Module 6 | Practice 6: birthwt dataset

Trang Tran

CPS, Northeastern University

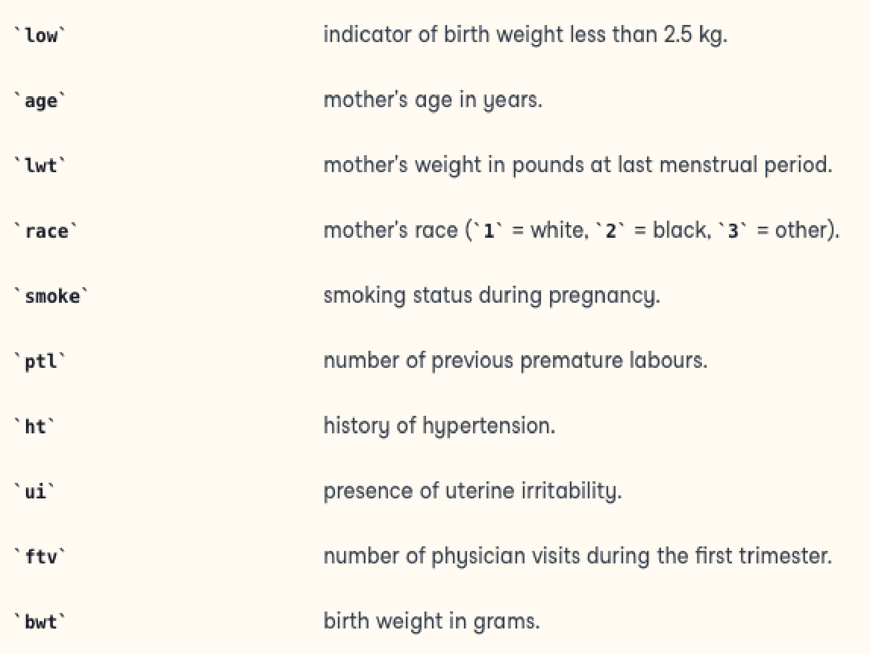
ALY6010 | Probability Theory and Introductory Statistics

Patrick McQuillan

Mar 31, 2023

Introduction

The “birthwt” dataset related to risk factors associated with low infant birth weight consists of 10 columns and 189 rows which have 10 numeric variables. Below is the data dictionary:



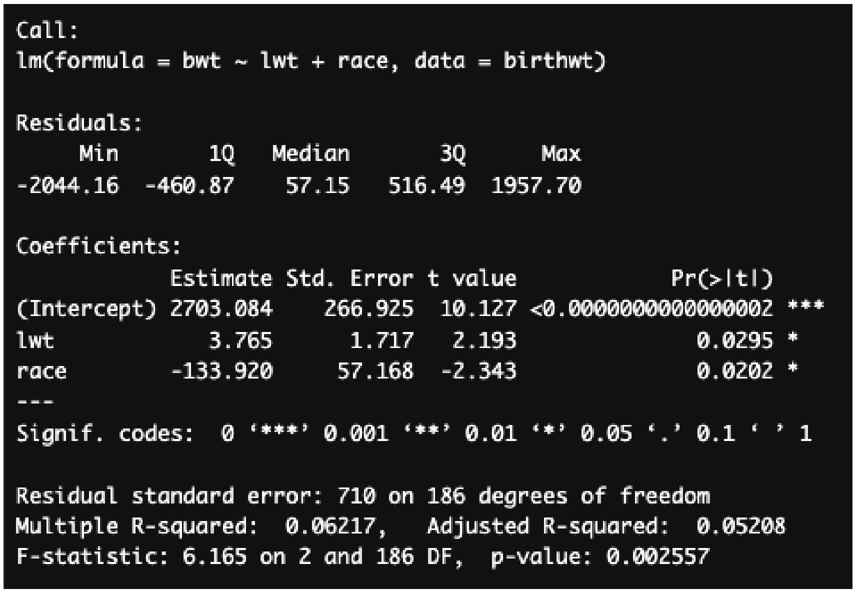
`low`	indicator of birth weight less than 2.5 kg.
`age`	mother's age in years.
`lwt`	mother's weight in pounds at last menstrual period.
`race`	mother's race (`1` = white, `2` = black, `3` = other).
`smoke`	smoking status during pregnancy.
`ptl`	number of previous premature labours.
`ht`	history of hypertension.
`ui`	presence of uterine irritability.
`ftv`	number of physician visits during the first trimester.
`bwt`	birth weight in grams.

