# A Statistical Report to Assess the Relationship Between Birth Weight and A Series of Measured Variables

Module: Introductory Data Analysis (MT5762)
Project 2

Team: **Gryffindor**

Word Count: 4177

# Table of Contents

**Executive Summary**

This analysis focuses on identifying relationships between birth weight of babies and the rest of the data. The data are part of a larger group of studies named “Child Health and Development Studies (CHDS)”. Following an initial cleaning and some exploratory analysis of basic correlations present, two linear models were created using alternative model selection techniques; forward and backwards selection. Due to the fundamental assumptions of normality and constant variance made by linear models, various assumptions tests were undertaken on each model. Two model validation tests and bootstrapping of the data allowed the model with the best predictive ability to be selected, which in this study was model 2. Through the various tests, it was found that baby weight was indeed affected by almost all variables to differing degrees with the variable gestation showing the greatest degree of correlation. In agreement with similar studies, the mothers smoking habitats were found to have a negative impact on baby weights with the model suggesting a lower baby weight as more likely if the mother smoked during pregnancy. Alternatively, mother’s height and parity were found to have a positive effect on baby weight. This Study found our selected model to have relatively significant statistical explanatory power but from a practical point of view its predictive power is low. To further the research and models predictive ability, more socio-economic variables could be included.

# 1. Introduction

## *1.1. Background*

The economic and social implications of low birth weights (LBW) amongst newborn babies are problematic (Walker *et al.,* 2009). According to the World Health Organisation (2012), LBW is defined as an infant mass of less than 2500 g. LBW can lead to contribute to post-natal medical problems and even lower literacy levels (Walker *et al.,* 2009).

External factors, such as smoking, during pre-natal development have raised concerns in the past. A study by MacMahon *et al.,* (1965), highlighted cigarette smoke as a genetic mutagen which can be detrimental to a foetus and cause LBW. However, whilst infants in the study did appear to have LBW's due to the mothers smoking, no other detrimental effects were noted (MacMahon *et al.,* 1965). More recent studies have indicated contradictory views, for example, Lawoyin, (2001) and Adegboye *et al.* (2010) stated that such maternal factors such as smoking during pregnancy may lead to LBW, birth asphyxia or premature births – all concerns relating to increased infant mortality. Abrevaya (2006) also stated that both social and economic inequality can be detrimentally affected by low birthweights as mothers of a lower socio-economic backgrounds are more at risk of their baby having a low birth weight.

Primary studies have provided identification of key factors that contribute to LBW with aid from numerous retrospective studies. However, many social, economic and biological factors can contribute to LBW, thus it has been difficult to isolate or identify which factors are more important than others (Makhija *et al.,* 1998; Negi *et al.,* 2006). With large data sets available, recording both births and information about the mother and fathers health and background, it is important for studies to continue to examine the relationships between various social, economic and biological factors and low birth weights as finding causal factors could help alleviate the social and economic costs of low birth weights (Makhija *et al.,* 1989; Walker *et al.,* 2009). As with similar studies (e.g. Visscher *et al.,* 2003; Abrevaya, 2006) a linear model will most likely be extremely important for attempting to determine causal relationships, thus it is important that models are designed and tested to work best with all variables.

### *1.2. Aims and Research Questions*

Given the importance of understanding the relationships between social, economic and biological variables and birth weight, this study will use a data set from the Child Health and Development Studies (CHDS) containing such variables to assess relationships. Following initial data exploration to identify interesting relationships between variables, a linear model will be developed and tested to determine the best linear model to assess relationships between the potential drivers of LBW babies. This report will provide an insight into the following research question:

***"What Relationships Are There Between the Measured Variables and The Birth Weight of Babies?"***

## 2. Methodology

All the statistical computing and analysis within this report was performed using the RStudio® V1.2.1335 (R Core Team, 2019) and additional packages which have been referenced at the end of the report.

### *2.1. Cleaning the Data*

The data set provided by the CHDS was initially checked for inconsistencies such as unknown values. Filtering the data helped determine the number of unknown values present for each variable. Once identified, unknown values had their values re-assigned to NA.

The way categorical variables had been assigned values within the data set meant that within the categories for race, for example, 0-5 all represented white. Therefore, in categorical variables where this occurred, values were merged so that each category only had one value assigned to it. However, these values were then re-assigned to the name they represented, for example in the variable race, where 0 represented white, this was changed to white enabling categorical and numerical variables to be completely separated.

The variable date was represented by a single number which was changed to standard date format. Based on the data, it was assumed that 0 was 1$^{st}$ January 1958. In addition, in order to ensure consistency between units of weight, all variables were converted to kilograms as birth weight had been measured in ounces, whilst the mother's and father's weights were in

pounds. Finally, the increments for income were changed to increments of 2500 between the values one and nine.

### *2.2. Exploratory Analysis*

To gain an overview of the data, a series of summary statistics were generated for each variable using the summary function in R (R Core Team, 2019). To determine relationships between numerical variables only, a subset of the main data set was created including only such variables. Correlations between these numerical variables were initially explored by producing two correlation matrices where correlations not significant to a 99% confidence level were excluded. The variable birth weight was then plotted against age, height and weight of mother and father and gestation length in scatter plots. The points were colour coded based on categorical variables such as if the mother smoked or the race of the mother or father. Four further plots were then created examining the numerical variables against baby weight. The data was filtered and points colour coded based on the number of cigarettes smoked by mums who smoked through pregnancy, mums who had never smoked coded by race, the number of cigarettes smoked by mums who smoked until pregnancy and mums who used to smoke but not now colour coded by race.

### *2.3. Partitioning the Data*

When developing a linear model, its ability to predict is important (Fushiki, 2012). It is important the model is developed and followed by a series of validation tests to determine how accurately the model can predict (Kassambara, 2018). Following the method proposed by Shao (1993), the entire data set was split into subsets. This study created 20% and 80% subsets based on the number of rows. The 80% subset was used to develop and fit the model and 20% was set to validate the model's predictive abilities (Shao, 1993; Kassambara, 2018).

### *2.4. Creating the Linear Model*

To create a suitable linear model which describes the variables affecting birth weight, two models were created with the 80% subset using two alternative model selection methods and the best model selected (Benedetti and Brown, 1978; Shao, 1993; Kassambara, 2018).

### *2.4.1. Model One*

Model one was created using the backwards variable elimination method for model selection (Austin, 2008). This involves beginning with a model including all variables and one by one eliminating these from the starting model until a pre-determined significance level for each variable within the model is reached, this is usually the Akaike Information Criterion (AIC) and is used to select the best model (Austin, 2008; Brusco and Steinley, 2015). The use of a linear model makes two fundamental assumptions: linearity and constant variance. A series of diagnostic tests were used to test these assumptions on all the covariates. In all tests, a significance level of 0.05 was used (α).

Firstly, a two-tailed ANOVA test was conducted to determine if a significant difference was present between the means of the model covariates. The test used the following null ($H_0$) and alternative ($H_a$) hypotheses:

***H*$_0$**: There is no significant difference between the means of the covariates
***H*$_a$**: There is a significant difference amongst the means of the covariates

To assess the extent of multicollinearity within the data, a variance inflation factor (VIF) test was conducted (O'Brian, 2007; Yoo *et al.,* 2014). A test output of above five, would indicate it was highly likely that the variance of the regression coefficient was inflated due to multicollinearity in the model (Yoo et al., 2014).

Next, it was tested if the residuals in the model were normally distributed in order to validate the assumption of normality. The Shapiro-Wilk normality test was conducted with the following $H_0$ and $H_a$, (Rochon *et al.,* 2012):

***H*$_0$**: The model residuals are normally distributed.
***H*$_a$**: The model residuals are not normally distributed.

A non-constant error variance (NCV) test was then undertaken to measure whether the variance of error is dependent on the values of the independent variables. The test used the following $H_0$ and $H_a$ *(*Zaman, 1995; Breusch and Pagan, 1979; Rosopa *et al.,* 2013*)*:

***$H_0$***: The error variance of the model residuals is constant (heteroscedasticity).

***$H_a$***: The error variance of the model residuals is not constant (homoscedasticity).

The final test conducted was the Durbin-Watson (DW) test which examined autocorrelation within the residuals (Chen, 2016; Salamon *et al.,* 2019). Independence in the residuals must be tested for, as linear models assume independence between variables. The test used the following $H_0$ and $H_a$ (Nerlove and Wallis, 1966; Salamon *et al.,* 2019*)*:

***$H_0$***: The model residuals are not autocorrelated.

***$H_a$***: There is autocorrelation present between the residuals of the model or another variable

Following the linear model assumption tests, confidence intervals for the variables were produced in addition to various plots to visualise the residual distribution of the residuals e.g. QQ-plots which could visually support the assumption tests. Finally, the residual distribution of the covariates and the partial residual distribution per covariate were plotted.

#### *2.4.2. Model Two*

The difference between the production of model one and model two lies in the use of either the forward or the backwards selection method. Model two was created using the forward model selection method. The same assumption testing, hypotheses and diagnostic plots were also undertaken for model two.

#### *2.4.3. Selecting the Best Model*

Once two models had been developed, it was important to select the model which was best at describing the relationships between the dependent and explanatory variables. The AIC calculated for each model was the basis of selection. As recommended by statistical literature, the model with the lowest AIC value was (Zuur *et al*., 2009).

### *2.5. Model Validation*

As previously stated, assessing the model's predictive ability is important (Shao, 1993). More specifically, it is important to assess how good the model is at predicting values of variables not used In the creation of the model (Kassambara, 2018). Model validation tests

use a small subset of the main data set not used in the creation of the model to assess the predictive ability through determining the error in the predictions made by the model (Fushiki, 2011). There is a wide array of different validation methods, two were used in this study.

#### *2.5.1. Cross-Validation Set Approach*

In this method, 80% of the main data set was used to create the model and the remaining 20% used to validate the model (Kassambara, 2018). The error in the prediction is stated as the root mean square error (RMSE) which can be used to compare the model and chose the model with the lowest RMSE (Kassambara, 2018).

#### *2.5.2. Five-fold Cross-validation*

This method is suggested by Kassambara (2018) as a more rigorous validation test. The predictive ability of the model is evaluated by feeding the whole data set through the validation function which divides it into 5 random subsets, It is suggested by Kassambara (2018) that 5-fold validation is enough to provide good results which minimise bias. The model would be tested on one subset, then repeated on all 5 subsets and an average prediction error calculated (Kassambara, 2018). This method is preferred to many others as it is less computationally intensive and is useful for larger data sets such as the one in the study (Fushiki, 2011).

#### *2.5.3. Selecting the Best Model*

Based on the cross-validation tests, the model with the lowest prediction error is typically chosen as this shows it has the greatest predictive ability (Shao, 1993; Ronchetti *et al.,* 2012).

### *2.6. Bootstrapping*

Bootstrapping enables distributions of the sample populations to be determined without making assumptions about the data (Fox, 2015; Fox and Weisberg, 2002). This involves repeatedly re-sampling a sample of a specified size using the main data set as a sampling population (Fox, 2015). The re-sampled data is then fitted to the model, this is repeated 1000 times which will give 1000 estimators or the parameters. It was suggested by Efrom and Tibshirani (1993) that 1000 samples are more than satisfactory to provide accurate outputs. Furthermore, a quantile function will be used to get confidence intervals for 95%.

