The Study of the Determinants of Birth Weight

Siqi Wang

December 2014

1 Introduction

There have been numerous studies that have shown links between birth weight and later-life conditions, including diabetes, obesity, tobacco smoking and intelligence. In this paper, we are going to study the determinants of birth weight, in order to provide suggestions to possibly reduce the risk of the diseases that caused by low birth weight.

Section 2 introduces the source and elements of the data set. Section 3 gives the univariable analysis of variables in the data set. In Section 4 the simple linear regression model will be discussed. We will bring out and interprete our final model in Section 5. Section 6 makes the conclusion.

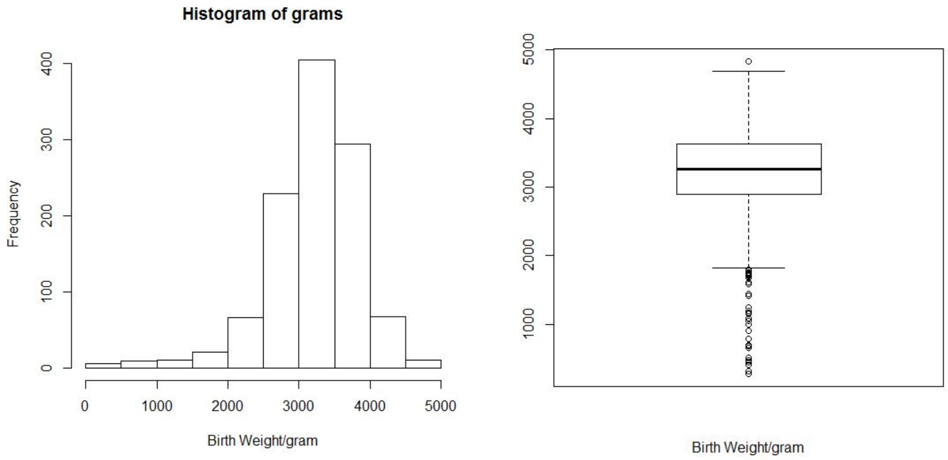
2 Data Set

Our data is a random 5% sample of all births occurring in Philadelphia in 1990 and it is from Elo, Rodríguez and Lee's paper [**?**]. The sample has 1115 observations on five variables, which are *gestate* (gestational age in weeks), *educ* (Mother's year of education), *black* (dummy variable, it is TRUE if mother is black; it is FALSE if mother is not black), *smoke* (dummy variable, it is TRUE if mother smokes during pregnancy, it is FALSE if mother does not smoke during pregnancy) and *grams* (birth weight in grams).

3 Univariable Analysis

3.1 Birth Weight (*grams*)

Here we look at the variable *grams* first.



(a) (b)

Figure 1: (a) Histgram of *grams*; (b) Boxplot of *grams*

The range of *grams* is from 284 to 4830, with median as 3267 and mean as 3220. From Figure 1(a), we can see that the distribution is skewed to the left, which indicates that the transformation is required. The Figure 1(b) shows that there are some outliers in the graph. The minimium value is 284 grams. According to the Guinness World Records [**?**], the lightest birth in the world is 260 grams. Thus the minimium value is a valid birth weight. We will remain this data point in our analysis.

3.2 Gestational Age (*gestate*)

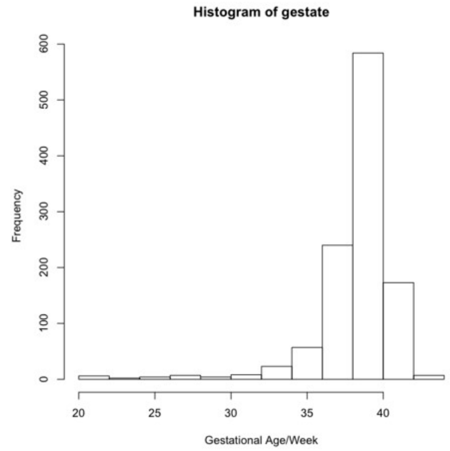


Figure 2: Histogram of gestate

From Figure 2 we observe that the distribution is highly skewed to the left, which suggests possible transformation. The range of the data is from 20 to 43, with median as 39 and mean as 38.84. According to our research [?], a birth occured before the start of the 37th week is defined as premature birth. Thus our data contains premature birth cases. By this definition, 111 data points are defined as premature birth records.

