Homework 6

Akshay Kumar Varanasi (av32826)

This homework is due on Mar. 12, 2019 at 4:00pm. Please submit as a PDF file on Canvas.

For this homework, you will work with a dataset collected by John Holcomb from the North Carolina State Center for Health and Environmental Statistics. This data set contains 1409 birth records from North Carolina in 2001.

```
NCbirths <- read_csv("http://wilkelab.org/classes/SDS348/data_sets/NCbirths.csv
")
```

```
## Parsed with column specification:
## cols(
##   Plural = col_integer(),
##   Sex = col_integer(),
##   MomAge = col_integer(),
##   Weeks = col_integer(),
##   Gained = col_integer(),
##   Smoke = col_integer(),
##   BirthWeightGm = col_double(),
##   Low = col_integer(),
##   Premie = col_integer(),
##   Marital = col_integer()
## )
```

```
head(NCbirths)
```

```
## # A tibble: 6 x 10
##   Plural   Sex MomAge Weeks Gained Smoke BirthWeightGm   Low Premie Marital
##    <int> <int>  <int> <int>  <int> <int>         <dbl> <int>  <int>   <int>
## 1      1     1     32    40     38     0         3147.     0      0       0
## 2      1     2     32    37     34     0         3289.     0      0       0
## 3      1     1     27    39     12     0         3912.     0      0       0
## 4      1     1     27    39     15     0         3856.     0      0       0
## 5      1     1     25    39     32     0         3430.     0      0       0
## 6      1     1     28    43     32     0         3317.     0      0       0
```

The column contents are as follows:

- **Plural**: 1=single birth, 2=twins, 3=triplets.
- **Sex**: sex of the baby 1=male 2=female.
- **MomAge**: Mother's age (in years).

- **Weeks**: Completed weeks of gestation.
- **Gained**: Weight gained during pregnancy (in pounds).
- **Smoke**: Mother is a smoker: 1=yes, 0=no.
- **BirthWeightGm**: Birth weight in grams.
- **Low**: Indicator for low birth weight, 1=2500 grams or less, 0=otherwise.
- **Premie**: Indicator for premature birth, 1=36 weeks or sooner, 0=otherwise.
- **Marital**: Marital status: 0=married or 1=not married.

Problem 1: (5 pts)

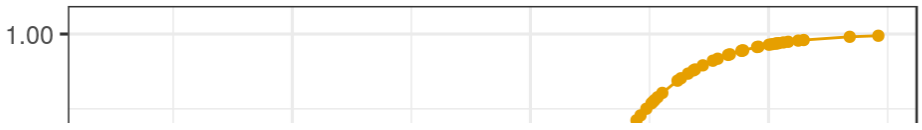
a. (1 pt) Make a logistic regression model that predicts premature births (`Premie`) from birth weight (`BirthWeightGm`), plural births (`Plural`), and weight gained during pregnancy (`Gained`) in the `NCbirths` data set. Show the summary (using `summary`) of your model below.

