**Generalized Linear Models**
**Final Project**
**Maria Mordvova 345232821**
**Submitted to: Samuel Oman**
**31 January 2022**

## I Description of the data

### *Background*

I was asked to analyze the subset of 139 observations from a birthweight data collected by Baystate Medical Center, Springfield, Mass during 1986. It is well-studied that infant mortality is higher for low birth-weight babies. Moreover, number of 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. That is why, we would like to examinate which factors, if at all, effect the babies' birth weight.

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. [1]

### *Data description*

The dataset consists of the following 10 variables:
**low:** indicator of birth weight less than 2.5kg (gives 1 if so)
**age:** mother's age in years
**bwt:** birth weight in grams
**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 labours

Nevertheless, we need to note that most of given variables are dummy ones (get 1 – if the event occurred, 0 - otherwise). Moreover, the explanatory variable "*race*" is a categorical variable, which will further transform into dummy variable as well.

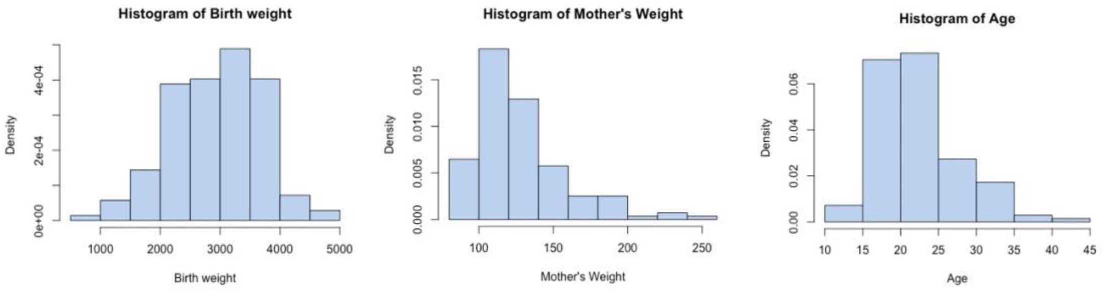
## II Exploratory analysis

We will start with preliminary analysis of data. Firstly, we will take a look on the variable "low", its frequency is: 97 babies – 0 and 42 babies – 1. Meaning, that according to the observations most of the babies have normal weight. Later on, we will try to predict frequency of "low" variable by regression.

[1] GLIM, S.Oman

### *Histograms for the continuous variables*

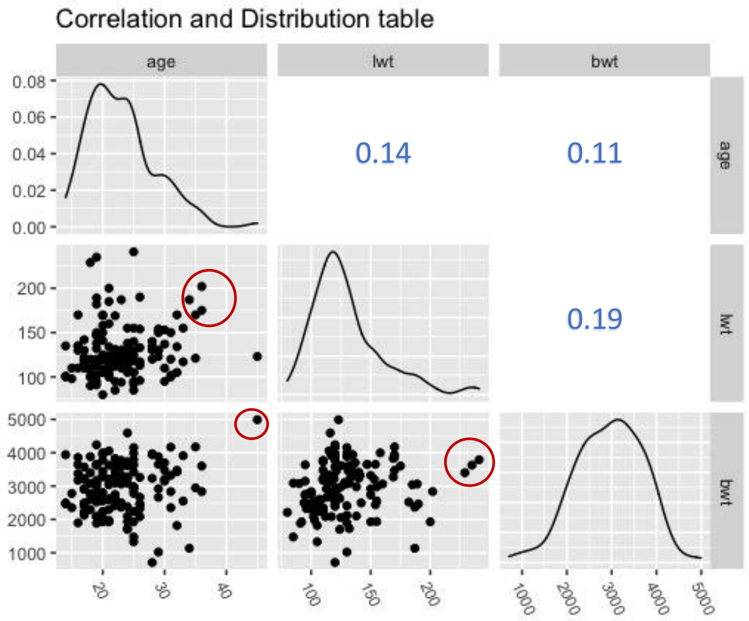
Now, we will do the histogram for the continuous variables to see how they distributed. Variable '*bwt*' has a nice distribution, but variables '*lwt*' – mother weight and '*age*' have heavy right tail (we will fix it later in the paper by applying log function transformations).



```
> my_summary(our_data$bwt,'Birth weight')
$sumry
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    709    2426    2977    2952    3524    4990

> my_summary(our_data$lwt,"Mother's Weight")
$sumry
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   80.0   110.0   121.0   129.4   140.0   241.0

> my_summary(our_data$age,'Age')
$sumry
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   19.00   22.00   22.97   25.00   45.00
```

### *Correlation Table, Joint Distribution of Continuous and Marginal Distribution*

The following figure displays the pairwise correlation between continuous explanatory variables:



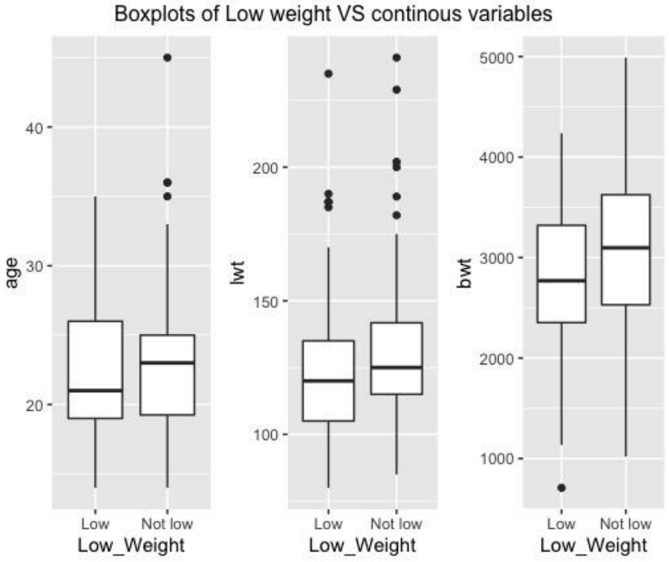
It is clear to see that the *marginal distributions* have the same shape as the histograms presented in the previous part (II).

