Assessing Key Drivers of Birthweight

Vidhi Patel

BIS 623

Abstract

This study aims to understand the causes of Low Birth Weight (LBW) and develop a robust methodology for birth weight prediction. LBW, characterized by the World Health Organization as a birth weight of less than 2500 grams, is a critical determinant of health and is associated with higher mortality risk and long-term health impairments. The research utilized a dataset of 4342 data points, encompassing variables such as baby's sex, birth length, mother's weight at delivery, birth head circumference, and mother's weight gain during pregnancy. The data was cleaned, normalized, and univariate analysis as well as multivariate analysis using hybrid forward-backward stepwise regression was conducted to identify statistically significant variables that help explain birth weight. Ridge regression was also done to develop a prediction model for birthweight. Based on our analysis, key variables that can impact birthweight include mother's pre-pregnancy weight, birth head circumference, number of live births prior to current pregnancy, mother's race and father's race. It is important to note, limitations of our analysis stem from the fact that we do not know the population origin for the sample in our dataset.

Introduction

The World Health Organization characterizes birth weight of less than 2500 grams as Low Birth Weight (LBW).[1] Birth weight is widely accepted as a critical determinant of health as it reflects maternal health, nutrition, healthcare accessibility and socioeconomic status.[1,2] LBW is associated with a higher risk of mortality (20 times more than normal birth weight) as well as long-term neurological impairments, immaturity of multiple organ systems and increased susceptibility to chronic diseases.[3,4] Globally, it is estimated that 15-20% of all births have LBW.[5] Moreover, reducing the burden of LBW benefits healthcare systems, providers and households, especially in low-income countries.[6] Understanding the causes behind LBW is essential to help guide measures and enhance infant health outcomes.

However, the biggest issues in preventing LBW is the lack of monitoring and inadequate infant weight measurements. In low-middle income countries, more than half infants aren't weighed.[7] Undoubtedly, thoroughly understanding factors contributing to LBW as well as being able to accurately predict birth weight is to assess health trends. Through this paper, we attempt to do just that. We attempt to examine the causes that influence birth weight and build a robust methodology for birth weight prediction.

Methods

Our preliminary data set consisted of 4342 data points. Along with birth weight (pounds) [bwt], variables consisted of sex of baby [babysex], birth length (cm) [blength], mother's weight at delivery (pounds) [mweight], birth head circumference [bhead], family monthly income (in hundreds, rounded) [fincome], father's race, gestational age in weeks [frace], presence of malformations that could affect weight [malform], mother's age at menarche (years) [menarche], mother's height (inches) [mheight], mother's age at delivery (years) [momage], mother's race [mrace], number of live births prior of this preganancy [parity], previous number of low birth weight babies [pnumlbw], number of prior small for gestational age babies [pnumgsa], mother's prepregnancy BMI [ppbmi], mother's prepregnancy weight (pounds) [ppwt], average number of cigarettes smoked per day during pregnancy [smoken] and mother's weight gain during preganancy (pounds) [wtgain].

Initial data cleaning consisted of converting the outcome of interest, bwt, to pounds as well as factorizing babysex, father race and mother race. Additionally, we removed any columns that only had 1 unique value (pnumsga and pnumlbw) and rows with bwt as 0 or NA. Doing this removed any rows or columns that added no information to the understanding of birth weight predictors. We then looked at a summary table along with histograms of numerical variables to gain a basic understanding of the data. For variables that had heavy right tails, we used log transformation to normalize their distribution. We also converted continuous variables that had distinct clusters in their histogram to categorical variables.

We then examined a correlation plot (Figure 1) to identify any predictor variables that were correlated to each other as well as with bwt. Those that had a high correlation between each other (excluding bwt) were dropped to minimize autocorrelation. We also conducted univariate analysis using two-tailed t-tests for continuous and one-way ANOVA (adjusted to unequal variance) for categorical variables. After thoroughly assessing the data, we conducted hybrid forward - backward stepwise regression to understand the relationship between each variable and birth weight while accounting for other variables. Doing so, we were able to identify statistically significant variables (two-tailed test for continuous and chi-squared for categorical) that help explain birth weight to a certain degree. The adjusted R-squared value was used to draw conclusions from the final inference model.