## 3. Results

### *3.1. Cleaning the Data*

Table 1 presents the number of unknown values for each variable. Numerous variables had very few unknown values, however, father's height, father's weight and income all had a high frequency of unknown values.

*Table 1: Frequency of unknown values per variable out of a total of 1236 values for each variable*

| Variable | Unknown Values |
|---|---|
| Plurality | 0 |
| Outcome | 0 |
| Date | 0 |
| Gestation | 13 |
| Sex | 0 |
| Birth Weight | 0 |
| Parity | 0 |
| Mothers Race | 13 |
| Mothers Age | 2 |
| Mothers Education | 1 |
| Mothers Height | 22 |
| Mothers Pregnancy Weight | 36 |
| Fathers Race | 15 |
| Fathers Age | 7 |
| Fathers Education | 13 |
| Fathers Height | 492 |
| Fathers Weight | 499 |
| Marital | 2 |
| Income | 124 |
| Does Mother Smoke | 10 |
| If Mother Quit, how long ago? | 9 |
| Number of Cigarettes smoked per day | 10 |

### *3.2. Exploratory Analysis*

Figures 1 and 2 show a correlation matrix for the numerical variables in the dataset. There is little evidence of correlations between the variables. The two strongest relationships present are between the age of the mother and age of the father and the height of the father and weight of the mother.

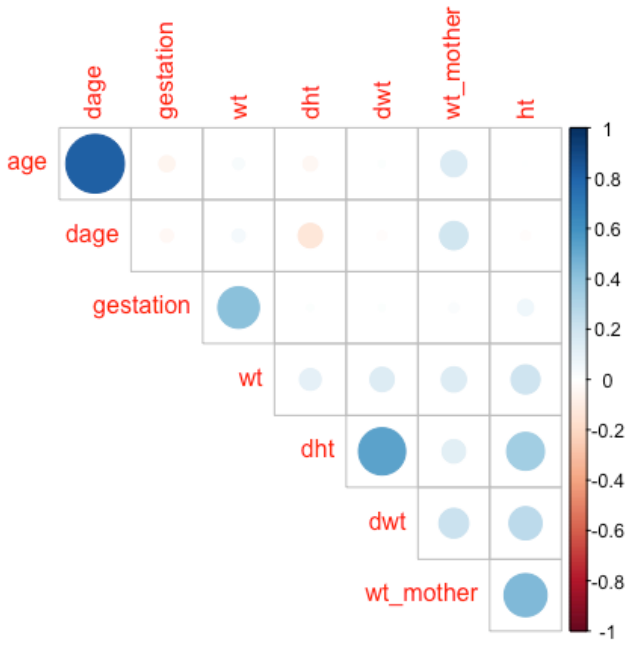


*Figure 1: Correlation matrix (1) showing the strength of the correlation between the numerical variables within the data. Correlations which were not significant to a 99% confidence level have been excluded.*

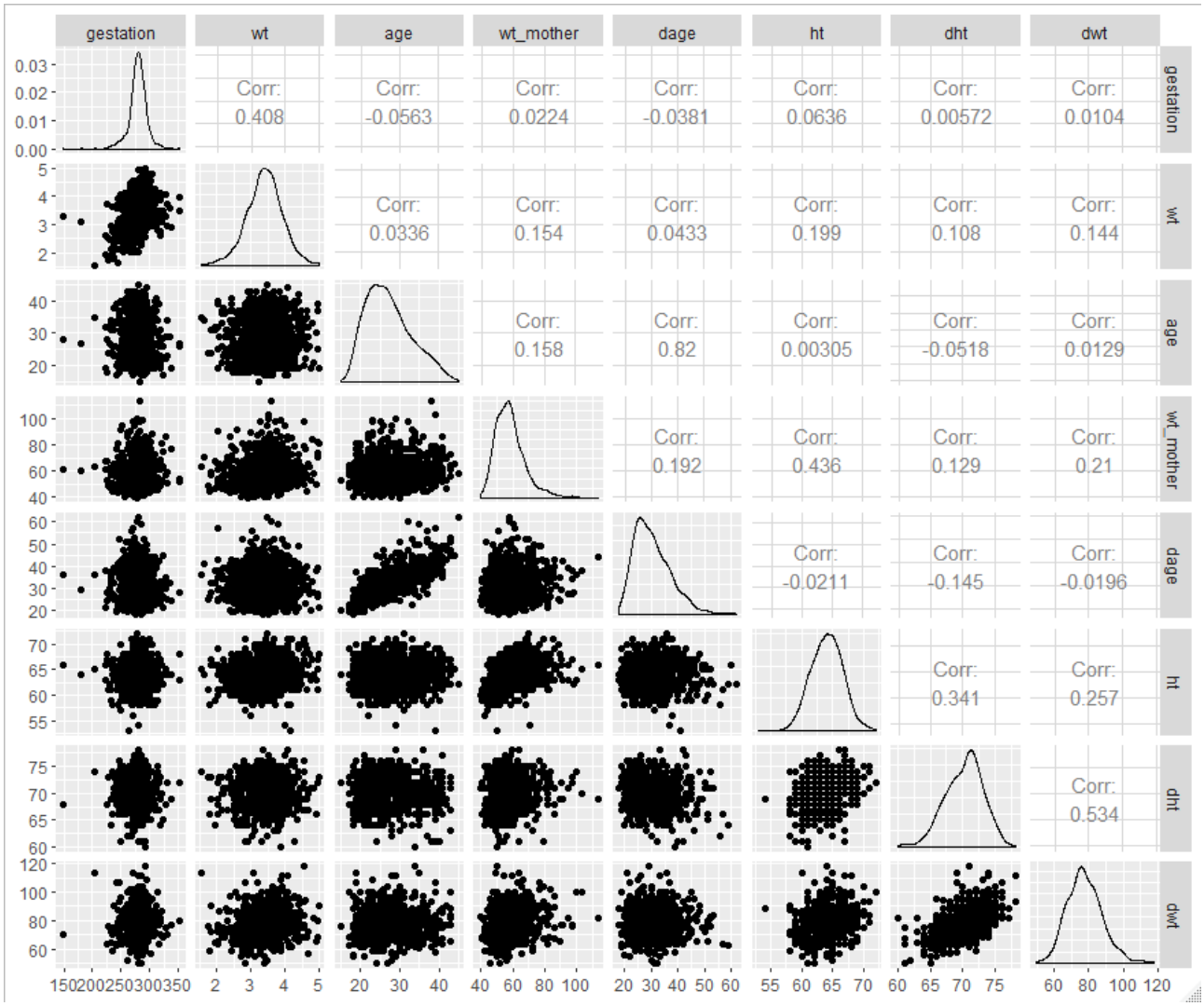


*Figure 2: Correlation matrix (2)*

Figures 3, 4 and 5 show how the seven numerical variables plot against baby weight, with the points colour coded based on if the mother smoked, the race of father and the race of the mother respectively. There are very few relationships evident between the variables and birth weight. The only variable where there is evidence of a relationship with birth weight is gestation and to a lesser extent the mother's weight. As seen above, gestation and mother's weight seem to be positively correlated with the weight of the baby. However, the latter relationship is still relatively weak as there are points spread over a wide area.

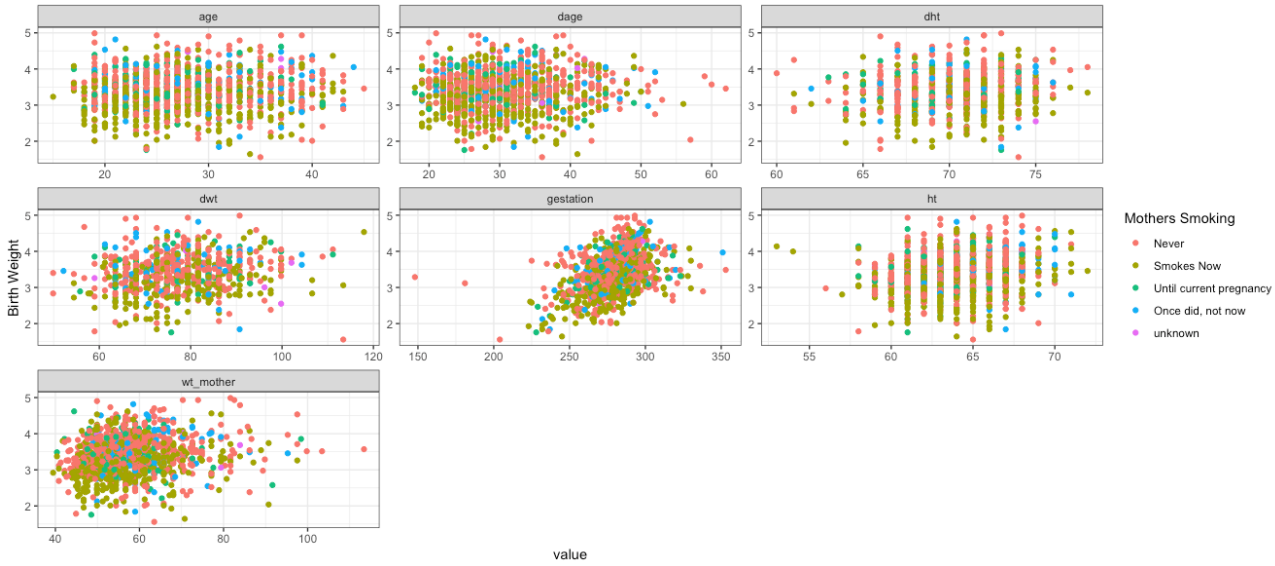


*Figure 3: Scatter plot for birth weight against 7 numerical variables: age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight). The points are colour coded based on mothers smoking.*

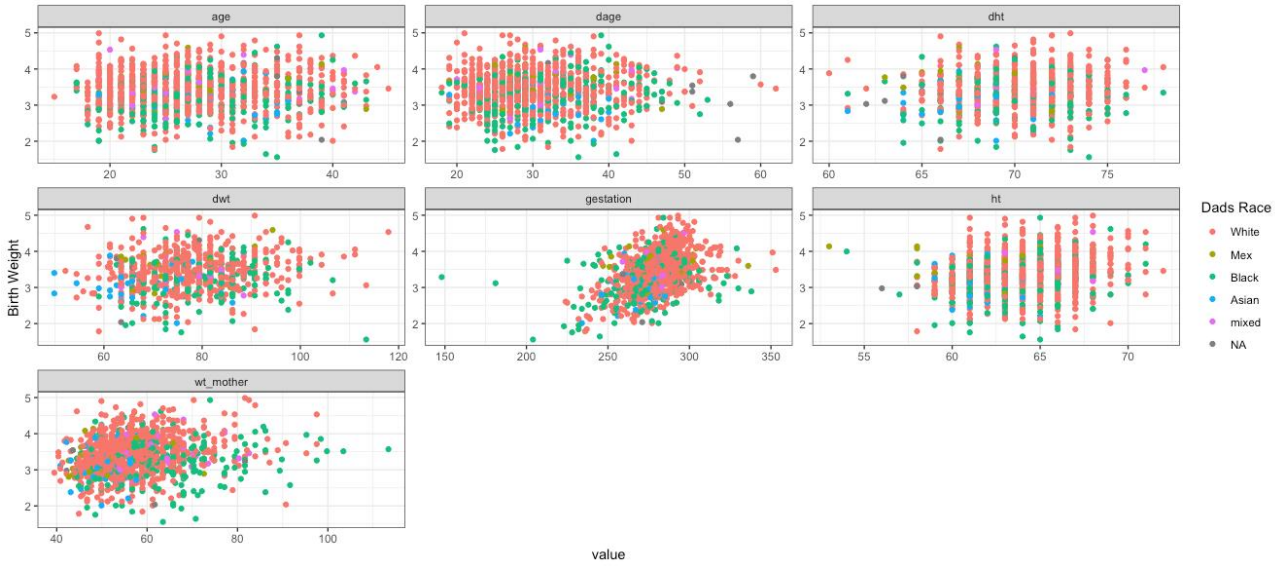


*Figure 4: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight). The points are colour coded based on the fathers' race.*

