MATH 189 Project 1

Group 29

January 2021

1 Introduction

Babies need a long time to become matured in utero before they finally meet the world, which means they may be influenced by many variables such as the living habits of their mothers. It is now increasingly believed that smoking by pregnant women may result in a negative influence on babies' birth weight. More specifically, it may be responsible for 150g of reduction of birth weight. This is an impending issue such that it is believed to impact babies' health condition. Hence, people want to figure out the specific difference in weight between babies born to mothers who smoked during pregnancy and those did not, and whether the difference should be high-valued.

Noticing these two questions are relatively complex, we separate them into several specific questions which can be classified into two parts. The first part responds to the first question. It includes summarizing distributions of baby birth weight numerically by using the given datasets, comparing the distributions graphically through visualizations such as histograms and boxplots, and comparing frequency of low-birth-weight babies to assess the reliability of estimates. Then, the second part responding to the second question is mainly consist of two sections: a series of tests to find confounders and assess statistical significance of the differences we found in three questions in the first part. Meanwhile, we analyze and respond to two additional questions to narrow down the argument and point out hints for further research. Last but not least, we briefly introduce the statistical methods used in this project and summarize all sections mentioned above.

2 Data

The dataset we use is *babies23.txt* provided by professor Jelena Bradic. The set contains 23 columns and 1236 observations. For these 23 variables, the discrete numerical variables include id, date, gestation, wt (birth weight), parity, age, ht (mother's height), wt (pregnancy weight), dht (father's height), dwt (father's weight), dage (father's age), the regular categorical variables include pluralty, outcome, sex, race, drace (father's race), marital, smoke, and the ordinal categorical variables include ed (mother's education), ded (father's education), inc (family income), time, number.

Since the pluralty, outcome, and sex has only one value for their observations, we omit these three variables from the dataset. For the rest of the variables, we mainly use smoke and wt to answer our questions, and use the remaining variables to figure out their influence on babies' birth weight.

The categorical variable smoke has five smoking status: never smoke, smoke now, smoke until current pregnancy, and smoke once but already quit. In the following report, the "never smoke" will be categorized as non-smokers, and all of the other smoking status will be put into the smoker's category.

3 Background

The research conducted by Butler and Goldstein concluded that "in a British population cigarette smoking during pregnancy increased the late fetal plus neonatal mortality 28 percent and reduced birth weight by 170 g, and these differences persist even after allowing for a number of 'mediating' maternal and social variables" (Butler et al., 1972, p. 127). These "mediating" maternal and social variables refer to social class, maternal age, parity, and maternal height. In addition, allowance is made for sex, gestational maturity in completed weeks, and perinatal mortality. The study concluded that the result is similar to the analysis made with and without theses variables. In addition, the research shows that the amount of smoking before pregnancy does not affect fetal survival considering the amount of smoke of mother after the fourth month of pregnancy (Butler et al., 1972, p. 128).

According to the study conducted by Vangen compared birth weights and perinatal survival of ethnic groups in Norway, "birthweight differences between these groups were not clearly related to perinatal mortality. Although mean birthweight varied by as much as 350 g between the groups, it could not explain ethnic group differences in perinatal mortality" (Vangen et al., 2002, p. 656). In addition, this study also suggests that "birthweight differences do not contribute to the substantial differences in perinatal mortality by ethnicity in Norway" (Vangen et al., 2002, p. 659).

4 Investigation

4.1 Numerical Analysis

To get a general sense of the babies' birth weight, we need to gather some numerical information regarding the distribution of babies' birth weight for both that smoker and non-smoker group.

