

**School of Mathematics and Statistics**

**MSc Data-Intensive Analysis**

**MSc Applied Statistics and Datamining**

**MT5762 INTRODUCTORY DATA ANALYSIS**

GROUP: DRUNKEN MASTER 2

ID Numbers: 180012191; 180025784; 110013122; 180029941, 180024795

**What relationships are there between the measured variables and the birth weight of babies?**

*Producing a model that describes potential drivers of low birth-weight babies.*

Tutor: Dr. Carl Donovan

**Executive Summary**

The present report is a data analysis project, which focusses mainly on fitting linear models. It intends to determine what relationships are there between the measured variables (Ethnicity, Age, Smoking habits (…)) and the birth weight of babies[[1]](#footnote-1).

In order to do so it is required to produce a model that describes potential drivers of low birth-weight babies.

**INDEX**

**Content**

[1 INTRODUCTION 1](#_Toc528931349)

[1.1 Variables in Data File Babies 3](#_Toc528931350)

[2 METHODS 4](#_Toc528931351)

[2.1 Data Cleaning 4](#_Toc528931352)

[2.2 Data Exploration 5](#_Toc528931353)

[2.2.1 Baby Birth Date Weight 5](#_Toc528931354)

[2.2.2 Gestation Period vs. Birth Weight 5](#_Toc528931355)

[2.2.3 Scatterplot of Mother’s Weight vs. Baby´s Weight 6](#_Toc528931356)

[2.2.4 Analysis of wt.1 (mother's weight) 7](#_Toc528931357)

[2.2.5 Scatterplot of father's weight against baby's weight 7](#_Toc528931358)

[2.2.6 Mother´s Smoking Habits 8](#_Toc528931359)

[2.3 Model Fitting 8](#_Toc528931360)

[2.3.1 Data Model 8](#_Toc528931361)

[2.3.2 Fitting a First Order Interaction Model 10](#_Toc528931362)

[2.4 Check Assumptions 13](#_Toc528931363)

[2.5 Other Tested Models 13](#_Toc528931364)

[3 FIVE-FOLD CROSS VALIDATION (Final Model) 14](#_Toc528931365)

[3.1.1 Five-Fold Cross-Validation 14](#_Toc528931366)

[3.1.2 Mean Square Error (MSE) 15](#_Toc528931367)

[3.2 Results Bla 2 16](#_Toc528931368)

[3.3 Resultas Bla 3 16](#_Toc528931369)

[3.3.1 Results 31 17](#_Toc528931370)

[3.3.2 Results 32 17](#_Toc528931371)

[3.3.3 Results 33 17](#_Toc528931372)

[4 DISCUSSION 19](#_Toc528931373)

[4.1 Discussion 1 19](#_Toc528931374)

[4.1.1 Discussion 11 19](#_Toc528931375)

[4.1.1.1 Discussion 111 20](#_Toc528931376)

[4.1.2 Discussion 12 20](#_Toc528931377)

[4.1.2.1 Discussion 121 20](#_Toc528931378)

[4.1.2.2 Discussion 122 21](#_Toc528931379)

[4.2 Discussion 2 21](#_Toc528931380)

[4.2.1 Discussion 21 22](#_Toc528931381)

[4.2.2 Discussion 22 22](#_Toc528931382)

[5 CONCLUSIONS AND RECOMMENDATIONS/DISCUSSION SUMMARY 23](#_Toc528931383)

[6 BIBLIOGRAPHY 24](#_Toc528931384)

# INTRODUCTION

As anyone ever said “life is a gamble” to you? Such a statement reflects the feeling that our lives are surrounded by unpredictable, or “random”, events (Wild & Seber, 2000, p.1).

The aim of the present report is to produce an analysis and discuss some results that can answer the question “what relationships are there between the measured variables and the birth weight of babies?”

The data used in this report is part of a larger group of studies from the Child Health and Development Studies (CHDS), which *“are prospective longitudinal studies on medical and social aspects of pregnancies and on the health and development of children”[[2]](#footnote-2)*.

