MATH 289C

Exploratory Data Analysis: Case Study 1

Saurabh Kulkarni
PID: A53099844

Mitesh Gadgil
PID: A53095373

UCSD, WINTER 2016

Table of Contents

1. Objective

Our objective is to answer the questions: **What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?**

2. Data Collection

Data collected for our study is enlarged portion of the mentioned CHDS data. The data consists of all pregnancies that occurred between 1960 and 1967 among women in Kaiser Health Plan in Oakland, California.

Population of Interest: All women who delivered babies recently
Sample of Interest: The women in the study are all those that were enrolled in Kaiser Heath Plan has obtained prenatal care in San Francisco area and delivered in any of the Kaiser hospitals in Northern California. Our study is comprised of 1236 babies: All boys, single births (no twins), all lived at least 28 days
Population to which results can be generalized: Adult women who will give birth to boys as their first child.

Both text-files: babies.txt and babies23.txt are essentially the same dataset. The dataset babies.txt is a more concise and reduced version of the dataset, focussing primarily on baby bodyweight, gestation time and mother's height and weight. The more detailed dataset focuses on more parameters mentioned later.

Healthy Baby Parameters

Healthy Gestation Range: 259 up to 294 days

Healthy Baby Body Weight: 88 up to 141 ounces

However the gestation range is not as important a factor as body weight. This is because despite having an abnormal gestation period, if the body weight is normal the baby can be considered to be healthy.

Concise Dataset

Variable	Description	Type of Variable
Bwt	Birth weight in ounces (999 unknown)	Numerical, Discrete
gestation	Length of pregnancy in days (999 unknown)	Numerical, Discrete
Parity	0= first born, 9=unknown	Numerical, Discrete
Age	mother's age in years	Categorical, Discrete
Height	mother's height in inches (99 unknown)	Numerical, Discrete
Weight	Mother's prepregnancy weight in pounds (999 unknown)	Numerical, Discrete
Smoke	Smoking status of mother (0=not now, 1=yes now, 9=unknown)	Categorical, Discrete

So for the smaller dataset we need to verify whether the body weight of the baby, which is an indicator of the babies health, is dependent in any way on whether the mother smokes or not. The key variables into question are "bwt" and "smoke". The other variables "gestation", "parity",

"height", "weight" may come into the picture as confounding variables. For eg: If the body weight of the baby is low and the gestation time is low as well, this can be due to the premature birth and not because of smoking in particular.

Elaborate Dataset

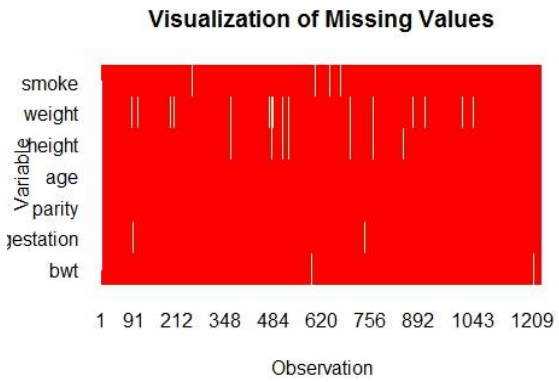
Variable	Description	Type of Variable
Id	Identification number	Numerical, Discrete
plurality	5 = single fetus	Categorical
outcome	1 = live birth that survived at least 28 days	Numerical
Date	Birth date where 1096 = January 1,1961	Categorical
Wt	Birth weight in ounces (999 unknown)	Numerical, Discrete
gestation	Length of pregnancy in days (999 unknown)	Numerical, Discrete
Sex	Sex of the baby (1=male, 2=female, 9=unknown)	Categorical
Parity	0= first born, 99=unknown	Numerical, Discrete
Age	mother's age in years	Categorical
Ht	mother's height in inches (99 unknown)	Numerical, Discrete
Wt	Mother's prepregnancy weight in pounds (999 unknown)	Numerical, Discrete
Race	Mother's race 0-5= white, 6=mex 7=black 8=asian 9=mixed 99=unknown	Categorical
Ed	Mother's Education (0= less than 8th grade, 1 = 8th -12th grade - did not graduate, 2= HS graduate--no other schooling , 3= HS+trade, 4=HS+some college 5= College graduate, 6&7 Trade school HS unclear, 9=unknown)	Categorical
Drace	father's race, coding same as mother's race.	Categorical
Dage	father's age, coding same as mother's age	Categorical
Ded	father's education, coding same as mother's education	Categorical
Dht	father's height, coding same as for mother's height	Numerical, Discrete
Dwt	father's weight coding same as for mother's weight	Numerical, Discrete
Marital	1=married, 2= legally separated, 3= divorced, 4=widowed, 5=never married	Categorical
Smoke	Smoking status of mother (0=never, 1=yes now, 2=until pregnancy, 3=once did, not now 9=unknown)	Categorical
Time	If mother quit, how long ago? 0=never smoked, 1=still smokes, 2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago, 5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago, 8=10+yrs ago, 9=quit and don't know, 98=unknown, 99=not asked	Categorical