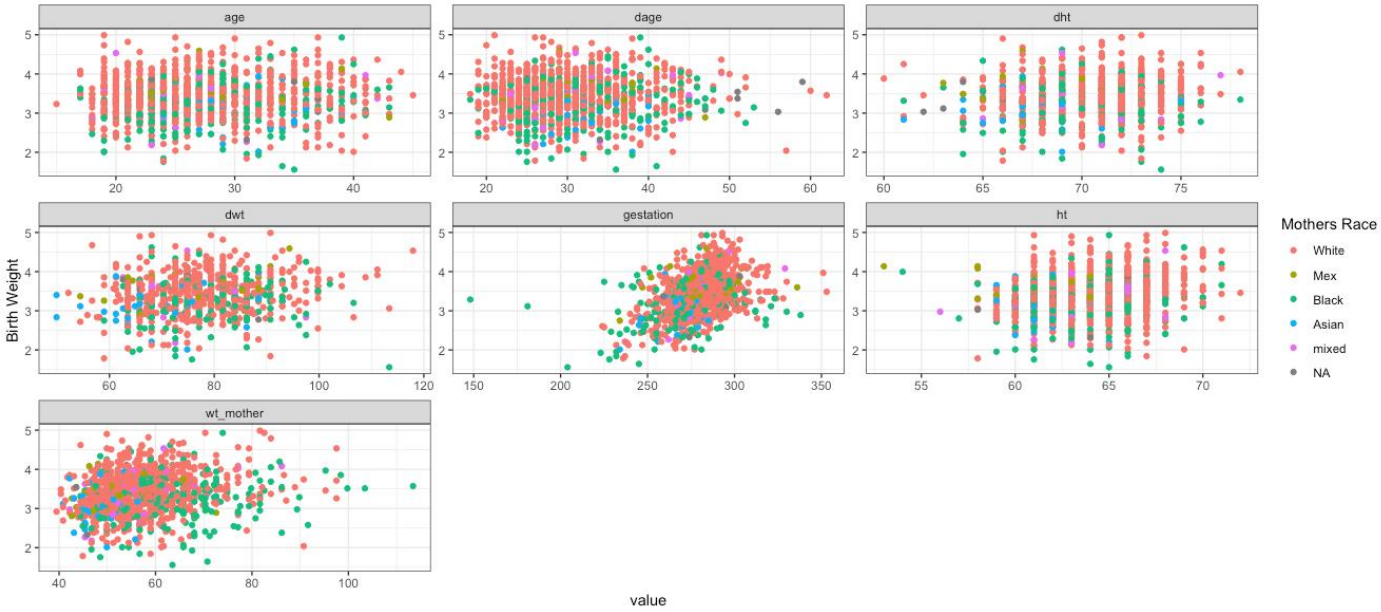


*Figure 5: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (father's weight), gestation, ht (height), wt_mother (Mothers weight). The points are colour coded based on the mothers' race.*

Figures 6-9 were filtered and so contained less data. Similarly, to Figures 3-5, there was very little which could be inferred from these plots as there were very few relationships evident between the variables and the data was very widely spread. In all four Figures, the plot for gestation against birth weight showed evidence for a weak positive relationship with gestation period increasing with increasing birth weight. In Figures 6 and 7 birth weight and mothers weight showed a very weak positive relationship, but the same variables showed a negative relationship in Figure 8 and no relationship in Figure 9. However, none of the relationships are significant due to the large spread of the points.

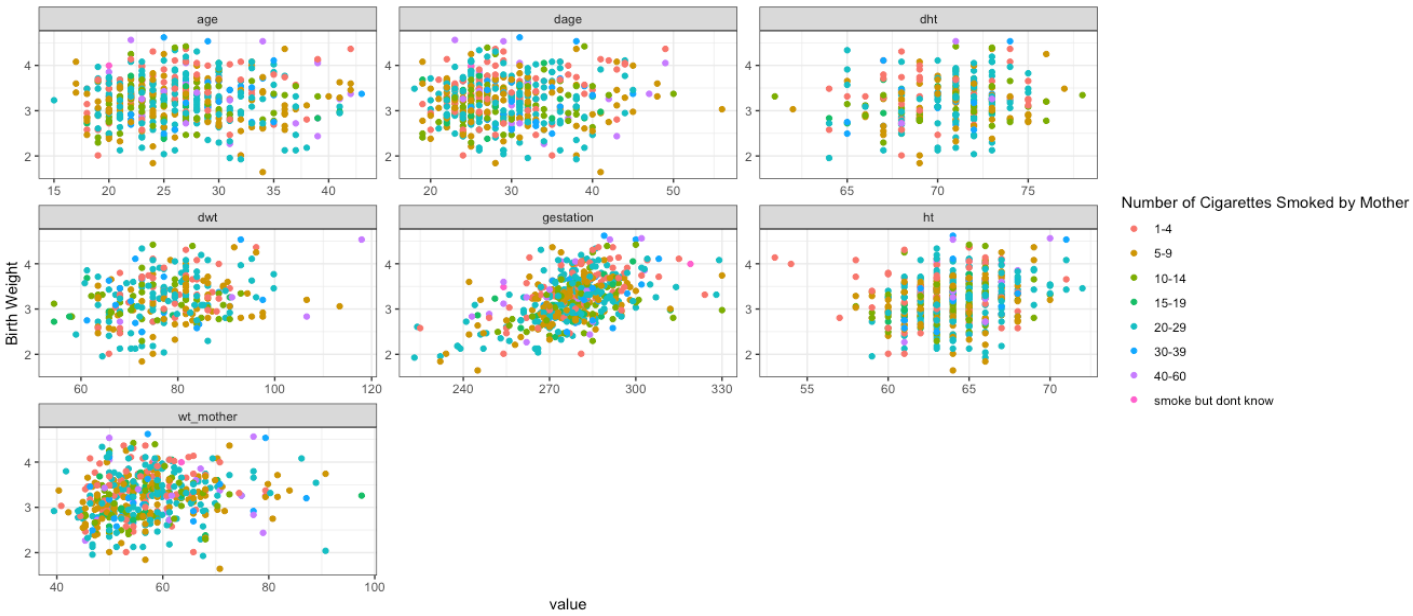


*Figure 6: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight). The points are colour coded based on the number of cigarettes smoked*

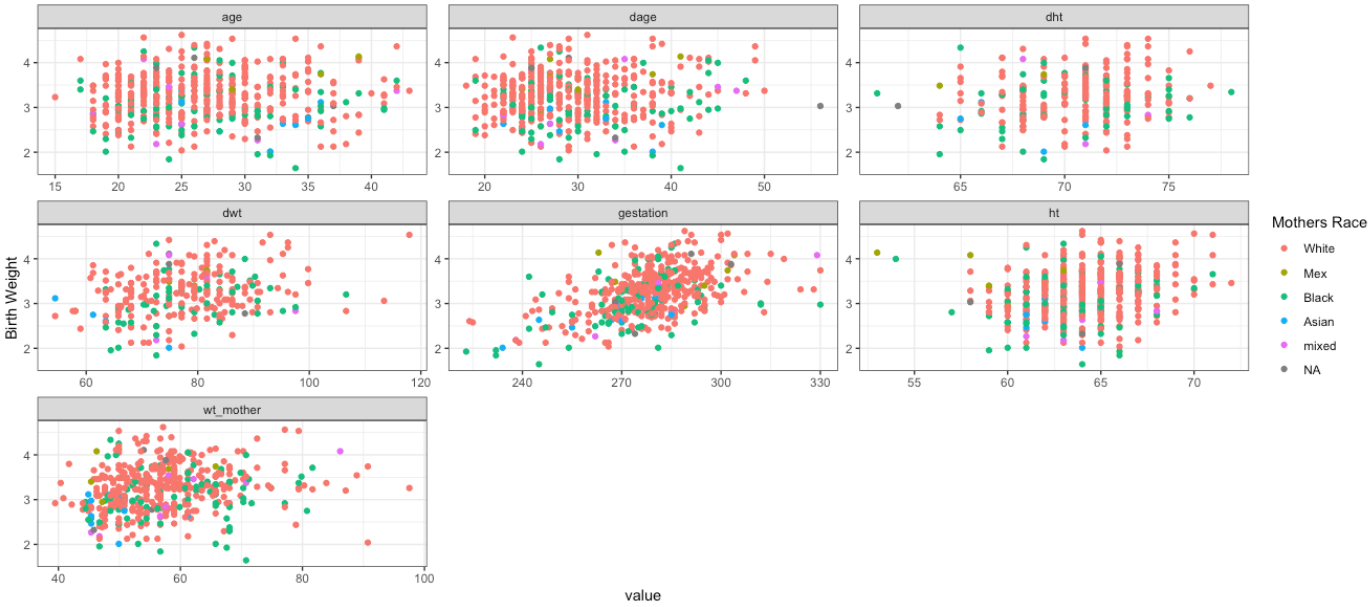


*Figure 7: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight) for mothers who had never smoked. The points are colour coded based on the race of the mother.*

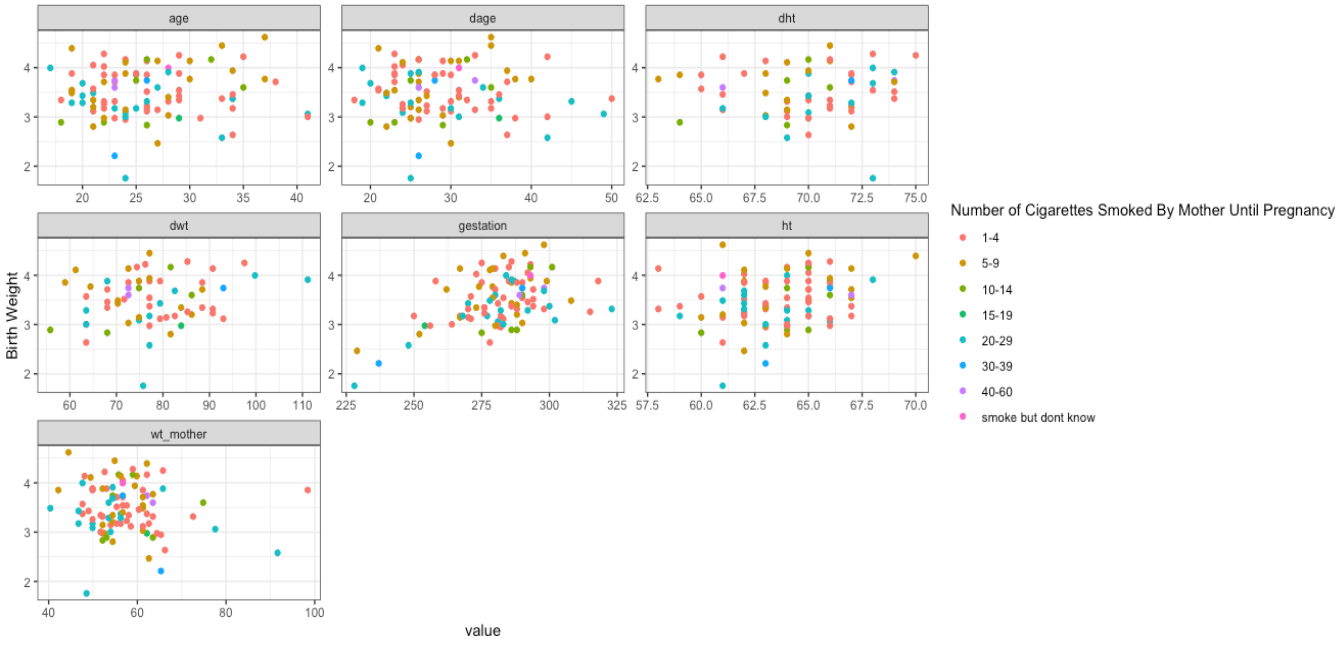


*Figure 8: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight) for mothers who smoked until pregnancy. The points are colour coded b*