Figure 1: Data Dictionary

I used [summary] and [skim] functions to get an overview of the dataset. After using the correlation function to check the whole correlation relationships of the dataset, I decided to choose these variables to make further regression analysis: “bwt” (birth weight in grams), “lwt” (mother's weight in pounds at last menstrual period), and “race” (mother's race (1 = white, 2 = black, 3 = other)).

Part I

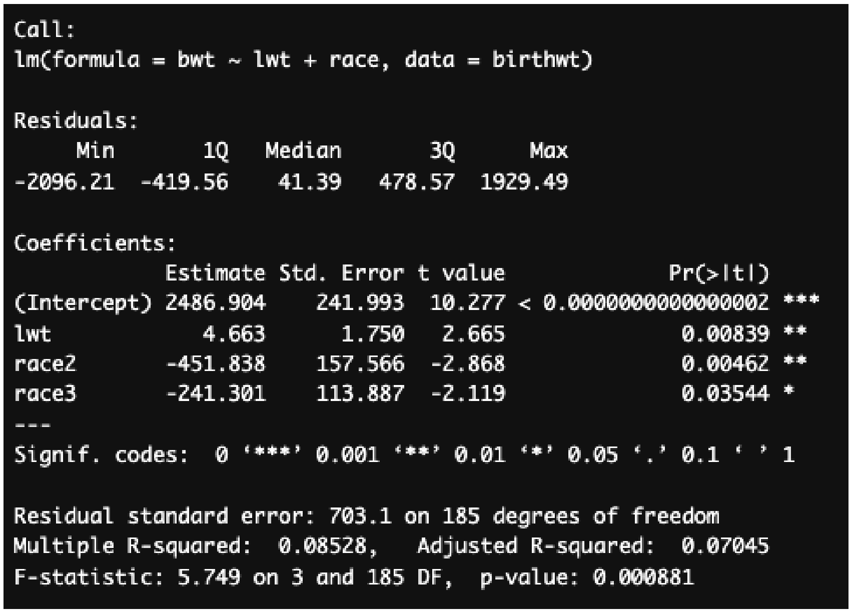
Regression Model



```
Call:
lm(formula = bwt ~ lwt + race, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max 
-2044.16  -460.87    57.15   516.49  1957.70 

Coefficients:
            Estimate Std. Error t value           Pr(>|t|)    
(Intercept) 2703.084    266.925  10.127 <0.0000000000000002 ***
lwt            3.765      1.717   2.193             0.0295 *  
race        -133.920     57.168  -2.343             0.0202 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 710 on 186 degrees of freedom
Multiple R-squared:  0.06217,   Adjusted R-squared:  0.05208 
F-statistic: 6.165 on 2 and 186 DF,  p-value: 0.002557
```

Figure 2: Regression Model without dummy variables

Firstly, I run a regression model between the dependent variable “bwt” and predictor variables: “lwt” and “race” as seen in Figure 2 besides. Since the p-value of “lwt” is less than the significance level of 0.05 (0.0295 < 0.05), we can conclude that the mother’s weight is a statistically significant predictor of birth weight. Also, they have a positive relationship which is reflected through the regression coefficient of 3.765. Because “race” is a categorical variable labeled by numeric data: 1 = white, 2 = black, 3 = other, I convert the “race” column into a factor type. Then we rerun the regression model with automatically created dummy variables of race as shown in Figure 3 below. In this case, “white” or label 1 of race is our baseline value. Each one-pound increase in maternal weight is associated with an average increase of 4.663 in birth weight. Since the p-value (0.00839) is less than 0.05, “lwt” is a statistically significant predictor of “bwt”. To interpret the dummy variables, we look at:



```
Call:
lm(formula = bwt ~ lwt + race, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.21  -419.56    41.39   478.57  1929.49

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 2486.904    241.993  10.277 < 0.0000000000000002 ***
lwt            4.663      1.750   2.665              0.00839 **
race2       -451.838    157.566  -2.868              0.00462 **
race3       -241.301    113.887  -2.119              0.03544 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 703.1 on 185 degrees of freedom
Multiple R-squared:  0.08528,   Adjusted R-squared:  0.07045
F-statistic: 5.749 on 3 and 185 DF,  p-value: 0.000881
```

