# GENERALIZED LINEAR MODELS- 52542

## Data on Baby Birth Weights

האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

From: Amit Yaron

i.d.: 205560618

To: Professor Samuel Oman

# Contents

## Background of the data

Infant mortality is higher for low birth-weight babies.

A 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.

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.

### Description of the data

The dataset consists of the following 10 variables

1. **low:** Indicator of birth weight less than 2.5kg .
2. **age:** Mother's age in years.
3. **lwt**: Mother's weight in pounds at last menstrual period.
4. **race**: Mothers' race ("white", "black", "other").
5. **smoke:** Smoking status during pregnancy
6. **ht**: History of hypertension
7. **ui:** Presence of uterine irritability
8. **ftv:** Number of physician visits during the first trimester
9. **ptl:** Number of previous premature labor's
10. **bwt:** Birth weight in grams

for practical reasons, I change the categorical variable "race" to indicator "white" if the mothers' race is white or black and other.

**Preliminary Analysis of Medfly Data:**

## The Research Question

Which variables increase the risk of low birth weight?

I will focus on how the mother's actions and health during pregnancy affect the birth weight.

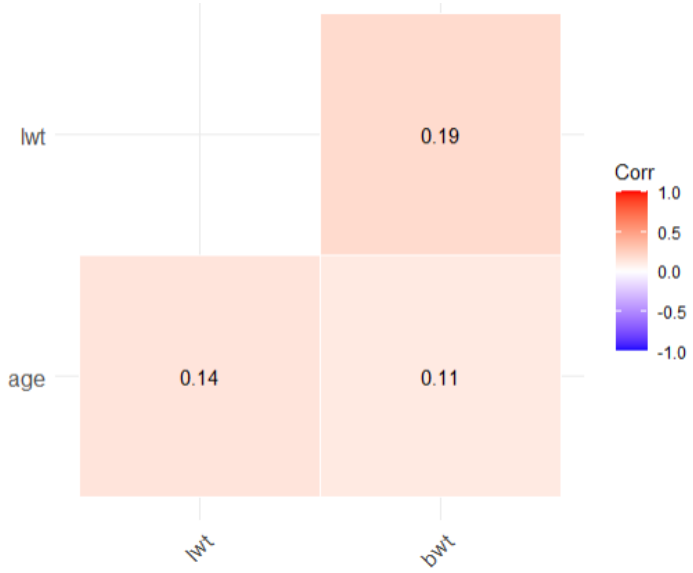
On GLM model we have the Explanatory variables and Explained variable.

The Explained variable will be the **low:** the Indicator of birth weight less than 2.5kg .

## Summary of the Variables

```
      low              age              lwt             white            smoke              ht                ui               ftv              ptl        
 Min.   :0.0   Min.   :14   Min.   : 80   Min.   :0.0   Min.   :0.0   Min.   :0.00   Min.   :0.0   Min.   :0.0   Min.   :0.0  
 1st Qu.:0.0   1st Qu.:19   1st Qu.:110   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.00   1st Qu.:0.0   1st Qu.:0.0   1st Qu.:0.0  
 Median :0.0   Median :22   Median :121   Median :1.0   Median :0.0   Median :0.00   Median :0.0   Median :0.0   Median :0.0  
 Mean   :0.3   Mean   :23   Mean   :129   Mean   :0.5   Mean   :0.4   Mean   :0.06   Mean   :0.2   Mean   :0.5   Mean   :0.2  
 3rd Qu.:1.0   3rd Qu.:25   3rd Qu.:140   3rd Qu.:1.0   3rd Qu.:1.0   3rd Qu.:0.00   3rd Qu.:0.0   3rd Qu.:1.0   3rd Qu.:0.0  
 Max.   :1.0   Max.   :45   Max.   :241   Max.   :1.0   Max.   :1.0   Max.   :1.00   Max.   :1.0   Max.   :1.0   Max.   :1.0  
      bwt       
 Min.   : 709  
 1st Qu.:2426  
 Median :2977  
 Mean   :2952  
 3rd Qu.:3524  
 Max.   :4990  
```

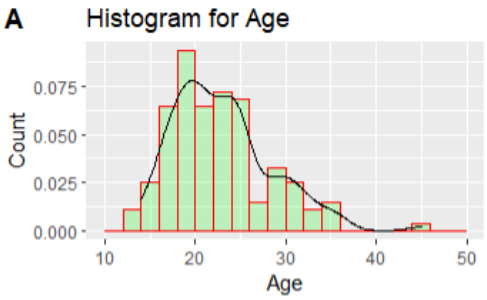
## Correlation matrix of continuous variables

