STAT 426: Project 2

Spring 2021, by Shuyu Jia (shuyuj2)

Due: Wednesday, May 12 by 11:59 PM

1. Introduction

Low birthweight is when a baby is born weighing less than 5 pounds, 8 ounces (2.5 kg). The average newborn weight is about 8 pounds (3.6 kg). Some babies with low birthweight are healthy, even though they are small. But being low birthweight can cause serious health problems for some babies. A baby with low birthweight may have trouble eating, gaining weight, and fighting off infections. Some low-birthweight babies may have long-term health problems, too. About 1 in 12 babies (about 8 percent) in the United States is born with low birthweight.

There are two main reasons why a baby may be born with low birthweight. One is premature birth. The earlier a baby is born, the lower her birthweight may be. The other is fetal growth restriction (FGR). Some growth-restricted babies may have low birthweight because their parents are small, others may because something slowed or stopped their growth in the womb.

In this project, we would like to find out the factors that are associated with low infant birthweight and then use these factors to predict low birthweight.

2. Data

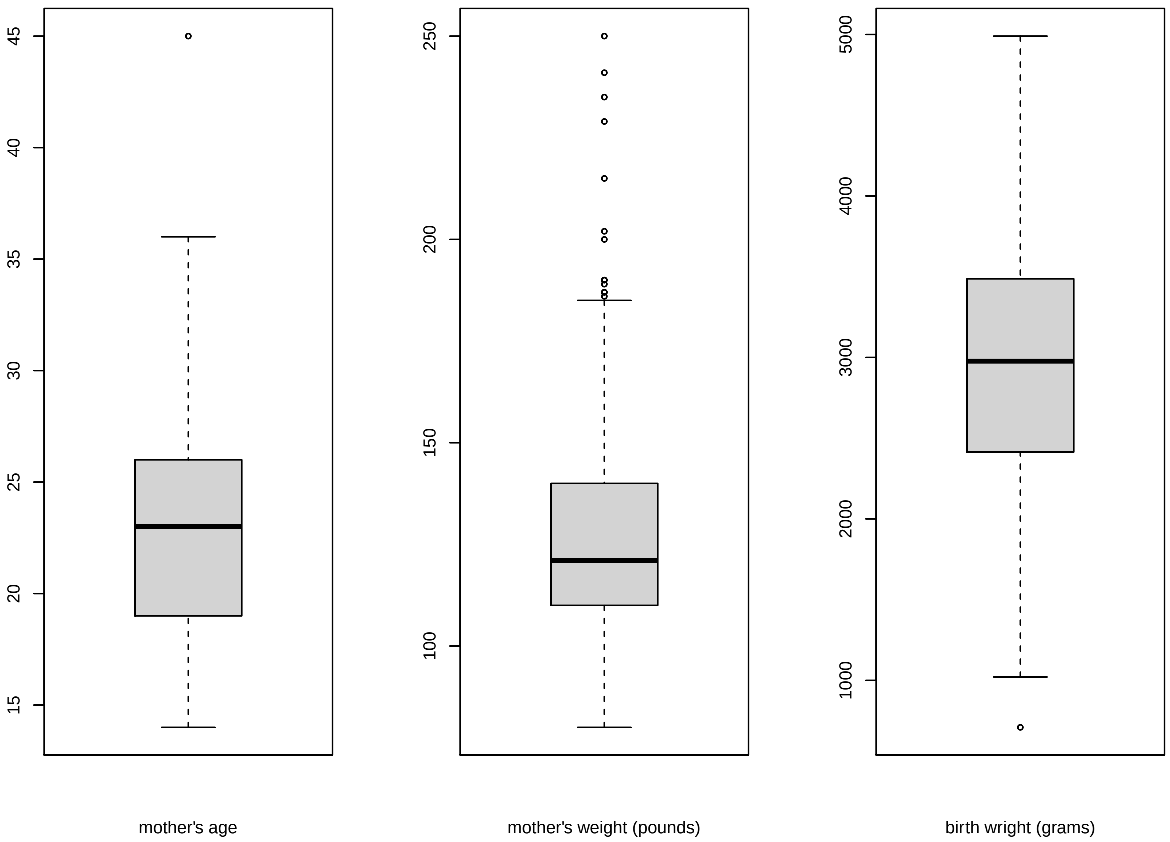
This dataset includes birthweights and several variables concerning the mother for 189 births at Baytowne Medical Center in Springfield, Massachusetts in 1986. The description of variables are as follows:

- `low` — indicator of birth weights less than 2.5 kg (binary variable)
- `age` — mother's age in years (continuous variable)
- `lwt` — mother's weight in pounds at last menstrual period (continuous variable)
- `race` — mother's race (1 = white, 2 = black, 3 = other) (categorical variable)
- `smoke` — smoking status during pregnancy (binary variable)
- `ptl` — number of previous premature labours (ordinal variable)
- `ht` — history of hypertension (binary variable)
- `ui` — presence of uterine irritability (binary variable)
- `ftv` — number of physician visits during the first trimester (ordinal variable)
- `bwt` — birth weight in grams (continuous variable)

```r
library(MASS)
data(birthwt)
summary(birthwt[,c("age", "lwt", "bwt")])
```

```
##       age             lwt             bwt
##  Min.   :14.00   Min.   : 80.0   Min.   : 709
##  1st Qu.:19.00   1st Qu.:110.0   1st Qu.:2414
##  Median :23.00   Median :121.0   Median :2977
##  Mean   :23.24   Mean   :129.8   Mean   :2945
##  3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3487
##  Max.   :45.00   Max.   :250.0   Max.   :4990
```

```
par(mfrow=c(1,3))
boxplot(birthwt[, "age"], xlab = 'mother\'s age')
boxplot(birthwt[, "lwt"], xlab = 'mother\'s weight (pounds)')
boxplot(birthwt[, "bwt"], xlab = 'birth wright (grams)')
```



```
table(birthwt$low)
```

```
##
##   0   1
## 130  59
```

```
prop.table(table(birthwt$low))
```

```
##
##         0         1
## 0.6878307 0.3121693
```

```
table(birthwt$race)
```

```
## 
##  1  2  3 
## 96 26 67
```

```
prop.table(table(birthwt$race))
```

```
## 
##         1         2         3 
## 0.5079365 0.1375661 0.3544974
```

```
table(birthwt$smoke)
```

```
## 
##   0   1 
## 115  74
```

```
prop.table(table(birthwt$smoke))
```

```
## 
##         0         1 
## 0.6084656 0.3915344
```

```
table(birthwt$ptl)
```

```
## 
##   0   1   2   3 
## 159  24   5   1
```

```
prop.table(table(birthwt$ptl))
```

```
## 
##           0           1           2           3 
## 0.841269841 0.126984127 0.026455026 0.005291005
```

```
table(birthwt$ht)
```

```
## 
##   0   1 
## 177  12
```

```
prop.table(table(birthwt$ht))
```

```
## 
##          0          1 
## 0.93650794 0.06349206
```

```r
table(birthwt$ui)
```

```
##
##   0   1
## 161  28
```

```r
prop.table(table(birthwt$ui))
```

```
##
##         0         1
## 0.8518519 0.1481481
```