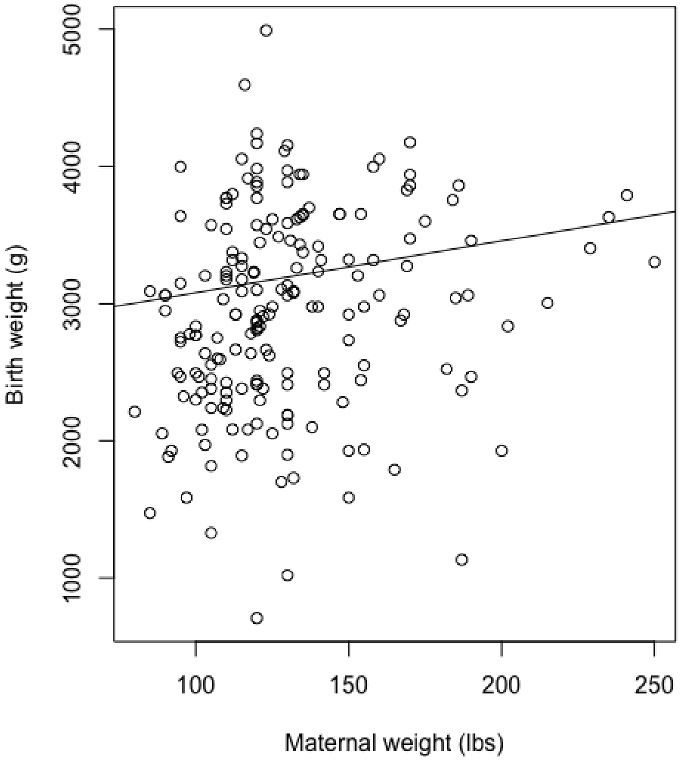
Figure 3: Regression Model with dummy variables

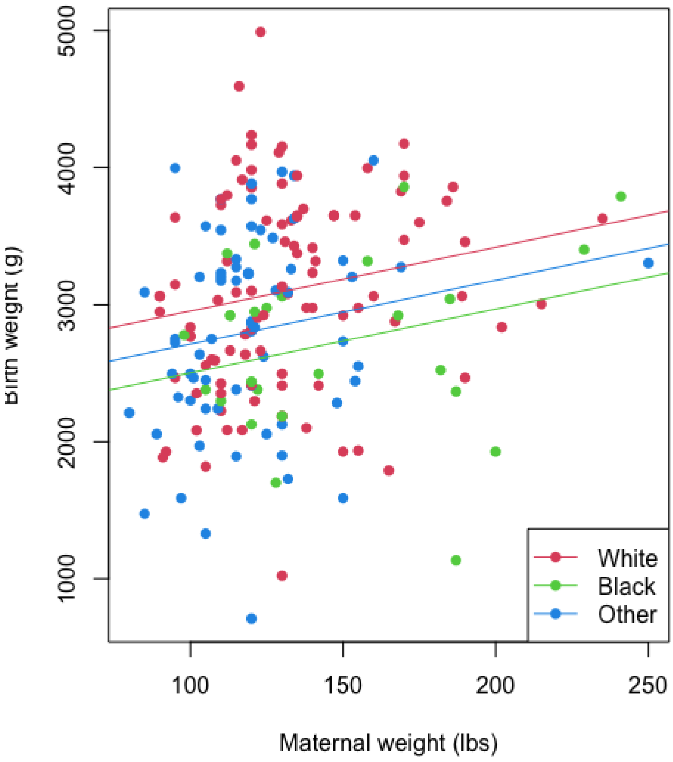
- “Race2”: A black mother, on average, has a birth weight of 451.838 grams less than a white individual. Since the p-value (0.00462) is much less than 0.05, this difference is statistically significant.
- “Race3”: A other-race mother (not white nor black race), on average, has a birth weight of 241.301 grams less than a white individual. Since the p-value (0.03544) is still less than 0.05, this difference is statistically significant.

