האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

Generalized Linear Models

Final Report

Introduction

Infant mortality is higher for low birth-weight babies. Several factors during pregnancy can greatly alter the probability of a woman carrying her baby to term and, consequently, delivering a baby of normal birth weight. Data on 189 births were collected at Baystate Medical Center, Springfield, Mass. during 1986. The dataset contains an indicator of low infant birth weight as a response and several risk factors associated with low birth weight. The actual birth weight is also included in the dataset.

In this paper I conducted a brief analysis of low-birthweight data. The purpose was to study the effect of several risk factors, related to the mother, on the birthweight of the newborn baby. I fitted generalized linear models and tried to find the most appropriate one which will best predict whether a baby will be born with low birthweight or not.

Submitted to: Professor Samuel Oman

Student: Shira Esudri

Date: February 2021

Description of Data - Birth weight

The dataset consists of the following 10 variables:

low: indicator of birth weight less than 2.5kg

age: mother's age in years

lwt: mother's weight in pounds at last menstrual period

race: mother's race ("white", "black", "other")

smoke: smoking status during pregnancy

ht: history of hypertension

ui: presence of uterine irritability

ftv: number of physician visits during the first trimester

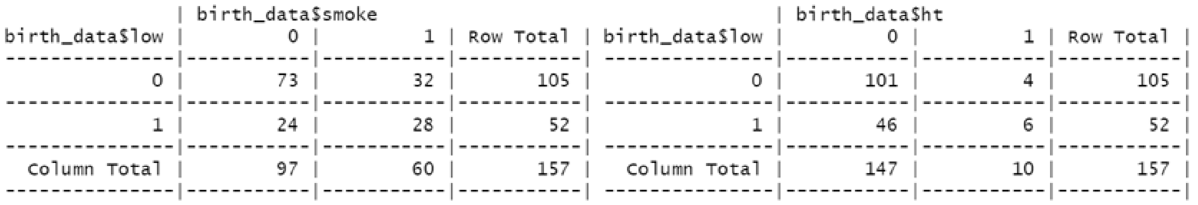
ptl: number of previous premature labors

bwt: birth weight in grams

There are 157 observations in the data, 75 observations of white women, 22 observations of black women and 60 observations of "Other" (nor white or Black) women. I split the race variable into three categorial variables ("White", "Black", "Other"). After the exploratory analysis, I removed the "Black" variable to avoid multicollinearity. An explanation for this remove will be provided soon. The actual birth weight (BWT) is also included in the dataset, but it does not add any important information to the analysis because it already has a variable indicator type (our explanatory variable) which shows whether the weight of the baby is considered low or not. Therefore, after the exploratory analysis, it will be removed from the data.[1] The data consist of continuous variables and categorical variables: The continuous variables are "Mother's Age" and "Mother's Weight (in pounds)". The categorical variables are "Smoke", "Hypertension", "Uterine Irritability"," FTV", "Prev Premature Labors", "Other", "White". The explained variable - Y is "Low".

Preliminary analysis

Cross Table frequencies for low birth weight with categorial variables



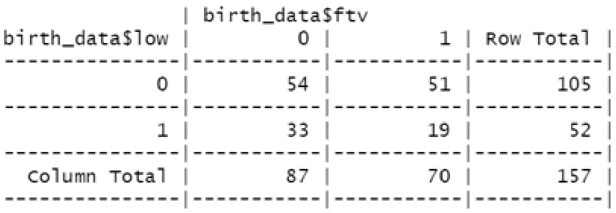
```
               | birth_data$smoke                                   | birth_data$ht
birth_data$low |         0 |         1 | Row Total | birth_data$low |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------| ---------------|-----------|-----------|-----------|
             0 |        73 |        32 |       105 |              0 |       101 |         4 |       105 |
---------------|-----------|-----------|-----------| ---------------|-----------|-----------|-----------|
             1 |        24 |        28 |        52 |              1 |        46 |         6 |        52 |
---------------|-----------|-----------|-----------| ---------------|-----------|-----------|-----------|
  Column Total |        97 |        60 |       157 |   Column Total |       147 |        10 |       157 |
---------------|-----------|-----------|-----------| ---------------|-----------|-----------|-----------|
```

The Smoke variable - There are more nonsmoking mothers (97) then smoking mothers (60). Of the non-smoking mothers 25% babies were born with low birth weight and of the smoking mothers 47% babies were born with low birth weight. It can be concluded that the smoking variable may influence whether the baby will be born with low birth weight.

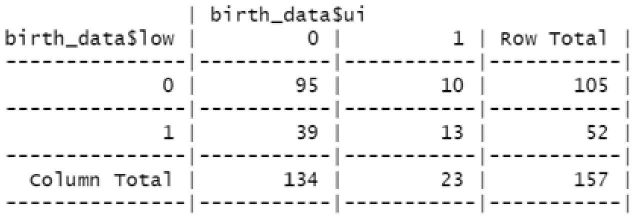
The Hypertension variable (HT) - There are 147 observations with no hypertension and only 10 observations with hypertension. In terms of percentage, of the non-hypertension mothers 31% of the babies were born with low birth weight. Of the hypertension mothers 60% of the babies were born with low birth weight. From the data we can conclude that Hypertension can influence the baby's born

[1] This variable will return to be relevant at the end of the paper when I will try to fit a linear regression model.

weight. keep in mind that there are very few observations of women with HT, therefore this conclusion should be taken with limited caution.



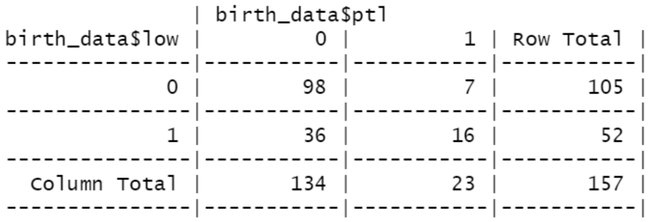
```
                | birth_data$ftv 
birth_data$low |         0 |         1 | Row Total | 
---------------|-----------|-----------|-----------|
             0 |        54 |        51 |       105 | 
---------------|-----------|-----------|-----------|
             1 |        33 |        19 |        52 | 
---------------|-----------|-----------|-----------|
  Column Total |        87 |        70 |       157 | 
---------------|-----------|-----------|-----------|
```



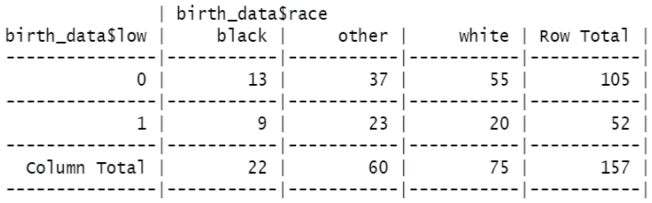
```
               | birth_data$ui 
birth_data$low |         0 |         1 | Row Total | 
---------------|-----------|-----------|-----------|
             0 |        95 |        10 |       105 | 
---------------|-----------|-----------|-----------|
             1 |        39 |        13 |        52 | 
---------------|-----------|-----------|-----------|
  Column Total |       134 |        23 |       157 | 
---------------|-----------|-----------|-----------|
```

The FTV variable – The FTV variable is number of physician visits during the first trimester. Of the mothers with FTV variable with value = 1, 27% of the babies were born with low birth weight and of the mothers with FTV variable with value = 0, 38% of the babies were born with low birth weight. There is a difference but it is not big so we will have to do more tests before determining whether the FTV variable has great effects on the weight of the baby.