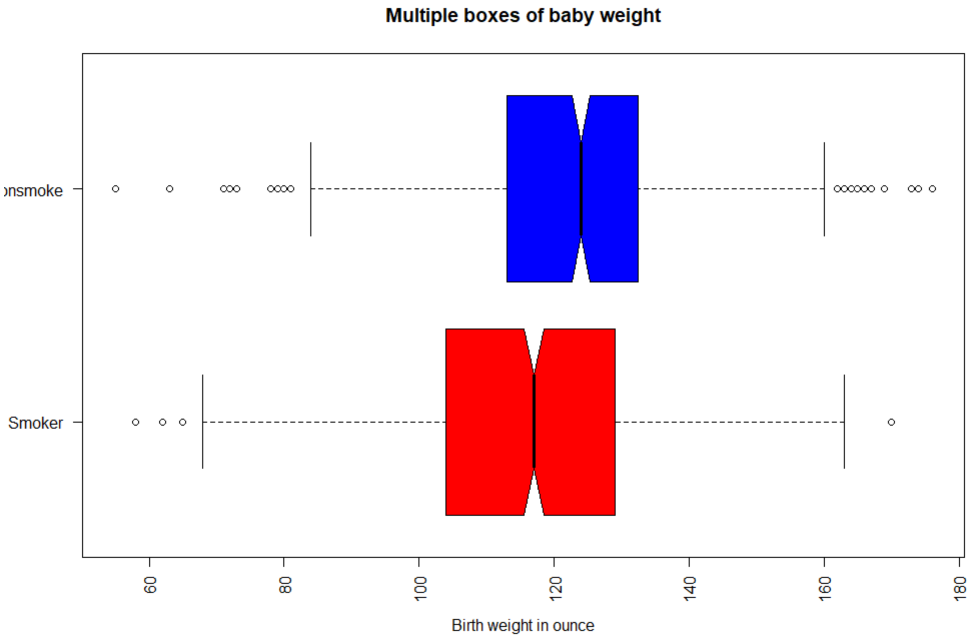
For the data set of mothers who smoke, it has a mean weight of 115.92 ounces, a lower quartile of 104 ounces, and a standard deviation of 18.64. For babies of non-smoker, it has a mean weight of 122.78 ounces, a lower quartile of 113 ounces, and a standard deviation of 17.11. Therefore, the data for smokers is centered around a higher value than that of non-smokers. Also, since the variance of babies' weight for smokers is greater than that of the non-smoker, smokers' data has a greater variation from the center. From this simple glimpse of the data, we may grab the idea that in general non-smokers' babies have greater weight, but it still needs further discussion.

4.2 Graphical Analysis

In the graphical analysis section, we collect, visualize and carefully analyze the data by forming and comparing distributions of babies' weight whose mothers either smoke or not. We can obtain not only the general views of the distributions but also an obvious visual contrast between those representative values of variables.

4.2.1 Boxplot

Before performing any rigorous test on data difference, we want to obtain a general view of how a mother with a smoking habit may affect her baby's birth weight. We first divide the data of baby birth weight into two groups based on whether their mother smoke or not. Then we draw a boxplot for each group. From the graph, we can conclude that baby birth weight would likely be higher if the baby's mother does not smoke. Also, it is surprising to observe that the upper fence of non-smoker's baby weight is lower than that of the smoker's.

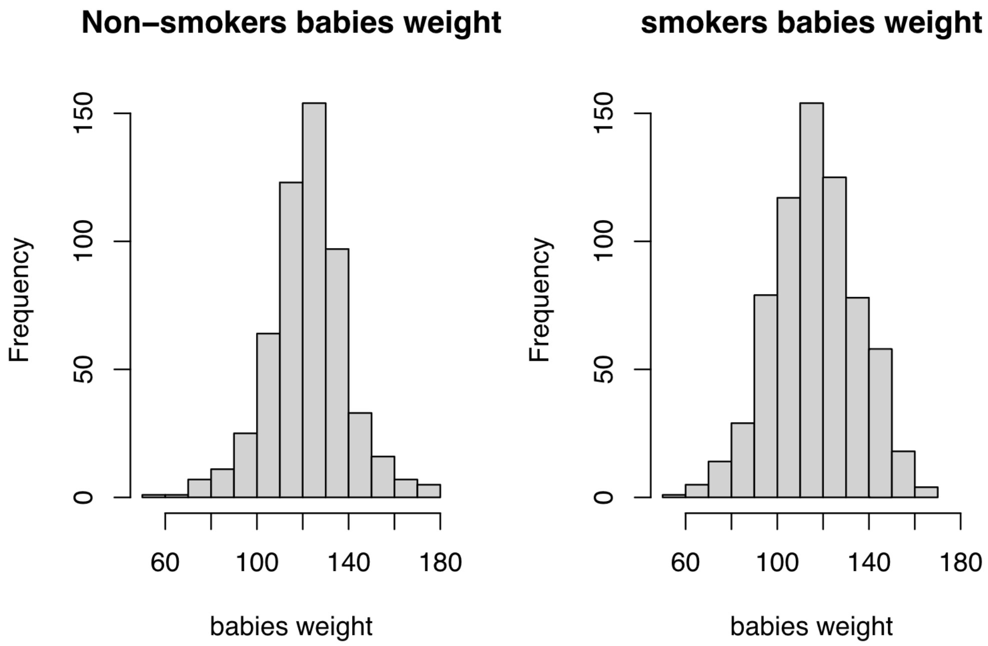


Caption: In the graph above, we observe that lower and upper quantiles and lower fence of non-smoker's babies weight are higher than those of the smoker's babies weight. In addition, since the notches of boxplots don't overlap at all, there is a strong evidence that nonsmoking mother's babies weight median is higher than those born by smoking mothers. It is also important to notice that each boxplot has a number of outliers. The non-smoking one has more lying outside the fences.

