

**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

**An Examination of the Influences on Low Birth-Weight Babies**

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

Tutor: Dr. Carl Donovan

**Executive Summary**

The present report focuses on fitting linear models to determine the effect of different variables on the birth weight of babies.

**INDEX**

**Content**

[1 INTRODUCTION 1](#_Toc528948736)

[2 METHODS 2](#_Toc528948737)

[2.1 Data Cleaning 2](#_Toc528948738)

[2.2 Data Exploration 2](#_Toc528948739)

[2.3 Model Fitting 2](#_Toc528948740)

[3 RESULTS 3](#_Toc528948741)

[3.1 Data Exploration 3](#_Toc528948742)

[3.1.1 Correlation of All Variables with Birth Weight 3](#_Toc528948743)

[3.1.2 Gestation Period 4](#_Toc528948744)

[3.1.3 Mother’s Height 4](#_Toc528948745)

[3.1.4 Mother’s Weight 5](#_Toc528948746)

[3.1.5 Father’s Weight 6](#_Toc528948747)

[3.1.6 Mother´s Smoking Habits 6](#_Toc528948748)

[3.2 8](#_Toc528948749)

[3.3 Model Fitting 8](#_Toc528948750)

[3.3.1 Model with all variables 8](#_Toc528948751)

[3.3.2 Fitting a First-Order Interaction Model 10](#_Toc528948752)

[3.4 Other Tested Models 14](#_Toc528948753)

[4 FIVE-FOLD CROSS VALIDATION 17](#_Toc528948754)

[4.1.1 Five-Fold Cross-Validation 17](#_Toc528948755)

[4.1.2 Mean Square Error (MSE) 17](#_Toc528948756)

[4.2 Predictions 18](#_Toc528948757)

[5 DISCUSSION 19](#_Toc528948758)

[5.1 Discussion 1 20](#_Toc528948759)

[5.1.1 Discussion 11 20](#_Toc528948760)

[5.1.1.1 Discussion 111 20](#_Toc528948761)

[5.1.2 Discussion 12 21](#_Toc528948762)

[5.1.2.1 Discussion 121 21](#_Toc528948763)

[5.1.2.2 Discussion 122 21](#_Toc528948764)

[5.2 Discussion 2 21](#_Toc528948765)

[5.2.1 Discussion 21 23](#_Toc528948766)

[5.2.2 Discussion 22 23](#_Toc528948767)

[6 CONCLUSIONS AND RECOMMENDATIONS/DISCUSSION SUMMARY 24](#_Toc528948768)

[7 BIBLIOGRAPHY 25](#_Toc528948769)

# 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 present report analyses 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”[[1]](#footnote-1)*.

Previous studies indicate that there are many potential drivers of low birth-weight (LBW) babies. According to Kramer (1987), “*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*”[[2]](#footnote-2).

The data set we are analysing in this report contains most of the variables mentioned by Kramer above and will be discussed later.

“*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[[3]](#footnote-3)*” (Kramer, M, 1987, p.664).

Data cleaning, analysis and plotting were produced in the R programming language using the software R-Studio version 3.5.1 (R Core Team, 2018).

# METHODS

## Data Cleaning

The data were cleaned to remove unknown values that were being presented as numerical within the data set. All variables had been classified as integers within the programming software so the numerical ones were changed to numerical to allow analyses to be performed on them.

## Data Exploration

Exploratory analyses were performed on the data to investigate the potential for the existence of relationships between the variables and birth weight. Correlation values were obtained and used to select which variables to explore. These variables were visualised with scatterplots, giving an indication of the strength of the relationship. The categorical variable of mother’s smoking habits was plotted as a boxplot.

## Model Fitting

Linear models were fitted using linear regression, analysis of variance (ANOVA) and the Akaike Information Criterion (AIC). Nominal variables were removed from the data before fitting to ensure they did not affect the result (ID number, for example). Assumptions of the models were checked. Normality was assessed by plotting the residuals along a quantile-quantile (QQ) plot and by plotting a histogram. A Shapiro-Wilk normality test was also performed. Independence was assessed through a scatterplot of the residuals against the fitted values. Constant variance was tested by performing a Breusch-Pagan test and a Durbin-Watson test was performed to test for autocorrelation.

# RESULTS

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

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).