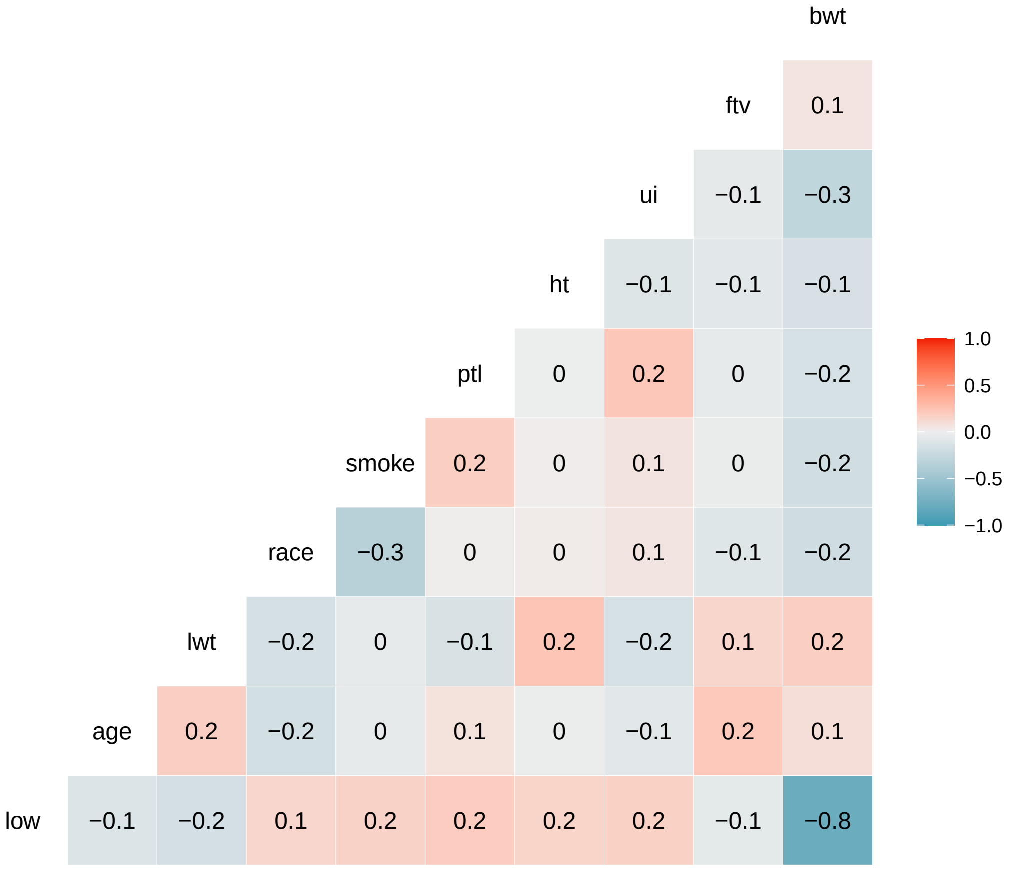
```r
table(birthwt$ftv)
```

```
##
##   0   1   2   3   4   6
## 100  47  30   7   4   1
```

```r
prop.table(table(birthwt$ftv))
```

```
##
##           0           1           2           3           4           6
## 0.529100529 0.248677249 0.158730159 0.037037037 0.021164021 0.005291005
```

```r
library(GGally)
ggcorr(birthwt, label = TRUE)
```



Since the binary variable `low` is generated from the continuous variable `bwt`, it is expected for them to have a strong correlation, and we should not use `bwt` to predict `low`. Except for `low` and `bwt`, no other pair of variables has any strong correlation.

3. Research Question One

What factors are associated with low infant birthweight (less than 2.5 kg)?

First, we consider all 8 variables and all possible 2-way interactions, and perform a backward selection.

```r
birthwt$race = as.factor(birthwt$race)
full_mod = glm(low ~ (age + lwt + race + smoke + ptl + ht + ui + ftv)^2,
               family = binomial, data = birthwt)
back_mod = step(full_mod, direction = "backward", trace = 0)
summary(back_mod)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv + age:ptl + age:ht + age:ftv + lwt:smoke + lwt:ht + lwt:ui +
##     race:ht + smoke:ht + ptl:ht + ptl:ui + ht:ftv, family = binomial,
##     data = birthwt)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8758  -0.6983  -0.3561   0.6875   2.5783
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.423e+00  2.251e+00   0.632 0.527223
## age          1.308e-01  6.277e-02   2.084 0.037166 *
## lwt         -5.062e-02  1.721e-02  -2.941 0.003268 **
## race2        7.079e-01  6.012e-01   1.177 0.239020
## race3        4.269e-01  5.095e-01   0.838 0.402096
## smoke       -4.650e+00  2.344e+00  -1.983 0.047347 *
## ptl          4.280e+00  2.527e+00   1.694 0.090316 .
## ht          -5.662e+02  4.051e+04  -0.014 0.988850
## ui          -2.577e+00  2.402e+00  -1.073 0.283347
## ftv          3.742e+00  1.134e+00   3.298 0.000973 ***
## age:ptl     -1.403e-01  1.029e-01  -1.364 0.172672
## age:ht      -4.562e+01  3.229e+03  -0.014 0.988728
## age:ftv     -1.707e-01  5.255e-02  -3.249 0.001158 **
## lwt:smoke    4.647e-02  1.911e-02   2.431 0.015045 *
## lwt:ht       3.385e+00  2.444e+02   0.014 0.988949
## lwt:ui       3.661e-02  2.004e-02   1.827 0.067750 .
## race2:ht     8.809e+02  6.074e+04   0.015 0.988430
## race3:ht     1.232e+03  8.593e+04   0.014 0.988560
## smoke:ht     6.204e+02  4.338e+04   0.014 0.988588
## ptl:ht       1.386e+02  1.623e+04   0.009 0.993184
## ptl:ui      -1.449e+00  7.908e-01  -1.832 0.066978 .
## ht:ftv       3.662e+02  2.764e+04   0.013 0.989431
## ---
## 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: 165.67  on 167  degrees of freedom
## AIC: 209.67
##
## Number of Fisher Scoring iterations: 20
```

