Project 3 - STAT 3022

Henrique Lispector *ID: 4839516 lispe001@umn.edu*

Due date: April 22nd, 2016

Project Description

Analyse the date birthwt from the library MASS. Load the library MASS. By using help file for the data, help(birthwt) first figure out what each variable means. Ignore the variable bwt. You goal is to understand whether all the remaining variables can determine the birth of an underweight infant. Fit a logistic regression model Determine what is the response variable and which ones are the covariates. Find out which should be considered as categorical and which ones as quantitative. See if they are correctly coded in R. Do all the initial EDA and data pre processing necessary before the model fit. After fitting the model, determine which covariates are significant and which ones are not. You do not have to use any model selection(covariate selection) technique for this problem.

Dataset Description

This data frame contains the following variables:

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.

Since the goal of this project is to understand whether birth of an underweight infant can be determined and explained by all the other variables, "low" is our response variable while all the other variables are the covariates (predictors).

Loading the data

```
library(MASS)
data("birthwt")
head(birthwt) #Views first rows of dataset
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

Step 1 - Data Pre-Processing

To see the data types of the variables in the dataframe, we use the following command:

```
lapply(birthwt, class)
```

```
## $low
## [1] "integer"
##
## $age
## [1] "integer"
##
## $lwt
## [1] "integer"
##
## $race
## [1] "integer"
##
## $smoke
## [1] "integer"
##
## $ptl
## [1] "integer"
##
## $ht
## [1] "integer"
##
## $ui
## [1] "integer"
##
## $ftv
## [1] "integer"
##
## $bwt
## [1] "integer"
```

All the variables are classified by default with the integer data type. However, there are many categorical variables, such as "race", "smoke", "ptl", "ht", "ui", and "ftv". Therefore, their data types need to be changed to categorical. Even though the response variable is categorical, we do not convert it to factor, otherwise this model will not work. Also, since the description mentions that the variable "bwt" should not be considered, "btw" will not be included in the adjusted birthwt dataset.

```
adj_birthwt <- with(birthwt, data.frame(data.frame(low=low, age=age, lwt=lwt, race=as.factor(race),
                                                   smoke=as.factor(smoke), ptl=as.factor(ptl),
                                                   ht=as.factor(ht), ui=as.factor(ui),
                                                   ftv=as.factor(ftv))))
lapply(adj_birthwt, class)
```

```
## $low
## [1] "integer"
##
## $age
## [1] "integer"
##
## $lwt
## [1] "integer"
##
## $race
## [1] "factor"
##
## $smoke
## [1] "factor"
##
## $ptl
## [1] "factor"
##
## $ht
## [1] "factor"
##
## $ui
## [1] "factor"
##
## $ftv
## [1] "factor"
```

With the new "adj_birthwt" adjusted dataset, we can move on to the exploratory data analysis.

Step 3: Exploratory Data Analysis

Let us do a quick numerical summary of the dataset:

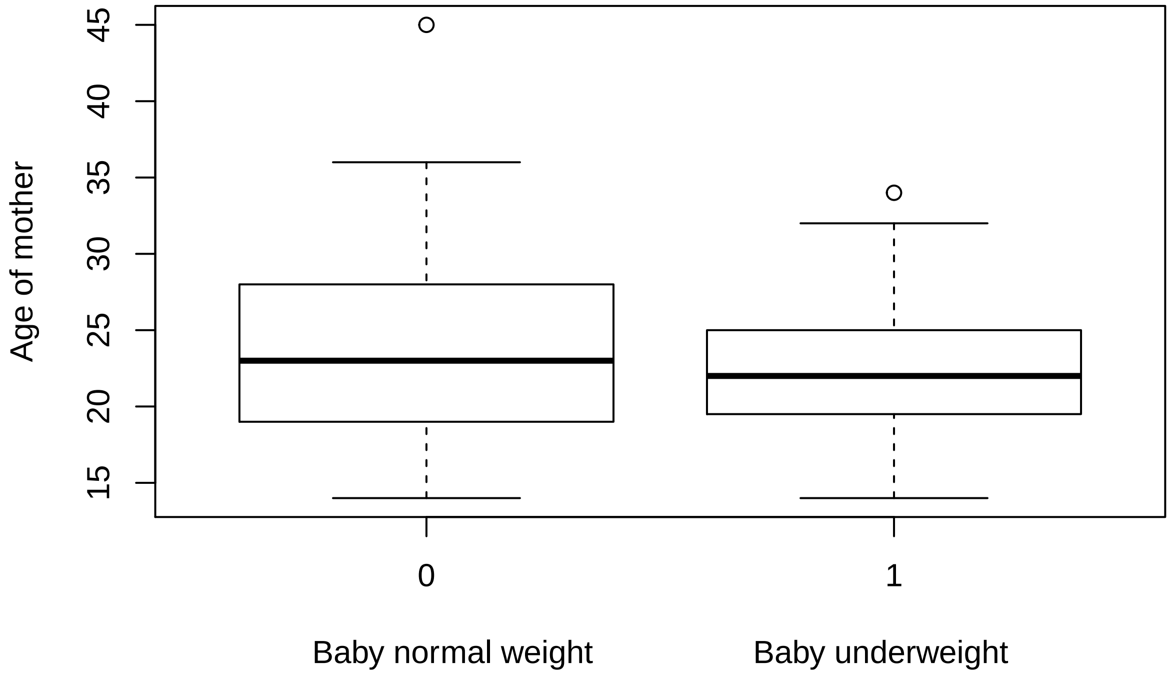
```
summary(adj_birthwt)
```

