EQC7009 BIOSTATISTICS Project Paper

Yuting Liu S2030500

Sem 2021/2022

Research question

This study will explore the main factors that affect low weight in infants.

Preprocessing, deletion of RACE_ID, SMOKE_ID, PTL_ID, HT_ID, UI_ID, FTV_ID and LWD_ID of the original data.

Import data

```
getwd()

## [1] "D:/桌面"

setwd("D:/桌面")library(readxl)
data1 <- read_excel("D:/桌面/data1.xlsx")
```

Summary of the data

check the dataset type

```
class(data1$LOW)

## [1] "numeric"

class(data1$AGE)

## [1] "numeric"

class(data1$LWT)

## [1] "numeric"

class(data1$RACE)

## [1] "numeric"

class(data1$SMOKE)

## [1] "numeric"

class(data1$PTL)

## [1] "numeric"
```

```
class(data1$HT)

## [1] "numeric"

class(data1$UI)

## [1] "numeric"

class(data1$FTV)

## [1] "numeric"

class(data1$BWT)

## [1] "numeric"

class(data1$LWD)

## [1] "numeric"
```

change into factor and check the dataset type again

```
data1$LOW <- as.factor(as.numeric(data1$LOW))
data1$RACE <- as.factor(as.numeric(data1$RACE))
data1$SMOKE <- as.factor(as.numeric(data1$SMOKE))
data1$PTL <- as.factor(as.numeric(data1$PTL))
data1$HT <- as.factor(as.numeric(data1$HT))
data1$UI <- as.factor(as.numeric(data1$UI))
data1$LWD <- as.factor(as.numeric(data1$LWD))

class(data1$LOW)

## [1] "factor"

class(data1$AGE)

## [1] "numeric"

class(data1$LWT)

## [1] "numeric"

class(data1$RACE)

## [1] "factor"

class(data1$SMOKE)

## [1] "factor"

class(data1$PTL)

## [1] "factor"
```

```
class(data1$HT)
## [1] "factor"
class(data1$UI)
## [1] "factor"
class(data1$FTV)
## [1] "numeric"
class(data1$BWT)
## [1] "numeric"
class(data1$LWD)
## [1] "factor"
summary(data1)
ID                LOW    AGE                 LWT
 Min.   :  4.0   1:130   Min.   :14.00   Min.   : 80.0
 1st Qu.: 68.0   2: 59   1st Qu.:19.00   1st Qu.:110.0
 Median :123.0           Median :23.00   Median :121.0
 Mean   :121.1           Mean   :23.24   Mean   :129.8
 3rd Qu.:176.0           3rd Qu.:26.00   3rd Qu.:140.0
 Max.   :226.0           Max.   :45.00   Max.   :250.0
 RACE   SMOKE   PTL     HT      UI           FTV
 1:96   1:115   1:159   1:177   1:161   Min.   :0.0000
 2:26   2: 74   2: 24   2: 12   2: 28   1st Qu.:0.0000
 3:67           3:  5                   Median :0.0000
                4:  1                   Mean   :0.7937
                                        3rd Qu.:1.0000
                                        Max.   :6.0000
      BWT       LWD
 Min.   : 709   1:147
 1st Qu.:2414   2: 42
 Median :2977

 Mean   :2945

 3rd Qu.:3475

 Max.   :4990
##
```

From the output, we can see that the total number of babies was 189. Of these, 130 were low-weight babies and 53 were non-low-weight babies. With regard to maternal age, the maximum age is 45 years and the minimum age is 14 years. about the weight of mother at last menstrual, the largest number is 250, the smallest number is 80. About the race, White is 96, Black is 26, Others is 67. There are 115 mums here who did not smoke while pregnant and 74 who smoked during pregnancy. There are 177 moms here who have no history of hypertension and 12 who have a history of hypertension. There are 161moms here who have no presence of Uterine irritability and 28 who have presence of Uterine irritability. About the birth weigh, the largest number is 4990 grams, the mean is 2945 grams, and the smallest number is 709 grams.