## Data Exploration

### Correlation of All Variables with Birth Weight

To investigate which variables were likely to affect birth weight, the correlation values were calculated. The strongest correlations were gestation period, mother’s height, mother’s weight and father’s weight. None of the correlations are particularly strong but they indicate that relationships may exist. Further exploratory analysis was performed on these variables. The correlation calculation did not include categorical variables, so relationships may exist that are not found here. The correlation values are shown in Figure.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | |  | **Correlation** | | **Gestation Period** | 0.40 | | **Mother’s Height** | 0.22 | | **Mother’s Weight** | 0.17 | | **Father’s Weight** | 0.15 |   Figure 1:**Correlations between variables and birth weight** |

### Gestation Period

A scatterplot of gestation period and birthweight was created to visualise the relationship. As can be seen in Figure, there birth weight appears to increase as gestation period gets longer. This is in line with the correlation value of 0.40 that was found.

|  |
| --- |
| Figure 2: Scatterplot of gestation period against baby weight |

### Mother’s Height

The second-strongest correlation with birth weight was seen with mother’s height (correlation of 0.22). Figure represents this as a scatterplot but does not indicate a strong relationship between the variables.

|  |
| --- |
| Figure 3: Scatterplot of Mother´s Height vs. Birth Weight |

### Mother’s Weight

Figure shows a scatterplot of mother's weight against birth weight. There does not appear to be a strong relationship between the variables. This was expected as the correlation between the variables was 0.17.

|  |
| --- |
| Figure 4: Scatterplot of mother´s weight and birth weight |

### Father’s Weight

The final variable visualised was father’s weight. Its correlation with birth weight was 0.15 so a clear relationship through visualisation was not expected. Figure shows the relationship as a scatterplot and does not indicate a large effect.

|  |
| --- |
| Figure 5: A scatterplot of father’s weight against birth weight |

### Mother´s Smoking Habits

Exploratory analysis was performed on mother’s smoking habits. This was a categorical variable with factors: never smoked, smokes now, smoked until pregnancy, and once smoked (long before pregnancy). Previous studies have suggested that maternal smoking during pregnancy causes low birth-weight in babies (Pereira, Da Mata, Figueiredo, de Andrade, & Pereira, 2017). Therefore, the relationship between mother’s smoking habits and birth weight were explored and visualised using boxplots (Figure). These show a smaller median for birth weight of babies whose mothers currently smoke but it is still within the interquartile range of the other levels of smoking. Therefore, the effect may not be significant.

|  |
| --- |
| Figure 6: Birth weight per level of mother´s smoking habits |

## Model Fitting

### Model with all variables

A model was fitted using all variables except ID and data. The stepwise AIC backwards selection method was chosen. This calculated the AIC score using all variables then removed the variable which caused the largest decrease in AIC score. This was repeated until removing any of the variables caused an increase in the AIC score. The final AIC score was 3359.82, providing a model with the following variables (Table 1).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Coefficients of the variables in the model   |  |  | | --- | --- | | Variable | Coefficient | | **(Intercept)** | -98.99311 | | **Gestation period** | 0.454 | | **Mother’s Previous Pregnancies** | 0.74966 | | **Mother’s Height** | 1.26968 | | **Father’s Race** | -0.56526 | | **Father’s Weight** | 0.07689 | | **Mother’s Smoking Habits** | 2.15663 | | **No. of Cigarettes Smoked by Mother** | -2.16762 | |

An ANOVA was performed to check whether the variables have contributed to the predictive ability of the model. The p-values were all < 0.05, which suggests that the variables selected for the model contribute to the model’s predictive ability.

The assumptions of the model were checked. For normality, a Shapiro-Wilk normality test was performed and returned a p-value of 0.09. As this is greater than 0.05, the null hypothesis (that the data are normally distributed) is not rejected. From the QQ plot of residuals of the model (Figure 8) and Shapiro-Wilks normality test, we could conclude that the residuals of the model come from a normal distribution.

To test for linearity, the residuals have been plotted against the fitted values (Figure). Although the graph is not perfect, it shows the linearity of the model.

For heteroskedasticity, a Breusch-Pagan test was performed. Its null hypothesis is that there is constant error variance. The p-value is < 0.05 so the null hypothesis is rejected. This indicates that heteroskedasticity exists. It can also be seen from the graph of residuals against fitted data (Figure 9).