The UI variable - We notice that there are much more mothers without presence of uterine irritability (134) then mothers with presence of uterine irritability (23). Of the mothers with UI, 57% babies were born with low weight and of the mothers without UI, only 29% babies were born with low weight. It therefore appears that UI does affect baby's birth weight. Again, keep in mind that there are very few observations of women with UI versus observations of women without UI and therefore this conclusion should be taken with limited caution.



```
               | birth_data$ptl 
birth_data$low |         0 |         1 | Row Total | 
---------------|-----------|-----------|-----------|
             0 |        98 |         7 |       105 | 
---------------|-----------|-----------|-----------|
             1 |        36 |        16 |        52 | 
---------------|-----------|-----------|-----------|
  Column Total |       134 |        23 |       157 | 
---------------|-----------|-----------|-----------|
```

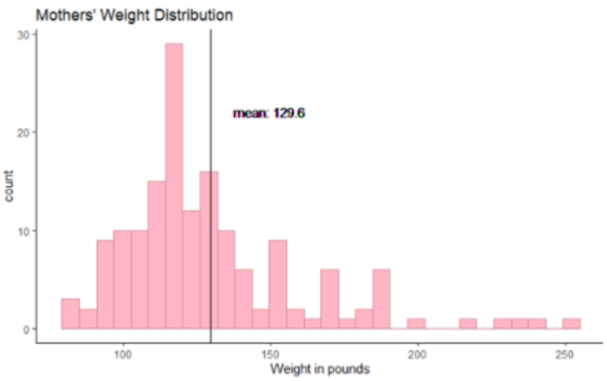


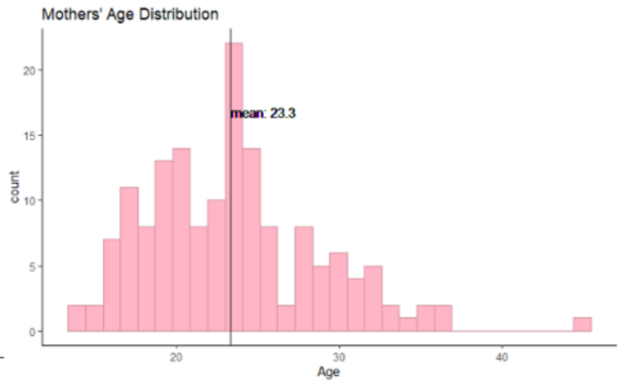
```
               | birth_data$race 
birth_data$low |     black |     other |     white | Row Total | 
---------------|-----------|-----------|-----------|-----------|
             0 |        13 |        37 |        55 |       105 | 
---------------|-----------|-----------|-----------|-----------|
             1 |         9 |        23 |        20 |        52 | 
---------------|-----------|-----------|-----------|-----------|
  Column Total |        22 |        60 |        75 |       157 | 
---------------|-----------|-----------|-----------|-----------|
```

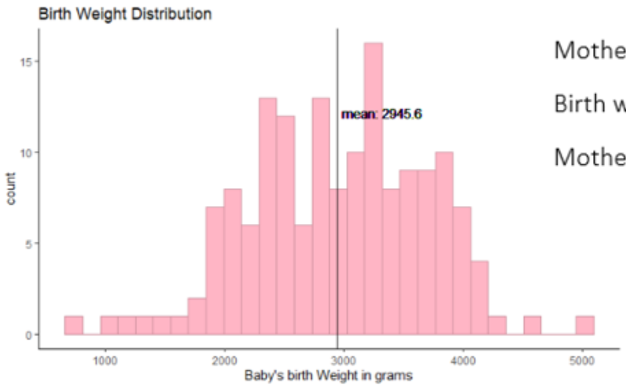
The PTL variable - There are much more mothers with no PTL (134) then mothers with PTL (23). Of the mothers that had previous premature labors 70% of the babies were born with low weight, which make sense because previous premature labors may indicate premature birth (if a baby is born prematurely it is likely that he will be born at a relatively low birth weight). Of the mothers that did not have previous premature labors 27% of the babies were born with low weight. Therefore, its look like the PTL variable has influence on the birth weight . But here too, there is small amount of observation and therefore we should keep that in mind before determining.

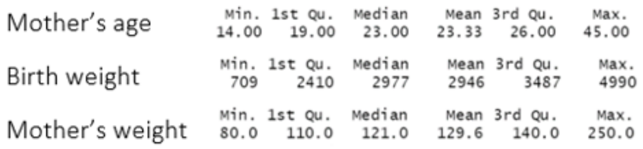
The Race variable - We see that 41% of the black babies are born with low birth weight, 38% of the "other" babies are born with low birth weight and 27% of the white babies are born with low birth weight. Based on these data I decided to separate the race variable into two categorial variables, "White" and "Other" to avoid multicollinearity in the model. I chose to split this way because there are very few observations of black women compare to the rest. In addition, about 40% of "other" and "black" babies are born with low weight so it may be possible to treat them as one group. Furthermore, both groups have similar socioeconomic status in the US, which could affect birth weight. However, if "others" includes Orientals who tend to be shorter, it may also affect birth weight. Therefore, we should pay attention to this fact while building and analyzing the model.

Histogram distributions of continuous variables









```
Mother's age        Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
                   14.00   19.00   23.00   23.33   26.00   45.00
Birth weight        Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
                     709    2410    2977    2946    3487    4990
Mother's weight     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
                    80.0   110.0   121.0   129.6   140.0   250.0
```

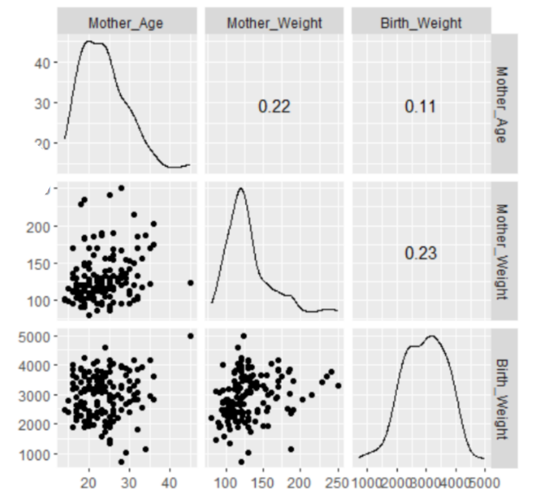
The distribution of the **Mother's Age** is between 14 and 35 years old with Average of 23.3 years old. Note that there is a small jump at 45 years old, an observation that we will check later.

The distribution of the **Mother's Weight** is between 80(+-) and 200 pounds with a few observations of 200-250 pounds (long right tail). Average is 129.6 pounds, and the main mass of observations is between 100 and 150. That is, it can be seen from this distribution that the particularly high-weight observations raise the overall average.

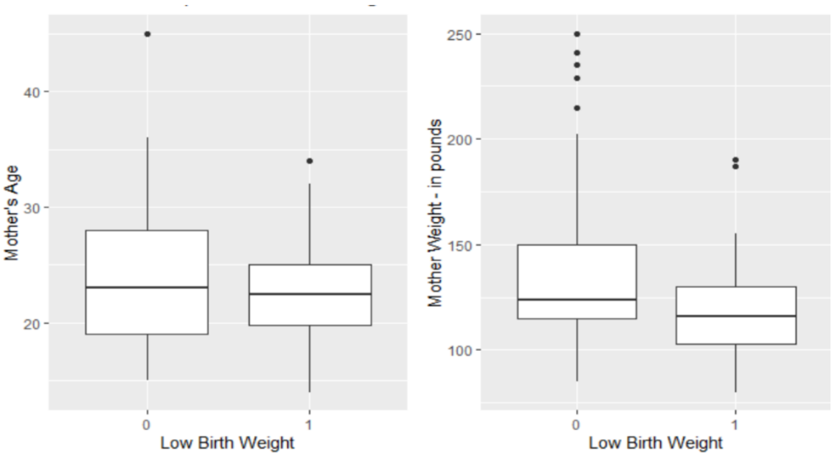
The distribution of the **Baby's Birth Weight** is between 1-5kg., while most babies weigh between 2-4kg. In general, the baby weight distribution seems to be close to normal around the average – 3kg.

