Executive summary

Zhimin Liu, Yunshuo Zhang, Yi Yang, Dahui Liu, Xinyi Jiang
This version was compiled on November 9, 2021

Abstract
Low Birth Weight(LBW) is one of the main predictors of infant mortality. This study undertakes analysis of various risk factors associated with LBW using Multiple Linear Regression Model, Akaike information criterion (AIC) and Cross Validation(CV). Racial difference, Smoking, Hypertension, Uterine irritability and Mother's weight are concluded as the most significant variables in the final reduced model. Limitations are clarified to enhance further research, and some suggestions related to prenatal care are offered to mothers to improve infant health.

Introduction Low Birth Weight(LBW) refers to babies weighing less than 5 pounds and 8 ounces at birth, which is one of the main predictors of infant low IQ, stunted growth, and even mortality. According to UNICEF(2015), nearly 15% of babies worldwide are born with low birth weight. There are many known risk factors, the most important of which are 3 types: socio-economic factors, medical risks before or during gestation and maternal lifestyles.
In this study, the question of interest is: **which specific risk factors really contribute to predicting LBW?**

Data set. The dataset was collected at Baystate Medical Center in Massachusetts in 1986 from 189 infants and their mothers(Hosmer, David W., et al., 2013).

3 types of risk factors are measured to analyze LBW. The first type of risk factor is socio-economic, which includes maternal `race`, `age` and `weight`. The second is medical risks before or during gestation, such as `uterine irritability`, `hypertension`, `physician visits` and `previous premature labours`. The last is maternal lifestyles, which is `smoking` status.

Birth weight in grams is measured as outcome, and values under 2500 grams are encoded as low birth weight in the indicator `low` in the dataset(Appendix 1).

IDA To ensure data validity, Initial Data Analysis (IDA) was conducted before model selection. Firstly, categorical variables such as `race`, `smoke`,`hypertension`,`uterine_irritability` were converted from integers(eg., 0,1,2) to factors with correct levels and descriptive labels. For example, race which was initially encoded as 1,2,3 is factorized as 'white', 'black', and 'other', to ensure that linear model function could treat such data correctly. Secondly, terminology such as `ui`, `lwt`,`ht` were renamed as "uterine.irr", "mother.weight", "hypertension" to facilitate comprehension. Thirdly, indicator of LBW is removed from linear model to avoid duplication of dependent variable.

Hypothesis In the multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{g-1} X_{g-1} + \beta_g X_g + \ldots + \beta_{p-1} X_{p-1} + \varepsilon \quad \text{(full model)}$$

Does the subset of all these 8 independent variables contribute to predicting birthweight? Or, would do just as well if these variables were dropped and the full model is reduced to:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_g X_{g-1} + \varepsilon \quad \text{(reduced model).}$$

Null Hypothesis: The initial assumption is that there is no relation, which is expressed as: H0: $\beta 1 = \beta 2 = \ldots = \beta p\text{-}1 = 0$.

Alternative Hypothesis: At least one of the independent variables is useful in explaining/predicting Y, expressed as: H1: At least one βi is $\neq 0$.

Assumptions. The 4 assumptions are checked on both full model and reduced model. For the full model, 4 assumptions are checked(Appendix 2): 1.Linearity: There is no obvious pattern in the residuals vs fitted values plot. The blue line on it also does not show any frowny face or smiley face. 2.Independence: Observations are not related to one another. 3.Homoskedasticity: There are not some fanning out of the residuals in the residuals vs fitted plots. The spread looks reasonably constant over the range of fitted values. Points are randomly distributed around 0. 4.Normality: QQ plot shows the majority of the points lie close to the diagonal lines which indicate

that the normality assumption is reasonably well satisfied.

Therefore, since all four assumptions are satisfied, we can use these data to do further model selection.

For the reduced model(Appendix 3), there does not seem to be any patterns or fan shapes. Linearity and homoskedasticity assumptions are fitted. In the normal QQ plot, the points are all reasonably close to the diagonal line. Therefore normality assumption is still satisfied. In conclusion, the reduced model also passes assumption checking.

