Risk Factors Associated with Low Infant Birthweight

Sophia Yuen[a], Gisella Sutanto[a], Xiaoying Wei[a], James Finlay[a], and Angus Vrantsis[a]

[a]013E03, University of Sydney

This version was compiled on November 5, 2023

What are the risk factors associated with a lower infant birth weight? This report aims to solve this with linear regression in order to help mothers give birth to healthier babies. It was found that the following are negatively associated with an infant's birth weight when the mother: 1. has a history of uterine irritability, 2. is non-white, 3. smokes, 4. has a history of hypertension, 5. has had one or two premature deliveries, and 6. her age.

Introduction. Studies have shown that there is an association between a newborn's weight and its immediate health. A lower weight at birth predisposes a newborn to long-term health complications, which may result in a premature death. In this report, multiple linear regression was used to identify the various factors that influence an infant's birth weight. The results of this study will allow researchers to develop strategies to help ensure that babies are born at a healthier weight, leading to improved infant health outcomes.

Data Description. The data we used was collected at Baystate Medical Center in 1986 and there were 189 samples. The variables in our data is shown below in Table 1.

Bwt was the dependent variable we aimed to predict. The variable low was removed from our calculations since it is essentially telling us the same thing as low.

Table 1. Variables in bwt Data

Variable	Description
bwt	Birth weight in grams.
low	Indicator if birth weight is less than 2.5 kg.
lwt	Mother's weight in pounds at her last menstrual period.
age	Mother's age in years.
race	Mother's race (1 = white, 2 = black, 3 = other).
smoke	Did the mother smoke during pregnancy?
ptl	Number of previous premature labors.
ht	Does the mother have a history of hypertension?
ui	Was there presence of uterine irritability?
ftv	Number of physician visits during the first trimester

Initial Assumptions. Before we performed linear regression, we checked the linearity and normality of our numerical variables which were the Mother's age (seen in Figure 1) and weight at last menstrual period (lwt) (seen in Figure 2). Our initial test of assumptions (seen in Appendix 1) showed that there was an outlier where a mother was over 40. This value was removed to better fit our assumptions.

Following the removal of the outlier, the linearity assumption is well-met by both variables. The normality assumption shows some concerns in both variables where some points deviate from the theoretical line but our sample size is large enough for us to rely on the central limit theorem. Hence all assumptions are met.

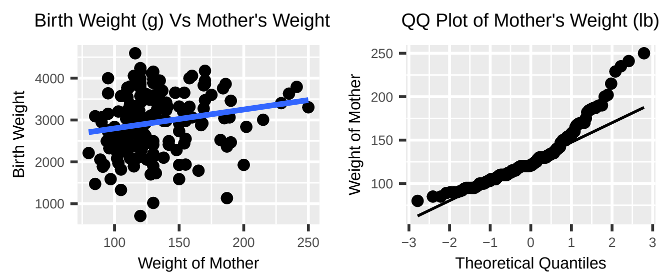


Fig. 1. Testing Linearity and Normality of lwt

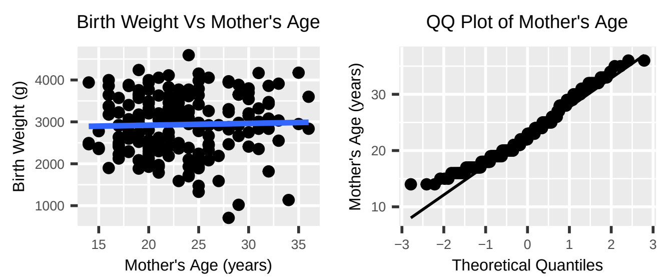


Fig. 2. Testing Linearity and Normality of age

Model Selection. Our initial model was produced with the Stepwise Method, both backward and forward (Appendix 2 & Appendix 3). The result from both selections resulted in the same model. The variables age and ftv (number of physician visits in the first trimester) were deleted.

$$\begin{aligned} \text{bwt} = {} & 2834.32 - 542.51(\text{ui}_{\text{yes}}) - 445.61(\text{race}_{\text{black}}) - \\ & 310.8(\text{race}_{\text{other}}) - 330.18(\text{smoke}_{\text{yes}}) - 573.97(\text{ht}_1) + \\ & 4.31(\text{lwt}) - 294.43(\text{ptl}_1) - 15.49(\text{ptl}_2) + \\ & 1266.34(\text{ptl}_3) \end{aligned}$$

It must be noted that the Stepwise Method has its limitations as it does not consider the theoretical significance of each variable. It favours statistically significant variables and lacks the context of the issue at hand, thus it may drop variables that are actually meaningful (Smith,2018). Hence it is important to supplement the Stepwise Method with further research.