3.3 Years of Education (*educ*)

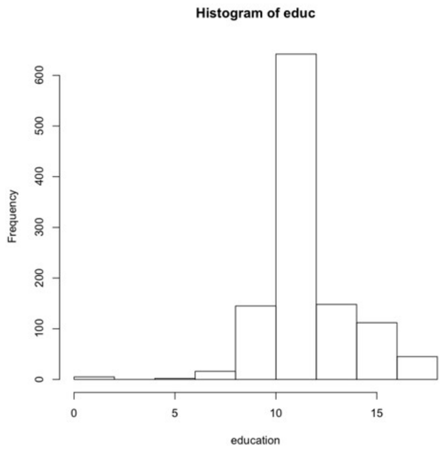


Figure 3: Histogram of education

Figure 3 tells us that the data has an asymmetric spread. The range of the data is from 0 to 17, with median as 12 and mean as 12.27. Since 0 is contained in the range, it should be careful when doing the transformation.

3.4 Dummy Variables

	True	False
Black	662	453
Smoke	269	846

Table 1: Summary of the dummy variables

Table 1 shows the count of each dummy variable.

We can visualize the relationship between variables from Figure 4, the scatter matrix plot.

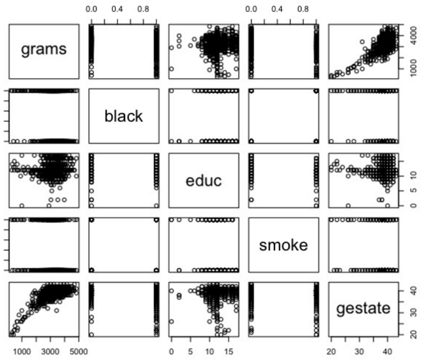


Figure 4: Scatter Plot Matrix

4 Diagnostics of Simple Linear Model

Now we will introduce the simple linear model as the baseline model that uses *gestate*, *educ*, *black* and *smoke* to predict *grams*. The model's formula is

$$grams = 156.512 \times gestate + 9.572 \times educ - 168.968 \times black - 174.813 \times smoke - 2834.486 \quad (1)$$

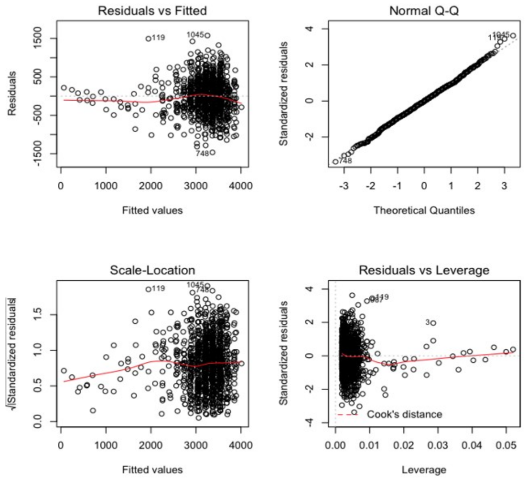


Figure 5: Diagnostic plot of the baseline model

We see that the residual plot has no discernible pattern, which indicate that the linearity assumption is satisfied. The standardized residual plot has a fairly horizontal fitted line which indicates that the model obeys the constant variance assumption. The Normal Q-Q plot has a straight fitted line which means the normality assumption is satisfied. According to Table 2, the variance inflation factors of four predictors are all between 1 and 1.1, so we conclude that multicollinearity assumption is satisfied. Therefore, the simple linear model is valid.

gestate	educ	black	smoke
1.0513	1.0741	1.0509	1.0733

Table 2: Variance Inflation Factors of the baseline model

The coefficients of *gestate*, *black* and *smoke* are significant at 5% significance level, while the coefficient of *educ* is insignificant. Thus, we would need to consider transformation or deleting *educ* in the model selection process. Overall, as a valid model with R^2 to be 0.53, the simple linear model is considered to be acceptable.

5 Final Model

5.1 Model Description

After testing several models through processes such as power transformation, Weighted Least Square, and Akaike' Information Criterion/Bayesian Information Criterion, we determined our final model as follow,

$$\log(grams) = 0.2426 - 0.0041 \times gestate^2 + 0.3604 \times gestate - 0.0522 \times black - 0.0576 \times smoke \quad (2)$$