We also built a model to predict birth weight using the variables in the cleaned dataset. To do this, we split the data into a testing (30%) and training dataset (70%). We standardized continuous variables so each had a mean of 0 and a standard deviation of 1 unit. Using the training dataset, we conducted ridge regression to adjust multicollinearity between variables and arrived at a robust final model. With this model, we predicted birth weight in the testing dataset. The RMSE and R-squared values were used to draw conclusions for the final prediction model.

For both the inference and prediction models, we looked at the variance inflation factor (VIF) for each predictor variable to correct for multicollinearity. We also used residuals plot, Q-

Q plot and Cook's distance plot to ensure HEIL Gauss constraints were met. A significance value of p< 0.05 was used. All analysis was done in R.

Results

During our preliminary analysis, we found that there is about an equal distribution of male (51.4%) and female babies (48.6%). Mother and father races were mostly white (father: 48.9%; mother: 49.4%) or black (father: 44.0%; mother: 43.9%) compared to any other race. We noticed that three variables were heavily right skewed and so log transformation was conducted to gain a normal distribution: momage, ppwt and delwt. The variables blength and menarche mostly clustered around three separate groups so we converted them to categorical variables with three categories as follows: blength [short: less than 47 cm; medium: between 47 to 53 cm; long: greater than 53cm] and menarche [early: less than 10 years; normal: between 10 to 15 years; late: more than 15 years].

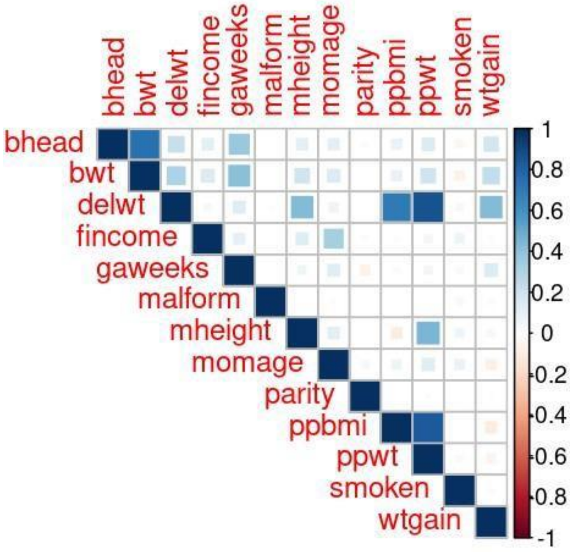

Figure 1: Correlation Plot

When looking at the correlation plot (Figure 1), we found that blength and bhead have high correlation and equal amount of correlation to bwt. Additionally, delwt highly correlated with ppbmi and ppwt and moderately correlated with mheight and wtgain. Because ppbmi and ppwt are highly correlated, we drop ppbmi. We also dropped blength and delwt. Based on the correlation plot, we also noticed that delwt and gaweeks had some correlation to bwt and mheight, momage, ppbmi, ppwt and wtgain had low correlation with bwt. These variables could prove to be of importance in the inference and prediction models.

X Variable	**X Variable Mean**	**Birthweight**	**p-value**
Continuous Variables (Welch Two Sample T-test)			
Birth Head Circumference (bhead)	33.65	6.86597	< 2.2e-16
Family Monthly Income (fincome)	44.11		< 2.2e-16
Presence of Malformations (malform)	0.0035		< 2.2e-16
Mother’s Height (mheight)	63.49		< 2.2e-16
Mother’s Age at Delivery (momage)	20.30		< 2.2e-16
Number Of Live Births Prior of This Pregnancy (parity)	0.0023		< 2.2e-16
Mother’s Pre-Pregnancy Weight (ppwt)	123.49		< 2.2e-16
Cigarettes Smoked Per Day During Pregnancy (smoken)	4.15		< 2.2e-16
Mother’s Weight Gain (wtgain)	22.08		< 2.2e-16
Gestational Age in Weeks (gaweeks)	39.43		< 2.2e-16

Mother’s Age at Menarche (menarche)	12.51		< 2.2e-16
Categorical Variables (One-Way ANOVA)			
	F-Statistic		
Sex (babysex)			< 2.2e-16
Mother’s Race (mrace)	100.34		< 2.2e-16
Father’s Race (frace)	139.79		< 2.2e-16