3. Analysis: Basic Dataset

Objective of this analysis to verify the association between baby bodyweights and the We have used numerical and graphical techniques to find out if there is any correlation or dependence among the variables in the concise dataset (i.e. babies..txt).

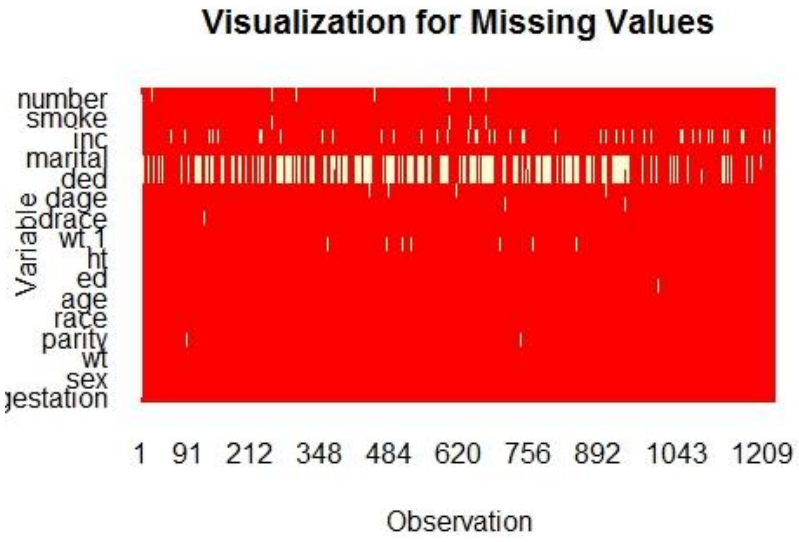
Missing Variables and cleaning the dataset

We know that there are a lot of variables missing in the dataset (saved as '9' or '99' or '999' etc.) This image shows a visualization of the missing variables in this dataset. The x axis is the sample number while the y axis is the variable. If the variable y of sample x is present the pixel coloured red else it is white. This tells us how much of the data is missing and how much is actually available for analysis