```
##       low              age             lwt        race   smoke   ptl
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   1:96   0:115   0:159
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   2:26   1: 74   1: 24
##  Median :0.0000   Median :23.00   Median :121.0   3:67           2:  5
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8                  3:  1
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0
##  ht      ui      ftv
##  0:177   0:161   0:100
##  1: 12   1: 28   1: 47
##                  2: 30
##                  3:  7
##                  4:  4
##                  6:  1
```

The summary of the categorical variables return each category name followed by the number of units in that category. Even though the response variable is categorical, can be interpreted from its summary result is that there are more zeros than ones, since the mean is less than 0.5, which indicates that the cases in which a child is born underweight occur less than the opposite.

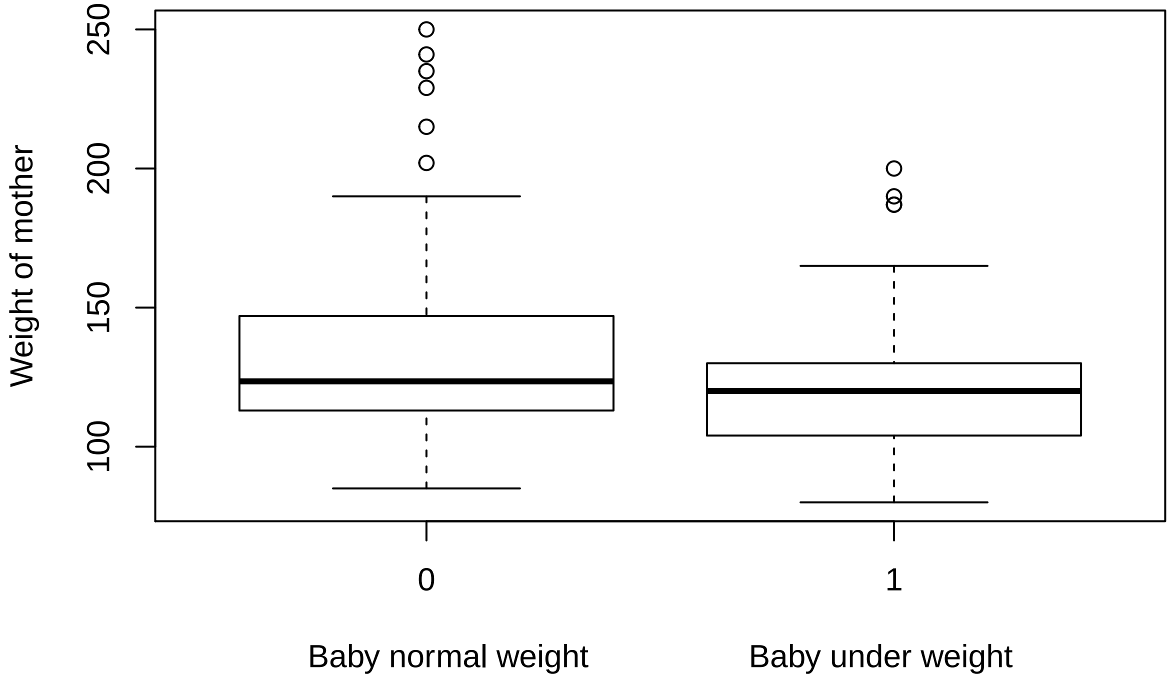
Let us graph boxplots of the quantitative variables to get a visual look of them:

```
boxplot(age~low, data = adj_birthwt, ylab="Age of mother",
        xlab="Baby normal weight                  Baby underweight")
```



The box plot above shows that when the baby is under weight, the median of mother's age is slightly lower than when the baby is born at a normal weight. We can also notice that babies with normal weights come from a wider range of older mothers relative to the mothers of underweight babies.

```
boxplot(lwt~low, data = adj_birthwt, ylab="Weight of mother", xlab="Baby normal weight
```



The box plot above shows that when the baby is under weight, the median of mother's weight is lower than when the baby is born at a normal weight. We can also notice that babies with normal weights come from a range of mothers with more weight relative to the mothers of underweight babies.

In order to find the average mother's age of underweight babies vs. the average mother's age of normal weight babies we use the tapply function.

```
with(adj_birthwt, tapply(age,low, mean))
```

```
##        0        1
## 23.66154 22.30508
```

This means that the average age of mothers of normal weight babies is 23.661, while the average age of mothers of underweight babies is of 22.305.

Similarly, to find the average mother's weight of underweight babies vs. the average of mother's weight of normal weight babies, we do the following:

```
with(adj_birthwt, tapply(lwt,low, mean))
```

```
##        0        1
## 133.3000 122.1356
```

The output tells us that mothers of underweight babies are around 11 pounds skinnier than the mothers of normal weight babies.

For all the categorical variables, the following function is used:

```
with(adj_birthwt, table(low, race))
```

```
##    race
## low  1  2  3
##   0 73 15 42
##   1 23 11 25
```

It is possible to notice from the table above that the odds of a baby being normal weight if their mother is white is around 3:1, while with a black mother the odds is close to 1:1. Which suggests that race might be an important factor, in this dataset, when it comes to understanding the birth of underweight infants.

```
with(adj_birthwt, table(low, smoke))
```

```
##    smoke
## low  0  1
##   0 86 44
##   1 29 30
```

The table above shows that the odds of having an underweight infant is much higher for a smoking mother than for a non-smoking mother. This also suggests that this category might be significant for our response variable.

```
with(adj_birthwt, table(low, ptl))
```

```
##    ptl
## low   0   1   2   3
##   0 118   8   3   1
##   1  41  16   2   0
```

The main insight from this table is that the odds of a child being born underweight increase tremendouosly if the mother has had one previous premature labour, compared to none. Which also suggests that this factor might be significant a significant predictor of an underweight child.

```
with(adj_birthwt, table(low, ht))
```

```
##    ht
## low   0   1
##   0 125   5
##   1  52   7
```

Even though the sample size of hypertension mothers is small, this table shows that the odds of having an underweight child increases if the mother has hypertension.

```
with(adj_birthwt, table(low, ui))
```

```
##    ui
## low   0   1
##   0 116  14
##   1  45  14
```

This table demonstrates that the odds of having an underweight infant increases if the mother has uterine irritability.

