birth weight of babies born to smokers. On average, babies born to mothers who smoked during pregnancy weight less than babies who born to mothers who did not smoked. We also bring up a new hypothesis regarding the effect of maternal age on low baby birth weight. The new study indicates that there's no significant association between maternal age and low birth weight. This result is consistent with the conclusion of *The Association between Maternal Age and Low Birth Weight Offspring, NHANES 2007-2008 by* Dianna Johnson.

5. Appendix

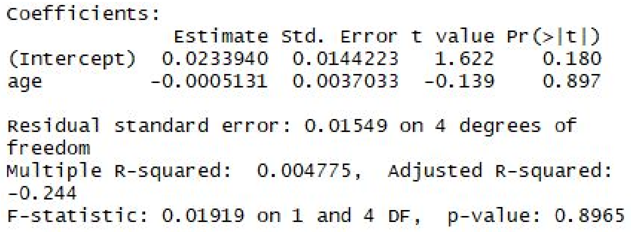
1)Below is the specific summary of our first regression.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.072934   0.022516   3.239   0.0317 *
age         -0.007167   0.005782  -1.240   0.2829
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02419 on 4 degrees of
freedom
Multiple R-squared:  0.2776,    Adjusted R-squared:
0.09694
F-statistic: 1.537 on 1 and 4 DF,  p-value: 0.2829
```

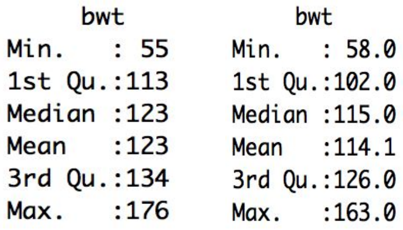
2)Below is the specific summary of our second regression.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0233940  0.0144223   1.622    0.180
age         -0.0005131  0.0037033  -0.139    0.897

Residual standard error: 0.01549 on 4 degrees of
freedom
Multiple R-squared:  0.004775,  Adjusted R-squared:
-0.244
F-statistic: 0.01919 on 1 and 4 DF,  p-value: 0.8965
```

3)Below is the summary of babies birth weight of nonsmokers and smokers



```
      bwt             bwt
Min.   : 55     Min.   : 58.0
1st Qu.:113     1st Qu.:102.0
Median :123     Median :115.0
Mean   :123     Mean   :114.1
3rd Qu.:134     3rd Qu.:126.0
Max.   :176     Max.   :163.0
```

(nonsmokers’ bwt) (smokers’ bwt)

6. Reference

1. Dolan, Siobhan. “Low Birthweight.” *March of Dimes*, Oct. 2014, www.marchofdimes.org/complications/low-birthweight.aspx.
2. Johnson, Dianna. “The Association between Maternal Age and Low Birth Weight Offspring, NHANES 2007-2008.” *Thesis / Dissertation ETD*, 2014.