**Linear Regression Models Describing Infant Birth Weight**

**Executive summary**

Low birth weight is one of the most common indicators of poor infant health. While it is not this report’s objective to establish causation or identify main factors, we are trying to find variables which might correlate with low birth weight. A total of 1236 babies from a US hospital were evaluated with applied statistical methods including  analysis of variance (ANOVA) tests and simple linear regression models.

ANOVAs were used to compare the infant weights based on different smoking behaviours of the mothers, as well as race and social factors for the parents. The linear regression model intends to firstly describe the relationship of weights and other variables to create a prediction model for children’s birth weights based on significant variables.

Model selection methods included using the Akaike Information Criterion, and sing  backward selection using Adjusted R squared values and type II ANOVA to determine the best model. Cross validation was used to select the final model. The variables in  the final model were subjected to bootstrapping to determine confidence intervals for each parameter.

The final model included relationships with infant weight and pregnancy duration, the parents’ race, the parents’ weight, the mother’s height, and whether or not the mother smoked during the pregnancy. How much the mother smoked was not found to significantly affect the infant’s weight, only whether or not they did.

All these findings are highly limited in its application because of the limited sampled population, as it is very small and collected in 1964, and includes data only for single, live, male births. Further research in this topic could deepen our understanding of these connection but would fail to help in preventing low birth weights. Nonetheless, the findings in this report allow to point research into an adequate direction. The contribution of this analysis is to identify mother’s smoking habits and ethnicity as possible determinants of children birth weight along with other attributes that might affect birth weight.



[**Executive summary 1**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.dq15lbslcs9c)

[**Introduction 3**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.2ooarkws53vm)

[**Data 3**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.18rsdt1n53dz)

[**Exploratory Analysis 4**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.vu50gtjjv622)

[Density 4](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.emncoxw8poel)

[Ethnicity 6](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.u21xxcn15qrk)

[Mother’s Weight 7](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.6l4ntesh5xcr)

[Mother’s Education 8](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.4gzgwat7o3fn)

[**3.Methods 10**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.b7tlzx1alux9)

[3.1 Linear Regression 10](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.n3rsxv6cr1h2)

[3.1.1 Model Selection 10](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.ea1invazcdq7)

[3.1.2 AIC 11](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.rnm3jtfrpi18)

[3.1.3 Adjusted R-squared 14](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.w88vmch090tk)

[3.1.4 Mean Squared Error 15](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.8ieycjgqp1hn)

[**4. Results 15**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.7md9uaccywvi)

[4.1 Assumptions 18](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.hgslriv48lka)

[4.1.1 Linearity Assumption 18](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.xv31wl6f67t0)

[4.1.2 Normality Assumption 18](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.pgzmfq3cb387)

[4.1.3 Other Assumption tests 20](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.wvfkegc1t0du)

[**Discussion 20**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.e6c0bru205k0)

[**References 21**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.58idmqgdw9cw)

[**Appendices 21**](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.bgyhp8k3xig0)

[Appendix 1 - Data cleaning 21](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.6uvc5rgfb8p)

[Appendix 2 - AIC model generation 22](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.mvkwb9wmdj0z)

[Appendix 3 - R2 model generation 23](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.8oooog6yiopn)

[Appendix 4 - Mean squared error function for Model Selection 24](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.wvosf1wii1yo)

[Appendix 5 - Model assessment (Results) 24](https://docs.google.com/document/d/1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30/edit#heading=h.9gu9ijstt8fz)

**Introduction**

Previous studies shown low birth weight “is the single most important factor affecting neonatal mortality and a significant determinant of post-neonatal mortality” (Neggers & Crowe, 2013). Naturally, there is an elevated interest of parents to avoid and therefore also to explore the determinants of low birth weights. In this study we are aiming to test certain factors and their relationship to birth weights. This report investigates the effects of mothers’ attributes on the weight of newborn babies, in order to make possible future suggestions about prevention of low birth weights.

**Data**

The data set used for this report consisted of 1236 observations of single live male births from a US hospital in the 1960s. After reformatting the data and removal of all rows containing missing data, 608 observations remained of the following variables:

* **wt**        - Baby's birth weight (Ounces)
* **mwt**       - Mother's weight (Pounds)
* **dwt**       - Father's weight (Pounds)
* **date**      - Baby's birth date (Date)
* **gestation** - Difference in pregnancy length from average (Days)
* **parity**   - Number of mother's previous pregnancies (Number of pregnancies)
* **race**      - Mother's ethnicity (5 Level Factor: White/Mexican/Black/Asian/Mixed)
* **drace**    - Father's ethnicity (Same as above)
* **age**       - Mother's age (Age)
* **dage**      - Father's age (Age)
* **ed**        - Mother's level of education (7 level Factor: Less than 8th grade/8th -12th grade - did not graduate/HS graduate-no other schooling/HS+trade/HS+some college/College graduate/Trade school HS unclear)
* **ded**       - Father's level of education (Same as above)
* **ht**        - Mother's height, rounded down (Inches)
* **dht**       - Father's height, rounded down (Inches)
* **marital**   - Status of parent's relationship (5 level factor: Married/Legally Separated/Divorced/Widowed/Unmarried)
* **inc**       - Income of family in $2500 increments, (\$2500 dollars)
* **smoke**     - Whether the mother smokes (4 level Factor: Never/Smokes Now/Until Pregnancy/Once did)
* **time**      - When the mother last smoked (9 level Factor: Never/Still/Until Pregnancy/<1 yr/1-2 Yrs/2-3/3-4 Yrs/5-9 Yrs/10+ Yrs/Quit)
* **number**    - Number of cigarettes smoked per day by mother while she smoked (8 level Factor: Never smoked/1-4/5-9/10-14/15-19/20-29/30-39/40-60)

The raw data also contained variables for the sex of the infant, the plurality of the pregnancy, and the outcome or whether or not the infant survived for at least 28 days. However, the values for these variables were, for all rows were the same. Because of this, the data used is only for male infants which survived at least 28 days from single births.

