Project 3

Bailey Perry

April 22, 2016

Introduction:

The goal of this project is to analyze the data birthwt from the library MASS, and to understand whether all of the variables, aside from bwt, can determine the birth of an underweight infant. This will be determined by using the logistic regression model.

Import Dataset and Summarize Data

```
require(faraway)
```

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

```
## Warning: package 'faraway' was built under R version 3.2.4
```

```
require(MASS)
```

```
## Loading required package: MASS
```

```
data(birthwt)
help(birthwt)
```

```
## starting httpd help server ...
```

```
##  done
```

The help command allows the user to access the names of all of the variables along with sources, and useful references.

Based on this, the 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.

Step 1: Data Pre-Processing

```
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"
```

Since all of the variables are currently categorized as integers, it is necessary to review their meaning and determine any necessary corrections to the dataset.Upon review, it is necessary to change the variables: race, smoke, ht, and ui to be recognized as a factor in R.

```
bwt_adj <- with(birthwt, data.frame(data.frame(low=low, age=age, lwt=lwt, race=factor(race, labels = c(

lapply(bwt_adj, class)
```

```
## $low
## [1] "integer"
##
## $age
## [1] "integer"
##
## $lwt
## [1] "integer"
##
```

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

From the code above, the variables of class 'factor' are: race, smoke, ht, and ui as intended. The variables age, lwt, pt, and ftv are classified as integers. Since the response variable is low and it is coded as an integer, although it is technically a categorical variable, we do not change it to factor, as demonstrated in class. Now the data matches the intentions of the project, so we can use them in our analysis and formulas correctly.

Step 2: Exploratory Data Analysis

```
summary(bwt_adj)
```

```
##       low              age              lwt           race    smoke
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   white:96   0:115
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   black:26   1: 74
##  Median :0.0000   Median :23.00   Median :121.0   other:67
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0
##       ptl         ht      ui           ftv
##  Min.   :0.0000   0:177   0:161   Min.   :0.0000
##  1st Qu.:0.0000   1: 12   1: 28   1st Qu.:0.0000
##  Median :0.0000                   Median :0.0000
##  Mean   :0.1958                   Mean   :0.7937
##  3rd Qu.:0.0000                   3rd Qu.:1.0000
##  Max.   :3.0000                   Max.   :6.0000
```

```
#bwt_adj
```

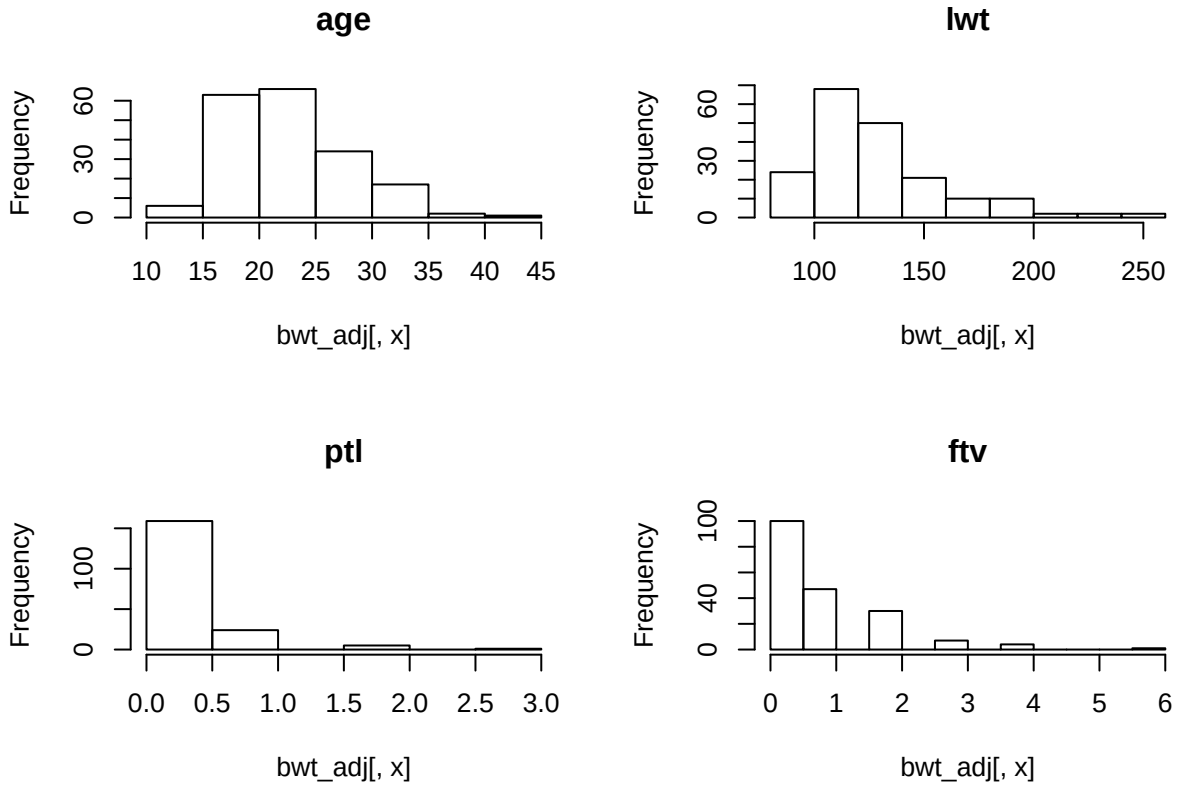
Based on the summary output, the user can see the important numerical values for each of the quantitative variables, and can also get insight to the categorical variables.