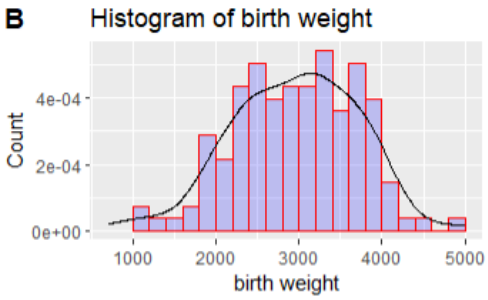


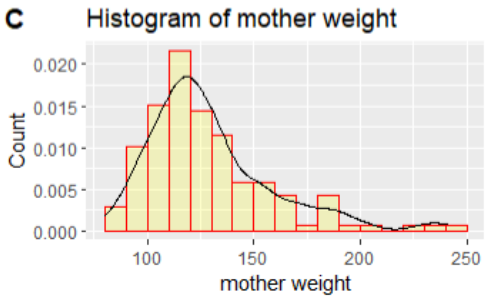

It is seen that there is no strong correlation between the continuous variable.

So age, lwt , bwt will be on the GLM model because there not suspicion of multinational culinary.

## Histograms of the Variables




**Histograms of continuous variables:**

By examining the histograms and the density graphs.

I can say that the variable 'bwt' is distributed approximately normal.

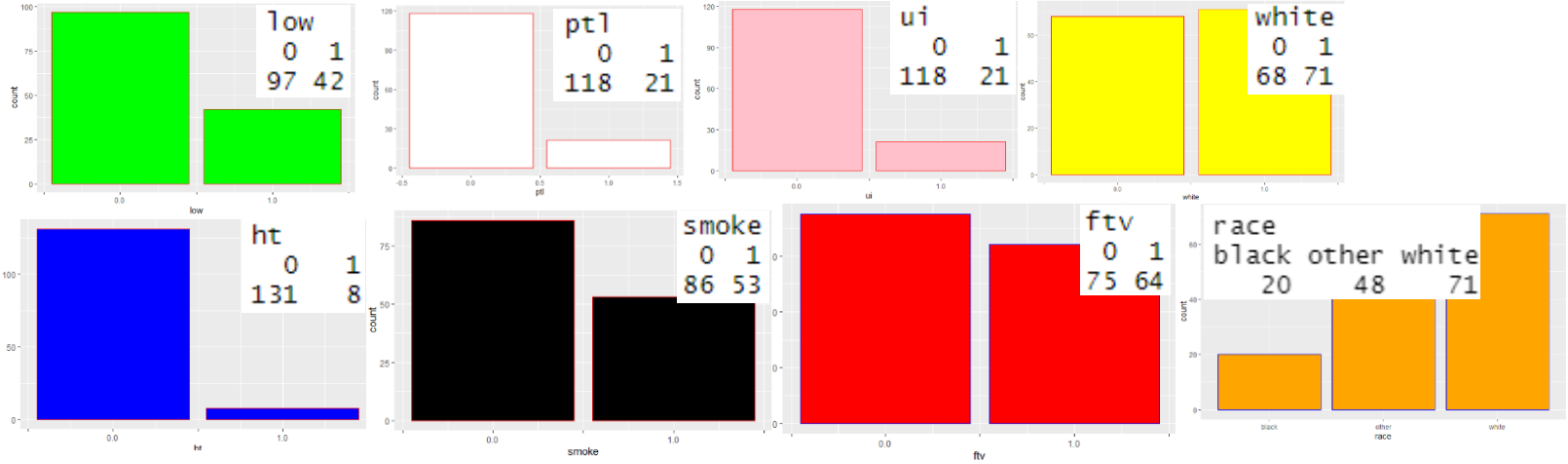
$$bwt \sim N(2952, 753.76)$$

The variables 'lwt' and 'age' are approximately divided chi-squared.

$$lwt \sim \chi_n^2 \quad age \sim \chi_n^2$$

It can see also that most of the mothers are between the ages of 17-30 (also observation that we suspect is unusual in terms of the age range of this current mothers sample), and most of them weighed between 100-140 k.

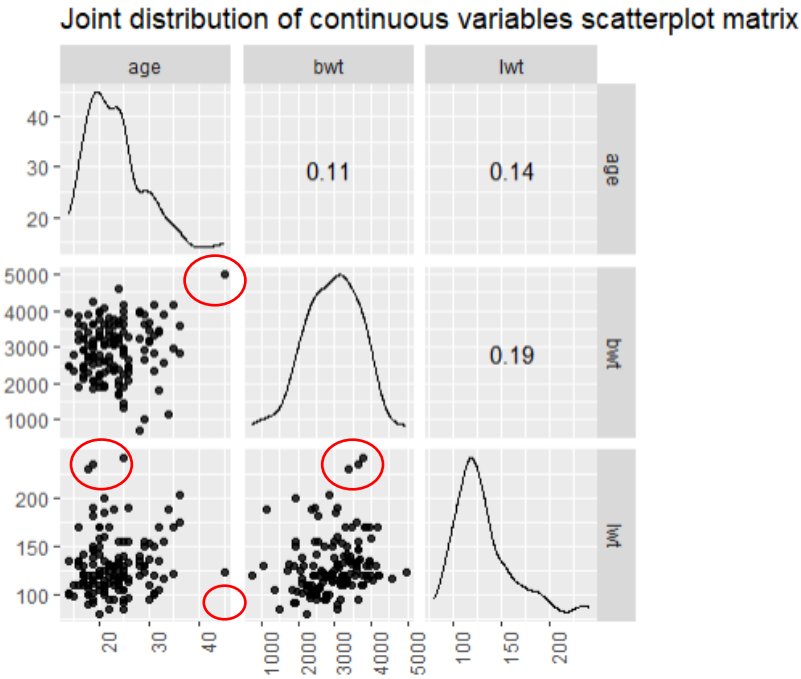
## proportion of the indicator variables