As we are hoping to extrapolate descriptive models from this data about births in general, it is of particular importance to note that the data set contains only male infants from a US hospital in the 1960s. As several variables considered are largely social measures, the time and place of these measurements are relevant and may limit how generalizable the derived models are. The sex of the child may similarly affect the relationships with the biological variables.

On conducting cleaning of the provided data a number of questions arose.

* Were only male babies sampled in the research?
* What does the ‘999’ in the variable Gestation stand for?
* What race does the numeric 10 represent in the categories race?
* In the marital column  0 has been used to specify what status of relationship of parents?
* If we assign values  to the family yearly income in $2500 increments the variables already given to us have incorrect values. Since we reach the interval ‘12,500 - 14,999’ by the time we reach numeric 5. What is the correct interval for each numeric assigned?

All analysis was performed using R Studio (RStudio Team,2015; R Core Team, 2018).

**Exploratory Analysis**

An initial descriptive statistical summary is conducted to explore potential relationships in the dataset. This section looks at the frequencies of baby’s weights across factors, and examines some specific variables with relationships suggested by preliminary plots. The exploratory data analysis was produced with the help of R software and plot features specifically.

We assumed each baby is distinct therefore, assuming each baby to be independent.

**Density**

In this section we are having a brief summary of densities across our dataset. The first density plot, Figure 1, describes the distribution of infants by their mother’s race. The distributions for each group are roughly symmetric. In the plot there seems to be a difference in mean of weights, for example between white and asian ethnic groups, this will be tested in further sections using inferential tests.

|  |
| --- |
| https://lh5.googleusercontent.com/jAmDS34BTPyX7w7-gxWiPAryxpUkPXHWR9wOt9g0PrxIkTJAKyoGloYFUp6OZosTvhtN6bCTK1jL0CGhwKAtyJEHdU227jqjdN-BndnSzopuANE29AIC3CwKEAeiMOMizp5RGqmx |
| Figure 1: Histogram of Birth Weights by Ethnicity |

|  |
| --- |
| https://lh4.googleusercontent.com/_MvjLCrVwmEvOg1V2Opm4wBJXVlWtLdWnKatUsP19g5sSczQY32ndNbW4mWTiJXQbG-3p5rnwkEGJeHpyzMeF3MWPWU4h_oXBfVvZkG8YsVAWSju1Ey7cRoFPhWy1kDR97-s0RZ2 |
| Figure 2: Histogram of Birth weights by Mothers Smoking Status.  Never, Once did-not now, Smoking now and until current pregnancy |

We expected to see relationships between infant weight and the mother’s smoking habits, and the distributions for each smoking group is shown in Figure 2. It is clear that due to missing values in the dataset the number of NAs are taking a considerable portion. At the same time it suggests that the mean weight for infants whom mothers have never smoked is higher than mothers who smoke during the pregnancy. Also, the number of mothers who have never smoked and mothers who are smoking now are much higher in our dataset than other groups, according to figure 2.

Finally, the last density plot in this section shows the distribution of the infant’s weight across this dataset which symmetric and seems to follow a normal distribution.

|  |
| --- |
| https://lh6.googleusercontent.com/1auqQS5b4BKOjZ_5r8YE1qQzxt9InmV8TlcmbIh0DXi_7h3p2_stajDJcT95EC7JF0JTGu2pd94vAIuQ3nGE5wlanNPMIvRz1j-q03tft9a3641OWybRL5mp_HMOLkFlS5bJm2M8 |
| Figure 3: Histogram of Infants Weight |

**Ethnicity**

Table 1 shows a summary babies weights across ethnic groups of their mothers. The mean birth weights for ethnic White and Mexican ethnic groups show a central tendency since their means are higher and close to each other. There is similarity of spread for each group with the figures obtained.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mother Race** | **Mean** | **Standard Deviation** |
| **1** | Asian | 110.4318 | 15.99331 |
| **2** | Black | 113.2377 | 19.08851 |
| **3** | Mexican | 124.15 | 14.14313 |
| **4** | Mixed | 119.8 | 20.14324 |
| **5** | White | 121.6414 | 17.69685 |
| **6** | NA | 116.8462 | 16.73741 |

Table 1: Mean and Standard deviation per mothers race

Figure 4 illustrates the distribution of the birth weights by race. The distributions for each group are roughly symmetric and unimodal. In the histogram there seems to be a difference in mean newborn weights across white and black or asian ethnic groups, this will be tested in further sections using inferential tests.

The scatter plot below also shows that the vast majority of babies measured were white.

|  |
| --- |
| https://lh4.googleusercontent.com/P8ux26sTqXkMxtB2eGwW_J7nZStIuPK6VEj5mykmeLWE_reX4l9ZqTLDKOupyQ985uF4qGqRQta7kO-0l5FYUu0lpJMJIdGc3WDgVqfcg1_SuW5NIxu4IQjM6WoFzZa-mbXLRYSr |
| Figure 4: Scatter Plot of Mothers and Babies Weight by Ethnicity |

**Mother’s Weight**

Investigating and exploring relationships between the weights of mothers and babies we managed to plot relationship between weight of mothers and babies delivered by below scatterplot. In below figure, the response variable, baby’s weight, is displayed against the mother’s weights. There is very weak relationship as there is no clear evidence that the points are tightly clustered, and several outliers interfere the overall pattern.

|  |
| --- |
| https://lh3.googleusercontent.com/45Y4q59hnzyF022GO_aRIjwLz0wrNDUpqrA_RXbhnnNR_t1t1GycucYncGADNDvVR3DAlwEaFKXvs-KwoMGTJCLzH2SudbzfhSeuXuAhS_NM-AgmqVitYGZPZ-3bhB8AI-0xtoV1VmnVDnKUSA |
| Figure 5: Weights of Mothers and Babies |

**Mother’s Education**

As a part of exploration we had the assumption of if mothers have higher education they might have infant with a higher mean weight which might show a possible correlation that describe healthier infants.