Model Selection For the full model, it includes all parameters and corresponding coefficients. But since not all variables have p value less than 0.05, insignificant variables should be further removed from the full model to avoid overfitting.

Therefore, Akaike information criterion (AIC) was applied to select appropriate variables for better model. For each variable in the full model, AIC investigate effect of removing it and including it. Insignificant variables, such as `physician.visits`(p=.71), `mother.age`(p=.64), `previous.prem.labor`(p=.76) have the lowest AIC, which indicates that the information loss by dropping these 3 variables are the least among all variables. Finally, AIC generated a reduced model of infant birth weight with 5 most informative factors.

Model Performance In terms of **in-sample performance**, reduced model is shown to have 1% higher adjusted r^2 than full model (Appendix 4). Although the explanatory power of reduced model is not strongly better than full model, it does show a outstanding out-of-sample performance.

To ensure model validity, **out-of-sample performance** was compared between full model and reduced model using Cross Validation(CV). Firstly, dataset is divided into 10 folds, 18 folds of 10 observations and 1 fold of 9 observations. Secondly, one fold is left out from both models as test set for the remaining 9 training sets to make predictions on and errors are calculated. Finally, root mean square(RMSE) and mean absolute errors(MAE) are compared between full model and reduced model, by aggregating the errors over 10 folds. As is shown in Appendix 5, reduced model shows an obviously lower RMSE and MAE, which strongly proves its model validity.

Therefore, through AIC and CV, a reduced model with better explanatory power is finally shown as:

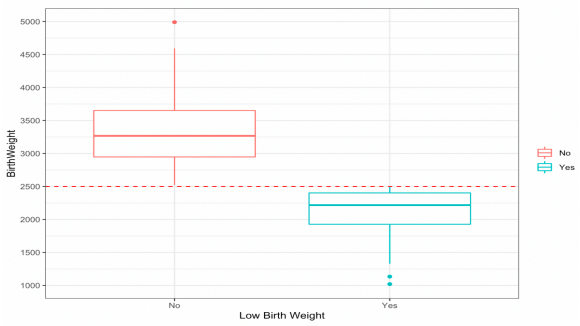
$$\widehat{\text{birthwt.grams}} = 2837.26 + 4.24(\text{mother.weight}) - 475.06(\text{race}_{\text{black}}) - 348.15(\text{race}_{\text{other}}) - 356.32(\text{mother.smokes}_{\text{Yes}}) - 585.19(\text{hypertension}_{\text{Yes}}) - 525.52(\text{uterine.irr}_{\text{Yes}})$$

Results The results show that infant birth weight is associated with variables race, mother smokes, hypertension, and uterine iritatability. Babies from black or other demographic groups are more likely to have lower weight than white, 16% and 13% respectively. Smoking reduces birthweight by 12%. Hypertension reduces birthweight by 25%. Uterine iritatability reduces birthweight by 22%.

Limitations. Firstly, this study is limited to dated data, as observations were initially made in 1989. Therefore, more contemporary data should be collected in the future to explain LBW in the modern day. Secondly, sampling bias is also a concern of this data set collected from the United States, because incidence of LBW differ a lot between developing countries(5-10%) and developed countries(10%-20%) (UNICEF, 2015). Therefore, more universal data should be collect to increase model generality.