The *joint distribution* shows us that the biggest part of the observations is situated close to the center (mean) of each continuous variable with slight deviation from the mean. However, we can observe some values that are pretty far away from the mean (have extreme values) – circled in red – they might be outliers, we will check them in the following part of the paper by plotting the residual graph.

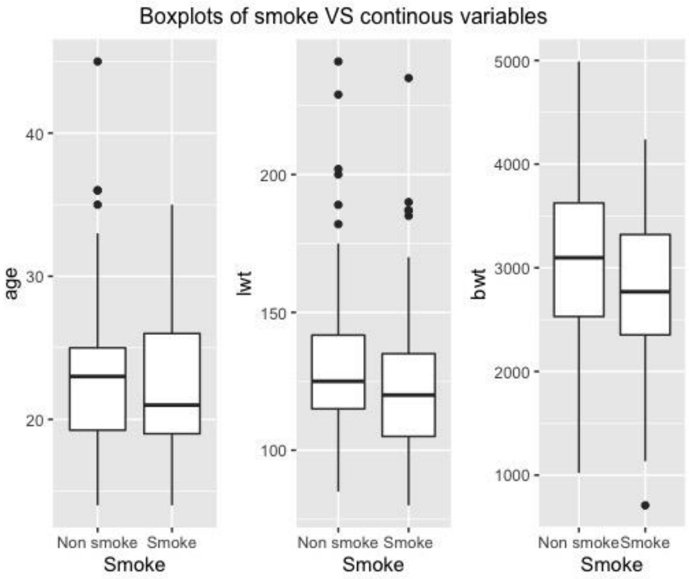
By analyzing correlation table, we can clearly see that the continuous variables have a *positive low correlation between each other*. It is easy to find the logical explanation, since in real world this explanation variables measure totally different and mostly unconnected things, that is why we do not expect some strong connection between mother's age ('*age*'), mother's weight ('*lwt*') and birth weight ('*bwt*'). The fact, that is not obvious is that we tend to think that mother's weight '*lwt*' would have kind of high correlation with baby weight '*bwt*' – that is, by common sense, women with bigger weight would probably have a heavier child. However, by the correlation table we understand that this is a false conclusion, because the weight of the baby, can be also affected by other factors (smoke, father's weight, family pathologies and etc.), which would probably have more strong effect on baby's weight. We will further examine, which explanatory variables have bigger impact on baby weight.

***Boxplots***

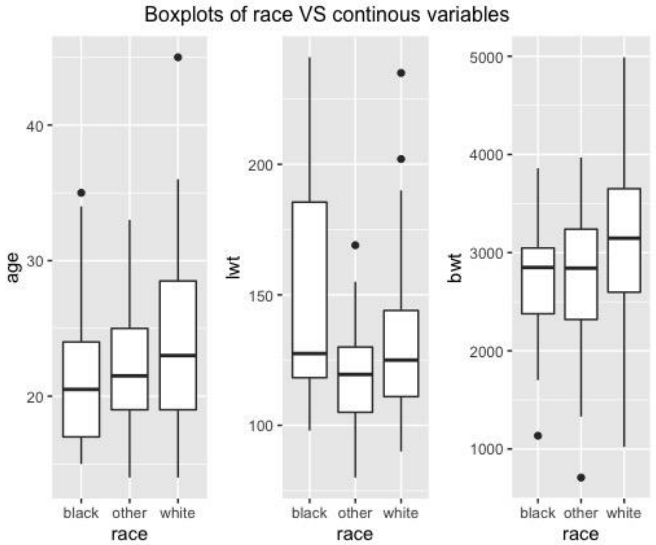
Furthermore, using boxplots we will examine the distribution of the continuous variables with respect to the discrete variables: *'low', 'smoke'* and *'race'*. However, it is needed to be noted that in some categories some of the groups presented in a lesser observation, for e.g., in the 'race' we have 80 observations among white race and only 20 among black race, therefore, we will analyze boxplots together with cross tables to understand the picture in the fuller manner.



- By analyzing the above boxplots, we can see that the mean age of a mother who gave birth to low weighted baby is slightly lower (approx. 2 years) than the mean age of a mother who gave birth to baby with not low weight. Nonetheless, the distribution of low and not low weight over the '*Age*' variable is quite similar.
- Moreover, we see that the average of mother's weight who gave birth to low and not low weighted baby are almost the same as their distributions.
- The 3rd boxplot does not hold any new worth information since it holds '*low*' variable distribution.



- 1st boxplots shows that the mean of smoking women is lower with regards to age comparing to non-smoking women, meaning that more women smoke in early ages.
- 2nd boxplots shows that the mean between smoking and non-smoking women with regards to women's weight are almost similar as well as their distribution.
- From the 3rd boxplot we can draw the important conclusion: that non-smoking women have baby with mean weight higher comparing to smoking women. The distribution of '*bwt'* among smoke and non-smoke women are the same.

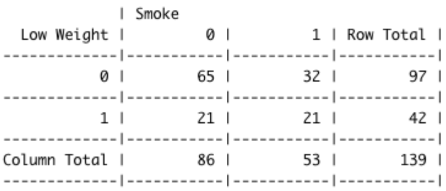


- From 1st boxplot it can be seen that the mean age of black mothers is lower comparing to white/other races, which are similar. Hence, according to our data we can make a conclusion that black women tend to give birth in earlier ages comparing to white/other races.
- 2nd boxplot shows that the women weight means among all of the races quite the same. However, from the distribution we can clearly see that black mothers are tend to have biggest weight (might be due to nutrition preferences, traditional cuisines and etc. or from the previous boxplot fact that black women tend to give birth in earlier ages). Regarding other races it is hard to make any conclusion since it includes variety of socio-economic and other factors, that result in high variance in this group.
- However, the 3rd plot shows that the baby weight mean is lowest among black race, the highest mean has white race and between them lies baby weight mean among other races.