- **Most of the mothers have baby birth weight** with more than 2.5kg.
- The mothers who visit the doctor in the first trimester of pregnancy and the mothers who does not visit the doctor in the first trimester are almost equal, with a small difference.
- **Most of the mothers does not have any history of hypertension** or previous premature labours.
- The proportion of the mothers who smoke and who does not smoke it about (3: 2).
  I know form external sources of knowledge, there are studies that show a relationship between smoking while pregnant and history of hypertension or infant health. **But here I don't see the relationship in our graph**

So I conclude that taking a sample in which unbalanced variables can then impair the reliability of the statistical model

## Joint distribution of continuous variables:




**I see there is no strong correlation between the continuous variables**
because Pearson correlation is very low:
0.11,0.14,0.19 .
Also, there is no suspicion of multinational culinary, because I see that the distribution of variables does not have a clear trend but is more like a random cloud.

here we can see side points we Suspicious it outlier in the next sections we check it

**The joint distribution of categorical variables with Frequency tables and cross tables**

I use here Frequency tables for categorical variables and cross tables

| white\smoke | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 50 | 18 | 68 |
| 1 | 36 | 35 | 71 |
| Total col | 86 | 53 | 139 |

| White\Smoke | 0 | 1 |
|---|---|---|
| 0 | 0.3597122 | 0.1294964 |
| 1 | 0.2589928 | 0.2517986 |

| white\low | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 42 | 26 | 68 |
| 1 | 55 | 16 | 71 |
| Total col | 97 | 42 | 139 |

| White\low | 0 | 1 |
|---|---|---|
| 0 | 0.3021583 | 0.1870504 |
| 1 | 0.3956835 | 0.1151079 |

| low\smoke | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 65 | 32 | 97 |
| 1 | 21 | 21 | 71 |
| Total col | 86 | 53 | 139 |

| low\smoke | 0 | 1 |
|---|---|---|
| 0 | 0.4676259 | 0.2302158 |
| 1 | 0.1510791 | 0.1510791 |

| low\ht | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 94 | 3 | 97 |
| 1 | 37 | 5 | 71 |
| Total col | 131 | 8 | 139 |

| low\ht | 0 | 1 |
|---|---|---|
| 0 | 0.676259 | 0.0215827 |
| 1 | 0.1510791 | 0.0359712 |

| low\ui | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 85 | 12 | 97 |
| 1 | 33 | 9 | 71 |
| Total col | 118 | 21 | 139 |

| low\ui | 0 | 1 |
|---|---|---|
| 0 | 0.6115108 | 0.0863309 |
| 1 | 0.2374101 | 0.0647482 |

| low\ptl | 0 | 1 | Total row |
|---|---|---|---|
| 0 | 89 | 8 | 97 |
| 1 | 29 | 13 | 71 |
| Total col | 118 | 21 | 139 |

| low\ptl | 0 | 1 |
|---|---|---|
| 0 | 0.6402878 | 0.057554 |
| 1 | 0.2086331 | 0.0935252 |