Low is the response variable we will be looking at to understand the effect of the other variables on determining whether a baby will be underweight or not at birth. The smoke variable summary tells us that 74 women smoked during pregnancy and 115 women did not. The analysis for history of hypertension (ht) and presence

of uterine irritability (ui) have the same interpretation with their respective summary outputs. All three of these are binary response variables.

Calling the data to review it in a table form that showcases all of the variables would be useful for a smaller dataset, but it is a very large set of 189 rows with 10 columns, so finding patterns is much more difficult and tedious. So, this will be done in later portions of the analysis.

```
x <- c(2,3,6,9)
par(mfrow= c(2,2))
lapply(x, function(x){hist(bwt_adj[,x], main = colnames(bwt_adj)[x])})
```




```
## [[1]]
## $breaks
## [1] 10 15 20 25 30 35 40 45
##
## $counts
## [1]  6 63 66 34 17  2  1
##
## $density
## [1] 0.006349206 0.066666667 0.069841270 0.035978836 0.017989418 0.002116402
## [7] 0.001058201
##
## $mids
## [1] 12.5 17.5 22.5 27.5 32.5 37.5 42.5
##
## $xname
```

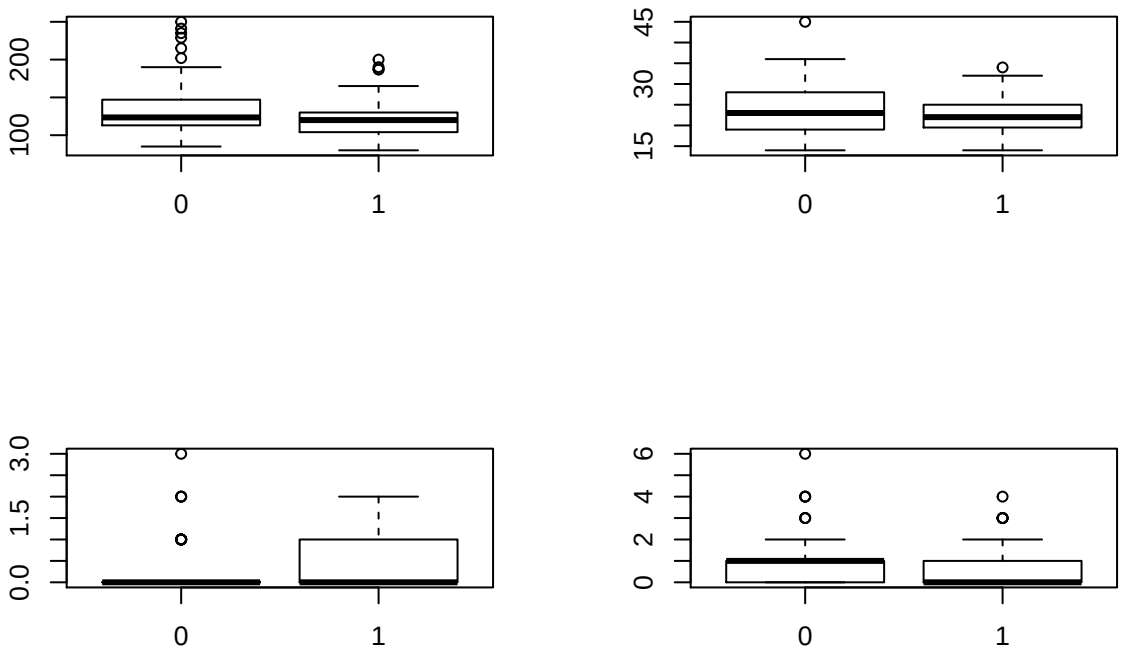
```
## [1] "bwt_adj[, x]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## [[2]]
## $breaks
##  [1]  80 100 120 140 160 180 200 220 240 260
##
## $counts
## [1] 24 68 50 21 10 10  2  2  2
##
## $density
## [1] 0.0063492063 0.0179894180 0.0132275132 0.0055555556 0.0026455026
## [6] 0.0026455026 0.0005291005 0.0005291005 0.0005291005
##
## $mids
## [1]  90 110 130 150 170 190 210 230 250
##
## $xname
## [1] "bwt_adj[, x]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## [[3]]
## $breaks
## [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0
##
## $counts
## [1] 159  24   0   5   0   1
##
## $density
## [1] 1.68253968 0.25396825 0.00000000 0.05291005 0.00000000 0.01058201
##
## $mids
## [1] 0.25 0.75 1.25 1.75 2.25 2.75
##
## $xname
## [1] "bwt_adj[, x]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## [[4]]
```

```
## $breaks
##  [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
##
## $counts
##  [1] 100  47   0  30   0   7   0   4   0   0   0   1
##
## $density
##  [1] 1.05820106 0.49735450 0.00000000 0.31746032 0.00000000 0.07407407
##  [7] 0.00000000 0.04232804 0.00000000 0.00000000 0.00000000 0.01058201
##
## $mids
##  [1] 0.25 0.75 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75
##
## $xname
## [1] "bwt_adj[, x]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

The histogram function is only applied to the quantitative variables in the set, while the command would not work for the categorical variables (produces an error). The function does not tell us anything about the relationship with the response variable, but it is useful for getting a visual of the variability within each of the covariates. Based on the histogram output, it is clear that ptl and ftv have the most skew, and are skewed to the right. The histograms for age and lwt do not seem to be extreme violations of symmetry, but also do not match it perfectly.

