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

Word Count: 2,752

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. The data used in this report are a part of a larger group of studies from the Child Health and Development Studies (CHDS). Models were built, tested and selected using linear regression, analysis of variance (ANOVA), the Akaike Information Criterion (AIC) and bootstrapping. Five-fold cross validation was used for further analysis and prediction. The selected model states that increases in gestation period, the number of the mother's previous pregnancies, the mother's height and the father's weight cause an increase in birth weight. It also states that birth weight increase with a reduction in the number of cigarettes smoked per day by the mother.

INDEX

Contents

1 INTRODUCTION 1
2 METHODS 2
2.1 Data Cleaning 2
2.2 Data Exploration 2
2.3 Model Fitting 2
2.4 Bootstrapping 3
2.5 Five-Fold Cross Validation 3
3 RESULTS 3
3.1 Data Exploration 3
3.1.1 Correlation of All Variables with Birth Weight 3
3.1.2 Gestation Period 4
3.1.3 Mother's Height 5
3.1.4 Mother's Weight 6
3.1.5 Father's Weight 7
3.1.6 Mother´s Smoking Habits 7
3.2 Model Fitting 9
3.2.1 Model A 9
3.2.2 Bootstrapping Model A 11
3.2.3 Model B 12
3.2.4 Bootstrapping Model B 16
3.3 Other Tested Models 16
4 FIVE-FOLD CROSS VALIDATION 19
4.1.1 Five-Fold Cross-Validation 19
4.1.2 Mean Square Error (MSE) 20
4.2 Predictions 20
5 DISCUSSION 21
6 BIBLIOGRAPHY 23
7 APPENDICES 24

1 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].

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].

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*

[1] https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3016.1988.tb00218.x

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=1

not only of the higher birth rate in these countries but also of their LBW[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).

2 METHODS

2.1 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.

2.2 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.

2.3 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

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491072/?page=2

values. Constant variance was tested by performing a Breusch-Pagan test and a Durbin-Watson test was performed to test for autocorrelation.

2.4 Bootstrapping

Bootstrapping was performed on the selected models to obtain 95% confidence intervals for the regression coefficients. Bootstrapping is a resampling method that involves taking repeated samples from the same data set with replacement. This generates a number of samples that can be used for further analyses.

2.5 Five-Fold Cross Validation

Five-fold cross validation was performed on models A and B to further assess which model should be selected. The process is explained in the Results section.

3 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).

3.1 Data Exploration

3.1.1 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 1.

	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**

3.1.2 Gestation Period

A scatterplot of gestation period and birthweight was created to visualise the relationship. As can be seen in Figure 2, 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.

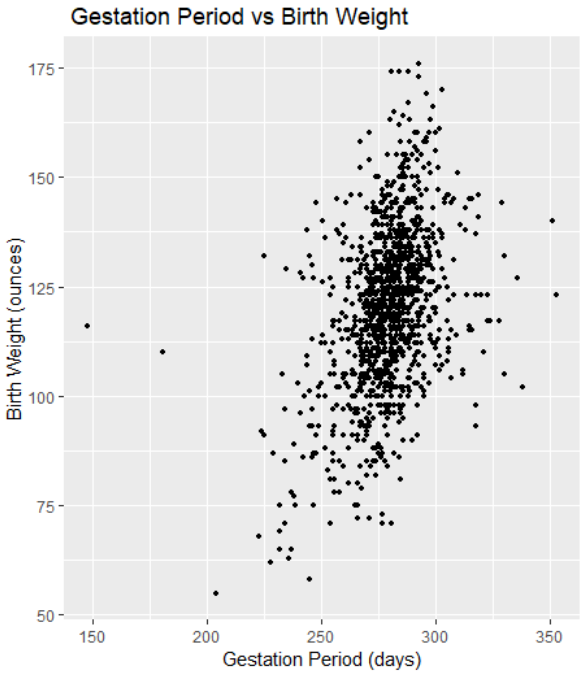

Figure 2: Scatterplot of gestation period against baby weight

3.1.3 Mother's Height

The second-strongest correlation with birth weight was seen with mother's height (correlation of 0.22). Figure 3 represents this as a scatterplot but does not indicate a strong relationship between the variables.

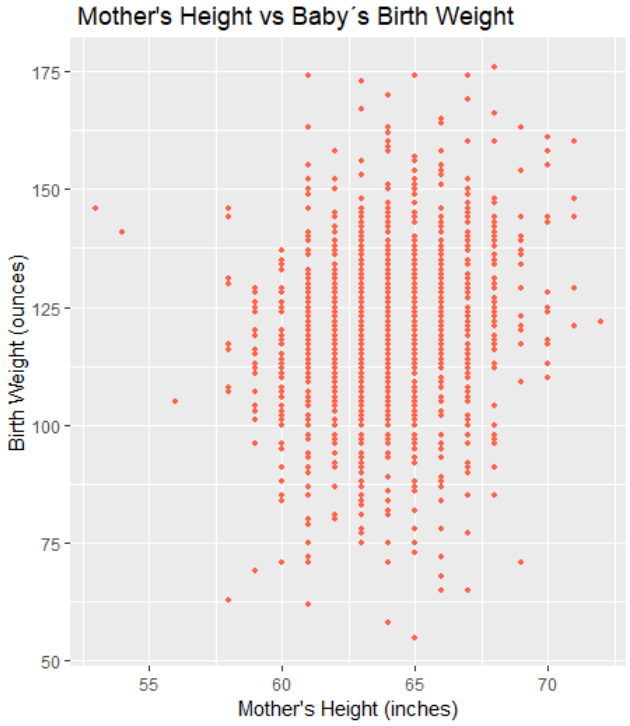

Figure 3: Scatterplot of Mother´s Height vs. Birth Weight

3.1.4 Mother's Weight

Figure 4 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.

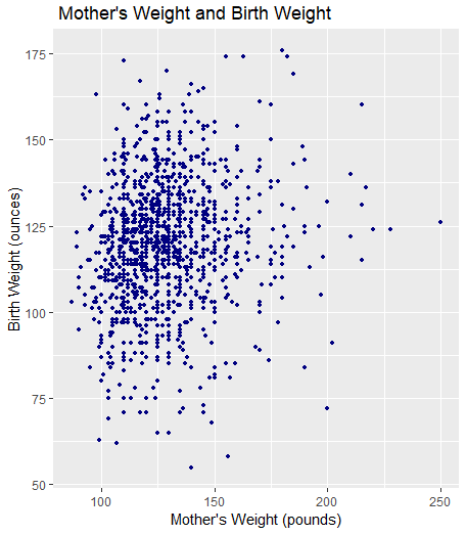

Figure 4: Scatterplot of mother´s weight and birth weight

3.1.5 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 5 shows the relationship as a scatterplot and does not indicate a large effect.

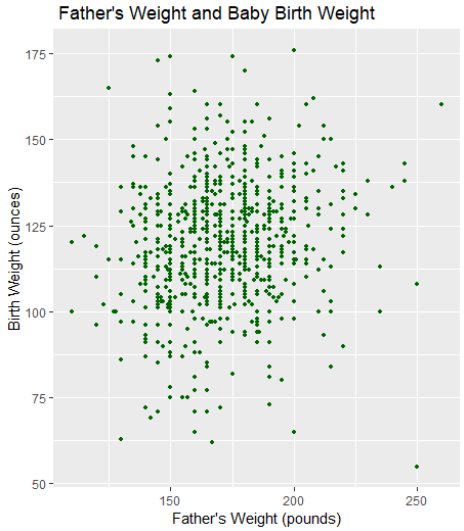

Figure 5: A scatterplot of father's weight against birth weight

3.1.6 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 6). 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.