To test for autocorrelation, a Durbin-Watson test was performed. The null hypothesis states that the residuals are uncorrelated. This returned a p value of 0.54, so we fail to reject the null hypothesis in this case. Also, a DW statistic close to 2 indicates that the residuals are uncorrelated. For this model, the test returned a statistic of 1.84.

Collinearity was tested by using variance inflation factors. Since all of the variance inflation factors were less than 10, collinearity is not considered to be an issue.

|  |
| --- |
| Figure 7: Histogram of Residuals |

|  |
| --- |
| Figure 8: Normal Q-Q |

|  |
| --- |
| Figure 9: Residuals vs Fitted |

### Fitting a First-Order Interaction Model

In this model, first-order interactions between two variables were examined. All variables were used and first-order interactions between every pair of variables in the data were calculated. This created over 200 variables. Stepwise AIC backward selection was performed which reduced the variables to between 50 and 60. The collinearity of this model was then examined. It was observed that there were a considerable number of variables which GVIF number was larger than 10. The variable with the largest VIF value was removed and the test was performed again. This was repeated until all values were below 10.

AIC backwards selection was performed again since many of the previous variables had been removed. After model selection, 12 variables remained and the AIC score of the model was 3358.58. The collinearity was checked again and all VIF values were less than 10. The coefficients of the model are shown in (Table 12).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: **Coefficients of the variables in the first-order interaction model**   |  |  | | --- | --- | | Variable | Coefficient | | **(Intercept)** | -83.21759 | | **Gestation period** | 0.44988 | | **Mother’s Height** | 1.04438 | | **Father’s Education** | -1.33282 | | **Father’s Weight** | 0.07541 | | **Family Yearly Income** | -0.50323 | | **Time Since Mother Quit Smoking** | 1.99309 | | **No. of Cigarettes Smoked by Mother** | -2.02635 | | **Family Yearly Income: Mother’s Previous Pregnancies** | 0.15018 | | **Mother’s Weight: Mother’s Education** | 0.01692 | | **Mother’s Education: Father’s Race** | -0.20793 | | **Time Since Mother Quit Smoking: Mother’s Education** | -0.64961 | | **Father’s Education: Mother’s Smoking Habits** | 0.61855 | |

The assumptions of the model were assessed in the same way as the previous model. For normality, the Shapiro-Wilks normality test returned a p-value of 0.22, so the null hypothesis is not rejected. From the QQ plot of the residuals of the model (Figure) and Shapiro-Wilks test, it cannot be concluded that the data come from a normal distribution. Figure shows the linearity of the model by plotting the residuals against the fitted values. For heteroskedasticity, we use Breusch-Pagan test returned a p value of 0.16 so the null hypothesis is not rejected.

The p-value of the Durbin-Watson test was 0.08, suggesting that there is no correlation of the residuals. It also returned a DW statistic of 1.83. Therefore, the model passed all assumptions.

|  |
| --- |
| **Figure 10: Histogram of Residuals** |

|  |
| --- |
| Figure 11: Normal Q-Q |

|  |
| --- |
| Figure 12: Residuals vs Fitted |

## Other Tested Models

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

The other interaction - effect models tested were:

* **Parity** and **mother’s weight** against **baby weight**
* **Mother’s weight and income** against **baby weight**
* **Smoke and mother’s weight** against baby weight

These models were fitted and their AIC scores were for each were really high as compared to dataModel, hence they were not chosen for the final model selection.

AIC scores for the fitted models were:

* **Parity** and **mother’s weight** against **baby weight** = 10547.07
* **Mother’s weight and income** against **baby weight** = 9494.029
* **Smoke and mother’s weight** against **baby weight =** 10457.45

Although, on further testing we found some interesting results. Despite these AIC scores the models passed the model diagnostic tests.

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

Normality test for the model – Smoke and mother’s weight against baby weight, passes the test.

|  |
| --- |
| Figure 14: Normal Q-Q Plot |

Normality test for the model – Income and mother’s weight against baby weight, passes the test.

|  |
| --- |
| **Figure 15: Normal Q-Q Plot** |

Normality test for the model – Parity and mother’s weight against baby weight, passes the test.

As far as the Durbin-Watson tests are concerned, the p-values for each model were:

* **Parity** and **mother’s weight** against **baby weight** = 0.0468
* **Mother’s weight and income** against **baby weight** = 0.046
* **Smoke and mother’s weight** against **baby weight =** 0.674

Clearly the autocorrelation in these models are either very insignificant or not present at all.

