STAT2402 Analysis of Observations
Assignment 2 Report

Johan Carlo A. Ilagan, 23832843

18 October 2024

Executive Summary

This study is aimed to create a model highlighting the relationship between low birth weight and maternal variables. It revolves around 8 explanatory variables - age, weight, race, smoking status, previous premature labours, history of hypertension, presence of uterine irritability, number of physician visits. The data has 187 rows, and is used to gain descriptive statistics to initially analyse the data. To develop the model, logistic regression for binomial data was applied, which formulated three models. First model involved all variables without interaction terms, second model used `stepAIC()` function, and third model used manual reduction of unsignificant predictors. Through ANOVA, the third model was classified as the best model out of the three and was used to determine the relationships. Maternal age had a negative relationship with the log odds of low birthweight, while race, smoking status, and history of hypertension positively correlated with the log odds of low birthweight.

1. Introduction

Identifying cases of low birth weight as early as possible is very crucial for immediate medical care, as it can greatly influence the mother's health, and especially the child's development. It can almost always be associated with short-term health problems to the baby, such as low oxygen levels, breathing problems, nervous system problems, as well as long-term problems like blindness, deafness, and impediments to growth (Stanford Medicine Children's Health, n.d.). The World Health Organisation (WHO) defines low birth weight as having a weight of less than 2.5 kilograms (WHO, n.d.). Over the years, numerous research and data have been made to determine possible causes or symptoms that could lead to low birth weight.

Studies made by Anil et al (2020) and Diabelková et al (2022) both discuss the significance of preterm birth to the likelihood of low birth weight. Meanwhile, AIHW (n.d.) explains that age and drug use among others are also variables that affect low birth weight. Wubetu et al (2021) supported this by using descriptive statistics and multiple linear regression to determine its correlation with age and maternal weight. Arabzadeh et al (2024) on the other hand, focused solely on drug use as a cause for low birth weight, using the meta-umbrella R package as their tool for analysis. Lastly, Adugna and Worku (2022) confirmed how hypertension, body mass index (BMI), preterm birth, and more are significant contributors to low birth weight. Most of these studies involve the use of multivariable logistic regression, which Kalan et al (2021) defines as a method that is used for analysing data with one dependent variable and multiple independent variables. While simple logistic regression is used for when there is one binary outcome and just one independent variable, multivariable logistic regression is used to find the best equation that predicts the success of the binary outcome with multiple independent variables. This will lead us to the equation,

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

In this analysis, we will use a linear model to determine the relationship between maternal variables and the birth weight of the baby, specifically whether or not it is considered low birth weight. The study will contain a methodology, followed by the results and discussion, where we will discuss findings based on the linear model and will come up with a conclusion.

2. Methodology

As the dependent variable for this set of data only has two outcomes (1 if $bwt < 2.5kg$ and 0 if $bwt \geq 2.5kg$), we can consider this as binary outcomes. In this study, the data will be analysed through logistic regression for binomial data. We will first examine the data and anaylse its descriptive statistics (min, mean, median, max, standard deviation). We will then plot graphs looking for any initial patterns with low birth weight and independent variables (maternal variables). The model will not include any interaction terms, as there are no evidence that any variable depends on another variable. After the logistic regression model is formed, it will be reduced to the significant terms only. The final model will explain what variables in the data correlate to low birth weight, and it will allow us to form the final equation for the log odds of low birth weight. All computations, graphs, and the formulation of logistic regression model is done in the R statistical environment. Significance level for this study is set as $\alpha = 0.05$.