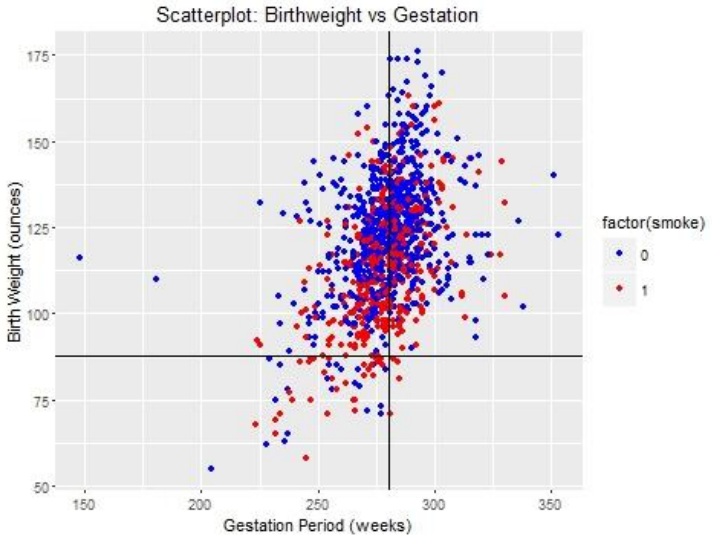
Clearly missing values of important variables like body weight, gestation time are important for our analysis and those particular points are not useful for our analysis as they will only create outliers. Hence we need to clean the dataset and remove the data-points where the crucial variables are missing.

When we look at the more elaborate dataset: babies23.txt we see the following missing variables: Marital status and father's education are inconclusive variables as many are missing. Also we already know from the sampling that parity is always 0 and sex is always 1



Scatterplots

Body weight vs Gestation



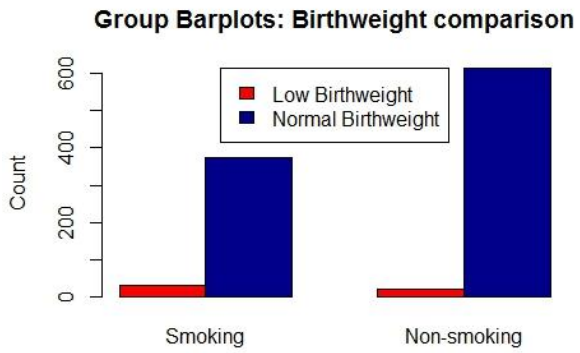
Observation

The scatterplot shows the locations of data-points in the XY-plane where X signifies Gestation time and Y signifies bodyweight. Typically, the healthy baby should fall in the normal range of both gestation time and body weight. The two lines signify the boundary conditions for healthy baby. The vertical line shows the minimum number of gestation period for a healthy baby, and the horizontal line shows the minimum body weight for a healthy baby. Also we can see that there are more data points in the top right and bottom left quadrants. This is obvious as mothers who tend to have expected gestation time have healthier babies and when the gestation time is low, naturally the baby weight will be low.

Hence based on the scatterplot we can say: ***The "gestation" and the "bwt" is positively associated***

Bar Graphs

Health of the Baby (given by body weight)



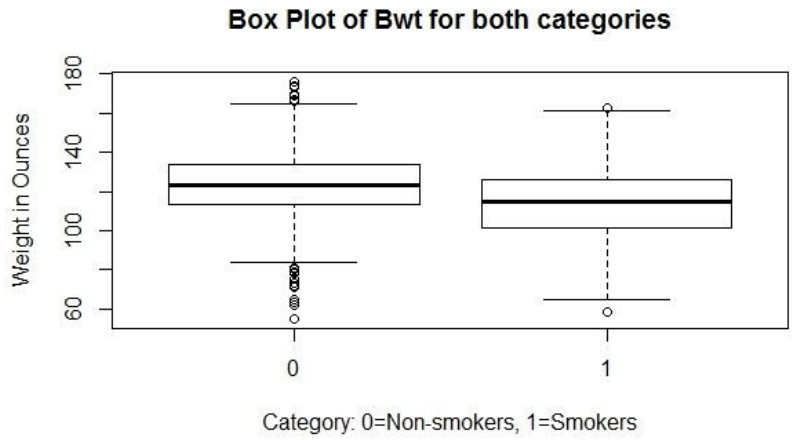
Observation

The above bar graph is an important statistic as it gives the ratio of unhealthy babies to healthy babies in smokers and non-smokers group. We can see that the ratio of unhealthy babies is higher among smokers than non-smokers.