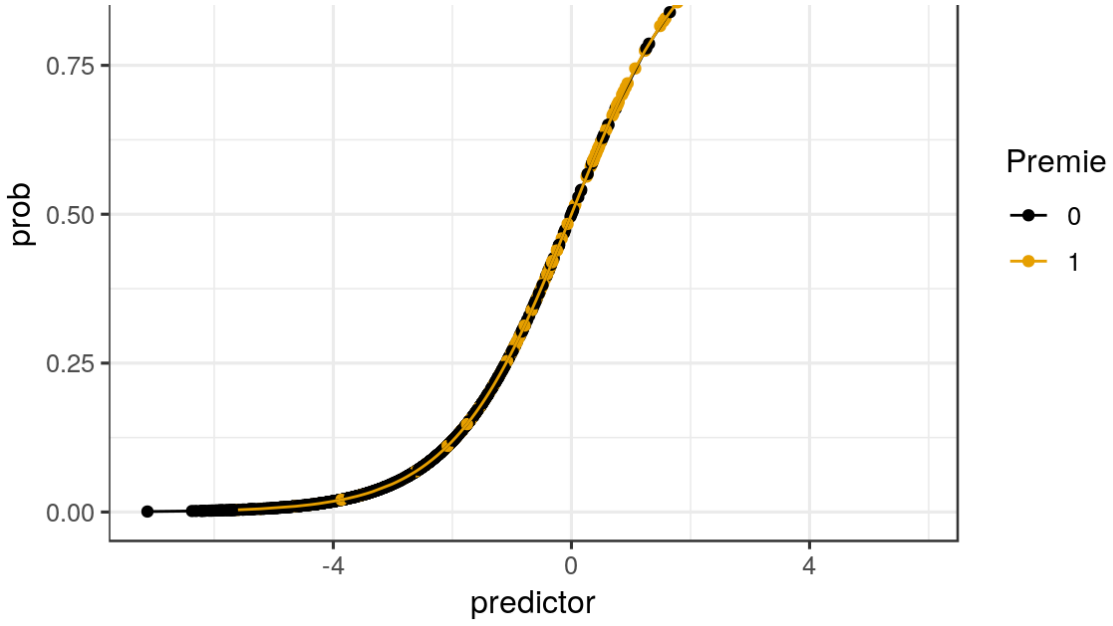
```
# Creating a logistic regression model

# model to use:
# Premie ~ BirthWeightGm + Plural + Gained
glm_out <- glm(
Premie ~ BirthWeightGm + Plural + Gained,
data = NCbirths,
family = binomial
)

# Summary of that model
summary(glm_out)
```

```
##
## Call:
## glm(formula = Premie ~ BirthWeightGm + Plural + Gained, family = binomial,
##     data = NCbirths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9120  -0.4410  -0.2964  -0.1693   3.3480
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.8586938  0.7962407   6.102 1.05e-09 ***
## BirthWeightGm -0.0025932  0.0002102 -12.335  < 2e-16 ***
## Plural         0.6713701  0.3779965   1.776   0.0757 .
## Gained         0.0130966  0.0073036   1.793   0.0729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1080.56  on 1408  degrees of freedom
## Residual deviance:  745.54  on 1405  degrees of freedom
## AIC: 753.54
##
## Number of Fisher Scoring iterations: 6
```

b. (1 pt) Make a plot of the fitted probability as a function of the linear predictor, colored by the indicator of premature births.

```
# Creating the dataframe with predictors, probability and Premie
lr_data <- data.frame(
predictor=glm_out$linear.predictors,
prob=glm_out$fitted.values,
Premie=factor(NCbirths$Premie)
)

# Plotting the fitted probabilty as a function of linear predictor colored by t
he indicator of premature births
ggplot(lr_data, aes(x = predictor, y = prob, color = Premie)) +
  geom_point() +
  geom_line()+
  scale_color_colorblind()
```





c. (3 pts) Use the probability cut-off of 0.50 to classify a birth as premature or non-premature. Calculate the **true positive rate** and the **false positive rate** and interpret these rates in the context of the `NCbirths` dataset. Your answer should mention something about premature births and the three predictors in part a.

```
 # cutoff of 0.5
cutoff <- 0.5

# True positive
premature_true <-
  lr_data %>%
  filter(prob > cutoff & Premie=="1") %>%
  tally()

# False negative
premature_false <-
  lr_data %>%
  filter(prob <= cutoff & Premie=="1") %>%
  tally()
```

```
# True negative
nopremature_true <-
  lr_data %>%
  filter(prob <= cutoff & Premie=="0") %>%
  tally()

# False positive
nopremature_false <-
  lr_data %>%
  filter(prob > cutoff & Premie=="0") %>%
  tally()
```

```
# True positive rate = TP/(TP+FN)
tpr <- premature_true/(premature_true + premature_false)

# True negative rate = TN/(TN+FP)
tnr <- nopremature_true/(nopremature_true + nopremature_false)
tpr
```

```
##           n
## 1 0.3646409
```

```
# False positive rate= FP/(FP+TN)
fpr <- nopremature_false/(nopremature_true + nopremature_false)
fpr
```

```
##            n
## 1 0.01547231
```

```
# We can verify by fpr=1-tnr
# 1-tnr
```