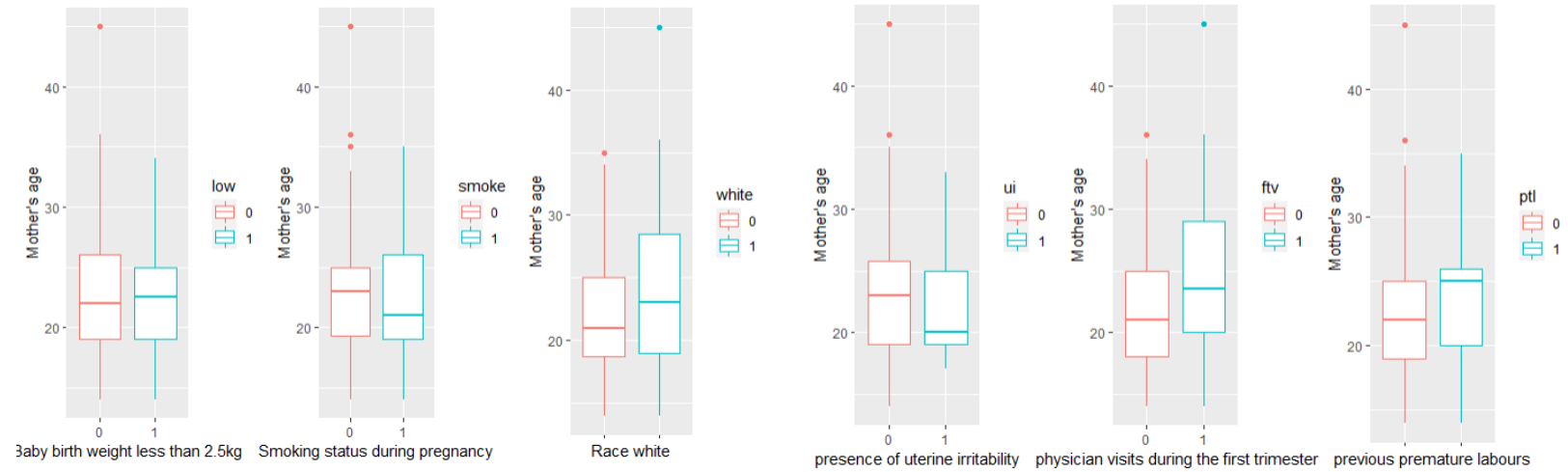
We can see there is a strong connection between a healthy lifestyle include not smoke, without a history of hypertension, without uterine irritability, and baby birth weight bigger than 2.5 km. which makes sense.

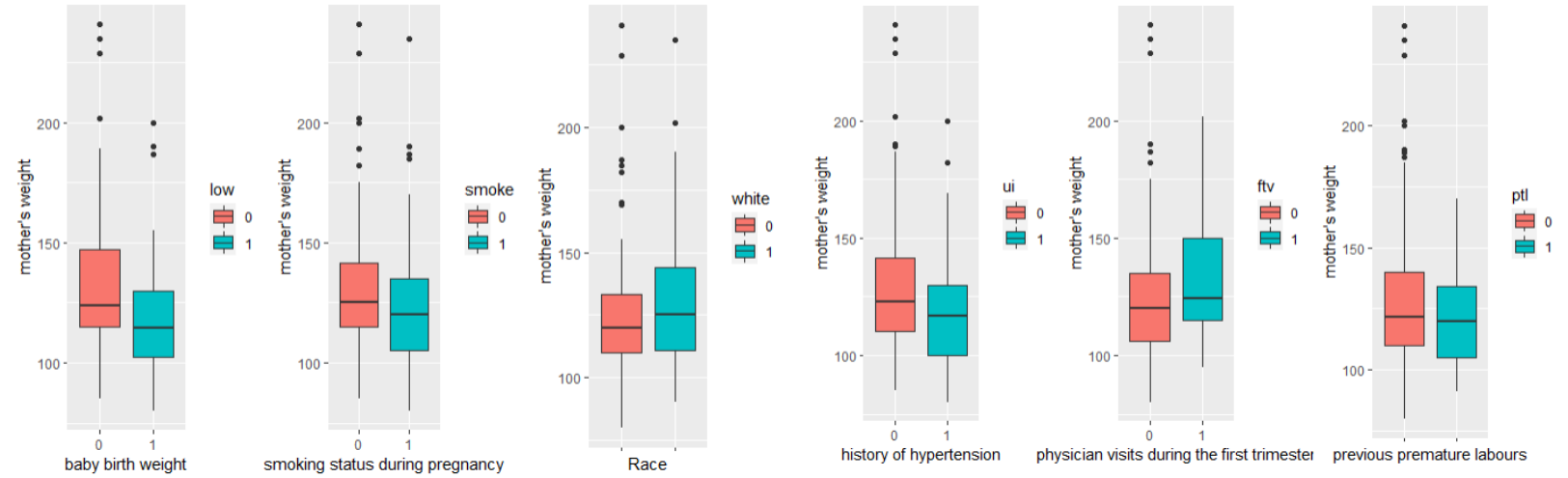
Most women who smoke during pregnancy are white. Also, there is more mother with non-race white with low baby birth Compared to white mothers.

