Project3

Andreas Hochrein *ID: 4855928* *hochr007@umn.edu*

April 22, 2016

1. Preparation

```
require(MASS) # Load library MASS
```

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

```
summary(birthwt)  # Get information about dataset
```

```
##       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
```

```
# What does each variable mean?
help(birthwt)
```

Based on the variable definitions, I think the response variable in the logistics regression model that I will fit is the variable low since it is a count. Thus, the model will fit the relationship between the odds of low birthweight and the given 8 covariants (all the other variables besides bwt). This model could answer relevant questions since low birthweight is considered a medical issue and thus understanding the variables it correlates with, particularly related to behavior during the pregnancy, are of interest to many.

2. Data Preprocessing

```r
# Check the data types of the variables
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"
```

```r
# All variables are classified as quantitative right now. The variables low, race, smoke, ht, ui are al
# I will not include variable bwt in the processed dataset since we will not consider it.

birthwt_f <- with(birthwt, data.frame(low_f=as.factor(low), age=age, lwt=lwt, race_f=as.factor(race), sm

lapply(birthwt_f, class)
```

```
## $low_f
## [1] "factor"
##
## $age
## [1] "integer"
##
## $lwt
## [1] "integer"
##
## $race_f
## [1] "factor"
##
```

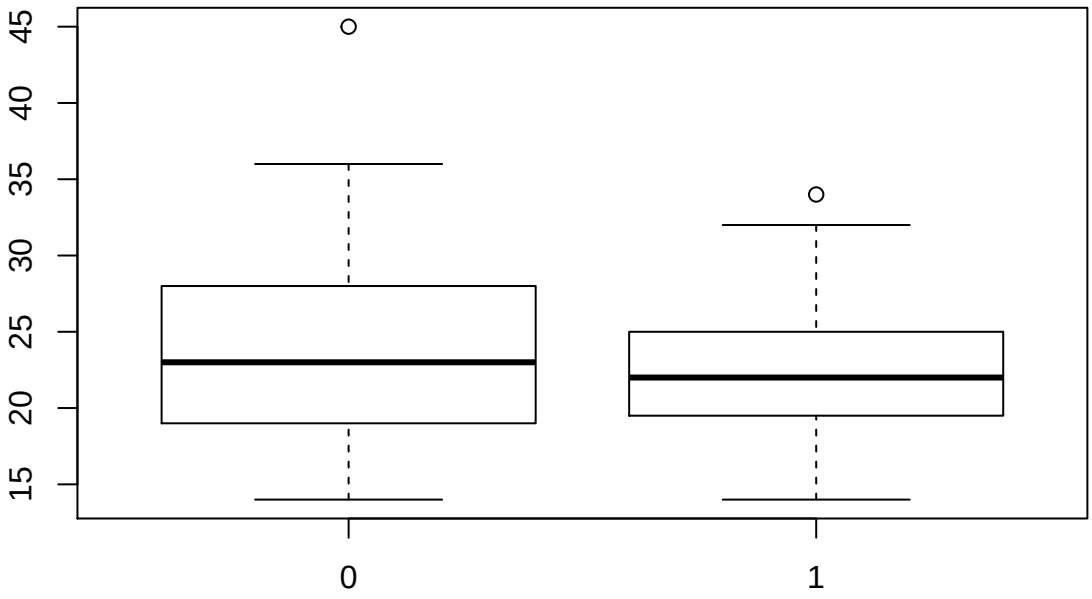
```
## $smoke_f
## [1] "factor"
##
## $ptl
## [1] "integer"
##
## $ht_f
## [1] "factor"
##
## $ui_f
## [1] "factor"
##
## $ftv
## [1] "integer"
```