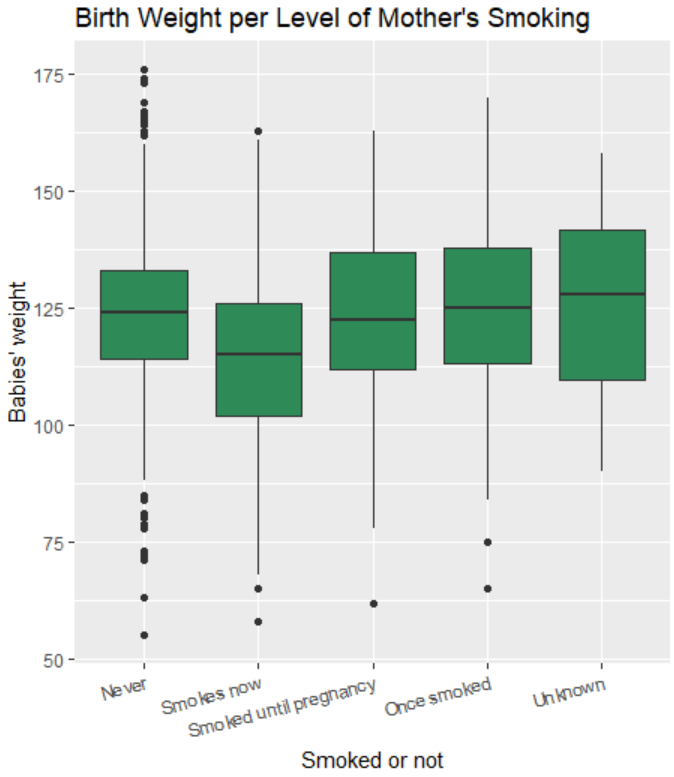

Figure 6: Birth weight per level of mother´s smoking habits

3.2 Model Fitting

3.2.1 Model A

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 9). 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.

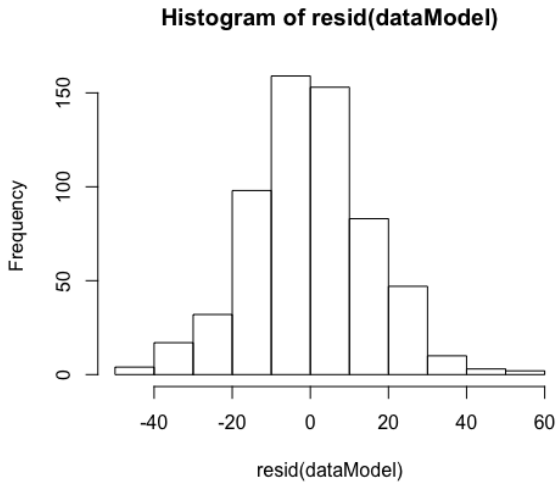

Figure 7: Histogram of Residuals

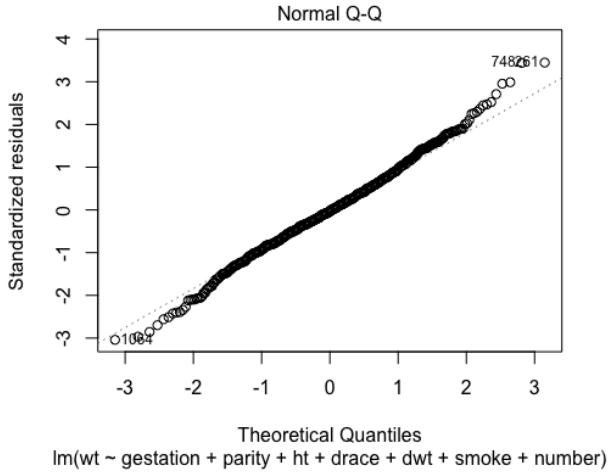

Figure 8: Normal Q-Q

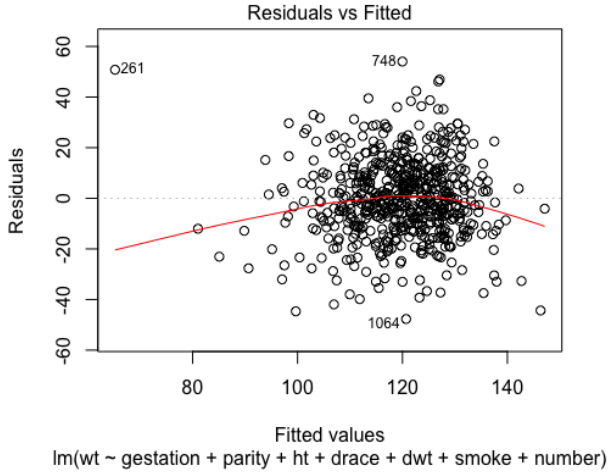

Figure 9: Residuals vs Fitted

3.2.2 Bootstrapping Model A

To obtain 95% confidence intervals for the regression coefficients, bootstrapping was performed on the both models. These obtained the results shown in Table 2.

Table 2: **95% confidence intervals for the regression coefficients**

Variable	2.5% CI	97.5% CI
(intercept)	-134.03	-52.43
Gestation period	0.3192	0.5501
Mother's Previous Pregnancies	0.0258	1.3838
Mother's Height	0.801	1.759
Father's Race	-1.0259	-0.2617
Father's Weight	0.016	0.1334
Mother's Smoking Habits	0.375	3.41
No. of Cigarettes Smoked by Mother	-2.632	-1.335

3.2.3 Model B

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 3).

Table 3: **Coefficients of the variables in the Model A**

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 11) and Shapiro-Wilks test, it cannot be concluded that the data come from a normal distribution. Figure 12 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.

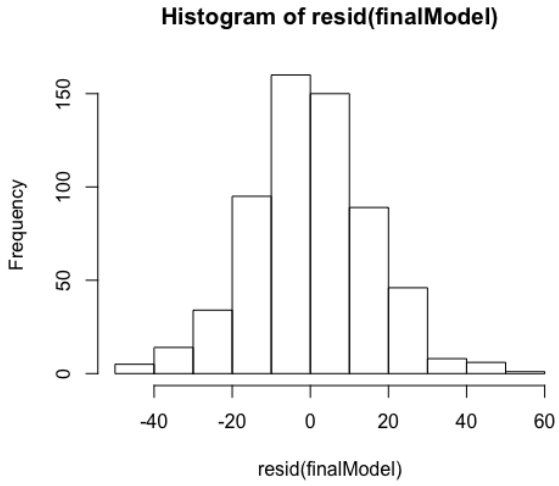

Figure 10: Histogram of Residuals

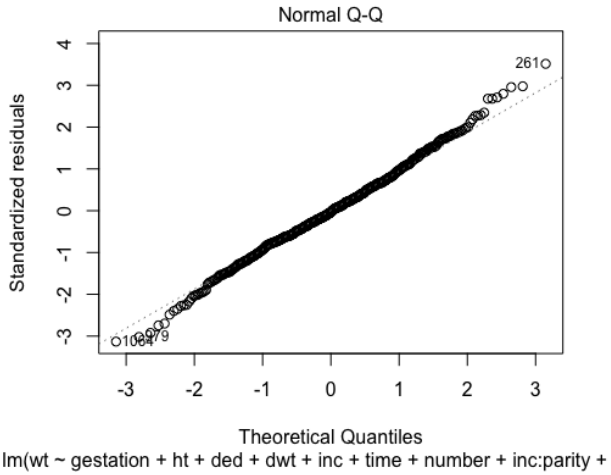

Figure 11: Normal Q-Q

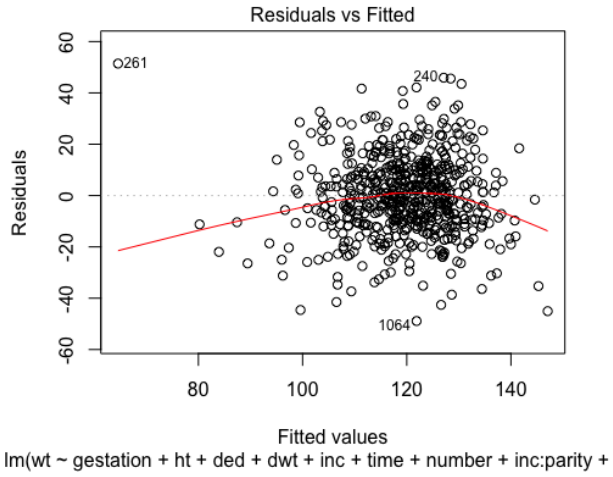

Figure 12: Residuals vs Fitted

3.2.4 Bootstrapping Model B

Bootstrapping was performed on the model to obtain 95% confidence intervals for the regression coefficients. These are shown in Table 4.