It is known that there are many potential drivers of Low-Birth Weight (LBW) babies. According to Kramer (Kramer, M, 1987, p.663), “*factors with well-established direct causal impacts on intrauterine growth*” and consequently LBW, “*include infant sex, racial/ethnic origin, maternal height, pre-pregnancy weight, paternal weight and height, maternal birth weight, parity, history of prior low-birth-weight infants, gestational weight gain and caloric intake, general morbidity and episodic illness, malaria, cigarette smoking, alcohol consumption, and tobacco chewing*”[[3]](#footnote-3).

The data set we are analysing in this report contains most of the variables referred above.

“*Of the 127 million infants born in the world in 1982, 20 million (16%) were estimated to weigh less than 2500g., and over 90% of these infants were born in developing countries, a function not only of the higher birth rate in these countries but also of their LBW[[4]](#footnote-4)*” (Kramer, M, 1987, p.664).

This report was produced using Microsoft Office (Word and Excel)[[5]](#footnote-5) and the Data Frames and its respective manipulation, statistical analysis and plotting were produced in R programming language using the software R-Studio version 3.5.1 available on Apps Anywhere[[6]](#footnote-6).

## Variables in Data File Babies

The variables that were subject to analysis in this report are listed and described in Figure 1: Variables in Data File Babies.

|  |
| --- |
| Figure 1: Variables in Data File Babies  Source: Adopted from:  <https://github.com/ziqsu/MT5762-Project2/blob/master/description%20of%20babies%20data.txt> |

# METHODS

“All models are wrong, but some models are better than others.” (Crawley, 2015, p.4)

## Data Cleaning

Data cleaning deals with data problems once they have occurred. Error-prevention strategies can reduce many problems but cannot eliminate them. We present data cleaning as a three-stage process, involving repeated cycles of screening, diagnosing, and editing of suspected data abnormalities (Van den Broeck, Argeseanu Cunningham, Eeckles, & Herbst, 2005).

|  |
| --- |
| Figure 2: Data-Cleaning Framework  (Illustration: Giovanni Maki)  <https://doi.org/10.1371/journal.pmed.0020267.g001> |

In order to start working and exploring the given data (babies23.data) and to prevent possible errors, the data was checked and cleaned. The data-cleaning framework followed is illustrated in **Figure 2**.

During this process unknown, not clear or not needed, or irrelevant data aspects were cleaned and some existing factor variables were transformed into numeric (Appendix 1 – R code).

## Data Exploration

This part of the report will be dedicated to exploring and determining potential relationships within the data set. The data exploration was produced using R Software.

### Baby Birth Date Weight

A quick exploratory analysis of the histogram (Figure 3) shows that the data is approximately normally distributed and the baby weight values are cantered about 120 ounces. The estimate of the mean is 119.8 ounces.

|  |
| --- |
| Figure 3: Density Histogram of Birthday Weight |

### Gestation Period vs. Birth Weight

Exploring the data related with gestation period and birth weight one can observe (Figure 4) that there is an increase in birth weight as gestation period increases. From the correlation point of view this are the variables that present the strongest relationship with Birth Weight.

|  |
| --- |
| Figure 4: Scatterplot Mother´s Weight against Baby Weight |

### Scatterplot of Mother’s Weight vs. Baby´s Weight

The following scatterplot (Figure 5) does not indicate a strong effect between the variables.

|  |
| --- |
| Figure 5: Mother´s Weight vs. Baby´s Birth Weight |

### Analysis of wt.1 (mother's weight)

Looking at the scatterplot of mother's weight against baby's weight (Figure 6) it can be seen that it does not indicate a strong effect between the variables.

|  |
| --- |
| Figure 6: Mother´s Weight and Baby´s Weight |

### Scatterplot of father's weight against baby's weight

The following scatterplot, father´s weight against baby´s weight, also does not indicate a strong relationship between the variables.