Correlations and Scatterplots between Low birth weight and the continuous variables

According to this figure there is no strong correlations between the continuous variables and it looks like that there are no multicollinearity. In the scatter plots we see that there is no clear trend and that there are some abnormal observations which we will examine later.



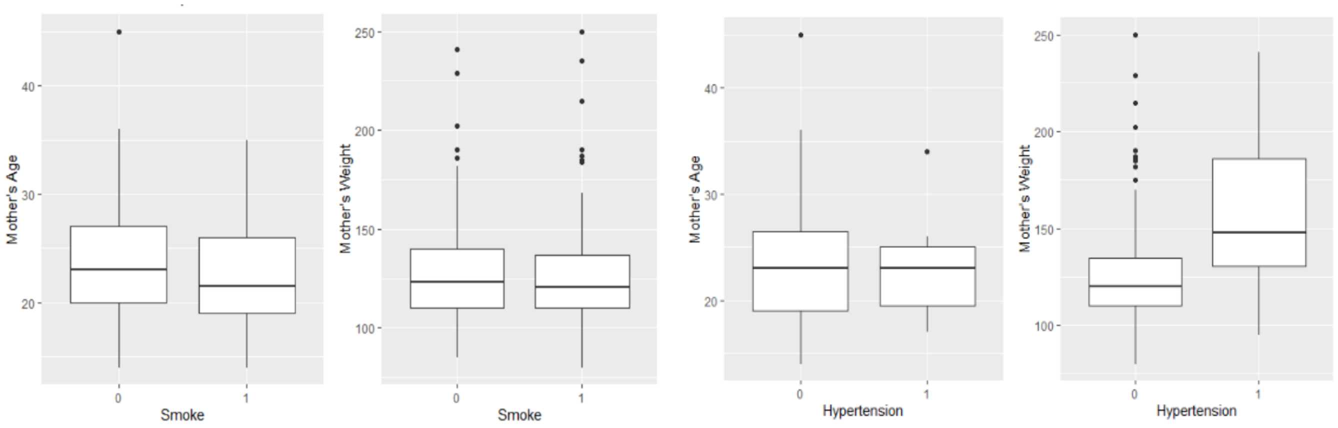
Boxplot of Low birth weight and continuous explanatory variables



Mother's Age VS. Low birth weight- The average age of the mother seems to be in the early 20's no matter if the baby is born with low birth weight or not but we see that the distribution of the mother's age with the non-low birth weight is wider than the distribution of the mother's age with low birth weight. The mother's age does not seem to play a major role in the baby's birth weight. There is unusual observation at the age of 45 that gave birth to non-low birth weight baby.

Mother's Weight VS. Low birth weight - The average weight of the mother is around 120-125 pounds no matter if the baby is born with low birth weight or not. We can see that the distribution of mother weight with non-low birth weight has long right tail. In addition, there is a difference in the extreme observations. The upper limit of the mother's weight who gave birth to low-birth weight babies is lower than the upper limit of the mother's weight who gave birth to non-low-birth-weight babies. Although the average of the two distributions is similar, it may be inferred that low weight women are more likely to give birth to low-birth-weight babies.

Boxplot of continuous and categorical explanatory variables

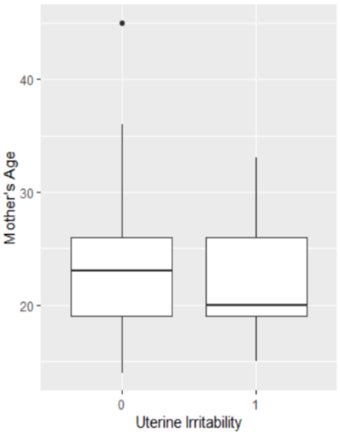


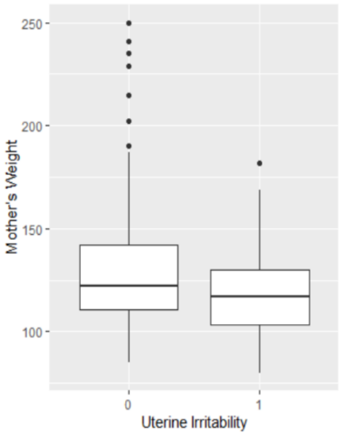
Mother's Age VS. Smoke - The distribution and average are similar for smoking women with mother's age and for non-smoking with mother's age, respectively. We cannot see clear relation between the two distributions. The 45 years old woman is nonsmoker.

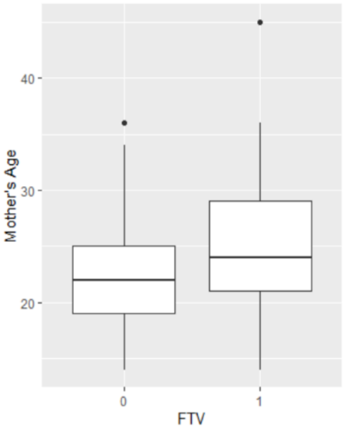
Mother's Weight VS. Smoke - The distribution is similar for smoking women with mother's weight and for non-smoking with mother's weight. They both have long right tail and similar average. Here we also see that apparently there is not clear relation between the two distributions.

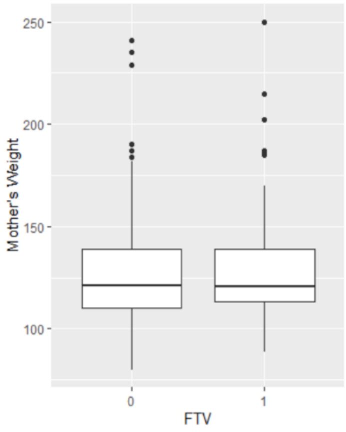
Mother's Age VS. Hypertension - The average age is quite similar in both distributions. Despite this, it seems that the age range of women without HT is wider than the age range of women with HT. In addition, there is an unusual observation in each of the distributions: a 35-year-old woman with HT and a 45-year-old woman without HT.

Mother’s Weight VS. Hypertension – Different distribution and average weight. The non-hypertension with mother’s weight distribution has long right tail. In addition, there seems to be some connection between HT and mother’s weight. Most of the observations of women without HT is between 110-135 pounds and most of the observations of women with HT is between 135-180 pounds. Meaning, Hypertension is more common in women with higher weight. Therefore, I will examine later the possibility of a correlation between the mother’s weight and Hypertension.







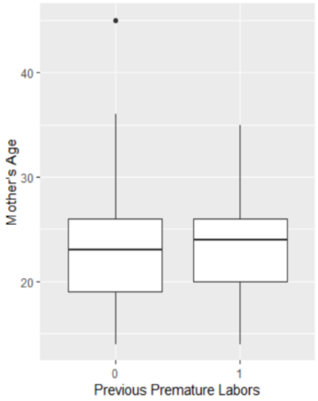


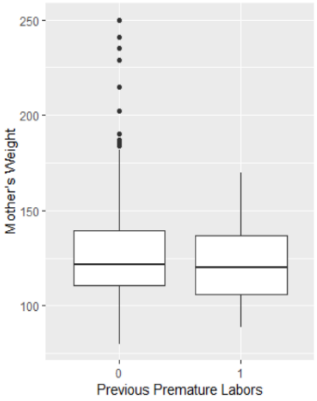
Mother’s Age VS. Uterine Irritability - The average age is not so similar, and we can see that the 45 years old woman did not had UI. There does not appear to be a relation between mother’s age and UI.