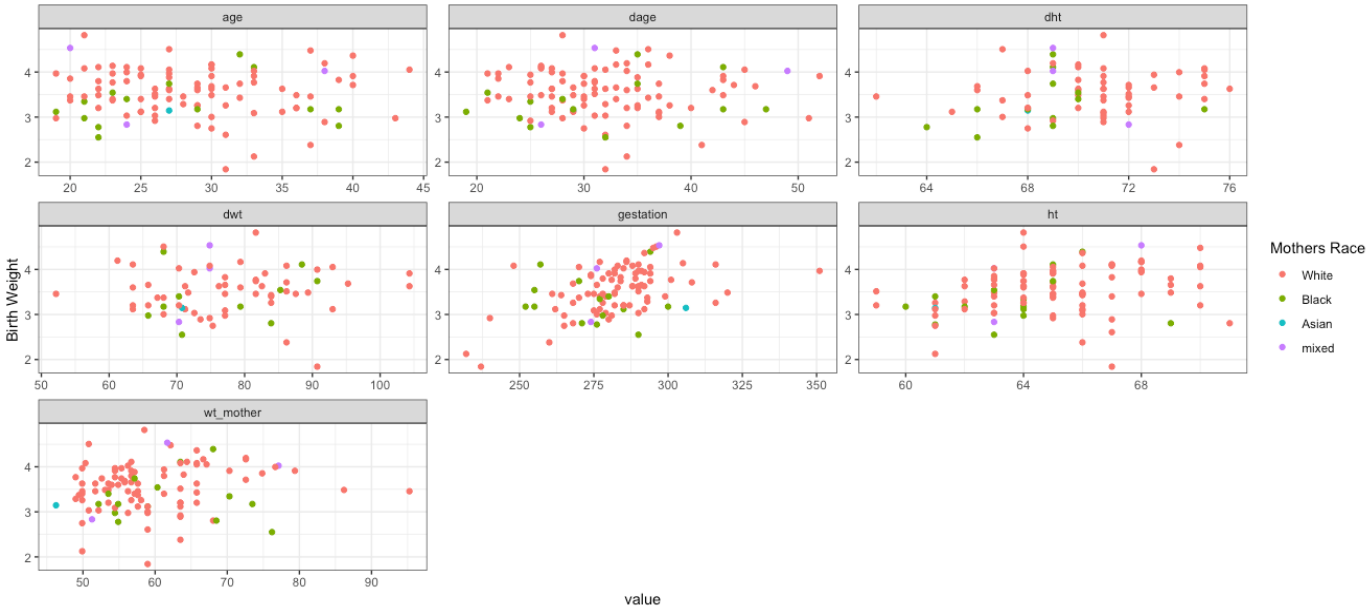


*Figure 9: Scatter plot for birth weight against 7 numerical variables age (mothers age), dage (fathers age), dht (fathers height), dwt (fathers weight), gestation, ht (height), wt_mother (Mothers weight) for mothers who once smoked but not now. The points are colour coded*

### 3.3. Creating a Linear Model

#### 3.3.1. Model 1

Using the backward selection method to determine which explanatory variables to include in the model, model one included the following variables:

- *Gestation* (Figure 10)
- *Parity* (Figure 11)
- *Mothers height* (Figure 12)
- *Mothers race* (Figure 13)
- *Fathers height*
- *Number of cigarettes smoked.* (Figure 14)