This clearly indicates some negative association between the health of a baby and whether the mother smoked or not (where health decreases if the mother smoked during pregnancy)

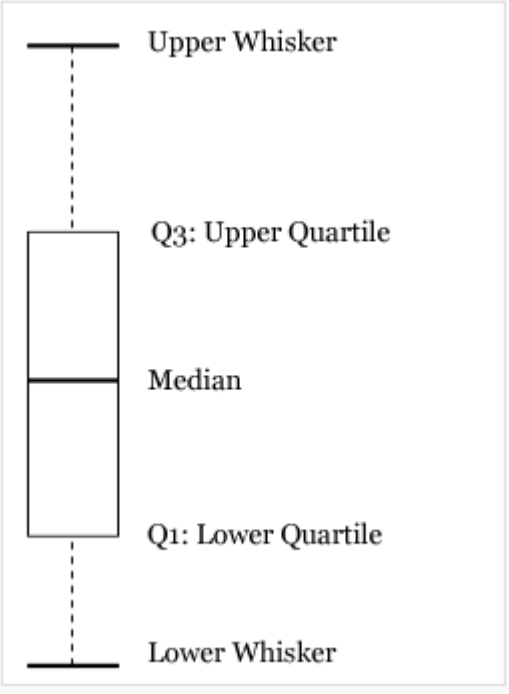
Box Plot

Smokers and Non-smokers



Observation

The box-plot is an important descriptor of the dataset to be analysed. There are a few terms that we should be familiar with, in order to interpret a box plot.



1. Lower Quartile (Q1): 25% of the data samples have value less than this number.

2. Upper Quartile (Q3): 75% of the data samples have value less than this number.

3. Median: There are exactly as many data samples with a value greater than the median, as there are samples with a value less than the median.

4. Lower whisker: this is the minimum value of the dataset or the value 1.5*IR less than Q1, where IR is the inter-quartile range (Q3-Q1), if the minimum value is lower than this value. All points with values less than the lower whisker are deemed outliers.

5. Upper whisker: This is analogous to the lower whisker except that it is the maximum value of the dataset or 1.5*IR above Q3.

Looking at the box plots of weights of babies born to smoking and non-smoking mothers, we can get an overview of the characteristics of the values in the two groups.

Firstly, we observe that the plot for weights of babies born to non-smoking mothers is vertically higher than the plot for smoking mothers. This suggests that the weights of babies born to non-smoking mothers is generally more than those of smoking mothers.

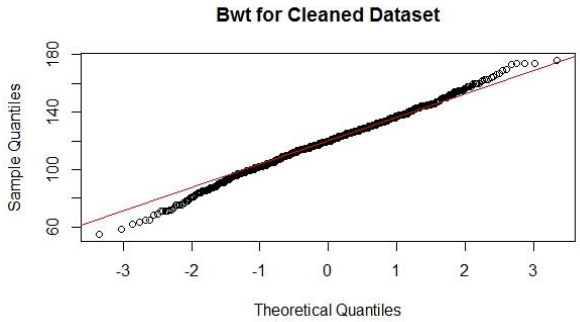
Next, the median for weights of babies born to non-smoking mothers is greater than the smoking group, which reaffirms the fact of heavier and healthier babies being born to non-smoking mothers compared to smoking mothers. The inter-quartile range for the non-smoking

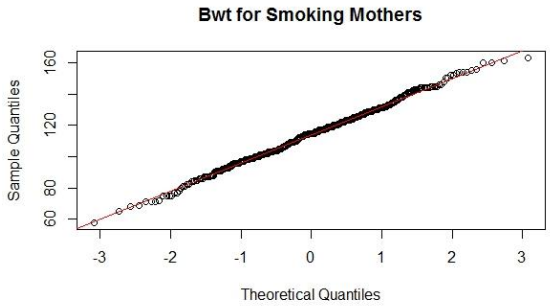
group is significantly lesser than the smoking group i.e. weights of babies born to non-smoking mothers are clustered around the median more heavily than the smoking group. Given this difference in inter-quartile ranges, there is a difference in the positioning of lower and

