Modeling Birth Weight of Newborns and Associated Factors

Joanna Chen *School of Public Health, Yale University*

Importance: Low birth weight is a severe issue related to newborns' mortality and morbidity.

Objective: To determine the factors associated with birth weight and to model the birth weight.

Evidence: The dataset contains 4342 babies, 51.4% male, 48.6% female. I used 10 models including two variations of multiple linear regression, Lasso regression, Ridge regression, two variations of polynomial regression, stepwise regression and three interaction models. I selected the optimal one using 5-fold cross validation.

Findings: Results show that the polynomial model performs the best (RMSE:256.49 after removing outliers). The model estimates show that the largest factor to birth weight is the number of gestational weeks, followed by mother's weight gain during the pregnancy. They have a nonlinear positive effect on the birth weight. Also, mother's height has quite a large impact on the birth weight.

Conclusion and Relevance: This study shows that birth weight can be modeled by several factors during pregnancy, including gestational weeks, mother's weight gain, mother's height.

Introduction

Low birth weight is a severe issues related to newborns' mortality and morbidity. It refers to babies who weigh less than 2500 grams at birth. Low birth weight babies may have problems with their organs and future developmental delays(Jin, 2015). It may help with prevention of low birth weight by determining the factors associated with birth weight. In this study, the objective was to determine the factors contributed to newborns' birth weight and build a regression model to find relationships between the factors and children's birth weight as well as predict future birth weights.

Methodology

Data Preparation

The definition of each variable is shown in Table 1. The initial dataset includes 20 variables which are characteristics from the infant, father and mother with 4342 observations. After importing the data, I convert *babysex*, *frace*, *malform* and *mrace* from integers to factors because they are categorical data. There is no missing data in the dataset. I drop variables *pnumlbw* and *pnumsga* as all values are 0 so they give us no information, drop 'parity' because only 10 non-zero observations and the rest are zero, drop *delwt* and *ppwt* as the difference between the two is already implied in *wtgain*. I delete one observation where *menarche* is equal to 0. There are some negative *wtgain* observations. I do not delete them because there are women who are overweight may lose weight in early pregnancy due to stored fat being used to power pregnancy growth. Figure 1 shows that negative *wtgain* is highly correlated with high pre-pregnancy weight. Also, a lot of mothers eat healthier foods and do more exercise when they become pregnant, so the healthier lifestyle could lead to weight loss (Ludwig, 2010). The final data after cleaning has 4341 observations. The characteristics were summarized in Table 2. Table 2 shows that the male and female have approximately the same proportion. The largest proportion of the parents' race is White (about half of the whole population), followed by Black (44%). Asian has the smallest proportion (1%) in the study.

I used R Studio version 3.6.1 for the data management and analysis. Exploratory data analysis was performed on the dataset. In this step, I generate the summary values of each variable. The response variables is birth weight. A normality check on the response variables with histogram is performed. In addition, I check the correlation between each variables and select variance inflation factor (VIF) until all predictors have VIF less than 10 for multicollinearity checking purpose. Finally, I checked for the missing value and zero values in each column.

Models

I then explore different combinations of variables and apply them to 10 different models to compare, analyze, and determine a relatively accurate one. The models include multiple linear regression, Lasso regression, Ridge regression, polynomial regression, stepwise regression and interaction models. Multiple linear regression (MLR) describes the response variables by linearly combining the explanatory variables using some coefficients. I regress a full model first, remove the individual term with highest p-values and refit the model with only significant terms. I continue the process until all terms remain with p-values below the threshold of 0.1. This is because low p-values imply that changes in the predictor are related to changes in the response variable which shows the variable is statistically significant and can be a meaningful addition to the model. I compare the fit of the full model against the new, reduced model to determine which one is better by comparing adjusted R-square. I also plot the model residuals against fitted values to see whether the model fit the data well or not.

Then, I apply polynomial models. I come up with a set of possible polynomial models, successively add the polynomial terms into the models models in increasing order, fit to the data, and look at the significance of regression coefficients as well as adjusted R-square. I keep the order of polynomial increasing until the term with highest order is non-significant. Moreover, when the adjusted R-squares are similar, I choose the most parsimonious one rather than the best-fitting one in order to prevent the over-fitting and put it in later cross validation for further selection.

