Identifying Risk Factors Responsible for Differences in Birth Weight

Bohan Zhang[a], **Michael Podbury**[a], **Bingkun Zhou**[a], **Jingqin Jiang**[a], **and Zhenyao Dou**[a]

[a]The University of Sydney, NSW, 2008

This manuscript was compiled on November 13, 2021

This study is aimed to investigate how the mother's health state influences the infant birth weight through developing a multiple linear regression model for prediction. The final model is formulated by applying the Akaike Information Criterion(AIC) and manually removing inappropriate predictors based on justifying the significance level. Results illustrate that the higher mothers' last menstrual weight aided in the increase in birth weight. However, non-white race, uterine irritation, smoking, and hypertension can cause a decreasing impact on birth weight, which is consistent with prior research. The adjusted r-squared of the final model is 0.195 which reflects that the goodness-of-fit in the regression model may not be as expected, also suggesting a limited extent in explaining the total variance of observed birth weight.

birth weight | maternal health | multiple linear regression

Introduction

Background. Birth weight is strongly associated with a newborn's immediate and long term health. Low birth weight (< 2500 grams) babies may be more prone to certain health issues, including sickness, infection, death in infancy, as well as susceptibility to chronic diseases later in life. High birth weight (>4000 grams) is linked to an increased risk of birth injuries, infancy mortality and long-term secretion disorders. Thus, the medical profession and society have been widely concerned for decades about the management of birth weight, with research indicating that infant birth weight is largely influenced by maternal pre-pregnancy health behaviour and characteristics. For instance, birth weight has innate racial disparities and is positively associated with the maternal level of obesity or emaciation before pregnancy. Additionally, smoking during pregnancy and tobacco exposure have been shown to significantly reduce the weight of newborns; advanced (>35 yr) or low (<16 yr) maternal age, and women with a history of hypertension or uterine irritability, are also more likely to deliver infants with a lower weight. Also, inadequate prenatal care can further raise the probability of premature labor and aberrant birth weight. The purpose of this study is to build a multiple linear regression model to predict infant birth weight given maternal health conditions, and explore which factors are significantly linked.

Data Set. The *birthwt* data set is sourced from Baystate Medical Center in 1986. However, the details of who collected the data and what was the method of collection are unclear. It contains 189 observations of birth with 10 variables, two of which are related to the baby's *birth_weight* (dependent variable), while the other eight variables describe maternal health and behavior conditions. Data cleaning includes: attributes renamed for readability; discrete counts transformed into factors; and binary indicators converted into Boolean values (True or False). For the two **quantitative** independent variables: mother's *age* and weight in pounds at last menstrual period, *weight_last_menstrual*, outliers were removed to improve statistical power. Remaining qualitative variables include: race, with three groups, white as 0, black as 1 and other as 2; *smoking_status* during pregnancy; a history of *hypertension*; and presence of uterine irritability, *uterine_irr*. The number of premature labour counts, as *premature_labor* is categorized into binary groups, 0 or 1+; and number of physician visits during the first trimester, as *physician_visits* is factorized into three groups, 0, 1, or 2+. Since the spread of *physician_visits* is skewed heavily towards 0, variables 1 and 2+ are combined due to consisting very few observations and to even the shape of spread.

Analysis

Assumptions. To guarantee validity, assumption checks were performed before and after variable selection for both full model and final model. Continuous data not fitting the linearity assumption are removed, including the age variable, which the attempt of Log transformation was still unable to satisfy a linear relationship with the dependent variable. Nevertheless, a Log transform of *birth_weight* does strengthen the linear relationship between w.r.t. the *weight_last_menstrual* variable. Thus, a Log transform of the dependent variable (*birth_weight*) is retained in the following study. Categorical data not meeting assumptions of equal variance across groups or normal distribution should be removed as well. Following this data preprocessing, the remaining qualitative variables all satisfy these assumptions. That is, the health status or behaviour of one mother should not be affected by another, thus independence is justified. In Fig.1, the residual plot of the final model shows a residual spread that is almost symmetrically distributed and without clear patterns. This represents that the dispersion of the residual is consistent across the range of fitted *birth_weight* values, thus satisfying the homoskedasiticity assumption. The residual qqplot also demonstrates that most points closely follow to the 45-degree line, with slight or negligible deviation at the ends, and which upon relying on the CLT(Central Limit Theorem) - normality can be assumed.

Model Selection. Because the *age* variable could not properly address the linearity assumption regardless of whether log transformation were applied, it was eliminated completely. Parameter selection of the final model is determined by the forward and backward stepwize Akaike information criterion (AIC). Variables are added to the null model in the forward method and removed from the full model in the backward method, until AIC no longer decreased - thus developing a model with predictors that minimise the AIC value. It shows that both method excluded the age variable which doesn't

fit the regression assumption. Interestingly, the p-value of *smoking_status* variable (0.053) is greater than the significance level (0.05), which indicates that whether or not the mother smokes has no significant influence on the birth weight. However, this contradicts the findings of several previous investigations and additionally, attempting to manually delete the variable decreases the model's r-squared value, thus reducing the model's predictive effectiveness. We have determined to include *smoking_status* in the final model.