Table 4: **Coefficients of the variables in Model B**

Variable	2.5% CI	97.5% CI
(Intercept)	-126.81	-40.85
Gestation period	0.3196	0.5711
Mother's Height	0.512	1.635
Father's Education	-2.61	-0.204
Father's Weight	0.0084	0.1412
Family Yearly Income	-1.1509	0.1632
Time Since Mother Quit Smoking	0.308	3.777
No. of Cigarettes Smoked by Mother	-2.735	-1.377
Family Yearly Income:Mother's Previous Pregnancies	-0.0253	0.3242
Mother's Weight:Mother's Education	0.0063	0.0275
Mother's Education:Father's Race	-0.3166	-0.0928
Time Since Mother Quit Smoking:Mother's Education	-1.2099	-0.1072
Father's Education:Mother's Smoking Habits	-0.0375	1.3283

3.3 Other Tested Models

Additional interaction-effect models were fitted to observe other effects that variables had on birth weight. However, the AIC scores from these models higher were higher than that of Model A. Therefore, they were not chosen for the final model selection. The models and their AIC scores were as follows:

- **Mother's previous pregnancies** and **mother's weight** against **birth weight**
  - AIC = 10547.07

- **Mother's weight and family yearly income** against **birth weight**
  - AIC = 9494.029
- **Mother's smoking habits and mother's weight** against **birth** weight
  - AIC = 10457.45

Despite these AIC scores the models passed the all model diagnostic tests and assumption checking, indicating that an effect between the variables selected exists.

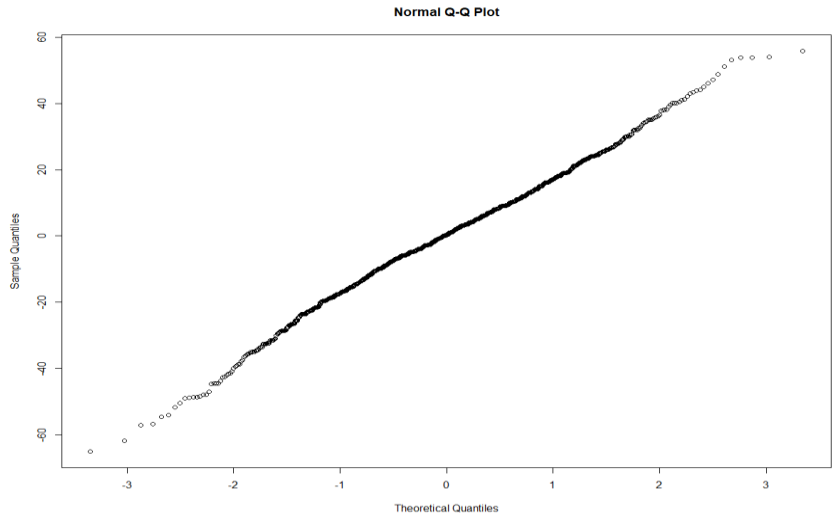

Figure 13: Normal Q-Q plot

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

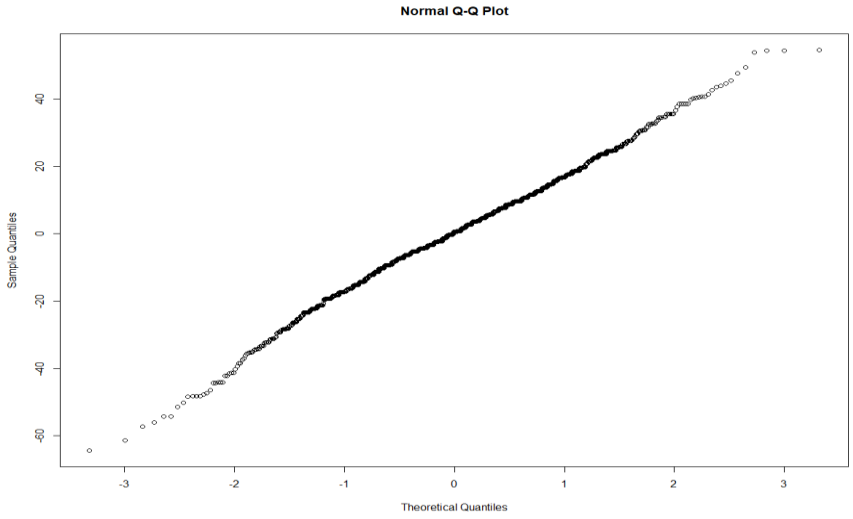

Figure 14: Normsl Q-Q Plot

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

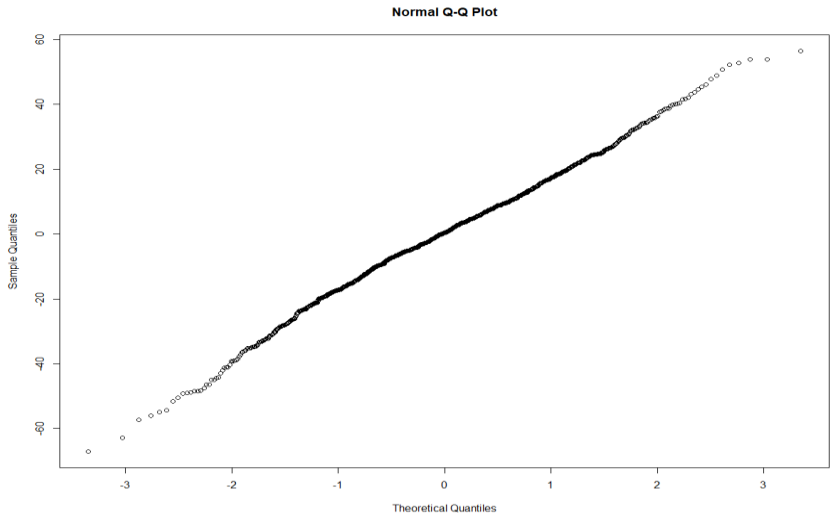

Figure 15: Normal Q-Q Plot

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

The autocorrelation in these models are either insignificant or not present at all.

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.

4 FIVE-FOLD CROSS VALIDATION

4.1.1 Five-Fold Cross-Validation

Model selection has been conducted through examining the AIC score. However, it can be difficult to determine if these improvements in scores result from the captures of better relationships within our model or if the model is being overfitted. To clarify this aspect, *k*-Fold Cross Validation (James, Witten, Hastie, & Tibshirani, 2014) is used.

In the cross-validation, the training set is divided into sub-samples, and each sub-sample is saved as the data for the verification of model while the other *k-1* groups of samples are 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 obtained. The advantage of this method is that it repeatedly uses randomly generated sub-samples for training and validation. 10-Fold Cross Validation is the most commonly used[4].

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

[4] https://machinelearningmastery.com/k-fold-cross-validation/

4.1.2 Mean Square Error (MSE)

Mean Square Error (MSE) is used to evaluate the quality of an estimator (parameter) or a predictor (some random variable). In other words, it is the average of the square of the errors. MSE satisfies the equation as below:

$$\text{MSE}(T) = \text{var}(T) + (\text{bias}(T))^2$$

$$\text{where bias}(T) = \text{E}(T) - \theta$$

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

4.2 Predictions

The results of the 5-fold cross validation are shown in Figures 16 and 17 (the whole output is in Appendix 2). They display the plots of the cross-validation predicted value of models A and B. It is hard to judge whether Model A or Model B is better because the five regression lines seem parallel in both plots. Thus, the focus of the output should be the comparison of overall MS (mean square) of both Model A and Model B. From the output, the overall MS of Model B is 268 whilst for Model A it is 255. Therefore, it can be predicted that Model A is more suitable for this case than Model B.