As for the Lasso regression and Ridge regression, I code the category variables as binary first, then use *cv.glmnet()* to find the value of best tuning parameter λ and use it to fit the model. I then perform stepwise regression of both direction and choose the model with the smallest Akaike information criterion (AIC). I determine the interaction terms to include in a multiple regression model based on the common sense that *babysex*, *bhead*, *blength* should have interaction effect on each other. I combine different combinations of these terms into interaction terms and leave the rest as main effects. Moreover, I drop the non-significant terms and use the fewest terms as I can to prevent over-fitting.

A summary of these selected models are shown in Table 3. They will be put in the cross validation later.

Validation Strategy

I apply 5-fold cross validation to validate these models. Specifically, I split the whole dataset randomly into 5 partitions. I use four of them as the training set and one as the test set. Using the model derived from the training set, I predict birth weight and compare the predicted value with the real value from the test data. The root mean squared error (RMSE) is calculated as the prediction error. This procedure is repeated 100 times and mean and variances of this prediction error are calculated. Given these prediction errors over different models, I choose the model with the lowest RMSE as my final model.

After the model is selected, I fit the model with the data excluding the leverage point, outliers and influential points to see whether there is any change of the fit. The leverage point is determined using typical rules of thumb $\frac{3(p+1)}{n}$, where n is the number of observations and p the number of predictor variables. The outlier is determined by points exceeding 3 standard deviations and the influential values is determined by visual inspection of plots of Cook's distance.

Results

Comparison among different models

The initial cross validation results are shown in Table 4. It shows that *lasso*, *ridge*, *poly1*, *poly2*, *stepwise* all have relatively small values compared to the other models. Among these models, *poly1* has the smallest error (RMSE: 273.74). These results suggest that the relationship between birth weight and the other explanatory variables are more likely to be nonlinear. The cross validation results after removing the infuential points are shown in Table 5, from which we see that all the models improve their performance. However, the accuracy comparison stays the same with *poly1* (RMSE: 256.49) still performing the best.

Interpreting estimated model

The optimal model selects two variables that are closely related to birth weight, mother's weight gain during pregnancy (*wtgain*) and gestational age in weeks (*gaweeks*). The estimated parameters are reflected in the following relationship. The contributions of *gaweeks* and *wtgain* are shown in Figure 2 and Figure 3 respectively. The details of the model are included in Table 6.
Equation 1. *poly1* before removing influential points

$$\begin{aligned} bwt = {} & 28 \times babysex2 + 130 \times bhead + 74 \times blength + 0.3 \times fincome + 7 \times malform1 \\ & - 3 \times menarche + 12 \times mheight + 0.8 \times momage - 135 \times mrace2 - 76 \times mrace3 \\ & - 99 \times mrace4 + 8 \times ppbmi - 5 \times smoken + 4 \times wtgain + 0.07 \times wtgain^2 \\ & - 0.001 \times wtgain^3 - 337 \times gaweeks + 10 \times gaweeks^2 - 0.09 \times gaweeks^3 - 2219 \end{aligned}$$

When we interpret the model variables, each variable interpretation is under the assumption that other variables remain constant. We are not going to repeat this in each individual variable interpretation.
The positive factors that contribute to birth weight are *wtgain*, *gaweeks*, *bhead*,*blength*,*babysex*,*fincome*, *mheight*,*ppbmi*, *malform*,*momage* which intuitively have a positive effect on the birth weight. The female baby has more birth weight than the male baby, which may require further study. On the contrary, negative factors include *menarche* and *smoken*. The heaviest babies were found to be when the race of the mother is White, followed by Asian, Puerto Rican, and Black babies tend to be the lightest.

Model checking

After fitting the final model, influential points are removed and the model is fit again. The reason I don't remove outliers and high leverage points is that there are 64 points with high leverage and 14 with large residuals. I do not want to lose potential valuable information. I get the following result.