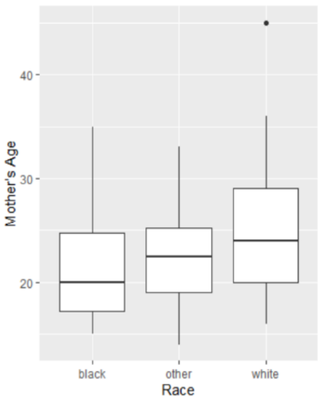
Mother’s Weight VS. Uterine Irritability – The average weight is quite similar, but the distribution of non-Uterine Irritability with mother’s weight has long right tail, unlike the second distribution that seems close to symmetric and has only one extreme observation.

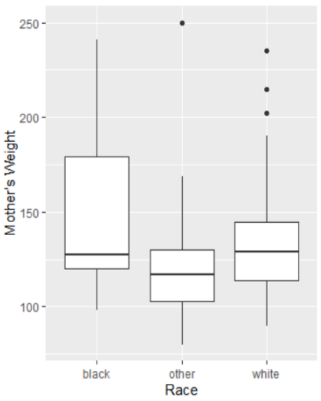
Mother’s Age VS. FTV – The average weight is quite similar. There are two extreme observations, one on each distribution (the 45 years old woman had FTV). Here, too, there is room to examine correlation between mother’s age and FTV when constructing the model.

Mother’s Weight VS. FTV – Similar distribution and average, both have long right tail. There does not appear to be a clear relation between the variables.









Mother’s Age VS. Previous Premature Labors - Similar distribution and average. The 45 years old woman did not have PTL. There does not appear to be a clear relation between the two.

Mother’s Weight VS. Previous Premature Labors – Similar average. The non- Previous Premature Labors with mother’s weight distribution has long right tail. The tail suggests a connection, which also fits in with the def of the variable.

Mother’s Age VS. Race – Different average age for each race and the 45 years old woman is white women. There is some connection between mother’s age and race, which also makes sense socio-economically.

Mother's Weight VS. Race – Different distribution for each race but quite similar average, around 120-130 pounds, as we saw in the histogram. It seems that most observations of Black women are in higher weight relatively to the other races and that most observations of Other women are in lower weight relatively to the other races. There are some unusual observations in Other and in White.

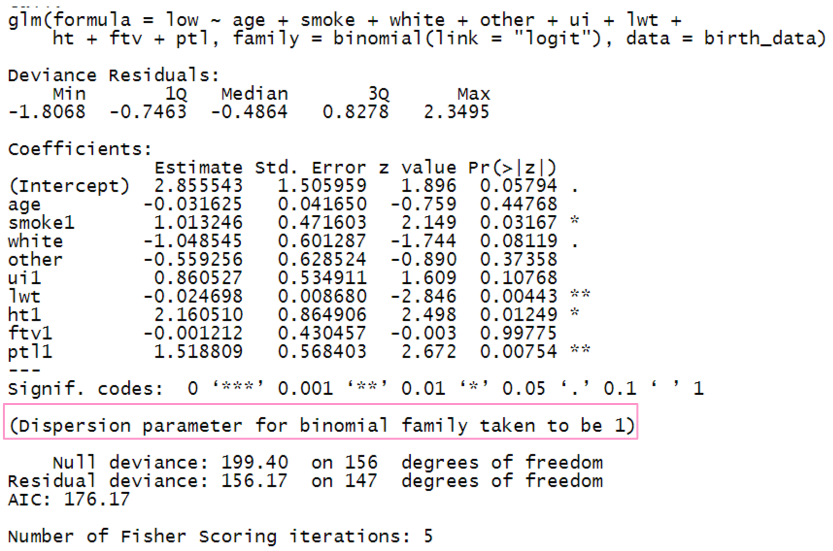
Fitting GLM Model

Due to the structure of the data, I used generalized linear model with logit link function to fit a proper model. The definition of logit: given that the p is a probability of success, that is, the probability of the response taking a value of 1, the odds of success is defined as p/(1 − p). the log-odds are ln(p/(1-p)). It is the linear predictor $X\beta$ of the binary logistic model that plays an important role in logistic modeling.[2] The logit link of the logistic model, is the canonical link, meaning that it stems directly from the binomial family.[3] The logit function is the logistic linear predictor[4]:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

In our case, the "p is a probability of success" is the probability that the baby will be born with low birth weight, under 2.5Kg.

Fit logistic regression model that contains all the explanatory variables



```
glm(formula = low ~ age + smoke + white + other + ui + lwt +
    ht + ftv + ptl, family = binomial(link = "logit"), data = birth_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8068  -0.7463  -0.4864   0.8278   2.3495

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.855543   1.505959   1.896  0.05794 .
age         -0.031625   0.041650  -0.759  0.44768
smoke1       1.013246   0.471603   2.149  0.03167 *
white       -1.048545   0.601287  -1.744  0.08119 .
other       -0.559256   0.628524  -0.890  0.37358
ui1          0.860527   0.534911   1.609  0.10768
lwt         -0.024698   0.008680  -2.846  0.00443 **
ht1          2.160510   0.864906   2.498  0.01249 *
ftv1        -0.001212   0.430457  -0.003  0.99775
ptl1         1.518809   0.568403   2.672  0.00754 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 199.40  on 156  degrees of freedom
Residual deviance: 156.17  on 147  degrees of freedom
AIC: 176.17

Number of Fisher Scoring iterations: 5
```

First, it can be seen in the full model that not all coefficients are significant. The two variables that are explained in the highest level of significant are PLT and Mother's weight with significant level of 0.01. After them, Smoke and HT with significant level of 0.05. This is in line with what we saw in the exploratory analysis (that those variables have influence on the outcome, which is low birthweight). In general, we can say that Smoke, HT and PLT has positive affect and White and LWT has negative affect.

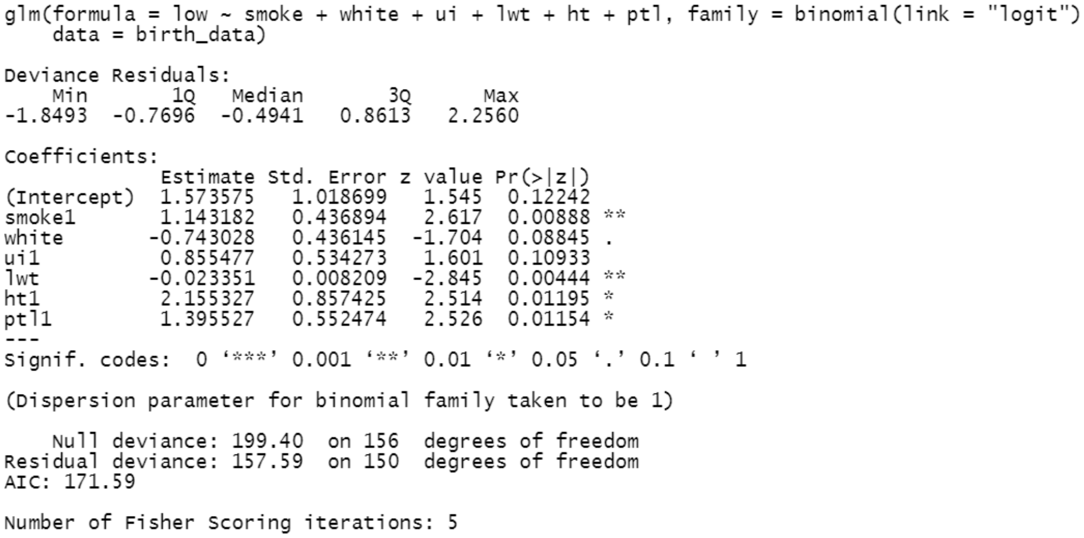
The AIC index of the full model is 176. I will try to improve the model by lowering the AIC and use the significant variables. As we learned, the AIC prevents the addition of too many variables (Overfitting) to the model and is therefore a good index for improving GLM models.