Model the low birth rate by using logistic regression analysis

This part will build the logistic regression model. the process included two steps. Step 1 is build the basic model(model_1 and model_7), then according to the p-value to delete the variables, and the smallest AIC value will be the best basic model. Step 2 try to use pfor interaction to make the AIC’s value more smaller, and compare with all the model, choose the best model.

model_1 to model_6, will build with the variables “LOW”, “AGE”, “LWT”, “RACE”, “SMOKE”, “PTL”, “HT” , “UI” and “BWT”.

model_7 to model_12, will build with the variables “LOW”, “AGE”, “LWD”, “RACE”, “SMOKE”, “PTL”, “HT” , “UI” and “BWT”.

model_1

```
model_1 <- glm(LOW~LWT+AGE+RACE+SMOKE+PTL+HT+UI+FTV, family = binomial
(), data = data1)
summary(model_1)

##
## Call:
## glm(formula = LOW ~ LWT + AGE + RACE + SMOKE + PTL + HT + UI +
##     FTV, family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6675  -0.7786  -0.5064   0.9044   2.2366
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.020e+03  4.531e+03  -0.225  0.82181   
## LWT         -1.866e-07  8.349e-08  -2.235  0.02539 * 
## AGE         -4.753e-07  4.500e-07  -1.056  0.29090   
## RACE2        1.151e+00  5.404e-01   2.130  0.03315 * 
## RACE3        7.448e-01  4.596e-01   1.621  0.10508   
## SMOKE1       8.662e-01  4.158e-01   2.083  0.03724 * 
## PTL1         1.574e+00  5.265e-01   2.990  0.00279 **
## PTL2         3.044e-01  9.838e-01   0.309  0.75699   
## PTL3        -1.458e+01  8.827e+02  -0.017  0.98682   
## HT1          1.874e+00  7.175e-01   2.611  0.00901 **
## UI1          8.581e-01  4.749e-01   1.807  0.07076 . 
## FTV          1.996e-07  2.079e-06   0.096  0.92354   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.44  on 177  degrees of freedom
## AIC: 216.44
## 
## Number of Fisher Scoring iterations: 13
```

From the table we can see that AGE, RACE, PTL, FTV p-value >0.05, not significant, so we can delete those variables one by one , and see the AIC value bigger or smaller. choose the AIC's value smallest one.

Model_2(-AGE)

```
model_2 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+FTV, family = binomial(), d
ata = data1)
summary(model_2)
```

```
## 
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + FTV, 
##     family = binomial(), data = data1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8634  -0.7737  -0.5140   0.9270   2.2024  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.912e+02  4.514e+03  -0.175  0.86086   
## LWT         -1.982e-07  8.280e-08  -2.394  0.01666 * 
## RACE2        1.245e+00  5.370e-01   2.319  0.02038 * 
## RACE3        7.914e-01  4.529e-01   1.748  0.08055 . 
## SMOKE1       8.817e-01  4.124e-01   2.138  0.03251 *
## PTL1         1.461e+00  5.096e-01   2.868  0.00413 **
## PTL2         2.698e-01  9.823e-01   0.275  0.78359
## PTL3        -1.475e+01  8.827e+02  -0.017  0.98667
## HT1          1.893e+00  7.207e-01   2.626  0.00864 **
## UI1          8.922e-01  4.705e-01   1.896  0.05793 .
## FTV         -1.599e-07  2.048e-06  -0.078  0.93777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 193.58  on 178  degrees of freedom
## AIC: 215.58
##
## Number of Fisher Scoring iterations: 13
```

model_2's AIC value is 215.58, smaller than model_1, continue to delete.

Model_3(-AGE-FTV)

```
model_3 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI, family = binomial(), data
= data1)
summary(model_3)

##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI, family = bin
omial(),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8644  -0.7707  -0.5171   0.9271   2.2084
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.391e+02  1.812e+02  -2.424  0.01537 *
## LWT         -1.988e-07  8.241e-08  -2.412  0.01588 *
## RACE2        1.249e+00  5.352e-01   2.333  0.01962 *
## RACE3        7.967e-01  4.474e-01   1.781  0.07493 .
## SMOKE1       8.854e-01  4.094e-01   2.163  0.03057 *
## PTL1         1.458e+00  5.074e-01   2.873  0.00406 **
## PTL2         2.738e-01  9.808e-01   0.279  0.78007
## PTL3        -1.474e+01  8.827e+02  -0.017  0.98667
## HT1          1.898e+00  7.175e-01   2.645  0.00816 **
## UI1          8.942e-01  4.696e-01   1.904  0.05691 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 193.59  on 179  degrees of freedom
## AIC: 213.59
##
## Number of Fisher Scoring iterations: 13
```

model_3's AIC value is 213.59, smaller than model_2, continue to delete.