Therefore, we tried to reorganise the factored variable of mother’s education (med) in a way that levels are sorted from lower to higher education. Below plot shows the relation that mother with high school, trade, and college educations have less variation in their babies’ weights. This could be due to less measurements. However, no difference in the mean is immediately apparent. Further testing will examine this further.

|  |
| --- |
| https://lh5.googleusercontent.com/FIq0sIXznnibJEQSaw6Phc6iYupcPP9g5jeGty8FVPCkRk14nzxLqym-zQoxJgTnHzpEtfph_gfOMYoK7Z0XJJ-NrAFTfb6sMHiJwp8wneUv3H9vTQhAoOLYfYN-D1Vttbhmshov |
| Figure 6: Scatter Plot of Mothers Education and Infants Weight |

|  |
| --- |
| https://lh4.googleusercontent.com/Lt2wbm-Z0B9gBPX3L_E4NbooWQcptkf9Tl0EZhEmk8dTH-Lb6DqAB_GzzH-FFMAwBo55sPRfpIr-7sXIzn_lmZ4k-GavxI-to2z8CFKvCgxeKi2e1sjdNJRa4F3CxwrjjvRI4-dx |
| Figure 7: BoxPlot of Mothers Education and Infants Weight  Red points represent mean. |

**3.Methods**

**3.1 Linear Regression**

**3.1.1 Model Selection**

Various models were developed for the relationship between babies’ weight and the variables contained within the data-set. 20% of the data was held off as a validation data-set to make a final assessment on the models using their mean squared error.

**3.1.2 AIC**

The variables were first split into three categories: Smoking, Health, and Race & Social factors. Each was individually examined and AIC used to judge the models. The smoking variables included the variables for smoking status of mother, number of cigarettes smoked, and time the mother quit smoking. Health included variables for parents’ heights, weights, ages, number of mother’s previous pregnancies, and gestation length. Race and social factors included variables for parents’ ethnicities, education, income, and marital status.

For smoking, as time and number had various subcategories of levels in the ‘smoke’ variable, a one-way ANOVA was first performed for infant weight against ‘smoke’ alone. A significant difference in weight was seen between categories (df = 3, p <<< 0.001), but Tukey’s Honest Significant Difference (HSD) test revealed that differences were only between mothers that smoked during the current pregnancy and those that did not.

Filtering the dataset to look only at mothers that smoked during the pregnancy, an ANOVA revealed no significant differences in infant weight based on the number of cigarettes smoked (df = 7, p = 0.266).

|  |
| --- |
| https://lh4.googleusercontent.com/j0dxZYUDWOvIrFBL8e1-7YBgL0VFJKUbT7ii-QyTdtVFch72aOuXuzuze4FEY5xwPNHRkmX-KpQ9EVw4SAToDlceW3BHQMIlp5ZOvO1L7rF-DCbxCMgOHvzRZDLWYNBGOnJup0zR |
| Figure 8: Boxplots of infants weights grouped by smoking status of mothers |

Based on this analysis, the only information of interest from the variables ‘smoke,’ ‘number,’ and ‘time’ was whether or not the mother smoked during the pregnancy. Because of this, a new factor variable ‘smoker’ was added to the dataset to represent this, with a boolean value of ‘true’ or ‘false.’

All health variables were modelled using linear regression, and variables were added or removed until the model with the best Akaike information criterion (AIC) was found. AIC maximises the likelihood of a model while also reducing the number of parameters. This eliminates collinear or insignificant parameters. Also considered were possible interactions between each parents’ height and their weight. This was because a larger weight would be considered more indicative of possible obesity for shorter individuals. These interactions were removed by the stepwise optimization of AIC however. The final linear regression model included relationships between infant weight and gestation length, mother and father’s weights, and mother’s height.

|  |
| --- |
| https://lh5.googleusercontent.com/ltZWMTqIuSA_ilxHutFu-YGEiECfXdW1rdx8bd3D_cLUH58mzAL1pGH6cE_ks5moU30vygtzOX6UlgEF8pM5sZT2DfafDKIsOy9vwAS2O5TtMA1ijU1LYlNjzgVy15ptRccPxT90 |
| Figure 9: Scatter plots of Infant weight against several health variables. |

All race and social variables were modelled using linear regression and, again, parameters added or removed to achieve the best AIC. Only the father’s ethnicity remained in the final model. Mother’s race was likely removed due to its very high collinearity with father’s race, shown in table **X**. A one-way ANOVA with Tukey’s HSD was used to examine the difference in infant weights between ethnicities, with results shown in table 3.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Father’s Race** | | | | |
|  |  | Asian | Black | Mexican | Mixed | White |
| https://docs.google.com/drawings/d/sjc2vP8_Tupr1Vju0hOPuLg/image?w=26&h=106&rev=27&ac=1&parent=1ltWb6UROtAZ8m3OPb7_toTnE175teBF-HZU02JnCp30 | Asian | 17 | 0 | 0 | 0 | 2 |
| Black | 0 | 97 | 0 | 3 | 0 |
| Mexican | 0 | 0 | 14 | 0 | 2 |
| Mixed | 0 | 1 | 1 | 5 | 3 |
| White | 3 | 2 | 1 | 3 | 333 |

Table 2: Table of counts of combinations of parents’ ethnicities..