To improve the model, I chose to remove variables that in the preliminary analysis did not show major effect on the explained variable - Low birth weight. As we have seen above, FTV and Age variables do not affect so much on the birth weight and therefore I will try to remove them. After that I decided to

[2] *Hilbe, Joseph M. (2009), Logistic Regression Models, CRC Press, p.3.*
[3] *Hilbe, Joseph M. (2009), Logistic Regression Models, CRC Press, p.69.*
[4] *Hilbe, Joseph M. (2009), Logistic Regression Models, CRC Press, p.83.*

remove the Other variable as well due to its small level of significant. Last, I choose to keep the UI variable, despite the low level of its significance in the model due to its importance in the explanatory analysis. To strengthen my claims, I also used “STEP” function in R to see which combination of variables gives the best model (i.e., the model that manages to lower the AIC the most). Then, I will examine whether the coefficients correspond to what we saw in the preliminary analysis.

The model after removing the variables



```
glm(formula = low ~ smoke + white + ui + lwt + ht + ptl, family = binomial(link = "logit")
    data = birth_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8493  -0.7696  -0.4941   0.8613   2.2560

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.573575   1.018699   1.545  0.12242
smoke1       1.143182   0.436894   2.617  0.00888 **
white       -0.743028   0.436145  -1.704  0.08845 .
ui1          0.855477   0.534273   1.601  0.10933
lwt         -0.023351   0.008209  -2.845  0.00444 **
ht1          2.155327   0.857425   2.514  0.01195 *
ptl1         1.395527   0.552474   2.526  0.01154 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 199.40  on 156  degrees of freedom
Residual deviance: 157.59  on 150  degrees of freedom
AIC: 171.59

Number of Fisher Scoring iterations: 5
```

This submodel contains the most significant variables of the full model, meaning the variables that are likely to have the most influence on Low birth weight. In addition, the AIC index stands on 171.6 which is lower compare to the full model that includes all the variables. Together we can say that the model is indeed improved (significant variables and AIC are the two indications for the quality of the model).

The most significance variables are smoking (has positive affect on the outcome) and LWT (has negative affect), with a significance level of 0.01. The variables PLT and HT have a significance level of 0.05, both are positive, and their coefficients are relatively high compare to the other variables. Meaning they both increase the risk as expected. The variable White has a significance level of 0.1 and negative, meaning it is lowering the risk (as we saw before, fewer white babies were born with low birth weight). In general, the signs and meanings of the coefficients are in line with the above preliminary analysis.

The variable UI is not significant therefore I was debating whether to remove it. But we saw in the exploratory analysis that even though there are few observations of women with UI, this variable has major influence on low birth weight.[5] Therefore, I decided not to remove it from the model. To strengthen my claim, I tried to remove it and saw that it fails to lower the AIC index and that there was not any improvement in the significance of the explanatory variables or in the model.

To further improve the model, I tried to conduct interactions. In the exploratory analysis we saw that there is room to examine two main interactions. The first, an interaction between FTV and the Mother's age and the second, an interaction between the Mother's weight and Hypertension. After examining the two interactions, together and separately, I decided not to add them to the model. That is because they do not improve the model and disrupt the level of significance. For all these reasons I decided to choose this model to be the selected submodel.

[5] “Of the mothers with UI 57% babies were born with low weight and of the mothers with no UI only 29% babies were born with low weight.” As written in the exploratory analysis .

Model Checking Comparison between the full model and submodel

First, compare the output of the full model and the submodel

Full model

coefficients	Estimate full model	Z value full model	Pr(>|z|) Full model
Intercept	2.86	1.896	0.058 .
age	-0.03	-0.759	0.447
smoke	1.01	2.149	0.031 *
white	-1.04	-1.744	0.081 .
other	-0.55	-0.89	0.373
Ui	0.860	1.609	0.107
lwt	-0.025	-2.846	0.004 **
ht	2.16	2.498	0.012 *
ftv	-0.001	-0.003	0.998
ptl	1.52	2.672	0.007 **

Submodel

Estimate submodel	Z value submodel	Pr(>|z|) submodel
1.573	1.545	0.122
1.143	2.617	0.009 **
-0.074	-1.704	0.088 .
0.855	1.601	0.109
-0.023	-2.845	0.004 **
2.155	2.514	0.012 *
1.396	2.526	0.012 *

In the submodel most of the variables are significant in level below 10%, which indicates the quality of the model. In addition, the coefficients did not change much, and their signs remained the same, i.e., the meaning of their effect on the explained variable did not change.

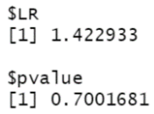
Second, the dispersion parameter

In the case of binomial family, the assumption about the dispersion parameter is that it is equal to 1.[6] According to the formula for dispersion parameter, the value of the submodel's dispersion parameter is: (Residual deviance)/df = 157.57/157 = 1.002. This result confirms the assumption.

Third, hypothesis testing

I would like to check the null hypothesis that the submodel's structure is the correct one. For that, I used likelihood ratio test. The likelihood-ratio test assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods. If the constraint (the null hypothesis) is supported by the observed data, the two likelihoods should not differ by more than sampling error.[7] Thus, the likelihood-ratio test tests whether this ratio is significantly different from one. To test whether the selected submodel is the correct model I used LR test (when the "glm_model" is the full model and the "model" is the submodel).

lr.test(glm_model,model):



```
$LR
[1] 1.422933

$pvalue
[1] 0.7001681
```

Due to that result, we will not reject the null hypothesis. That is, we will not reject the hypothesis that the submodel is indeed the correct model.

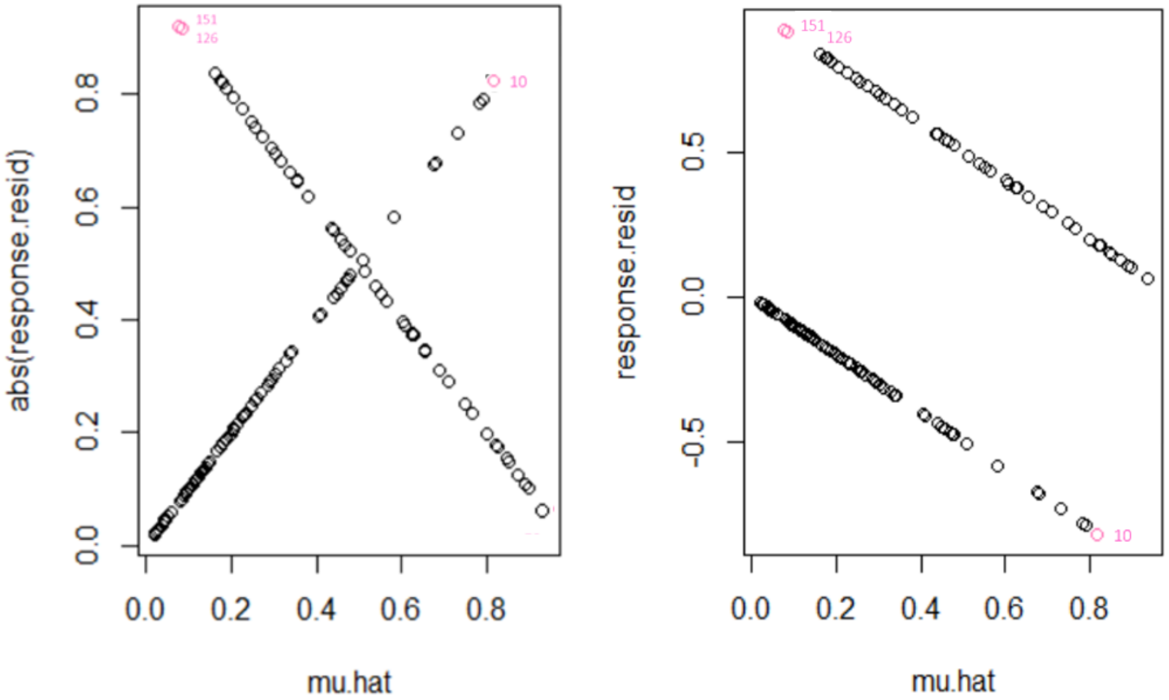
[6] *Hilbe, Joseph M. (2009), Logistic Regression Models, CRC Press, p.134.*