4.2.2 Histogram

In order to obtain a deeper understanding of the distributions, we draw a histogram for each of the groups and compare them. Conclusively, the graph implies that non-smoking mothers are more likely to bear higher birth weight babies comparing to smoking mothers.

fig2: We can see in the graph that both distributions of frequencies are unimodal and almost symmetric. The mode of babies' birth weight when mothers smoke is around within the bin of 110-120 ounces. The frequency is around 150 times.On the other hand, the mode of babies' birth weight when mothers don't smoke is within the bin of 120-130 ounces. The frequency is also around 150 times. Each corresponding median falls into the "mode" bin.



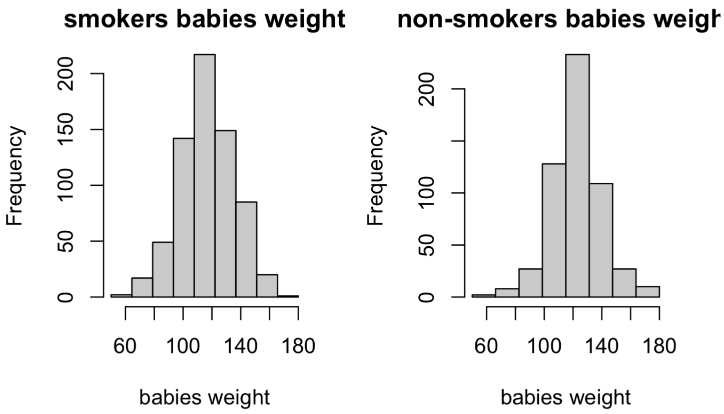
4.3 Analysis of Incidence

In this section, we will first estimate the incidence via histogram then calculate the actual value use the data.

4.3.1 Estimate incidence with Histogram

Babies who are weighted lower than 5 pounds and 8 ounces (88 ounces) are classified as low birth weight. We first try to estimate the incidence of the low birth weight baby by histogram.

To get the best selection of bins, we use the formula $bin = n^{\frac{1}{3}}$. Since the smoker has 682 sets of data, we plan to divide it into 9 bins; non-smoker has 544 sets of data, so it is divided into 8 bins.To grab a general sense on how will this plot look like, please refer to the histogram in section 4.2.2.



To estimate the incidence of the low birth weight, we have the following formula: incidence = $\sum$ heights to the left of 88

With this formula, we first calculate the incidence when the boundary for low birth weight is 88 ounces. For smoker, the incidence of low birth weight is estimated to be 52.9, and for non-smoker is 19.1.(for more detailed process of the calculation, please refer to the code)

Since the data set is relative small and the optimal number of bins we calculated is also small, we can't ensure the incidence we calculated is accurate. Next section will use the actual data of incidence to do a more accurate analysis.

4.3.2 incidence of Low-Birth-Weight Babies from data

If we set the boundary of low birth weight to 88 ounces according to the definition, then for smokers, there will be 42 out of 682 incidences of the low birth weight, which has a probability of 0.06158358. For non-smokers, there will be 16 out of 544 incidences of the low birth weight, which has a probability of 0.02941176.

$$0.06158358 * 0.02941176 = 2.09384206 \tag{1}$$

After comparing the results, we notice that low birth weight frequency of non-smokers' babies is less than half of the frequency smokers'. That means a mother who smokes will be more than twice likely to give birth to a low birth weight baby.

4.3.3 Reliability of the Estimates

To check the reliability of the result above, we change the definition of low birth weight by 1%, both increase and decrease, and observe their influences on low birth weight frequency. We can then check the reliability of the estimates by checking if the new frequencies are in proportion to the original one.

If we increase the original weight boundary of the low-birth-weight by 1%, the new boundary we have is

$$88 * 1.01 = 88.88 ounces \tag{2}$$

Then for smokers, there will be 46 out of 682 incidence of the low birth weight, which has a probability of 0.06744868. When comparing to the original frequency, the new frequency we get is 9.5% larger than the original one as calculated below.

$$0.06744868/0.06158358 - 1 = 0.09523805 \tag{3}$$

For non-smokers, there are 17 out of 544 incidences of the low birth weight, which has a probability of 0.03125. When comparing to the original frequency, the new frequency we get is 6.25% larger than the original one as calculated below.

$$0.03125/0.02941176 - 1 = 0.06250017 \tag{4}$$

If we decrease the reliability by 1%, the new boundary we have is

$$88 * 0.99 = 87.12 ounces \tag{5}$$

Then for smokers, there will be 42 out of 682 incidences of the low birth weight, and the new probability is 0.06158358, which is exactly the same as the original one. For non-smokers, there will be 16 out of

544 incidence of the low birth weight, and the new probability is 0.02941176. which is also exactly the same as original one.

