**Maternal Determinants of Birth Weight in the US: An Application of Multiple Regression**

Sheikh Arafat

**Abstract**: Birth weight is not only one of the most critical determinants of perinatal mortality but also considered as an indicator of the health status of a society. Identifying the factors that affect birth weight is an important task as this will help us to organize the medical care and lifestyle of a mother. Consequently, in this project, we have examined the effects of several maternal factors on birth weight, including mother’s age, mother’s pre-pregnancy body mass index (BMI), mother’s pre-pregnancy weight, mother’s weight gain during pregnancy, and cigarette smoking before and during pregnancy. We have observed that along with other maternal-related factors mother’s smoking status plays a vital role in modeling birth weight. The mean weight of a smoker mother’s baby will be around 200 grams less than the non-smoker mother’s baby. As a consequence, maternal factors need to be strictly monitored besides the social, psychological, demographic, and environmental factors.

**Introduction:** Birth weight is a known risk factor for newborn mortality [1]. It is also known to cause chronic diseases later in life and increase the risk of type 2 diabetes, hypertension, and cardiovascular diseases [2]. It is also considered a measure of intrauterine malnutrition. Psychological stress and social disadvantage may play a role in the development of low birth weight [3], [4]. It is also linked to the consumption of alcohol and cigarette. However, there is general agreement about the importance of any of these factors. Studies have shown that the environment can affect the birth weight of children [1]. Nevertheless, the effects of maternal factors on birth weight are significant [5].

Obese and overweight women are at increased risk of having too heavy or too light a baby. Obesity and excessive weight gain on the contrary can lead to adverse maternal and fetal outcomes [6]. As a consequence, it is important to estimate the effects so that preventive actions such as diet charts and nutrition care can be provided to the mother. Other factors that could influence birth weight include the mother’s age and weight gain during pregnancy. Moreover, a variety of socioeconomic, medical, and psychosocial factors are known to be associated with a higher risk of low birth weight [7].

In this study, we will estimate the effects of these maternal factors on birth weight by fitting a multiple regression model. In the next section, we will provide a brief description of our data set and the summary statistics of our considered variables. The methodology section explains all the techniques that are applied for the analysis of the data. The analysis section accommodates the discussion of the results of the graphical and data analysis summary. Finally, the results section summarizes this project and provides recommendations.

**Data description:** We have used the data set named “2020 Natality” published by the Centers for Disease Control and Prevention’s National Center for Health Statistics (NCHS) [8]. This data set has 3,619,826 live births across the United States in 2020. Each row corresponds to a live birth. We have taken 7500 random samples from the data set. We are dealing with six variables where the birth weight is the response variable measured in grams. The average birth weight is 3255 grams. We have five response variables including one qualitative predictor variable with three levels which implies whether the mother doing smoking before and during pregnancy or not and the status is unknown. The percentage of the recorded responses are 94%, 5%, and 1% respectively. The remaining predictor variables are quantitative such as mother’s age (median age is 26 years), mother’s weight (mean weight is 175 pounds), mother’s weight gaining during pregnancy (mean weight gaining is 31 pounds), and mother’s pre-pregnancy body mass index (mean BMI is 29 pound/in²).

**Methodology:** For this project, our methodology is based on multiple linear regression with interaction terms. We will explore the relationship between predictor variables and the response variables. Various assumptions will be checked for a linear regression model with better performance. Assumptions include linearity assumption, independence assumption, normality assumption, and assumption of constant variance of error. Unusual observations within the dataset are evaluated and possible action (deleting from dataset or keeping in the dataset) will be considered.

Formal hypothesis tests will be applied to the dataset so we can get concrete statistics of tests to decide whether there is a linear relationship between the response variable (baby body weight) and terms (predictors and their interactions). Box-cox transformation method is applied to the dataset to solve the problem of non-normality within the dataset and we possibly can get a regression model which can better express the relationship between predictors and response variables.

In order to select a set of efficient predictors, we will apply model selection (forward and backward) for the regression models. For possible multicollinearity within predictor variables, VIF will be computed from data and the VIF statistics will be used to evaluate the severity of multicollinearity within the data for predictors. Partial F-test is applied for checking whether reduced model or full model performs better, and for deciding which model should be chosen as our final model.