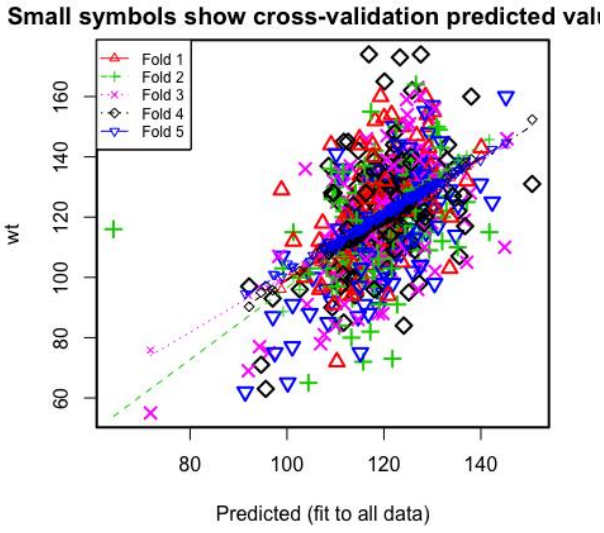

Figure 16: Cross-Validation predicted values for FinalModel

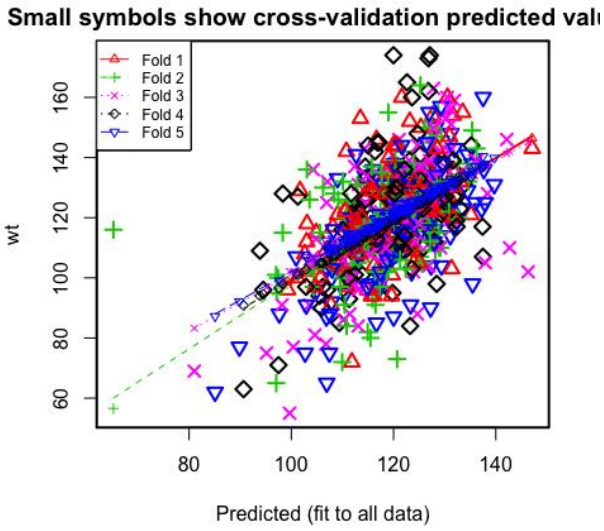

Figure 17: Cross Validation Predicted Values for DataModel

By using the 'get_mse' function of R-Studio, the MSE value of Model B is 258 while the MSE value of Model A is 248. Although the difference between these values is not large, it can be assessed that Model A is better than Model B in this case.

5 DISCUSSION

Models were fitted to understand the effects of certain variables on birth weight from the given data. The final model chosen to make conclusions was Model A. This model states that the parameters affecting baby weight are gestation period, mother's previous pregnancies, mother's height, father's race, father's weight, mother's smoking habits, no. of cigarettes smoked by mother.

The effects of the variables in Model A on birth weight are as follows:

- As the gestation period increases by unit, birth weight increases by 0.454 ounces.
- As the number of previous pregnancies increases, birth weight has shown to increase by 0.7496 ounces.
- As the mother's height increases by unit, birth weight increases by 1.2696 ounces.
- As father's weight increases by unit, birth weight increases by 0.07689 ounces.
- As the number of cigarettes smoked per day of the mother reduces, birth weight has shown to increase.

- The categorical variable of father's race has shown to have a positive increase on birth weight

It can be concluded from the model that taller mothers and heavier fathers give birth to heavier babies. Also, birth weight increases with increases in gestation period and increases in the number of previous pregnancies. The model also shows that birth weight is reduced by the number of cigarettes smoked per day by the mother.

The categorical variable of mother's smoking habits has shown to have a positive increase on baby weight. By statistical observations we can primarily see that if the mother has never smoked or if the mother smokes now, the baby's weight is higher. This finding contradicts previous work by Pereira, Da Mata, Figueiredo, de Andrade, & Pereira (2017) who found that smoking during pregnancy decreases birth weight. It is recommended that this finding is investigated further.

6 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.