Model_4(-AGE-FTV-PTL)

```
model_4 <- glm(LOW~LWT+RACE+SMOKE +HT+UI, family = binomial(), data = d
ata1)
summary(model_4)

##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + HT + UI, family = binomial
(),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7396  -0.8322  -0.5359   0.9873   2.1692
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.278e+02  1.731e+02  -2.471  0.01347 *
## LWT         -1.937e-07  7.874e-08  -2.459  0.01392 *
## RACE2        1.325e+00  5.215e-01   2.540  0.01108 *
## RACE3        9.262e-01  4.304e-01   2.152  0.03140 *
## SMOKE1       1.036e+00  3.926e-01   2.639  0.00832 **
## HT1          1.871e+00  6.909e-01   2.709  0.00676 **
## UI1          9.050e-01  4.476e-01   2.022  0.04317 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 204.22  on 182  degrees of freedom
## AIC: 218.22
##
## Number of Fisher Scoring iterations: 4
```

model_4's AIC value is 218.22, bigger than model_3, model_4 is unsuitable, continue to delete.

Model_5(-AGE-RACE-FTV)

```
model_5 <- glm(LOW~LWT+SMOKE +HT+UI+PTL, family = binomial(), data = da
ta1)
summary(model_5)

##
## Call:
## glm(formula = LOW ~ LWT + SMOKE + HT + UI + PTL, family = binomial(),

##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7304  -0.7536  -0.6102   0.8856   2.0774
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.277e+02  1.749e+02  -2.445  0.01449 *
## LWT         -1.939e-07  7.957e-08  -2.437  0.01483 *
## SMOKE1       5.397e-01  3.527e-01   1.530  0.12599
## HT1          1.949e+00  7.122e-01   2.736  0.00621 **
## UI1          8.823e-01  4.671e-01   1.889  0.05892 .
## PTL1         1.546e+00  4.975e-01   3.107  0.00189 **
## PTL2         1.935e-01  9.806e-01   0.197  0.84358
## PTL3        -1.500e+01  8.827e+02  -0.017  0.98644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 200.27  on 181  degrees of freedom
## AIC: 216.27
##
## Number of Fisher Scoring iterations: 13
```

model_5's AIC value is 216.27, bigger than model_3, model_5 is unsuitable, continue to delete.

Model_6(-AGE-RACE-FTV-PTL)

```
model_6 <- glm(LOW~LWT+SMOKE +HT+UI, family = binomial(), data = data1)
summary(model_6)

##
## Call:
## glm(formula = LOW ~ LWT + SMOKE + HT + UI, family = binomial(),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6635  -0.8060  -0.6646   1.0806   1.9433  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.162e+02  1.666e+02  -2.499  0.01247 * 
## LWT         -1.887e-07  7.577e-08  -2.491  0.01274 * 
## SMOKE1       6.530e-01  3.357e-01   1.945  0.05174 . 
## HT1          1.922e+00  6.827e-01   2.816  0.00487 **
## UI1          8.963e-01  4.429e-01   2.023  0.04303 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 212.83  on 184  degrees of freedom
## AIC: 222.83
## 
## Number of Fisher Scoring iterations: 4
```

model_6's AIC value is 222.83, much more bigger than model_3, model_6 is unsuitable. Therefore, the model_3 is the most suitable model so far.