The following commands create boxplots between the low variable and the quantitative variables in the dataset: lwt, age, ptl, and ftv.

```
par(mfrow=c(2,2))
boxplot(lwt~low, data=bwt_adj)
boxplot(age~low, data= bwt_adj)
boxplot(ptl~low, data=bwt_adj)
boxplot(ftv~low, data=bwt_adj)
```



Based on the boxplots, some general trends (on average) include: Women with the indicator for low birth weight have a slightly lower median weight in pounds at last menstrual period. Additonally, there is less variability in weight in pounds at last menstrual period for the mothers with the low birth weight indicator.

Similar to above, the women with the indicator for low birth weight had a similar median age as the women without the indicator for low birth weight. The women with the indicator had slightly less overall variability in age, and they also had a smaller interquartile range (IQR) which means that 50% of the women with the indicator were approximately between ages 20-25. On the other hand the IQR for the women without the indicator showed that they were approximately between ages 18-28.

The boxplot for the number of previous premature labours shows that women with the indicator for low birth weight had history of previous premature labours since there are 25% of the women with values above zero (based on the quartile). Comparatively, women without the indicator did not show a history of previous premature labors, other than a few outliers.

The boxplot with regard to the number of physician visits during the first trimester shows that women without the indicator for low birth weight had more physician visits in the first trimester (on average) than the women with the indicator.

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

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

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

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

```
with(bwt_adj, tapply(ptl,low,mean))
```

```
##         0         1
## 0.1307692 0.3389831
```

```
with(bwt_adj, tapply(ftv,low,mean))
```

```
##         0         1
## 0.8384615 0.6949153
```

This output gives the respective mean values for each of the quantitative variables and the values corresponding to women with the indicator (1) and women without the indicator (0). These values agree with the statements made above.

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