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Father’s Race** | **Difference** | **Lower 95% CI** | **Upper 95% CI** | **Adjusted p-value** |
| Black-Asian | 5.520000 | -6.8217551 | 17.861755 | 0.7369989 |
| Mexican-Asian | 15.162500 | -1.7371442 | 32.062144 | 0.1024440 |
| Mixed-Asian | 17.759091 | -1.1543613 | 36.672543 | 0.0774109 |
| White-Asian | 12.005882 | 0.4128221 | 23.598943 | 0.0381755\* |
| Mexican-Black | 9.642500 | -3.9240776 | 23.209078 | 0.2944270 |
| Mixed-Black | 12.239091 | -3.7663059 | 28.244488 | 0.2243716 |
| White-Black | 6.485882 | 0.7541199 | 12.217645 | 0.0175025\*\* |
| Mixed-Mexican | 2.596591 | -17.1379418 | 22.331124 | 0.9963933 |
| White-Mexican | -3.156618 | -16.0458440 | 9.732609 | 0.9626353 |
| White-Mixed | -5.753209 | -21.1886501 | 9.682233 | 0.8458558 |

Table 3: One-way Anova with Tukey’s HSD

The parameters included in these three models were then included in a final model, which predicted infant weight based on mother and father’s weight at pregnancy, the mother’s height, the duration of the gestation, the father’s race, and whether the mother smoked or not during the pregnancy ( F(9, 477) = 22.77, p << 0.001), with an R2 of 0.3005. A table of coefficients is given in table 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Change in infant weight in ounces per unit of variable** | **Standard Error** | **p-value** |
| Baseline (Non-smoking mother, Asian father) | -99.1111392 | 22.83716 | <<0.001 |
| Mother’s weight in pounds | +0.0796445 | 0.03987 | 0.04633 |
| Father’s weight in pounds | +0.0668948 | 0.03235 | 0.03919 |
| Mother’s height in inches | +0.8451491 | 0.31560 | 0.00766 |
| Gestation length in days | +0.4997548 | 0.04912 | <<0.001 |
| Black father | +2.3228534 | 4.02772 | 0.56440 |
| Mexican father | +13.1834253 | 5.34283 | 0.01396 |
| Mixed-race father | +9.6437409 | 6.11209 | 0.11527 |
| White father | +7.2041918 | 3.78359 | 0.05750 |
| Mother smokes during pregnancy | -7.9174635 | 1.52880 | <<0.001 |

Table 4: Table of coefficients for the final model given by AIC on the test data

**3.1.3 Adjusted R-squared**

Using the Adjusted R squared and backward selection to determine a statistically significant model for factors that determine baby weights.

We start with a full model, including each independent variable in the model. We then run a type II ANOVA which determines the effect of each independent variable as a whole to the dependent variable. Taking the first update to the model as an example.  We removed Mother’s age which has the highest p-value (0.9979), which indicates that it has no statistical significance in determining the weight of her baby. We also see a significantly higher adjusted R-squared ( from 0.2876 to 0.2892) indicating that 28.92% of the change in weight can now be explained by our model.

The R squared measures the percentage of change in the weight that can be explained by the independent variables. This is done by taking the standard error (distance of data points from the regressed line) and dividing it by the total variance within the weights(distance of weights from the mean weight). Our full model has a higher R square than the updated model. This does not necessarily mean that the independent variables are all significant determinants of babies weight. As you add more and more variables to a model the R squared will increase as each variable  helps explain away the squared errors. Thus we use the adjusted R squared which has a component that helps determine if the variable we added or subtracted was statistically significant in determining the babies weight. If we remove significant determinants the percentage of the change in weights will go down and vice versa. Thus, we can conclude that mother’s age did not significantly impact the babies weight according to the observations given to us.

Continuing this process we use p-values to determine which variable to remove until we reach the point where removing or adding variables lowers our adjusted R-squared.

**3.1.4 Mean Squared Error**

Finally, to select a model from the different models developed we fitted the regression line we modelled from the validation data set to the testing data set and calculated the mean squared error. In addition to the final AIC model described, an extra model including any interactions between included parameters was chosen based on the same stepwise AIC optimization. Out of this interaction-including model, the AIC based model, and the R2 based model, we picked the model with the least mean squared value (206.52), which was the AIC based model without interactions.

**4. Results**

Our final model to predict male infant weights generated on the basis of mother’s weight, father’s weight, gestation, mother’s height, father’s race and whether or not the mother was a smoker.