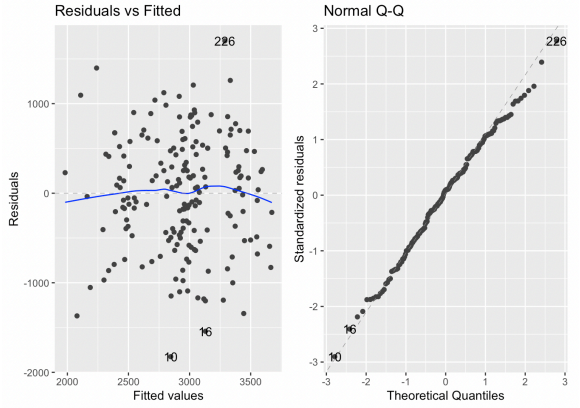
Conclusion. Through our analysis and research on the data, we found that the infant birth weight is correlated with 5 risk factors which are the weight of the mother, race smokes, hypertension, and uterine irritability. Therefore, we should provide more care and support to pregnant mothers regardless of the race so that the babies could have a good growth environment.

Appendix 1



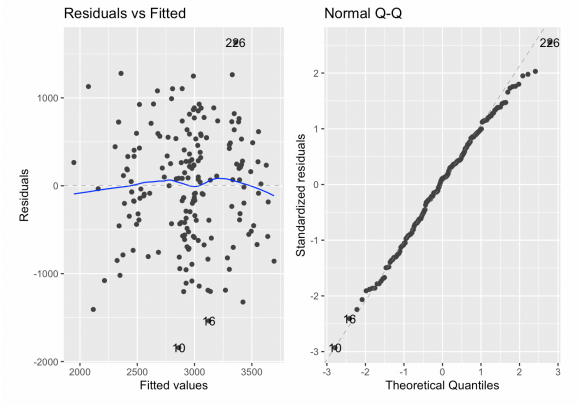
IDA: Low Indicator and Birth Weight

Appendix 2



Assumption Checking: Full Model

Appendix 3



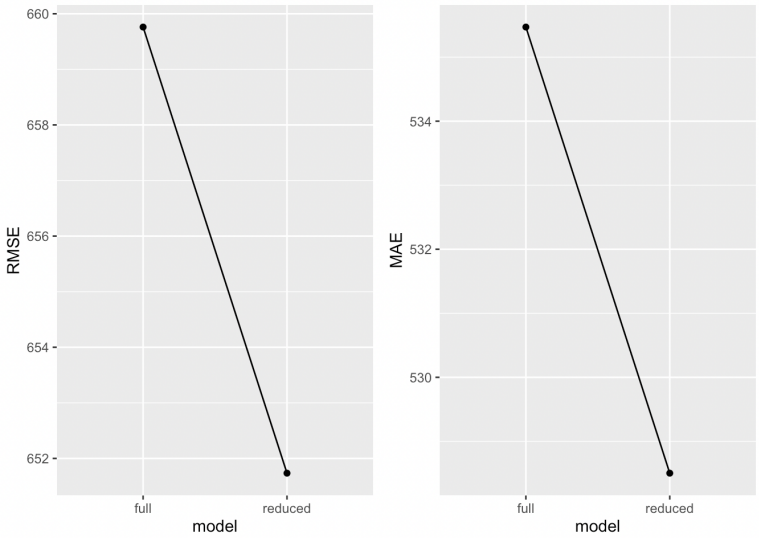
Assumption Checking: Reduced Model

Appendix 4

	Full Model		Reduced Model	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	2927.9619	**<0.001**	2837.2639	**<0.001**
mother age	-3.5699	0.711		
mother weight	4.3540	**0.013**	4.2415	**0.012**
race [black]	-488.4275	**0.001**	-475.0576	**0.001**
race [other]	-355.0771	**0.002**	-348.1504	**0.002**
mother smokes [Yes]	-352.0445	**0.001**	-356.3209	**0.001**
previous prem labor	-48.4020	0.636		
hypertension [Yes]	-592.8274	**0.004**	-585.1931	**0.004**
uterine irr [Yes]	-516.0810	**<0.001**	-525.5239	**<0.001**
physician visits	-14.0581	0.763		
Observations	189		189	
R^2 / R^2 adjusted	0.243 / 0.205		0.240 / 0.215	
AIC	2996.567		2991.153	

In-sample performance: full model vs reduced model

Appendix 5



Out-of-Sample performance using CV: full model vs reduced model