Multiple Regression Analysis on Birthweight (Executive Summary)

Taoran Bu, Yiren Cao, Mingze Gong, Rajat Lal, Yu Qiu[a]

[a]M13_early_3, USYD

This version was compiled on November 18, 2020

Continuing with our presentation, we summarise all critical findings into this report and dig deep into the interaction effect between certain variables to find a more appropriate model.

Data description.

Data source. The data was collected from Baystate Medical Center, Springfield, Massachusetts in 1986. The dataset is accessible in Rstudio via the package 'MASS' under the name 'birthwt'.

Background research. The purpose of this statistical report is to bring light to what particular risk factors help predict infant birth weight. As shown below are ranges of expected birth weights of babies, provided by University of Michigan. These ranges can be potential indicators to further understand results which show correlation between risk factors and babies born with weights outside of the expected range, suggesting a potential negative external influence on birth weight.

The average birth weight for babies is around 7.5 lb (3.5 kg), although between 5.5 lb (2.5 kg) and 10 lb (4.5 kg) is considered normal. In general:

- Boys are usually a little heavier than girls.
- First babies are usually lighter than later siblings.
- Large parents generally have large babies, while small parents generally have small babies."

According to Victorian Government organization Better Health, on average the mother's who smoked during pregnancy were on average delivering babies weighing 150 to 200 grams less than normal. Through our multiple regression models we would like to find out if smoking is in fact a predicting factor for birth weight of babies, in particular does it have a negative coefficient.

Data discovery. The variables in the dataset which we will be testing against birth weight of the child are:

- Numerical variables:
  - mother's age (age)
  - mother's weight (lwt)
  - number of premature labors (ptl)
  - number of physician visits (ftv)
  - birth weight of the child (bwt)
- Categorical variables:
  - low (birth weight below or above 2.5kg)
  - race of mother (race)
  - whether mother smokes (smoke)
  - mother's hypertension history (ht)
  - mother's presence of uterine irretability (ui)

Pre Analysis Observations. Mean birth weight of babies (bwt) is approximately 2.9kg, which is 600 grams below the expected mean of 3.5kg. This is within the normal range of weights of 2.5kg - 4.5kg.

Mother's weight had a mean of 58.8 kg compared to the expected mean of 74.3 kg in 1999–2000. This may be an influencing factor for the lower range of birth weights of the babies.

Questions.

1. If smoking is in fact a predicting factor for birth weight of babies, in particular does it have a negative coefficient.
2. Does mother's birth weight in the last menstrual period predict baby birth weight?
3. What are all the significant predicting factors of baby birth weight?

Data Analysis & Model Generate. With the data without 'low' variable, we first performed a stepwise variable selection from the full model. With the dataset, we perform a F-test every time to remove the least informative variable, using the AIC value of each model and F test result.

	AIC	F value
	2458.2	
age	2456.3	0.1377
lwt	2462.7	6.2934
race	2470.0	7.7776
smoke	2467.4	10.9317
ptl	2456.4	0.2253
ht	2465.1	8.5857
ui	2470.2	13.8077
ftv	2456.3	0.0915

		AIC	F value
		2456.3	
age	1	2454.5	0.1880
lwt	1	2460.7	6.2360
race	2	2468.0	7.7861
smoke	1	2465.5	10.9473
ptl	1	2454.5	0.2176
ht	1	2463.1	8.5375
ui	1	2468.3	13.8266

	AIC	F value
	2454.5	
lwt	2458.7	6.0776
race	2466.2	7.8440
smoke	2463.5	10.8142
ptl	2452.8	0.2802
ht	2461.1	8.4665
ui	2466.3	13.7292

	AIC	F value
	2452.8	
lwt	2457.3	6.4093
race	2464.6	7.9452
smoke	2462.7	11.8652
ht	2459.5	8.5918
ui	2466.0	15.2268

Fig. 1. Stepwise Backward selection for data without low

As it shown above, variable 'ftv'(physical visit number) is removed , followed by 'age'(mother's age), and then 'ptl'(number of previous premature labors) in the end.