The fitted regression line turns out to be:

Birthweight = 2486.904 + 4.663 *(maternal weight) – 451.838*(black) – 241.301*(other)

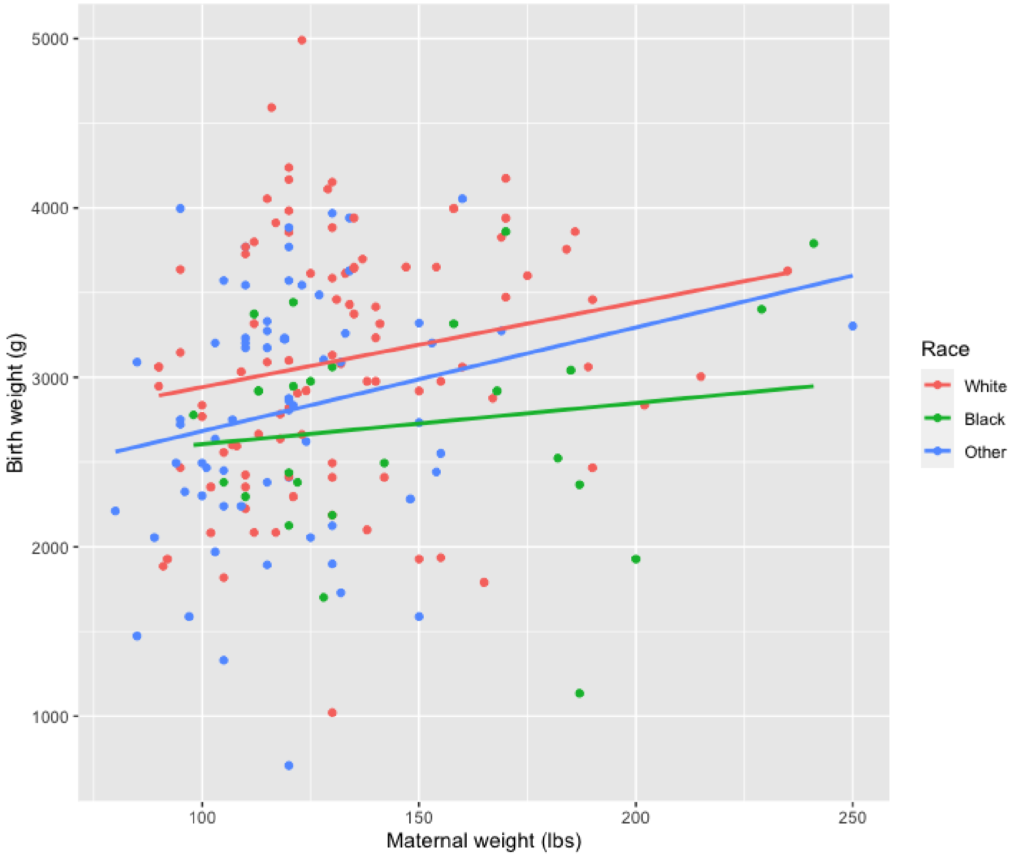
The two graphs below present the original regression line and the regression line with the dummy variables of race. We see a slight difference in the line's slope between the two types of regression lines. There are three regression lines colored based on three races in the dataset, and of course with the same slope.





Part II

In this part, we continue to create a scatterplot with separate regression lines for race subsets. The plot below shows three different lines in slope representing three race subsets. We see the strongest positive relationship between maternal weight and birth weight in the other-raced mothers (blue line). Meanwhile, with black mothers, the relationship between maternal weight and birth weight is the weakest (green line). Linear regression using data subsets brings more meaningful insight into analyzing a mother's race affecting the whole relationship in this dataset.


5000
4000
3000
2000
1000
Birth weight (g)
100
150
200
250
Maternal weight (lbs)
Race
White
Black
Other

References

1. Stackoverflow. *Plot regression lines in r with multiple dummy variables*. Retrieved March 31, 2023. https://stackoverflow.com/questions/67556086/plot-regression-lines-in-r-with-multiple-dummy-variables
2. RDocumentation. *birthwt: Risk Factors Associated with Low Infant Birth Weight*. Retrieved March 31, 2023. https://www.rdocumentation.org/packages/MASS/versions/7.3-55/topics/birthwt
3. Statology. *How to Create Dummy Variables in R (Step-by-Step).* Retrieved March 31, 2023. https://www.statology.org/dummy-variables-in-r/