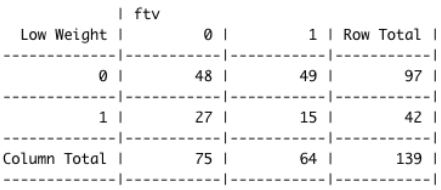
The paradox that we face is that black women tend to have bigger weight, but do not have the highest weighted babies leads us to the conclusion that there are might be factors that have more significant influence on the baby weight rather than mother's weight.

### *Cross tables*

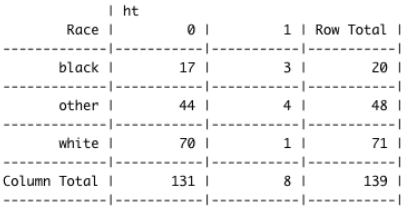
By using cross-tables we will compare the joint distribution of discrete variables regarding each of them:



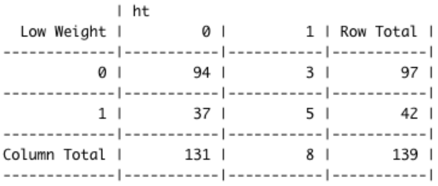
```
             | Smoke
  Low Weight |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        65 |        32 |        97 |
-------------|-----------|-----------|-----------|
           1 |        21 |        21 |        42 |
-------------|-----------|-----------|-----------|
Column Total |        86 |        53 |       139 |
-------------|-----------|-----------|-----------|
```



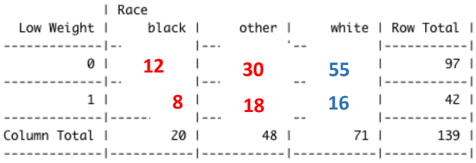
```
             | ftv
  Low Weight |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        48 |        49 |        97 |
-------------|-----------|-----------|-----------|
           1 |        27 |        15 |        42 |
-------------|-----------|-----------|-----------|
Column Total |        75 |        64 |       139 |
-------------|-----------|-----------|-----------|
```



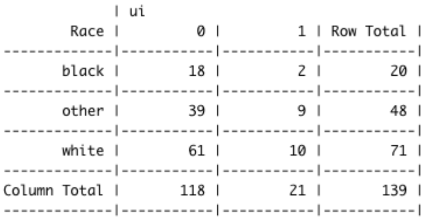
```
             | ht
        Race |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
       black |        17 |         3 |        20 |
-------------|-----------|-----------|-----------|
       other |        44 |         4 |        48 |
-------------|-----------|-----------|-----------|
       white |        70 |         1 |        71 |
-------------|-----------|-----------|-----------|
Column Total |       131 |         8 |       139 |
-------------|-----------|-----------|-----------|
```



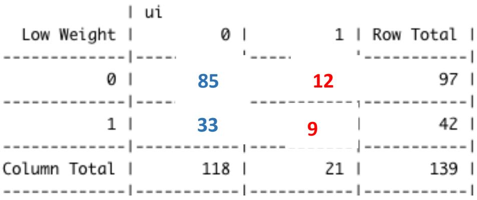
```
             | ht
  Low Weight |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        94 |         3 |        97 |
-------------|-----------|-----------|-----------|
           1 |        37 |         5 |        42 |
-------------|-----------|-----------|-----------|
Column Total |       131 |         8 |       139 |
-------------|-----------|-----------|-----------|
```



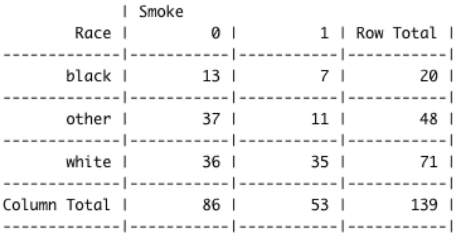
```
             | Race
  Low Weight |     black |     other |     white | Row Total |
-------------|-----------|-----------|-----------|-----------|
           0 |        12 |        30 |        55 |        97 |
-------------|-----------|-----------|-----------|-----------|
           1 |         8 |        18 |        16 |        42 |
-------------|-----------|-----------|-----------|-----------|
Column Total |        20 |        48 |        71 |       139 |
-------------|-----------|-----------|-----------|-----------|
```



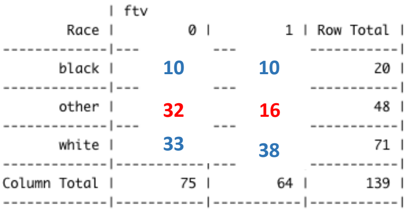
```
             | ui
        Race |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
       black |        18 |         2 |        20 |
-------------|-----------|-----------|-----------|
       other |        39 |         9 |        48 |
-------------|-----------|-----------|-----------|
       white |        61 |        10 |        71 |
-------------|-----------|-----------|-----------|
Column Total |       118 |        21 |       139 |
-------------|-----------|-----------|-----------|
```



```
             | ui
  Low Weight |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        85 |        12 |        97 |
-------------|-----------|-----------|-----------|
           1 |        33 |         9 |        42 |
-------------|-----------|-----------|-----------|
Column Total |       118 |        21 |       139 |
-------------|-----------|-----------|-----------|
```



```
             | Smoke
        Race |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
       black |        13 |         7 |        20 |
-------------|-----------|-----------|-----------|
       other |        37 |        11 |        48 |
-------------|-----------|-----------|-----------|
       white |        36 |        35 |        71 |
-------------|-----------|-----------|-----------|
Column Total |        86 |        53 |       139 |
-------------|-----------|-----------|-----------|
```



```
             | ftv
        Race |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
       black |        10 |        10 |        20 |
-------------|-----------|-----------|-----------|
       other |        32 |        16 |        48 |
-------------|-----------|-----------|-----------|
       white |        33 |        38 |        71 |
-------------|-----------|-----------|-----------|
Column Total |        75 |        64 |       139 |
-------------|-----------|-----------|-----------|
```