|  |
| --- |
| Figure 7: Father´s Weight vs. Baby´s Weight |

### Mother´s Smoking Habits

Although 'smoke' had no correlation with birth weight, common sense says that there would be an effect here between factors of smoking. The boxplots show smaller mean for 'smokes now' but it is still within the interquartile range of the other levels of smoking. Therefore, the effect may not be significant.

|  |
| --- |
| Figure 8: Birth Weight per level of Mother´s Smoking |

## Model Fitting

### Data Model - Fitting a Model Using Provided Variables

We use all the variable provided in the data except id and data in our analysis to find a linear model that have the best fit. We choose AIC backwards selection method by using step function. After we get the lowest AIC scores (3359.82), we get the model with following variables (table 1).

|  |
| --- |
| Table 1: Summary |

We also use Anova function to check whether the variable have contributed to the predictive ability of the model, all the p value is less than 0.05, which means all the variables that we selection contribute to the model’s predictive ability. We trace down the extreme residuals in the model. Then we want to check the assumptions about the model. For normality, we use Shapiro Wilk normality test which null hypothesis is the population is normally distributed which our p value is 0.09, we fail to reject the hypothesis. From QQ plot of residuals of the model below and Shapiro Wilk normality test, we could conclude that the model fit normality.

|  |
| --- |
| Figure 9: Histogram of Residuals |

|  |
| --- |
| Figure 10: Normal Q-Q |

For linearity, we have plotted a residual against fitted values graph, even though our graph is not perfect, it is a well-behaved graph that show the linearity of the model. For heteroscedasticity, we use Breusche-Pagan test (ncvTest) which null hypothesis is constant error variance, p value is 0.05 which reject the null hypothesis. So heteroscedasticity do exist and we could also see from the graph of residual against fitted data. For autocorrelation, we use Durbin Watson statistic which null hypothesis is linear regression residuals are uncorrelated, p value is 0.54 which means that we fail to reject the null hypothesis. For collinearity, we use variance inflation factors to find collinearity in our model. Since all of our variance inflation factor is less than 10, we don’t need to worry about collinearity in our model. (The name of the model is dataModel in r file, you could find all the r file in appendix modelselection-Su.r)

|  |
| --- |
| Figure 11: Residuals vs Fitted |

### Fitting a First Order Interaction Model

In this model. We only looked at first order interactions between two variables. There is a possible that second order interaction between three or more variables maybe useful to analysis. We use all the variables and first order interaction between every two variables in the data except **id** and **date**, so we have roughly 200 to 300 variables to start with. Then we will proceed, choose AIC backwards selection approaches using step function. After this stage, we eliminated a high number of variables, being now reduced to some 50 – 60 variables. Then we examine the collinearity of the first order Model. It is observed that there are a considerable number of variable which GVIF number is larger than 10, which prompt us to the following step.

1. We find the maximum number of GVIF; if it is larger than 10 then we remove it;

2. Repeat checking collinearity using vif function get the maximum - repeat the step 1;

After we remove all the collinearity that variance inflation number is larger than 10, we use AIC backwards selection again since we already form a new model by deleting a lot of collinear variables. After model selection, we are left with 12 variables, the AIC scores of the model is 3358.58. We check the collinearity again and the variance inflation factor, now all VIF are less than 10. After we get the model below, we check for the assumption.

|  |
| --- |
| Table 2: Final Models Coefficients |

For normality, we use Shapiro Wilk normality test which null hypothesis is the population is normally distributed which our p value is 0.22, we fail to reject the hypothesis.