Even for the ncv test the p-values for the models were as follows:

* **Parity** and **mother’s weight** against **baby weight** = 0.03291
* **Mother’s weight and income** against **baby weight** = 0.11154
* **Smoke and mother’s weight** against **baby weight =** 0.43456

The tests show that there is heteroscedasticity in two models but one model does not have it present.

# FIVE-FOLD CROSS VALIDATION

### 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 *k*-Fold Cross Validation (James, Witten, Hastie, & Tibshirani, 2014).

In the cross-validation, the training set is divided into sub-samples, and each single sub-sample will be saved as the data for the verification of model while the other *k-1* groups of sample will be used for training. Cross-validation is repeated *k* times, of which each sub-sample is verified once. The average number of results or other combinations are used, and a single estimate is finally obtained. The advantage of this method is that it repeatedly uses randomly generated sub-samples for training and validation. Actually, 10-Fold Cross Validation is the most commonly used[[4]](#footnote-4).

In our experiment, *k* has a specific value, 5, the reference to the model with be 5-Fold Cross-Validation.

### 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)), in other words,is the average of the square of the errors. MSE satisfies the equation as below:

MSE (*T*) = var (*T*) + (bias (*T*))2

where bias(*T*) = E(*T*) - **

Usually, if the MSE of one model is larger, the error of this model will be larger.

## Predictions

Cross-validation for linear regression can be easily achieved in R Studio. The function ‘cv.lm’ can predict the accuracy for multiple linear regression. By using the function ‘cv.lm’, the output will be displayed, including the analysis of variance table and the observations in five test sets. The whole output can be found in Appendix 2. The output gives the plots of the cross-validation predicted value of two models as below in Figure 15 and 16. It is quite hard to say whether **dataModel** or **finalModel** is the better because the five regression lines all seems parallel in both plots.

Thus, the focus of the output should be the comparison of overall ms (mean square) of both finalModel and dataModel. From the output, the overall ms of finalModel is 268 whilst which of dataModel is 255. It can be predicted that dataModel is a bit more suitable for this case than finalModel.

|  |
| --- |
| Preditedcrossvalidation-final.png  Figure 16: Cross-Validation predicted values for FinalModel |

|  |
| --- |
| Predictedobserveddata.png  Figure 17: Cross Validation Predicted Values for DataModel |

Also, by using get\_mse function, the output turns out that the MSE value of finalModel is 258 while the MSE value of dataModel is 248. Although the discrepancy between these two models is not so large, it is clear that dataModel is better than finalModel in this case.

# 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.

# 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.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications*. (S. Publishing, Ed.) (7th ed.). New York.

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

Pereira, P., Da Mata, F., Figueiredo, A., de Andrade, K., & Pereira, M. (2017). Maternal Active Smoking During Pregnancy and Low Birth Weight in the Americas: A Systematic Review and Meta-Analysis. *Nicotine & Tobacco Research*, *19(5)*, pp.497-505.

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**

**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

**N**

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

**W**

wt = birth weight in ounces

**Appendix 2 - Output**







**Appendix 3 – Data Cleaning**

#load in the data

babies.data <- read.table(file.choose(), header = TRUE)

babies.data

#observations from data set:

# pluralty is always 5

# outcome is always 1

# there are values of 999 for gestation but readme doc does not clarify if

# these are unknown - CLEANED ANYWAY

# all subjects are male

# for race, I'm unsure why white is assigned six values (0-5) - one unknown

# two unknown ages (mother) - CLEANED

# one unknown education (mother) - CLEANED

# many unknown heights (mother) - CLEANED

# many unkown weights (mother) - CLEANED

# five unknown fathers' races as well as values of 10? - 99s CLEANED -

# many unknown fathers' ages - CLEANED

# many unknown fathers' educations - CLEANED

# many unknown fathers' heights - CLEANED

# many unknown fathers' weights - CLEANED

# no explanation of 0 in marital status - assume unknown?

# many unknown incomes - CLEANED

# ten unknown smokers - CLEANED

# nine unknown quitting times, one not asked - CLEANED

# ten unknown number of cigarettes smoked - CLEANED

##### cleaning the data as per unknown values above #####

clean.data <- babies.data

clean.data$gestation[clean.data$gestation == "999"] <- NA

clean.data$age[clean.data$age == "99"] <- NA

clean.data$ed[clean.data$ed == "9"] <- NA

clean.data$ht[clean.data$ht == "99"] <- NA

clean.data$wt[clean.data$wt == "99"] <- NA

clean.data$drace[clean.data$drace == "99"] <- NA

clean.data$dage[clean.data$dage == "99"] <- NA

clean.data$ded[clean.data$ded == "9"] <- NA

clean.data$dht[clean.data$dht == "99"] <- NA

clean.data$dwt[clean.data$dwt == "999"] <- NA

clean.data$inc[clean.data$inc == "98"] <- NA

clean.data$smoke[clean.data$smoke == "9"] <- NA

clean.data$time[clean.data$time == "99"] <- NA

clean.data$time[clean.data$time == "98"] <- NA

clean.data$number[clean.data$number == "98"] <- NA

clean.data$wt.1[clean.data$wt.1 == "999"] <- NA

#make some factors numeric

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 5)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 7)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 10)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 12:13)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 15)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 17:18)

####### Exploration of the birthweight data #######

library(ggplot2)

#Histogram shows that the baby weight values appear to be normally distributed

plot\_hist.wt <- ggplot(data = clean.data, aes(x = wt, y = ..density.. )) +

geom\_histogram(binwidth = (5), colour = "black", fill = "steelblue") +

ggtitle(" Density Histogram of Birth Weight ") +

xlab(" Birth weight in ounces ")+ ylab(" Density ")+ theme\_dark()

plot\_hist.wt

#Create data frame of baby weight summary statistics

BabyWeight <- c(summary(clean.data$wt))

wt.df <- as.data.frame(BabyWeight)

#Investigate how each variable correlates with baby weight

#These show what variables are likely to have an effect on baby weight

#The highest ones are gestation, mother's height (ht), mother's weight (wt.1)

#and father's weight (dwt)

CorrelationValue <- cor(clean.data, clean.data$wt, use = "complete.obs")

cor.df <- as.data.frame(CorrelationValue)

cor.df

#

library(corrplot)

corr\_plot <- corrplot(wt.df,type = "upper", method = "square", insig = "blank",

order = "hclust", tl.col = "black")

#We will now explore each variable in turn

########## Exploration of gestation ##########

##The scatterplot shows data that indicate an increase in birth weight as

#gestation period increases

plot\_gest <- ggplot(clean.data, aes(x = gestation,y = wt)) +

geom\_point(size = 1) +

xlab(" Gestation Period (days) " ) + ylab(" Birth Weight (ounces) ") +

ggtitle(" Gestation Period vs Birth Weight ")

plot\_gest

########## Analysis of ht (mother's height) ##########

##scatterplot of mother's height against baby's weight

##This does not indicate a strong effect between the variables.

scat.mht <- ggplot(clean.data, aes(ht, wt)) +

geom\_point(size = 1, colour = "tomato1") +

xlab(" Mother's Height (inches) " ) + ylab(" Birth Weight (ounces) ") +

ggtitle(" Mother's Height vs Baby´s Birth Weight ")

scat.mht

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

#scatterplot of mother's weight against baby's weight

##This does not indicate a strong effect between the variables.

scat.mwt <- ggplot(clean.data, aes(wt.1, wt)) +

geom\_point(size = 1, colour = "navyblue") +

xlab(" Mother's Weight (pounds) " ) + ylab(" Birth Weight (ounces) ") +

ggtitle(" Mother's Weight and Birth Weight ")

scat.mwt

########## Analysis of dwt (father's weight) ##########

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

##This does not indicate a strong relationship between the variables.

scat.dwt <- ggplot(clean.data, aes(dwt, wt)) +

geom\_point(size = 1, colour = "darkgreen") +

xlab(" Father's Weight (pounds) " ) + ylab(" Birth Weight (ounces) ") +

ggtitle(" Father's Weight and Baby Birth Weight ")

scat.dwt

########## Exploration of smoke ##########

#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

#nterquartile range of the other levels of smoking

#Therefore, the effect may not be large.