7 APPENDICES

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 of Five-Fold Cross Validation

```
<html><head></head><body><pre style="word-wrap: break-word; white-space: pre-wrap;">###Final
Model
Analysis of Variance Table

Response: wt
              Df Sum Sq Mean Sq F value  Pr(&gt;F)
gestation      1  32816   32816  127.17 &lt; 2e-16 ***
age            1    810     810    3.14   0.077 .
ht             1   8685    8685   33.66 1.1e-08 ***
dht            1    223     223    0.87   0.353
dwt            1   1693    1693    6.56   0.011 *
gestation:age  1   3963    3963   15.36 9.9e-05 ***
gestation:dht  1    769     769    2.98   0.085 .
Residuals    600 154825     258
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



fold 1
Observations in test set: 121
               22    24  27    30     34     44    47    76    77    119 124    125 144
162    175
Predicted   121.56 126.0 115 119.9 113.96 113.76 130.2 113.8 120.0 114.93 131 122.66 118
116.17 122.88
cvpred      120.41 124.7 115 118.9 113.75 113.46 129.3 113.3 118.8 113.61 130 121.55 117
116.03 122.12
wt          115.00 122.0 146 114.0 119.00 111.00 155.0 127.0 153.0 121.00 121 117.00 144
111.00 120.00
CV residual  -5.41  -2.7  31  -4.9   5.25  -2.46  25.7  13.7  34.2   7.39  -9  -4.55  27
-5.03  -2.12
              177    179   184 190    222 238   244   245    257    265   309    316    327
342
Predicted   121.1 121.24 121.8 107 114.17 116 128.9 123.7 120.88 113.22 115.3 111.65 123.30
122.94
cvpred      119.6 121.07 121.4 107 113.59 115 128.7 122.5 120.46 111.99 113.9 111.31 121.75
121.69
wt          134.0 112.00 124.0  96 121.00 126 154.0 150.0 128.00 110.00 101.0 104.00 117.00
129.00
CV residual  14.4  -9.07   2.6 -11   7.41  11  25.3  27.5   7.54  -1.99 -12.9  -7.31  -4.75
7.31
              345    347    367     369   381     402   403    404   441   455     480    496
499
Predicted   123.7 123.50 129.21 128.252 114.2 119.619 107.4 127.90  98.9 121.9 127.239 117.88
125.78
cvpred      122.4 123.14 128.19 128.739 113.2 119.476 107.9 128.18  96.0 121.3 126.682 117.35
124.32
wt          105.0 122.00 135.00 129.000  96.0 120.000 118.0 127.00 129.0 142.0 127.000 114.00
132.00
CV residual -17.4  -1.14   6.81   0.261 -17.2   0.524  10.1  -1.18  33.0  20.7   0.318  -3.35
7.68
               504   511   514   535    537 544    548   570    585    586   600   607   608
620
Predicted   123.32 121.6 116.5 115.1 117.83 121 119.93 118.3 112.38 110.16 112.3 119.5 125.2
114.75
cvpred      122.57 120.5 116.1 114.6 117.08 120 117.68 117.6 111.95 110.25 111.4 118.7 124.2
114.32
wt          129.00 154.0 129.0 134.0 123.00 109 111.00 128.0 115.00 103.00  97.0 105.0 135.0
123.00
CV residual   6.43  33.5  12.9  19.4   5.92 -11  -6.68  10.4   3.05  -7.25 -14.4 -13.7  10.8
8.68
               625   639    647   656   664    670   685   696    738   760    780    786
793    799
Predicted   106.40 118.2 106.65 117.1 133.6 123.48 101.3 129.1 123.96 122.0 116.26 111.75
125.5 112.26
cvpred      106.68 117.6 104.89 115.9 133.7 121.89 101.8 126.9 122.65 121.5 115.59 111.32
125.4 111.52
wt          108.00 152.0 112.00 129.0 103.0 116.00 112.0 150.0 129.00 110.0 110.00 110.00
123.0 117.00
CV residual   1.32  34.4   7.11  13.1 -30.7  -5.89  10.2  23.1   6.35 -11.5  -5.59  -1.32
-2.4   5.48
               800   817   850    865    877    885    904   905    914    930 932    934
953    957
Predicted   120.55 121.4 119.3 123.25 124.95 119.37 120.60 110.3 112.79 113.84 113 121.34
```

```
130.57 126.60
cvpred      119.69 120.9 118.1 122.02 124.26 119.04 120.64 110.4 111.93 113.35 113 119.93
132.88 125.67
wt          125.00 125.0 160.0 113.00 122.00 123.00 115.00  72.0 109.00 104.00  94 112.00
123.00 130.00
CV residual   5.31   4.1  41.9  -9.02  -2.26   3.96  -5.64 -38.4  -2.93  -9.35 -19  -7.93
-9.88   4.33
             960    962    981   986    988   990    992    999  1000   1007  1009 1023  1025
1027
Predicted   127 119.20 122.41 117.9 116.29 115.0 114.97 113.27 115.2 129.72 129.3  112 109.1
140.09
cvpred      126 118.24 121.23 118.5 115.84 114.3 113.77 113.29 114.3 129.19 128.2  112 105.8
140.39
wt          147 122.00 123.00 104.0 118.00 144.0 117.00 110.00 100.0 137.00 143.0   94 144.0
143.00
CV residual  21   3.76   1.77 -14.5   2.16  29.7   3.23  -3.29 -14.3   7.81  14.8  -18  38.2
2.61
              1035  1056  1057  1062  1079   1094   1097   1101   1102  1111   1117 1122
1126  1130
Predicted   128.6 109.7 116.8 120.9 116.8 111.74 122.64 124.85 111.54 118.4 112.59  116
119.60 117.6
cvpred      129.5 109.9 116.5 120.7 115.9 109.65 121.39 123.62 111.66 117.4 112.42  116
118.12 117.5
wt          160.0  91.0 129.0  94.0 126.0 108.00 123.00 129.00 119.00 129.0 120.00  128
128.00 140.0
CV residual  30.5 -18.9  12.5 -26.7  10.1  -1.65   1.61   5.38   7.34  11.6   7.58   12
9.88  22.5
              1132   1137   1139   1144   1166   1188  1199  1201  1231
Predicted   103.7 119.25 112.16 122.11 109.72 137.09 133.8 107.2 120.2
cvpred      103.5 118.87 112.23 121.28 109.84 136.79 133.1 105.5 119.1
wt          100.0 122.00 114.00 117.00 112.00 132.00 120.0  97.0 132.0
CV residual  -3.5   3.13   1.77  -4.28   2.16  -4.79 -13.1  -8.5  12.9

Sum of squares = 26263    Mean square = 217    n = 121

fold 2
Observations in test set: 122
                1    18     29    32     41  50     57     64    70    81    98   107    116
120
Predicted   113.46 101.3 125.89 113.5 119.29 122 128.55 104.08 118.1 131.3 137.1 129.6 116.55
102.62
cvpred      114.31  99.8 128.31 114.1 121.91 124 128.89 104.61 118.1 134.5 140.4 131.1 116.75
95.88
wt          120.00 115.0 125.00  93.0 129.00 145 124.00 101.00 142.0 149.0 125.0 155.0 118.00
100.00
CV residual   5.69  15.2  -3.31 -21.1   7.09  21  -4.89  -3.61  23.9  14.5 -15.4  23.9   1.25
4.12
               122   160    164    206    208 221    234   248    258   259   261    273
274    282
Predicted   124.16 117.3 128.08 112.54 128.13 121 120.82 111.8 118.35 127.8  64.2 133.64
112.46 114.14
cvpred      124.86 118.6 130.21 113.14 128.61 123 120.94 106.9 118.93 129.4  47.9 135.62
112.19 115.46
wt          118.00 134.0 135.00 109.00 119.00 107 125.00 122.0 125.00 114.0 116.0 127.00
104.00 113.00
CV residual  -6.86  15.4   4.79  -4.14  -9.61 -16   4.06  15.1   6.07 -15.4  68.1  -8.62
-8.19  -2.46
              290    294   301 321    324     325    334    339    341    348   396    411
413   415
Predicted   120.29 127.27 113.3 123 108.59 135.445 123.52 109.44 119.11 129.11 129.5 131.67
123.8 121.5
cvpred      121.21 127.18 113.4 123 106.68 137.397 124.28 108.68 119.66 129.67 130.9 133.18
123.8 121.9
wt          117.00 121.00  80.0 110 104.00 138.000 120.00 111.00 116.00 133.00 115.0 132.00
138.0 132.0
CV residual  -4.21  -6.18 -33.4 -13  -2.68   0.603  -4.28   2.32  -3.66   3.33 -15.9  -1.18
14.2  10.1
               422    432    436    447   451   465    495    524    531    558 559    561
581   605
Predicted   127.51 116.18 123.28 119.00 111.6 120.7 126.74 104.63 119.86 112.67 119 111.34
123.26 127.1
cvpred      127.72 116.85 123.28 119.69 109.6 120.3 128.07 101.42 120.02 111.13 121 109.37
123.84 128.7
wt          119.00 112.00 122.00 113.00 120.0 109.0 130.00 103.00 116.00 105.00 103 105.00
118.00 109.0
CV residual  -8.72  -4.85  -1.28  -6.69  10.4 -11.3   1.93   1.58  -4.02  -6.13 -18  -4.37
```

```
CV residual  -6.72  -4.65  -1.28  -6.09  10.4 -11.5   1.93   1.58   4.02  -6.15  -18  -4.57
-5.84 -19.7
                617   640   646    667   677   681    694   695    712    735 758   766   767
778   813
Predicted   132.0 122.6 132.9 115.00 110.6 125.8 124.77 114.1 126.88 127.57 117 119.1 130.0
125.5 109.1
cvpred      132.4 123.6 135.8 116.53 110.8 128.8 128.78 114.3 127.59 128.07 118 119.6 130.9
128.7 108.8
wt          120.0 150.0 122.0 113.00 100.0 115.0 134.00 128.0 120.00 136.00 155  91.0 147.0
112.0 136.0
CV residual -12.4  26.4 -13.8  -3.53 -10.8 -13.8   5.22  13.7  -7.59   7.93  37 -28.6  16.1
-16.7  27.2
                816   819   837   842   858   873    898    916   917   922   924    945   966
967
Predicted   121.21 115.5 123.4 135.2 129.3 126.6 125.98 118.27 130.9 118.6 113.3 113.40 124.7