Model_7(all variables-LWD)

```
model_7 <- glm(LOW~LWD+AGE+RACE+SMOKE+PTL+HT+UI+FTV, family = binomial
(), data = data1)
summary(model_7)
```

```
## 
## Call:
## glm(formula = LOW ~ LWD + AGE + RACE + SMOKE + PTL + HT + UI + 
##     FTV, family = binomial(), data = data1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8186  -0.7513  -0.4779   0.8555   2.1956  
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)   
## (Intercept)     -2.105e+03  4.298e+03  -0.490  0.62431   
## LWD-2208988800   9.373e-01  4.129e-01   2.270  0.02320 * 
## AGE             -5.587e-07  4.469e-07  -1.250  0.21123   
## RACE2            9.997e-01  5.217e-01   1.916  0.05534 . 
## RACE3            7.406e-01  4.555e-01   1.626  0.10400   
## SMOKE1           8.094e-01  4.113e-01   1.968  0.04911 * 
## PTL1             1.642e+00  5.248e-01   3.128  0.00176 **
## PTL2             3.495e-01  9.977e-01   0.350  0.72615   
## PTL3            -1.474e+01  8.827e+02  -0.017  0.98667   
## HT1              1.427e+00  6.579e-01   2.169  0.03005 * 
## UI1             7.956e-01  4.794e-01   1.660  0.09698 .
## FTV            -3.937e-07  1.974e-06  -0.199  0.84191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.89  on 177  degrees of freedom
## AIC: 216.89
##
## Number of Fisher Scoring iterations: 13
```

model_7's AIC value is 216.89. Delete some variables(RACE, AGE, PTL and FTV) not significant one by one.

Model_8(-AGE)

```
model_8 <- glm(LOW~LWD+RACE+SMOKE+PTL+HT+UI+FTV, family = binomial(), d
ata = data1)
summary(model_8)
```

```
##
## Call:
## glm(formula = LOW ~ LWD + RACE + SMOKE + PTL + HT + UI + FTV,
##     family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9381  -0.7131  -0.4444   0.8496   2.1745
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.877e+03  4.264e+03  -0.440   0.6597
## LWD-2208988800  9.727e-01  4.105e-01   2.370   0.0178 *
## RACE2           1.099e+00  5.200e-01   2.114   0.0345 *
## RACE3           8.022e-01  4.467e-01   1.796   0.0725 .
## SMOKE1          8.228e-01  4.074e-01   2.020   0.0434 *
## PTL1            1.503e+00  5.047e-01   2.978   0.0029 **
## PTL2            3.115e-01  9.965e-01   0.313   0.7546
## PTL3           -1.494e+01  8.827e+02  -0.017   0.9865
## HT1             1.425e+00  6.644e-01   2.145   0.0320 *
## UI1             8.462e-01  4.726e-01   1.791   0.0734 .
## FTV            -8.488e-07  1.930e-06  -0.440   0.6601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 194.50  on 178  degrees of freedom
## AIC: 216.5
##
## Number of Fisher Scoring iterations: 13
```

model_8’s AIC value is 216.5, smaller than model_7, continue to delete.

Model_9(-AGE-FTV)

```
model_9 <- glm(LOW~LWD+RACE+SMOKE+PTL+HT+UI, family = binomial(), data
= data1)
summary(model_9)
```

```
##
## Call:
## glm(formula = LOW ~ LWD + RACE + SMOKE + PTL + HT + UI, family = bin
omial(),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9783  -0.7160  -0.4289   0.8457   2.2053
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.3397     0.4120  -5.679 1.36e-08 ***
## LWD-2208988800   0.9723     0.4100   2.371  0.01772 *
## RACE2            1.1092     0.5183   2.140  0.03235 *
## RACE3            0.8223     0.4427   1.857  0.06324 .
## SMOKE1           0.8342     0.4048   2.061  0.03931 *
## PTL1             1.4902     0.5040   2.957  0.00311 **
## PTL2             0.3358     0.9941   0.338  0.73552
## PTL3           -14.8923   882.7435  -0.017  0.98654
## HT1              1.4485     0.6610   2.191  0.02843 *
## UI1              0.8594     0.4708   1.825  0.06795 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 194.69  on 179  degrees of freedom
## AIC: 214.69
##
## Number of Fisher Scoring iterations: 13
```

model_9’s AIC value is 214.69, smaller than model_8, continue to delete.

