BABIES' BIRTH WEIGHT PREDICTION

2. ABSTRACT

Babies' birth weight is not only the concern of an individual but also a prime concern of countries. The weight of the newborn babies can help specialists give professional suggestions to parents. Thus, the study focuses on predicting babies' birth weight. The objective is to investigate the factors which affect the babies' birth weight. The research also evaluates how smoking affects birth weight of babies and how to predict the birth weight for medical advice.

The study was conducted using multiple linear regression model and R language. The result suggests that gestation, parity, height, weight, and smoke influence babies' birth weight except for the mother's age, based on the data. Specifically, smoking mothers deliver smaller babies. On this basis, the prediction model for babies' birth weight can be derived. .

3. PROBLEM AND MOTIVATION

Baby's birth weight is an important factor to determine the delivery method for moms when the baby is matured. In addition, the association between the birth weight and later-life conditions (for instance, diabetes, obesity, tobacco smoking and intelligence) is proved by various studies. Low birth weight also links with the neonatal infection and infant mortality. The babies' birth weight becomes the concern of families individually and countries generally.

The data was collected in 1960 – 1967, the period of post-World War Two. The shortage of food and the diets might influence the decrease of weight during pregnancy of these mothers. However, the important concepts (for example, genetic factors) and medical technologies (such as pre-existing disease and gestational diabetes diagnostic) that produce significant effects on birth weight, had not been developed during the 1960s. Due to the shortages, the data collected lack those key factors, which can be used to provide a better explanation on birth weight in modern decades.

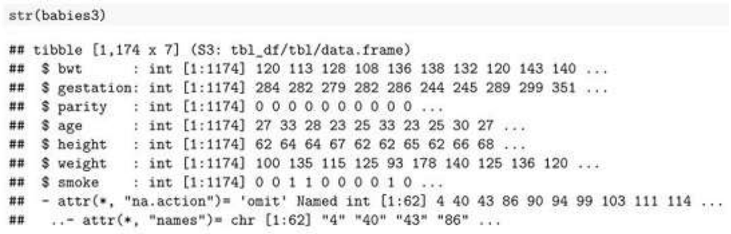
Hence, this research does not deeply provide medical advice that works for all cases. In fact, it demonstrates a prediction for reference which can be used by medical professionals in the healthcare industry. It helps mothers be well-prepared and protect themselves and their babies.

4. DATA DESCRIPTION

The data set was collected from all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area by The Child Health and Development Studies.The original data set includes null value which was eliminated. The data frame contains 1236 rows and 8 variables. However, we removed the "case" variable and were left with 7 variables as shown below (Figure 4a).

- weight of the infants in ounces (bwt)
- length of pregnancy in days (gestation)
- mother's age in years (age)
- mother's height in inches (height)
- whether the child was the first born (parity)
- mother's pregnancy weight in pounds (weight)
- whether the mother was a smoker (smoke).

Among these variables, Weight of the infant (btw) is the response, and the rest are predictors. After cleaning the null value, the set with 1179 cases was being used.



```
str(babies3)

## tibble [1,174 x 7] (S3: tbl_df/tbl/data.frame)
##  $ bwt      : int [1:1174] 120 113 128 108 136 138 132 120 143 140 ...
##  $ gestation: int [1:1174] 284 282 279 282 286 244 245 289 299 351 ...
##  $ parity   : int [1:1174] 0 0 0 0 0 0 0 0 0 0 ...
##  $ age      : int [1:1174] 27 33 28 23 25 33 23 25 30 27 ...
##  $ height   : int [1:1174] 62 64 64 67 62 62 65 62 66 68 ...
##  $ weight   : int [1:1174] 100 135 115 125 93 178 140 125 136 120 ...
##  $ smoke    : int [1:1174] 0 0 1 1 0 0 0 0 1 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:62] 4 40 43 86 90 94 99 103 111 114 ...
##   ..- attr(*, "names")= chr [1:62] "4" "40" "43" "86" ...
```

Figure 4a: The structure of the data set.