Table 1: Univariate Analysis

Based on the univariate analysis (Table 1), it appears that all variables in the dataset post-processing have a mean that is significantly different from the mean of birthweight. This implies that all of these variables influence birthweight outcomes. For our inference model, we conducted a hybrid stepwise regression and found that baby sex of female compared to male (p <. 05), mother's pre-pregnancy weight (pounds) (p < 2e-16), birth head circumference (p < 2e-16), mother’s height (p < 2e-16), gestational age in weeks (p < 2e-16), average number of cigarettes smoked per day during pregnancy (p < 2e-16), weight gain during pregnancy (p < 2e-16) and mother’s race being black (p < 2e-16) or Puerto Rican (p < 7e -9) compared to white were all significantly associated with birth weight. Having a female baby compared to male, mother's pre-pregnancy weight (pounds), birth head circumference, mother’s height, gestational age in weeks are all positively associated with birth weight. In contrast, average number of cigarettes smoked per day during pregnancy, weight gain during pregnancy and mother’s race being Black, Asian or Puerto Rican compared to White were all negatively associated with birth weight. Overall, we found that about 62% of the variation in birth weight around its mean is explained by the variables mentioned above (Table 2).

Variable	Coefficient	p-value	VIF
intercept	-1.394e+01	< 2e-16	1.044007
babysex (female)	5.040e-02	0.0184	1.315062
ppwt	5.903e-01	5.93e-14	1.321627
bhead	4.364e-01	< 2e-16	1.335978
mheight	2.385e-02	1.68e-07	1.206374
gaweeks	4.364e-02	< 2e-16	1.091863
smoken	-1.474e-02	< 2e-16	1.064398
wtgain	1.212e-02	< 2e-16	1.237329
mother’s race (black)	-3.833e-01	< 2e-16	
mother’s race (asian)	-1.690e-01	0.1148	
mother’s race (puerto rican)	-2.799e-01	7.39e-09	
Residual standard error: 0.6891 on 4331 degrees of freedom **Multiple R-squared**: 0.6284, **Adjusted R-squared**: 0.6275 **F-statistic**: 732.4 on 10 and 4331 DF; **p-value**: < 2.2e-16			

Table 2: Inference Model Output

During variable selection, we also inputted an interaction term between gestational age in weeks and weight gain during pregnancy but found that it was not significant enough to add to the inference model. The VIF factor for each of these variables indicates no presence of multicollinearity. Furthermore, when examining the residual and Q-Q plots, we found that the residuals are mostly randomly distributed, and the Q-Q plot indicates a normal distribution. It is important to note that the residuals are slightly skewed to the left, however.

To create a prediction model for birth weight, we conducted ridge regression on all the variables present post-processing. The predictive performance of the model was evaluated on the testing dataset, revealing that approximately 59% of the total variance in birth weight could be explained by the included variables. This value suggests a moderate level of predictive capability, indicating that the model accounts for a substantial portion of the variability in birth weight. Additionally, the Root Mean Squared Error (RMSE) of 0.700 was observed, representing the average discrepancy between the predicted birth weights and the actual birth weights in the testing dataset. The magnitude of effect as well as the direction of association for each variable can be seen in Table 3.

Variable	**Magnitude**	**Direction of Association**
Mother's Race (Black)	-2.06e-01	Negative
Mother's Race (Puerto Rican)	-1.75e-01	Negative
Father's Race (Black)	-1.57e-01	Negative
Mother's Race (Asian)	-1.09e-01	Negative
Menarche (Late)	-7.24e-02	Negative
Presence of Malformations (malform)	-6.31e-02	Negative
Father's Race (Puerto Rican)	-4.07e-02	Negative
Mother's Age at Delivery (momage)	-3.84e-02	Negative
Cigarettes Smoked Per Day During Pregnancy (smoken)	-1.33e-02	Negative
Family Monthly Income (fincome)	9.64e-04	Positive
Mother's Weight Gain during Pregnancy (wtgain)	1.10e-02	Positive
Mother's Height (mheight)	2.79e-02	Positive
Sex (female)	2.87e-02	Positive
Father's Race (Other)	4.63e-02	Positive
Gestational Age in Weeks (gaweeks)	5.15e-02	Positive
Menarche (Normal)	5.80e-02	Positive
Father's Race (Asian)	1.18e-01	Positive
Number of Live Births Prior to this Pregnancy (parity)	1.74e-01	Positive
Birth Head Circumference (bhead)	3.97e-01	Positive
Mother's Pre-Pregnancy Weight (ppwt)	5.67e-01	Positive