```
##    race
## low white black other
##   0    73    15    42
##   1    23    11    25
```

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

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

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

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

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

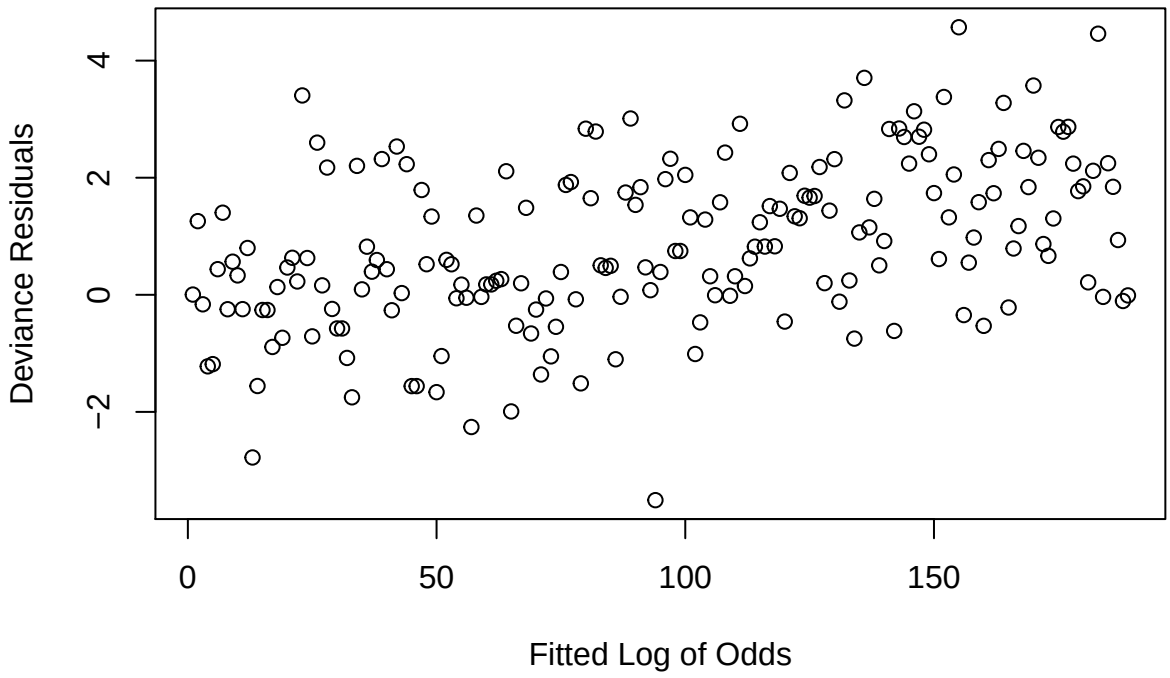
```
##    ui
## low   0   1
##   0 116  14
##   1  45  14
```

This output gives the respective numbers of women for each of the categorical variables (0 for without the variable (ie no smoking) and 1 for with the variable), paired with the values corresponding to women with the indicator (1) and women without the indicator (0). It creates a table to show where the women fall for each of the four categories, for example: no smoking, no indicator; smoking, no indicator; no smoking, indicator; and smoking, indicator. This was performed for all of the variables and they follow the same set-up. It is a good way to see which categories have high numbers, and make hypotheses about which variables correspond to underweight babies.

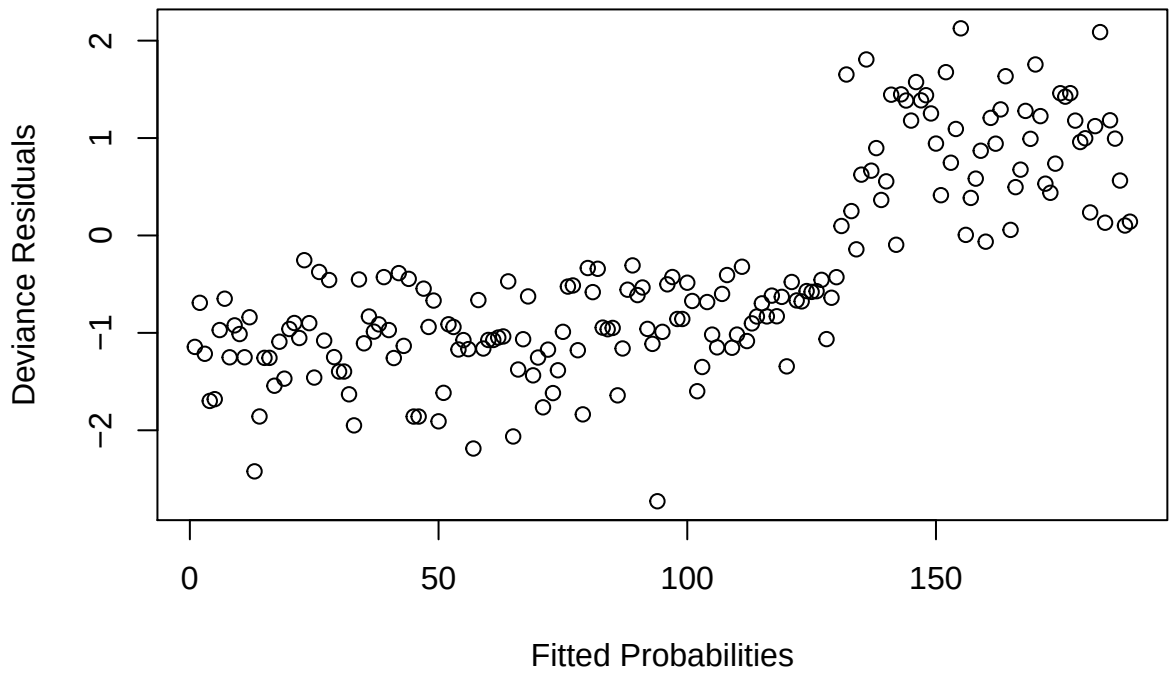
The overall trends suggest: White women are least likely to have an underweight baby and black women have the highest odds. The women who do not smoke are less likely to have underweight babies. The women who do not have hypertension seem to have much lower odds of having an underweight baby. And finally, the women who do not have uterine irritability are less likely to have underweight babies.

Step 3: Model Fitting and Diagnostics