Then if we increase the low-birth-weight boundary in weight by 1 percent, there is a considerable frequency increase for both non-smokers' and smokers' babies. This is worth noticing so that we compare the new probabilities of the frequency. For smokers, it is 0.06744868. For non-smokers. it is 0.03125. We can observe that the proportion increases from 2.09384206 to 2.15835776.

$$0.06744868/0.03125 = 2.15835776 \tag{6}$$

That means by increasing the boundary, low birth weight rate increases faster among smoking mothers. Conversely, that implies exactly that non-smoking mothers has a smaller chance to bear low birth weight child. Adding the analysis result that decreasing the boundary would not affect the original result, it is reliable to say non-smokers' babies are less likely to have low birth weight babies.

4.4 Importance of Difference

4.4.1 Two Sample t-test

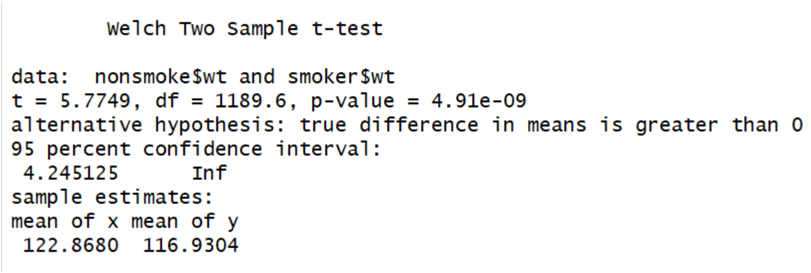
According to the description of data, the babies are assigned into two groups based on whether their mother smoke or not. To see if there is any significant statistical evidence that non-smoking mother's baby birth weight is greater, we do the Welch two-sample t-test to test the mean of one group is equal to the mean of other group. To do the Welch two-sample t-test, the population of mothers who don't smoke are denoted by x, and the population of mothers who smoke are denoted by y.

We set the hypothesis that

$$H0 : mean(x) = mean(y) \tag{7}$$

$$H1 : mean(x) > mean(y). \tag{8}$$

And the result is showed below.



```
        Welch Two Sample t-test

data:  nonsmoke$wt and smoker$wt
t = 5.7749, df = 1189.6, p-value = 4.91e-09
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.245125      Inf
sample estimates:
mean of x mean of y
 122.8680  116.9304
```

As the results shows that the p-value of t-test is 4.91e-09, which is smaller than 0.00001. At a given significance level alpha = 0.05, we reject the null hypothesis and conclude that there is strong statistical evidence that non-smoking mothers tend to bear greater birth weight babies. This finding statistically illustrated that our previous analytical results are correct. It makes our argument more rigorous and thus important. Since health condition is positively related to baby birth weight, we can say our finding is important to assess the argument's importance to baby's health condition.

4.4.2 Multivariate Regression

To decide if other variables will influence the birth weight, we use the multivariate linear regression to decide which independent variables are statistically significant to the babies' birth weight. We filter the sex, plurality, outcome since they only have one value, as well as id and date since they are irrelevant to the birth weight. The result is showed below.