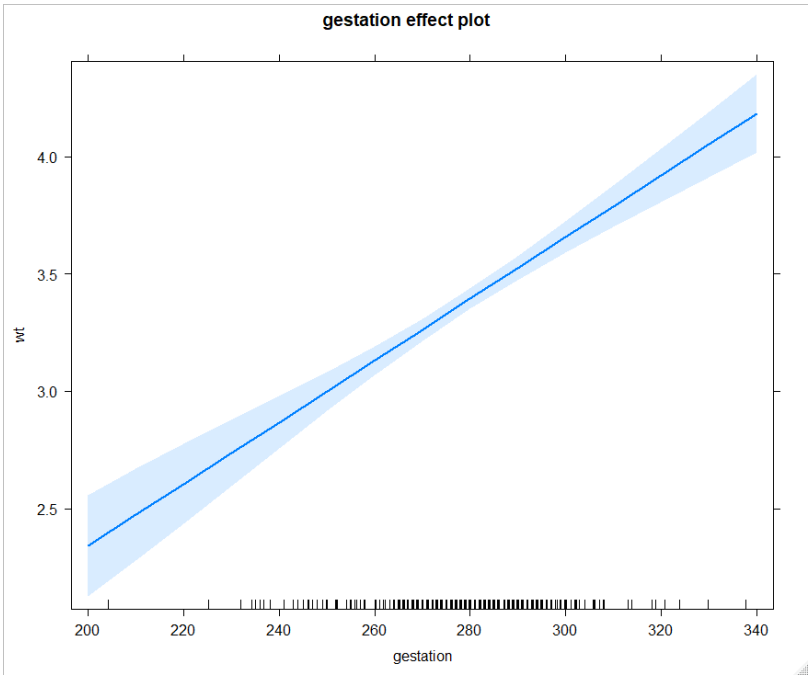


*Figure 10: Gestation period (in days) effect in baby's weight*

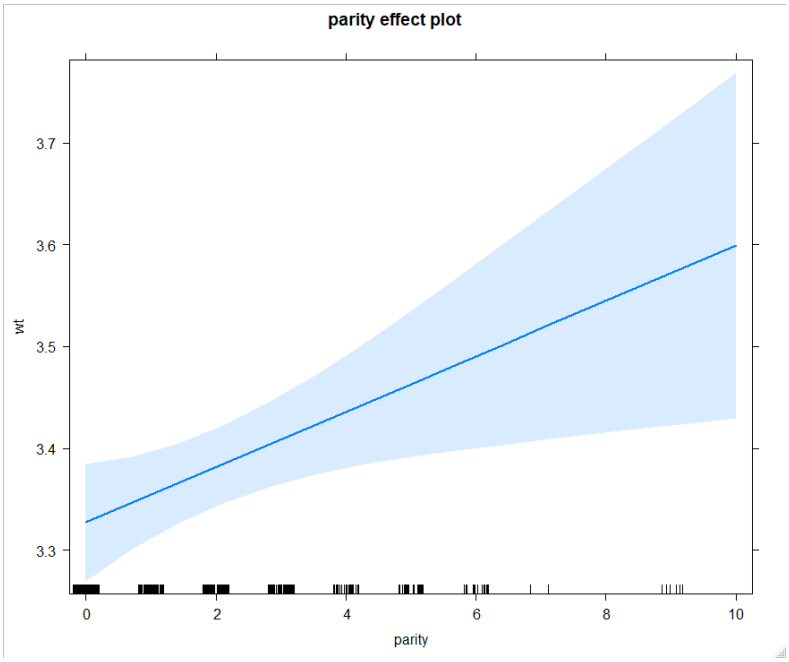


*Figure 11: Parity effect on baby's weight*

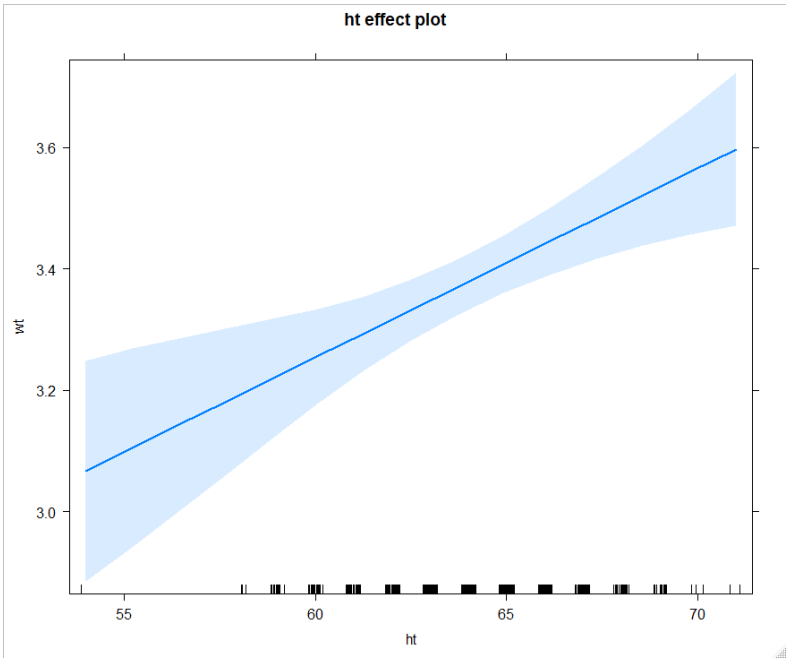


*Figure 12: Mother's height effect in baby's weight*

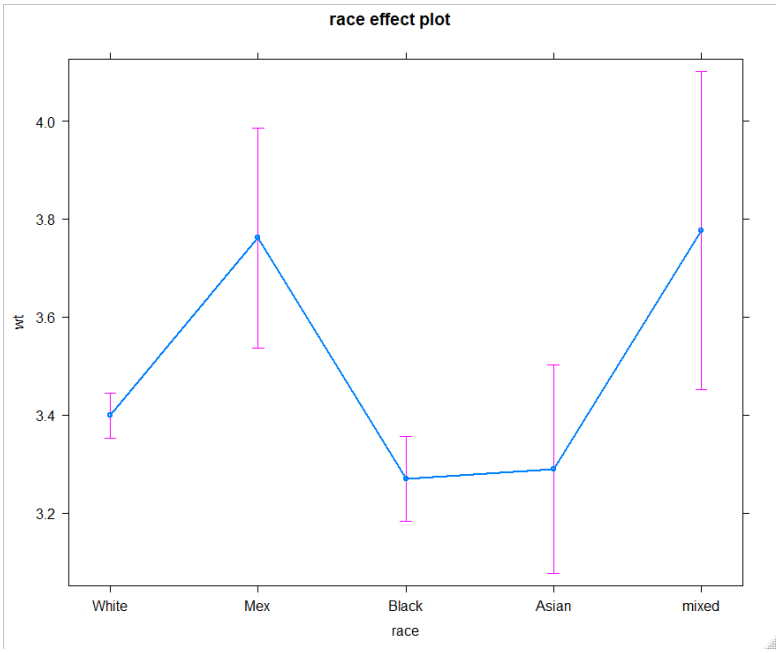


*Figure 13: Race effect on baby's weight*

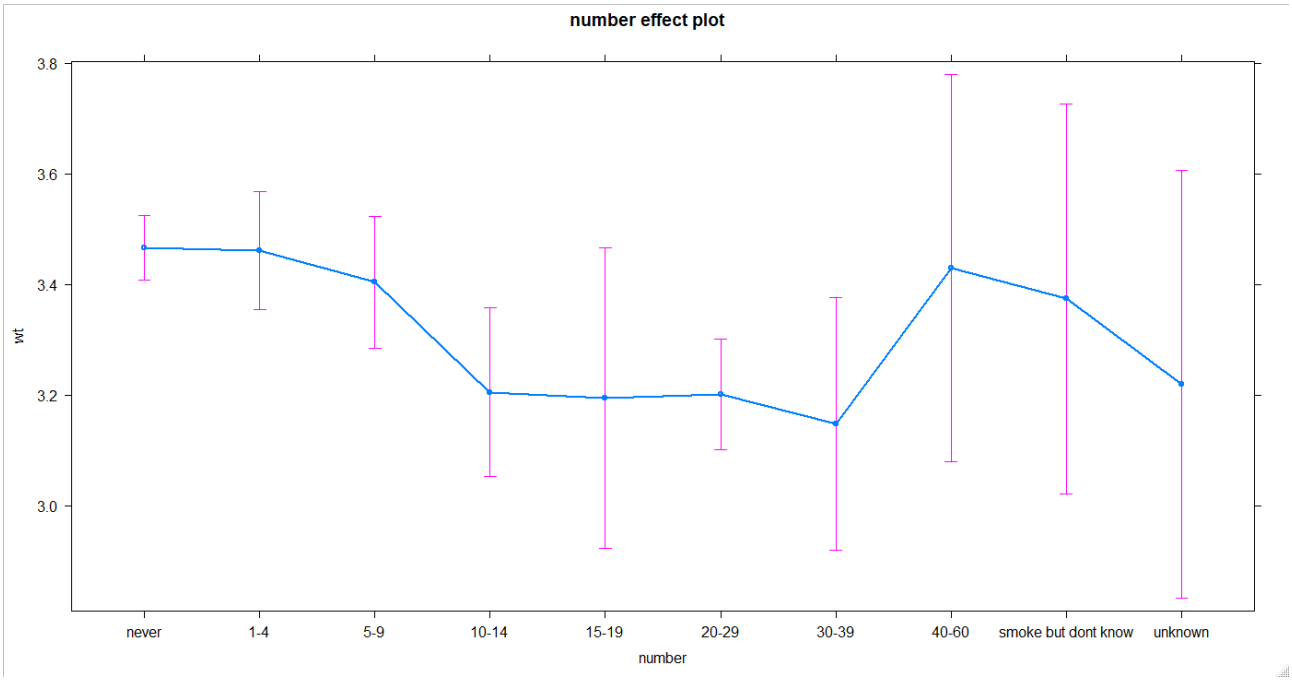


*Figure 14: Number of cigarettes effect on baby's weight*

#### *3.3.1.1. Model 1 Diagnostics*

The results of the ANOVA test on model one (Table 2) show the p value for gestation, parity, mothers' race, mothers height and the number of cigarettes smoked is less than the 0.05 significance level (α) (Rana *et al*., 2013). Therefore, evidence to reject the null hypothesis that there is no difference between the means of the covariates is present. The p-value for the variable Fathers Height is greater than α thus, there is plausible evidence to support the null hypothesis.

*Table 2: ANOVA results for model one. F-values have been rounded to 2 decimal places*

| Factors | Degrees of Freedom | F-value | P-value |
|---|---|---|---|
| Gestation | 1 | 97.75 | < 2.2 e-16 |
| Parity | 1 | 8.92 | 0.0029728 |
| Race | 4 | 6.51 | 4.16 e-05 |
| Mothers Height | 1 | 19.24 | 1.421e -05 |
| Dads Height | 1 | 2.20 | 0.1390217 |
| Number of cigarettes smoked | 9 | 3.75 | 0.001415 |

The results of the VIF test (Table 3) show that all values are low (less than two in all cases). Therefore the probability of an inflated variance of a regression coefficient due to multicollinearity in the model is very low.

*Table 3: VIF test results for model one.*

| Factors | Degrees of Freedom | GVIF | GVIF^(1/(2*DF)) |
|---|---|---|---|
| Gestation | 1 | 1.058771 | 1.028966 |
| Parity | 1 | 1.100260 | 1.048933 |
| Race | 4 | 1.376156 | 1.040719 |
| Mothers Height | 1 | 1.206111 | 1.098231 |
| Dads Height | 1 | 1.199174 | 1.095068 |

Following a Shapiro-Wilks test, a p-value greater than the level of significance was present (p=0.2149 >α (0.05)). Therefore, the null hypothesis that the model residuals are normally distributed cannot be rejected.

The NCV test results (Table 4) show a p-value greater than the 0.05 significance level ($p=0.2149 > \alpha$), hence the null hypothesis for homoskedasticity can be rejected.

*Table 4: NCV test results*

| Model | Chi Square | Degrees of Freedom | p-value |
|---|---|---|---|
| 1 | 0.2927551 | 1 | 0.58846 |

The D-W test results (Table 5) show a D-W statistic less than 2 and a p-value greater than α – evidence suggesting the null hypothesis cannot be rejected and indicated the possibility that there is no autocorrelation between the model residuals.

*Table 5: Durbin-Watson test output*

| Model | Autocorrelation | D-W statistic | p-value |
|---|---|---|---|
| 1 | 0.07329175 | 1.850598 | 0.098 |

Table 6 you can see the confidence intervals for each variable included in model one.

*Table 6: 2.5% and 97.5% confidence intervals. Values have been rounded to 2 decimal places.*

| Factors | 2.5% | 97.5% |
|---|---|---|
| Intercept | -4.93 | -2.01 |
| Gestation | 0.01 | 0.016 |
| Parity | 0.01 | 0.05 |
| Race Mexican | 0.11 | 0.57 |
| Race Black | -0.24 | -0.04 |
| Race Asian | -0.36 | 0.08 |
| Race Mixed | 0.08 | 0.73 |
| Mothers Height | 0.02 | 0.05 |
| Dads Height | -0.00 | 0.03 |
| Numbers 1-4 | -0.13 | 0.12 |
| Numbers 5-9 | -0.20 | 0.07 |
| Numbers 10-14 | -0.43 | -0.10 |
| Numbers 15-19 | -0.55 | 0.01 |
| Numbers 20-29 | -0.38 | -0.15 |
| Numbers 30-39 | -0.56 | -0.08 |
| Numbers 40-60 | -0.40 | 0.32 |
| Numbers who smoke but don't know | -0.45 | 0.26 |
| Number unknown | -0.64 | 0.14 |

Figure 15a shows the fitted values against the residuals. The line fitted through the points seems to have a slope of 0 (relatively flat) suggesting the model captures the signal and only randomness (noise) remains. Figures 15b and 15c are Q-Q plots for the residuals which show the residuals fit well with the straight line suggesting the residuals are normally distributed. This is also supported by the histogram in Figure 15d and the Shapiro test results shown above.

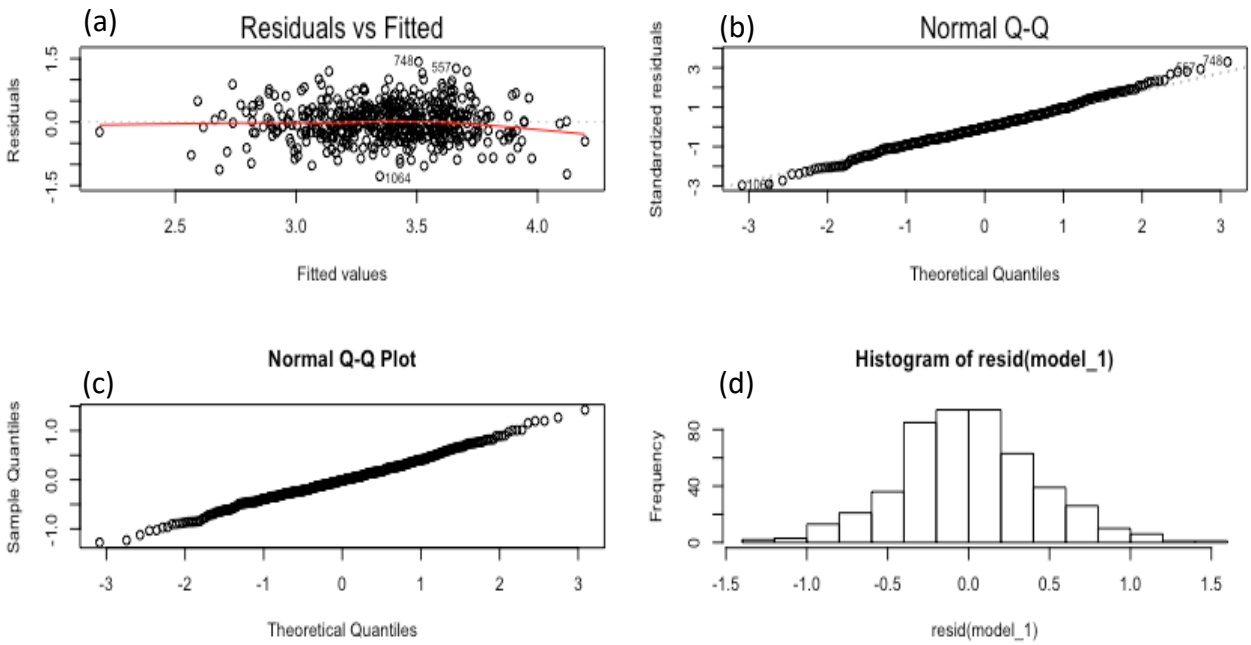


*Figure 15: A selection of model one residual plots. (a) Fitted vs residual plot, (b) and (c) Q-Q plots and (d) a histogram.*

Figure 16 plots partial residuals to examine the relationship between a given independent variable and the response variable given that other independent variables are included in the model. It seems that father's height is relatively insignificant due to the slope of the curve being relatively flat. This is also confirmed by the ANOVA test above which finds father's height to be insignificant. Other than that, all variables are significant in explaining the signal of the observed variable.

