# Report on birthwt Data Set

Saikat Bera, Presidency University, KOLKATA

December 2, 2022

**Abstract**

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

We will be using the birthwt dataset from the MASS library. We first load the dataset and run summary:

```
##       low              age              lwt             race       
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000  
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000  
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847  
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000  
##      smoke            ptl               ht                ui        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000  
##       ftv              bwt        
##  Min.   :0.0000   Min.   : 709  
##  1st Qu.:0.0000   1st Qu.:2414  
##  Median :0.0000   Median :2977  
##  Mean   :0.7937   Mean   :2945  
##  3rd Qu.:1.0000   3rd Qu.:3487  
##  Max.   :6.0000   Max.   :4990
```

This data frame contains the following columns:

- 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.

We fit a model predicting birth weight using mother's age, mother's weight, smoking status, self-reported race, and number of previous premature labors. The selected race for reference is the Caucasian group.

```
##    bwt age lwt smoke  race ptl
## 1 2523  19 182    no black   0
## 2 2551  33 155    no other   0
## 3 2557  20 105   yes white   0
## 4 2594  21 108   yes white   0
## 5 2600  18 107   yes white   0
## 6 2622  21 124    no other   0
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2300.22  -450.53    26.89   519.96  1702.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2853.8412   321.1317   8.887 6.05e-16 ***
## age           -0.4701     9.8749  -0.048 0.962079
## lwt            3.7001     1.7516   2.112 0.036014 *
## smokeyes    -373.5910   111.3413  -3.355 0.000965 ***
## raceblack   -503.3695   156.9314  -3.208 0.001582 **
## raceother   -387.7939   119.7022  -3.240 0.001423 **
## ptl         -131.1114   104.3283  -1.257 0.210466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 681 on 182 degrees of freedom
## Multiple R-squared:  0.1556,Adjusted R-squared:  0.1278
## F-statistic: 5.591 on 6 and 182 DF,  p-value: 2.405e-05
```

Keeping all the other variables fixed, babies of black mothers are born on average weighing more than babies of white mothers. The amount of more weight is given by the coefficient:

```
## raceblack
## -503.3695
```

Similarly, keeping all the other variables fixed, babies of other mothers are born on average weighing more than babies of white mothers. The amount of more weight is given by the coefficient:

```
## raceother
## -387.7939
```

By a similar argument, we can say that keeping all other variables fixed, babies of mothers who smoke are born on average weighing more than babies of white mothers. The amount of more weight is given by the coefficient:

```
## raceother
## -387.7939
```

We re-designed the model to find the coefficient of non-Caucasian vs. Caucasian:

```
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2310.03  -445.69     4.02   464.36  1680.52
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2469.1572   289.9116   8.517 5.89e-15 ***
## age                  0.1149     9.8247   0.012 0.990685
## lwt                  3.3413     1.6709   2.000 0.047018 *
## smokeyes          -385.7292   109.7989  -3.513 0.000558 ***
## raceNon-Caucasian  424.1822   107.4191   3.949 0.000112 ***
## ptl               -129.8287   104.1637  -1.246 0.214214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 680.1 on 183 degrees of freedom
## Multiple R-squared:  0.1534,Adjusted R-squared:  0.1303
## F-statistic: 6.632 on 5 and 183 DF,  p-value: 1.064e-05
```

The interpretation would be like this: Keeping all the other variables fixed, babies of Non-Caucasian mothers are born on average weighing more than babies of Caucasian mothers. The amount of more weight is given by this coefficient:

```
## raceNon-Caucasian
##          424.1822
```

We note that all the variables approximately retained their previous estimates and no new variables became insignificant. However, the standard error decreased slightly.

We might ignore the age variable as it is insignificant.

There are only a few unique values for number of previous premature labors; we might be better off treating this variable as categorical.

```
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2309.12  -404.23   -10.77   497.73  1684.53
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2478.318    223.597  11.084  < 2e-16 ***
## lwt                  3.520      1.624   2.168 0.031441 *
## smokeyes          -370.501    107.773  -3.438 0.000727 ***
## raceNon-Caucasian  394.183    104.051   3.788 0.000206 ***
## ptl1              -368.544    150.147  -2.455 0.015046 *
## ptl2              -124.204    307.063  -0.404 0.686328
## ptl3               800.590    679.328   1.179 0.240135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 670.7 on 182 degrees of freedom
## Multiple R-squared:  0.1809,Adjusted R-squared:  0.1539
## F-statistic: 6.701 on 6 and 182 DF,  p-value: 1.998e-06
```

Looking at the p-values, I would rather classify this variable in three factors: 0,1, and greater than 1.

```
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2310.01  -408.89   -20.12   489.46  1683.03
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2485.083    223.882  11.100  < 2e-16 ***
## lwt                  3.434      1.625   2.114 0.035909 *
## smokeyes          -365.327    107.863  -3.387 0.000865 ***
## raceNon-Caucasian  399.551    104.126   3.837 0.000171 ***
## ptl1              -369.797    150.379  -2.459 0.014858 *
## ptl2                25.660    283.323   0.091 0.927935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 671.8 on 183 degrees of freedom
## Multiple R-squared:  0.1739,Adjusted R-squared:  0.1513
## F-statistic: 7.702 on 5 and 183 DF,  p-value: 1.352e-06
```

The p-values of ptl1 and ptl2 are very high, indicating they are not significant. Thus, we can model taking this variable as: 0 and greater than 0.

```
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2315.61  -441.20   -16.55   501.57  1676.23
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2501.926    223.885  11.175  < 2e-16 ***
## lwt                  3.263      1.622   2.012 0.045674 *
## smokeyes          -365.623    108.051  -3.384 0.000874 ***
## raceNon-Caucasian  410.433    103.960   3.948 0.000112 ***
## ptl1              -291.747    137.728  -2.118 0.035493 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 673 on 184 degrees of freedom
## Multiple R-squared:  0.1664,Adjusted R-squared:  0.1483
## F-statistic: 9.185 on 4 and 184 DF,  p-value: 8.677e-07
```

I will choose the last model for regression. This is because among the previous models, this model would have almost all the variables that are significant. However, we are losing much information about the number of premature

labours.

Leaving race as non-Caucasian vs. Caucasian and premature labors as $> 0$ vs. 0, we can also take into account the number of physician visits during the first trimester. We model using the factor as 0 visits and greater than 0 visits.

```
##
## Call:
## lm(formula = bwt ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2337.10  -443.55   -10.29   493.66  1641.18
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2469.169    287.772   8.580 4.08e-15 ***
## age                  0.823     10.028   0.082 0.934683
## lwt                  3.207      1.656   1.936 0.054375 .
## smokeyes          -353.157    111.292  -3.173 0.001770 **
## raceNon-Caucasian  396.394    109.000   3.637 0.000359 ***
## ptl1              -299.031    140.216  -2.133 0.034294 *
## ftv1                51.761    104.347   0.496 0.620458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 676.1 on 182 degrees of freedom
## Multiple R-squared:  0.1677,Adjusted R-squared:  0.1403
## F-statistic: 6.113 on 6 and 182 DF,  p-value: 7.432e-06
```

**Reference:-**

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.