Results

Inferences.

$$\begin{aligned}\log(\text{birth\_weight}) =& 7.8442 - 0.1314(\text{race}_{\text{black}}) - 0.1152(\text{race}_{\text{other}}) \\ &+ 0.0020(\text{last\_menstrual\_weight}) \\ &- 0.2143(\text{uterine\_irritability}_{\text{TRUE}}) \\ &- 0.2679(\text{hypertension}_{\text{TRUE}}) \\ &- 0.1020(\text{smoke}_{\text{TRUE}}) + \epsilon\end{aligned}$$

The significant coefficients of our model, represent the mean change in the response variable for one unit of change in the predictor variable - while holding all other predictors in the model constant. Each predictor variable in our model has on average, the following effect on the response variable (holding other predictors constant): A black racial background is associated with a ~13% decrease in birth weight; A racial background other than black or white is associated with a ~12% decrease in birth weight; A one pound increase in mother's last menstrual weight is associated with a ~0.2%increase in birth weight; A history of uterine irritability is associated with a ~21% decrease in birth weight; A history of hypertension is associated with a ~27% decrease in birth weight; and lastly, a positive smoking status prior to giving birth is associated with a~10% decrease in birth weight.

Performance. We compared our final model with the full model, which contains all variables used as predictors; and the simple model, which contains the most significant predictor *uterine_irritability*. Out-sample performances are tested using the "Caret" package with 10 repeats of the 10-fold cross-validation method. This method is appropriate for smaller data sets such as *birthwt*, in attaining a more reliable reading for performance. As seen in our final model, results show an RMSE value of ~0.25 and an MAE value of ~0.19 - both lower than simple and full models. A lower RMSE and MAE value, implies higher accuracy as it minimises the difference between predicted and observed values.However, our model does record a low R-squared value of ~0.21. Additionally, in-sample R-squared and adjusted R-squared values also reveal low R-squared scores of ~0.22 and ~0.19. This indicates that our model only accounts for ~21% of the variation in the actual values for log birth weight. As these values are higher than both simple and full models, the final model is preferred in maximising the explanatory power of the model in relation to the response variable. Nonetheless, a low value conveys a weak relationship between our model and the response variable. This may mean that our model needs more variables to account for the remaining variation in data; or simply that the data itself is inherently unpredictable (too complex to be fully explained).

Although our model has a low R-squared score, a simple check for the p-value of the F-statistic shows *1.29e-07*, supporting our model as statistically significant in fitting the data compared to a model in the rejected H0 - where predictors are assumed to have no relationship with the log of birth weight. In addition, our model can still be used to draw conclusions about how changes in the predictor values can effect changes in the response variable (as shown above). For making predictions - the low R-squared value does hint towards lower precisison, i.e. having a wider prediction interval (PI). Consider row 100 from *birthwt_lg*; this mother is 100 pounds, white, smokes, has had no history of hypertension or uterine irritability. According to our model the log birth weight value should be ~7.73 with a 95% PI of [~7.23, ~8.23]. The actual value for row 100 is ~7.92 (also within PI). If we back-transform our model, the predicted birth weight in grams is the multiplication of all exponentiated predicted variables when multiplied with corresponding predictor values. Thus, row 100 is predicted using back-transformed coefficients to be ~2820.45 grams. The actual predicted birth weight value for row 100 is 2769 grams. This is a ~51 gram difference - not bad! However, this is only for one point chosen from the data set. More interestingly, the scaled PI after back-transforming is [~2589.89, ~2948.11] - this is an interval of ~358 grams.

Discussion

Limitations. In general: The low adjusted r-squared value indicates that the final model is only able to explain ~21% of total birth weight variance.The reliability of model is questionable since sample size is small (189 observations),which has limited the ability to predict infant birth weight by five predictors. With data collected in 1986 as outdated, and demography likely to have changed significantly - the data is likely inconsistent with the present.Better medical treatment, social welfare, and human rights could even out variables such as racial effect on birth weight.All observations are from only one medical center may cause selection bias, such that the model is incapable of effectively capturing the whole human population.There is no information about who and how the data is collected, so accuracy and credibility of the data is in doubt. Moreover, because of improper collecting techniques, samples may interact and affect each other, causing the data to fail to fulfill the model's independence assumption.

For specific variables: Only 12 of the 189 women had a history of hypertension, making it difficult to determine if birth weight is normally distributed for this group, and there will be deviations owing to insufficient sample size when measuring the influence of this variable.Furthermore, additional factors that have been discovered to have a significant impact on birth weight in previous research, such as baby's gestational age, indicator of gestational diabetes and average alcohol intake during the duration of the pregnancy are not addressed by.The inadequate consideration of potential factors may also reduce the predictive power of the final model.