Model_10(-AGE-PTV-PTL)

```
model_10 <- glm(LOW~LWD+RACE+SMOKE+HT+UI, family = binomial(), data = d
ata1)
summary(model_10)
```

```
##
## Call:
## glm(formula = LOW ~ LWD + RACE + SMOKE + HT + UI, family = binomial
(),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6334  -0.7118  -0.4513   1.0245   2.1609
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.2328     0.3975  -5.617 1.94e-08 ***
## LWD-2208988800     0.8924     0.3925   2.273   0.0230 *
## RACE2              1.1866     0.5076   2.338   0.0194 *
## RACE3              0.9516     0.4246   2.241   0.0250 *
## SMOKE1             0.9892     0.3880   2.550   0.0108 *
## HT1                1.4172     0.6386   2.219   0.0265 *
## UI1                0.8810     0.4465   1.973   0.0485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 206.04  on 182  degrees of freedom
## AIC: 220.04
##
## Number of Fisher Scoring iterations: 4
```

model_10's AIC value is 220.04, bigger than model_9, model_10 is unsuitable, continue to delete.

Model_11(-AGE-PTV-RACE)

```
model_11 <- glm(LOW~LWD+SMOKE+PTL+HT+UI, family = binomial(), data = da
ta1)
summary(model_11)
```

```
##
## Call:
## glm(formula = LOW ~ LWD + SMOKE + PTL + HT + UI, family = binomial(),

##     data = data1)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max  
## -1.9387  -0.7137  -0.5712   0.9226   1.9461  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -1.7305     0.2772  -6.244 4.26e-10 ***
## LWD-2208988800    1.0094     0.3967   2.545  0.01094 *  
## SMOKE1            0.4929     0.3533   1.395  0.16297    
## PTL1              1.5726     0.4963   3.169  0.00153 ** 
## PTL2              0.2641     0.9955   0.265  0.79079    
## PTL3            -15.2001   882.7435  -0.017  0.98626    
## HT1               1.5045     0.6478   2.322  0.02021 *  
## UI1               0.8622     0.4692   1.837  0.06614 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 200.74  on 181  degrees of freedom
## AIC: 216.74
## 
## Number of Fisher Scoring iterations: 13
```

model_11's AIC value is 216.74, bigger than model_9, model_11 is unsuitable, continue to delete.

Model_12(-AGE-PTV-RACE-PTL)

```
model_12 <- glm(LOW~LWD+SMOKE+HT+UI, family = binomial(), data = data1)
summary(model_12)
```

```
## 
## Call:
## glm(formula = LOW ~ LWD + SMOKE + HT + UI, family = binomial(), 
##     data = data1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.5755  -0.8158  -0.6237   1.1750   1.8617  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -1.5384     0.2599  -5.919 3.25e-09 ***
## LWD-2208988800    0.9351     0.3790   2.467   0.0136 *  
## SMOKE1            0.6091     0.3360   1.813   0.0698 .  
## HT1               1.4762     0.6270   2.354   0.0185 *  
## UI1               0.8942     0.4422   2.022   0.0432 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 213.99  on 184  degrees of freedom
## AIC: 223.99
##
## Number of Fisher Scoring iterations: 4
```

model_12's AIC value is 223.99, bigger than model_9, model_12 is unsuitable, continue to delete.

model_3 compare with model_9, 213.59 is smaller than 214.69. So choose the model_3

in order to make the model more accurate, try to use the "pfor interaction" to make the AIC value more smaller.

model_31(LWT*AGE)

```
model_31 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+(LWT*AGE), family = binomi
al(), data = data1)
summary(model_31)

##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + (LWT *
##     AGE), family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6508  -0.7821  -0.5064   0.9151   2.2186
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.684e+04  8.644e+05   0.100  0.91998
## LWT          3.998e-05  3.933e-04   0.102  0.91903
## RACE2        1.144e+00  5.417e-01   2.112  0.03467 *
## RACE3        7.344e-01  4.579e-01   1.604  0.10872
## SMOKE1       8.597e-01  4.138e-01   2.078  0.03774 *
## PTL1         1.580e+00  5.262e-01   3.003  0.00267 **
## PTL2         2.999e-01  9.838e-01   0.305  0.76048
## PTL3        -1.458e+01  8.827e+02  -0.017  0.98682
## HT1          1.867e+00  7.138e-01   2.615  0.00892 **
## UI1          8.583e-01  4.756e-01   1.805  0.07111 .
## AGE          3.953e-05  3.917e-04   0.101  0.91960
## LWT:AGE      1.820e-14  1.782e-13   0.102  0.91866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.44  on 177  degrees of freedom
## AIC: 216.44
##
## Number of Fisher Scoring iterations: 13
```