#Creating labels for the x axis

smoke.box.xlabels <- c("Never", "Smokes now", "Smoked until pregnancy",

"Once smoked", "Unknown")

smoke.box <- ggplot(clean.data, aes(factor(smoke), wt)) +

geom\_boxplot(fill = "seagreen4") +

labs(title = "Birth Weight per Level of Mother's Smoking",

x = "Smoked or not", y = "Babies' weight") +

scale\_x\_discrete(labels= smoke.box.xlabels) +

theme(axis.text.x=element\_text(angle=15, hjust=1))

smoke.box

**Appendix 4 - modelselection-Su.r**

  #setwd("~/Masters/")

library(tidyverse)

library(ggplot2)

library(car)

library(GGally)

library(effects)

#setwd("~/Masters/")

babies.data <- read.table("babies23.data", header = TRUE)

#since we are working in our directory, I change the directory that I think that

#people use this project can run it.

#observations from data set:

# pluralty is always 5

# outcome is always 1

# there are values of 999 for gestation but readme doc does not clarify if

# these are unknown - CLEANED ANYWAY

# all subjects are male

# for race, I'm unsure why white is assigned six values (0-5) - one unknown

# two unknown ages (mother) - CLEANED

# one unknown education (mother) - CLEANED

# many unknown heights (mother) - CLEANED

# many unkown weights (mother) - CLEANED

# five unknown fathers' races as well as values of 10? - 99s CLEANED -

# many unknown fathers' ages - CLEANED

# many unknown fathers' educations - CLEANED

# many unknown fathers' heights - CLEANED

# many unknown fathers' weights - CLEANED

# no explanation of 0 in marital status - assume unknown?

# many unknown incomes - CLEANED

# ten unknown smokers - CLEANED

# nine unknown quitting times, one not asked - CLEANED

# ten unknown number of cigarettes smoked - CLEANED

##### cleaning the data as per unknown values above #####

clean.data <- babies.data

clean.data$gestation[clean.data$gestation == "999"] <- NA

clean.data$age[clean.data$age == "99"] <- NA

clean.data$ed[clean.data$ed == "9"] <- NA

clean.data$ht[clean.data$ht == "99"] <- NA

clean.data$wt[clean.data$wt == "99"] <- NA

clean.data$drace[clean.data$drace == "99"] <- NA

clean.data$dage[clean.data$dage == "99"] <- NA

clean.data$ded[clean.data$ded == "9"] <- NA

clean.data$dht[clean.data$dht == "99"] <- NA

clean.data$dwt[clean.data$dwt == "999"] <- NA

clean.data$inc[clean.data$inc == "98"] <- NA

clean.data$smoke[clean.data$smoke == "9"] <- NA

clean.data$time[clean.data$time == "99"] <- NA

clean.data$time[clean.data$time == "98"] <- NA

clean.data$number[clean.data$number == "98"] <- NA

clean.data$wt.1[clean.data$wt.1 == "999"] <- NA

#make some factors numeric

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 5)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 7)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 10)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 12:13)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 15)

clean.data <- clean.data %>% mutate\_each(funs(as.numeric), 17:18)

####### Exploration of the birthweight data #######

#normally distributed

hist(clean.data$wt)

summary(clean.data$wt)

##################################

clean.data.naomit <- na.omit(clean.data)

# select data that does not contain id and data of birth

# consider this two factor does not have effect on baby birth weight

# on the real life

clean.data.naomit <- clean.data.naomit %>% dplyr::select(-id, -date)

#factor(clean.data.naomit$id)

dataModel <- lm(wt ~., data = clean.data.naomit)

summary(dataModel)

#try to use Anova

Anova(dataModel)

#model selection use AIC

dataModel <- step(dataModel)

Anova(dataModel)

#check about normality of dataModel's residual

qqnorm(resid(dataModel))

qqline(resid(dataModel))

#the qq plot looks great but the shapiro test, p value is large than 0.05,

# so the residual of the data Model is normal

shapiro.test(resid(dataModel))

hist(resid(dataModel))

# we track down the extreme residuals

bigResid <- which(abs(resid(dataModel))>5)

clean.data.naomit[bigResid,]

#plot residuals against fitted values

dataResid <- resid(dataModel)

plot(fitted(dataModel),dataResid, ylab= "Residuals", xlab = "Fitted Values")

#it looks good

#https://onlinecourses.science.psu.edu/stat501/node/277/

# do Breusche-Pagan test with respect to fitted model

ncvTest(dataModel)

# null hypothesis: constant error variance. "If we have constant error variance

#then the variation in the residuals should be unrelated to any coveriant."