In order to better understand the data and explain the results of the analysis, comprehensive data visualization will be performed on the dataset and the analysis result. Currently enhanced scatter plots will be the main part of our data visualization.

**Analysis:** We have sketched a pairwise plot for the births data set and tried to detect the possible association within the variables. We have observed that almost all the variables are showing a scatter pattern instead of any specific pattern except the Mother’s BMI and Mother’s Pre-pregnancy Weight. Figure 1 exhibits a positive association between these two predictors.

To check whether the data set has any outliers or not, we have created a box-plot of the response variable and observed that a lot of outliers exist in the data set (Ref. Figure 2). The values of birth weight which are below and above the 1.5IQR of the first and third quartile respectively are removed from the data set.

After preparing the data set, we fit the multiple regression model. We have a qualitative predictor variable with three categories and introduce two dummy or indicator variables in our model. We define the dummy or indicator variables as follows:

In this case, the smoking status No is the baseline level. Let, , , , , and, .

Now the model will look like

Our fitted model is following:

With this model, we actually have three different models. Appendix contains all the considered and fitted models.

The interpretations of these fitted parameters are

, is the true mean birth weight of a mother’s baby whose smoking status is no when all the other predictor variables equal zero (this is out of the scope of the model).

, is the change in the true mean birth weight as the Mother Age increases by 1 year, regardless of the mother’s smoking status.

, is the change in the true mean birth weight as the Mother BMI increases by 1 unit, regardless of the mother’s smoking status.

, is the change in the true mean birth weight as the Mother Pre-pregnancy Weight increases by 1 pound, regardless of the mother’s smoking status.

, is the change in the true mean birth weight as the Mother Weight Gain during pregnancy increases by 1 pound, regardless of the mother’s smoking status.

, is the change in the true mean birth weight of a mother’s baby whose smoking status is unknown relative to a mother’s baby whose smoking status is no when all the other predictor variables equal zero.

, is the change in the true mean birth weight of a mother’s baby whose smoking status is yes relative to a mother’s baby whose smoking status is no when all the other predictor variables equal zero.

, is the true mean birth weight of a mother’s baby whose smoking status is unknown when all the other predictor variables equal zero (this is out of the scope of the model).

, is the true mean birth weight of a mother’s baby whose smoking status is yes when all the other predictor variables equal zero (this is out of the scope of the model).

Now, to check the usefulness of the model, we have performed the F-test. The calculated value of the test statistic is 37.82 on 6 and 7231 degrees of freedom with a p-value of . As the p-value is less than any usual level of significance, we have rejected the null hypothesis which implies that the fitted model is useful to estimate the birth weight. After that, we have performed the individual t-test to check the significance of the individual predictors of the fitted model. Similarly, with the F-test, we have observed that all the p-values of the individual t-test provide smaller p-values than the usual level of significance. In addition to this, none of the 95% confidence intervals of the regression coefficients include zero which also implies that the regression coefficients values are different than zero. So, we can conclude that all the considered predictors of the fitted model are significant and required to model the birth weight. Though the fitted model is significant and all the considered predictors are useful, the multiple R-squared of the fitted model is 0.0304. This implies that 3.04% of the total variations in the response variable birth weight are explained by the linear relationship of the considered predictors.

Our fitted model holds the linearity and independence assumptions. The “Residuals vs Fitted” exhibits that the residuals are bouncing randomly around the  line which confirms linearity (*Ref*. Figure 3). We have taken the samples randomly which implies that there is no dependence. In addition to this, the “Residuals Sequence Plot” doesn’t show any pattern (*Ref*. Figure 4). The normality assumption isn’t satisfied because the “Normal Q-Q” plot exhibits a noticeable skew in the points at the top and bottom (*Ref*. Figure 3). In addition to this, the “Anderson-Darling Test” provides the same conclusion with a p-value of 0.001. The constant variance assumption isn’t satisfied due to the lack of consistency of the variance into the scale and location plot (*Ref*. Figure 3). Similar result is also observed from the “Breusch- Pagan Test” with the test statistic, BP=47.276, and a p-value of . Finally, the residual vs leverage plot states that there are significant number of high leverage points in the data set.