model_31’s AIC value is 216.44, bigger than model_3, model_31 is unsuitable

Model_32(LWT*RACE)

```
model_32 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+(LWT*RACE), family = binom
ial(), data = data1)
summary(model_32)
```

```
##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + (LWT *
##     RACE), family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8421  -0.7862  -0.4924   0.9130   2.1888
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.572e+02  2.628e+02  -1.359  0.17402
## LWT         -1.615e-07  1.196e-07  -1.351  0.17666
## RACE2        4.438e+01  3.900e+02   0.114  0.90941
## RACE3       -4.060e+02  4.661e+02  -0.871  0.38369
## SMOKE1       8.877e-01  4.131e-01   2.149  0.03163 *
## PTL1         1.494e+00  5.093e-01   2.934  0.00334 **
## PTL2         2.671e-01  9.791e-01   0.273  0.78501
## PTL3        -1.465e+01  8.827e+02  -0.017  0.98676
## HT1          1.806e+00  7.141e-01   2.529  0.01144 *
## UI1          8.945e-01  4.805e-01   1.861  0.06269 .
## LWT:RACE2    1.966e-08  1.775e-07   0.111  0.91182
## LWT:RACE3   -1.850e-07  2.120e-07  -0.873  0.38277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.57  on 177  degrees of freedom
## AIC: 216.57
##
## Number of Fisher Scoring iterations: 13
```

model_32's AIC value is 216.57, bigger than model_3, model_32 is unsuitable.

Model_33(LWT*SMOKE)

```
model_33 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+(LWT*SMOKE), family = bino
mial(), data = data1)
summary(model_33)

##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + (LWT *
##     SMOKE), family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8784  -0.7660  -0.5126   0.9175   2.2431
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.636e+02  2.790e+02  -2.378  0.01739 *
## LWT         -3.009e-07  1.269e-07  -2.371  0.01775 *
## RACE2        1.263e+00  5.363e-01   2.355  0.01854 *
## RACE3        7.104e-01  4.511e-01   1.575  0.11524
## SMOKE1       4.004e+02  3.485e+02   1.149  0.25058
## PTL1         1.511e+00  5.120e-01   2.951  0.00317 **
## PTL2         2.849e-01  9.782e-01   0.291  0.77085
## PTL3        -1.461e+01  8.827e+02  -0.017  0.98679
## HT1          1.805e+00  7.204e-01   2.506  0.01222 *
## UI1          9.807e-01  4.788e-01   2.048  0.04054 *
## LWT:SMOKE1   1.818e-07  1.585e-07   1.146  0.25162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.23  on 178  degrees of freedom
## AIC: 214.23
##
## Number of Fisher Scoring iterations: 13
```

model_33's AIC value is 214.23, bigger than model_3, model_33 is unsuitable.

Model_34(LWT*HT)

```
model_34 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+(LWT*HT), family = binomia
l(), data = data1)
summary(model_34)

##
## Call:
```

```
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + (LWT *
##     HT), family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8644  -0.7692  -0.5160   0.9295   2.2045
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.322e+02  2.012e+02  -2.148  0.03175 *
## LWT         -1.956e-07  9.154e-08  -2.137  0.03261 *
## RACE2        1.250e+00  5.352e-01   2.336  0.01950 *
## RACE3        7.981e-01  4.475e-01   1.783  0.07455 .
## SMOKE1       8.897e-01  4.129e-01   2.155  0.03119 *
## PTL1         1.455e+00  5.085e-01   2.862  0.00421 **
## PTL2         2.754e-01  9.807e-01   0.281  0.77884
## PTL3        -1.474e+01  8.827e+02  -0.017  0.98668
## HT1         -3.191e+01  4.350e+02  -0.073  0.94152
## UI1          8.957e-01  4.699e-01   1.906  0.05662 .
## LWT:HT1     -1.540e-08  1.981e-07  -0.078  0.93805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 193.58  on 178  degrees of freedom
## AIC: 215.58
##
## Number of Fisher Scoring iterations: 13
```

model_34's AIC value is 215.58, bigger than model_3, model_34 is unsuitable.