```
with(adj_birthwt, table(low, ftv))
```

```
##    ftv
## low  0  1  2  3  4  6
##   0 64 36 23  3  3  1
##   1 36 11  7  4  1  0
```

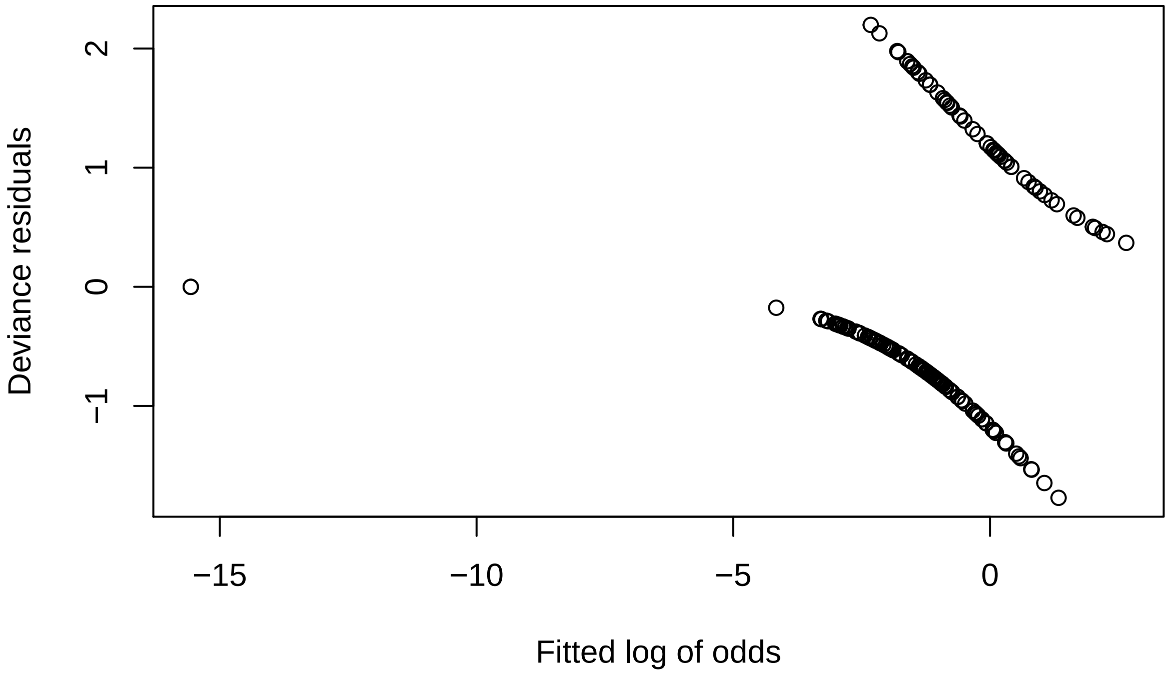
The table above reveals that the odds of having an underweight child do not change much with the increase of physician visits by a future mother during the first trimester.

Step 3: Fit the Model

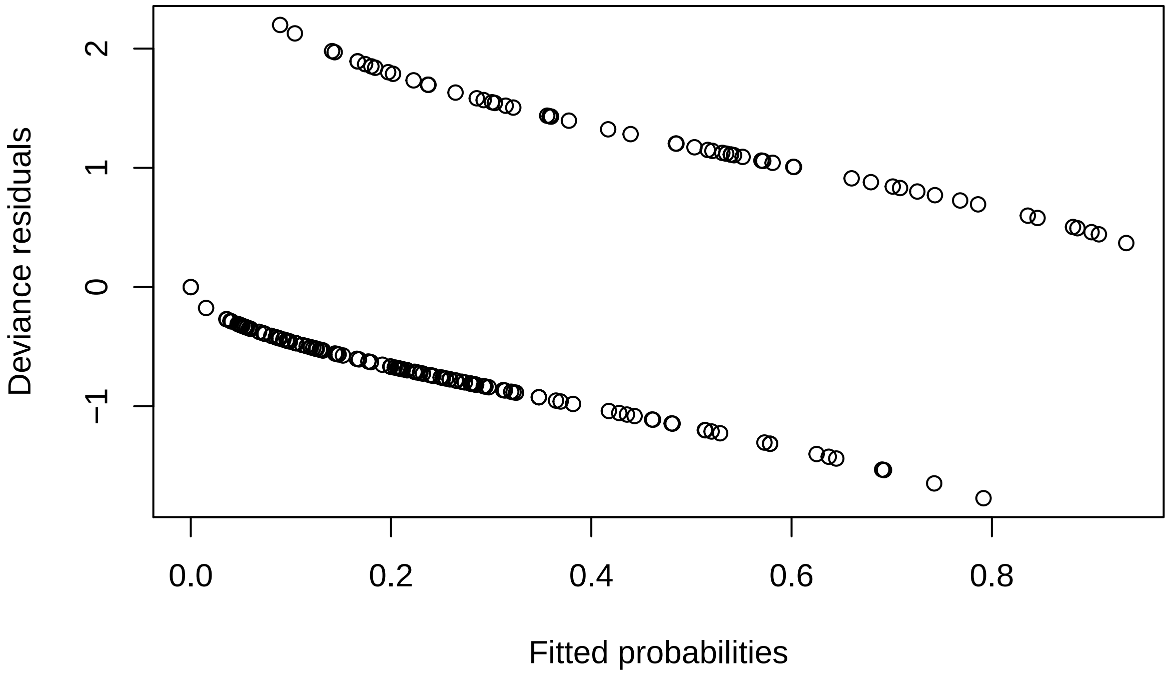
```
mod1 <- glm(low~age+lwt+race+smoke+ptl+ht+ui+ftv, data = adj_birthwt, family = binomial)
```

Model Diagnostics