Once this final set of parameters was decided on, a linear regression model was developed using the full dataset of 1236 observations. 697 values remained after removing any containing NAs for relevant parameters. The values resulting from this are shown in table 5, results from a Tukey’s HSD test on the ‘drace’ parameter of an Anova model on the same parameters are shown in table 6. The corrected p-value for differences in infant weight between father’s ethnicities was 0.000180 (F = 5.64) Table 7 shows empirical confidence intervals for the same parameters from bootstrapping with 1000 resamples.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Change in infant weight in ounces per unit of variable** | **Standard Error** | **p-value** |
| Baseline (Non-smoking mother, Asian father) | -82.75938 | 18.90686 | <<0.001 |
| Mother’s weight in pounds | +0.05834 | 0.03302 | 0.077774 |
| Father’s weight in pounds | +0.07462 | 0.02785 | 0.007565 |
| Mother’s height in inches | +1.04909 | 0.27040 | <0.001 |
| Gestation length in days | +0.40530 | 0.03749 | <<0.001 |
| Black father | +0.08382 | 3.39030 | 0.980282 |
| Mexican father | +10.93542 | 4.58882 | 0.017440 |
| Mixed-race father | +6.82061 | 4.79404 | 0.155269 |
| White father | +6.16732 | 3.21096 | 0.055182 |
| Mother smokes during pregnancy | -8.66788 | 1.24540 | <<0.001 |

Table 5. Table of coefficients for the final model from the full data-set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Father’s Race** | **Difference** | **Lower 95% CI** | **Upper 95% CI** | **Adjusted p-value** |
| Black-Asian | -1.281648 | -10.161096 | 7.597801 | 0.9948685 |
| Mexican-Asian | 11.381312 | -1.024913 | 23.787537 | 0.0897613 |
| Mixed-Asian | 5.385063 | -7.383170 | 18.153296 | 0.7777534 |
| White-Asian | 4.117073 | -4.320314 | 12.554460 | 0.6696698 |
| Mexican-Black | 12.662959 | 2.759227 | 22.566692 | 0.0045459 |
| Mixed-Black | 6.666710 | -3.686901 | 17.020322 | 0.3973481 |
| White-Black | 5.398721 | 1.479666 | 9.317775 | 0.0016729 |
| Mixed-Mexican | -5.996249 | -19.496882 | 7.504384 | 0.7428224 |
| White-Mexican | -7.264239 | -16.773644 | 2.245167 | 0.2258204 |
| White-Mixed | -1.267990 | -11.245071 | 8.709092 | 0.9968650 |

Table 6. Pairwise differences in infant weight between father’s ethnicities using the final model.

|  |  |  |
| --- | --- | --- |
| **Parameter Coefficient** | **95% Confidence Intervals** | |
| **Lower** | **Upper** |
| Baseline (Non-smoking mother, Asian father) | -119.31138 | -44.51521 |
|
| Mother’s weight in pounds | -0.008432798 | 0.120769295 |
| Father’s weight in pounds | 0.02085098 | 0.12956448 |
| Mother’s height in inches | 0.4813196 | 1.5765884 |
| Gestation length in days | 0.3299913 | 0.4757639 |
| Black father | -6.103952 | 6.478472 |
| Mexican father | 2.451231 | 19.778310 |
| Mixed-race father | -2.399897 | 16.039511 |
| White father | 0.1593435 | 12.4788042 |
| Mother smokes during pregnancy | -10.894073 | -6.249776 |

Table 7. Confidence intervals for parameters from the final model, obtained through bootstrapping with 1000 resamples.

**4.1 Assumptions**

**4.1.1 Linearity Assumption**

Linearity assumption is tested visually looking at the scatter plot of the dependent variable against the independent variable. From figure 5, the linear trend is not substantially strong with large deviates from the relationship line. The visual inspection however suggests at least a rough linear relationship and this assumption will be reevaluated as the linear model is fitted, with R² as a measure of fit.

**4.1.2 Normality Assumption**

In a linear regression the residuals or errors in observed data is assumed to be normally distributed. A Shapiro-Wilk normality test suggests non-normal residuals ( W = 0.99501, p = 0.02291), however this is common with large samples, and a normal QQ plot shows very normal looking residuals (see Figure 10). Plotting residuals against the fitted values shows normal errors, but with a very obvious downward curve for extremely high or low baby weights. Looking at the Pearson residuals for each parameter shows that this is because of the relationship with gestation duration (Figure 11), which is to be expected. The model fails to apply to extremely high or low gestation durations because they suggest a further complication which is likely to correlate with low birth weight.

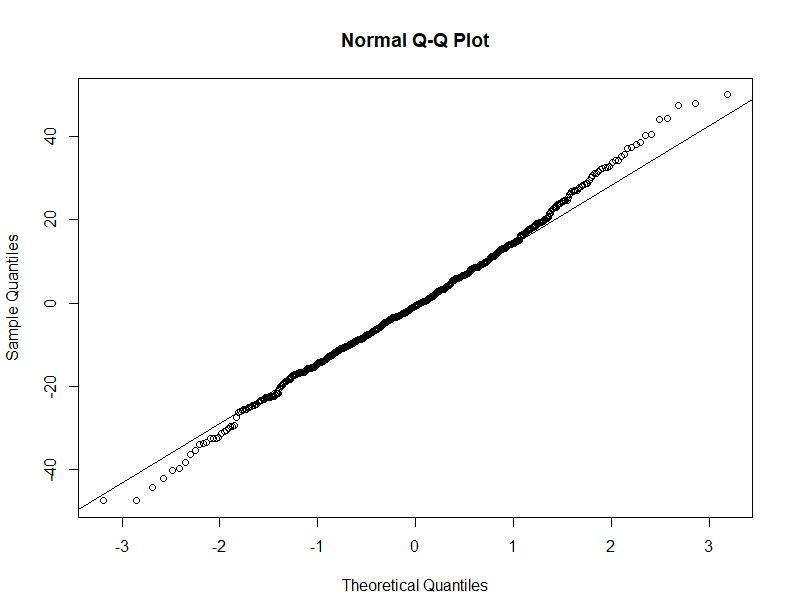


Figure 10. QQ Plot for residuals from the final regression model on the full data-set.