[7] *King, Gary (1989). Unifying Political Methodology: The Likelihood Theory of Statistical Inference. New York: Cambridge University Press. p. 84.*

Residuals diagnosis

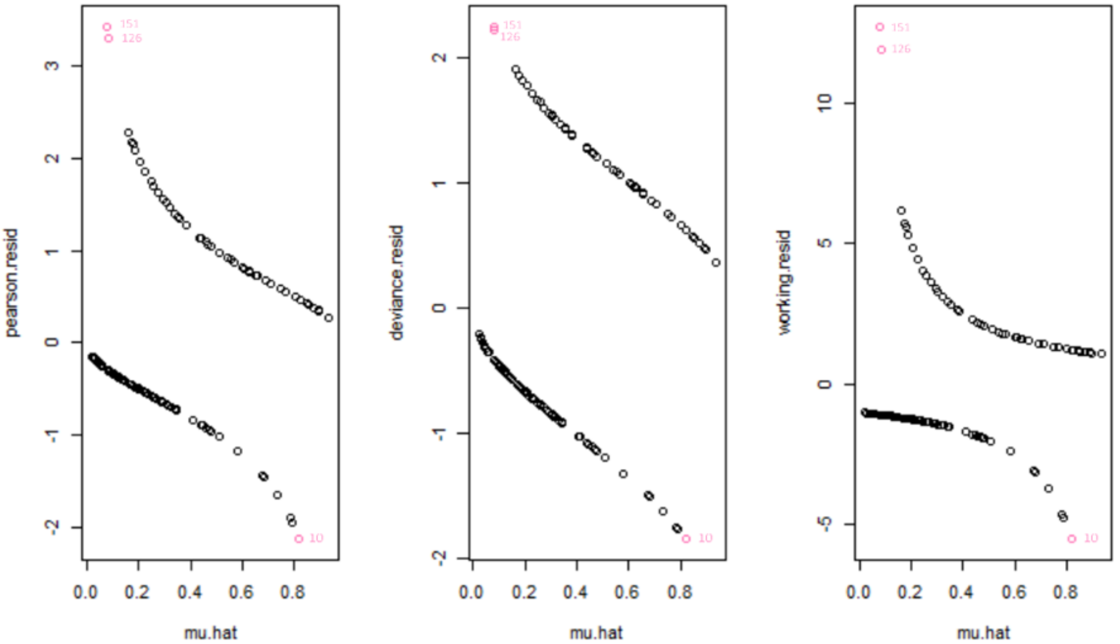
To further test the nature of the submodel and the assumptions I will analyze its residuals.

Response residuals (response & abs(Response))



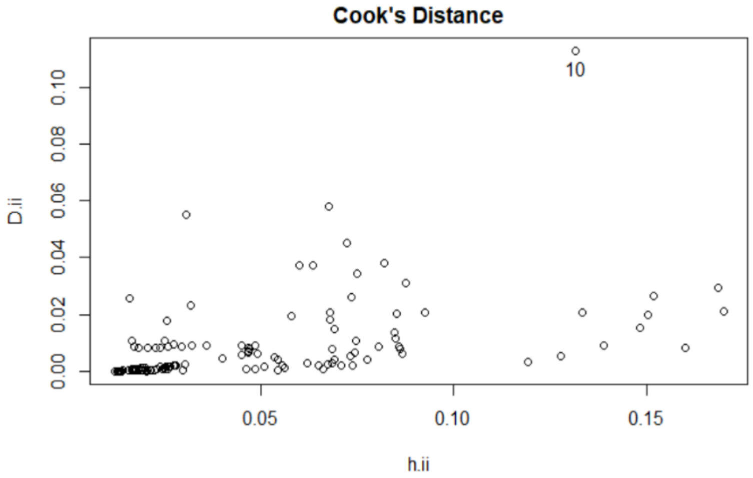
Deviance residuals, Pearson residuals and Working residuals against fitted values

Use deviance residuals, Pearson residuals and Working residuals to look for outliers



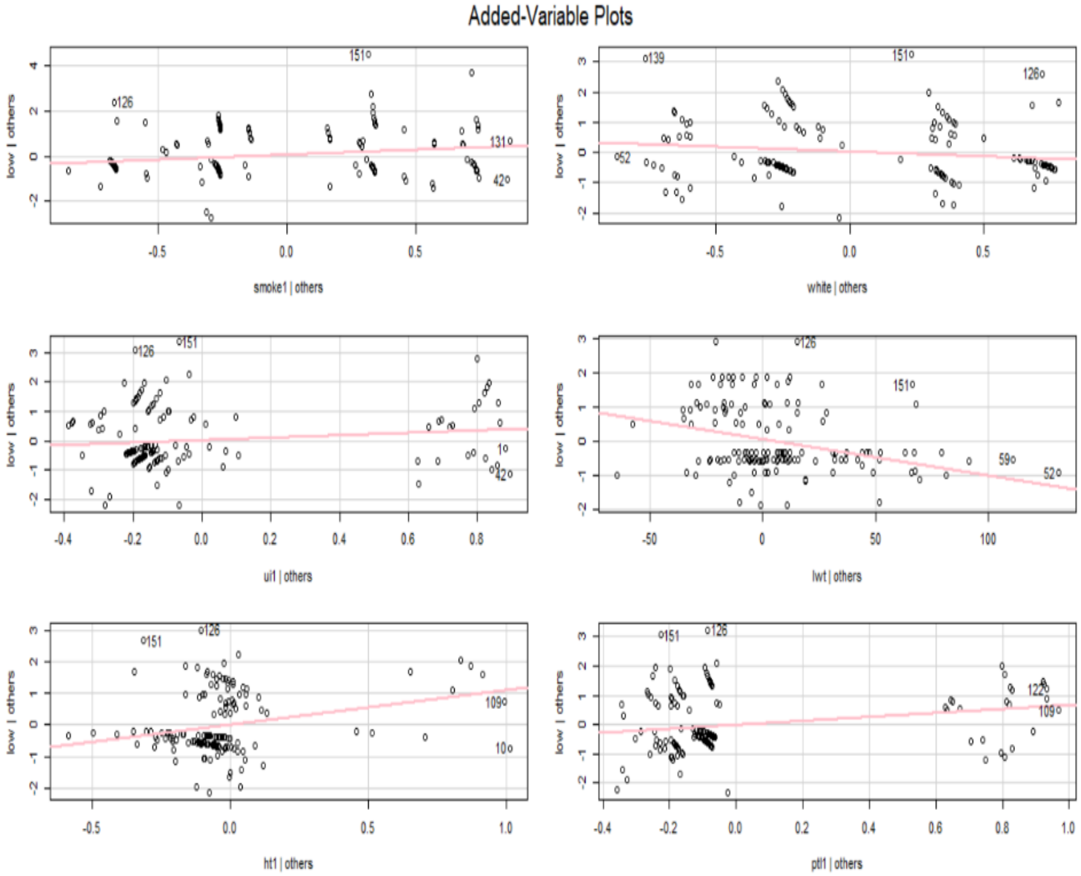
In general, the residuals look ok. We can see that there are two observations suspected to be abnormal at the top and one at the bottom, which will be examined soon. I tried to use variance-stabilizing transformation function of mu.hat (log of mu.hat), but the results were very close to the plots above and we could see the same three observations.