From the cross-tables we can derive several conclusions:

a) If we examine the race of the mother we can see that almost 50% of babies that were born from black mothers are low weighted (8 babies out of 20), while white women have low weighted babies only in around 22% (16 babies out of 71), for other races women this percentage varies around 37% (18 babies out of 48).

b) In case of presence of uterine irritability *ui*, it is more likely for baby to have lower weight. Our data shows that out of 21 babies who have *ui* 9 are low weighted, which makes it almost 50%, while among 118 babies without *ui* only 33 are low weighted, which is only 27%.

c) Black and White women are smoking in 50% of cases. In other races it’s around 23%.

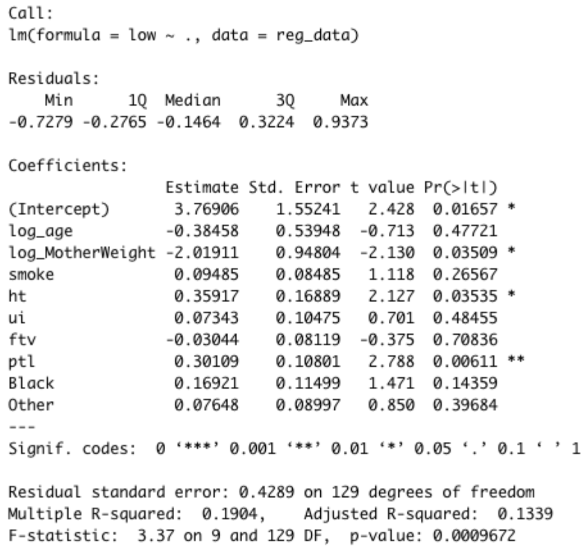
## III Model fitting

As we already explored in histograms for continuous variables, that before applying regression we need to proceed the log transformation on '*age'* and '*lwt'* columns in order to get rid of heavy right tails.

We also denote that in our model the white race categorical variable will be omitted, but will affect the intersection. It is more informative and easier to read and understand, rather than a model with this variable but without intersection.

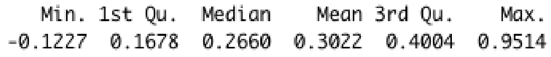
Certainly, we will not plug '*bwt'* given explanatory variable, because it is the exact information that we are trying to predict by regression.

### *Linear regression*



```
Call:
lm(formula = low ~ ., data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7279 -0.2765 -0.1464  0.3224  0.9373 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)   
(Intercept)       3.76906    1.55241   2.428  0.01657 * 
log_age          -0.38458    0.53948  -0.713  0.47721   
log_MotherWeight -2.01911    0.94804  -2.130  0.03509 * 
smoke             0.09485    0.08485   1.118  0.26567   
ht                0.35917    0.16889   2.127  0.03535 * 
ui                0.07343    0.10475   0.701  0.48455   
ftv              -0.03044    0.08119  -0.375  0.70836   
ptl               0.30109    0.10801   2.788  0.00611 **
Black             0.16921    0.11499   1.471  0.14359   
Other             0.07648    0.08997   0.850  0.39684   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4289 on 129 degrees of freedom
Multiple R-squared:  0.1904,	Adjusted R-squared:  0.1339 
F-statistic:  3.37 on 9 and 129 DF,  p-value: 0.0009672
```

Predicted values:



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.1227  0.1678  0.2660  0.3022  0.4004  0.9514 
```

Linear regression does not fit our data and we will try GLM.

### *Generalized linear model*

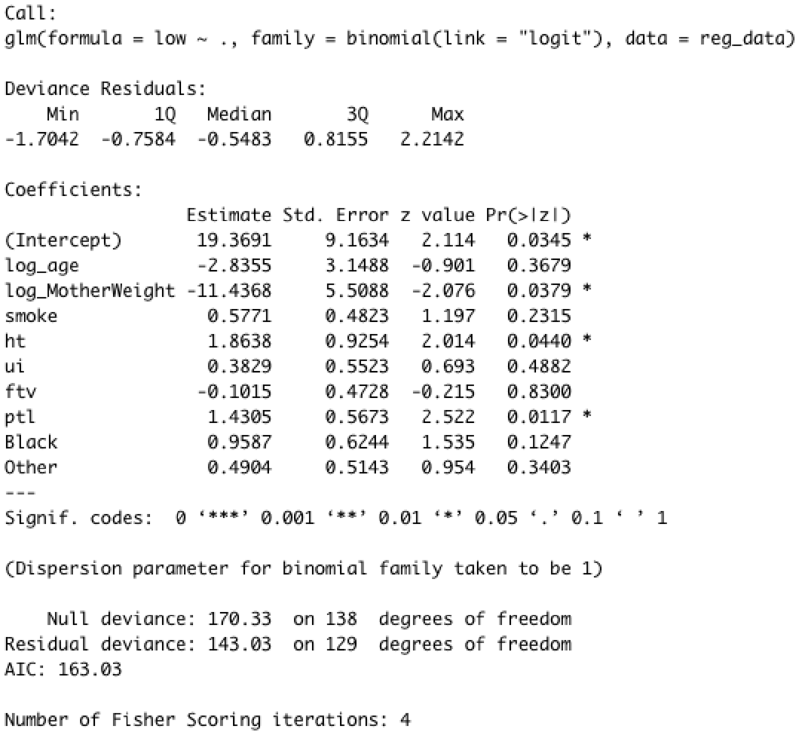
In our case we have 'low' – binary outcome variable, we will apply generalized linear model with a logit link function. That is, probability of baby i to be low weighted, given his covariates vector $x_i \in R^p$, is:

$$\pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

Meaning that log-odds for low weighted baby is: $\eta_i = log(\frac{\pi_i}{1-\pi_i}) = \sum_{j=1}^{p} x_{ij}\beta_j$.

We can consider to use probit link function, which is $\phi^{-1}(\pi_i)$. However, it is easier to interpret logit link function coefficient (The fitted values are the same in approximate values in both models).

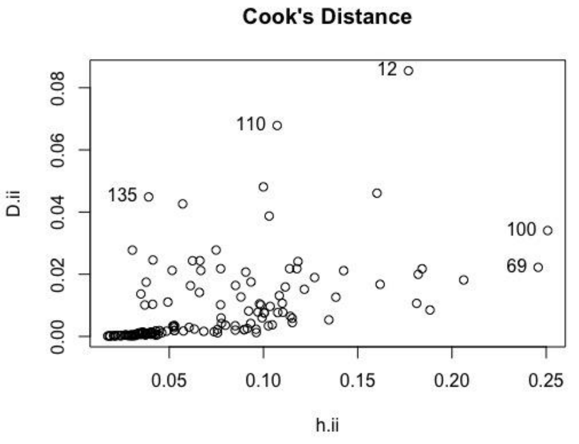
### *Generalized linear model logit regression*

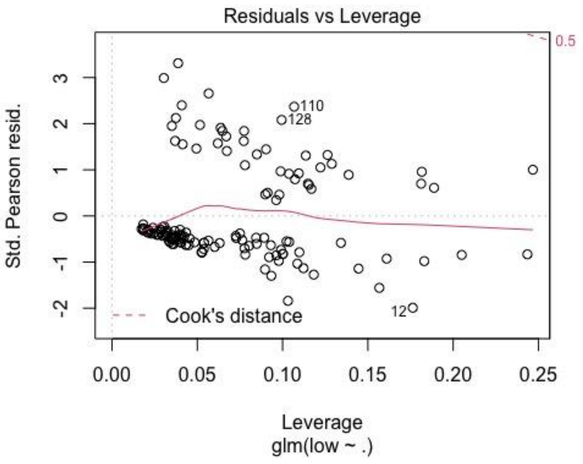


```
Call:
glm(formula = low ~ ., family = binomial(link = "logit"), data = reg_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7042  -0.7584  -0.5483   0.8155   2.2142  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)  
(Intercept)       19.3691     9.1634   2.114   0.0345 *
log_age           -2.8355     3.1488  -0.901   0.3679  
log_MotherWeight -11.4368     5.5088  -2.076   0.0379 *
smoke              0.5771     0.4823   1.197   0.2315  
ht                 1.8638     0.9254   2.014   0.0440 *
ui                 0.3829     0.5523   0.693   0.4882  
ftv               -0.1015     0.4728  -0.215   0.8300  
ptl                1.4305     0.5673   2.522   0.0117 *
Black              0.9587     0.6244   1.535   0.1247  
Other              0.4904     0.5143   0.954   0.3403  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.33  on 138  degrees of freedom
Residual deviance: 143.03  on 129  degrees of freedom
AIC: 163.03