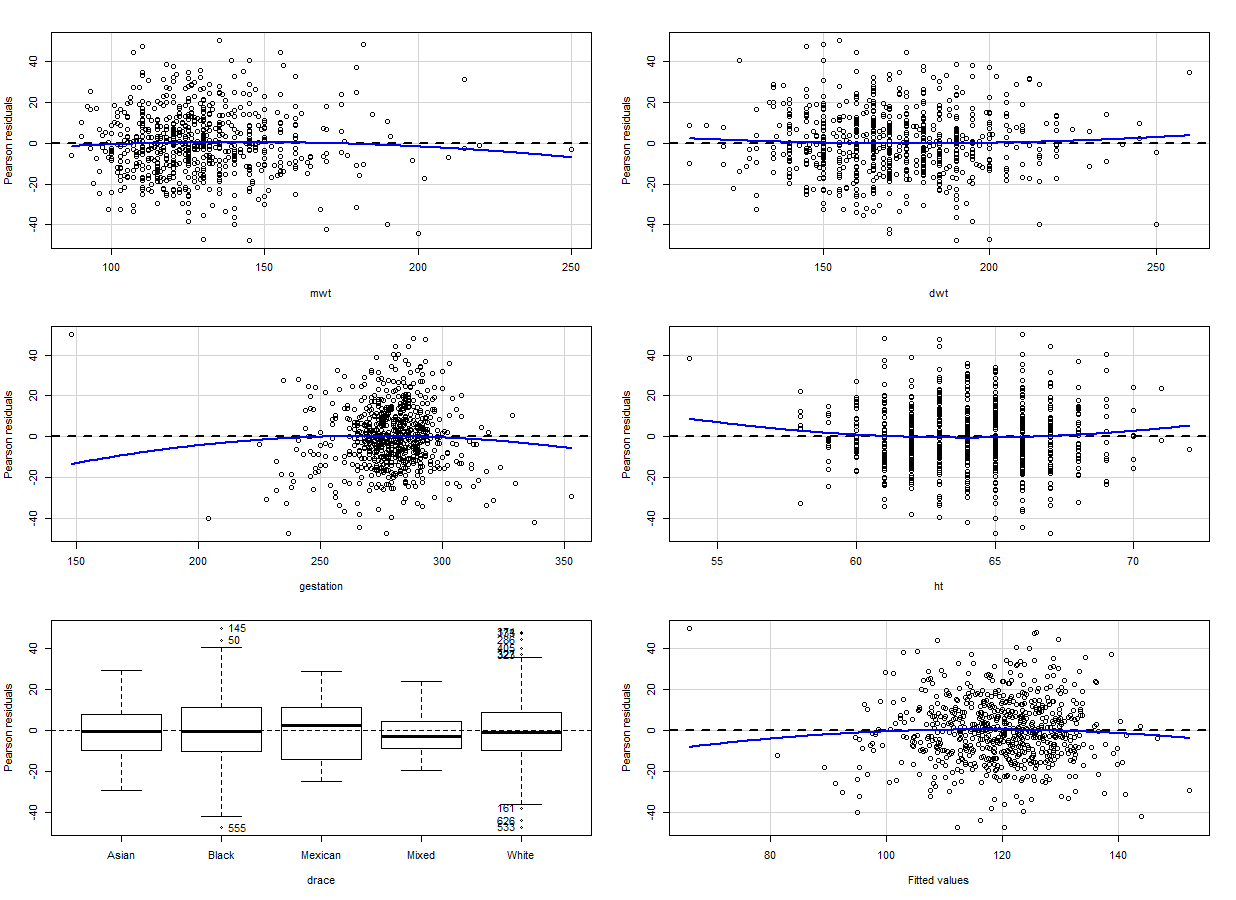


Figure 11. Individual graphs of residuals vs fitted values for the different parameters, and full model, for the final regression model on the full data-set.

**4.1.3 Other Assumption tests**

A studentized Breusch-Pagan test for heteroskedasticity (non-constant variance) weakly suggests that the variance is not constant across residuals (BP = 17.185, p = 0.04589). The variances of infant weight between father’s ethnicities were relatively similar (Lowest:  232, Highest: 363). A Durbin-Watson test showed that errors were uncorrelated and independent (D-W = 1.966058, p = 0.652)

**Discussion**

The final model was made by categorizing the available data according to the characteristics such as social, health and mother’s smoking status. The best predictors from each categories were integrated into a model. The model used the following variables  mother’s weight, father’s weight, gestation, mother’s height, father’s race and whether or not the mother was a smoker. The model passed most assumptions supporting its validity, however the residuals showed some weak evidence of heteroskedasticity, which should be further examined. It is also clear that the model does not apply accurately at extremely high or low gestation durations.

However, a larger problem is the limitations of the data-set as a whole. The data-set is taken from US hospitals in the 1960s, which gives a very specific cultural snapshot with which to examine the social variables. In addition, only male, single, live births were considered. These findings may not extend to female or multiple births, or provide support for predicting whether a birth will survive longer than 28 days.

Smoking mother’s had babies that weighed lighter than mother’s who did not smoke. Practically understandable as mother’s health is a key determinant of babies growth. Babies also, unsurprisingly, weigh more if their parents individually weigh more or are taller. A longer pregnancy also correlated with infants who weighed more, likely due to further development and growth by the infant before birth. The father’s race was seen to be a determinant of baby weights with babies born to Mexican father’s having the higher weight and babies born to black fathers having lowest birth weight. However, It is important to note that the mother and father’s race were extremely correlated, and so it may be more accurate to say there was a relationship with the parents’ race.

There was extremely strong support for four of the relationships found with male infant weight: whether the mother smokes during the pregnancy, the duration of the pregnancy, the mother’s height, and the father’s race.