# null hypothesis is rejected since the p value is less than 0.05

#MT5761 notes page 22

# need to write durbinWatsonTest on model

durbinWatsonTest(dataModel)

#null hypothesis: error are uncorrelated, fail to reject the null hypothesis

plot(dataModel, which = 1:2)

#collinearity

numericOnly <- clean.data.naomit %>% select\_if(is.numeric)

#use with caution, picture is sooo huge and difficult to generate

# and do harm to my computer and not useful because we have sooo many variabales

#ggpairs(numericOnly)

vif(dataModel)

# all number is less than 10, do not have to delete any variable

#calculate confidence interval of the model

confint(dataModel)

#add more effect plot if you want and select variable that you

# think is interested

#plot(effect(term="gestation", mod = dataModel))

#plot(effect(term="smoke", mod = dataModel))

#plot(effect(term="number", mod = dataModel))

cols\_to\_change = c(1, 2, 3, 4,6, 8, 9, 11, 14, 16, 19, 20:23)

for(i in cols\_to\_change){

class(clean.data[, i]) = "factor"

}

cols\_to\_change

#create a first order iteraction for every variable

firstorderModel <- lm(wt ~.\*., data = numericOnly)

summary(firstorderModel)

#model selection use AIC

firstorderModel <- step(firstorderModel)

summary(firstorderModel)

Anova(firstorderModel)

qqnorm(resid(firstorderModel))

qqline(resid(firstorderModel))

shapiro.test(resid(firstorderModel))

hist(resid(firstorderModel))

firstorderResid <- resid(firstorderModel)

plot(fitted(firstorderModel),firstorderResid, ylab= "Residuals", xlab = "Fitted Values")

ncvTest(firstorderModel)

durbinWatsonTest(firstorderModel)

plot(firstorderModel, which = 1:2)

# we exam the collinearity of the firstorderModel we find that there are a lot of

# variable that its GVIF number is larger than 10, so in the following step.

# 1. we find the maximum number of GVIF, if it is larger than 10,remove it

# 2. do the vif function again to check the collinearity and get the maximum repeat the step 1

# we do the above two steps until all the variable's collinearity GVIF is less than 10

# or we do not have a collinearity problem anymore

# following just the process of removing every variable that is collinear

k<-vif(firstorderModel)

k[which.max(k)]

alteredModel <-update(firstorderModel,.~.-ht:marital )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-race )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-smoke )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dht:race)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dage)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-age:marital)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-drace)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dht:inc)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-gestation:number)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-wt.1)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-ht:smoke)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-marital:dage )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-ed )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-parity )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-age:dwt )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-marital:race )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-age:race )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dwt:wt.1 )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-gestation:drace )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-ded:dwt )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dwt:dage )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-gestation:smoke )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-ded:time )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-marital:ed )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dage:race )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dwt:ed )

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-gestation:parity)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-ed:smoke)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-age:drace)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dwt:race)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-dwt:smoke)

p<-vif(alteredModel)

p[which.max(p)]

alteredModel <-update(alteredModel,.~.-inc:ed)

p<-vif(alteredModel)

p[which.max(p)]

#finally, we finish deleting collinear variable and we do a AIC do a backward

#model selection and get the finalModel

finalModel <- step(alteredModel)

#check final model colinearity and all of them are less than 10, it works.

vif(finalModel)

#get summary of finalModel

summary(finalModel)

#use qq plot and Shapiro-Wilk normality test to test the normality

# because the p value in Shapiro-Wilk normality test is larger than 0.05,

# the data is normal, the QQ plot show the same result

qqnorm(resid(finalModel))

qqline(resid(finalModel))

shapiro.test(resid(finalModel))

hist(resid(finalModel))

plot(finalModel, which = 1:2)

# do Breusche-Pagan test with respect to fitted model

ncvTest(finalModel)

# null hypothesis: constant error variance. "If we have constant error variance

#then the variation in the residuals should be unrelated to any coveriant."