Number of Fisher Scoring iterations: 4
```

By running the logit GLM regression we can see which coefficients are significant. For e.g., *log_MotherWeight*, *ht* and *ptl* variables have a significant influence on the response variable. For further improvement of our model, we will work with residuals and outliers.

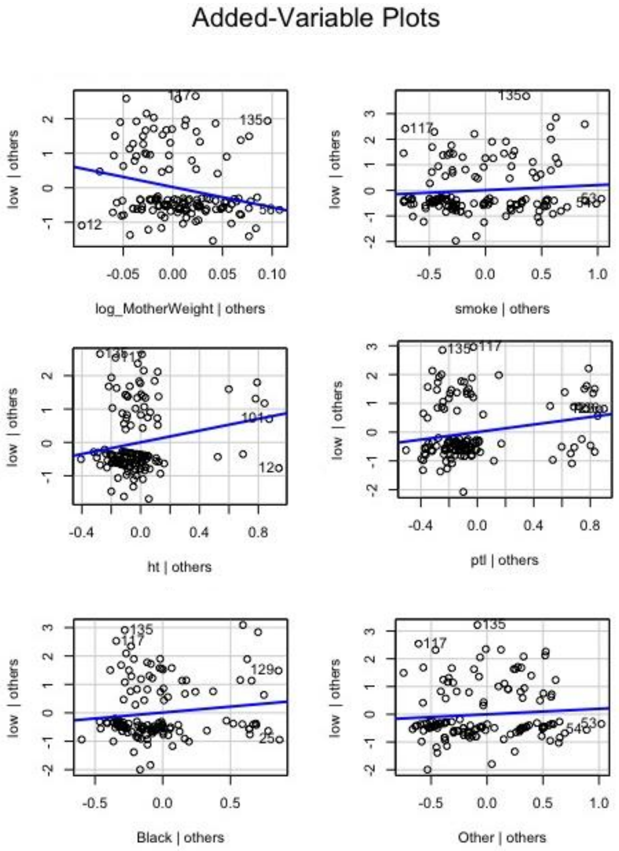
### *Residuals and Outliers, Cook distance*





The Cook’s distance plotted against the hat values can also help us to indicate outliers. From the above plot we can see, that observations 12,128,100,110,135 hold extraordinary values, therefore we also suspect them to be outlier. To check our assumption, we will proceed with added-variable plots.

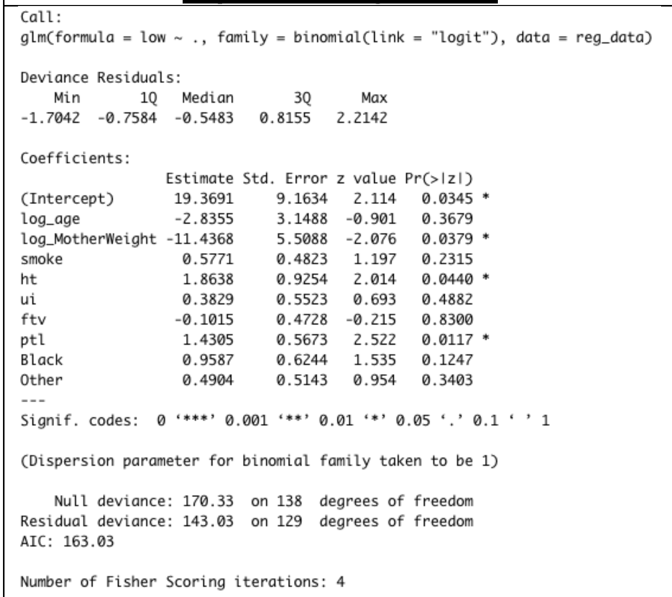
### *Added-variable plots*



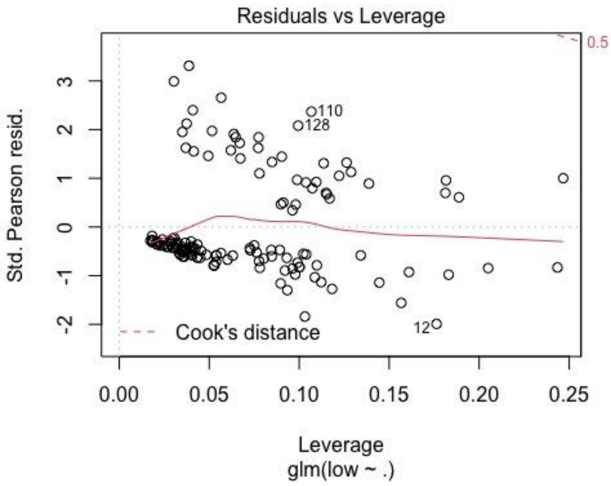
Combining together conclusions from residual analysis, Cook’s distance and AV plots I run the plots without suspected observation (and with different combination with/without some of them). Therefore, I came to the conclusion, that observations ***117, 135, 110*** are our outliners, without them we get the better model. That is why, we have decided to omit them in our analysis to find a better model fit.

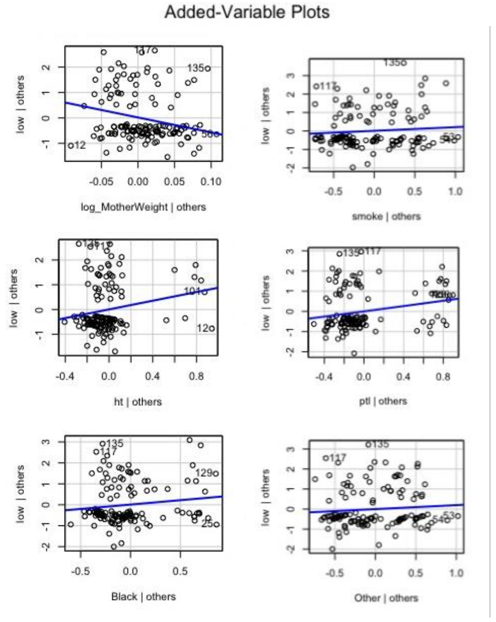
### *Generalized linear model logit regression after omitting outliers*

## ***Before omitting outliers***

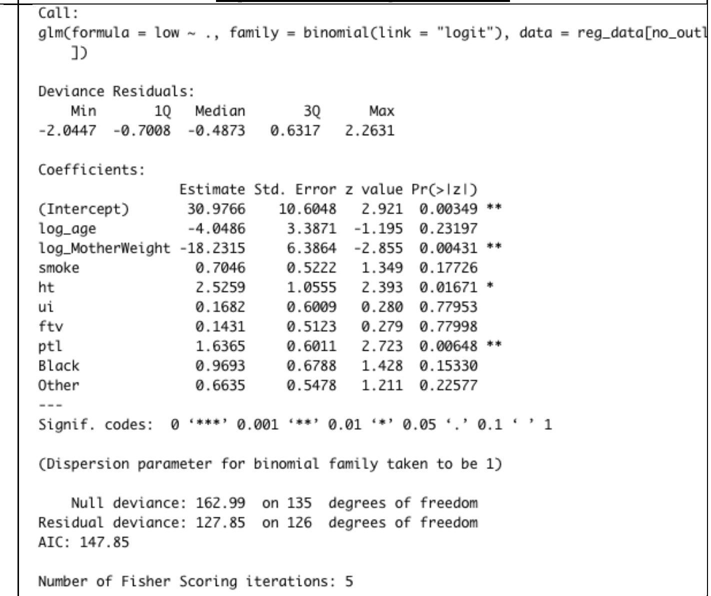


```
Call:
glm(formula = low ~ ., family = binomial(link = "logit"), data = reg_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7042  -0.7584  -0.5483   0.8155   2.2142

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         19.3691     9.1634   2.114   0.0345 *
log_age             -2.8355     3.1488  -0.901   0.3679
log_MotherWeight -11.4368      5.5088  -2.076   0.0379 *
smoke                0.5771     0.4823   1.197   0.2315
ht                   1.8638     0.9254   2.014   0.0440 *
ui                   0.3829     0.5523   0.693   0.4882
ftv                 -0.1015     0.4728  -0.215   0.8300
ptl                  1.4305     0.5673   2.522   0.0117 *
Black                0.9587     0.6244   1.535   0.1247
Other                0.4904     0.5143   0.954   0.3403
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.33  on 138  degrees of freedom
Residual deviance: 143.03  on 129  degrees of freedom
AIC: 163.03