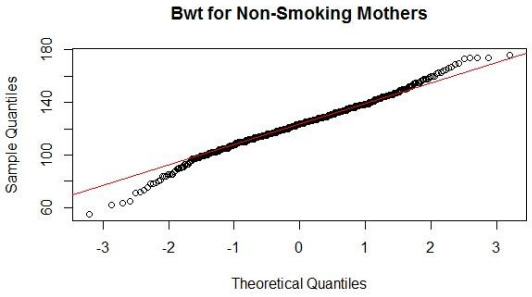
upper whiskers. Hence, we can see a number of outliers in the non-smoking group, while there are a lot less outliers in the smoking group.

From, this analysis we can imagine the dataset of non-smoking birthweights to have 3 clusters, 1 near the median in the inter-quartile range and 2 others in the outlier region beyond the whiskers. While, the smoking birthweights are slightly more scattered without dense clusters.

Quantile-Quantile Plot





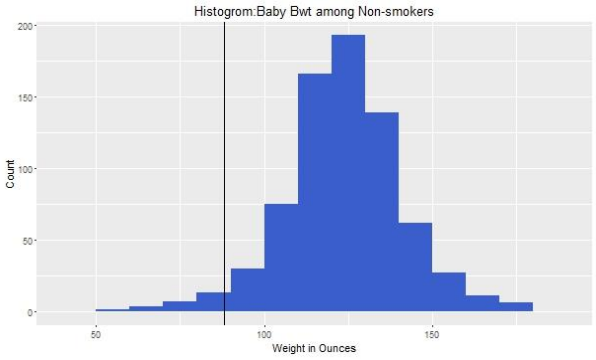


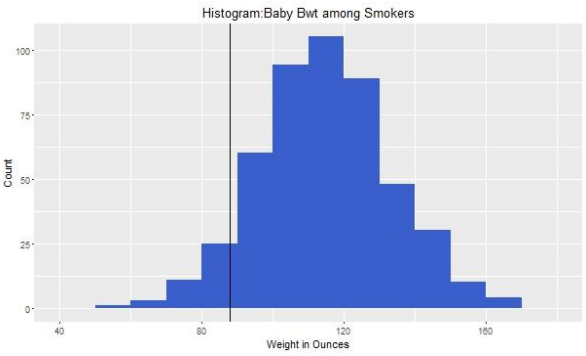
Observation

The quantile-quantile plots indicate how close the data is to normal distribution. By the reference line we can clearly see that all the plots are close to normal distribution. The smoking mother's plots especially clearly follows the normal distribution curve even at the outliers.

Histograms

Body weight for both categories





Observations

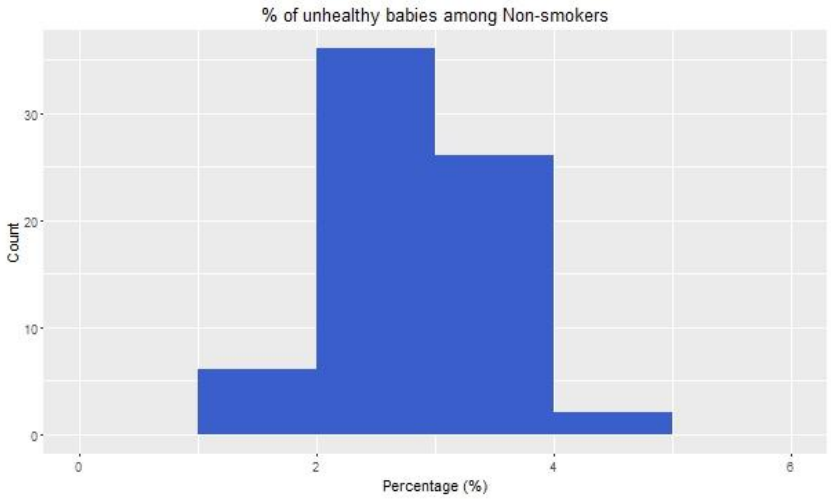
The above figures show the histograms of bwt among smokers and non-smokers. The vertical line in both graphs indicates the boundary condition for a healthy baby, i.e. bwt = 88. We can clearly see both histograms tend to follow a Gaussian like distribution. The important part to observe is that among smokers higher part of the curve falls under the boundary condition compared to non-smokers. Both histograms are symmetric, unimodal. The non-smokers graph has mode at around 125 and the smokers graph has mode around 115.