A forward selection as well as an exhaust search is performed to compare with the first one.

All of these selections give us the same 4 predictors (races are counted as one predictor).

By making a simple comparison with the models, and surprisingly both selections provided us with exactly the same model. Hence, this is the model we selected for this dataset.

In the end, after we obtain our model from forward/backward selection, a 10-fold cross validation with roughly 18 samples each fold is performed to see the strength of the selected model. Detail of the model validation is shown in the figure below.

	intercept	RMSE	Rsquared	MAE
1	TRUE	646.66	0.28	527.01

Table 1. Output performance check-10 Fold validation

Analysis. When we look into our dataset, we find that the predictor `low`, which indicates whether the birth weight is less than 2.5 kg and dependent variable `bwt`, the birth weight in grams, are likely to be measured in the same way, hence we exclude the `low` variable before building up our model.

First, we conduct preliminary assumption checking. In p1, the general correlation between each predictor and birthweight does not show obvious linear or non-linear pattern, which requires further investigation.

Second, we find approximately a straight line on the residual plot p2. Points are randomly distributed above and below across the line. So it satisfies the linearity assumption. We also find that the spread on the residual plot looks reasonably constant over the range, which indicates that it satisfies the homoscedasticity assumption.

Third, on the QQ plot p3, despite several departure in the upper and lower tail, the majority of the points lie quite close to the diagonal line. So normality is satisfied. In addition, since we have a fairly large sample size, we can rely on central limit theorem.

Because we don't see any violation of independence but we don't know exactly how the experiment was designed. Based on the current information we know, we assume it satisfies the assumption of independence between the errors.

We build up our first multiple regression model. To begin with, we performed a stepwise backward variable selection starting from the full model, where we remove the least informative variable step by step and conclude the final model f1 at the end. The forward and backward AIC as well as an exhaust search are also performed and they all lead to the model with the same 5 predictors as before. After building up the model, we perform the assumption re-check on it. On p4, we still find relatively a straight line on the residual plot and points are randomly distributed above and below all over the line. So the linearity is satisfied. The independence, homoscedasticity, normality assumptions also hold for the same reasons mentioned above. We find that GVIF for all predictors are smaller than 5, so it satisfies no multicollinearity assumption.

We do not include the interaction between our predictors in the current model and by conducting several trials, we find that the interaction between `lwt` and `ht` might be useful to predict the birth weight. Hence we include the interaction and build up our second model f2 using AIC forward and backward methods that give us the same result. Following the same procedure of assumption checking before f5, we conclude that the second model satisfied all assumptions.

Results. Finally, we get two models Equation1 and Equation2 at our disposal:

$$\begin{aligned}\text{bwt}_1 &= 2362.21 - 525.52(\text{ui}_{\text{yes}}) + 126.91(\text{race}_{\text{other}}) \\ &\quad + 475.06(\text{race}_{\text{white}}) - 356.32(\text{smoke}_{\text{yes}}) \\ &\quad - 585.19(\text{ht}_{\text{yes}}) + 4.24(\text{lwt}) + \epsilon\end{aligned} \tag{1}$$

$$\begin{aligned}\text{bwt}_2 &= 2509.39 - 536.56(\text{ui}_{\text{yes}}) + 140.13(\text{race}_{\text{other}}) \\ &\quad + 489.75(\text{race}_{\text{white}}) - 379.65(\text{smoke}_{\text{yes}}) \\ &\quad - 1791.37(\text{ht}_{\text{yes}}) + 3.08(\text{lwt}) \\ &\quad + 7.88(\text{ht}_{\text{yes}} \times \text{lwt}) + \epsilon\end{aligned} \tag{2}$$

As Allison (1977) suggested, if two variables were measured on a "numerical scale", it is quite common to test the "presence of interaction" through the product of these two variables. The difference of our two models above exactly derives from the inclusion of interaction between **lwt** and **ht**, which is actually one of our most important findings when conducting the research.