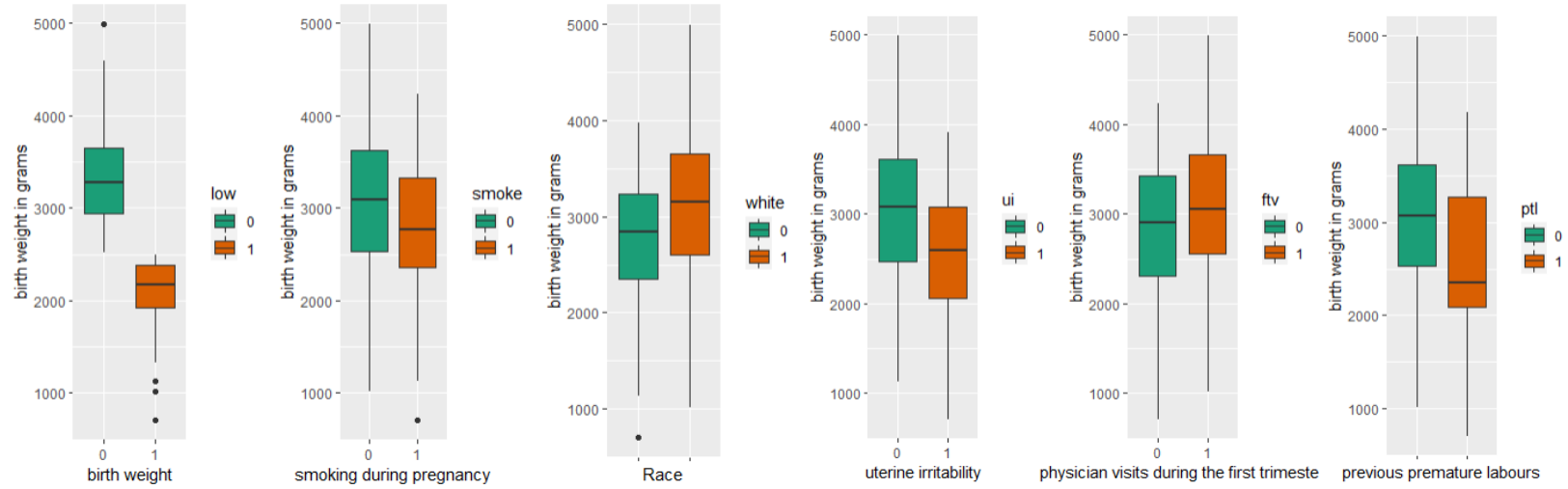
Low birth weights are found in most cases of women who have a history of hypertension (ht)

Finally, There were no significant findings in the other tables

## Boxplot about the relation between the continuous variables to the category variables



We can see from the boxplot those conclusions:

- The older mothers get those properties: with less baby birth less than 2.5 kg, smoking during pregnancy, non-white, without a history of hypertension, don't visit the physician during the first trimester, and finally with less uterine irritability.
- The mothers with higher weight get those properties: with less baby birth less then 2.5 kg, smoking during pregnancy, non-white, with a history of hypertension, don't visit the physician during the first trimester, and finally with previous premature labors.-
- The higher baby weight gets those properties: their mothers don't smoke during pregnancy, their race is white with less uterine irritability, and visit the physician during the first trimester finally with less uterine irritability

## Simple OLS regression:

I Estimate the explained variable "low": If the baby is under 2.5kg or not?

First I start with the response variable "bwt" baby birth weight

We can see the variables that seem to be the most influential are:

Mother's race(white or not ), mother's weight, smoking status during pregnancy, history of hypertension, and presence of uterine irritability.

The in Multiple R-squared, Adjusted R-squared represents that this model

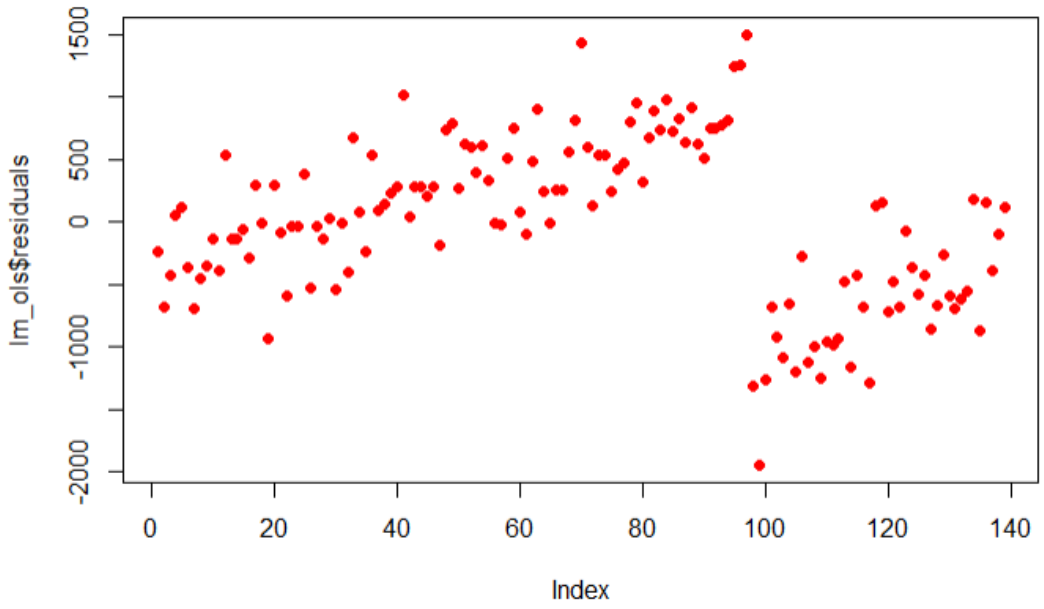
It, not the best model to answer the research questions but it's a good start.

```
Call:
lm(formula = bwt ~ age + lwt + white + smoke + ht + ui + ftv +
    ptl, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1941.48  -462.20    36.75   524.43  1499.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2339.998    349.435   6.697 5.82e-10 ***
age            6.106     11.250   0.543  0.58822
lwt            4.203      2.049   2.051  0.04228 *
white        318.137    125.415   2.537  0.01237 *
smoke       -277.849    130.644  -2.127  0.03533 *
ht          -662.172    267.217  -2.478  0.01449 *
ui          -459.712    163.618  -2.810  0.00573 **
ftv           40.585    125.496   0.323  0.74691
ptl         -262.070    169.552  -1.546  0.12462
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 672.1 on 130 degrees of freedom
Multiple R-squared:  0.2511,    Adjusted R-squared:  0.205
F-statistic: 5.449 on 8 and 130 DF,  p-value: 6.264e-06
```