From the summary we can see that there are 5 interaction terms (`age:ptl`, `age:ftv`, `lwt:smoke`, `lwt:ui`, `ptl:ui`) have reasonably low p-values. We included them along with all 8 variables and modify our full model, and then ran backward selection again.

```
full_mod = glm(low ~ age + lwt + race + smoke + ptl + ht + ui + ftv +
                     age*ptl + age*ftv + lwt*smoke + lwt*ui + ptl*ui,
               family = binomial, data = birthwt)
back_mod = step(full_mod, direction = "backward", trace = 0)
summary(back_mod)
```

```
##
## Call:
## glm(formula = low ~ age + lwt + smoke + ptl + ht + ui + ftv +
##     age:ftv + lwt:smoke + lwt:ui + ptl:ui, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9711  -0.7484  -0.4342   0.8595   2.4035
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.16560    1.80864   0.644 0.519276
## age          0.09169    0.05288   1.734 0.082941 .
## lwt         -0.03816    0.01360  -2.805 0.005025 **
## smoke       -2.71949    1.93569  -1.405 0.160045
## ptl          1.14162    0.50169   2.276 0.022874 *
## ht           1.85656    0.77280   2.402 0.016289 *
## ui          -2.17780    2.21743  -0.982 0.326036
## ftv          3.67297    1.06902   3.436 0.000591 ***
## age:ftv     -0.16693    0.04955  -3.369 0.000755 ***
## lwt:smoke    0.02819    0.01560   1.807 0.070706 .
## lwt:ui       0.03268    0.01820   1.796 0.072522 .
## ptl:ui      -1.51190    0.71129  -2.126 0.033538 *
## ---
## 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: 184.80  on 177  degrees of freedom
## AIC: 208.8
##
## Number of Fisher Scoring iterations: 5
```

Now we ran a forward selection.

```
null_mod = glm(low ~ 1, data = birthwt, family = binomial)
forw_mod = step(null_mod, scope = list(lower = null_mod, upper = full_mod),
                direction = "forward", trace = 0)
summary(forw_mod)
```

```
##
## Call:
## glm(formula = low ~ ptl + lwt + ht + race + smoke + ui + ptl:ui,
##     family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7960  -0.7893  -0.5225   0.9770   2.1961
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.220292   0.961386  -0.229  0.81876
## ptl          0.928233   0.444829   2.087  0.03691 *
## lwt         -0.015197   0.006893  -2.205  0.02748 *
## ht           1.857450   0.699082   2.657  0.00788 **
## race2        1.252008   0.527318   2.374  0.01758 *
## race3        0.848093   0.439162   1.931  0.05346 .
## smoke        0.927496   0.403130   2.301  0.02141 *
## ui           1.155409   0.510920   2.261  0.02373 *
## ptl:ui      -1.065660   0.674581  -1.580  0.11417
## ---
## 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: 199.50  on 180  degrees of freedom
## AIC: 217.5
##
## Number of Fisher Scoring iterations: 4
```

```
c(AIC(back_mod), AIC(forw_mod))
```

```
## [1] 208.7954 217.5004
```

The backward model has a lower AIC score, so we chose it as our best model. From the backward selected model, `lwt`, `ptl`, `ht`, `ftv`, `age:ftv`, `ptl:ui` are all significant factors that are associated with low infant birthweight at a level of 0.05.