Estimation and Numerical Analysis

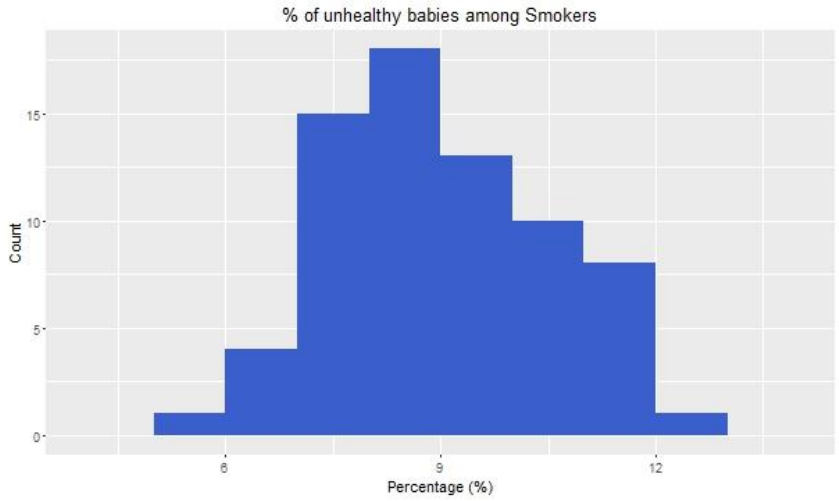
Here we separate data among smokers and non-smokers. Then we use 90% of the data (data is selected using random sampling) in each of the separated datasets to find mean of the baby weight, mean gestation time and frequency of incidence of unhealthy babies. 10% number of trials were taken (~70) to avoid repetition in sampling.

These means are essentially the sample means of the baby weight and gestation time as well as frequency. But as we have seen that the histograms shown above are Gaussian, the sample means are the maximum likelihood estimators of the actual mean.

Frequency of Incidence



The above graphs indicate the percentage incidence of the unhealthy babies among non-smokers. We can clearly see that the curve is Gaussian, Unimodal and has a mean at around 3%.



The above graphs indicate the percentage incidence of the unhealthy babies among smokers. We can clearly see that the curve is Gaussian, Unimodal and has a mean at around 9%. We further see that this curve is within the 6% to 12% range and clearly has a higher percentage of incidences than the smokers.

The above analyses clearly show that smoking is in fact affecting the health of the baby.

Estimators

We saw the mean estimator in the above histograms. There are other parameters which tell us a lot about the data. Standard Deviation, Skew and Kurtosis. Skew is $E(X^3)$ and Kurtosis is $E(X^4)$.

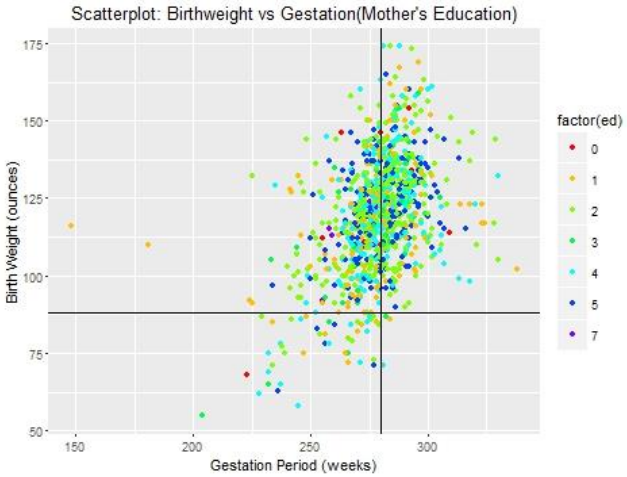
In order to verify the Gaussian nature, we perform kurtosis on both the subsets (smoking mothers and non-smoking). Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. The kurtosis of an ideal normal distribution is ~3. The sample estimators we found of this dataset are (variable is body weight):