Number of Fisher Scoring iterations: 4
```

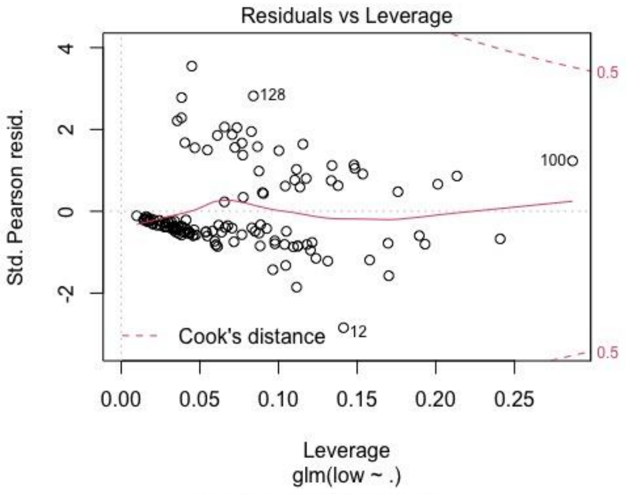


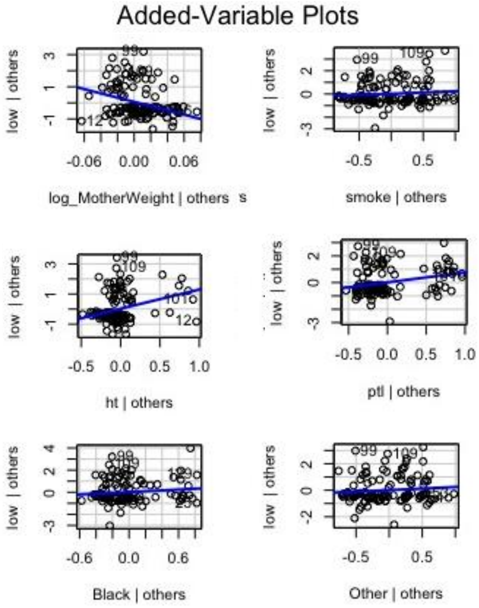


## ***After omitting outliers***



```
Call:
glm(formula = low ~ ., family = binomial(link = "logit"), data = reg_data[no_outl
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0447  -0.7008  -0.4873   0.6317   2.2631

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       30.9766    10.6048   2.921  0.00349 **
log_age           -4.0486     3.3871  -1.195  0.23197
log_MotherWeight -18.2315     6.3864  -2.855  0.00431 **
smoke              0.7046     0.5222   1.349  0.17726
ht                 2.5259     1.0555   2.393  0.01671 *
ui                 0.1682     0.6009   0.280  0.77953
ftv                0.1431     0.5123   0.279  0.77998
ptl                1.6365     0.6011   2.723  0.00648 **
Black              0.9693     0.6788   1.428  0.15330
Other              0.6635     0.5478   1.211  0.22577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.99  on 135  degrees of freedom
Residual deviance: 127.85  on 126  degrees of freedom
AIC: 147.85

Number of Fisher Scoring iterations: 5
```





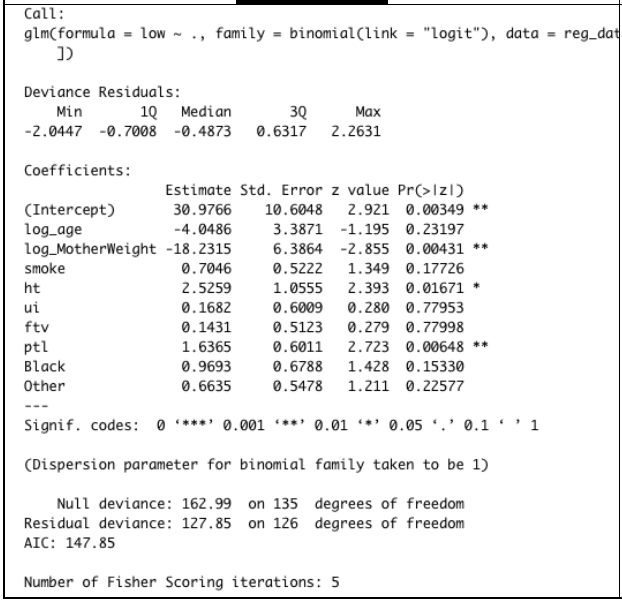
From the above comparison, we see that the coefficients did not change dramatically and the signs (positive/negative) remain the same for most of the coefficients. But what is more important, we can see how coefficients became more significant. Now we can clearly see which coefficients has the biggest significance. Moreover, the residuals vs leverages and AV plot looks a little bit better after removing outliers.

We can clearly see that AIC had a significant drop in 15 units from 163 to 147.85. Taking everything into consideration, we will choose model with 136 observations (omitting 3 observation) as the final model, so we will still have enough observations avoiding overfitting.

Furthermore, in order to fit our model even better we will get rid of less significant explanatory variables. We will use AIC backwards selection method, trying to minimize AIC step by step after every run. The advantage of AIC that it deals with the trade-off between the goodness of fit of the model and the simplicity of the model, consequently, it deals with both the risk of overfitting and the risk of underfitting.

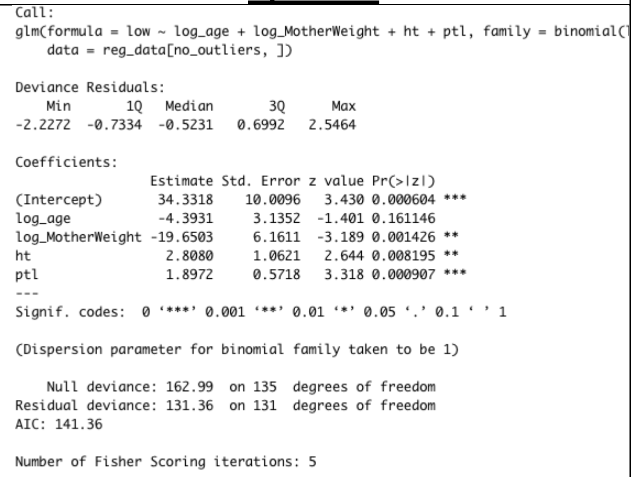
***Generalized linear model logit regression after AIC backwards selection***

***<u>Before AIC</u>***