4. Research Question Two

How well can these factors predict low birthweight?

Predictive power:

```
cor(birthwt$low, fitted(back_mod))
```

```
## [1] 0.4933314
```

```
as.numeric((logLik(back_mod) - logLik(null_mod)) / (0 - logLik(null_mod)))
```

```
## [1] 0.2125375
```

```
logit_mod = glm(low ~ age + lwt + smoke + ptl + ht + ui + ftv +
    age*ftv + lwt*smoke + lwt*ui + ptl*ui, family = binomial(link = "logit"), data = birthwt)
probit_mod = glm(low ~ age + lwt + smoke + ptl + ht + ui + ftv +
    age*ftv + lwt*smoke + lwt*ui + ptl*ui, family = binomial(link = "probit"), data = birthwt)
cloglog_mod = glm(low ~ age + lwt + smoke + ptl + ht + ui + ftv +
    age*ftv + lwt*smoke + lwt*ui + ptl*ui, family = binomial(link = "cloglog"), data = birthwt)
```

```
# AUC for logit link
pihat = fitted(logit_mod)
mean(outer(pihat[birthwt$low==1], pihat[birthwt$low==0], ">") +
           0.5 * outer(pihat[birthwt$low==1], pihat[birthwt$low==0], "=="))
```

```
## [1] 0.8046936
```

```
# AUC for probit link
pihat = fitted(probit_mod)
mean(outer(pihat[birthwt$low==1], pihat[birthwt$low==0], ">") +
           0.5 * outer(pihat[birthwt$low==1], pihat[birthwt$low==0], "=="))
```

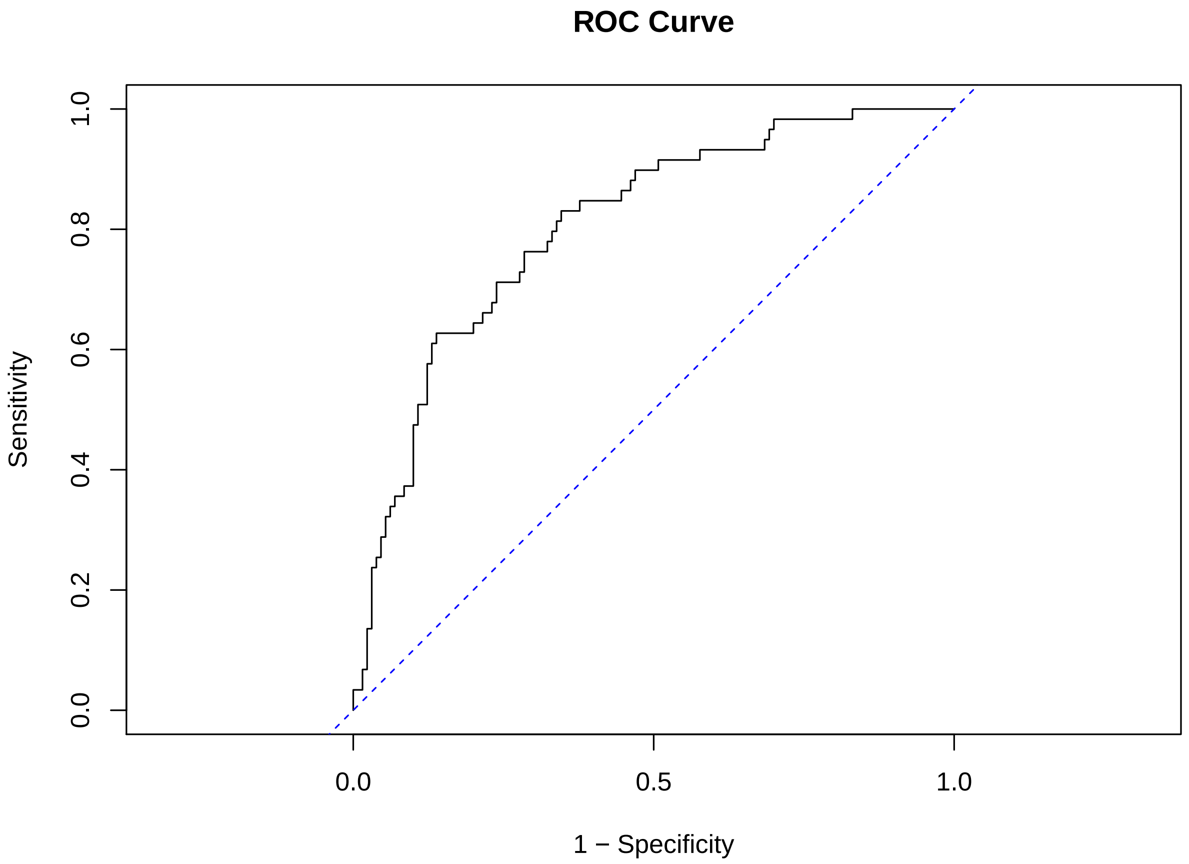
```
## [1] 0.8044329
```

```
# AUC for cloglog link
pihat = fitted(cloglog_mod)
mean(outer(pihat[birthwt$low==1], pihat[birthwt$low==0], ">") +
           0.5 * outer(pihat[birthwt$low==1], pihat[birthwt$low==0], "=="))
```

```
## [1] 0.796219
```