3. Results

For descriptive statistical analysis, we will analyse any patterns with their counts (if it is count data) or with their min, median, mean, max, and standard deviation (if it is continuous data). Using `summary()` and `sd()` functions, we will arrive with the following numbers:

'low': 'No'= 129, 'Yes' = 58

'age': Min = 14.0; Median = 22.0; Mean = 23.1; Max = 36.0; SD = 5.07

'lwt': Min = 80.0; Median = 121.0; Mean = 129.9; Max = 250.0; SD = 30.73

'race': 'White' = 95, 'Black' = 26, 'Other' = 66

'smoke': 'No' = 114, 'Yes' = 73

'ptl': Min = 0.0; Median = 0.0; Mean = 0.1925; Max = 3.0; SD = 0.49

'ht': 'No' = 175, 'Yes' = 12

'ui': 'No' = 160, 'Yes' = 27

'ftv': Min = 0.0; Median = 0.0; Mean = 0.7968; Max = 6.0; SD = 1.06

'bwt': Min = 1021; Median = 2977; Mean = 2946; Max = 4593; SD = 698.65

Looking at these set of numbers, we can see that all of the variables with count data have a big gap between the counts of each category, especially the dependent variable `lwt` with 129 for 'No' and 58 for 'Yes'. Another thing to note is the big standard deviation for `bwt`, which indicates data for the birthweight are spread out. Lastly, one distinguishable attribute is that median for both `ptl` and `ftv` are 0, but their means are far from 0, indicating how common the variable is at 0, but there are extreme cases that make the mean jump up. The graphs for numerical data are formed using boxplots, while the graphs for the count or categorical data are formed using mosaic plots. The boxplots do not show any relationship and will thus need to further be analysed. However, the mosaic plots did show a correlation with `smoke`, `ht`, `ui`, indicating that smoking, hypertension, and the presence of uterine irritability influence the likelihood of low birthweight (See Figure 1).

`M1` is a regression model including all terms, without including any interaction terms. The model (`M1`) is then reduced using the `stepAIC` function from the library `MASS`, with a backwards direction. This will be referred to as `M2`. Lastly, `M2`, will then further be reduced manually using the update function, removing variables that are unsignificant i.e. $\alpha > 0.05$, and will be referred to as `M3`. Based on the p-values from `M2`, variables removed to form M3 are `ptl` and `ui`.

The objective of this study is to find a relationship between the maternal variables and low birth weight. To get the model that we will utilise, we will use the `anova` function to compare the three models and find what is statistically the best and most significant model. Based on the p-values in the ANOVA test, M3 presents a great change in statistical significance (p-value) compared to M2 and will therefore be the model that we will conclude with. The model for log odds of low birth weight formed through the logistic regression for binomial data and ANOVA is as follows:

$$logit(\widehat{\pi}) = -0.018(lwt) + 1.26(race2) + 0.863(race3) + smoke1 + 1.757(ht1)$$

Using `residuals()` and `fitted()`, we plot the residuals against the fitted values to assess the model, in this case M3. The plot (see Figure 2) demonstrated two parallel lines, one for observations equal to 0 and the other one for observations equal to 1. The plot does not show any obvious violations of the model. Aside from that, using the deviance goodness-of-fit test, the p-value gained from that is 0.00014, which is sufficient evidence to conclude that M3 is the better model.

4. Discussion

Based on M1, significant predictors of low birth weight include `lwt`, `race`, `smoke`, and `ht` as they are the only variables that reached a p-value of below 0.05. Based on M3, where unsignificant predictors are removed, the equation is $logit(\widehat{\pi}) = -0.018(lwt) + 1.26(race2) + 0.863(race3) + smoke1 + 1.757(ht1)$. The effects of the maternal variables on the log odds of low birth weight are as follows:

1. The maternal age is the only variable that has a negative relationship with low birth weight. As age increases, the odds for low birth weight decreases by a factor of 0.018.
2. The coefficients for race varies on its categorical value. For `Black` (race = 2), the odds for low birth weight increases by a factor of 1.26. For `Other` (race = 3), the odds for low birth weight increases by a factor of 0.863
3. Another significant predictor is the smoking status of the mother. Since the coefficient is 1, we can say that if the mother smokes during pregnancy, the odds for low birth weight increases by a factor of 1.
4. Lastly, the model shows that having a history of hypertension tends to increase the likelihood of low birth weight by a factor 1.757.

Using this model and study, we can contribute to ongoing research who is looking to find significant causes of low birth weight, so that as much as possible we can reduce the cases or if not, minimise the bad effects it results to.