Conclusion. To guarantee variety and minimize selection bias, further researches should utilize larger and more recent data sets, and sample collecting should include medical centers from diverse locations. In addition, other maternal health factor categories should be explored for collection, and details such as sample data gathering techniques should be included. This can assist give a more comprehensive and rigorous perspective on which factors contribute to the birth weight.

Appendix

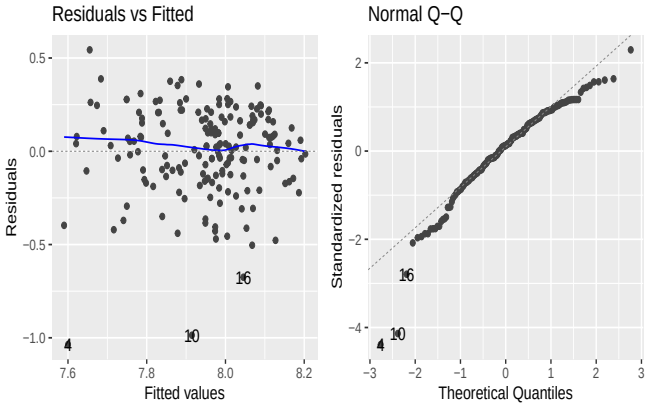

Fig. 1. Residuals plots for the final regression model

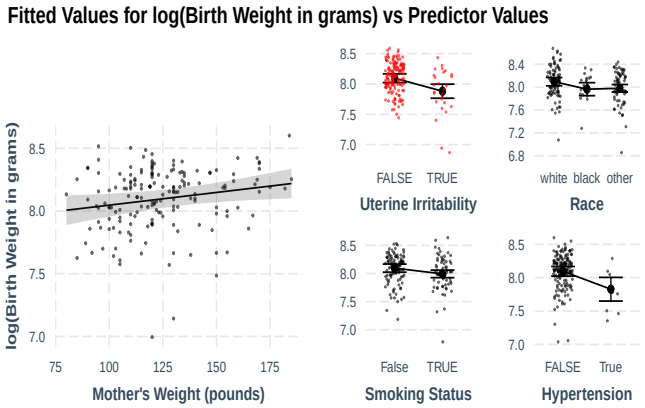

Fig. 2. Predictor effect plots for the final regression model

Table 1. Out of Sample Performance of Models

model	RMSE	Rsquared	MAE
Simple	0.2602087	0.1546183	0.2059141
Full	0.2510851	0.2052202	0.1947776
Final	0.2492008	0.2124256	0.1936373

Table 2. In Sample Performance of Models

model	RMSE	Rsquared	Adj-Rsquared	MAE
Simple	0.2624828	0.0921336	0.0868858	0.2046209
Full	0.2383641	0.2513098	0.2056580	0.1815581
Final	0.2429023	0.2225301	0.1947634	0.1852899

References

GitHub repository

[1] Bernstein, I., Mongeon, J., Badger, G., Solomon, L., Heil, S. and Higgins, S., 2005. Maternal Smoking and Its Association With Birth Weight. Obstetrics & Gynecology, 106(5, Part 1), pp.986-991.

[2] England, L., 2001. Measures of Maternal Tobacco Exposure and Infant Birth Weight at Term. American Journal of Epidemiology, 153(10), pp.954-960.

[3] Frederick, I., Williams, M., Sales, A., Martin, D. and Killien, M., 2007. Pre-pregnancy Body Mass Index, Gestational Weight Gain, and Other Maternal Characteristics in Relation to Infant Birth Weight. Maternal and Child Health Journal, 12(5), pp.557-567.

[4] Goisis, A., Remes, H., Barclay, K., Martikainen, P. and Myrskylä, M., 2017. Advanced Maternal Age and the Risk of Low Birth Weight and Preterm Delivery: a Within-Family Analysis Using Finnish Population Registers. American Journal of Epidemiology, 186(11), pp.1219-1226.

[5] González-Jiménez, J., & Rocha-Buelvas, A. (2018). Risk factors associated with low birth weight in the Americas: literature review. Revista De La Facultad De Medicina, 66(2), 255-260. doi: 10.15446/revfacmed.v66n2.61577

[6] Journal of Nurse-Midwifery, 1986. Maternal weight and birth weight Abrams B, Laros R: Prepregnancy weight, weight gain, and birth weight. AM 3 OBSTET GYNECOL 154:503, 1986. 31(5), p.242.

[7] Rahfiludin, M., & Dharmawan, Y. (2018). Risk Factors Associated with Low Birth Weight. Kesmas: National Public Health Journal, 13(2). doi: 10.21109/kesmas.v13i2.1719

[8] (2021). Retrieved 27 October 2021, from https://www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf

[9] Wilcox, A., 2001. On the importance—and the unimportance— of birthweight. International Journal of Epidemiology, 30(6), pp.1233-1241.

[10] Wiley-Blackwell. (2011, December 13). Mothers' weight before and during pregnancy affects baby's weight. ScienceDaily. Retrieved October 25, 2021 from www.sciencedaily.com/releases/2011/12/111213110521.htm