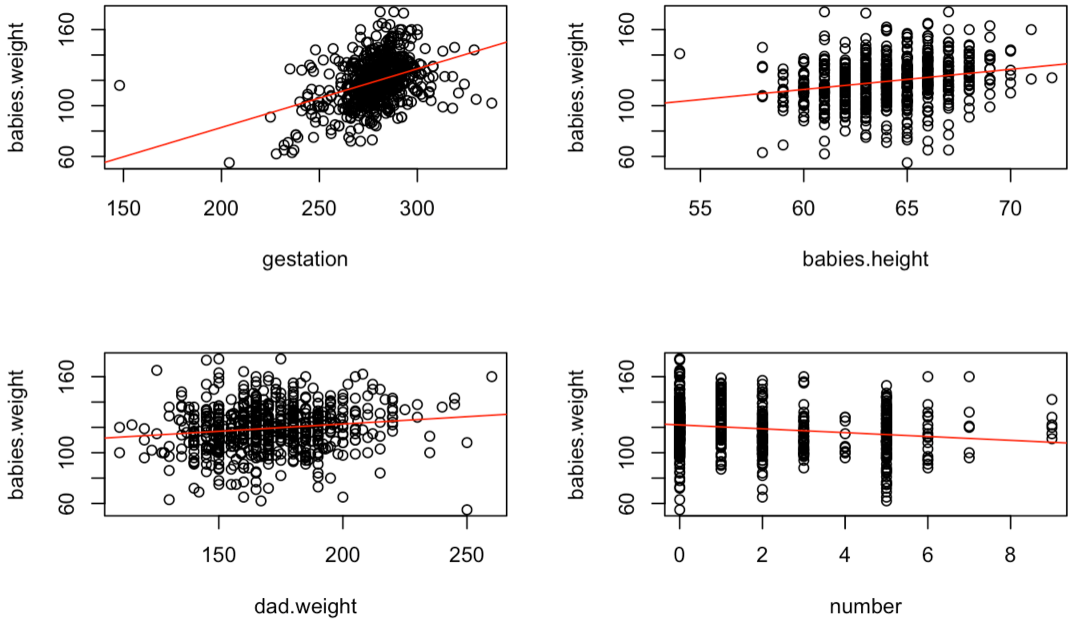
```
Call:
lm(formula = wt ~ ., data = babies.reg)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.336 -10.022  -0.717   9.627  48.679 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -87.17449   25.23872  -3.454 0.000591 ***
gestation     0.44881    0.04148  10.819  < 2e-16 ***
parity        0.68154    0.43292   1.574 0.115946    
race         -0.40033    0.28367  -1.411 0.158697    
age          -0.05679    0.21488  -0.264 0.791656    
ed            0.68512    0.59607   1.149 0.250853    
ht            1.13769    0.29901   3.805 0.000157 ***
wt.1          0.02984    0.03726   0.801 0.423505    
drace        -0.33618    0.28749  -1.169 0.242722    
dage          0.09208    0.17749   0.519 0.604101    
ded          -0.75643    0.52219  -1.449 0.147982    
dht          -0.03545    0.28337  -0.125 0.900477    
dwt           0.07545    0.03462   2.180 0.029667 *  
marital      -2.63065    3.05466  -0.861 0.389477    
inc          -0.23434    0.30905  -0.758 0.448581    
smoke         2.45420    1.67813   1.462 0.144140    
time         -0.19870    0.90444  -0.220 0.826188    
number       -2.20605    0.37409  -5.897 6.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.84 on 599 degrees of freedom
Multiple R-squared:  0.2765,	Adjusted R-squared:  0.256 
F-statistic: 13.47 on 17 and 599 DF,  p-value: < 2.2e-16
```

As we can see, within the 0.05 significance level, only gestation, babies’ heights, father’s weight, and number of cigarettes for smokers are showed to be significant toward the babies’ weight. We plot the linear regression for these four variables separately as showed below.



Caption: As gestation increases by 1, the babies’ weight will increase by 0.44881.In this graph, there is a strong and positive linear relationship. We also observe that babies’ height, dad’s weight also have positive linear relationship with babies’ weight, though not that strong. The number of cigarette smoked per day has a negative linear relationship as the correlation is relatively weaker.

According to both p-value and estimates, we observe that the gestation is the most significant one toward the babies’ weight. Also, By the graph and the estimates, we observe babies’ heights have strong relationship with the babies' weights, as they are almost increases in 1:1 portion. Since it is possible that a mother's habit may affect the above variables, we will list them as possible candidates of confounders. In addition, although dad's weight seems not have a very strong relationship with the heights because of the low estimates and the graph, the p-value tells us it is still a significant explanatory variable. Moreover, the numbers of cigarettes for the smokers also show a strong relationship with babies weight. As the mothers smoke more cigarettes per day, the babies birth weight will be much more smaller. In this section, we narrow down our argument by observing four possible distracting variables. This makes the argument more reliable and important to assess babies' health condition.