Equation 2. *poly1* after removing influential points.

$$\begin{aligned} bwt = {} & 30 \times babysex2 + 127 \times bhead + 80 \times blength + 0.4 \times fincome + 11 \times malform1 \\ & - 3 \times menarche + 12 \times mheight + 0.8 \times momage - 131 \times mrace2 - 73 \times mrace3 \\ & - 95 \times mrace4 + 8 \times ppbmi - 5 \times smoken + 4 \times wtgain + 0.07 \times wtgain^2 \\ & - 0.001 \times wtgain^3 - 314 \times gaweeks + 9 \times gaweeks^2 - 0.08 \times gaweeks^3 - 2493 \end{aligned}$$

We can see from the second equation that the coefficients are similar. We can extend the interpretation of the previous model equation to this one.

Discussion

I tried 6 different types of models. It seems that polynomial models work best and *wtgain* and *gaweeks* are the two most predictive variables. Hence, it suggests a nonlinear relationship among them (Figure 2 and Figure 3).

The results are also illuminating. From the equation, I see that birth weight, *bwt*, is affected by many different values. The largest contribution comes from *gaweeks*, which makes sense since the longer the pregnancy, the more mature the baby. Also, *wtgain*, which is the mother's weight gain, has a positive effect on the birth weight. This is expected, as the mother's weight gain can have a direct correlation to the baby's weight. Similar findings between weight gain and birth weight have been reported in [2]. On a similar note, the income of the father, *fincome*, has a slightly positive correlation, which could be due to higher income families can afford better nutrition. Other expected positive correlations are the size of the baby's head, the length of the baby, the pre-pregnancy BMI of the mother, the height of the mother, malformities. The length of the baby, *blength* and the head size of the baby, *bhead* are expected, as a larger head and a longer baby will have a greater weight. The same goes for the height of the mother, *mheight*, and the pre-pregnancy BMI, *ppbmi*, as a mother that is taller or heavier will intuitively have a larger baby. The same goes for a malformity, as a greater weight can be attributed to macrosomia. Findings of positive relationship between mother's age at delivery *momage* and weight gain in our data have also been reported in [3].

A few expected attributes that negatively affect the birth weight include the mother's age at menarche *(menarche)* and the number of cigarettes the mother smoked during pregnancy *(smoken)*. There are a few attributes that may require further exploration. For example, a female baby *(babysex2)*, is on average 28 times heavier than a male baby, if all other variables remains constant. However, I am unsure if there are further correlations between a female sexed baby and other attributes. Interestingly enough, the mother's age at menarche have negative effects on the weight of the baby and it may also require further exploration.

The RMSEs of CV for all the methods are in the range of 200 - 300 grams. More complicated models may help reduce this prediction error. Especially, nonlinear models, e.g. kernels, may be used in regression to improve the error.

There are some limitations in the study. For example, the distribution of race in the data seems to be imbalanced. This can be a cause of a misleading result and might have caused bias. Hence, including race as a major risk factor might not be adequate and need further study with bigger sample size and balance among race groups.

Conclusion

In this project, I applied 10 models on a baby's birth weight dataset of 4342 subjects. I used 5-fold cross validation with 100 random realizations to validate the prediction performance on birth weight given other explanatory variables. The results show that a 3rd order polynomial model on *gaweeks* and *wtgain* has the best prediction error (RMSE:256.49 after removing outliers). These results show that baby's birth weight can be modeled by several factors during pregancy, among which gestational weeks, mother's weight gain, mother's height all have quite positive effects on the birth weight. On the contrary, smoking, late menarche age have a slightly negative effect on the birth weight.

Acknowledgement

I would like to thank Professor Yize Zhao's for her guidance and help on my linear model's studying, as well as the TAs Wenlan Zang, Yiliang Zhang and Jinghao Sun, so I was able to learn and apply the knowledge on this project.

Appendices

Table 1. The definition of variables

Variable Name	Definition	Unit	Range	Median
babysex	baby's sex		{0, 1}	
bhead	baby's head circumference at birth	centimeter	[21, 41]	34.00
blength	baby's length at birth	centimeter	[20, 63]	50.00
bwt	baby's birth weight	gram	[595, 4791]	3147
delwt	mother's weight at delivery			
fincome	family monthly income	hundreds	[0, 96]	35.00
frace	father's race		{1,. . . ,4,8,9}	
gaweeks	gestational age in weeks	week	[17.7, 51.3]	39.90
malform	presence of malformations that could affect weight		{0, 1}	
menarche	mother's age at menarche	year	[5, 19]	12.00
mheight	mother's height	inch	[48, 77]	63.00
momage	mother's age at delivery	year	[12, 44]	20.00
mrace	mother's race		{1,. . . ,4,8}	
parity	number of live births prior to this pregnancy			
pnumlbw	previous number of low birth weight babies			
pnumgsa	number of prior small for gestational age babies			
ppbmi	mother's pre-pregnancy BMI		[13.1, 46.1]	21.03
ppwt	mother's pre-pregnancy weight			
smoken	average number of cigarettes smoked per day during pregnancy		[0, 60]	0.00
wtgain	mother's weight gain during pregnancy	pound	[-46, 89]	22.00