|  |
| --- |
| Figure 12: Histogram of Residuals |

|  |
| --- |
| Figure 13: Normal Q-Q |

|  |
| --- |
| Figure 14: Residuals vs Fitted |

From QQ plot of residuals of the model below and Shapiro Wilk normality test, we could conclude that the model fit normality. For linearity, we have plotted a residual against fitted values graph, even though our graph is not perfect, it is a well-behaved graph that show the linearity of the model. For heteroscedasticity, we use Breusche-Pagan test (ncvTest) which null hypothesis is constant error variance, p value is 0.16 which fail to reject the null hypothesis. So heteroscedasticity do not exist and we could also see from the graph of residual against fitted data. For autocorrelation, we use Durbin Watson statistic which null hypothesis is linear regression residuals are uncorrelated, p value is 0.08 which means that we fail to reject the null hypothesis. Now this model pass all the assumption. (The name of the model is finalModel, you could find all the r file in appendix modelselection-Su.r)

## Check Assumptions

To test for non-constant error variance we use the Breusche-Pagan test (H0: constant error variance) with respect to fitted model. If we have constant error variance then the variation in the residuals should be unrelated to any covariant." null hypothesis is rejected since the p value is less than 0.05.

With the Durbin Watson Test we check upon the correlation of the residuals. If the statistical value is 2 there is no autocorrelation. If the value is between 1.5 and 2.0 we should not concern about it. As our results are 1.84 for the data model and 1.83 for the final model we can assume that both of our models have no or little autocorrelation.

## Other Tested Models

Apart from data model, we tried fitting other models based on other criteria too. Using logic, we tried to fit certain interaction-effect models to observe the effect variables had on baby weight.

Interaction parity **mother weight** - **baby weight** and **mother income – baby weight** were fitted within the model and AIC showed each were really high. Despite the values they all passed the model diagnostic tests: Q-Q, Durbin-Watson and ncv.

# FIVE-FOLD CROSS VALIDATION (Final Model)

### Five-Fold Cross-Validation

We have, so far, looked upon various models in order to find a better performing one. However, it can be difficult to determine if these improvements in scores result from the captures of better relationships within our model or if we are just overfitting the model. In order to clarify this aspect we use validation techniques such as Five-Fold Cross Validation.

This procedure has a single parameter k which refers to a numbers of groups that the given data will be divided into –usually called k-fold cross-validation. When k has a specific values, such as 5, the reference to the model with be 5-Fold Cross-Validation[[7]](#footnote-7).

|  |
| --- |
| R:\WorkingDirectory\MT5762\MT5762-Project2\Predictedobserveddata.png  Figure 15: Cross-Validation predicted values |

From the plots, we cannot say which, **dataModel** or **finalModel,** is the better because the five regression lines all seems parallel in both plots. However from the output, the overall **ms** of dataModel is 255, whilst the value of finalModel is 268. From these values we can understand that the finalModel is little bit better than dataModel.

|  |
| --- |
| R:\WorkingDirectory\MT5762\MT5762-Project2\Predictedresidualfinal.png  Figure 16: Cross Validation Predicted Values |

|  |
| --- |
| R:\WorkingDirectory\MT5762\MT5762-Project2\Preditedcrossvalidation-final.png  Figure 17: Cross-Validation predicted values |

### Mean Square Error (MSE)

Mean Square Error (MSE) is used to evaluate the quality of an [estimator](https://en.wikipedia.org/wiki/Estimator) (parameter) or a predictor (some [random variable](https://en.wikipedia.org/wiki/Random_variable)). **MSE —** is the average of the square of the errors. The larger the number the larger the error[[8]](#footnote-8). The MSE value of **finalModel** is **258** and the value of **dataModel** is **248**.Therefore, **dataModel** seems a better model than **finalModel.**

## Predictions

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut varius laoreet consectetur. Nulla in vulputate lacus. In et volutpat ante. In ultrices turpis neque, id dictum erat auctor ac. Pellentesque mattis, magna rutrum rutrum placerat, orci augue molestie nunc, vel porttitor risus tellus non dui. Vivamus id convallis odio. Duis ut nulla id nulla gravida vulputate a condimentum tellus. Ut purus justo, tempus sed iaculis at, accumsan id sapien. Etiam eu massa vehicula, accumsan urna sed, varius magna. Maecenas dapibus arcu leo, et tempus leo tincidunt id.