3. Explanatory Data Analysis (EDA)

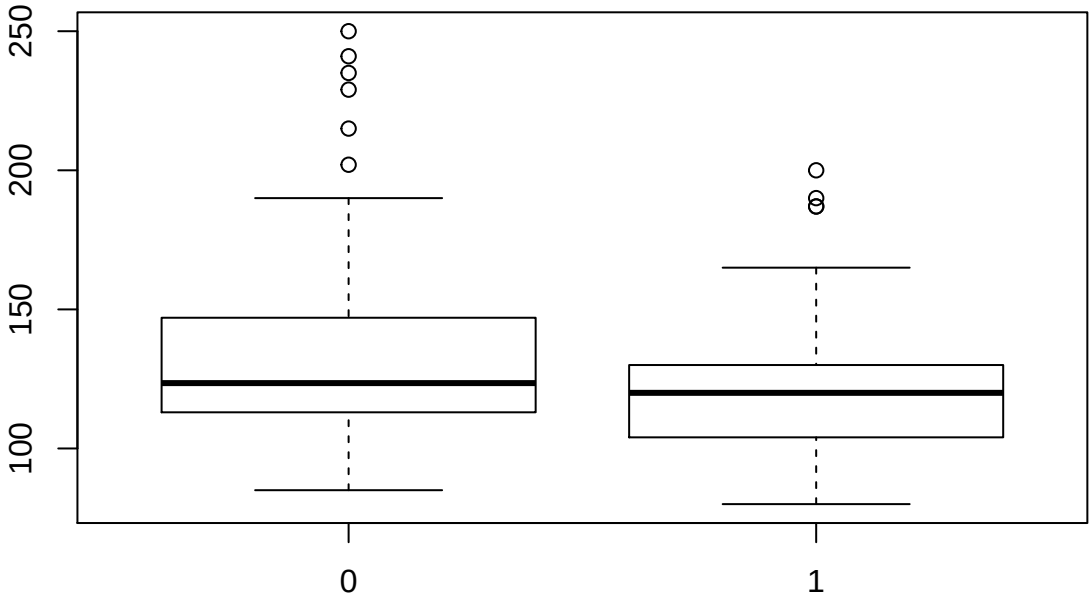
```
summary(birthwt_f)
```

```
##  low_f        age              lwt         race_f smoke_f      ptl
##  0:130   Min.   :14.00   Min.   : 80.0   1:96   0:115   Min.   :0.0000
##  1: 59   1st Qu.:19.00   1st Qu.:110.0   2:26   1: 74   1st Qu.:0.0000
##          Median :23.00   Median :121.0   3:67           Median :0.0000
##          Mean   :23.24   Mean   :129.8                  Mean   :0.1958
##          3rd Qu.:26.00   3rd Qu.:140.0                  3rd Qu.:0.0000
##          Max.   :45.00   Max.   :250.0                  Max.   :3.0000
##  ht_f    ui_f         ftv
##  0:177   0:161   Min.   :0.0000
##  1: 12   1: 28   1st Qu.:0.0000
##                  Median :0.0000
##                  Mean   :0.7937
##                  3rd Qu.:1.0000
##                  Max.   :6.0000
```

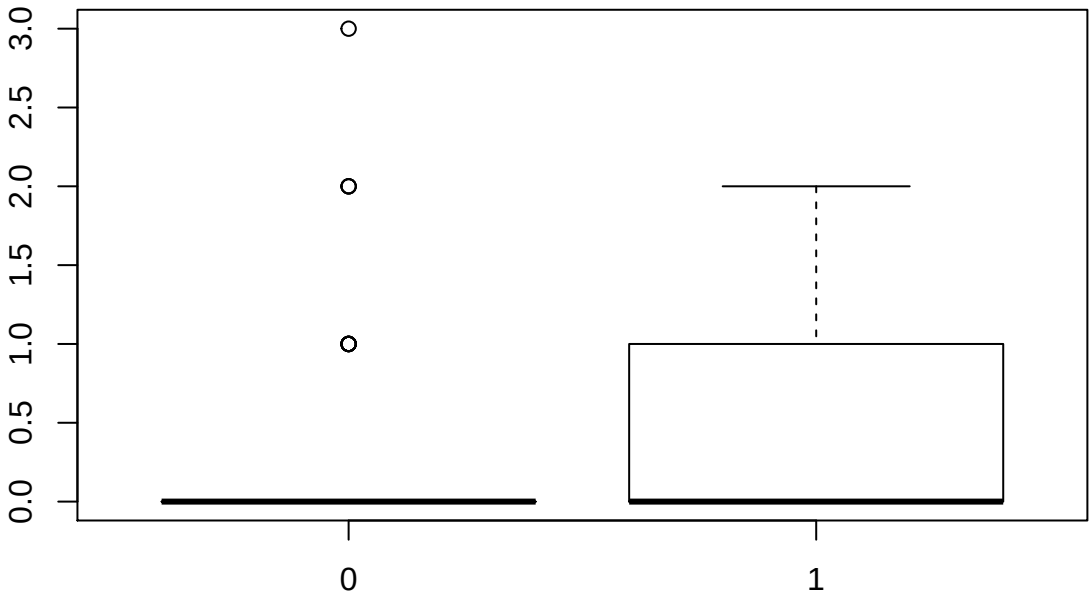
```
# Boxplots of individual quantitative covariants ~ response
boxplot(age ~ low_f, data=birthwt_f)
```