Cook's Distance

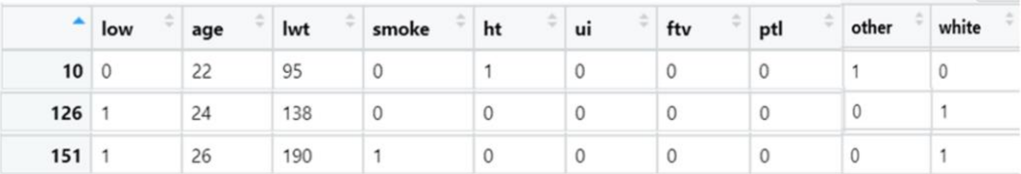


In Cook's Distance graph, which sometimes indicates abnormal observations as well, there is one observation that looks particularly unusual, number 10, and there might be few more scattered.

Add-Variables Plot



According to the Added-Variables Plots, observations number 126 and 151 appear in each of the graphs, which reinforces the suspicion that these are indeed abnormal observations. Also, observation number 10 are shown in "HT vs Low" graph which reinforces the suspicion as well. This graph shows additional observations that influence the linear line, such as observation number 139 and perhaps 42, but they do not appear consistently in all graphs and we have not seen them in the other residual analyzes so I decided to ignore them. After the residual diagnostic and plots, the three candidate observations are 126, 151,10. Before choosing whether to remove the observations and if so which observation to remove I will examine them.



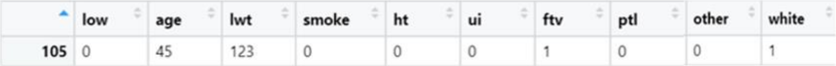
	low	age	lwt	smoke	ht	ui	ftv	ptl	other	white
10	0	22	95	0	1	0	0	0	1	0
126	1	24	138	0	0	0	0	0	0	1
151	1	26	190	1	0	0	0	0	0	1

Observation number 10 is a woman who is neither white nor black (other) in average age and below average weight that has hypertension and did not give birth to a low birthweight. We have learned that low weight mothers are more likely to give birth to low birthweight babies. But remember that women of a “other” race can be of Asian descent. If so, they tend be shorter and therefore in lower weight. On the other hand, 95 pounds is very low weight, much lower than the average weight in the data. That is, for the woman not to be considered low weight she must be extremely short.

Observation number 126 is a white woman in average age and above average weight with no other variable that can affect low birth weight. We have seen that among white women there is the least chance of giving birth to a low birthweight babies and the fact she is a little above the average weight and did not had any of the risk factors should also decrease the chances. Following that information, it is strange that she gave birth to a low birthweight baby and therefore it is possible that this is an abnormal observation. I will examine this claim below.

Observation number 151 is a white woman in average age and above the average weight. This woman has no other variable that can affect the birth of a low-weight baby beside smoking. The opposite, we have seen in the exploratory analysis that white women are less likely to give birth to low weight babies. In addition, In the distribution of low-birth weight with mother’s weights we saw that the higher weight of the mother the less chance she will give birth to a low birthweight baby. But we also saw in that distribution two unusual observations of high-weight women giving birth to low-birthweight babies. It seems to be one of those women. Therefore, I will check the option that this is an abnormal observation.

Note, During the preliminary analysis we noticed that there is an interesting observation (number 105). A white 45-year-old woman who gave birth to a non-low birth weight baby. we said that we would test to see if it affected the analysis and the model. However, this observation did not come up in the residual diagnostic above and when we examine the rest of its explanatory variables it does not look abnormal. Therefore, we will not remove it. But it was important to check it too to not miss important information.

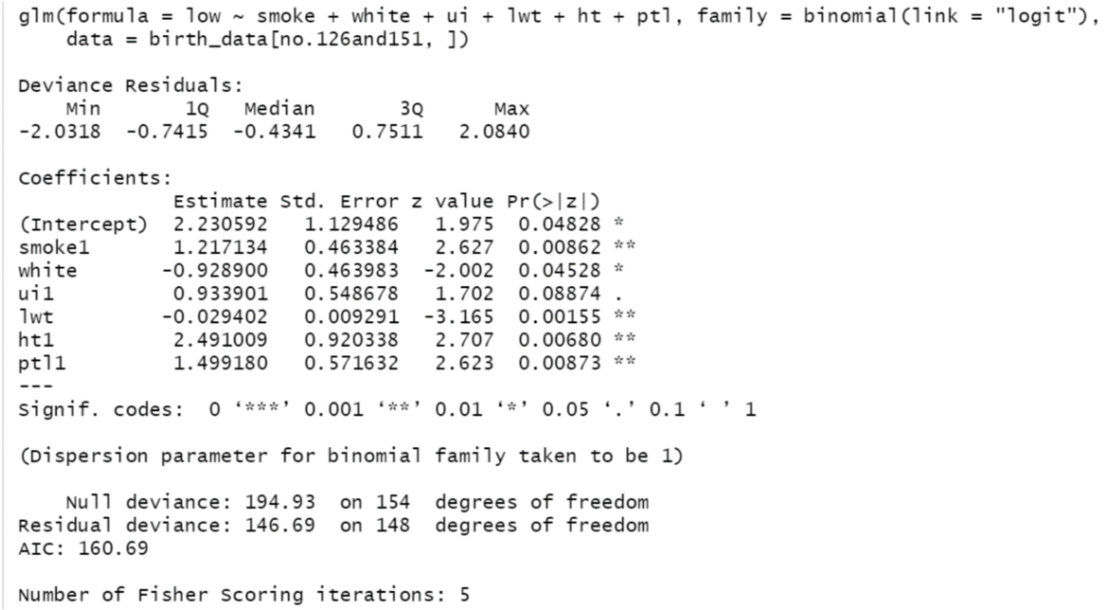


	low	age	lwt	smoke	ht	ui	ftv	ptl	other	white
105	0	45	123	0	0	0	1	0	0	1

After the examination I choose to remove two particularly unusual observations, number 126 and 151, that emerged from the residuals diagnostic and graphs analyzes above. I decided not to remove obseravtion number 10 for serveral reasons. First, the fact that we do not know the true origin of the woman and her BMI, therefore we cannot determine whether she is indeed in low weight. Second, I believe that this observation can be categorized as both exceptional and non-exceptional. For those reasons and the absence for more information I decided not to remove this observation to avoid losing essential data.

GLM model without abnormal observations

The output for the model without observation 126 and 151.



