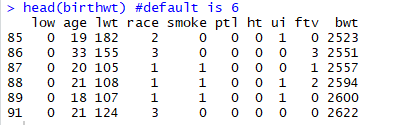
Data Science in Biomedical Research

**Kishan Sarpangala**

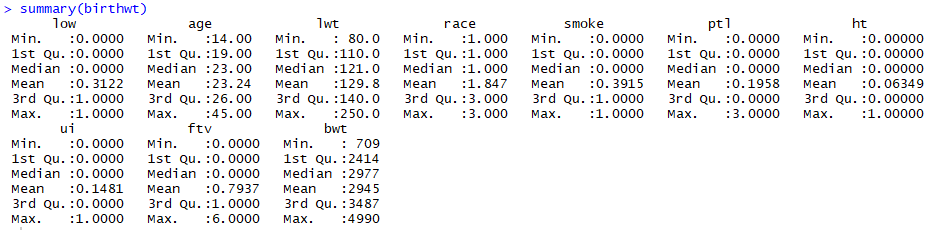
**Assignment 1**

**1. What is the size of the dataset? Show the top six rows of the dataset.**





head(birthwt) default is set at 6; if you want to increase the number of rows you can edit it to Head(birthwt,9)



**2. Understand the data. Describe the meaning of each variable.**

Birthwt is data which is also available from the MASS package. This data set includes the birth weight (in grams) of 189 newborn babies along with other characteristics like ( age, smoking status) of their mothers.

The data were collected at Baystate Medical Center, Springfield, MA, during 1986

**low**: indicator of birth weight less than 2.5 kg (0 = normal birth weight, 1 = low birth weight).

**age**: mother’s age in years.

**lwt**: mother’s weight in pounds at last menstrual period.

**race**: mother’s race (1 = white, 2 = African-American, 3 = other).

**smoke**: smoking status during pregnancy (0 = not smoking, 1 = smoking).

**ptl**: number of previous premature labors.

**ht**: history of hypertension (0 = no, 1 = yes).

**ui**: presence of uterine irritability (0 = no, 1 = yes).

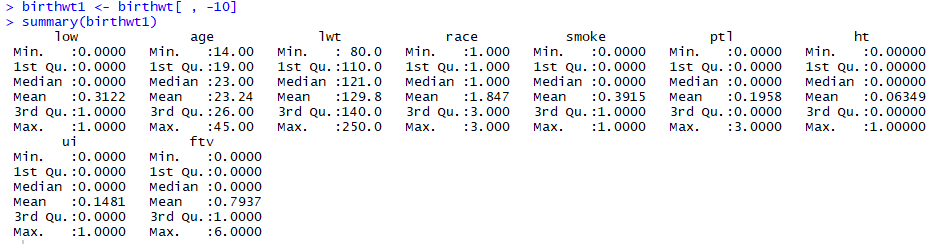
**ftv**: number of physician visits during the first trimester.

**bwt**: birth weight in grams.

**Note**:

1. Variables age, lwt, ptl, ftv, and bwt are numerical variables. Among these variables, ptl and ftv are count variables.
2. The variables low, race, smoke, ht, and ui are all categorical.
3. all categorical variables are coded with numerical values

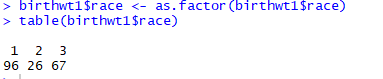
**3. The first column of the data is binary. If the baby has low weight it is indicated by 0. Otherwise 1 is used. The last column actually gives the birth weight. The first column was created from the last column by dichotomization. Omit the last column, which is the 10th column of the data. (< birthwt1 <- birthwt[ , -10] ) Determine the nature or data type or class of each variable. (apply(birthwt1, 2, class))**





* Image above mentions the data types or class of each variable.

**4. ‘race’ is ternary. Change it into a categorical (factor) variable. (< birthwt1$race <- as.factor(birthwt1$race)) Count how many whites, blacks, and others are in the study. (table(birthwt1$race))**

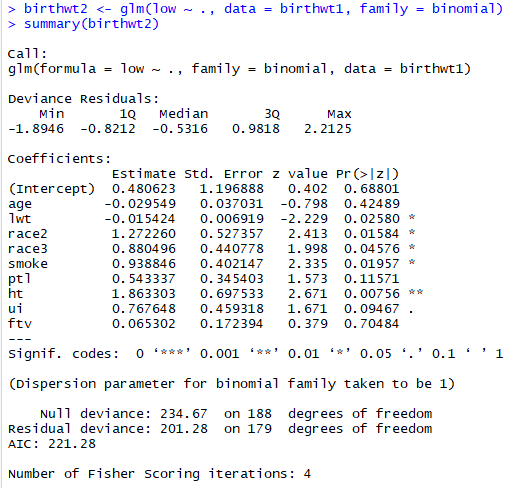


* White is 96
* African American is 26
* Other is 67

**5. Postulate a logistic regression model with ‘low’ as response variable and the remaining as predictors.**

A data frame with 189 observations; low is indicator of birth weight less than 2.5kg