121.80
cvpred      122.09 117.3 125.3 138.9 132.5 127.8 126.67 119.14 132.3 119.7 113.5 113.85 125.6
122.61
wt          120.00 100.0 113.0 110.0 109.0 164.0 129.00 129.00 150.0 108.0 115.0 119.00 138.0
120.00
CV residual  -2.09 -17.3 -12.3 -28.9 -23.5  36.2   2.33   9.86  17.7 -11.7   1.5   5.15  12.4
-2.61
                978   985  1011  1015  1019   1032  1034   1043  1044   1046  1048  1053  1064
1066
Predicted   117.8 125.9 116.5 110.9  99.2 124.17 109.3 126.50 141.8 115.09 118.3 115.9 121.8
104.4
cvpred      117.1 126.9 117.1 110.3  88.6 124.31 107.8 127.71 145.7 115.27 116.4 116.4 122.5
101.2
wt          102.0 139.0 131.0 125.0 103.0 129.00  97.0 122.00 115.0 108.00 131.0 102.0  73.0
65.0
CV residual -15.1  12.1  13.9  14.7  14.4   4.69 -10.8  -5.71 -30.7  -7.27  14.6 -14.4 -49.5
-36.2
             1068   1072 1090  1091  1103    1114  1127   1128  1135  1138  1145  1146 1149
1151
Predicted   118.1 108.20  123 121.9 110.6 117.032 109.4 126.61 109.9 116.0 131.7 118.9  116
118.0
cvpred      119.6 108.54  124 122.9 109.1 117.065 108.4 128.28 110.9 116.2 132.8 119.3  116
119.7
wt          102.0 103.00   91 112.0 126.0 118.000 126.0 127.00 130.0 137.0 143.0 106.0   72
97.0
CV residual -17.6  -5.54  -33 -10.9  16.9   0.935  17.6  -1.28  19.1  20.8  10.2 -13.3  -44
-22.7
              1155  1165  1171  1172  1195  1196   1211  1217  1221
Predicted   124.38 110.8 117.2 114.4 121.6 132.0 121.08 110.3 111.6
cvpred      124.97 112.2 117.5 114.2 122.4 132.2 120.96 107.9 112.3
wt          117.00  84.0  82.0 119.0 103.0 112.0 116.00  97.0 135.0
CV residual  -7.97 -28.2 -35.5   4.8 -19.4 -20.2  -4.96 -10.9  22.7

Sum of squares = 36640    Mean square = 300    n = 122

fold 3
Observations in test set: 122
                 2     8    12    20     36    38    54    55     65    68     73     83
102    154
Predicted   120.63 109.6 123.6 136.6 125.31 121.4 127.4 116.4 121.37 123.5 123.43 129.32
115.2 122.15
cvpred      119.76 109.5 122.5 133.5 125.69 121.6 126.6 116.5 120.93 123.5 123.51 129.17
114.5 122.88
wt          113.00 132.0 144.0 119.0 134.00 134.0 143.0 146.0 128.00 137.0 133.00 139.00
119.0 119.00
CV residual  -6.76  22.5  21.5 -14.5   8.31  12.4  16.4  29.5   7.07  13.5   9.49   9.83
4.5  -3.88
               166    171   173   193   195    198    211    217   225   262   268   296
302   313
Predicted   121.46 113.12 122.9  96.2 127.5 123.38 108.14 137.82 116.4 107.7 128.7 104.1
124.3 120.1
cvpred      122.22 113.37 122.9  96.5 127.1 123.46 107.15 137.54 116.8 107.9 126.8 104.4
124.1 120.1
wt          116.00 121.00 138.0  75.0 104.0 118.00 113.00 128.00 134.0  81.0 142.0  91.0
109.0 106.0
CV residual  -6.22   7.63  15.1 -21.5 -23.1  -5.46   5.85  -9.54  17.2 -26.9  15.2 -13.4
-15.1 -14.1
              323   329    340    344   350    358   363   370    372    378   383    387
389   410
Predicted   128.6 116.2 117.97 117.40 120.4 123.02 119.8 114.5 122.78 112.35 110.3 124.33
118.7 127.0
```

```
CV residual -12.7 -21.6   9.27 -13.6   3.92  -6.24  -0.0359   6.13  -0.423  12.7 -19.6

Sum of squares = 36111    Mean square = 296    n = 122

fold 5
Observations in test set: 121
                 9    16    37    45     48    52    66    69     74    92    112 115     126
131
Predicted   121.51 128.7 116.5  97.1 110.31 123.5 127.4 115.4 114.35 111.5 131.37 120 114.959
115.275
cvpred      121.67 128.6 117.2 101.1 111.07 124.2 128.3 116.5 114.98 112.7 131.48 121 115.761
116.137
wt          120.00 115.0 107.0  87.0 110.00 108.0 104.0 103.0 120.00 134.0 125.00  95 115.000
117.000
CV residual  -1.67 -13.6 -10.2 -14.1  -1.07 -16.2 -24.3 -13.5   5.02  21.3  -6.48 -26  -0.761
0.863
               141   147   148 149    156    167   174    180   183   188    189   197   201
215
Predicted   105.30 109.6 145.2 124 123.50 130.78 111.9 123.02 118.5 111.2  98.96 118.9 113.7
122.1
cvpred      107.38 111.1 144.8 125 126.24 131.69 112.9 123.75 119.6 111.9 100.27 120.2 114.6
122.6
wt          100.00 102.0 160.0 113 123.00 129.00 136.0 132.00  96.0 133.0 107.00  90.0 101.0
150.0
CV residual  -7.38  -9.1  15.2 -12  -3.24  -2.69  23.1   8.25 -23.6  21.1   6.73 -30.2 -13.6
27.4
               220   242   249   270   285   311     359    391   408   417   439   459
471   476
Predicted   118.53 121.0 128.2 125.9 127.6 135.3 118.555 133.62 114.5 139.9 122.1 108.5
131.93 104.7
cvpred      119.61 121.9 128.9 127.2 128.7 135.4 119.206 133.44 115.6 138.8 123.7 111.9
131.16 105.4
wt          115.00 114.0 141.0 102.0 104.0 124.0 120.000 129.00 118.0 131.0 135.0  85.0
136.00  88.0
CV residual  -4.61  -7.9  12.1 -25.2 -24.7 -11.4   0.794  -4.44   2.4  -7.8  11.3 -26.9
4.84 -17.4
              484   487    494    533    541    552    571   583    590   623     635     643
657   684
Predicted   119.7 117.2 119.76 130.12 119.90 121.83 116.41 121.8 117.30 101.2 126.457 123.264
126 115.9
cvpred      119.8 118.2 120.96 129.64 121.11 123.08 119.24 122.3 119.33 103.7 127.392 124.398
127 116.2
wt           98.0 133.0 115.00 123.00 115.00 129.00 117.00  98.0 117.00  91.0 127.000 124.000
125  91.0
CV residual -21.8  14.8  -5.96  -6.64  -6.11   5.92  -2.24 -24.3  -2.33 -12.7  -0.392  -0.398
-2 -25.2
            692   693    702   704    723   724   736   737    759   761   783   792   796
798   829
Predicted   119 122.5 121.22 122.1 113.38 124.9 109.2 134.8 127.85 114.7 129.3 119.2 128.2
125.87 101.1
cvpred      119 121.9 122.85 122.7 114.11 126.6 110.7 134.5 127.91 115.6 130.7 120.4 129.4
125.72 103.3
wt          136 130.0 116.00 109.0 121.00 116.0  92.0 114.0 121.00  87.0 148.0 131.0 123.0
127.00  77.0
CV residual  17   8.1  -6.85 -13.7   6.89 -10.6 -18.7 -20.5  -6.91 -28.6  17.3  10.6  -6.4
1.28 -26.3
              830   839    846    871    874   879   889 891   892    895 897   912    920
938
Predicted    91.4 125.2 111.73 126.97 127.22 120.8 142.3 125 117.4 126.55 108 127.5 114.89
117.01
cvpred       94.6 127.2 112.38 127.04 127.08 121.8 142.6 126 119.2 127.71 109 127.6 116.35
118.89
wt           62.0 143.0 111.00 133.00 133.00 100.0 125.0 116 131.0 137.00  96 138.0 115.00
115.00
CV residual -32.6  15.8  -1.38   5.96   5.92 -21.8 -17.6 -10  11.8   9.29 -13  10.4  -1.35
-3.89
                968    983    995 1008  1010   1012  1021  1028  1036    1050  1058   1061
1078
Predicted   117.3543 129.90 123.70  117 122.5 114.27  97.5 131.9 100.2 123.118 128.5 121.22
121.57
cvpred      118.9259 130.09 123.81  118 124.1 115.37 100.9 131.4 104.5 124.147 128.6 122.23
122.73
wt          119.0000 138.00 125.00  120 108.0 110.00  75.0 145.0  65.0 125.000 155.0 125.00
121.00
CV residual   0.0741   7.91   1.19    2 -16.1  -5.37 -25.9  13.6 -39.5   0.853  26.4   2.77
-1.73
```

```
CV residual  -8.04  12.5   6.21  -6.49  -9.24 -20  -4.84   4.1   4.91 -13.5  59.4  -2.25
-4.86  -9.03
               290    294 301 321    324   325    334    339    341   348    396    411   413
415 422
Predicted   121.45 119.38 115 129 112.05 132.6 113.36 111.55 121.42 125.0 118.45 134.64 126.4
110.1 132
cvpred      122.06 118.69 116 131 111.59 134.1 113.26 112.06 122.05 125.5 119.06 137.87 127.7
111.2 133
wt          117.00 121.00  80 110 104.00 138.0 120.00 111.00 116.00 133.0 115.00 132.00 138.0
132.0 119
CV residual  -5.06   2.31 -36 -21  -7.59   3.9   6.74  -1.06  -6.05   7.5  -4.06  -5.87  10.3
20.8 -14
              432    436   447    451    465    495    524   531    558   559   561    581
605   617
Predicted   120.9 123.94 123.8 113.71 120.15 130.51 108.07 123.9 113.00 119.0 125.7 126.99
125.2 132.7
cvpred      122.5 123.73 126.3 114.06 117.37 131.89 107.84 124.5 111.47 120.3 125.8 127.91
126.8 133.3
wt          112.0 122.00 113.0 120.00 109.00 130.00 103.00 116.0 105.00 103.0 105.0 118.00
109.0 120.0
CV residual -10.5  -1.73 -13.3   5.94  -8.37  -1.89  -4.84  -8.5  -6.47 -17.3 -20.8  -9.91
-17.8 -13.3
              640    646    667    677   681    694   695    712   735   758   766   767 778
813
Predicted   126.1 126.50 114.88 109.67 128.0 124.02 108.4 124.18 122.7 119.0 116.5 129.3 129
103.0
cvpred      126.7 126.57 114.46 109.98 129.7 124.97 106.9 125.43 121.6 119.3 117.5 129.6 132