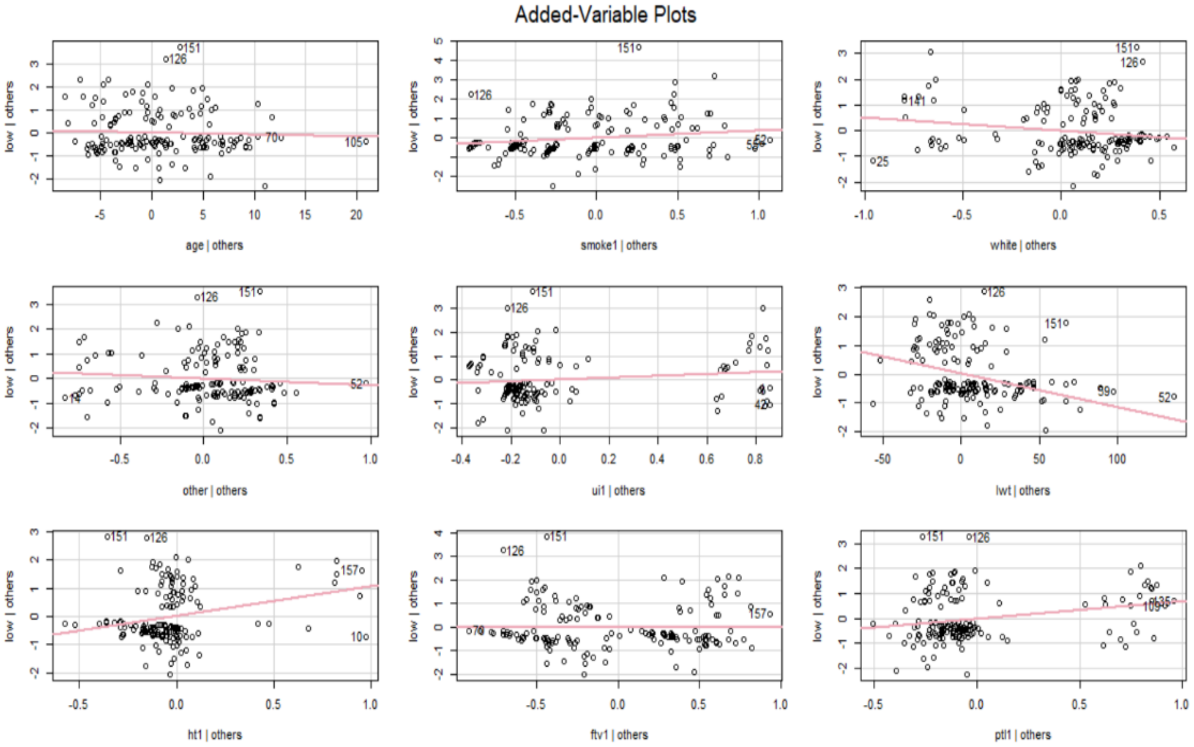
```
glm(formula = low ~ smoke + white + ui + lwt + ht + ptl, family = binomial(link = "logit"),
    data = birth_data[no.126and151, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0318  -0.7415  -0.4341   0.7511   2.0840

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.230592   1.129486   1.975  0.04828 *
smoke1       1.217134   0.463384   2.627  0.00862 **
white       -0.928900   0.463983  -2.002  0.04528 *
ui1          0.933901   0.548678   1.702  0.08874 .
lwt         -0.029402   0.009291  -3.165  0.00155 **
ht1          2.491009   0.920338   2.707  0.00680 **
ptl1         1.499180   0.571632   2.623  0.00873 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.93  on 154  degrees of freedom
Residual deviance: 146.69  on 148  degrees of freedom
AIC: 160.69

Number of Fisher Scoring iterations: 5
```

We see that the coefficients did not change much, and their signs (positive or negative - The direction of the slope) remain the same for all coefficients. That is, the meaning of the coefficients remained the same and we did not remove observations that had too much impact on our data or significantly changed the conclusions. In addition, most of the variables are more significant and the AIC is much lower, which indicates that the model without those observations is better.

In general, we are in no hurry to remove observations from the data. But in this case, we have quite a few observations, 157 in total, so dropping about 1.2% of the observations from the dataset is not very critical. For that matter, omitting such a small percentage can happen for a variety of other reasons therefore I do not think it will negatively affect the results or cause too great bias. In addition, we saw that those observation were unusual and that removing them from the model did not change the interpretation of the results. Therefore, the final model will be the model without the observations 151 and 126. Despite all, I will limit my remarks and present to the doctor the two improved models and let him choose the best in his opinion.

Examine the abnormal observations in the full model

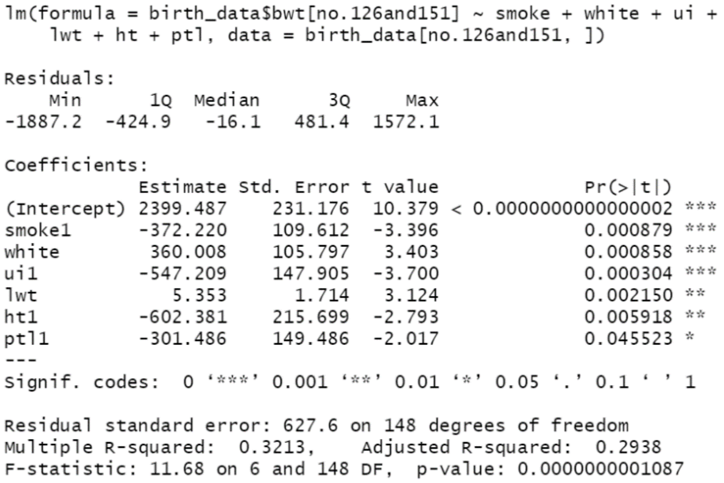
To check that the removed observations are indeed suspected to be abnormal, I checked the Added-Variable Plots of the full model to compare.



In these graphs we see even more clearly that observations 151 and 126 are shown in all the graphs and more importantly that they affect the straight line. This is positive because it appears that we did not removed necessary information in the transition and strengthens the suspicion that the observations we removed are indeed abnormal.

Fit linear regression model

OLS model when using the birth weight variable as continuous variable for the final model (without abnormal observations).



```
lm(formula = birth_data$bwt[no.126and151] ~ smoke + white + ui +
    lwt + ht + ptl, data = birth_data[no.126and151, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-1887.2  -424.9   -16.1   481.4  1572.1 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept) 2399.487    231.176  10.379 < 0.0000000000000002 ***
smoke1      -372.220    109.612  -3.396             0.000879 ***
white        360.008    105.797   3.403             0.000858 ***
ui1         -547.209    147.905  -3.700             0.000304 ***
lwt            5.353      1.714   3.124             0.002150 ** 
ht1         -602.381    215.699  -2.793             0.005918 ** 
ptl1        -301.486    149.486  -2.017             0.045523 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 627.6 on 148 degrees of freedom
Multiple R-squared:  0.3213,	Adjusted R-squared:  0.2938 
F-statistic: 11.68 on 6 and 148 DF,  p-value: 0.0000000001087
```

As we can see, the coefficients in the OLS model are all significant. In addition, the P-value, Multiple R-squared, Adjusted R-squared and the F-statistic shows that this model looks relatively good.

Conclusions

In this report I learned about different variables (risk factors) of women during pregnancy that may affect the birth of a low birthweight baby. Due to the structure of the data and the fact the explained variable was binary, I fitted generalized linear model. To construct the GLM model I examined frequency boxes, correlations, joint distributions, marginal distributions, and other checks on the explanatory variables to learn about the connections between the variables and their behavior.

To improve the model, I selected the most influential or significant explanatory variables and try different types of interaction. I chose the model whose AIC is the lowest and whose explanatory variables are significant. In addition, I did a few checks and comparison on the submodel (for example check the Dispersion parameter and Hypothesis testing) to confirm its quality. For better improvement of the model, I checked the residuals to look for outliers and decided to remove the most unusual observations, which indeed improve the model in terms of AIC and significant. In addition, I checked the decision to remove those observations by comparing the Added Variables Plots for the full model and the submodel. In the end, I fitted linear regression model for the final model using the birth weight variable as continuous, when the results of the indices showed that the model is relatively good.

The results of the final GLM model (without the abnormal observations) indicates that the variables “Smoke” “UI” “HT” and “PLT” is significantly associated with the probability that the baby will be born with low birthweight. That is, they increase the risk. The variables: “White” and “LWT” has negative effect, which means lowering the risk (Which is actually a good effect). The submodel and the model without the abnormal observations are consistent with the preliminary analysis of the variables. For example, the higher the mother’s weight the less risk and white mothers are less likely to give birth to low birthweight babies. On the other hand, if a woman has HT or PLT it increases her chances of giving birth to a low birthweight baby (Probably more than if the women Smokes or has UI).