```
plot(residuals(mod1) ~ predict(mod1, type="link"), xlab="Fitted log of odds",
                                                    ylab="Deviance residuals")
```



```
plot(residuals(mod1) ~ predict(mod1, type="response"), xlab="Fitted probabilities",
     ylab="Deviance residuals")
```



For both of the plots above, we are looking for irregularities, such as if any point is sticking out of the usual pattern. Since the model assumptions of the General Linear Model (GLM) are over the scope of this class, we are essentially looking for wheter the model fits all individual data points uniformly.

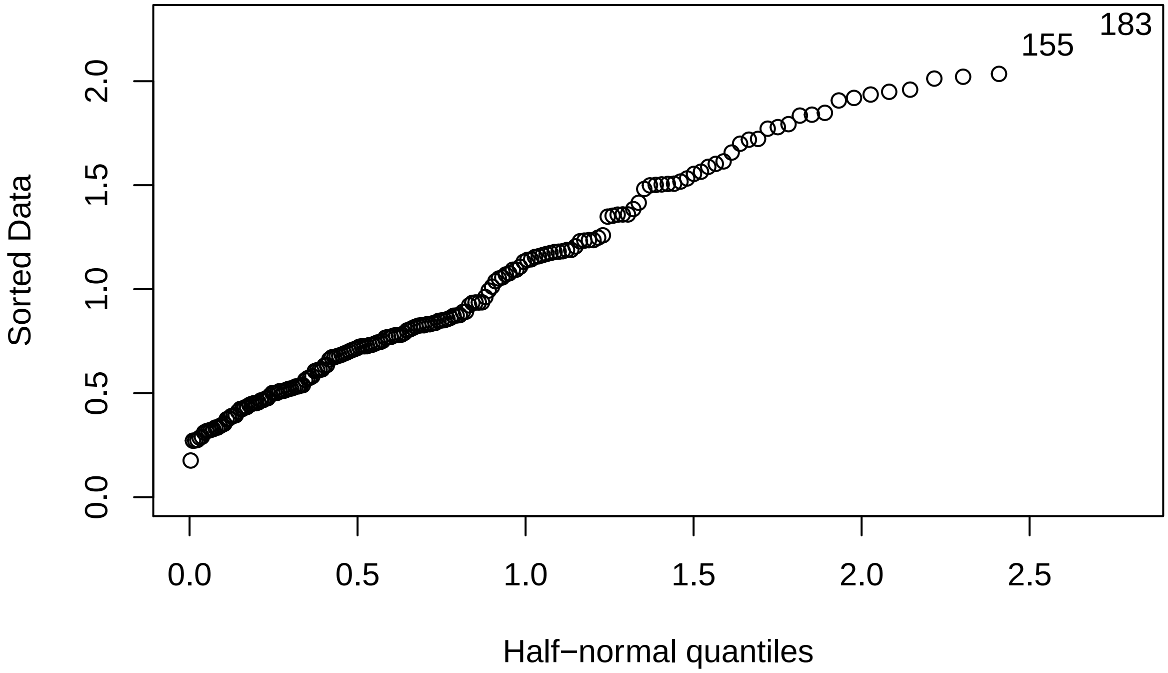
By looking at both plots, there does not seem to be a point standing out from the uniform patterns. So the model looks good so far.

We proceed to plot a half normal plot, which should be examined similarly as the examination of the Normal Q-Q Plot. The difference is that a half normal plot looks for outliers and not whether errors are normally distributed. It is important to note that there is no normality assumption for logistic regression.