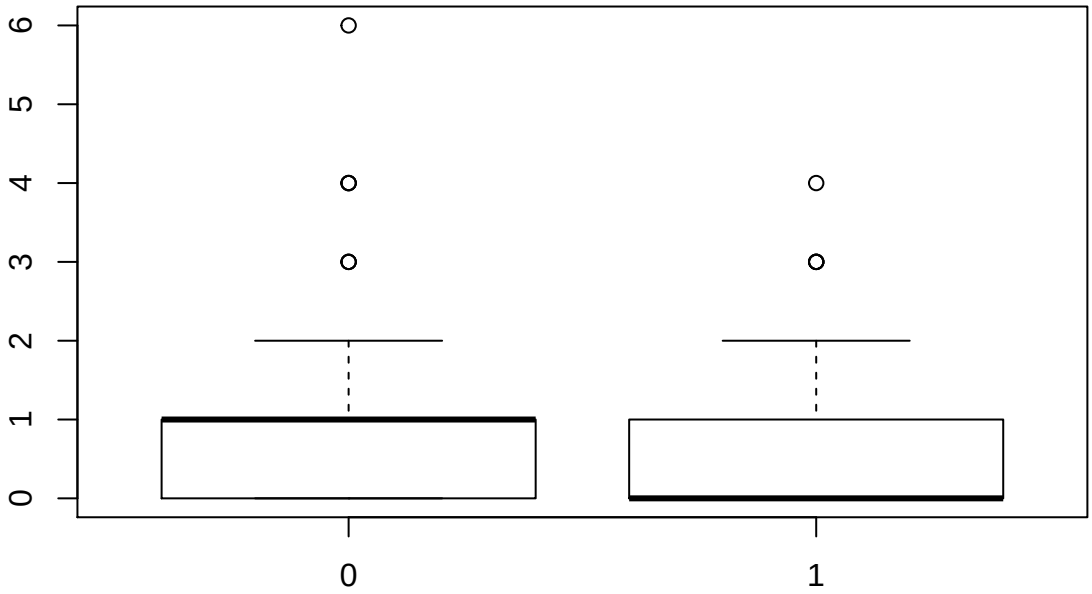

```
boxplot(lwt ~ low_f, data=birthwt_f)
```




```
boxplot(ptl ~ low_f, data=birthwt_f)
```




```
boxplot(ftv ~ low_f, data=birthwt_f)
```




```
# Means of covariants for low and normal weight at birth
with(birthwt_f, tapply(age,low_f,mean))
```

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

```
with(birthwt_f, tapply(lwt,low_f,mean))
```

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

```
with(birthwt_f, tapply(ptl,low_f,mean))
```

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

```
with(birthwt_f, tapply(ftv,low_f,mean))
```

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

The boxplots indicate that the mean age of mothers of low birthweight children is slightly lower and the spread is less than for other mothers. It also indicates that the mean weight of the mother of low birthweigt children is slightly lower than the weight of other mothers. The number of previous premature labours was one average about the same for both groups of mothers, but the mothers with low birthweight babies had a greater spread. The number of physicians visited during the pregnancy has a slightly lower mean for mothers of low birthweight children then other mothers.

```
# Table summary of categorical covariants and response
with(birthwt_f, table(low_f, race_f))
```

```
##      race_f
## low_f  1  2  3
##     0 73 15 42
##     1 23 11 25
```

```
with(birthwt_f, table(low_f, smoke_f))
```

```
##      smoke_f
## low_f  0  1
##     0 86 44
##     1 29 30
```

```
with(birthwt_f, table(low_f, ht_f))
```

```
##      ht_f
## low_f   0   1
##     0 125   5
##     1  52   7
```