Let’s fit a logistic regression model to the ‘birthwt1’ data



**6. Fit the model to the data. Write the prediction equation.**

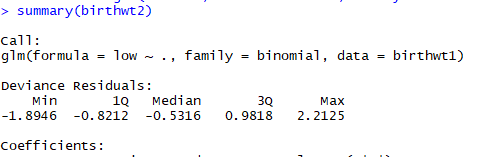
Pr(low) = exp(0.4806 + -0.029\*age + -0.0154\*lwt + 1.27226\*race2 + 0.8804\*race3+0.9388\*smoke+0.543337\*ptl+1.863303\*ht+0.767648\*ui+0.065302\*ftv)/

(1 + exp(0.4806 + -0.029\*age + -0.0154\*lwt + 1.27226\*race2 + 0.8804\*race3+0.9388\*smoke+0.543337\*ptl+1.863303\*ht+0.767648\*ui+0.065302\*ftv))

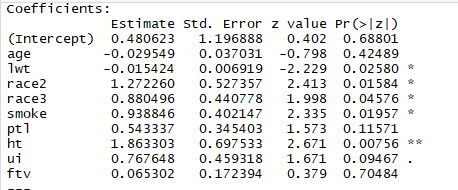
**Note**: age coefficient is zero so is lwt variable which is also set to zero

**7. Which predictors are significant? Interpret the coefficients associated with ‘race.’**

This information is available from the summary.



* More stars a variable has the more significant it is.
* Hypertension is most significant as we can notice from the image below. (two stars) (image in next page -Image1)
* We can notice that mother’s race, mother’s weight in pounds at last menstrual period and smoking status during pregnancy also plays a role in birth weight of newborn babies - as this is highlighted by one star designation. (image in next page -Image 1)
* R uses white race as the base; this is due to alpha-numerically- as W is the last letter in the alphabet.
* Race 2 is African-American while Race 3 is associated with other and W is associated with white.
* **African-American the coefficient value is 2.413 and for Other races the coefficient value is 1.998**



**Image 1**

**8. Interpret the coefficients.**

* It is exponential of the coefficients column of the model.
* Each coefficient is the log odds associated with only that variable. The intercept is the base log odds that everyone starts with (much like the intercept in a linear regression). Negative log odds decrease your risk. Positive log odds increase your risk. 0 is neutral.
* If we want to get the odds ratio back, we need to do exp(coefficient).

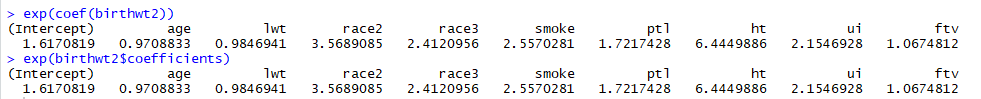
**For example**, the odds associated with each year increase in age are

exp(-0.02955) = 0.97088. That is, each year increases in age decreases your odds of giving to birth to an infant of low birth weight by 3%. Refer to the above image 1 and below image 2

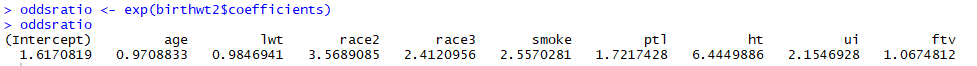


**Image 2**

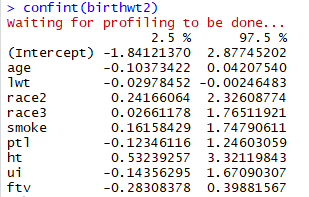
**Note**: exp(coef(birthwt2)) or exp(birthwt2$coefficients) both give same results.



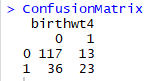
**9. Lay out the odds ratios along with their 95% confidence intervals.**



The command confint() can be employed with the coefficients of the model to obtain their 95% confidence interval for example



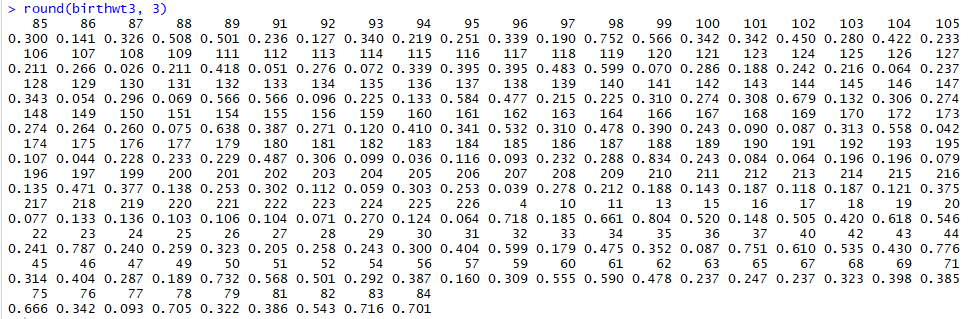
**10. Work out the confusion matrix. Determine the misclassification rate.**

