University of Waterloo

# Final Project Report

STAT 331 Fall 2018

*Group 91:*

Gnegyao Yuan(20613017)

()

# Contents

# 1 Summary

The goal of the STAT 331 final project is to explore the relation of healthy male single-fetus birth weight and some explanatory variables. This report will be divided into 4 main sections:
Summary, which will cover the main purpose of the report and give a brief explanation of how the project will analyze the data. Two candidate models will be produced in the model selection section by using the pre-fitting data diagnostic and automated model selection. Model diagnostics section will perform an in-depth comparison of the two candidates models by comparing different types of residual plots, leverage and influence measures and cross-validation(rPMSE). In the end, there will be a discussion section basing on the result of the most likely linear model we get from the previous sections to talk about several topics such like: "what is the most important factors associated with/influencing birth weight?'
After using serial statistical analysis way that we learned in STAT 331 course, we find that

# 2 Model Selection

## 2.1 Brief data overview

By view the summary of the data, we notice that there are several illegal data. The domain of "marital"(the mother's marital status) is 1 to 5, but it is clearly showing that there exist 0 in "marital," we replace all the 0 to NA since it is not available data(out of range).
For the categorical predictor "meth"(The self-reported ethnicity of the mother) and "feth"(The self-reported ethnicity of the father), all 0 to 5 is Caucasian meaning that they are in the same group, so we replace the 0-5(Caucasian) to 0, 6(Mexican) to 1, 7(African-American) to 2, 8(Asian) to 3, 9(Mixed) to 4, 10(Other) to 5

## 2.2 Transform categorical predictors

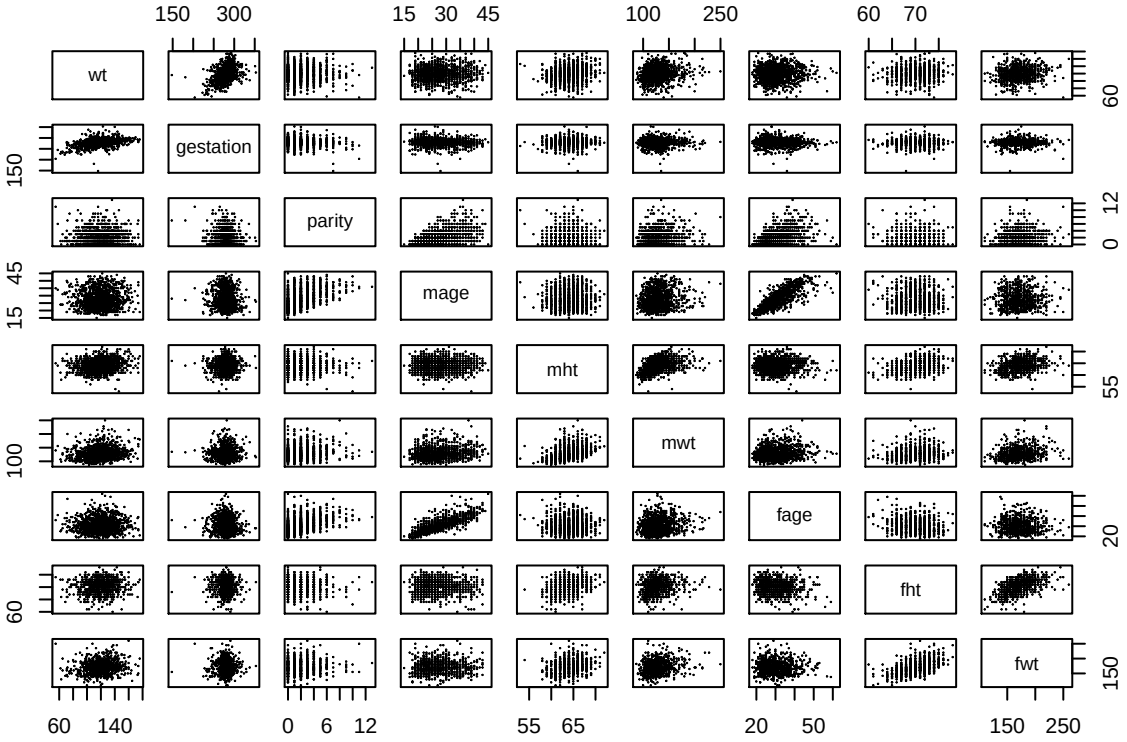
Since all the categorical predictors should not be treated as continuous variables, although they may look like continuous variables (such like 0,1,2,3,4....), we use "one-hot" encoding scheme to make new factors for them. For example, 'med' means the mother's education, whose domain is 0 to 7, where '0' level means 'elementary school' level, level '1' means 'middle school' level, level '2' means 'high school' .... etc, we just transfer numbers to factors with the same name(for example, number 1 to NEW factor '1'), after successfully transfer all the levels to new factors, dropping all the '0' levels for all categorical predictors by the requirement of one-hot"encoding scheme(we can do that because all predictors have '0' level(meth and feth didn't have, but we already transform them)), and give all the new factors a 0/1 binary variable to show that factor is applied to this data or not.
There is a trick in R code:"as.factor" function, it can automatically transfer variables to new factors, so we applied it on all the categorical predictors(meth, med,feth, fed, marital, smoke, time, number) to factor type instead of continuous variables.