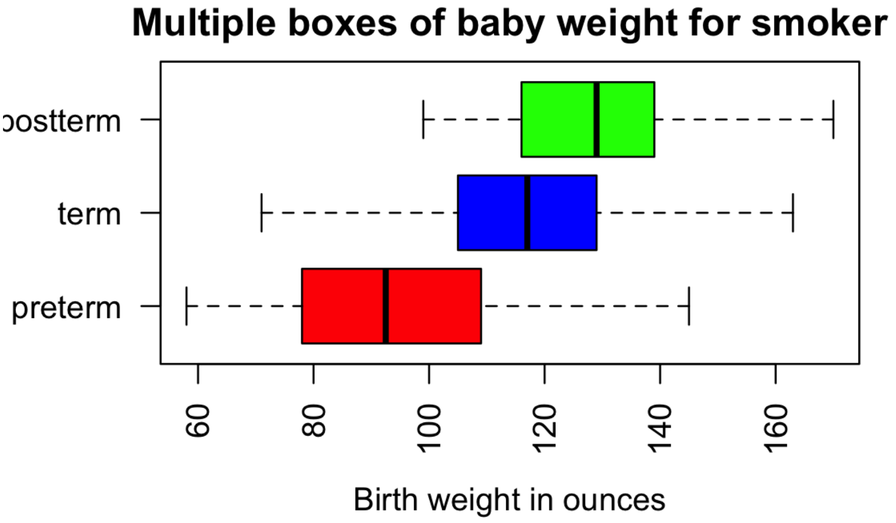
4.5 Additional Questions

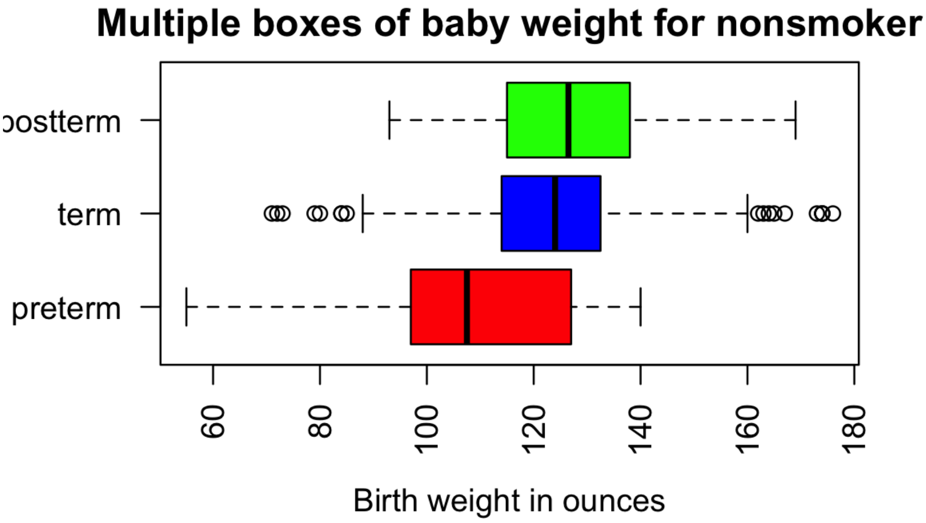
We know that a lower birth weight may indicate a relatively worse health condition for a baby. In the previous section, we have numerically and analytically showed that low birth weight somewhat associates with smoking. In this section, we will explore other sub-questions to strengthen our argument.

4.5.1 Gestation ⇒ Weight

According to the regression result from the previous section, the gestation is a vital factor that influences babies' birth weight. Therefore, we may address the question that how will the length of gestation influence the babies' birth weight excluding the effect of the smoke.

To exclude the effect of the smoke, we divide the data into two subsets: smoker and non-smoker, and plot box plot for each other set. We also divide gestation into three categories: pre-term, term, post-term. We define the pre-term to be 37 weeks (259 days), term to be time between 37 weeks and 42 weeks (259–293 days), and post-term to be time greater than 42 weeks (> 293 days).



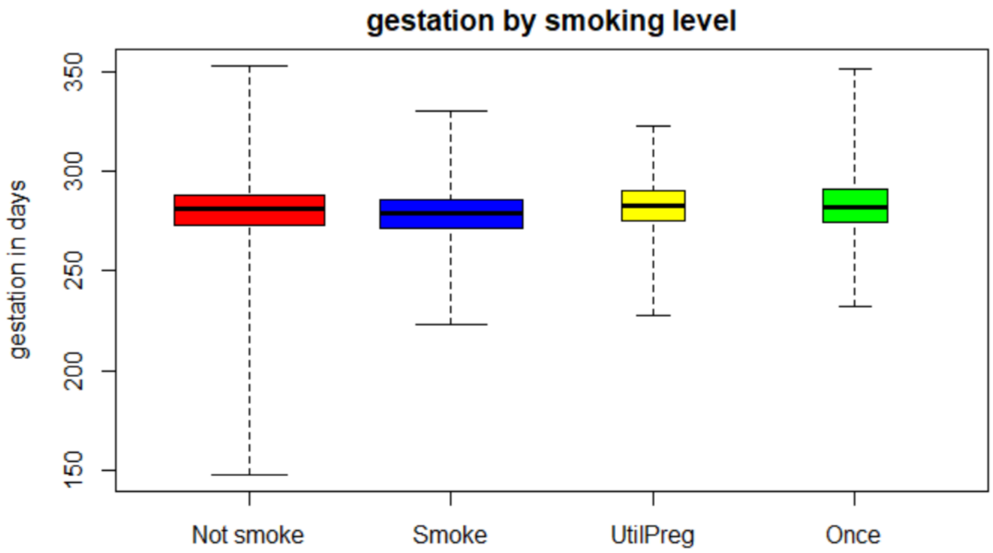


Caption: The first box plot applies the data for smoker, and the second box plot applies for non-smoker. In the first box plot, there is a clear distinction between different gestation length. The longer the gestation, the greater the median of the birth weight. This same pattern also apply for the non-smoker box plot, but the difference in median are relative smaller.

From this comparison, we believe gestation is also an important factor on babies' weight, and the longer the gestation, the greater the weight. Though the exact causality is still unclear, we are aware that the effect of the gestation on babies' weight is greater on smokers compared to non-smokers.