# DISCUSSION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut varius laoreet consectetur. Nulla in vulputate lacus. In et volutpat ante. In ultrices turpis neque, id dictum erat auctor ac. Pellentesque mattis, magna rutrum rutrum placerat, orci augue molestie nunc, vel porttitor risus tellus non dui. Vivamus id convallis odio. Duis ut nulla id nulla gravida vulputate a condimentum tellus. Ut purus justo, tempus sed iaculis at, accumsan id sapien. Etiam eu massa vehicula, accumsan urna sed, varius magna. Maecenas dapibus arcu leo, et tempus leo tincidunt id.

## Discussion 1

# CONCLUSIONS AND RECOMMENDATIONS/DISCUSSION SUMMARY

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut varius laoreet consectetur. Nulla in vulputate lacus. In et volutpat ante. In ultrices turpis neque, id dictum erat auctor ac. Pellentesque mattis, magna rutrum rutrum placerat, orci augue molestie nunc, vel porttitor risus tellus non dui. Vivamus id convallis odio. Duis ut nulla id nulla gravida vulputate a condimentum tellus. Ut purus justo, tempus sed iaculis at, accumsan id sapien. Etiam eu massa vehicula, accumsan urna sed, varius magna. Maecenas dapibus arcu leo, et tempus leo tincidunt id.

# BIBLIOGRAPHY

Crawley, M. J. (2015). *Statistics - An Introduction Using R*. ( lda John Wiley & Sons, Ed.) (Second Edi). Sussex.

Kramer, M, S. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization, 65 (5): 663-737 (1987)*. Obtido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=1

Van den Broeck, J., Argeseanu Cunningham, S., Eeckles, R., & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. Obtido de https://doi.org/10.1371/journal.pmed.0020267

Wild, C. J., & Seber, G. A. F. (2000). *Chance Encounters - A first Course in Data Analysis and Inference*. (J. W. & Son, Ed.) (1st ed.). USA: John Wiley & Son, Inc.

**APPENDIX**

Appendix 1 – Abbreviations

**Abbreviations**

**A**

AIC = Akaike´s Information Criterion

**D**

drace = father’s race, coding same as mother´s race

dage = father´s age, coding same as mothers age

ded = father´s education, coding same as mother´s education

dht = father´s height, coding same as mothers height

dwt = father´s weight, coding same as mothers weight

**E**

ed = mother´s education

**G**

GVIF – Variance Inflation Factor

**H**

ht = mother´s height in inches to the last completed inch

**I**

id = identification number

inc = family yearly income in $2500 increments

**L**

LBW = Low Birth Weigh

**M**

**N**

number = number of cigarettes smoked per day for past and current smokers

**O**

**P**

**Q**

**R**

**S**

**T**

**U**

**W**

wt = birth weight in ounces

**X**

1. <https://moody.st-andrews.ac.uk/moodle/course/view.php?id=8191> [↑](#footnote-ref-1)
2. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3016.1988.tb00218.x> [↑](#footnote-ref-2)
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=1> [↑](#footnote-ref-3)
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=2> [↑](#footnote-ref-4)
5. Windows 10 Home, Word and Excel 2013. [↑](#footnote-ref-5)
6. <https://appstore.st-andrews.ac.uk/> [↑](#footnote-ref-6)
7. <https://machinelearningmastery.com/k-fold-cross-validation/> [↑](#footnote-ref-7)
8. Brownlee, J. (2018) ‘A gentle introduction to k-fold cross-validation’, *Statistical Methods*, 23 May. Available at: https://machinelearningmastery.com/k-fold-cross-validation/ (Accessed: 1 November 2018)

   James, G., Witten, D., Hastie, T. And Tibshirani, R. (2014) *An introduction to statistical learning with applications in R*. 7th edn. New York: Springer Publishing. [↑](#footnote-ref-8)