```
require(faraway)
```

```
## Loading required package: faraway
```

```
halfnorm(rstudent(mod1))
```



The Half-normal plot looks okay, but theoretically could be better. There are some jumps between sequences of points. There are also a few points sticking out at the top, but this is common in this type of plot. Overall, the plot is not bad. Now we move on to the conclusion.

Conclusion

```
summary(mod1)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, family = binomial, data = adj_birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7713  -0.7839  -0.4859   0.8305   2.1983
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)   
## (Intercept)  8.986e-01  1.299e+00   0.692  0.48908   
## age         -3.889e-02  3.952e-02  -0.984  0.32508   
## lwt         -1.536e-02  7.483e-03  -2.052  0.04017 * 
## race2        1.110e+00  5.509e-01   2.016  0.04382 * 
## race3        6.698e-01  4.774e-01   1.403  0.16061   
## smoke1       7.073e-01  4.399e-01   1.608  0.10784   
## ptl1         1.865e+00  5.722e-01   3.259  0.00112 **
## ptl2         4.875e-01  1.008e+00   0.484  0.62858   
## ptl3        -1.553e+01  1.455e+03  -0.011  0.99148   
## ht1          1.805e+00  7.488e-01   2.410  0.01595 * 
## ui1          7.930e-01  4.780e-01   1.659  0.09712 . 
## ftv1        -5.598e-01  4.958e-01  -1.129  0.25879   
## ftv2        -8.141e-02  5.447e-01  -0.149  0.88120   
## ftv3         1.103e+00  8.648e-01   1.276  0.20213   
## ftv4        -9.225e-01  1.384e+00  -0.667  0.50496   
## ftv6        -1.291e+01  1.455e+03  -0.009  0.99292   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 188.60  on 173  degrees of freedom
## AIC: 220.6
##
## Number of Fisher Scoring iterations: 14
```

There are five coefficients that are significantly different from zero, and therefore statistically significant to understanding the birth of underweight infants. Among these coefficients, the predictor “lwt” is quantitative, while the other four (“race2”, “ptl1”, “ht1”, “ui1”) are categorical.

Interpretation of coefficient for “lwt”:

Suppose “lwt”" = x, and everything else is constant.

- log(p1/(1-p1)) = 0.8986 - 0.01536x + 1.11race2 + 1.865ptl1 + 1.805ht1 + 0.79ui1

Suppose now that “lwt” = x + 1, and everything else is constant.

- log(p2/(1-p2)) = 0.8986 - 0.01536(x+1) + 1.11race2 + 1.865ptl1 + 1.805ht1 + 0.79ui1

Consider the following:

- p1: probability of infant being underweight at lwt = x
- p2: probability of infant being underweight at lwt = x + 1

log p2/(1-p2) - log p1/(1-p1) = -0.01536
=> (p2/(1-p2))/(p1/(1-p1)) = exp(-0.01536) = 0.98476

p2/(1-p2) = 0.98476*(p1/(1-p1))

Therefore, when “lwt” increases by 1 unit, the odds of an infant being underweight becomes 0.98476 than before, everything else remaining the same.

Interpretation of coefficient for “race2” (categorical predictor)

We first note that race1 is the baseline for comparison in this category. Consider the following:

- p1: probability of infant being underweight from race 1.
- p2: probability of infant being underweight from race 2.

Everything else remaining the same, we have:

log p2/(1-p2) - log p1/(1-p1) = coeff. of race2 = 1.11

log((p2/(1-p2))/(p1/(1-p1))) = 1.11 => (p2/(1-p2))/(p1/(1-p1)) = exp(1.11) = 3.03458

=> p2/(1-p2) = 3.03458*(p1/(1-p1))

Therefore, the odds of an infant being underweight from race2 is 3 times higher than from race 1, everything else remaining the same.

Interpretation of coefficient for “ptl1” (categorical predictor)

We first note that ptl0 is the baseline for comparison in this category. Consider the following:

- p1: probability of infant being underweight from ptl0.
- p2: probability of infant being underweight from ptl1.

Everything else remaining the same, we have:

log p2/(1-p2) - log p1/(1-p1) = coeff. of ptl1 = 1.865

log((p2/(1-p2))/(p1/(1-p1))) = 1.865 => (p2/(1-p2))/(p1/(1-p1)) = exp(1.865) = 6.455936

=> p2/(1-p2) = 6.455936*(p1/(1-p1))

Therefore, the odds of an infant being underweight from a mother that had 1 previous premature labour is nearly 6.5 times higher than from a mother with no previous premature labours, everything else remaining the same.

Interpretation of coefficient for “ht1” (categorical predictor)

We first note that ht0 is the baseline for comparison in this category. Consider the following:

- p1: probability of infant being underweight from ht0.
- p2: probability of infant being underweight from ht1.

Everything else remaining the same, we have:

log p2/(1-p2) - log p1/(1-p1) = coeff. of ht1 = 1.805

log((p2/(1-p2))/(p1/(1-p1))) = 1.805 => (p2/(1-p2))/(p1/(1-p1)) = exp(1.805) = 6.079971

=> p2/(1-p2) = 6.079971*(p1/(1-p1))

Therefore, the odds of an infant being underweight from a mother with a history of hypertension is about 6 times higher than from a mother without hypertension history, everything else remaining the same.

Interpretation of coefficient for "ui1" (categorical predictor)

We first note that ui0 is the baseline for comparison in this category. Consider the following:

- p1: probability of infant being underweight from ui0.
- p2: probability of infant being underweight from ui1.

Everything else remaining the same, we have:

log p2/(1-p2) - log p1/(1-p1) = coeff. of ui1 = 0.793

log((p2/(1-p2))/(p1/(1-p1))) = 0.793 => (p2/(1-p2))/(p1/(1-p1)) = exp(0.793) = 2.210017

=> p2/(1-p2) = 2.210017*(p1/(1-p1))

Therefore, the odds of an infant being underweight from a mother with uterine irritability is 2.2 times higher than from a mother without uterine irritability, everything else remaining the same.

Wald Tests

For categorical attributes with more than two categories, we can also check if the attribute is significant or not to determining if an infant is underweight.