# null hypothesis is rejected since the p value is less than 0.05

# need to write durbinWatsonTest on model

durbinWatsonTest(finalModel)

#null hypothesis: error variances are uncorrelated, fail to reject the null hypothesis

#MT5761 notes page 22

Anova(finalModel)

#get the confidence interval

confint(finalModel)

**Appendix 5 - bootstrapping**

#load boot library

library(boot)

#PURPOSE: A bootstrapping function which generates 95% confidence intervals for

#regression coefficients when used as the 'statistic' argument in the function

#boot()

#INPUTS: The linear model, the data from which the model comes,

#the index parameters

#OUTPUT: The coefficients of the linear regression model

bst <- function(formula, data, indices){

d <- data[indices, ]

fit <- lm(formula, data=d)

return(coef(fit))

}

#The bootstrapping results are stored as 'results'

#1500 replications is the fewest that allow the boot() function to run

#I do not know why that is

results <- boot(data = clean.data, statistic = bst, R = 1250, formula = dataModel)

#View results as density histogram and qqplot

#The data are normally distributed for all variables

results

plot(results, index=1) # intercept

plot(results, index=2) # gestation

plot(results, index=3) # parity

plot(results, index=4) # ht

plot(results, index=5) # drace

plot(results, index=6) # dwt

plot(results, index=7) # smoke

plot(results, index=8) # number

# Get 95% confidence intervals

boot.ci(results, type="bca", index=1) # intercept

boot.ci(results, type="bca", index=2) # gestation

boot.ci(results, type="bca", index=3) # parity

boot.ci(results, type="bca", index=4) # ht

boot.ci(results, type="bca", index=5) # drace

boot.ci(results, type="bca", index=6) # dwt

boot.ci(results, type="bca", index=7) # smoke

boot.ci(results, type="bca", index=8) # number

finalModel.results <- boot(data = clean.data.naomit, statistic = bst, R = 1500, formula = finalModel)

# view results as density histogram and qqplot

finalModel.results

plot(finalModel.results, index=1) #intercept

plot(finalModel.results, index=2) # gestation

plot(finalModel.results, index=3) # ht

plot(finalModel.results, index=4) # ded

plot(finalModel.results, index=5) # dwt

plot(finalModel.results, index=6) # inc

plot(finalModel.results, index=7) # time

plot(finalModel.results, index=8) # number

plot(finalModel.results, index=9) # inc:parity

plot(finalModel.results, index=10) # wt.1:ed

plot(finalModel.results, index=11) # ed:drace

plot(finalModel.results, index=12) # time:ed

plot(finalModel.results, index=13) # ded:smoke

# get 95% confidence intervals

boot.ci(finalModel.results, type="bca", index=1) # intercept

boot.ci(finalModel.results, type="bca", index=2) # gestation

boot.ci(finalModel.results, type="bca", index=3) # ht

boot.ci(finalModel.results, type="bca", index=4) # ded

boot.ci(finalModel.results, type="bca", index=5) # dwt

boot.ci(finalModel.results, type="bca", index=6) # inc

boot.ci(finalModel.results, type="bca", index=7) # time

boot.ci(finalModel.results, type="bca", index=8) # number

boot.ci(finalModel.results, type="bca", index=9) # inc:parity

boot.ci(finalModel.results, type="bca", index=10) # wt.1:ed

boot.ci(finalModel.results, type="bca", index=11) # ed:drace

boot.ci(finalModel.results, type="bca", index=12) # time:ed

boot.ci(finalModel.results, type="bca", index=13) # ded:smoke

**Appendix 6 – Cross-Validation Test**

#install package "DAAG"

library(DAAG)

# 5 fold cross-validation for finalModel

# with both Observed and Residual

cv.lm(clean.data.naomit, form.lm = finalModel, plotit = "Observed", m=5)

cv.lm(clean.data.naomit, form.lm = finalModel, plotit = "Residual", m=5)

# 5 fold cross-validation for dataModel

# with both Observed and Residual

cv.lm(clean.data.naomit, form.lm = dataModel, plotit = "Observed", m=5)

cv.lm(clean.data.naomit, form.lm = dataModel, plotit = "Residual", m=5)

# From the plots, we cannot say if dataModel or finalModel is better

# because the five regression lines all seems parallal in both plots

# However from the output, the overall ms of dataModel is 255 whilst which of finalModel is 268

# which means finalModel is little bit better than dataModel

#################

# MSE for finalModel and dataModel

# model with smaller MSE is better

library(dvmisc)

get\_mse(finalModel,var.estimate = FALSE)

get\_mse(dataModel, var.estimate = FALSE)

# MSE of finalModel is 258 and dataModel is 248

# So, dataModel seems better than finalModel

1. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3016.1988.tb00218.x> [↑](#footnote-ref-1)
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=1> [↑](#footnote-ref-2)
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=2> [↑](#footnote-ref-3)
4. <https://machinelearningmastery.com/k-fold-cross-validation/> [↑](#footnote-ref-4)