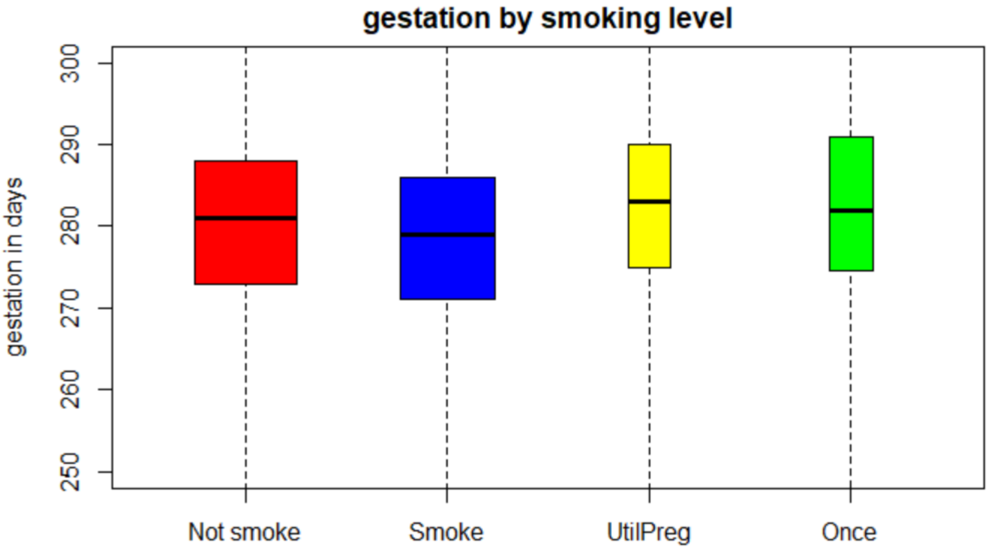
4.5.2 Smoke ⇒ Gestation

In lecture, we learned that babies born early will also have a lower survival rate. Hence, we can examine how a mother's habit of smoking affects gestation age in order to assess the significance of smoking to babies' health. This association is suggested by an article from CDC (see reference). Since we want to compare data among several levels of a categorical variable, we first divide the data into groups based on different levels of "smoke" variable. Then we draw parallel boxplots for these groups.



In the graph above, we first observe that the range of babies' gestation is the largest when his mother doesn't smoke. Since this boxplot contains both the maximum and minimum values among these four, we cannot make further interpretation. Furthermore, each of the boxplots illustrates a well-centralized distribution such that the center half of the data are squeezed in a relatively small range around 260 - 290 days.

From this graph, we can barely tell anything about these representative values. Hence, we need to take a closer look at it.



Caption: 1. In this graph, we can see that both median and the quantiles for those who always smoke are lower than those who don't smoke. 2. From the graph we also observe that those mothers who stop smoking when pregnant or smoke for once before their pregnancy have a higher values of median and quantiles. 3. we observe that the sample size for these two boxplots are both much smaller according to box width as expressed in the code

Let's zoom in and take a closer look. We can analyze the graph according to the above caption point to point. 1.This is a persuasive visual evidence that smoking leads to a shorter gestation. Since we know that gestation positively affects babies' health, this finding illustrates that gestation can be a confounding variable in our analysis. As a result, we can narrow down our argument to make it more specific and persuasive. 2.This is not in the same direction as our underlying assumption. 3.Yet the second point should not be scientifically counterintuitive because of point 3. But again, this finding will narrow down our argument to seeking difference between those mothers who never smoke and always smoke and consider gestation as a possible confouder, while incentivizing new research and discovery.

5 Conclusion

Responding to the first general question, three sub-claims are formed based on the analysis: baby birth weights will be comparably lower if their mothers smoked during their pregnancy; mothers who did not smoke are more likely to bear babies with higher birth weights; mothers who smoked during pregnancy will be about twice likely to give birth to a low weight infant. Meanwhile, responding to the second question, two discoveries are given: the first one reflects a strong statistical evidence that non-smoking mothers have a disposition to bear higher birth weight babies based on the two sample t-test made; the second is the variables including gestation, baby heights, and the numbers of cigarettes smoked per day are closely related to the baby birth weight, and the dad weight is a variable needed to be further explored. The results we find generally reflect the articles we researched, that is there is an intrinsic relationship between smoking and birth weight. According to the articles, it is well believed that smoking is strongly related to baby birth weight and neonatal mortality. This may support the significance of our research that smoking changes birth weight and consequently affects survival rate. In addition, our research supports that there are other factors we need to consider in which qualify our research result. For example, we also discover that smoking can make gestation shorter resulting in negatively influencing babies' health. This makes the research more persuasive and specific while allowing the possibility for further study. Hence, to conclude, smoking can increase the probability of mothers to give birth to comparably lower birth weight babies and some other variables need to be considered in the research.