Table 2. Summary of predictor characteristics

	N	**Percent**
Baby's Sex		
Male	2230	51.4%
Female	2111	48.6%
Father's Race		
White	2123	48.9%
Black	1910	44.0%
Puerto Rican	248	5.7%
Asian	46	1.1%
Other	14	0.3%
Malformations		
Absent	4326	99.7%
Present	15	0.3%
Mother's Race		
White	2147	49.5%
Black	1908	44.0%
Puerto Rican	243	5.6%
Asian	43	1.0%

Table 3. Models used in the cross validation

model name	Model
mlr1	bwt ~.-mrace
mlr2	bwt ~.-mrace-malform
lasso	bwt ~.
ridge	bwt ~.
poly1	bwt ~ . + $wtgain^2 + wtgain^3 + gaweeks^2 + gaweeks^3$
poly2	bwt ~ . + $wtgain^2 + wtgain^3 + menarche^2 + menarche^3$
stepwise	bwt ~.

model name	Model
inter	bwt ~ babysex * bhead +blength * bhead+ gaweeks + mheight + ppbmi + smoken + wtgain
inter2	bwt ~ babysex * bhead +blength * bhead+ babysex * blength + gaweeks + mheight + ppbmi + smoken + wtgain
inter3	bwt ~ blength*bhead+ gaweeks + mheight + ppbmi + smoken + wtgain

Table 4. RMSE of all the methods before removing influential points

	mlr1	mlr2	lasso	ridge	poly1	poly2	stepwise	inter	inter2	inter3
RMSE	279.95	279.87	274.39	274.66	273.74	273.79	273.83	281.89	283.04	282.25

Table 5. RMSE of all the methods after removing influential points

	mlr1	mlr2	lasso	ridge	poly1	poly2	stepwise	inter	inter2	inter3
RMSE	272.52	272.40	267.89	268.23	256.49	267.36	267.27	275.24	275.30	275.59

Table 6. Coefficients of the final model chosen

	Estimate	Std. Error	t value	P value
Intercept	-2.219e+03	8.665e+02	-2.561	0.010461 *
babysex2	2.784e+01	8.443e+00	3.298	0.000983 ***
bhead	1.297e+02	3.452e+00	37.571	< 2e-16 ***
blength	7.399e+01	2.021e+00	36.604	< 2e-16 ***
fincome	2.745e-01	1.785e-01	1.538	0.124165
gaweeks	-3.366e+02	7.264e+01	-4.634	3.69e-06 ***
malform1	6.667e+00	7.047e+01	0.095	0.924637
menarche	-3.362e+00	2.910e+00	-1.155	0.248005
mheight	1.246e+01	1.653e+00	7.535	5.93e-14 ***
momage	8.308e-01	1.216e+00	0.683	0.494521

	Estimate	Std. Error	t value	P value
mrace2	-1.349e+02	1.021e+01	-13.210	< 2e-16 ***
mrace3	-7.558e+01	4.261e+01	-1.774	0.076157 .
mrace4	-9.859e+01	1.935e+01	-5.096	3.62e-07 ***
ppbmi	8.300e+00	1.407e+00	5.899	3.93e-09 ***
smoken	-4.916e+00	5.844e-01	-8.413	< 2e-16 ***
wtgain	3.807e+00	1.104e+00	3.447	0.000572 ***
wtgain2	6.859e-02	3.826e-02	1.793	0.073093 .
wtgain3	-1.225e-03	4.359e-04	-2.810	0.004973 **
gaweeks2	9.519e+00	2.056e+00	4.629	3.78e-06 ***
gaweeks3	-8.519e-02	1.914e-02	-4.451	8.76e-06 ***