```r
require(aod) #Package required for doing wald test
```

```
## Loading required package: aod
```

```
##
## Attaching package: 'aod'
```

```
## The following objects are masked from 'package:faraway':
##
##     rats, salmonella
```

```r
#Testing if race is significant
wald.test(b = coef(mod1), Sigma = vcov(mod1), Terms = 4:5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 4.5, df = 2, P(> X2) = 0.1
```

```
#Terms=4:5 refer to rows 4 & 5 from summary(mod1) output
```

In this case, Wald's test tells us that Race is not a significant attribute in determining if an infant is underweight because the p-value = 0.1, which is greater than 0.05 . This is contradictory with the result from the logistic regression model, but this does not make the logistic regression result or interpretation for Race 2 wrong.

```
#Testing if number of premature labours is significant
wald.test(b = coef(mod1), Sigma = vcov(mod1), Terms = 7:9)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 10.6, df = 3, P(> X2) = 0.014
```

```
#Terms=7:9 refer to rows 7 & 9 from summary(mod1) output
```

From the result above, Wald's test tells us that the number of premature labours is significant in determining if an infant is underweight or not, as the p-value < 0.05 .

```
#Testing if number of physician visits during first trimester is significant
wald.test(b = coef(mod1), Sigma = vcov(mod1), Terms = 12:16)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 3.6, df = 5, P(> X2) = 0.61
```

```
#Terms=12:16 refer to rows 12 & 16 from summary(mod1) output
```

Finally, in this case, Wald's test tells us that the number of pysician visits during the first trimester of pregnancy is not significant in determining if the infant is udnerweight or not, as p-value $= 0.61 > 0.05$.