6 Further study

After this research, several extensions of this study should be easily carried out. First, we can relate to the research intrinsically that if smoking habit affects gestation while subsequently gestation affects baby birth weight, is there a direct relationship between smoking and baby birth weight? Or equivalently, is gestation a real confounding variable? Second, we can narrow down the argument once more by criticizing the effectiveness of the variable ”smoke”. More specifically, it doesn’t demonstrate recognition to how much a mother smokes. Hence, we can do some research based on the variable ”number” to assess how much she smokes.

7 Theory

7.1 Mean

Mean equals to the average value of data in your sample or population. Such value can always be basically categorized sample mean and population mean, and the sample mean is always used to estimate the situation of the whole population. The sample mean can be an unbaised estimator as it can be the least squares estimator according to the Gauss Markov theorem.

7.2 Histogram

Histogram is a straightforward visual estimator of probability density function. As we have a dataset, we can categorize the data points into a number of bins and use the number of data collected in each bin to reconstruct the PDF. In fact, histogram follows the philosophy of using sample data to estimate population distribution.

7.3 Two Sample t-testy density function. t

Welch two sample t-test is designed to test whether two samples have the equal mean. It has the following hypothesises:

$$H0 : mean(X1) = mean(X2) \tag{9}$$

$$H1 : mean(X1) > mean(X2) \tag{10}$$

It’s test statistics are defined as the following:

$$t = \frac{mean(X1) - mean(X2)}{\sqrt{\frac{s1^2}{N1} + \frac{s2^2}{N2}}} \tag{11}$$

The test statistic follows a t-distribution with a degree of freedom v

$$v \approx \frac{(\frac{s1^2}{N1} + \frac{s2^2}{N2})^2}{\frac{s1^4}{N1^2(N1-1)} + \frac{s2^4}{N2^2*(N2-1)})} \tag{12}$$

7.4 Multivariate Linear Regression

Multivariate linear regression is used to estimate the association between multiple explanatory variables and a response variable. The model comes in the following format:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{13}$$

Several important assumptions are made regarding to this model:

- The response variable is linearly dependent on all of covariates $x_1, ..., x_p$.
- All of the observations $y_1, ..., y_n$ are independent.

- x_j is uncorrelated to the error ϵ.
- The errorϵ is normally distributed.
- The covariates $x_1, ..., x_p$ are not highly correlated.
- The mean and variance of ϵ_i are the same for i.

8 Appendix

8.1 Contribution

ˆ Shuyang Zhang: did exploratory data analysis on each variable within the data set before further investigation; did data cleaning.

ˆ Zetong Lai: wrote the introduction and accomplished the conclusion part based on the analysis we have made. Also, write the "mean" part of the "Theory" section and did some adaption of the "analysis" section in the report. Review and proofread.

ˆ Yibei Cai: Did numerical analysis(4.1), part of the analysis of incidence(4.3), the analysis of gestation on weight, and part of the Theory(7.3,7.4).

ˆ Yuan Lin: conducted background research and completed the background section (part 3). Wrote the variable types in the data section (part 2).

ˆ Junqian Liu: wrote part of dataset and the basic structure of the report; did parts of the histogram, two sample t-test, and multivariate regression.

ˆ Yiteng Lu: did part of the basic structure; did 4.2.1(boxplot), 4.5.2(smoke and gestation), further study; part of 4.2.2(histogram), 4.3(incidence), and Theory(7.2); .

8.2 Citation

ˆ Butler, N. R., et al. "Cigarette Smoking in Pregnancy: Its Influence on Birth Weight and Perinatal Mortality." The BMJ, British Medical Journal Publishing Group, 15 Apr.1972, `www.bmj.com/content/2/5806/127.abstract?casa_token=ndJrsIXcazgAAAAA$\%$3AXLk_jBTXcNDuSbFnbQ6IpG-FfM--lUs5Td3f2t-ZJy0GcGdZwCFXUadQBITHmAE77NgSU4ef4RBD7w`

ˆ Quinn, Julie-Anne, et al. "Preterm Birth: Case Definition amp; Guidelines for Data Collection, Analysis, and Presentation of Immunisation Safety Data." Vaccine, Elsevier Science, 1 Dec. 2016, `www.ncbi.nlm.nih.gov/pmc/articles/PMC5139808/`.

ˆ Vangen, Siri, et al. "Heavier the Better? Birthweight and Perinatal Mortality in Different Ethnic Groups." OUP Academic, Oxford University Press, 1 June 2002, `academic.oup.com/ije/article/31/3/654/629787?login=true`.

ˆ "Welch's t-Test." Wikipedia, Wikimedia Foundation, 14 Jan. 2021, `en.wikipedia.org/wiki/Welch$\%$27s_t-test`.

item[ˆ] Professor Wenxin Zhou's power point from Math185