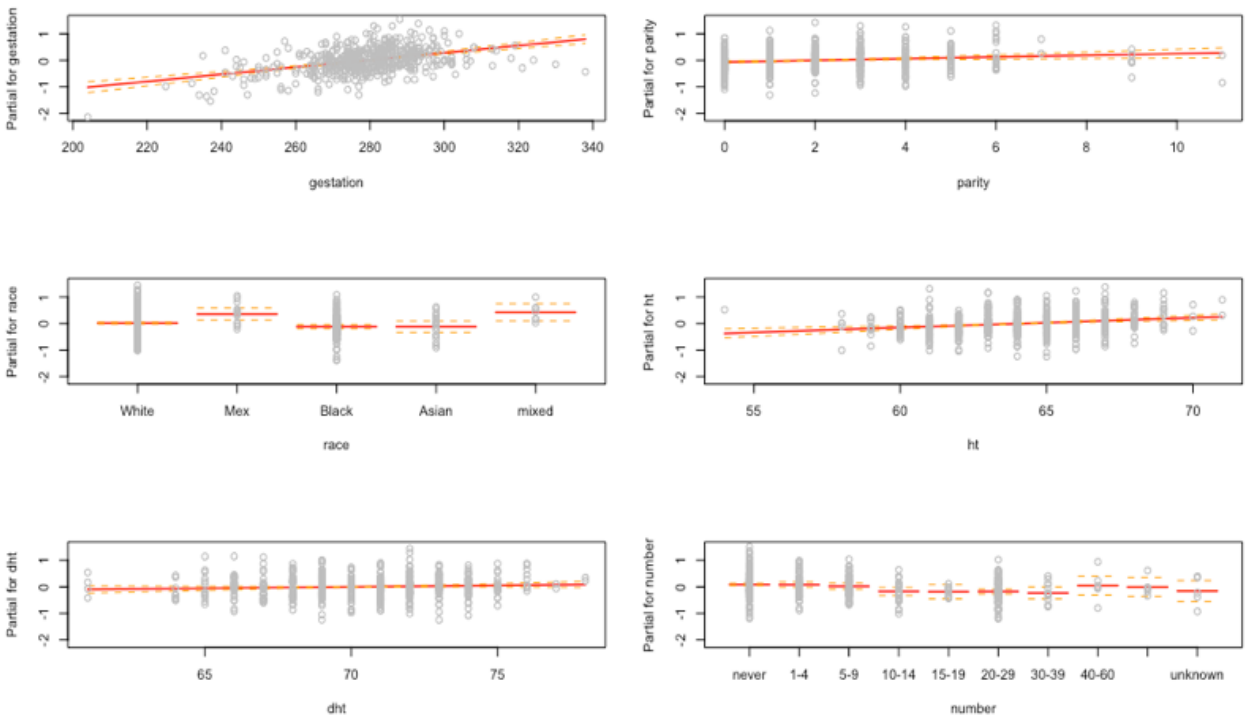


*Figure 16: Partial residuals*

Figure 17 shows the relationship between gestation and birth weight. With a p-value significantly less than 0.05, there is strong evidence to support the alternative hypothesis that gestation affects the weight of the baby. Also, most of the observations are around 250-300 days of gestation which makes practical sense as well.

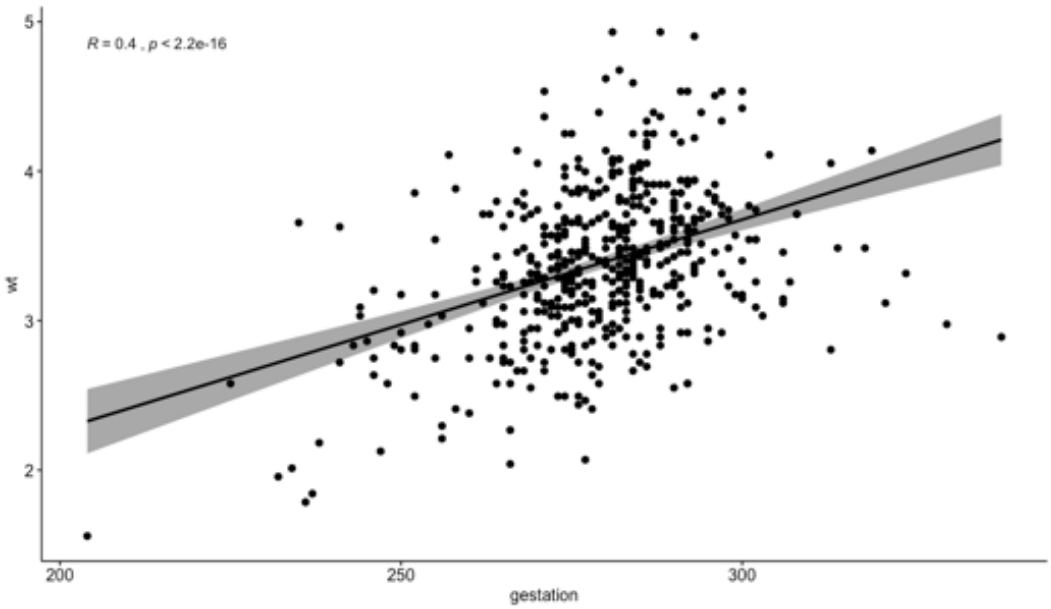


*Figure 17: Relationship between birth weight and gestation in model one*

### *3.3.2. Model 2*

Model 2 was created using the forward selection method for selecting explanatory variables, the following variables were included in the model:

- *Mother's race*
- *Mother's weight* (Figure 18)
- *Father's weight* (Figure 19)
- *Smoke* (Figure 20)
- *Gestation*
- *Parity*
- *Mother's height*

Model 2 includes a slightly different selection of variables than model one. Below the effects of those variables which are included in Model 2 but not in Model 1 are plotted:

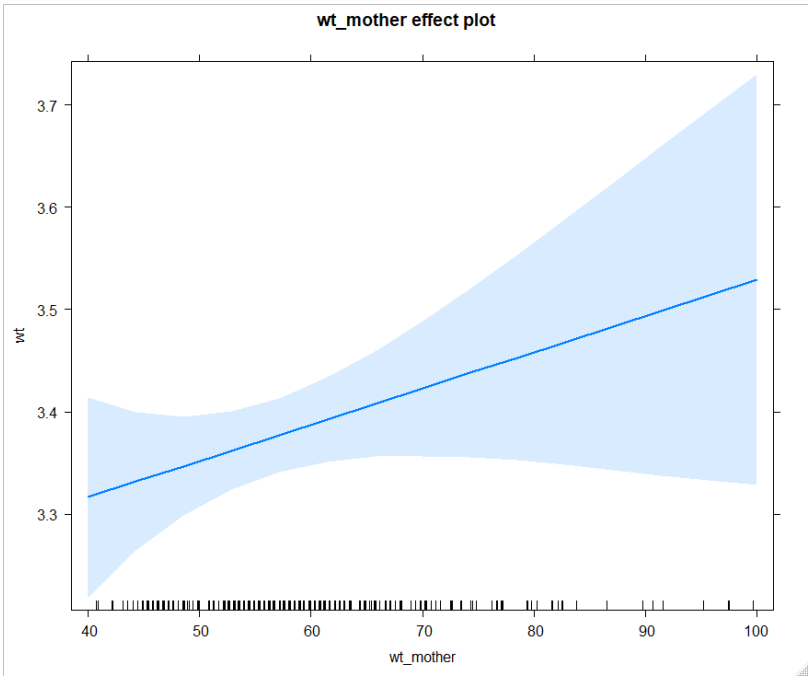


*Figure 18: Mother's weight effect on baby's weight*

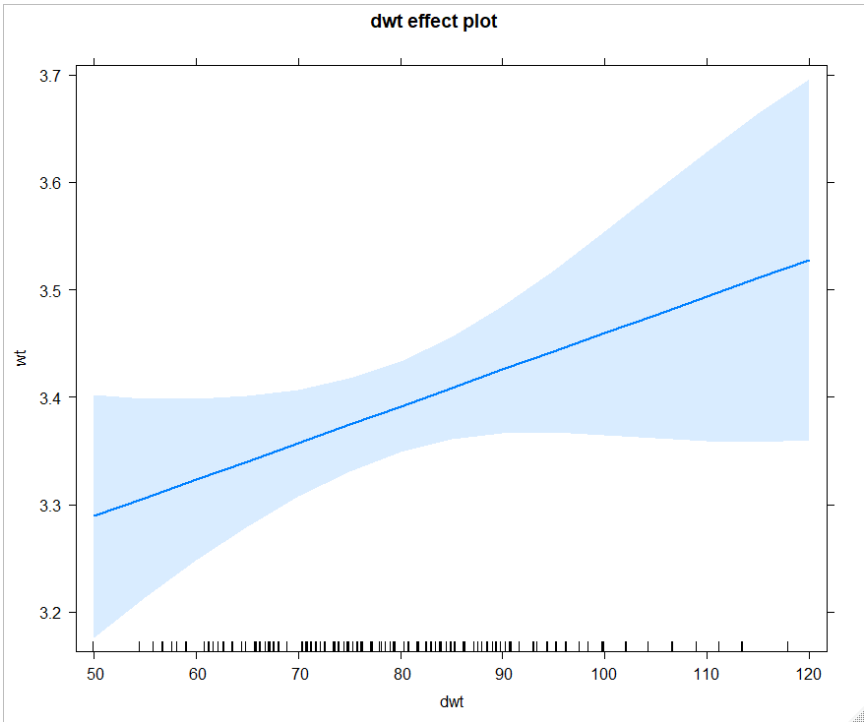


*Figure 19: Father's weight effect on baby's weight*

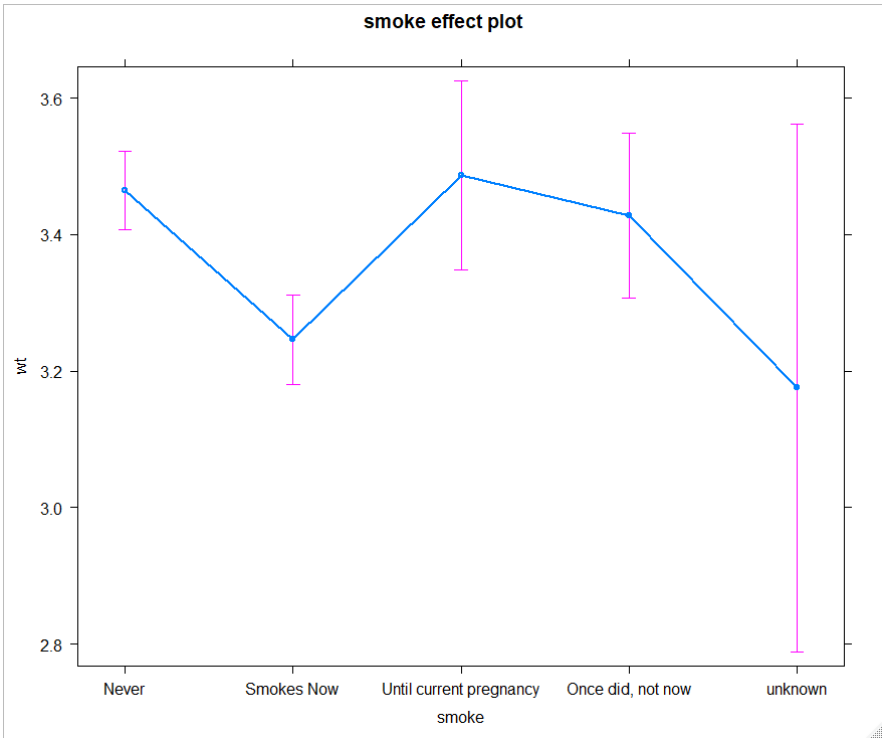


*Figure 20: Mother's smoking habit effect on baby's weight*

#### *3.3.2.1. Model 2 Diagnostics*

Table 7 shows the results of the ANOVA test on Model 2. The p-value for race, smoke, gestation, parity and mother's height are less than the significance level so the null hypothesis is rejected. The p-values for the weight of mother and weight of the father were greater than the significance level. However these were still included in the model as they act as explanatory variables and indicated that AIC becomes smaller.