**Conclusion**

We started our modelling taking different approaches, using practical and statistical methods to predict male baby weights according to the given variables.The final model was developed keeping in mind statistical and practical significance. The covariates to be included were selected by dividing them into categories on the basis of their characteristics (social, smoker, etc) and statistically tested. While at generating level many models seemingly showed higher statistical significance (Higher R squared), when fitted  into tests sets, it was clear that the practically selected model generated better results i.e gave a lower mean squared error. Relying purely on test statistics to determine models gave more data specific results than generalisable ones. Despite problems with the data set, we have provided strong evidence for relationships, for male babies who survive for at least 28 days after birth, between infant weight and mother’s height, parents’ race, pregnancy length, and whether or not the mother smoked during the pregnancy.

**References**

Neggers, Y. & Crowe, K. (2013) Low Birth Weight Outcomes: Why Better in Cuba Than Alabama? *J. Am. Board. Fam. Med.* 26 (2) 187-195 doi: 10.3122/jabfm.2013.02.120227

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/).

**Appendices**

**Appendix 1 - Data cleaning**

library(tidyverse)  
babiesRaw <- read.csv("babies23.data", sep="")  
FixNAsBabies <- function(data) {  
  # Remove one level factors  
  data <- data %>% select(-pluralty, -outcome, -sex, -id)  
  # Only male single births  
    
  # Reformat date  
  data$date <- as.Date(data$date, origin = '1961-01-01')  
  
  # Eliminate NAs from numeric values  
  data$wt[data$wt == 999] <- NA  
  data$gestation[data$gestation == 999] <- NA  
  data$parity[data$parity == 99] <- NA  
  data$age[data$age == 99] <- NA  
  data$ht[data$ht == 99] <- NA  
  data$wt.1[data$wt.1 == 999] <- NA  
  data$dage[data$dage == 99] <- NA  
  data$dht[data$dht == 99] <- NA  
  data$dwt[data$dwt == 999] <- NA  
  data$inc <- as.numeric(data$inc)  
  data$inc[data$inc == 98 | data$inc == 99] <- NA  
  data$inc <- as.numeric(data$inc)  
  
    
    data$race <- as.factor(recode(data$race, '0' = 'White', '1' = 'White', '2' = 'White',   
                          '3' = 'White', '4' = 'White', '5' = 'White', '6' = 'Mexican',  
                          '7' = 'Black','8' = 'Asian', '9' = 'Mixed', .default = NA\_character\_))  
    data$drace <- as.factor(recode(data$drace, '0' = 'White', '1' = 'White', '2' = 'White',   
                          '3' = 'White', '4' = 'White', '5' = 'White', '6' = 'Mexican',  
                          '7' = 'Black','8' = 'Asian', '9' = 'Mixed', .default = NA\_character\_))  
    data$ed <- as.factor(recode(data$ed, '0' = 'less than 8th grade',   
                        '1' = '8th -12th grade - did not graduate',   
                        '2' = 'HS graduate--no other schooling', '3' = 'HS+trade',  
                        '4' = 'HS+some college', '5' = 'College graduate', '6' = 'Trade school HS unclear',  
                        '7' = 'Trade school HS unclear', .default = NA\_character\_))  
    data$ded <- as.factor(recode(data$ded, '0' = 'less than 8th grade',   
                        '1' = '8th -12th grade - did not graduate',   
                        '2' = 'HS graduate--no other schooling', '3' = 'HS+trade',  
                        '4' = 'HS+some college', '5' = 'College graduate', '6' = 'Trade school HS unclear',  
                        '7' = 'Trade school HS unclear', .default = NA\_character\_))  
    data$marital <- as.factor(recode(data$marital, '1' = 'Married', '2' = 'Legally Separated', '3' = 'Divorced',  
                             '4' = 'Widowed', '5' = 'Never Married', .default = NA\_character\_))  
    data$smoke <- as.factor(recode(data$smoke, '0' = 'Never', '1' = 'Smokes Now',   
                           '2' = 'Until Current Pregnancy', '3' = 'Once Did, not now', .default = NA\_character\_))  
data$time <- as.factor(recode(data$time, '0' = 'Never smoked', '1' = 'Still Smokes',  
                          '2' = 'During current preg', '3' = 'Within 1 yr', '4' = '1 to 2 years ago',  
                          '5' = '2 to 3 yr ago', '6' = '3 to 4 yrs ago', '7'= '5 to 9yrs ago',   
                          '8' = '10+yrs ago', '9' = 'quit and dont know', .default = NA\_character\_))  
data$number <- as.factor(recode(data$number, '0' = 'Never', '1' = '1-4',  
                            '2' = '5-9', '3' = '10-14', '4' = '15-19',  
                            '5' = '20-29', '6' = '30-39', '7'= '40-60',   
                            '8' = '60 +', '9'= 'smoke but dont know', .default = NA\_character\_))  
  
  # Rename mother's weight  
  data <- rename(data, mwt = wt.1)  
  return(data)  
    
}  
babiesNAFixed <- FixNAsBabies(babiesRaw)  
babiesNoNA <- na.omit(babiesNAFixed)  
str(babiesNoNA)  
  
# Separating training data and validation dataset  
set.seed(140011358)  
length <- nrow(babiesNoNA)  
validation.vector <- sample(length, length/5)  
babiesNoNAAll <- babiesNoNA  
validation.data.set <- babiesNoNAAll[validation.vector,]  
babiesNoNA <- babiesNoNA[-validation.vector,]  
  
  
attach(babiesNoNA)

**Appendix 2 - AIC model generation**

smoking <- aov(wt ~ smoke, babiesNoNA)

summary(smoking)