The logit link has the highest AUC, so we chose logit link. The ROC curve is as follows:

```
pihat = fitted(logit_mod)
false.neg = c(0,cumsum(tapply(birthwt$low,pihat,sum)))
true.neg = c(0,cumsum(table(pihat))) - false.neg
plot(1-true.neg/max(true.neg), 1-false.neg/max(false.neg), type="l", main="ROC Curve", xlab="1 - Specifi
abline(a=0, b=1, lty=2, col="blue")
```



Evaluation metrics for prediction:

```
# max accuracy: 77% --- 50% sens, 90% spec
# 100% sens: 0.05 --- 17% spec
# 98% sens: 0.11 --- 30% spec
# 95% sens: 0.13 --- 32% spec
# 90% sens: 0.19 --- 50% spec
# 80% sens: 0.27 --- 67% spec
pi0 = 0.5
table(y = birthwt$low, yhat = as.numeric(fitted(back_mod) > pi0))
```

```
##    yhat
## y     0   1
##   0 116  14
##   1  30  29
```

```
# sensitivity
29 / 59
```

```
## [1] 0.4915254
```

```
# specificity
116 / 130
```

```
## [1] 0.8923077
```

```
# accuracy
(116+29) / 189
```

```
## [1] 0.7671958
```

5. Summary

- In our best model, `lwt`, `ptl`, `ht`, `ftv`, `age:ftv`, `ptl:ui` are significant factors that are associated with low infant birthweight at a level of 0.05
- The factors of our best model to predict low birthweight included `age`, `lwt`, `smoke`, `ptl`, `ht`, `ui`, `ftv`, `age:ftv`, `lwt:smoke`, `lwt:ui`, `ptl:ui`. The best link function is the logit link due to its highest area under the ROC curve. The AIC score of our best model is 208.8
- The highest accuracy (77%) is achieved when we used a cutoff of 0.5, with a 50% sensitivity and 90% specificity.
- 80% sensitivity can be achieved when we used a cutoff of 0.27, the corresponding specificity is 67%.
- 90% sensitivity can be achieved when we used a cutoff of 0.19, the corresponding specificity is 50%.
- 95% sensitivity can be achieved when we used a cutoff of 0.13, the corresponding specificity is 32%.
- 98% sensitivity can be achieved when we used a cutoff of 0.11, the corresponding specificity is 30%.
- 100% sensitivity can be achieved when we used a cutoff of 0.05, the corresponding specificity is 17%.

6. References

https://www.marchofdimes.org/complications/low-birthweight.aspx

https://en.wikipedia.org/wiki/Low_birth_weight