We can see from the graph above the OLS Y V.S lm residuals that there is trends in the graph

The point doesn't look like random clouds it Which means we need to use a better model.

## Formulation of a GLIM:

I Estimate the explained variable "low": If the baby is under 2.5kg or not ?

Model because bwt and low are dependent variables I remove the bwt from the Explanatory variables.

So the Logistic regression is the compatible model we use the link function $X\beta = ln(\frac{\mu}{1-\mu})$

Other type of regression it's also relevant here is the Probit Model

I compere the AIC models and I get that there is no real different between the results of the model

I choose to work with the Probit model

| AIC Logistic Model | AIC Probit Model |
|---|---|
| 161.91 | 161.77 |

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7379  -0.7874  -0.5562   0.8095   2.1753

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.624554   1.338173   1.214   0.2247
age         -0.044066   0.044193  -0.997   0.3187
lwt         -0.014498   0.008061  -1.798   0.0721 .
white       -0.662267   0.458723  -1.444   0.1488
smoke        0.670578   0.471706   1.422   0.1551
ht           1.982689   0.967321   2.050   0.0404 *
ui           0.384166   0.551185   0.697   0.4858
ftv         -0.070071   0.464353  -0.151   0.8801
ptl          1.396601   0.566318   2.466   0.0137 *
---
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: 143.92  on 130  degrees of freedom
AIC: 161.92

Number of Fisher Scoring iterations: 4
```

## Significantly Variables with Automatically Selection Algorithm

I use Forward Selection and Backward Selection algorithm to find the only significant variables

And get a better model base on AIC . ( lower values of AIC present a better model)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7466  -0.7863  -0.5725   0.8939   2.1930

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.886358   1.024982   0.865   0.3872
lwt         -0.015982   0.008016  -1.994   0.0462 *
white       -0.742158   0.443337  -1.674   0.0941 .
smoke        0.726264   0.446000   1.628   0.1034
ht           1.911848   0.942827   2.028   0.0426 *
ptl          1.317682   0.539636   2.442   0.0146 *
---
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: 145.71  on 133  degrees of freedom
AIC: 157.71

Number of Fisher Scoring iterations: 4
```

The significant variables are the: mother's weight, mothers races (white or not ), smoking status during pregnancy, history of hypertension, and previous premature labors.

That makes sense I keep here only the variables connected to mother healthy lifestyle.

## Interaction and transformation variables

I try to add the Interaction variable without losing the AIC criteria goodness.

Unfortunately, I don't find any Interaction that gives better value than 157.87

But I replace the variable lwt with log(lwt) and get a better AIC value: 157.33

The Best Model I achieve:

```
Call:
glm(formula = low ~ log(lwt) + white + smoke + ht + ptl, family = binomial(link = "probit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7083  -0.7826  -0.5663   0.8910   2.1607

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.5150     2.9672   1.859   0.0631 .
log(lwt)     -1.2862     0.6171  -2.084   0.0371 *
white        -0.4344     0.2575  -1.687   0.0916 .
smoke         0.4297     0.2617   1.642   0.1006
ht            1.0729     0.5318   2.018   0.0436 *
ptl           0.8087     0.3247   2.490   0.0128 *
---
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: 145.33  on 133  degrees of freedom
AIC: 157.33

Number of Fisher Scoring iterations: 5
```

We can see that the AIC has improved and drop to 157.33 most of the variable with P value less than 0.05

I don't touch in the model because I want to avoid overfitting.