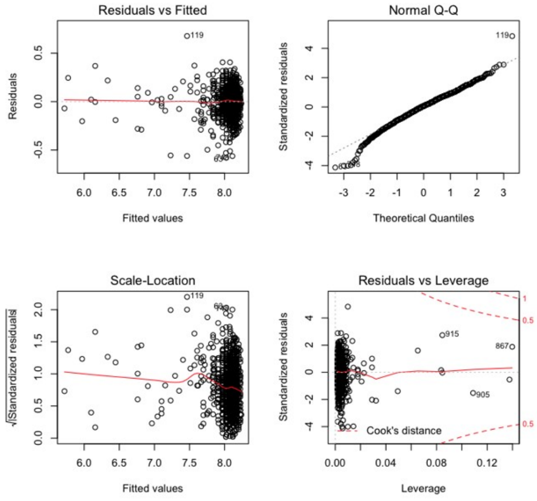
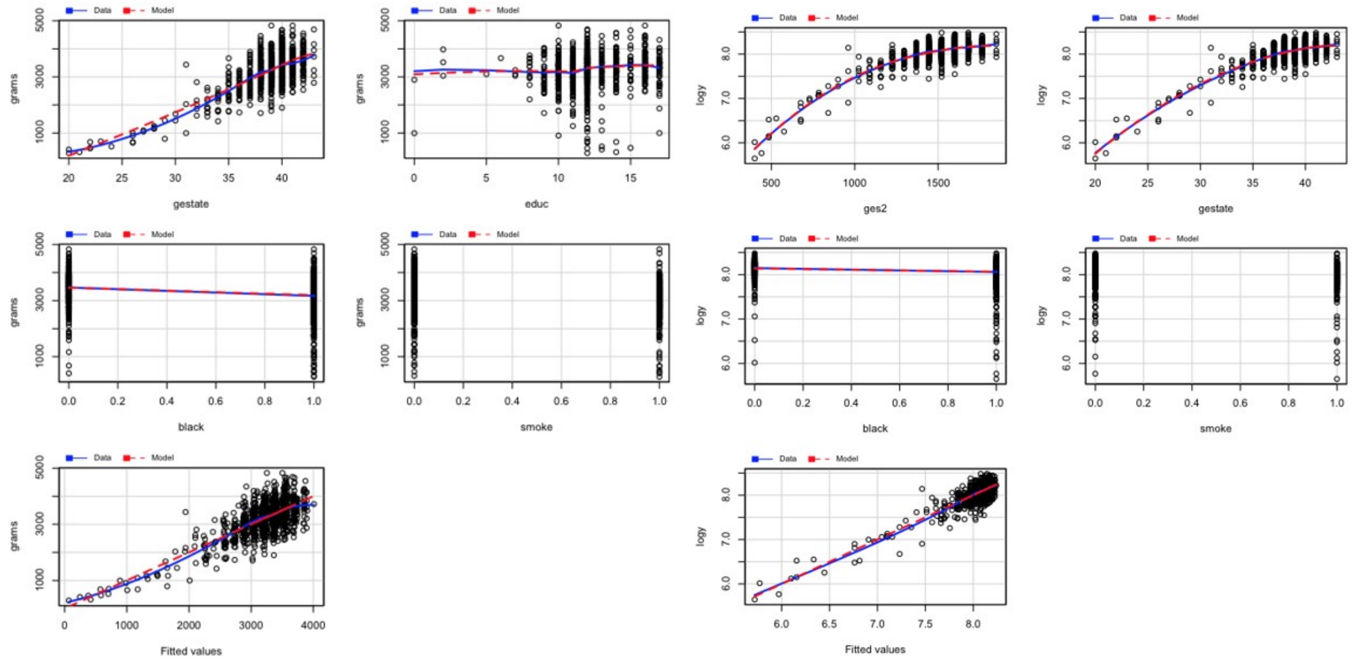


Figure 6: Diagnostic plot of the final model

In the final model we delete the variable *educ* and add the quadratic term $gestate^2$. Figure 5 tells us that, similar to our discussion in the previous section, the final model does not violate the

$gestate^2$	$gestate$	$black$	$smoke$
99.1863	98.8865	1.0399	1.0226

Table 3: Variance Inflation Factor of the final model



(a)

(b)

Figure 7: (a) MMPs of the Baseline model; (b) MMPs of the Final model

linearity, normality and constant variance assumptions. From Table 3, the VIF of $gestate^2$ and $gestate$ are high, that is because $gestate^2$ is derived from $gestate$. However, other variables' VIF are much lower than the cut-off value 5, so the multicollinearity assumption is satisfied and the model is valid.

All variable coefficients are statistically significant at 5% significance level and the adjusted R^2 is 0.736 which is pretty well. And if we look at the comparison of the marginal model plots of the baseline model and the final model in Figure 6, we can find that the final model fits the data better.