*Table 7: ANOVA results for model two. F-values have been rounded to 2 decimal places*

| Factors | Degrees of Freedom | F-value | P-value |
|---|---|---|---|
| Race | 4 | 6.27 | 6.379 e-05 |
| Weight of Mother | 1 | 2.12 | 0.1460778 |
| Weight of Dad | 1 | 3.02 | 0.0826929 |
| Smoke | 4 | 6.95 | 1.899 e-05 |
| Gestation | 1 | 93.99 | < 2.2 e-16 |
| Parity | 1 | 6.61 | 0.0104455 |
| Mothers Height | 1 | 12.25 | 0.005097 |

The VIF tests for model two (Table 8) show the values are less than two, suggesting there is no collinearity within the variables (Allison, 1999; Yoo *et al.,* 2014).

*Table 8: VIF test results for model two*

| Factors | Degrees of Freedom | GVIF | GVIF^(1/(2*DF)) |
|---|---|---|---|
| Race | 4 | 1.348022 | 1.038035 |
| Weight of Mother | 1 | 1.383816 | 1.176357 |
| Dads Weight | 1 | 1.131438 | 1.063691 |
| Smoke | 4 | 1.140104 | 1.016525 |
| Gestation | 1 | 1.045863 | 1.022675 |
| Parity | 1 | 1.133581 | 1.064698 |
| Mothers Height | 1 | 1.376836 | 1.173386 |

The Shapiro-Wilk test p-value for Model 2 is greater than the significance level ($p=0.09 > \alpha$), hence the null hypothesis is accepted as plausible (Allison, 1999; Yoo *et al.,* 2014).

The p-value for the NCV test (Table 9) is greater than the significance level, suggesting the null hypothesis is accepted as plausible.

*Table 9: NCV test results*

| Model | Chi Square | Degrees of Freedom | p-value |
|---|---|---|---|
| 2 | 0.1664701 | 1 | 0.68327 |

The D-W statistic (Table 10) was less than two and had a p-value greater than the significance level, thus the null hypothesis was not rejected.

*Table 10: Durbin-Watson test output*

| Model | Autocorrelation | D-W statistic | p-value |
|---|---|---|---|
| 2 | 0.04839412 | 1.900212 | 0.23 |

Table 11 shows the confidence intervals for all the variables in Model 2. Intercept is the variable *Race White.*

*Table 11: 2.5% and 97.5% confidence intervals. Values have been rounded to 2 decimal places.*

| Factors | 2.5% | 97.5% |
|---|---|---|
| Intercept | -3.99 | -1.45 |
| Race Mexican | 0.13 | 0.59 |
| Race Black | -0.23 | -0.03 |
| Race Asian | -0.33 | 0.12 |
| Race Mixed | 0.05 | 0.71 |
| Mothers Weight | -0.00 | 0.01 |
| Dads Weight | -0.00 | 0.01 |
| Smokes Now | -0.31 | -0.13 |
| Smoked Until Pregnant | -0.13 | 0.17 |
| Smoked once, not now | -0.17 | 0.10 |
| Smoke Unknown | -0.68 | 0.10 |
| Gestation | 0.01 | -0.02 |
| Parity | 0.01 | 0.05 |
| Mothers Height | -0.01 | -0.05 |

Figure 21a shows the fitted values against the residuals, the line fitted through the points appears to have a slope of 0 suggesting the model is capturing the signal and all that remains is the randomness (noise). Figures 21b and 21c show Q-Q plots for the residuals. The residuals fit well with the straight line and therefore suggest the residuals are normally distributed. This is also supported by the histogram in Figure 21d and the Shapiro-Wilk test results shown above.

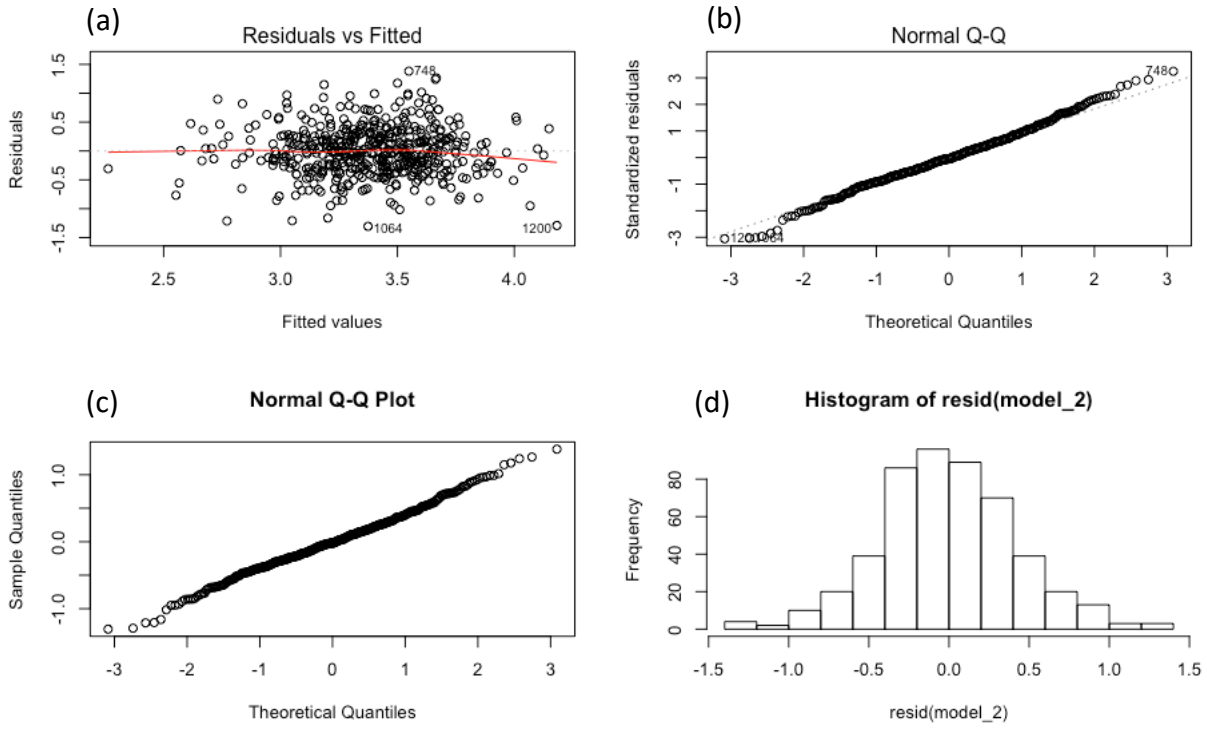


*Figure 21: A selection of model one residual plots. (a)Fitted vs residual plot, (b) and (c) Q-Q plots and (d) a histogram*

When plotting the partial residuals (Figure 22) for all covariates within the model, the weight of the mother and father were deemed insignificant, indicated by the relatively flat curve of the plots. This is also confirmed by the ANOVA test results shown above which find those variables being insignificant. The rest of the variables seem significant.

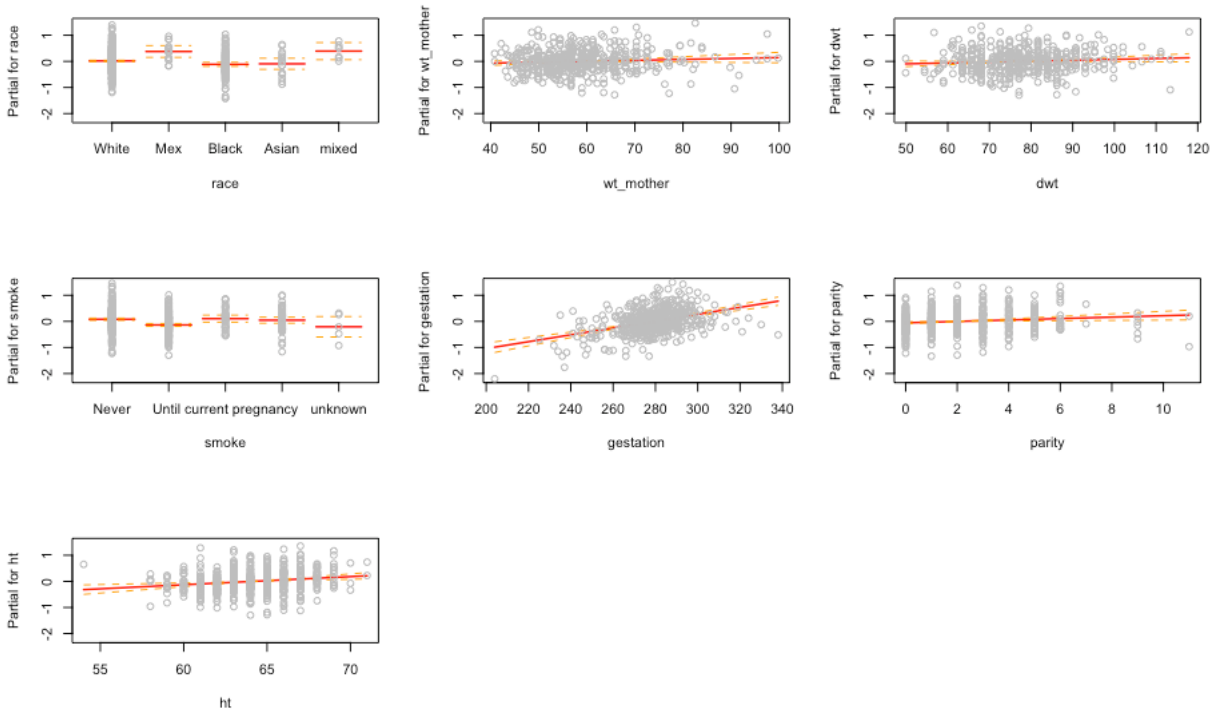


*Figure 22: Partial Residuals*

### *3.4. Selecting the Best Model*

A widely recognised criterion for selecting between models is AIC (Heinze *et al.,* 2018; Wong *et al.,* 2014). The AIC for model 2 is lower than model 1 (Table 12), therefore, model 2 would be selected as the best model to capture the response variable Oxygen.

*Table 12: AIC values used to choose between model 1 and 2*

| Model | AIC Value |
|---|---|
| 1 | 598.2940 |
| 2 | 592.8881 |

### 3.5. Model Validation

#### 3.5.1. Validation the Set Approach

Table 13 shows the Mean Square Error (MSE) results from the set approach validation test. Model 2 showed the lowest MSE value suggesting that its explanatory power is more accurate than the model in comparison (Kassambara, 2018).

*Table 13: MSE values from the validation set approach method*

| Model | MSE |
|---|---|
| 1 | 0.2134067 |
| 2 | 0.2075065 |

#### 3.5.2. 5-fold Cross-Validation

Table 14 shows the Root MSE (RMSE), R squared value, Mean Absolute Error (MEA), values from the 5-fold Cross-validation (Kassambara, 2018). The RMSE and MAE values for the models are very similar, however the values for model two are lower than for model one which suggests model two is a better model (Kassambara, 2018). The R-squared value for model two is higher which again suggests it is a better model.

*Table 14:RMSE, R-Squared and MAE values from the 5-fold Cross-Validation*

| Model | RMSE | R-Squared | MAE |
|---|---|---|---|
| 1 | 0.4499902 | 0.2568768 | 0.3545067 |
| 2 | 0.4493301 | 0.2615274 | 0.3527888 |

### 3.6. Bootstrapping

Table 15 shows the confidence intervals from the bootstrapping undertaken on model two. In comparison to the confidence intervals from Model 2 in Table 11, they are very similar.