With regard to questions aforementioned, we found that both smoke (**smoke**) and mother's weight at last menstrual period (**lwt**) indeed exist in our two models, while the former imposes a negative effect on a baby's weight and the other way around for the latter. This is quite reasonable especially when we refer to some seminal research (??reference).

	RMSE	Rsquared	MAE
bwt1	657.29	0.21	534.17
bwt2	644.94	0.22	528.31

Table 2. bwt1 and bwt2

Table2 presents the result of out of sample performance test. Table4 reports results of multiple regression and in sample performance test of these two models.

Discussion. Firstly, the sightly higher overall **adjusted** R^2 from Table4 implies that with a brand-new variable (**interaction** between **ht** and **lwt**) added, the ability to explain more variation from the variables may seemingly get improved although with a tiny increase.

Also, the interaction is significant at 90%, the same level as lwt itself which was kept by three model selection methods that we utilized before (Stepwise, AIC backward & forward). Since lwt has been included in the model, there is no decent reason to drop the interaction between **lwt** and **ht**.

Lastly, similar to in sample performance test, out of sample performance test also favors the model with interaction involved due to the lower **RMSE** and **MAE**.

Conclusion. The low R^2 in both models significantly indicates that either of the model we constructed before can pleasantly explain much variation brought by these models. We can definitely, however, choose a relatively good model, **bwt2** Equation**??**, based on the discussion above and results shown in the tables.

Nonetheless, we have to admit that there does exist some unavoidable limits even with the interaction effect considered.

- In Allison's seminal work (1977), we should notice that the **high-order interactions** (such as the product of three independent variables) should not be neglected and **a hierarchical testing for interaction in multiple regression* should be rigorously followed, both of which imply that our interaction test stopping at ht and lwt is far away from being perfectly finished.
- Product of variables is definitely not the only way to carry out the interaction test due to the fact that there are some other interaction models available in Allison's research.
- Of all 189 observations, there seems a bias during the sampling period when we found some extremely big difference in each group. Such bias may be one of the main reasons why we are exposed to these feeble models.

References

Allison, Paul D. "Testing for Interaction in Multiple Regression." *American Journal of Sociology* 83.1 (1977): 144–153.

Appendix

	Forward model		Backward model	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	2362.21	**<0.001**	2362.21	**<0.001**
ui [yes]	-525.52	**<0.001**	-525.52	**<0.001**
race [other]	126.91	0.422	126.91	0.422
race [white]	475.06	**0.001**	475.06	**0.001**
smoke [yes]	-356.32	**0.001**	-356.32	**0.001**
ht [yes]	-585.19	**0.004**	-585.19	**0.004**
lwt	4.24	**0.012**	4.24	**0.012**
Observations	189		189	
R^2 / R^2 adjusted	0.240 / 0.215		0.240 / 0.215	
AIC	2991.153		2991.153	

Table 3. AIC backward and forward

Table 4. Regression results

	Dependent variable:	
	bwt	
	bwt1	bwt2
	(1)	(2)
lwt	4.242**	3.080*
	(1.675)	(1.794)
htyes:lwt		7.880*
		(4.513)
raceother	126.907	140.128
	(157.594)	(156.897)
racewhite	475.058***	489.749***
	(145.603)	(145.034)
smokeyes	−356.321***	−379.650***
	(103.444)	(103.730)
htyes	−585.193***	−1,791.370**
	(199.644)	(718.757)
uiyes	−525.524***	−536.557***
	(134.675)	(134.073)
Constant	2,362.206***	2,509.395***
	(281.621)	(292.461)
Observations	189	189
R^2	0.240	0.253
Adjusted R^2	0.215	0.224
Residual Std. Error	645.940 (df = 182)	642.335 (df = 181)
F Statistic	9.600*** (df = 6; 182)	8.756*** (df = 7; 181)
Note:		*p<0.1; **p<0.05; ***p<0.01