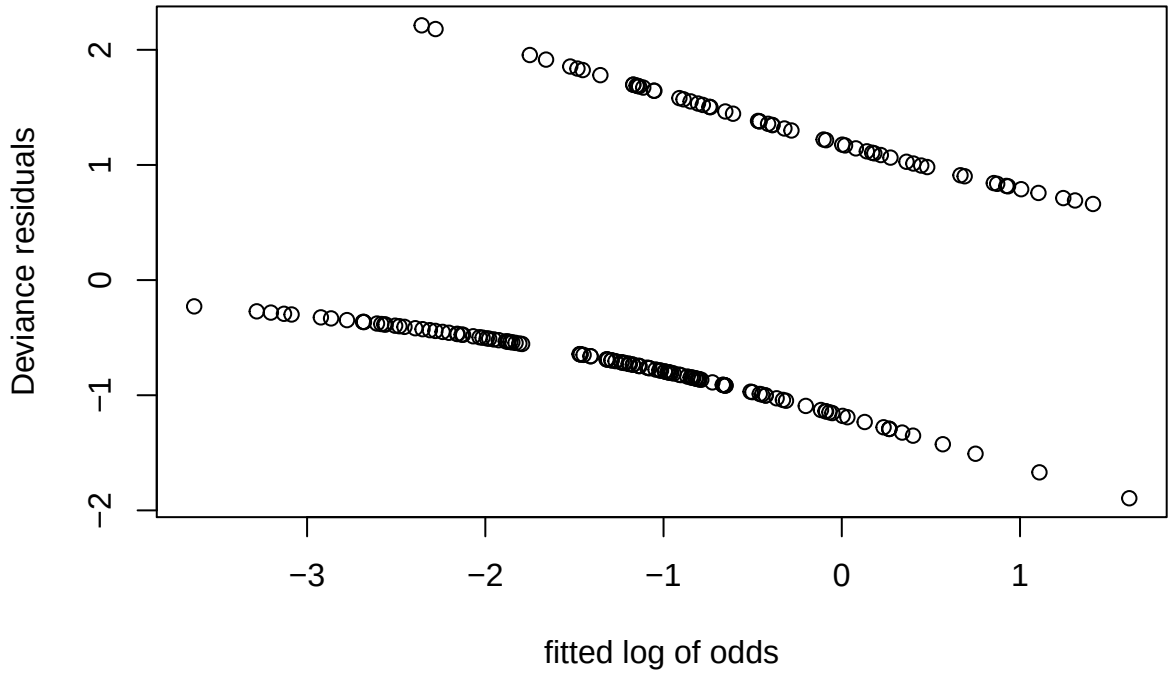
```
with(birthwt_f, table(low_f, ui_f))
```

```
##      ui_f
## low_f   0   1
##     0 116  14
##     1  45  14
```

The tables indicate that the proportion of mothers with lowweight children is seemingly significantly higher for black and other mothers than for white mothers. Also, smoking seems to increase the proportion of lowweight children. A history of hypertension seems to increase the proportion of children with low birthweight as well, however, the number of mothers in this category might be too low to make any reliable statement about this relationship. Also, uterine irritability seems to increase the proportion of lowweight children significantly.