TukeyHSD(smoking)

babiesSmokers <- filter(babiesNoNA, smoke == 'Smokes Now')

smokingNumber <- aov(wt ~ number, babiesSmokers)

summary(smokingNumber)

TukeyHSD(smokingNumber)

babiesNoNA$smoker <- ifelse(babiesNoNA$smoke == 'Smokes Now', TRUE, FALSE)

validation.data.set$smoker <- ifelse(validation.data.set$smoke == 'Smokes Now', TRUE, FALSE)

detach(babiesNoNA)

attach(babiesNoNA)

health <- lm(wt ~ mwt + dwt + ht + dht + gestation + parity + age + dage + smoker, babiesNoNA)

summary(health)

health.stepped <- step(health, direction = 'both')

summary(health.stepped)

health.int <- lm(wt~mwt+dwt+ht+dht+gestation+parity+age+dage+smoker+mwt\*ht+dwt\*dht)

summary(health.int)

health.int.stepped <- step(health.int, direciton = 'both')

summary(health.int.stepped)

social <- lm(wt ~ race + ded + drace + marital + inc + ed, babiesNoNA)

summary(social)

social.stepped <- step(social, direction = 'both')

summary(social.stepped)

car::vif(social)

social.aov <- aov(wt~drace+ed+ded+marital)

car::Anova(social.aov, type = c("III"))

summary <- lm( wt ~ mwt + dwt + ht + gestation + smoker + drace

              , babiesNoNA)

summary(summary)

summary.stepped <- step(summary, direction = "both") # This shouldn't do anything more

summary(summary.stepped)

summary.aov <- aov(summary.stepped)

TukeyHSD(summary.aov,which = 'drace')

summary.int <- lm( wt ~ (mwt + dwt + ht + gestation + smoker +

   drace)^2 babiesNoNA)

summary(summary.int)

summary.int.stepped <- step(summary.int, direction = "both")

summary(summary.int.stepped)

TukeyHSD(social.aov)

**Appendix 3 - R2 model generation**

#Backward Selection

#FullLinearModel

FullModel <- lm(wt ~ ., data = babiesNoNA)

summary(FullModel)

Anova(FullModel)

#Removing Variable with highest p value

Model<- update(FullModel, .~ . - age)

summary(Model)

Anova(Model)

#adjusted R squared increases

model1<- update(Model, .~. -race)

summary(model1)

Anova(model1)

#adjusted R squared increases

model2<- update(model1,.~. -dage)

summary(model2)

Anova(model2)

#adjusted R squared increases

model3 <- update(model2,.~. -marital)

summary(model3)

Anova(model3)

#adjusted R squared increases

model4 <- update(model3,.~. -ded)

summary(model4)

Anova(model4)

#adjusted R squared increases

model5<- update(model4,.~. -date)

summary(model5)

Anova(model5)

#adjusted R squared increases

model6<-update(model5,.~. -ed)

summary(model6)

Anova(model6)

#adjusted R squared increases

model7<- update(model6,.~. -smoke)

summary(model7)

Anova(model7)

model8 <- update(model7, .~. -inc)

summary(model8)

Anova(model8)

model9 <- update(model8,.~. -dht)

summary(model9)

Anova(model9)

##Removing any other variables reduces the R-squared

finalmodel <- model9

**Appendix 4 - Mean squared error function for Model Selection**

GetMeanSquaredError <- function(model, validation.data) {

 predicted <- predict(model,validation.data)

 ind.var <- colnames(model$model[1])

 actual <- validation.data[ind.var]

 residuals <- actual - predicted

 squared.resid <- residuals\*\*2

 return(mean((residuals[,ind.var])\*\*2))

}

**Appendix 5 - Model assessment (Results)**

BootstrapCoefNoCorr <- function(data, model, b = 999, alpha = 0.05) {  
  n <- nrow(data)  
  vars <- colnames(model$model)  
  coefs <- names(coef(model))  
  coefs[1] <- vars[1]  
  num.vars <- length(vars)  
  num.coefs <- length(coefs)  
  matr <- matrix(nrow = b+1, ncol = num.coefs)  
  matr[1,] <- coef(model)  
  colnames(matr) <- names(coef(model))  
  dependent.var <- vars[1]  
  independent.vars <- vars[2:num.vars]  
    
  pred <- predict(model)  
  resid <- resid(model)  
    
  for (i in 2:(b+1)) {  
    resample.index <- sample(1:n, n, replace = TRUE)  
    resample.dep.var <- pred + resid[resample.index]  
    dataframe <- data[,independent.vars]  
    dataframe$dep <- resample.dep.var  
    matr[i,] <- coef(lm(dep ~ mwt + dwt + gestation + ht + drace + smoker, dataframe))  
  }  
    
  coefs <- ncol(matr)  
  quantiles <- matrix(nrow = 2, ncol = coefs)  
  colnames(quantiles) <- colnames(matr)  
  rownames(quantiles) <- c('lower.ci', 'upper.ci')  
  for (i in 1:coefs) {  
    quantiles[,i] <- quantile(matr[,i], probs = c(alpha/2,(1-alpha/2)))  
  }  
  return(quantiles)  
}  
  
BootstrapCoefNoCorr(babiesNoNA, summary.stepped)

shapiro.test(resid(summary.stepped)

qqnorm(resid(summary.stepped))

qqline(resid(summary.stepped)

car::residualPlots(summary.stepped)

car::bptest

lmtest::bptest(summary.stepped)

babiesNAFixed %>% group\_by(drace) %>% summarise(var(wt))

car::durbinWatsonTest(summary.stepped)