Table 3: Ridge Regression Model Coefficients

Conclusion

Through this report, we sought to understand key drivers of birthweight using a dataset of over 4000 data points. We also assessed which factors have a statistically significant association with birthweight and built a model using ridge regression to predict birthweight. Based on our analysis, key variables that can impact birthweight include mother's pre-pregnancy weight, birth head circumference, number of live births prior to current pregnancy, mother’s race and father’s race. Additionally, having either parent of a minority race is almost always negatively associated with birth weight while, mother's weight related factors are almost always positively associated with birth weight.

Limitations of our analysis stem from the fact that we do not know the population origin for the sample in our dataset. Because of this, it is hard to draw any concrete conclusions that could be generalizable to a given population. Having imbalanced racial representations impacts the validity of our interpretation of the correlation between birthweight and mother/father's race. Additionally, we did not resolve some of the skewness present in our residuals and Q-Q plot which could impact the robustness of our interpretation of the inference model. Performing other types of regression analysis (i.e., LASSO) or more advanced modeling (i.e., mixed-effects, random forest, gradient boosting) for predictions could also prove to be more fruitful.

References

1. Trends in maternal mortality 2000 to 2017: estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division: executive summary. https://apps.who.int/iris/handle/10665/327596.

2. Lawn JE, Cousens S, Zupan J. 4 Million neonatal deaths: when? where? why? Lancet. 2005;365:891–900.

3. United Nations Children's Fund (UNICEF), World Health Organization. LOW BIRTHWEIGHT ESTIMATES Levels and trends 2000–2015. Lancet Glob Heal. 2019;7:e849–60.

4. Goyal N, Canning D. The association of in-utero exposure to ambient fine particulate air pollution with low birth weight in India. Environ Res Lett. 2021;16:054034.

5. Groen-Blokhuis MM, Middeldorp CM, Van Beijsterveldt CEM, Boomsma DI. Evidence for a causal association of low birth weight and attention problems. J Am Acad Child Adolesc Psychiatry. 2011;50:1247-1254.e2.

6. Mccormick MC, Brooks Gunn J, Workman Daniels K, Turner J, Peckham GJ. The health and developmental status of very low-birth-weight children at school age. JAMA. 1992;267:2204–8.

7. Lee, A. C., Katz, J., Blencowe, H., Cousens, S., Kozuki, N., Vogel, J. P., ... & Black, R. E. (2013). National and regional estimates of term and preterm babies born small for gestational age in 138 low-income and middle-income countries in 2010. The Lancet global health, 1(1), e26-e36.

8. Jamshed S, Khan F, Chohan SK, et al. Frequency of Normal Birth Length and Its Determinants: A Cross-Sectional Study in Newborns. Cureus. 2020;12(9):e10556. Published 2020 Sep 20. doi:10.7759/cureus.10556

9. Lacroix AE, Gondal H, Shumway KR, Langaker MD. Physiology, Menarche. In: StatPearls. StatPearls Publishing, Treasure Island (FL); 2022. PMID: 29261991.