```
logmod1 <- glm(low~age+lwt+race+smoke+ptl+ht+ui+ftv, data = bwt_adj, family=binomial)
par(mfrow = c(1,1))
plot(residuals(logmod1)-predict(logmod1, type="link"), xlab = "Fitted Log of Odds", ylab = "Deviance Re
```



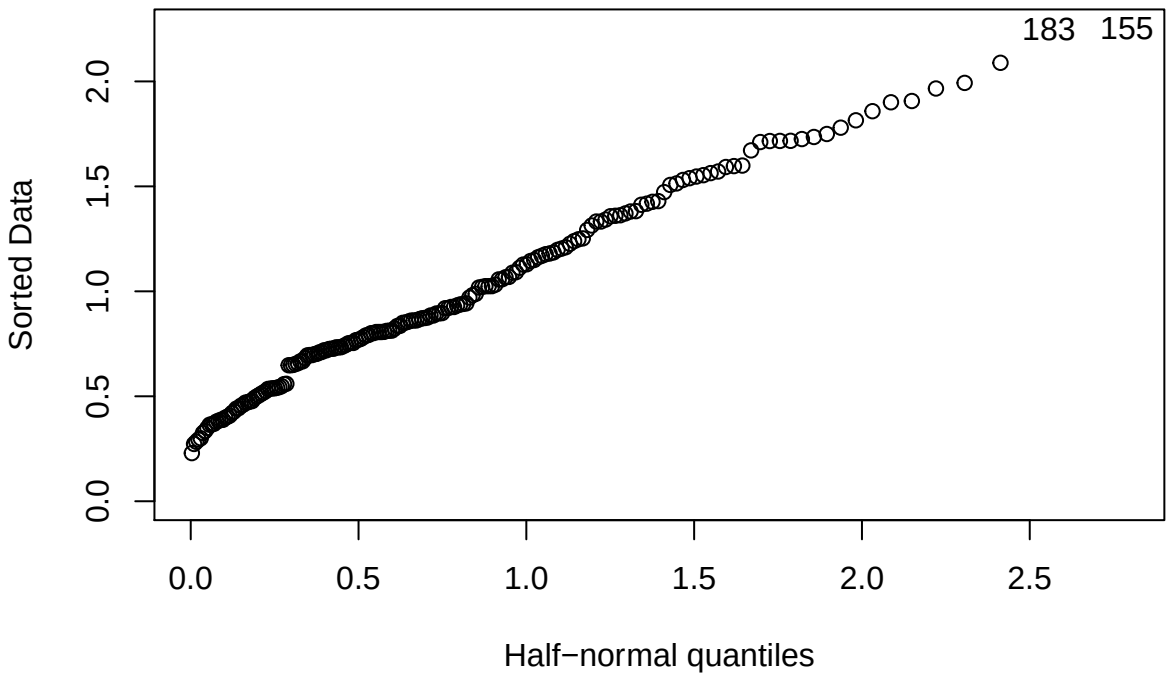