```
Call:
glm(formula = low ~ ., family = binomial(link = "logit"), data = reg_dat
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0447  -0.7008  -0.4873   0.6317   2.2631

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       30.9766    10.6048   2.921  0.00349 **
log_age           -4.0486     3.3871  -1.195  0.23197
log_MotherWeight -18.2315     6.3864  -2.855  0.00431 **
smoke              0.7046     0.5222   1.349  0.17726
ht                 2.5259     1.0555   2.393  0.01671 *
ui                 0.1682     0.6009   0.280  0.77953
ftv                0.1431     0.5123   0.279  0.77998
ptl                1.6365     0.6011   2.723  0.00648 **
Black              0.9693     0.6788   1.428  0.15330
Other              0.6635     0.5478   1.211  0.22577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.99  on 135  degrees of freedom
Residual deviance: 127.85  on 126  degrees of freedom
AIC: 147.85

Number of Fisher Scoring iterations: 5
```

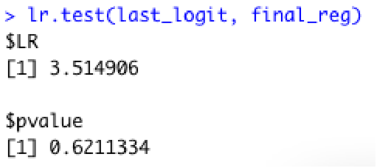
***<u>After AIC</u>***



```
Call:
glm(formula = low ~ log_age + log_MotherWeight + ht + ptl, family = binomial(
    data = reg_data[no_outliers, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2272  -0.7334  -0.5231   0.6992   2.5464

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       34.3318    10.0096   3.430 0.000604 ***
log_age           -4.3931     3.1352  -1.401 0.161146
log_MotherWeight -19.6503     6.1611  -3.189 0.001426 **
ht                 2.8080     1.0621   2.644 0.008195 **
ptl                1.8972     0.5718   3.318 0.000907 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.99  on 135  degrees of freedom
Residual deviance: 131.36  on 131  degrees of freedom
AIC: 141.36

Number of Fisher Scoring iterations: 5
```

The start AIC was 147.4, in the final run as we see the AIC is 141.36, which gave us great difference in ~6. Nevertheless, we also checked the model manually: checking the regression after every step of R AIC selection process to make sure that none of the significant variables were automatically omitted by R without solid reasons. However, the given combination turned out to be the best fit with optimal AIC, p_values of the explanatory variables. Moreover, we checked the model for interactions as well, however, I didn't receive any significant ones, therefore, we leave the model without any interactions.

Furthermore, we need to justify that we don't reject the hypothesis that our sub-model is the better fit compared to the full model. So, we compare the full model to sub-model using *likelihood ratio test*, which gave us the following output:



```
> lr.test(last_logit, final_reg)
$LR
[1] 3.514906

$pvalue
[1] 0.6211334
```

As we can see our p_value ≫ 0.1≫0.05 as well as LR test output is big enough to "legally" justify that *our sub-model is better for our data than the full model.*

Next step we need to check that the assumption about binomial family, that the estimated dispersion parameter $\phi$ is 1 was held in our case as well. Therefore, we proceed:

$\hat{\phi} := \sum(Pearson\ residuals)^2\ /df$ = 131.36/131 = 1.003≈ 1. Hence, the assumption is held.

# IV Conclusion

In this paper we were examining which factors (explanatory variables) and to what extend affect chances for mother to have a low-weighted baby (less 2.5 kg). I started the paper with the preliminary exploratory analysis of the data, including histograms for explanatory continuous variables to see their distribution, cross-tables, correlations between the variables. Further, we proceeded with formulation of a GLIM which appears appropriate for the problem. Following with model fitting, including testing relevant hypotheses, comparing different sub-models and analyzing residuals and outliers. Therefore, we came to the final conclusion and now we can see the coefficients of the final selected model:

| **Coefficient** | *Intercept* | *log_age* | *log_MotherWeight* | *ht* | *ptl* |
|---|---|---|---|---|---|
| **Est.Value** | 34.3318 | -4.3931 | -19.6503 | 2.808 | 1.8972 |

Meaning*: log* $\left(\frac{\pi_i}{1-\pi_i}\right)$ *= 34.3318 – 4.3931****Log_age*** *– 19.6503****Log_MotherWeight*** *+*
*+2.808**$\mathbf{I}_{\{patient\ i\ has\ hypertension\}}$ *+ 1.8972**$\mathbf{I}_{\{patient\ i\ had\ premature\ labors\}}$

This model gives us understanding that the effect of **2.718 units change in *log_age*** (other variables stay the same) **decreasing** the odds to born a low-weighted baby by **4.3931**. Regarding categorical variables, for a patient *with hypertension* the log odds increase by **2.808** and for a patient, that *had premature labors* log odds increase by **1.8972**.

Moreover, as we noticed in preliminary analysis and now we see it by the coefficient signs here as well, that increase in mother age and mother weight decrease the chances to born low weighted baby, while the presence of hypertension or premature labors will increase the chances to born a low weighted baby.

This model, based on 136 observations, can help us to recognize on the stage of a women pregnancy the probability of a baby to be born low-weighted and therefore to provide the needed treatment and help the baby to be born with normal weight (which will higher its chances to survive).