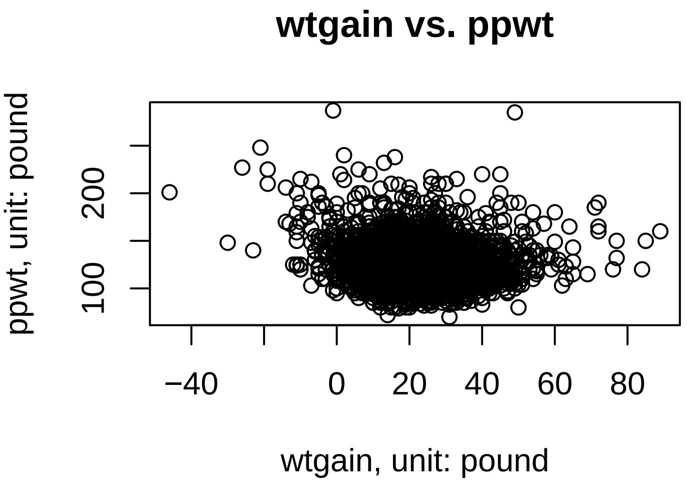

Figure 1: wtgain vs. ppwt

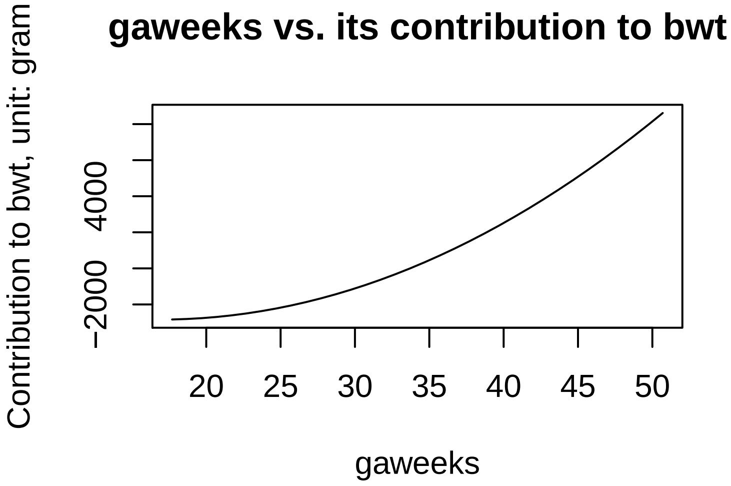

Figure 2: gaweeks vs. its contribution to bwt

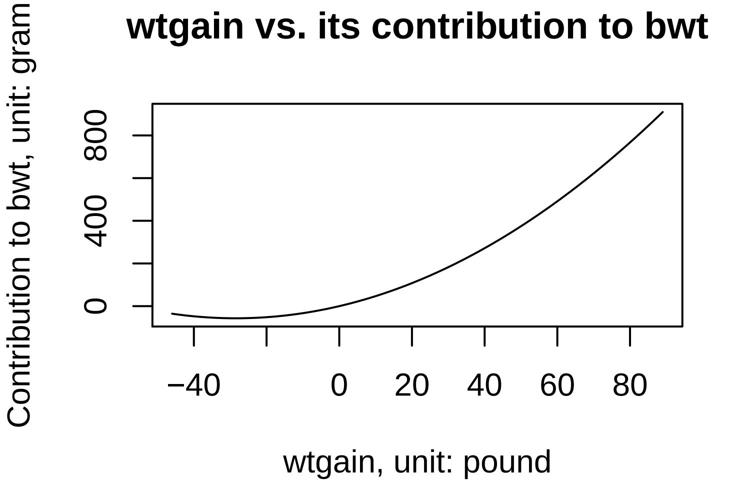

Figure 3: wtgain vs. its contribution to bwt

References

[1] Jin, Jill. 2015. “Babies With Low Birth Weight”, JAMA.

[2] Ludwig, DS, Currie, J. 2010. “The Association between Pregnancy Weight Gain and Birthweight: A Within-family Comparison”, The Lancet, Elsevier.

[3] Restrepo-Méndez, MC et al. 2015. “The Association of Maternal Age with Birthweight and Gestational Age: A Cross-Cohort Comparison”, Paediatr Perinat Epidemiol.