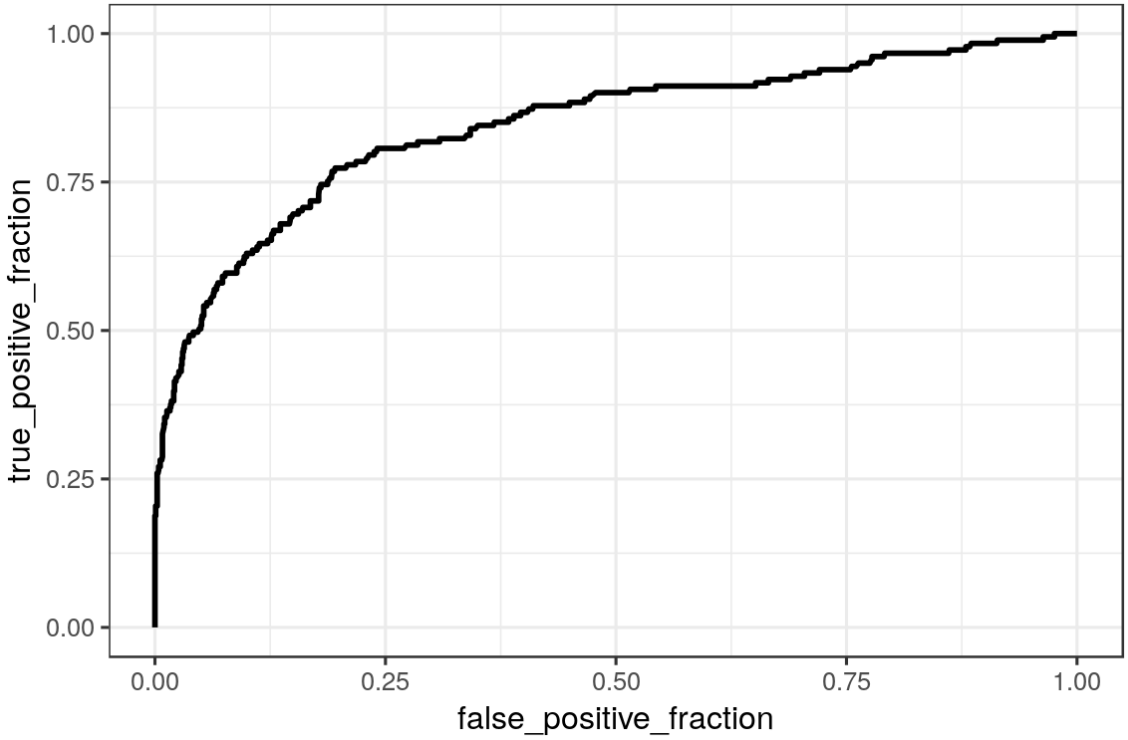
True positive rate is 0.3646 which is less, that means predictors don't tell that well if a baby is Premature or not. False positive rate is 0.01547 which is less which means predictors are really good in telling if baby is not premature. Looking at the estimates, we see that only Plural has significant positive value and other two are almost equal to zero so Premature birth mostly depends on Plural and from TPR and FPR values we conclude predictors especially number of births(Plural) may not predict Premature births well but are good at predicting non premature births.

Problem 2: (5 pts)

a. (1 pt) Plot an ROC curve for the model that you created in problem 1a. Does the model perform better than a model in which you randomly classify a birth as premature or non-premature? Explain your

answer in 2-3 sentences.

```
# Data frame for ROC containing predicted and known truth for Model 1
df <- data.frame(
predictor = predict(glm_out, NCbirths),
known_truth = NCbirths$Premie,
dataname = 'Model1'
)


# Calculating ROC for Model 1
p<-ggplot(df, aes(d = known_truth , m = predictor)) +
geom_roc(n.cuts = 0)
p
```




```
calc_auc(p)
```

```
##   PANEL group     AUC
## 1     1    -1 0.84453
```

Yes, the model performs better than randomly classify as curve is higher than line y=x. Another way to tell is AUC here is 0.84453, which is greater than 0.5 so it is better than randomly classifying.

b. (4 pts) Use the mothers' marital status (`Marital`) and the mothers' age (`MomAge`) as a new set of

predictor variables for premature births, and create a logistic regression model. Plot an ROC curve for your newly-created model and, on the same plot, add an ROC curve from your model in problem 1a. What can you conclude from your plot? Which model performs better and why? Support your conclusions **with AUC values for each model**.