Appendix

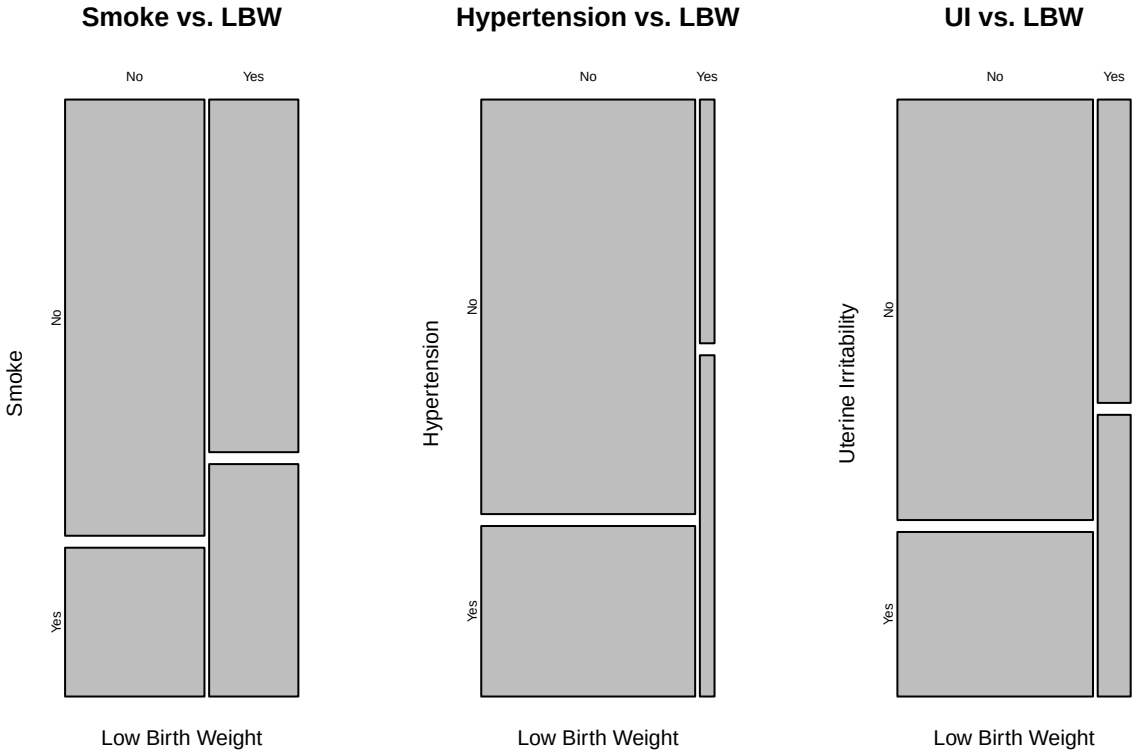

Figure 1. Mosaic Plots of `smoke`, `ht`, and `ui` against `low`

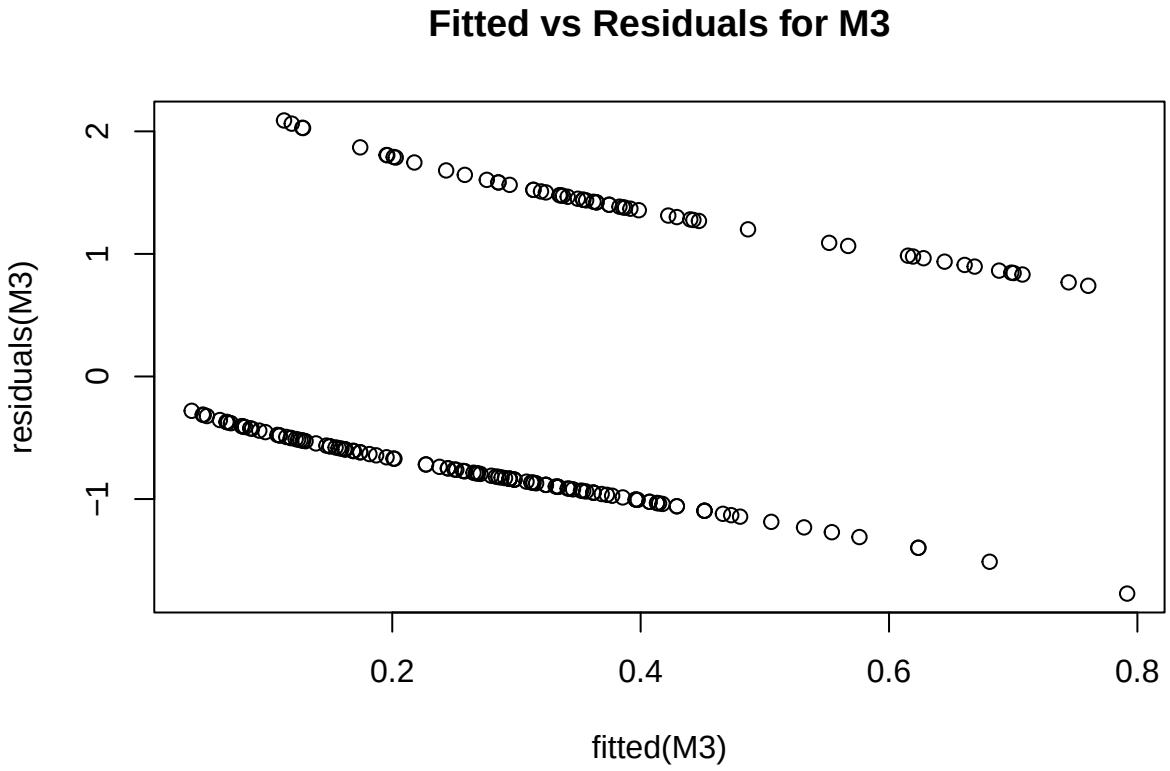

Figure 2. Fitted Values vs Residuals for M3

References

Adugna, D. G., & Worku, M. G. (2022). Maternal and neonatal factors associated with low birth weight among neonates delivered at the University of Gondar comprehensive specialized hospital, Northwest Ethiopia. Frontiers in Pediatrics, 10. https://doi.org/10.3389/fped.2022.899922

Arabzadeh, H., Doosti-Irani, A., Kamkari, S., Farhadian, M., Elyasi, E., & Mohammadi, Y. (2024). The maternal factors associated with infant low birth weight: an umbrella review. BMC Pregnancy and Childbirth, 24(1). https://doi.org/10.1186/s12884-024-06487-y

Australian Institute of Health and Welfare. (2020). Australia's children, Birthweight. Australian Institute of Health and Welfare. https://www.aihw.gov.au/reports/children-youth/australias-children/contents/health/birthweight

Diabelková, J., Rimárová, K., Urdzík, P., Dorko, E., Houžvičková, A., Andraščíková, Š., Drabiščák, E., & Škrečková, G. (2022). Risk factors associated with low birth weight. Central European Journal of Public Health, 30, S43–S49. https://doi.org/10.21101/cejph.a6883

K. C., A., Basel, P. L., & Singh, S. (2020). Low birth weight and its associated risk factors: Health facility-based case-control study. PLoS ONE, 15(6), e0234907. https://doi.org/10.1371/journal.pone.0234907

Kalan, M. E., Jebai, R., Zarafshan, E., & Bursac, Z. (2020). Distinction between two statistical terms: multivariable and multivariate logistic regression. Nicotine & Tobacco Research, 23(8). https://doi.org/10.1093/ntr/ntaa055

Stanford Medicine Children's Health. (n.d.). Low Birth Weight. Stanford Medicine Children's Health. https://www.stanfordchildrens.org/en/topic/default?id=low-birth-weight-90-P02382

World Health Organization. (2023). Low birth weight. Www.who.int. https://www.who.int/data/nutrition/nlis/info/low-birth-weight

Wubetu, A. D., Amare, Y. E., Haile, A. B., & Degu, M. W. (2021). Newborn Birth Weight and Associated Factors Among Mother-Neonate Pairs in Public Hospitals, North Wollo, Ethiopia. Pediatric Health, Medicine and Therapeutics, 12, 111–118. https://doi.org/10.2147/phmt.s299202