As we have seen that the model has non-constant variance and the residuals aren’t also normally distributed. We will perform the box-cox transformation to improve the models’ behavior. As our response variable is positive so there is no restriction to use it. Here, we have found that for , we will get the maximum log-likelihood (*Ref*. Figure 5). We have to consider the nearest value to the optimum and which is 1. So, the box-cox transformation isn’t also providing a better result for this case. The histogram of the birth weight shows that the distribution is almost normally distributed so the log transformation will not be an effective one (*Ref*. Figure 6).

We have performed the stepwise regression method by considering the forward and backward selection process. We have got the full model as the best model with the AIC of 89318.73 and BIC of 89367.19. As we have seen from the pairwise plot that there is a possible association between the Mother BMI and Mother Pre-Pregnancy Weight, so we have performed the multicollinearity test. The multicollinearity is detected by the Farrar Chi-Square test and Theil's Method. Therefore, to determine the degree of multicollinearity, we have estimated the variance inflation factor (VIF) and observed that for almost all the predictors the VIF is very close to 1 except Mother BMI and Mother Pre-Pregnancy Weight. The Mother Pre-Pregnancy Weight has the highest VIF which is 6.53.

As a consequence, we will remove this predictor from the model and perform the partial F test to check the performance of the reduced model without the Mother Pre-Pregnancy Weight. The calculated value of the partial F test statistic is 17.76 with 1 degree of freedom and the p-value is .As the p-value is less than the level of significance, so we will reject the null hypothesis.

We have also fit the model with the interaction terms. As we have five significant predictors, the all possible combined interactions model contains a lot of terms and parameters. We performed the model selecting process by considering the forward and backward steps. This provides us with several suggested models. However, all the suggested models are quite long and don’t support the parsimonious model selection criteria.

**Results:** From the analysis section, we can conclude that the full model with all the considered predictors will be the better choice for modeling the newly born baby’s birth weight. Though the Adjusted of our fitted model is only 3.04% because the birth weight is not only dependent on the maternal factors but also the social, psychological, demographic, and environmental factors that play a vital role in modeling birth weight. We have observed that the model with interaction term contains more than 20 regression coefficients with an adjusted R-square of 0.06. Though the adjusted R-square is quite larger than the without interaction model, the model is no longer parsimonious. Finally, we have decided to use the full model without the interaction terms.

The main takeaway of this project is that only the maternal factors are not sufficient to model the birth weight and we need to consider all the other above-mentioned factors. Though we have observed that all the considered predictors including mother’s age, mother’s pre-pregnancy body mass index (BMI), mother’s pre-pregnancy weight, mother’s weight gain during pregnancy, and cigarette smoking before and during pregnancy plays a significant role.

**Acknowledgment**

We would like to thank Dr. Miljkovic for her valuable feedback and suggestions regarding this project.

**References**

[1] D. V. Mavalankar, R. H. Gray, and C. R. Trivedi, “Risk factors for preterm and term low birthweight in Ahmedabad, India,” *Int. J. Epidemiol.*, vol. 21, no. 2, pp. 263–272, Apr. 1992.

[2] B. L. Horta, C. G. Victora, A. M. Menezes, R. Halpern, and F. C. Barros, “Low birthweight, preterm births and intrauterine growth retardation in relation to maternal smoking,” *Paediatr. Perinat. Epidemiol.*, vol. 11, no. 2, pp. 140–151, 1997.

[3] M. C. McCormick, “The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity,” *http://dx.doi.org/10.1056/NEJM198501103120204*, vol. 312, no. 2, pp. 82–90, Jan. 2010.

[4] P. S. Mishra, D. Sinha, P. Kumar, S. Srivastava, and R. Bawankule, “Newborn low birth weight: do socio-economic inequality still persist in India?,” *BMC Pediatr.*, vol. 21, no. 1, p. 518, Nov. 2021.

[5] V. Marsoosi, A. Jamal, and L. Eslamian, “Pre-pregnancy weight, low pregnancy weight gain, and preterm delivery,” *Int. J. Gynecol. Obstet.*, vol. 87, no. 1, pp. 36–37, Oct. 2004.

[6] S. Sukalich, M. J. Mingione, and J. C. Glantz, “Obstetric outcomes in overweight and obese adolescents,” *Am. J. Obstet. Gynecol.*, vol. 195, no. 3, pp. 851–855, Sep. 2006.

[7] I. Siramaneerat, F. Agushybana, and Y. Meebunmak, “Maternal Risk Factors Associated with Low Birth Weight in Indonesia,” *Open Public Health J.*, vol. 11, no. 1, pp. 376–383, Sep. 2018.