## Testing the Best Model

I compare the best model above as the Probit Model that include all variables

Model 1 is the best model

Model 2 is the full model

```
Likelihood ratio test

Model 1: low ~ log(lwt) + white + smoke + ht + ptl
Model 2: low ~ age + lwt + white + smoke + ht + ui + ftv + ptl
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -72.665
2   9 -71.883  3 1.5639     0.6676
```

| | Model 1 | Model 2 (full model) |
|---|---|---|
| AIC | 157.33 | 161.77 |
| BIC | 174.93 | 188.175 |
| Residual deviance | 145.33 | 143.92 |

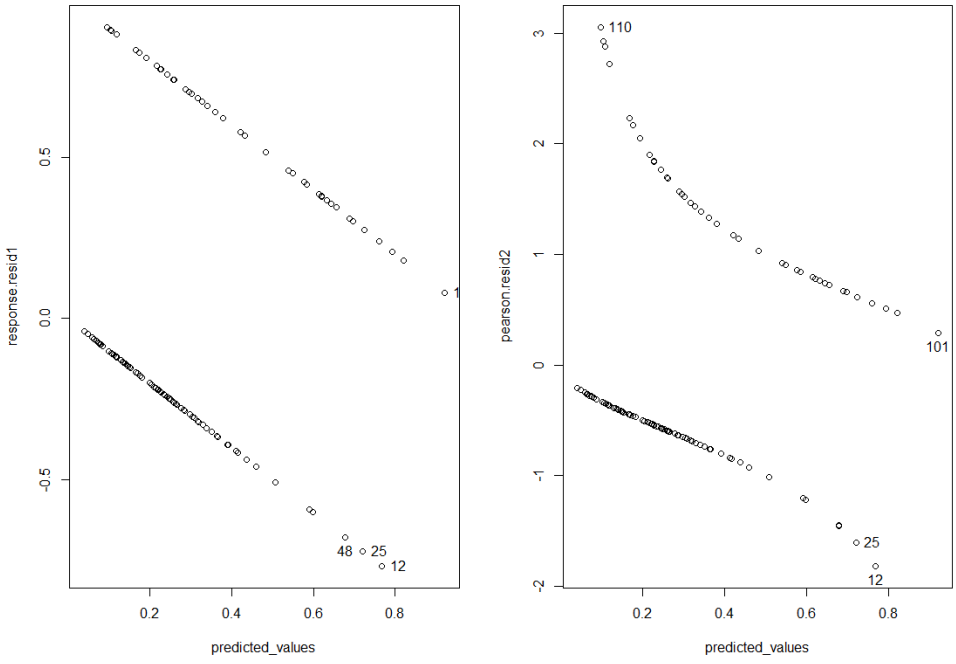
The value of the Residual deviance is higher, but according to the AIC value, as can be seen below, the model I obtained is better.

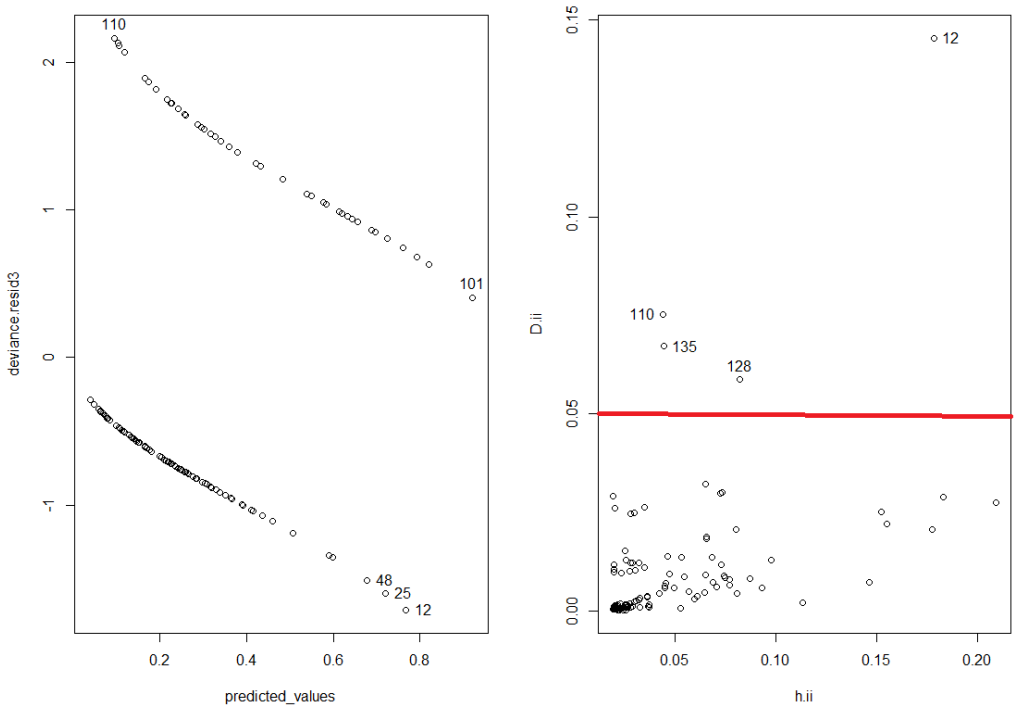
## Estimate phi and check for overdispersion

$$\phi = \frac{Persons\ X^2 Statistic}{n-p} = \frac{X^2}{n-p} = \frac{\sum_i (y_i - \hat{\mu})^2/\hat{\mu}}{n-p}$$