As we conducted more research to enhance our model, it was found that risk of low infant birthwight decreased until the maternal age of 36 (Wang et al.,2020) (Our data stops at 36). Considering this, we chose to add back the age variable, resulting in the following model.

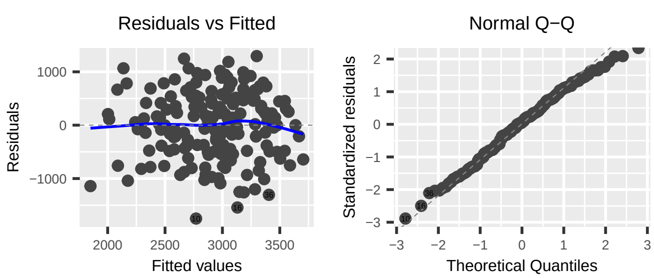
$$
\begin{aligned}
bwt = {} & 2878.98 - 2.21(age) - 544.75(ui_{yes}) - \\
& 453.36(race_{black}) - 315.19(race_{other}) - 333.07(smoke_{yes}) - \\
& 576.66(ht_1) + 4.38(lwt) - 288.34(ptl_1) - \\
& 15.09(ptl_2) + 1274.61(ptl_3)
\end{aligned}
$$

Interpretation of Final Model. The final model we have selected is based on mostly categorical variables, but our model also includes numerical variables like the mother's weight (lwt) and her age. It is the factors that are negatively associated with birth weight which demand the highest attention from practitioners as they can be readily managed in a clinical preventative setting.

The parameters that are significant at the 0.05 p-value and have a negative association with birth weight are: 1) if a mother has a history of uterine irritability, 2) if a mother is black, 3) if a mother is race 'other', 4) if a mother smokes, and 5) if a mother has a history of hypertension. The medical factors and pathologies aforementioned are supported by other literature. In addition, the race factors of 'black' and 'other' are most likely proxies for socioeconomic conditions that a mother is subject to.

In summary, regarding the factors with a negative association with birth weight, these are all seemingly cogent.

Checking Assumptions. It is important to confirm the validity of our model assumptions: 1) the model is linear in parameters and is correctly specified, 2) the residuals are homoscedastic, 3) the residuals are normally distributed and have an expected value of zero, and 4) the values of the residuals have independent distributions. As shown in the two plots in the following figure, the final model appears to meet these conditions.



Assessing In Sample Performance. Table 2 shows the rsquared values for the model developed by the Stepwise Method with and without age. It can be seen that the values are almost identical. The values indicate that 27% of the variation in birth weight is explained by each model. Although the difference is very miniscule, the model with age performs slightly better than without. However, further assessment is required to ensure that our model does not suffer from overfitting.

Table 2. R squared value comparison

Variable	Description
R squared of with age	0.2758835.
R squared of without age	0.27105771.

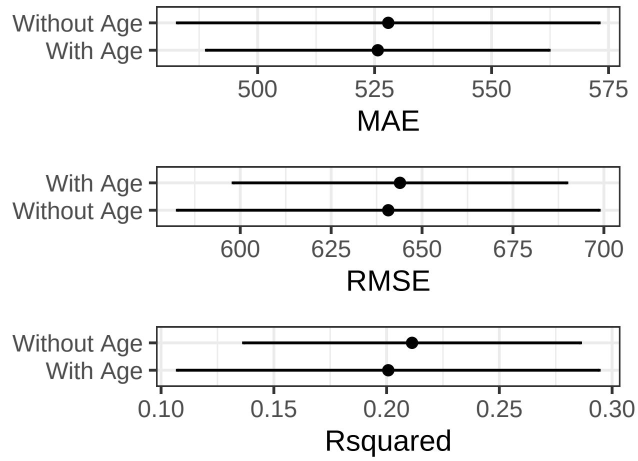


Fig. 3. Results from Cross-Validation

Assessing Out of Sample Performance. To protect our model against overfitting, we performed cross-validation where we split our data into ten groups. We built our model using 9 of the 10 groups and tested our model on the remaining group. For further accuracy, we repeated this 10 times and averaged out the error rates. We did this for the model that we got after performing stepwise and for the model where we added age back in and compared the performances of each.