[8] “NVSS - Birth Data.” [Online]. Available: https://www.cdc.gov/nchs/nvss/births.htm. [Accessed: 27-Nov-2021].

**Appendix A:**

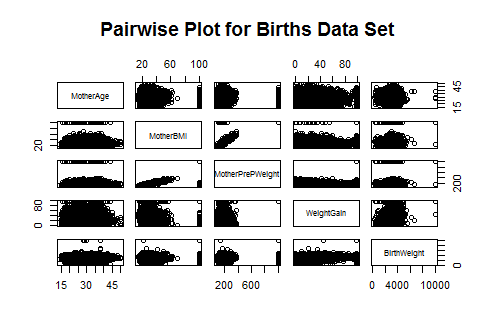


Figure 1: Pairwise Plot for Births Data Set.

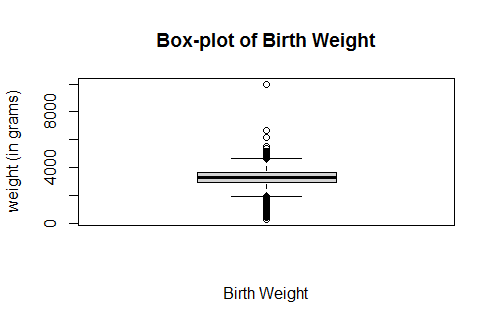


Figure 2: Box-plot of the response variable, Birth Weight.

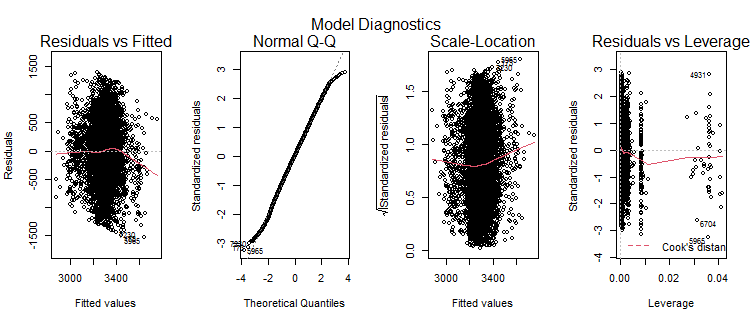


Figure 3: Fitted Regression Models Diagnostics.

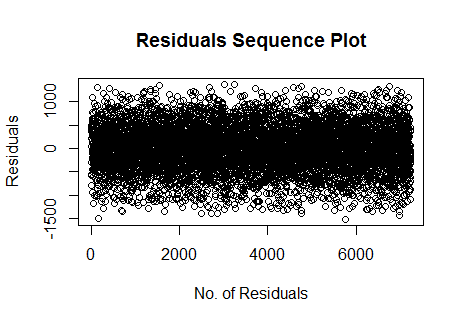


Figure 4: Residuals Sequence Plot of the Fitted Model.

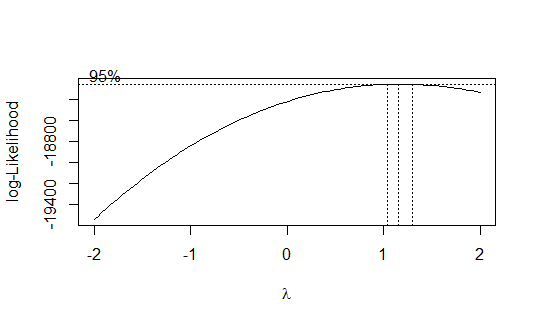


Figure 5: Plot of the log-likelihood against the transformation parameter .

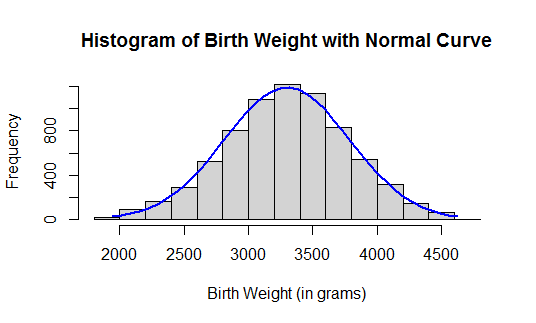


Figure 6: Histogram of Birth Weight with Normal Curve.