We have five (5) numerical variables which are shown as the following:

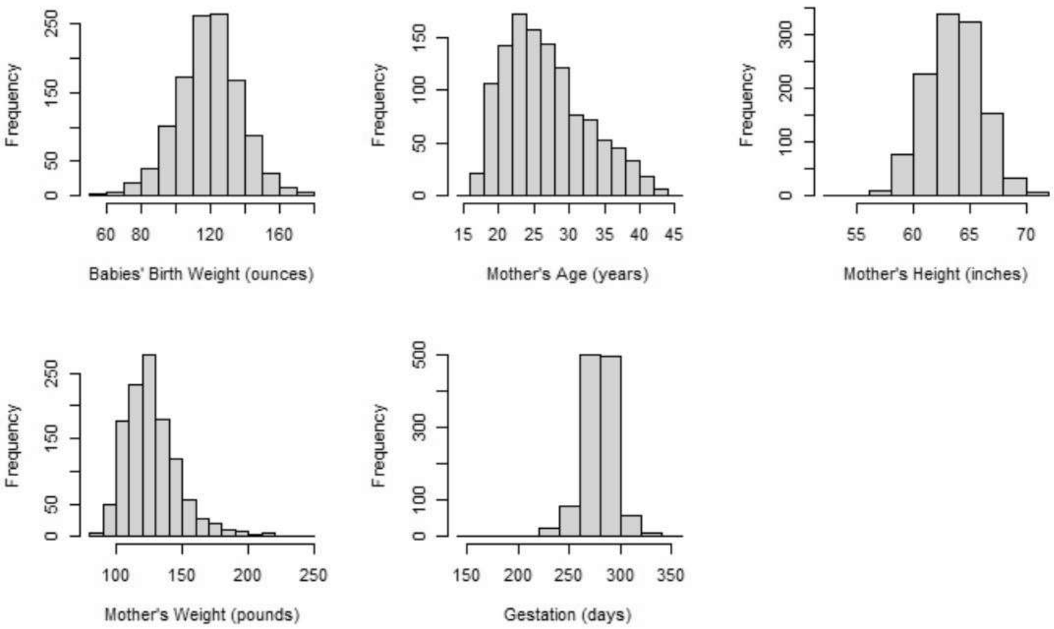


Figure 4b: Histogram of Numerical Variables

5. QUESTIONS OF INTEREST

The goal of this study is to model the weight of babies (in ounces) based on the predictor variables: gestation, age, height, parity, weight, and smoke. Thus, the question of interest concentrates on which of these factors have an effect on the baby's birth weight and whether the effect of one variable depends on the value of another variable. Hence, the focus is to find if there is any association between variables. The following factors will be analyzed: how smoking affects babies' birth weight, whether a prediction model can be used to predict a baby's birth weight, and how confident the model is at predicting birth weight within a certain range.

6. STATISTICAL ANALYSIS AND MAIN OUTCOME

Multiple Linear Regression models were used to estimate baby's birth weight from predictors listed in Section 4 above.

6.1 Exploratory Analysis and Variable Selection

The added variable plots (Figure 6.1a) are used for initial screening whether there is linear relationship between any predictors and response.