101.7
wt          150.0 122.00 113.00 100.00 115.0 134.00 128.0 120.00 136.0 155.0  91.0 147.0 112
136.0
CV residual  23.3  -4.57  -1.46  -9.98 -14.7   9.03  21.1  -5.43  14.4  35.7 -26.5  17.4 -20
34.3
               816   819   837   842   858   873    898   916   917    922   924    945   966
967
Predicted   124.99 110.8 127.6 129.2 128.2 125.3 124.65 117.8 127.4 115.05 105.6 116.25 123.9
110.6
cvpred      124.23 111.6 129.2 131.8 128.9 126.4 124.82 118.4 127.1 115.78 104.6 117.88 123.9
109.6
wt          120.00 100.0 113.0 110.0 109.0 164.0 129.00 129.0 150.0 108.00 115.0 119.00 138.0
120.0
CV residual  -4.23 -11.6 -16.2 -21.8 -19.9  37.6   4.18  10.6  22.9  -7.78  10.4   1.12  14.1
10.4
              978   985  1011  1015    1019   1032  1034  1043  1044  1046  1048  1053  1064
1066
Predicted   116.4 120.3 112.9 110.9 104.158 128.57 117.3 130.8 128.8 119.7 118.9 118.7 120.7
97.1
cvpred      116.3 119.8 113.5 111.3 102.398 130.23 118.8 132.7 129.5 120.5 117.9 120.2 121.1
94.8
wt          102.0 139.0 131.0 125.0 103.000 129.00  97.0 122.0 115.0 108.0 131.0 102.0  73.0
65.0
CV residual -14.3  19.2  17.5  13.7   0.602  -1.23 -21.8 -10.7 -14.5 -12.5  13.1 -18.2 -48.1
-29.8
             1068    1072 1090   1091  1103    1114  1127   1128  1135  1138   1145  1146
1149   1151
Predicted   118.6 104.424  110 114.38 103.6 116.685 109.2 121.12 106.2 115.7 136.24 121.3
110.0 104.36
cvpred      119.5 102.669  111 114.19 103.3 117.423 110.7 121.06 105.3 116.3 136.06 121.8
109.1 103.71
wt          102.0 103.000   91 112.00 126.0 118.000 126.0 127.00 130.0 137.0 143.00 106.0
72.0  97.00
CV residual -17.5   0.331  -20  -2.19  22.7   0.577  15.3   5.94  24.7  20.7   6.94 -15.8
-37.1  -6.71
             1155  1165  1171   1172  1195  1196   1211  1217  1221
Predicted   114.18 110.8 114.9 116.92 122.4 127.1 115.21 109.5 117.7
cvpred      113.88 110.7 114.7 117.65 121.3 127.8 113.33 108.5 119.1
wt          117.00  84.0  82.0 119.00 103.0 112.0 116.00  97.0 135.0
CV residual   3.12 -26.7 -32.7   1.35 -18.3 -15.8   2.67 -11.5  15.9

Sum of squares = 33966    Mean square = 278    n = 122

fold 3
Observations in test set: 122
                2   8  12     20     36    38  54    55     65    68  73    83    102    154
166
Predicted   123.17 107 128 121.95 128.97 118.3 131 121.1 121.97 118.9 121 135.9 120.83 123.61
```

```
-4.68   0.19
             1189 1194  1200   1215   1218 1220   1234   1235
Predicted   113.1  127 146.3 125.01 142.15  143 129.18 123.80
cvpred      113.3  127 145.1 124.64 141.26  142 128.93 123.23
wt           84.0  139 102.0 118.00 146.00  110 130.00 125.00
CV residual -29.3   12 -43.1  -6.64   4.74  -32   1.07   1.77

Sum of squares = 32990    Mean square = 270    n = 122

fold 4
Observations in test set: 122
                6    23     33     39     49    56    58     61     62     67    95     96
100
Predicted   120.8 120.9 128.43 120.31 123.48 117.0 117.1 112.83 129.75 106.14 109.4 126.63
108.74
cvpred      120.1 120.7 128.81 120.43 123.44 116.7 116.4 112.22 129.13 104.94 108.9 126.57
107.44
wt          136.0 137.0 130.00 122.00 122.00 124.0 145.0 107.00 124.00  97.00  85.0 135.00
105.00
CV residual  15.9  16.3   1.19   1.57  -1.44   7.3  28.6  -5.22  -5.13  -7.94 -23.9   8.43
-2.44
               118 128   129   146    151   169    176   187   192   196   200    209    214
224
Predicted   128.90 106 122.1 123.0 121.74 119.0 129.87 114.5 120.9 115.9 126.1 128.08 126.52
118.258
cvpred      128.63 105 121.8 122.8 121.03 118.1 129.36 114.9 120.7 114.5 124.8 127.61 126.02
117.467
wt          131.00  94 109.0 136.0 126.00 131.0 122.00 137.0 136.0 130.0 137.0 131.00 117.00
117.000
CV residual   2.37 -11 -12.8  13.2   4.97  12.9  -7.36  22.1  15.3  15.5  12.2   3.39  -9.02
-0.467
              240   241    246   250    264    279     303    305   315    331   346   362
380    385
Predicted   126.8 135.2 119.76 116.4 126.74 133.79 113.316 131.66 111.4 117.40 111.3  97.5
113.9 132.40
cvpred      125.4 133.5 119.74 115.3 125.85 133.59 111.956 131.69 110.8 117.44 111.3  96.6
113.4 131.76
wt          173.0 144.0 111.00 142.0 125.00 131.00 111.000 136.00 100.0 116.00  93.0  71.0
101.0 127.00
CV residual  47.6  10.5  -8.74  26.7  -0.85  -2.59  -0.956   4.31 -10.8  -1.44 -18.3 -25.6
-12.4  -4.76
              414   470   485 497   506 512    516     518    528     556   557   567    588
604
Predicted   123.1 137.5 94.51 119 127.4 101 107.96 117.437 122.77 129.302 127.1 114.5 133.00
117.084
cvpred      123.1 137.4 94.57 118 126.5 101 106.88 117.111 122.11 129.266 126.3 113.8 132.52
116.198
wt          139.0 107.0 96.00 105 115.0 127 110.00 117.000 132.00 129.000 174.0 133.0 123.00
117.000
CV residual  15.9 -30.4  1.43 -13 -11.5  26   3.12  -0.111   9.89  -0.266  47.7  19.2  -9.52
0.802
               610    611   615   618   644   703    708   713    716   734   748    749
769    790
Predicted   114.08 108.33 123.6 116.9 112.8 125.7 113.16 119.8 128.27 123.2 120.0 130.13
118.0 118.71
cvpred      113.39 107.51 123.5 116.3 112.2 126.2 112.99 119.5 127.93 122.2 119.2 129.37
117.6 118.72
wt          123.00 105.00 103.0 145.0 123.0 110.0 121.00  95.0 131.00  84.0 174.0 127.00
128.0 127.00
CV residual   9.61  -2.51 -20.5  28.7  10.8 -16.2   8.01 -24.5   3.07 -38.2  54.8  -2.37
10.4   8.28
              801   812    814   815   824    827   843   855   862    864    901   909
915    948
Predicted   127.8 115.0 129.42 122.6 125.3 122.23 126.8 114.6 120.2 112.94 121.01 105.9
116.14 129.78
cvpred      127.7 114.5 129.36 122.3 124.9 121.06 125.8 113.9 120.2 112.68 120.84 105.1
115.84 129.91
wt          139.0 144.0 121.00 165.0 141.0 130.00 162.0 126.0 124.0 122.00 111.00  90.0
110.00 122.00
CV residual  11.3  29.5  -8.36  42.7  16.1   8.94  36.2  12.1   3.8   9.32  -9.84 -15.1
-5.84  -7.91
              959   973   974   984   993   997  1001  1003   1006   1017   1020   1029
1040 1041
Predicted   124.95 124.7 127.7 118.4 132.2 112.5 128.4 137.4 119.84 122.18 119.35 123.92
125.80  122
```

```
-9.09
               798   829   830   839   846   871    874   879   889   891   892    895   897
912
Predicted   128.11  89.8  85.1 132.0 115.1 123.5 122.74 117.3 137.8 128.1 117.9 126.74 105.0
120.1
cvpred      128.62  92.1  87.2 132.1 115.5 122.6 123.13 117.4 137.8 128.3 118.7 128.28 106.3
119.6
wt          127.00  77.0  62.0 143.0 111.0 133.0 133.00 100.0 125.0 116.0 131.0 137.00  96.0
138.0
CV residual  -1.62 -15.1 -25.2  10.9  -4.5  10.4   9.87 -17.4 -12.8 -12.3  12.3   8.72 -10.3
18.4
                920    938     968    983    995   1008   1010   1012  1021  1028  1036
1050  1058
Predicted   122.47 116.57 116.934 135.37 120.26 111.74 112.99 112.19 102.7 126.5 106.9
124.621 126.9
cvpred      122.28 117.02 118.773 135.52 122.09 113.43 114.83 112.95 107.7 125.8 109.6
125.952 127.6
wt          115.00 115.00 119.000 138.00 125.00 120.00 108.00 110.00  75.0 145.0  65.0
125.000 155.0
CV residual  -7.28  -2.02   0.227   2.48   2.91   6.57  -6.83  -2.95 -32.7  19.2 -44.6
-0.952  27.4
              1061   1078   1082 1084  1088  1093   1110   1115   1119   1129  1131  1141
1152  1156
Predicted   121.09 123.81 119.96  132 119.8 129.2 110.10 111.07 117.48 128.50 117.4 135.5
118.8 127.9
cvpred      121.75 125.32 120.77  131 120.7 130.6 111.97 111.26 118.35 129.85 118.6 134.3
121.4 128.8
wt          125.00 121.00 118.00  117 107.0 157.0 108.00 111.00 113.00 134.00 102.0  98.0
106.0 112.0
CV residual   3.25  -4.32  -2.77  -14 -13.7  26.4  -3.97  -0.26  -5.35   4.15 -16.6 -36.3
-15.4 -16.8
               1169   1175  1177  1183   1187   1206  1209  1210  1212  1213   1228
Predicted   129.764 121.13 107.5 114.9 108.23 112.10 112.7 121.6 107.4 129.5 107.95
cvpred      130.804 122.45 110.1 116.4 109.54 112.86 112.4 122.4 109.9 129.1 108.68
wt          130.000 124.00  88.0  97.0 116.00 114.00 141.0 144.0  75.0 138.0 103.00
CV residual  -0.804   1.55 -22.1 -19.4   6.46   1.14  28.6  21.6 -34.9   8.9  -5.68

Sum of squares = 28947    Mean square = 239    n = 121

Overall (Sum over all 121 folds)
 ms
255
</pre></body></html>
```