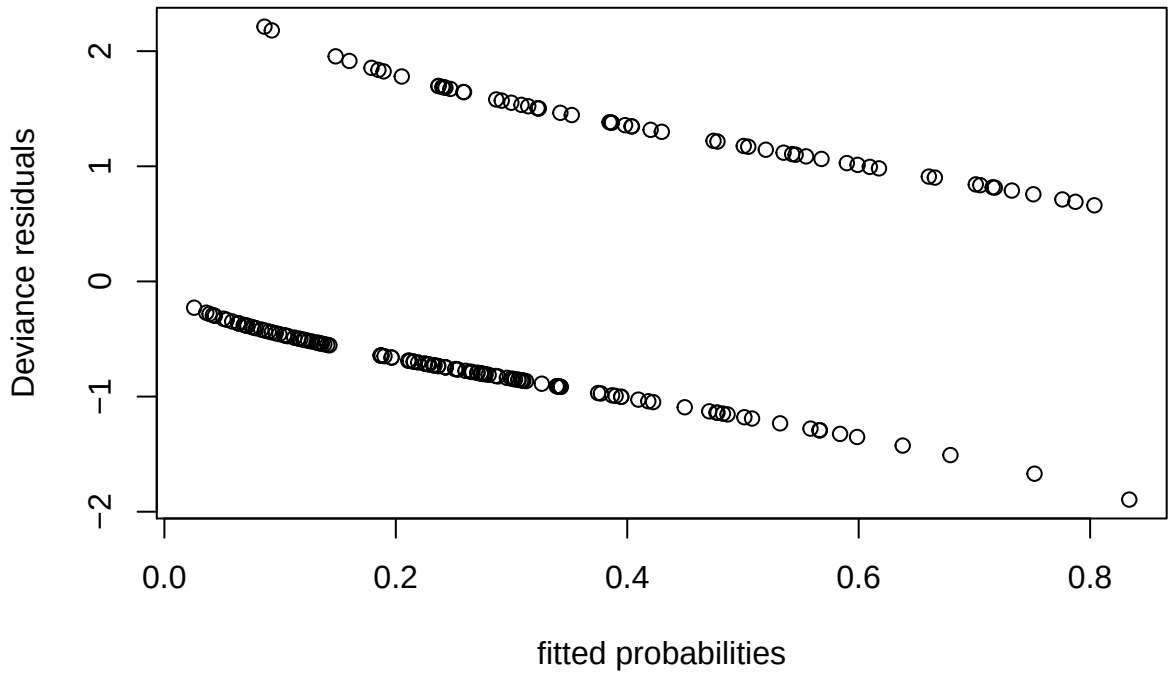
4. Fit the Logistic Regression Model and do Model Diagnostics

```
# Model Fitting

logmod <- glm(low_f~age+lwt+race_f+smoke_f+ptl+ht_f+ui_f+ftv, data =birthwt_f, family=binomial)

# Diagnostics

# Does model fit all individual data points uniformly well?

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

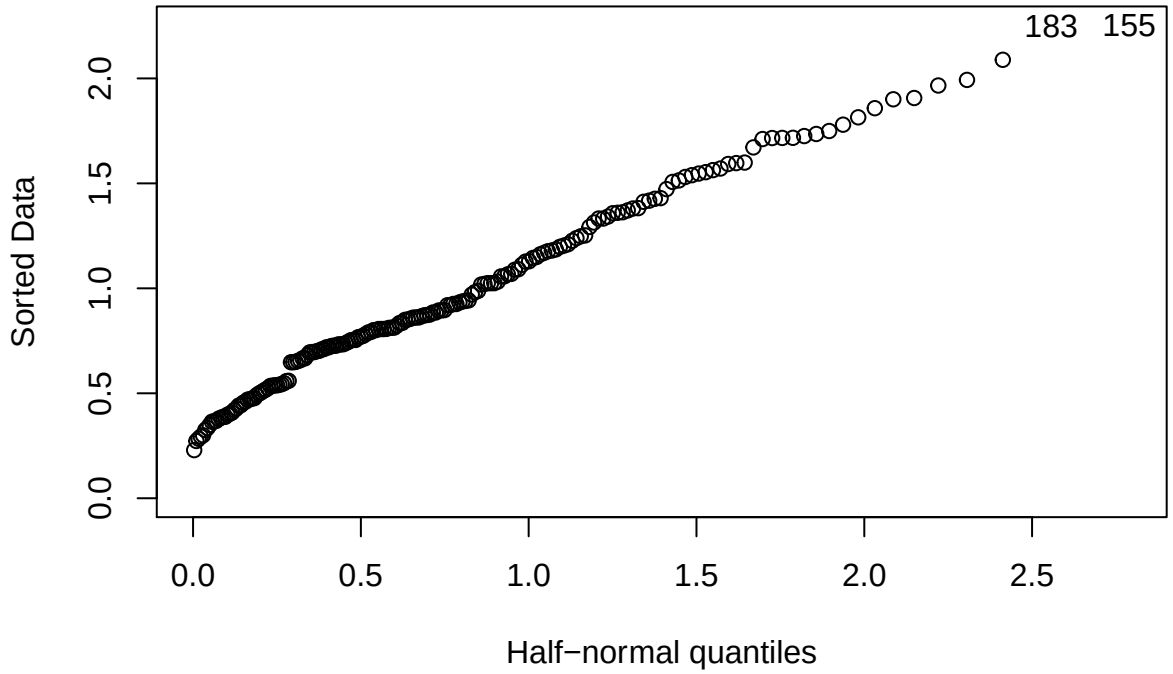