*Table 15: 2.5% and 97.5% confidence intervals for the bootstrapping of model two*

| Factors | 2.5% | 97.5% |
|---|---|---|
| Intercept | -4.10 | -1.37 |
| Race Mexican | 0.16 | 0.57 |
| Race Black | -0.23 | -0.02 |
| Race Asian | -0.33 | 0.11 |
| Race Mixed | 0.15 | 0.62 |
| Mothers Weight | -0.00 | 0.01 |
| Dads Weight | -0.00 | 0.01 |
| Smokes Now | -0.31 | -0.13 |
| Smoked Until Pregnant | -0.12 | 0.16 |
| Smoked once, not now | -0.19 | 0.10 |
| Smoke Unknown | -0.80 | 0.18 |
| Gestation | 0.01 | 0.02 |
| Parity | 0.01 | 0.05 |
| Mothers Height | -0.01 | -0.05 |

## 4. Conclusion

Based on this analysis, it can be stated that baby's weight is affected by almost all the variables in the given dataset except for date, father's age and mother's age (correlation, ANOVA results). The gestation period appears to be the most correlated variable in estimating the weight of the baby. Furthermore, with a strong positive correlation of 0.40, it can be inferred that babies with low birth weight are more likely to be observed when the mother's gestation period is shorter (less than 9 months). This statistical observation is also consistent with the practical literature (Gathwala et al., 2008). As the gestation period increases, the weight of the baby increases moderately. With the effect plots from models 1 and 2 there is a positive effect on baby weight in instances where the mother's race is Mexican or mixed. On the contrary, if the race of the mother is Asian, black or white, the baby tends to weigh less relative to the other 2 races with some margin for error in the race Asian. This observation is also consistent with the existing literature that finds race to be a factor (Conley and Bennett, 2000). Furthermore, the results indicated that mother's smoking habits during the pregnancy has a negative effect on baby's weight – a conclusion previously stated by MacMahon *et al.,* (1965). If the mother smokes during the pregnancy, the baby is more likely to weigh less relative to if the mother has never smoked or did but not during the pregnancy. Existing literature also support this finding (Meyer, 1978). In addition, the mother's height and parity both have a positive effect on the baby's weight.

Two models were constructed during the analysis following two different methods and came up with the following explanatory variables for each:

M1: wt ~ gestation + parity + race + ht + dht + number
M2: wt ~ race + wt_mother + dwt + smoke + gestation + parity + ht

Both models passed all the diagnostics with relative ease however model 2 was selected as the final model based on the AIC measurements between the two. From the 5-fold cross validation of model 2, an R-Squared value of 0.2615 with RMSE=0.4493 and MSE=0.3520 was identified. This low-medium explanatory power of the model with relatively high margin for error indicates that the dependent variable (weight of the baby) can be identified 26% of the time, based on having these explanatory variables.

## 5. Discussion

Although one could question the practical significance of this model, from a statistical point of view, the model has significant explanatory power. Moreover, despite the widely searched high R squared value in modelling researches, it is noted that R squared statistic measures the significance of the slope of the regression rather than measure the goodness of fit (Fonticella, 1999). Although a high R squared value is desirable, other factors are also important before rejecting a model such as the pattern in the residuals and the significance from the ANOVA (Fonticella, 1999). With that being said, model 2 checks both points.

Further to the investigation, in many cases, especially in social or behavioural studies low R-square is common since not all relevant predictors are accounted for from the given dataset. This suggests that the dependent variable is affected by other factors in addition to the ones considered in this analysis (Ferenc,1999).

To enhance research within this study area, it is recommended socioeconomic variables are added, such as education and economic status of the family as part of the analysis. Previous studies on low-birth weight babies suggested that socioeconomic background of the parents plays an important role (Conley and Bennett, 2000). Thus, via incorporating them as potential explanatory variables, the practical significance in the predictive power of the model is possible to increase.

## 6. References

*Literature*

**Abrevaya, J.** (2006). Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach. *Journal of Applied Econometrics*, 21(4), pp.489-519.

**Adegboye, A., Rossner, S., Neovius, M., Lourenço, P. and Linné, Y.** (2010). Relationships Between Prenatal Smoking Cessation, Gestational Weight Gain and Maternal Lifestyle Characteristics. *Women and Birth*, 23(1), pp.29-35.

**Allison, P.** (1999). *Multiple Regression: A Primer*. 2nd ed. Pennsylvania: SAGE Publications, pp.43-61.

**Austin, P.C.** (2008). Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in medicine*, 27(17), pp.3286-3300.

**Benedetti, J.K. and Brown, M.B.** (1978). Strategies for the selection of log-linear models. *Biometrics*, pp.680-686.

**Breusch, T.S. and Pagan, A.R.** (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pp.1287-1294.

**Brusco, M., Steinley, D.** (2015). Phychometrics: Combinatorial Data Analysis. in: Wright, J.D. (ed). International Encyclopaedia of the social and behavioural sciences. [Online] Amsterdam: Elsevier, pp.431-435. Available at: https://www.sciencedirect.com/referencework/9780080970875/international-encyclopedia-of-the-social-and-behavioral-sciences#book-info [Accessed on: 4/11/19].

**Chen, Y.** (2016). Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression. *PLOS ONE*, 11(1), pp.146-165.

**Conley, D. and Bennett, N.** (2000). Is Biology Destiny? Birth Weight and Life Chances. *American Sociological Review*, 65(3), p.458.

**Efron, B. and Tibshirani, R.J.** (1994). An introduction to the bootstrap. *Boca Raton: CRC press*, pp. 168-177.

**Fonticella, R.** (1999). The UsqUrzess of the R2 Statistic by Ross Fonticella, ACAS.

**Fox, J.** (2002). Bootstrapping regression models. *An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book. Sage, Thousand Oaks, CA.* [Online] Available at: *http://cran. r-project. org/doc/contrib/Fox-Companion/appendix-bootstrapping. Pdf* [Accessed on: 4/11/19]

**Fox, J.,** (2015). Applied Regression Analysis and Generalised Linear Models. *Thousand Oaks: Sage Publications*, pp. 647-669

**Fushiki, T.** (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), pp.137-146.

**Gathwala, G., Singh, B. and Balhara, B.** (2008). KMC Facilitates Mother Baby Attachment in Low Birth Weight Infants. *The Indian Journal of Paediatrics*, 75(1), pp.43-47.

**Heinze, G., Wallisch, C. and Dunkler, D.** (2018). Variable Selection - A Review and Recommendations for the Practicing Statistician. *Biometrical Journal*, 60(3), pp.431-449.

**Kassambara, A.** (2018). *Regression Model Validation – Cross-Validation Essentials in R.* [Online] Available at: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r [Accessed on: 4/11/19]

**Lawoyin, T.** (2001). Risk Factors for Infant Mortality in a Rural Community in Nigeria. Journal of the Royal Society for the Promotion of Health, 121(2), pp.114-118.

**MacMahon B, Alpert M, Salber EJ.,** (1966). Infant Weight and Parental Smoking Habits. *Am J Epidemiol,* 82(1), pp.247–261

**Makhija, K., Murthy, G.V.S., Kapoor, S.K. and Lobo, J.** (1989). Socio-biological determinants of birth weight. *The Indian Journal of Paediatrics*, 56(5), pp.639-643.

**Meyer, M.** (1978). How does maternal smoking affect birth weight and maternal weight gain?. American Journal of Obstetrics and Gynecology, 131(8), pp.888-893.

**Moksony, F**. (1999). Small Is Beautiful: The Use and Interpretation of R2 in Social Research. Szociologiai Szemle, pp. 130-138.

**Moore, D.S., McCabe, G.P., Craig, B.A.** (2017). *Introduction to the practice of statistics*. 9th ed. New York: Macmillan Education, pp. 643-697.

**Negi, K.S., Kandpal, S.D. and Kukreti, M.** (2006). Epidemiological factors affecting low birth weight. *JK Science*, 8(1), pp.31-4.

**Nerlove, M. and Wallis, K.F.** (1966). Use of the Durbin-Watson statistic in inappropriate situations. *Econometrica: Journal of the Econometric Society*, pp.235-238.

**O'brien, R.** (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5), pp.673-690.

**Pavlou, M., Ambler, G., Seaman, S.R., Guttmann, O., Elliott, P., King, M. and Omar, R.Z.** (2015). How to develop a more accurate risk prediction model when there are few events. *BMJ*, *351*, pp.38-68.

**Rana, R., Singhal, R. and Singh, V.** (2013). Analysis of Repeated Measurement Data in the Clinical Trials. *Journal of Ayurveda and Integrative Medicine*, 4(2), pp.77-82.

**Rochon, J., Gondan, M. and Kieser, M.** (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC medical research methodology*, 12(1), p.81.

**Rosopa, P., Schaffer, M. and Schroeder, A.** (2013). Managing heteroscedasticity in general linear models. *Psychological Methods*, 18(3), pp.335-351.

**Salamon, S., Hansen, H. and Abbott, D.** (2019). How real are observed trends in small correlated datasets?. *Royal Society Open Science*, 6(3), pp.181-89.

**Shao, J.** (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422), pp.486-494.

**Visscher, W.A., Feder, M., Burns, A.M., Brady, T.M. and Bray, R.M.** (2003). The impact of smoking and other substance use by urban women on the birthweight of their infants. *Substance use & misuse*, 38(8), pp.1063-1093.

**Walker, M.B., Tekin, E. and Wallace, S.** (2009). Teen Smoking and Birth Outcomes. *Southern Economic Journal*, 75(3), pp.892-907.

**Williams, R.** (2015). *Heteroskedasticity.* [Online] Available at: https://www3.nd.edu/~rwilliam/stats2/l25.pdf [Accessed on: 4/11/19].

**Wong, Y., Li, C. and Chen, B**. (2014). Evolution of Network Biomarkers from Early to Late Stage Bladder Cancer Samples. *BioMed Research International*, 14(1), pp.1-23.

**World Health Organisation** (2012). International Statistical Classification of Diseases and Related Health Problems, Volume Two. *10th ed*. Malta: World Health Organisation, pp.140-164.

**Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q.P. and Lillard Jr, J.W.** (2014). A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, 4(5), p.9-15.

**Zaman, A.** (2000). Inconsistency of the Breusch-Pagan test. *Journal of Economic and Social Research*, 2(1), pp.1-11.

**Zuur, A., Ieno, E., Walker, N., Saveliev, A. and Smith, G.** (2009). *Mixed Effects Models and Extensions in Ecology with R*. 1st ed. New York, NY: Springer New York.

*R Packages*

**Bartoń, K.** (2019). MuMIn: Multi-Model Inference. R package version 1.43.6. https://CRAN.R-project.org/package=MuMIn

**Fox, J. and Weisberg, S.** (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks, CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

**Fox, J. and Weisberg, S.** (2019). An R Companion to Applied Regression, 3rd Edition. Thousand Oaks, CA. http://tinyurl.com/carbook

**Kassambara, A.** (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.3.https://CRAN.R-project.org/package=ggpubr

**Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Brenton Kenkel, the R Core Team *et al.,*** (2019). Caret: Classification and

Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

**R Core Team** (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

**Schloerke, A., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J.** (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally

**Wei, T. and Simko, V.** (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

**Wickham, H and Lionel Henry**. (2019). Tidyr: Tidy Messy Data. R package version 1.0.0. https://CRAN.R-project.org/package=tidyr

**Wickham, H, François, R., Henry, L and Müller, K.** (2019). Dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

**Wickham, H.** (2017). Tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

**Wickham, H.** Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

**Yan, Y.** (2016). MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. https://CRAN.R-project.org/package=MLmetrics