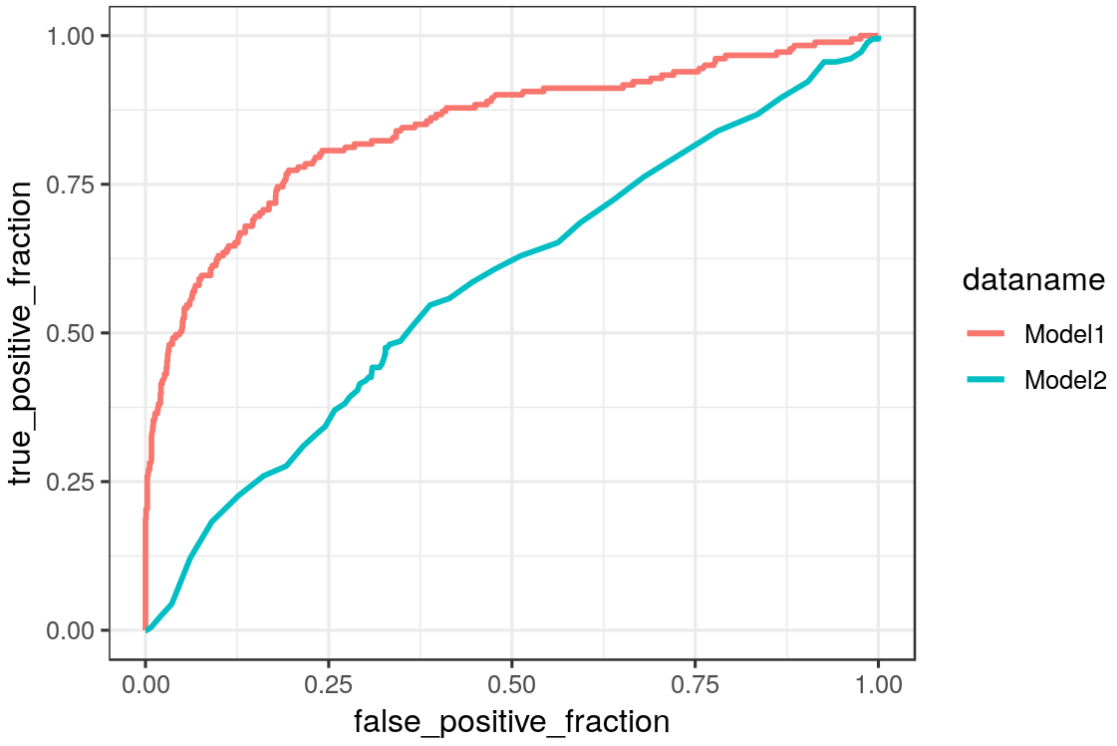
```
# Creating a logistic regression model

# Premie ~ Marital + MomAge
glm_out2 <- glm(
Premie ~ Marital + MomAge,
data = NCbirths,
family = binomial
)

# Summary of the model
summary(glm_out2)
```

```
##
## Call:
## glm(formula = Premie ~ Marital + MomAge, family = binomial, data = NCbirths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6481  -0.5963  -0.4757  -0.4604   2.1912
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.81255    0.44177  -4.103 4.08e-05 ***
## Marital      0.50796    0.18399   2.761  0.00577 **
## MomAge      -0.01147    0.01498  -0.765  0.44420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1080.6  on 1408  degrees of freedom
## Residual deviance: 1067.4  on 1406  degrees of freedom
## AIC: 1073.4
##
## Number of Fisher Scoring iterations: 4
```

```
# Data frame for ROC containing predicted and known truth for Model 2
df2 <- data.frame(
predictor = predict(glm_out2, NCbirths),
known_truth = NCbirths$Premie,
dataname = 'Model2'
)

# Calculating ROC for Model 2
p2<-ggplot() +
geom_roc(data=df2, mapping= aes(d = known_truth , m = predictor),n.cuts = 0)
calc_auc(p2)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5812623
```

```
# Combined plot with both the ROC curves
p3<-ggplot() +
geom_roc(data=df, mapping= aes(d = known_truth , m = predictor, color = datanam
e),n.cuts = 0) +
geom_roc(data=df2, mapping =aes(d = known_truth , m = predictor, color = datana
me),n.cuts = 0)


p3
```



We can see that model developed in 1a is better than the current model as the AUC value of that model is 0.844 whereas for this, it is 0.581. Higher the value better the model.