## 2.3 Drop off NAs and prediction missing data

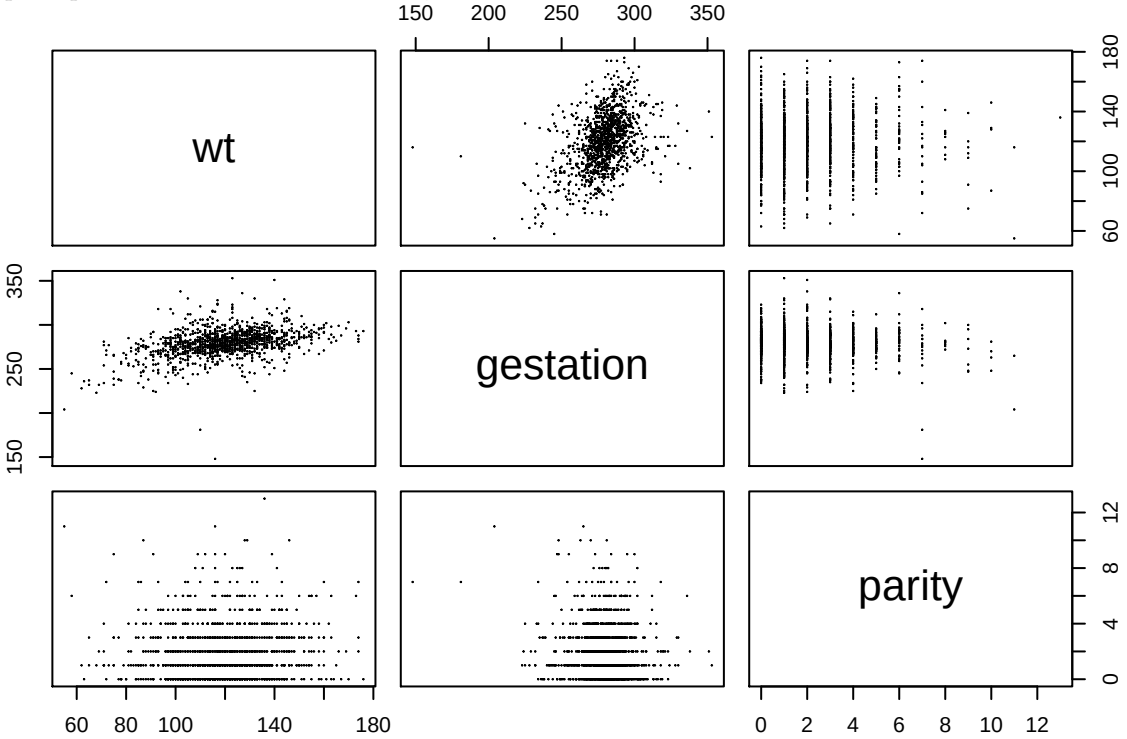
It is clear that there is still lots of NA data point in our data frame, there are several methods covered in STAT 331 can produce missing data points. However, we use MICE here.
MICE can help us to impute missing values which are drawn from a distribution specifically designed for each missing data point. Don't like replace all NA variables by mean of the data; MICE can also include the 'var' in the data prediction, which can help lessen the bias and make the data close to the original.

## 2.4 Revisiting the category variables

## 2.5 Visually inspect data




By using pairs plot of all countinuse variables,




```
## [1] TRUE
```

```
## [1] TRUE
## [1] "timestill smokes" "timeothers"
## [1] TRUE
```