I calculate the phi I get 1.050 is almost 1 so I conclude there is no overdispersion

## Remove Outliers from the Best Model






In the graphs above we can see the outliers with difference measures： response, Pearson, deviance, and cook distance. The outlier is the isolated point from the trend of the other points

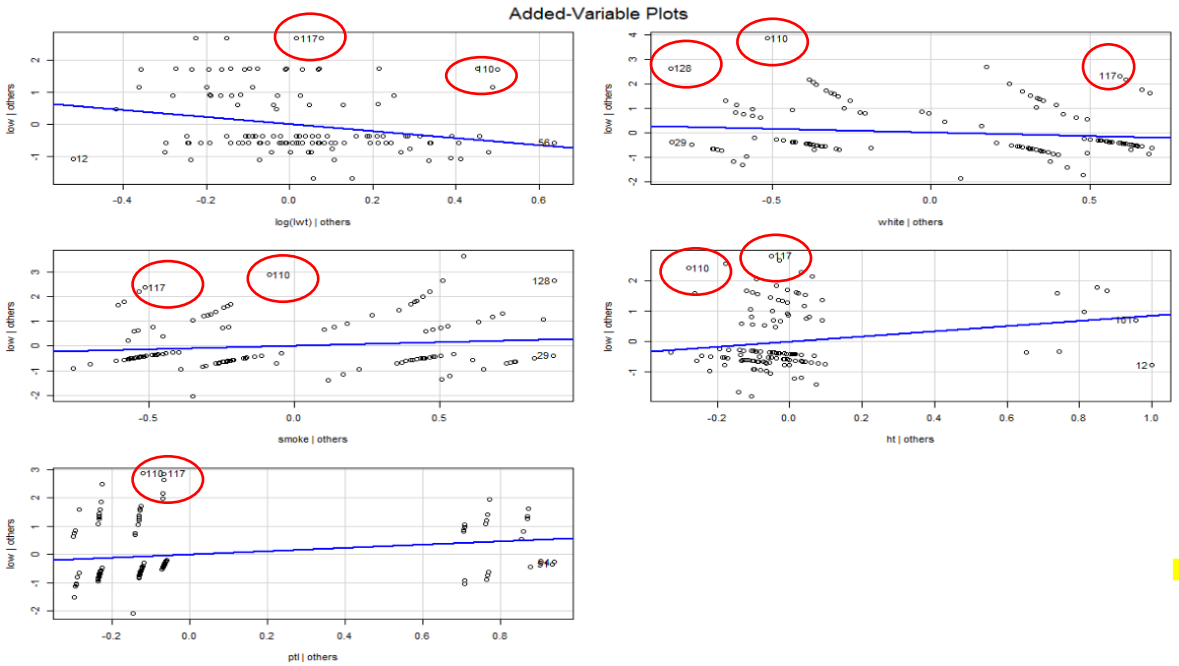
Or the Trend edges. In the graph cook distance v.s leverage it the graph ( Dii v.s hii)

I decided to point to observation as an outlier if it higher than 0.0.5

The outlier here are the points：12,110,135,128,101,25,1

The outlier arises from mistakes in recording the data. If the percentage of the outlier it high we should record the data again.

Also, We should be careful when we omit outlier And see how it affects the best model for that I use Added - Variable Plot to estimate the effect of each variable on the low indicator baby weight.



I decided to remove the outlier that appears in all measurement.

we can see that point 117,110,128 appear as an outlier on measurement : response, Pearson, deviance, and cook distance, and also by the Added -Variable Plot method they affect each explaining variable in the model so the final model without the 3 points looks like that:

```
Call:
glm(formula = low ~ log(lwt) + white + smoke + ht + ptl, family = binomial(link = "probit"),
    data = data[no_outliers, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0427  -0.7213  -0.5174   0.7405   2.4969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.5539     3.3899   2.818  0.00483 **
log(lwt)     -2.1592     0.7079  -3.050  0.00229 **
white        -0.3230     0.2636  -1.225  0.22045
smoke         0.3956     0.2705   1.463  0.14357
ht            1.4335     0.5660   2.533  0.01131 *
ptl           0.9098     0.3336   2.727  0.00639 **
---
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: 131.28  on 130  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 143.28

Number of Fisher Scoring iterations: 5
```

The AIC measurement of the Final model is significantly smaller compared to starting model that base on the selection algorithm.

In addition, the phi that checks overdispersion is 1.047 it almost equally 1 and it better results then the model with the outliers.

## Summary of the Project

My conclusion from the final model is that there are 3 Significant explanatory variables:

1. Mother weight
2. Number of previous premature labors
3. History of hypertension

Those variables have a significant influence on the probability for a baby to weigh under 2.5 kg.

Also, there are 2 important explanatory variables with a positive influence on the probability: smoking and race but we can see that they are not significant explanatory variables because of the value of the $P(>|z|)$

So an over-weight mother with a history of hypertension and previous premature labor has a high probability to give birth to a baby with low weight .