```
plot(residuals(logmod1)-predict(logmod1, type="response"), xlab = "Fitted Probabilities", ylab = "Devia
```




```
require(faraway)
```

In both plots above, the goal is to look for and identify irregularities, such as any point sticking out from the normal pattern of the plot. Since the model assumptions of the GLM are too complicated to verify graphically, the user is essentially looking for whether the model fits all individual points uniformly well.

```
halfnorm(rstudent(logmod1))
```



The half normal plot should be examined and interpreted similary to the Normal QQ Plot. The difference is that the half normal plot looks for outliers and not whether errors are normally distributed. There is no normality assumption for logistic regression analysis. As seen above, the plot identifies two points as potential outliers, point 183 and 155. These points are found in the upper right corner of the plot, as all of these types of points are when looking at the half normal plot.

Step 4: Conclusion

```
summary(logmod1)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, family = binomial, data = bwt_adj)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8946  -0.8212  -0.5316   0.9818   2.2125
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.480623   1.196888   0.402  0.68801
## age         -0.029549   0.037031  -0.798  0.42489
```

```
## lwt         -0.015424   0.006919  -2.229  0.02580 *
## raceblack    1.272260   0.527357   2.413  0.01584 *
## raceother    0.880496   0.440778   1.998  0.04576 *
## smoke1       0.938846   0.402147   2.335  0.01957 *
## ptl          0.543337   0.345403   1.573  0.11571
## ht1          1.863303   0.697533   2.671  0.00756 **
## ui1          0.767648   0.459318   1.671  0.09467 .
## ftv          0.065302   0.172394   0.379  0.70484
## ---
## 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: 201.28  on 179  degrees of freedom
## AIC: 221.28
##
## Number of Fisher Scoring iterations: 4
```

```
require(aod)
```

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

```
## Warning: package 'aod' was built under R version 3.2.5
```

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

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

```
wald.test(b = coef(logmod1), Sigma=vcov(logmod1), Terms = 4:5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.1, df = 2, P(> X2) = 0.028
```

Many of the coefficients are significantly different from zero. There was one quantitative variable that was significantly different from zero, and there were five categorical variables that were as well. The summary function allows the user to determine the significance and interpretation for the quantitative variables. For the categorical variable race, it is slightly more complicated, and Wald's test needs to be performed.

Coefficient Interpretation for Significant Covariates:

When the lwt variable increases by one unit (1 lb added to weight since last menstrual cycle), odds of having the indicator for a baby less than 2.5 kg (response variable, "low") becomes 0.98469 than before, everything else remaining the same.

First it is noted that for race, white is the baseline for comparison; for smoke, no smoking is the baseline for comparison; and for ht, no history of hypertension is the baseline for comparison. Also, it is noted that based on the summary output and the Wald test performed above, all of the categorical variables, except ui, are statistically significant and different from zero since the summary output and Wald test found p-values less than 0.05.

Race Interpretation: When the woman's race is black, the odds of having a baby that is underweight is 3.5689 that of a white woman having a baby that is underweight (assuming all else the same). When the woman's race is other, the odds of having a baby that is underweight is 2.412096 that of a white woman having a baby that is underweight (assuming all else the same).

Smoke Interpretation: When the woman is a smoker, the odds of having a baby that is underweight is 2.55703 that of a woman who is a non-smoker having a baby that is underweight (assuming all else the same).

Ht Interpretation: When the woman has a history of hypertension, the odds of having a baby that is underweight is 6.444989 that of a woman with no history of hypertension having a baby that is underweight (assuming all else the same).