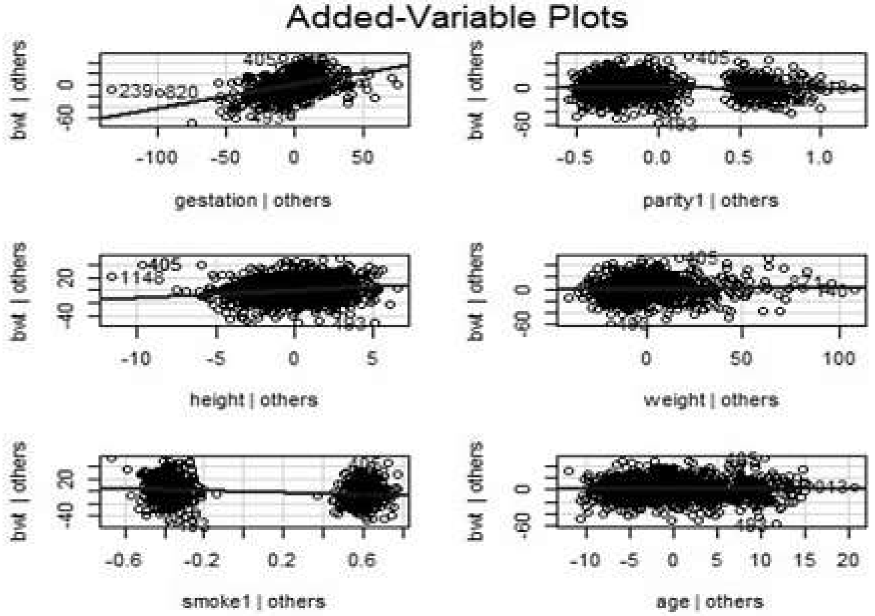
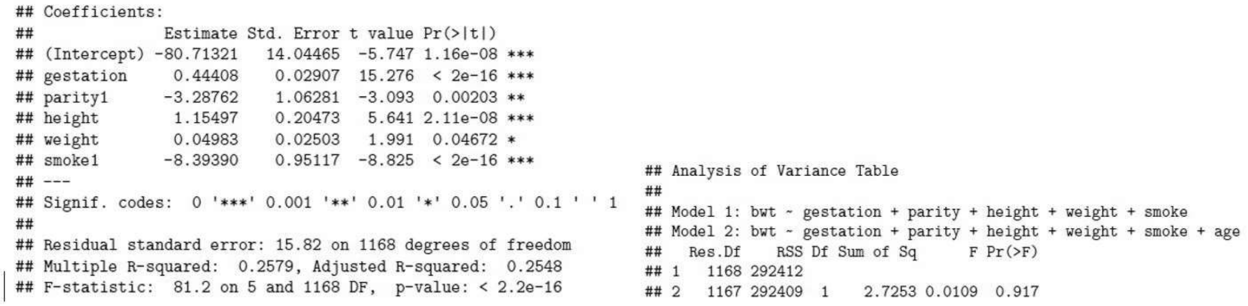


Figure 6.1a: Added-Variable Plots

Then backward selection with AIC criteria, forward selection with AIC criteria, and hybrid selection with AIC criteria ended with same five (5) variables selected for prediction of birth weight—gestation, parity, height, weight and smoke. We will adopt these five variables to initialize modeling.

At this point we shall assume all the necessary assumptions hold. In later Section 6.3, we will check these assumptions for the final model established. Individual t-test is used to test the existence of a linear association between response and five (5) selected individual predictors. As showed in Figure 6.1b, all the five predictors are statistically significant and useful to predict response.



```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80.71321   14.04465  -5.747 1.16e-08 ***
## gestation     0.44408    0.02907  15.276  < 2e-16 ***
## parity1      -3.28762    1.06281  -3.093  0.00203 **
## height        1.15497    0.20473   5.641 2.11e-08 ***
## weight        0.04983    0.02503   1.991  0.04672 *
## smoke1       -8.39390    0.95117  -8.825  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.82 on 1168 degrees of freedom
## Multiple R-squared:  0.2579, Adjusted R-squared:  0.2548
## F-statistic:  81.2 on 5 and 1168 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: bwt ~ gestation + parity + height + weight + smoke
## Model 2: bwt ~ gestation + parity + height + weight + smoke + age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1168 292412
## 2   1167 292409  1    2.7253 0.0109  0.917
```

Figure 6.1b: R Summary of Initial Model

Since variable age is removed by the stepwise selection, we are still interested to investigate. The partial F-test is used to compare model with age and without age (Figure 6.1b).

$H_0 : \beta_{age} = 0$, which is reduced model without age.

$H_1 : \beta_{age} \neq 0$, which is the full model with age.