From this it can be seen that both models are relatively on par with each other having very similar error rates. This indicates that including age in our model has not created a problem of overfitting that was not there before. From this, we have concluded to keep the model with age as our final model, since it provides us extra information and does not seem to perform worse.

Limitations of our Model. Linear regression is commonly used since it is easy to understand and implement however it can be an oversimplification of the relationships and some assumptions may not hold.

The data was collected in 1986 and there have been several medical advancements since then so our results may be outdated. It was sourced from only one hospital, which may violate our independence assumption. For example, some mothers could have blood relations or have the same doctor which could cause a bias.

Our model is also limited to the measurements that were collected in this data. There are various other variables that could also influence birth weight that were not considered such as the mother's income or any previous diseases she has had.

Birth weight is influenced by a multitude of complex factors that cannot be explained by a simple linear equation. More sophisticated statistical techniques and research is required.

Conclusion. The final model produced has some explanatory power regarding an infant's birthweight. Further, most of the model's parameters are supported by other research in terms of their expected effect regarding birthweight. More data is required to improve the model, but in addition to this, future research could refine some of the data being collected: the 'race' factors could be replaced with information that has a clearer link to economic status such as a mother's household income. The final model shows that certain aspects of a mother's health and factors that are likely related to her economic status are associated with a baby's birthweight.

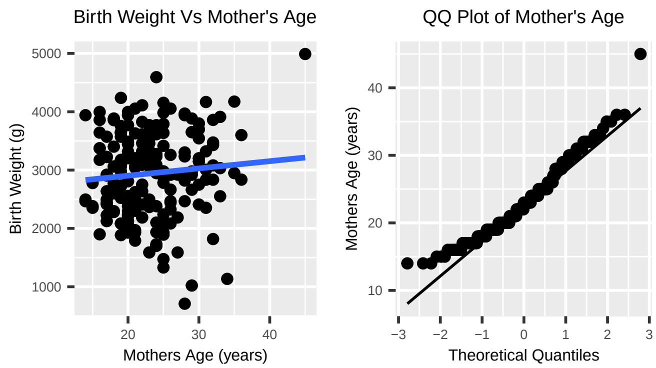


Fig. 4. Appendix 1

Table 3. Appendix 2: Backward Selection

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2783.761	238.506	11.672	0.000
uiyes	-534.401	133.959	-3.989	0.000
ptl1	-291.754	141.127	-2.067	0.040
ptl2	-8.839	288.945	-0.031	0.976
ptl3	1272.486	644.096	1.976	0.050
ht1	-569.763	193.787	-2.940	0.004
lwt	4.451	1.633	2.725	0.007
raceblack	-426.927	141.875	-3.009	0.003
raceother	-284.062	110.166	-2.578	0.011
smokeyes	-307.734	102.569	-3.000	0.003

References. Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0143-6 Wang, S.,

Yang, L., Shang, L., Yang, W., Qi, C., Huang, L., Xie, G., Wang, R., & Chung, M. (2020). Changing trends of birth weight with maternal age: a cross-sectional study in Xi'an city of Northwestern China. BMC Pregnancy and Childbirth, 20(1). https://doi.org/10.1186/s12884-020-03445-2

Table 4. Appendix 3: Forward Selection

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2783.761	238.506	11.672	0.000
uiyes	-534.401	133.959	-3.989	0.000
ptl1	-291.754	141.127	-2.067	0.040
ptl2	-8.839	288.945	-0.031	0.976
ptl3	1272.486	644.096	1.976	0.050
ht1	-569.763	193.787	-2.940	0.004
lwt	4.451	1.633	2.725	0.007
raceblack	-426.927	141.875	-3.009	0.003
raceother	-284.062	110.166	-2.578	0.011
smokeyes	-307.734	102.569	-3.000	0.003