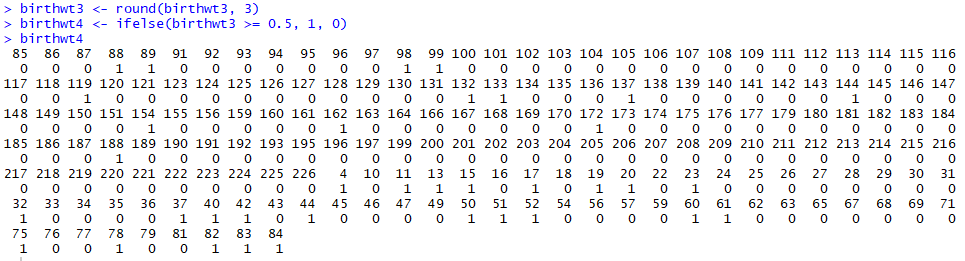


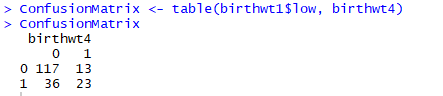
**Misclassification rate = 49/189 = 0.259259 🡪 25.9%**

**Steps to Generate the Confusion Matrix:**





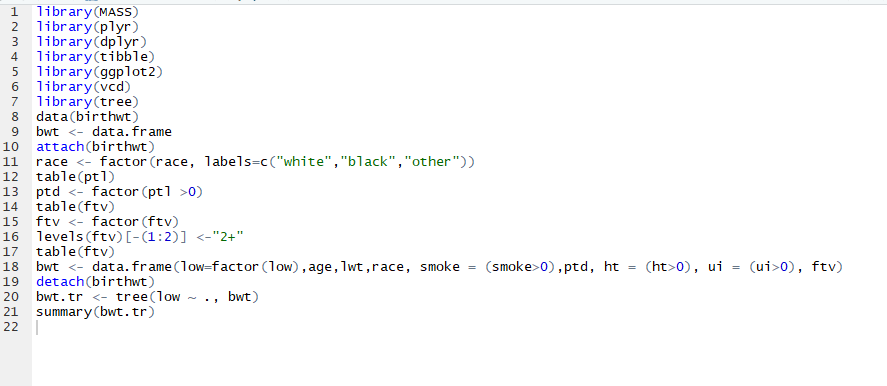




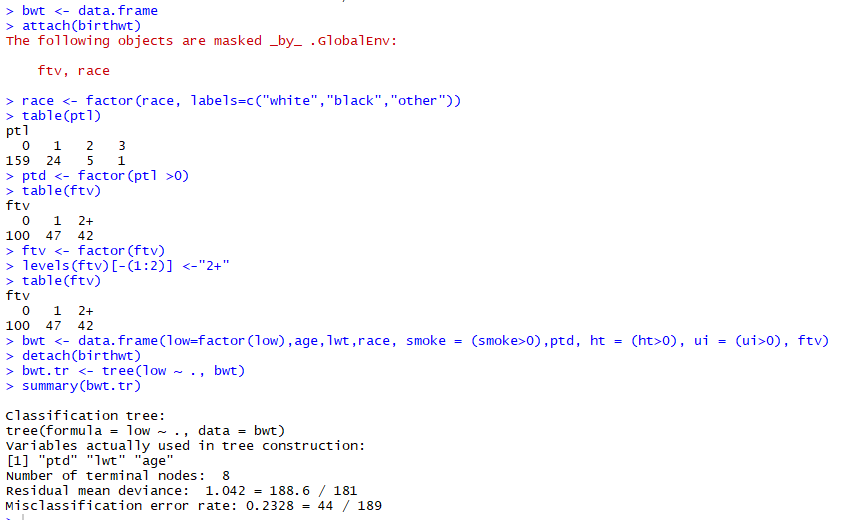
**Misclassification rate:**

Another way to calculate misclassification rate is

Using following code which uses tree library



Low birth weight miss classification rate script output



Misclassification error rate comes out to be 23.28% earlier it was calculated around 25.9%

**11. Summarize your findings in a few lines.**

* **Confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
* **Odds ratio** can be used to estimate the relative risk in a case-control study. Calculating a confidence interval provides you with an indication of how reliable your odds ratio is (the wider the interval, the greater the uncertainty associated with your estimate).
* **Model validation** is possibly most important step in model building sequence. It is also overlooked.
* **Logistic regression** is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
* Through this exercise found that **miscalculation rate of low birth rate** is around 23% to 25% depending on the flow of the script or the approach used to calculate.
* **Removing predictor variables from a model will almost always make the model fit less well** (i.e. a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant.
* Could there be a better model to depict the above data? Can we improve the model to fit the data better? These are some questions we can probably need to answer in future.