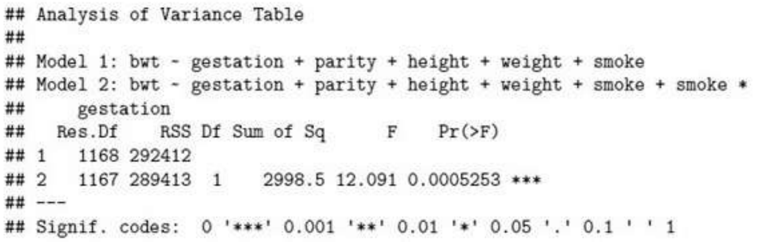
The F- statistic is 0.0109 and p-value is large as 0.917, which failed to reject the null hypothesis with reduced model. Thus, there is no evidence to support the alternative hypothesis. This means we adopt the reduce model without age.

Multicollinearity has been checked by correlation matrix and variance inflation factors (VIF) values (Appendix 1). None of the correlation between two predictors is greater than 0.5 and all the VIFs are below 5, which confirms no multicollinearity issues in the variables selected.

6.2 Interaction Term

Next, we considered whether the effect on the response variable of one predictor depends on the value of the other. Interactions between weight and height, smoke and other four (4) predictors have been checked on statistical significance.

Model comparison by ANOVA table shows the interaction between smoke and gestation is statistically significant (Figure 6.2).



```
## Analysis of Variance Table
##
## Model 1: bwt ~ gestation + parity + height + weight + smoke
## Model 2: bwt ~ gestation + parity + height + weight + smoke + smoke *
##     gestation
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1168 292412
## 2   1167 289413  1    2998.5 12.091 0.0005253 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6.2. ANOVA Table for Model Comparison—Interaction Term

$H_0 : \beta_{13} = 0$, which is the reduced model without interaction term.

$H_1 : \beta_{13} \neq 0$, which is the full model with interaction term.

The F-statistic value is 3.477 and p-value is much smaller than 0.001, which rejected the null hypothesis with reduced model. Thus, the result suggests the full model with interaction term between smoke and gestation. After adding interaction term into the model, the adjusted R has been increased from 0.2548 to 0.2618, which indicates more variability can be explained by the model with interaction term.

It also indicates the interaction between smoke and gestation will affect the baby's birth weight, since smoking will lead baby premature issue that is related to birth weight.

6.3 Diagnostic Checks

We next looked at regression diagnostics to check the validity of all aspects of our multiple linear regression model.

The diagnostic begins by examining the Residuals vs. Fitted Plot and Q-Q Plot (Figure 6.3). The points in the residual versus fitted plot are randomly scattered around zero with no discernible pattern, which indicates constant variance assumption is met. The normal Q-Q plot shows that the standardized residuals follow an approximate normal distribution and confirms linearity.

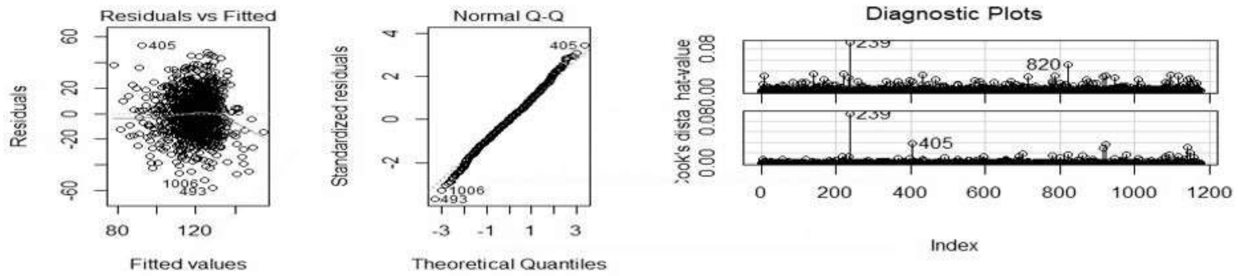


Figure 6.3: Residual vs Fitted Plot and Q-Q Plot

Finally, the Cook's distance (Figure 6.3) also confirms there is no influential point that must be removed, since the obvious Number 239, which is far away from other points, is still below the cutoff 0.5. Thus, no data point will be removed and transformation is not needed.