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 – Testing Other Interaction-Effect Models

```
# Fitting
interaction
models for
certain
variables
against
baby weight
```

```
# Linear model between race and mother weight against baby weight

race_wt.1 <- lm(wt ~ race*wt.1, data = data)
summary(race_wt.1)
anova(race_wt.1)

# Linear model between mother's weight and smoking against baby weight
smoke_wt.1 <- lm(wt ~ smoke*wt.1, data = data)
summary(smoke_wt.1)
anova(smoke_wt.1)

# Linear model between mother's weight and parity against baby weight
parity_wt.1 <- lm(wt ~ parity*wt.1, data = data)
summary(parity_wt.1)
anova(parity_wt.1)

# Linear model between mother's weight and time against baby weight
```

```
time_wt.1 <- lm(wt ~ time*wt.1, data = data)
summary(time_wt.1)
anova(time_wt.1)

# Linear model between mother's weight and income against baby weight
inc_wt.1 <- lm(wt ~ inc*wt.1, data = data)
summary(inc_wt.1)
anova(inc_wt.1)

#Checking AIC scores for each model
AIC(race_wt.1)
AIC(smoke_wt.1)
AIC(parity_wt.1)
AIC(inc_wt.1)

# Model diagnostics for each model built,
# Error shape and distribution of model between race and mother weight
against baby weight
qqnorm(resid(race_wt.1))
shapiro.test(resid(race_wt.1))
hist(resid(race_wt.1))

# Error shape and distribution of model between mother's weight and
smoking against baby weight
qqnorm(resid(smoke_wt.1))
shapiro.test(resid(smoke_wt.1))
hist(resid(smoke_wt.1))

# Error shape and distribution of model between mother's weight and parity
against baby weight
qqnorm(resid(parity_wt.1))
shapiro.test(resid(parity_wt.1))
hist(resid(parity_wt.1))

# Error shape and distribution of  model between mother's weight and
income against baby weight
qqnorm(resid(inc_wt.1))
shapiro.test(resid(inc_wt.1))
hist(resid(inc_wt.1))

# Error spread of model between race, mother weight against baby weight
race_resid <- resid(race_wt.1)
plot(fitted(race_wt.1), race_resid, ylab = 'residuals', xlab = 'Fitted
values')
```

```
# Error spread of model between mother's weight and smoking against baby
weight
smoke_resid <- resid(smoke_wt.1)
plot(fitted(smoke_wt.1), smoke_resid, ylab = 'residuals', xlab = 'Fitted
values')

# Error spread of model between mother's weight and parity against baby
weight
parity_resid <- resid(parity_wt.1)
plot(fitted(parity_wt.1), parity_resid, ylab = 'residuals', xlab = 'Fitted
values')

# Error spread of model between mother's weight and income against baby
weight
inc_resid <- resid(inc_wt.1)
plot(fitted(inc_wt.1), inc_resid, ylab = 'residuals', xlab = 'Fitted
values')

# Error independence of model between race, mother weight against baby
weight
library(car)
durbinWatsonTest(race_wt.1)

# Error independence of  model between mother's weight and smoking against
baby weight
durbinWatsonTest(smoke_wt.1)

# Error independence of mother's weight and parity against baby weight
durbinWatsonTest(parity_wt.1)

# Error independence of model between mother's weight and income against
baby weight
durbinWatsonTest(inc_wt.1)

# Ncv test for the models
ncvTest(smoke_wt.1)
ncvTest(inc_wt.1)
ncvTest(parity_wt.1)
```

Appendix 6 - 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 7 – 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
```