Appendix

```
## ----------------------------------------------------------------------------
library(tidyverse)
library(dplyr)
library(corrplot)
library(MASS)
library(ggfortify)
library(car)
library(glmnet)
load('BIS623_FinalProjectData.rda')



## ----------------------------------------------------------------------------
#convert bwt from grams to pounds
data$bwt = (data$bwt)/453.6

#factorize categorical data
data$babysex = factor(data$babysex)
data$frace = factor(data$frace)
data$mrace = factor(data$mrace)

#drop columns with 0 - if it doesn't make sense
colsrmv = c("bhead", "blength", "delwt", "gaweeks", "menarche", "mheight","momage",
"ppbmi", "ppwt")

data = subset(data, select = sapply(data[colsrmv], function(x) length(unique(x))) > 1)

# remove rows with 0 birth weight and na
data = data[!(data$bwt == 0 | is.na(data$bwt)), ]



## ----------------------------------------------------------------------------
#quick glance at data
head(data)

# summary statistics
summary(data)
```

```
## ----------------------------------------------------------------------------
# drop columns with same value
data = data[, -c(15, 16)]


## ----------------------------------------------------------------------------
par(mar = c(1, 1, 1, 1))
# analyze distribution of numeric variables
numeric_cols <- data[, sapply(data, is.numeric)]
par(mfrow = c(ceiling(sqrt(ncol(numeric_cols))), ceiling(sqrt(ncol(numeric_cols)))))
for (col in names(numeric_cols)) {
  hist(numeric_cols[[col]], main = col)
}
dev.off()



## ----------------------------------------------------------------------------
# transform non-normal columns
data2 <- data
data2$momage = log(data2$momage)
data2$ppwt = log(data2$ppwt)
data2$delwt = log(data2$delwt)

# convert columns with clusters into categories
data2$menarche = factor(ifelse(data2$menarche < 10, "Early",
                        ifelse(data2$menarche >= 10 & data2$menarche <= 15, "Normal",
"Late")))
data2$blength = factor(ifelse(data2$blength < 47, "Short",
                        ifelse(data2$menarche >= 47 & data2$menarche <= 53, "Normal",
"Long")))


## ----------------------------------------------------------------------------
numeric_cols <- data2[, sapply(data2, is.numeric)]
# correlations
corrplot(cor(numeric_cols),method = "square", type = "upper")
```

```
## ---------------------------------------------------------------------------
# drop colinear columns
data3 = data2[, -c(5, 3, 15)]


## ---------------------------------------------------------------------------
# conduct t-test and ANOVA for variables
t.test(data3$bhead, data3$bwt)
t.test(data3$fincome, data3$bwt)
t.test(data3$malform, data3$bwt)
t.test(data3$mheight, data3$bwt)
t.test(exp(data3$momage), data3$bwt)
t.test(data3$parity, data3$bwt)
t.test(exp(data3$ppwt), data3$bwt)
t.test(data3$smoken, data3$bwt)
t.test(data3$wtgain, data3$bwt)
t.test(data3$gaweeks, data3$bwt)
t.test(data$menarche, data3$bwt)
aov(data3$bwt ~ data3$babysex)
oneway.test(data3$bwt ~ data3$frace, var.equal = FALSE)
oneway.test(data3$bwt ~ data3$mrace, var.equal = FALSE)


## ---------------------------------------------------------------------------
# hybrid stepwise regression
back <- lm(bwt ~ . + gaweeks * ppwt, data = data3)
step <- stepAIC(back, direction = "both", trace = FALSE)
summary(step)

# residual and Q_Q plot
autoplot(step, which = 1:6)


## ---------------------------------------------------------------------------
# final inference model
regplot = lm(bwt ~ babysex+ppwt+bhead+mheight
                +gaweeks+smoken+wtgain+mrace, data = data3)
summary(regplot)
autoplot(regplot, which = 1:6)
```

```
# VIF for inference model
vif(regplot)


## ----------------------------------------------------------------------------
# split data into training and testing
set.seed(12)
train_indices <- sample(1:nrow(data3), 0.7 * nrow(data3))
train_data <- data3[train_indices, ]
test_data <- data3[-train_indices, ]

x_train = model.matrix(bwt~.,train_data)[,-1]
y_train = train_data$bwt

x_test = model.matrix(bwt~.,test_data)[,-1]
y_test = test_data$bwt


## ----------------------------------------------------------------------------
# ridge regression
ridge = glmnet(x_train, y_train, alpha = 0)
summary(ridge)

cv_ridge = cv.glmnet(x_train, y_train, alpha = 0)

eval_results =  function(real, pred, df) {
  SSE <- sum((pred - real)^2)
  SST <- sum((real - mean(real))^2)
  RSq <- 1 - SSE / SST
  RMSE = sqrt(SSE/nrow(df))
  data.frame(
    RMSE = RMSE,
    RSq = RSq
  )
}

ridge_final = glmnet(x_train,y_train,alpha = 0, lambda = cv_ridge$lambda.min)
# get coeff of model
coef(best_model)
```

```
# predict on test data
pred_test =  predict(ridge, s = cv_ridge$lambda.min, newx = x_test)

# get RMSE and R_2
eval_results(y_test, pred_test, test_data)
```