In addition to the diagnostics above, the transformation did not improve the model and it is much more complicated to interpret (Appendix 1), time consuming for prediction, the original model and data have been used to establish the final model.

6.4 Final Model Established and Summary

Final model established:

$$\widehat{bwt} = -59.07 + 0.37\text{gestation} - 3.30\text{parity} + 1.15\text{height} + 0.05\text{weight} - 66.83\text{smoke} + 0.21\text{gestation*smoke}$$

Based on the model shown above, variable gestation, parity, height, weight, and smoke including interaction between smoke and gestation, affect the birth weight. Smoke is a key factor to be considered, as shown in Figure 6.4. If the mother is a smoker with her first baby, the baby's birth weight is smaller than that of a baby whose mother is a non-smoker, when all the other predictors are held constant. Similarly, smoke affects the weight of non-first baby in the same way on mom who is smoking.

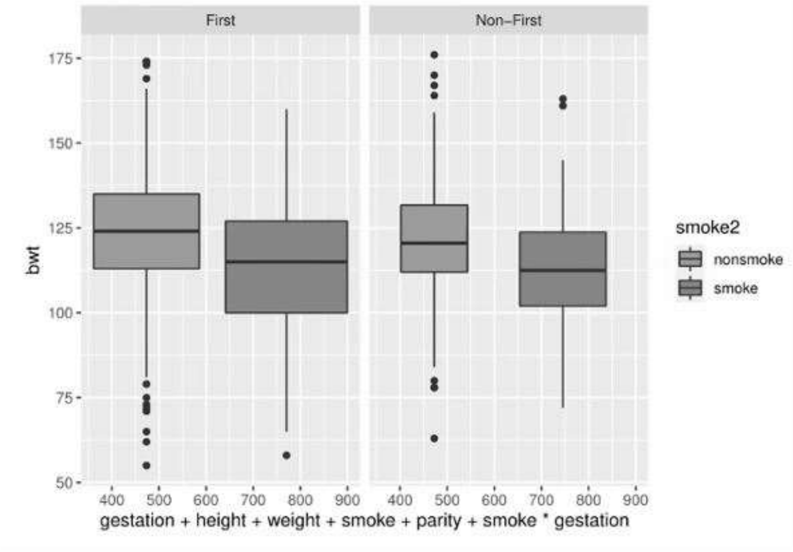


Figure 6.4: Effect of Smoking Boxplot

7. CONCLUSION

Based on the data analysis, gestation, parity, height, weight, and smoke were found to have an effect on the babies' birth weight, except for the Mother's age. The results show that there is no correlation between the predictor variables, which is beneficial for the final model because it indicates the model does not over-fit, and this can help improve predictions. Additionally, smoke and gestation have an interaction effect on the babies' birth weight since the interaction is statistically significant, which leads to including the interaction effect of smoke and gestation in the final model. It was also found that mothers who smoke deliver smaller babies than mothers who do not smoke, whether it is her first child or not.

Furthermore, the final model can be used to predict a baby's birth weight. It serves as a reference model to guide doctors and mothers in deciding the best method of delivery, whether she should have natural delivery or C-section, based on the baby's predicted birth weight. For instance, by using the final model to make a prediction, we are 95% confident that a baby's birth weight will be between 91 and 153 ounces, when the mother is 62 inches, 150 pounds, 280 days pregnant, is a nonsmoker, and the baby is her first child. However, only about 26% of variability in the baby's birth weight can be explained by the model that was established based on the limited data, which is missing factors such as diet, race, genetic information, underlining disease, and whether mom has gestational diabetes. Those key factors are medically proved to affect baby's birth weight so further data collection and study should be involved to explain more on the variability, and to improve the modeling.

8. APPENDICES

Appendix 1: R Code (20 pages)