Model_35(LWT*UI)

```
model_35 <- glm(LOW~LWT+RACE+SMOKE+PTL+HT+UI+(LWT*UI), family = binomia
l(), data = data1)
summary(model_35)
```

```
##
## Call:
## glm(formula = LOW ~ LWT + RACE + SMOKE + PTL + HT + UI + (LWT *
##     UI), family = binomial(), data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8412  -0.7787  -0.5154   0.8918   2.2664
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.124e+02  2.065e+02  -2.482  0.01306 *
## LWT         -2.321e-07  9.391e-08  -2.472  0.01344 *
## RACE2        1.194e+00  5.423e-01   2.203  0.02762 *
## RACE3        7.948e-01  4.499e-01   1.767  0.07729 .
## SMOKE1       9.473e-01  4.180e-01   2.266  0.02345 *
## PTL1         1.402e+00  5.134e-01   2.731  0.00631 **
## PTL2         2.508e-01  9.641e-01   0.260  0.79474
## PTL3        -1.452e+01  8.827e+02  -0.016  0.98688
## HT1          2.002e+00  7.389e-01   2.709  0.00674 **
## UI1          3.771e+02  4.317e+02   0.874  0.38237
## LWT:UI1      1.711e-07  1.963e-07   0.871  0.38350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 192.85  on 178  degrees of freedom
## AIC: 214.85
##
## Number of Fisher Scoring iterations: 13
```

model_35's AIC value is 214.85, bigger than model_3, model_35 is unsuitable.

Find the best model for this data

compare different model's AIC value

Different Model's AIC

MODEL NAME	AIC VALUE	
model_1	216.44	
model_2	215.58	
model_3	**213.59**	**the best model*
model_4	218.22	
model_5	216.27	
model_6	222.83	
model_7	216.89	
model_8	216.5	
model_9	214.69	
model_10	220.04	
model_11	216.74	
model_12	223.99	
model_31	216.44	
model_32	216.57	

MODEL NAME	AIC VALUE
model_33	214.23
model_34	215.58
model_35	214.85

finally, the model is $y(LOW) = 0.0303 - 0.0171x(LWT) + 1.24x(RACE2) + 0.7967x(RACE3) + 0.8853x(SMOKE) + 1.4578x(PTL1) + 0.2738x(PTL2) - 14.744x(PTL3) + 1.8982x(HT) + 0.8942x(UI)$

Hosmer – Lemeshow (H-L) Test

```
library(ResourceSelection)

## Warning: 程辑包'ResourceSelection'是用 R 版本 4.1.3 来建造的

## ResourceSelection 0.3-5   2019-07-22

hl<-hoslem.test(model_3$y,fitted(model_3),g=10)
hl

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_3$y, fitted(model_3)
## X-squared = 5.7238, df = 8, p-value = 0.6781
```

the p-value=0.6781>0.05, that is mean the model is suitable.

Interpret and discuss

after compare the value of difference model's AIC, the study has a interpret:

1 unit increase of X(LWT) will cause 0.0171 decrease of Y on average, 1 unit increase of X(RACE2) will cause 1.24 increase of Y on average, 1 unit increase of X(RACE3) will cause 0.7967 increase of Y on average, 1 unit increase of X(SMOKE) will cause 0.8853 increase of Y on average, 1 unit increase of X(PTL1) will cause 1.4578 increase of Y on average, 1 unit increase of X(PTL2) will cause 0.2738 increase of Y on average, 1 unit increase of X(PTL3) will cause 14.744 decrease of Y on average, 1 unit increase of X(HT) will cause 1.8982increase of Y on average, 1 unit increase of X(UI) will cause 0.8942 increase of Y on average,

Conculsion

the age of mother and the time of physician visits during first trimester has not deep impression on the infant's weight.

the weight of mother at last menstrual period(LWT), Race, smoking status during pregnancy(SMOKE), history of premature labour(PTL), history of hypertension(HT) and presence of uterine lrritability(UI) are the major factors affecting the baby's weight.

the history of premature labour(PTL) over2 times, have a very serious negative impact on a infant's weight. Therefore, women who have the habit of premature delivery should pay more attention to their own health.