Skew = -0.14 for non-smoking mothers
Skew = -0.073 for smoking mothers
Kurtosis = 4.01 for non-smoking mothers
Kurtosis = 2.95 for smoking mothers
Std Dev = 17.21 for non-smoking mothers
Std Dev = 18.37 for smoking

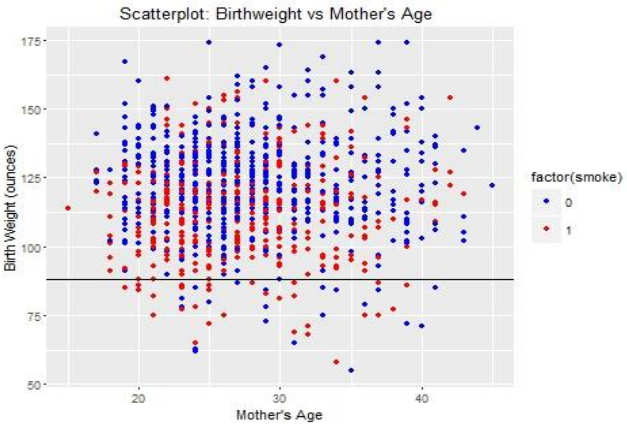
The kurtosis and skew values indicate that the data for non-smokers need not be Gaussian but the data for smokers is definitely Gaussian in nature. This indicates that by central limit theorem, the smoking mother's samples are as such as they are drawn from a distinct population/group. Hence this indicates that smoking indeed is a factor affecting the baby body weight.

4. Advanced Analysis

In addition to the dataset with lesser variables namely babies..txt, we have linked the bigger dataset and performed exploratory analysis on these other attributes as well. Few such attributes are the mother's age, mother's education and the income of the family.



As can be seen from the above scatterplot, the mother's education seems to have a positive effect on both the gestation period as well as the birth weight of the baby. This can be attributed to better care taken during pregnancy owing to proper knowledge and good practices being followed. Thus, in general the mother's education has a positive effect on the health of the baby.



From the above scatterplot, we try to investigate the relationship between the mother's age and the weight of the baby born to her. We can see that there is no clear relationship as most points in the plot appear random and don't follow a particular pattern. The low birthweight points below the horizontal mark aren't clustered at some interval but are more or less uniformly distributed.

This kind of an analysis can help us find out if other variables can better explain the trends in low birthweight that we actually observe. Such analysis was made difficult by the lower number of samples for each categorical variables since each categorical variable took on values of upto 7 types.

With a lesser number of categories and a larger dataset, mosaic analysis can be performed to gauge the relative occurrence of various combinations of factors and its effect. Such an analysis also makes use of statistical significance tests and helps reach statistically significant conclusions.

Conclusions

Using the complete datasets given we are trying to answer the question: What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

The concise dataset is a more useful for quick analysis to see the kind of association between the weights of the babies and maternal smoking. We clearly observe a positive association between the two variables. Although the scatter-plot is not as conclusive, the histogram and the frequency of incidence clearly show that smoking affects the body weight of the baby in a bad way.

The advanced analysis made use of the bigger dataset with variables that allowed exploration into the possible confounding variables like family's income, mother's education and whether they have an influence on the health of the child born. This analysis cannot be used for making significant conclusions since, the number of samples of each category of the variable are small.

But by purely basing our results on the dataset given and the graphical analysis we have done and the numerical values of skew and kurtosis.