```
plot(residuals(logmod)~predict(logmod, type="response"), xlab="fitted probabilities", ylab="Deviance res

# Half normal plot for Outliers
require(faraway)
```

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




```
halfnorm(rstudent(logmod))
```



In both of the first two models, all points fit the pattern. Thus, we can assume that all individual data points fit uniformily well. The half normal plot is a approx straight line. Thus, we can assume that there are probably no outliers that we need to be concerned about. Thus, none of the diagnostic plots give us much reason to worry about the assumptions in our model.

5. Model Analysis

To determine which quantitative covariants are significant and which not, we can simply look at the p-values of z-tests. For the categorical variables we need to conduct wald tests.

```
# z-test for quantitative covariants
summary(logmod)
```

```
##
## Call:
## glm(formula = low_f ~ age + lwt + race_f + smoke_f + ptl + ht_f +
##     ui_f + ftv, family = binomial, data = birthwt_f)
##
## 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 *
## race_f2      1.272260   0.527357   2.413  0.01584 *
## race_f3      0.880496   0.440778   1.998  0.04576 *
## smoke_f1     0.938846   0.402147   2.335  0.01957 *
## ptl          0.543337   0.345403   1.573  0.11571
## ht_f1        1.863303   0.697533   2.671  0.00756 **
## ui_f1        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
```

Observing the p-values, we can see that of the quantitative covariants only lwt (weight of mother) appears to have a significant correlation with children with low birthweight. Age, ptl, ftv appear not to be significantly different from 0 in this model.

```
# Waldtest for categorical covariants
require(aod)
```

```
## Loading required package: aod
##
## Attaching package: 'aod'
##
## The following objects are masked from 'package:faraway':
##
##     rats, salmonella
```

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

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

```
wald.test(b = coef(logmod), Sigma = vcov(logmod), Terms = 6) # smoke
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.02
```

```
wald.test(b = coef(logmod), Sigma = vcov(logmod), Terms = 8) # ht
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.1, df = 1, P(> X2) = 0.0076
```

```
wald.test(b = coef(logmod), Sigma = vcov(logmod), Terms = 9) # ui
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 2.8, df = 1, P(> X2) = 0.095
```

At a .05 confidence level, race, smoke, and ht (history of hypertension) are significantly different 0, ui (presence of uterine irritability) is not.

Thus, overall the covariants that appear to be significantly correlated with low birthweight of babies are lwt (weight of mother), race, smoke, and ht (history of hypertension).

As determined in the project description, we will not do any model selection here. Thus, the model analysis concludes with simply pointing out which covariants are significant and which are not.

6. Interpretation

For the quantitative covariant lwt, if the mother's weight is 1 pound higher, the odds of having a low weight birth becomes approx (e^-.015424) = .98469 of the odds before, given everything else stays constant.

Being a mother of race 2 (black) and race 3 (other than white or black) increases the odds of given birth to a low weight baby in our model to 3.5689 and 2.41210 of the odds for white mothers, given all other factors stay constant, respectively.

Smoking during the pregnancy increases the odds of giving birth to a low weight baby to 2.55703 of the odds of giving birth to a low weight baby for nonsmoking mothers, everything else staying constant.

Having a history of high blood pressure as a mother increases the odds of having a low weight baby to 6.44499 what the odds would be otherwise in our model, with everything else staying constant. This covariant needs to be considered with care since the sample group of mothers with ht was so small.

All other variables cannot be interpreted since they do not appear to be significantly different from 0 in our model.