We are aware that the relationship between predictors with the premature birth weight may be different with predictors' relationship with full term birth weight. Thus we tried to introduce a dummy variable *preterm* which indicates *very preterm* or extremer cases. Mayo Clinic [**?**] defines that *very preterm* means born at less than 32 weeks of pregnancy. We only indicating severe preterm case because we want to see a stronger effects on birth weight due to gestational age. However, the model after adding this dummy variable does not change the coefficients of exsiting variables dramatically (thus the model includes $log(grams)$, $gestate^2$, *gestate*, *smoke*, *black* and *preterm*). All of the coefficients of original predictors change less than 0.05. And the coefficient of *preterm* is -0.04 which means it would causes about 4% decreases of birth weight if the baby borns in less than 32 weeks. The partial F-test between this model and the model which also adds interaction term on *preterm* has p-value 0.33 which suggest us not to use interaction term. These results reflects that premature birth would not have different effect (or more accurately, dramatic different effects) to the relationship between birth weight and gestational age. In Gage, Fang and Stratton's work [?] they analyze the relationship between birth weight and premature gestational age and full term gestational age together which supports our thought.

5.2 Interpretation of the Final Model

According to equation (2), we can interprete our final model as

- Holding everything else constant, the birthweight with $(x + 1)$ gestational age has a $(35.63 - 0.82x)\%$ difference with the the birth weight with x gestational age.
- If the mother is black, the baby's birth weight will be approximately 5.22% less.
- If the mother smokes, the baby's birth weight will be approximately 5.76% less.

Since the final model consists a polynomial term of *gestate*, the change rate of birth weight due to gestational age is not constant.

5.3 Leverage Points

Given the sample size as 1115 and number of predictors as 4, according to the rule of determining leverage poingts introduced by the book *A Modern Approach to Regression with R*, we set the cut off level of hat values to be 0.008 that points with hat values larger than 0.008 are defined to be leverage points. With the range of standardized residuals to be -4 to 4, the points whose standardized residuals fall out of the range are identified as outliers. As a result, three points are considered to be outliers, which are 119, 590 and 1084 (these are also leverage points by the definition). Those points have both small gestational age and low birth weight. Since our model haven shown that the birth weight is possitively correlated with gestitional age, those three points may carry useful information for building our model. As a result, we decide to keep the outliers in our final model.

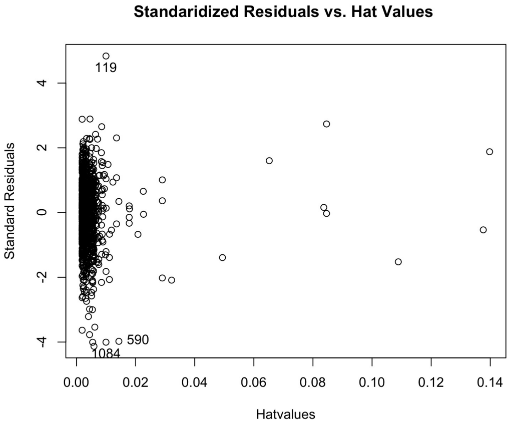


Figure 8: Leverage Points of the Final Model

5.4 Limitations & Possible Improvements

There is no perfect model. Although our final model seems better than other models, it still has its own shortcomings:

- According to our research, mother's age and parents weight are important factors to the birth weight. Unfortunately our dataset does not include the information of these two factors.
- The dataset does not record the survival information of these data. As we analyzed previously, most of the outliers are premature birth weight. If we can know wether these baby are survived, we can give a better solution to the outlier.
- The dataset is collected in 1990 which is almost 25 years ago. Demographic information and life standard has had great changes. Thus it might not be accurate to use this model to predict birth weight nowadays.

If we can obtain a more updated dataset which includes more important factors and information of the baby, we may build a better model to do the prediction.

6 Takeaways

Besides proving that the birth weight is positively related with the gestational age (assume the largest gestational age is less than 45 weeks), the results demonstrate that smoking during pregnancy may lead to a lower birth weight. Therefore, according to the diseases that may be caused by low birth weight, pregnant women should be advised to abandon smoking during pregnancy.

References

[1] I.T.Elo, G.Rodríguez and H.Lee. *Racial and Neighborhood Disparities in Birthweight in Philadelphia.* Annual Meeting of th Population Association of America, Washington D.C., 2001.

[2] Gage, Timothy B., Fu Fang, and Howard Stratton. *Modeling the Pediatric Paradox: Birth Weight by Gestational Age.* Biodemography and social biology 54.1 (2008): 95–112.

[3] http://www.mayoclinic.org/diseases-conditions/premature-birth/basics/definition/con-20020050. *Premature Birth.* Mayo Clinic.

[4] http://www.guinnessworldrecords.com/world-records/lightest-birth/. *Lightest Birth.* Guinness World Records, 2005.