Project 2

Rohit Kamat rgk359

For the project, I worked on a dataset collected by John Holcomb from the North Carolina State Center for Health and Environmental Statistics. This dataset contains 1409 birth records from North Carolina in 2001. The goal for this project was to analyze the births dataset using several statistical approaches I have learned, in two parts.

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

train_fraction <- 0.8 # fraction of data for training purposes
set.seed(123)  # set the seed to make your partition reproductible
train_size <- floor(train_fraction * nrow(NCbirths)) # number of observations in trainin
g set
train_indices <- sample(1:nrow(NCbirths), size = train_size)

train_data <- NCbirths[train_indices, ] # get training data
test_data <- NCbirths[-train_indices, ] # get test data

head(NCbirths)
```

```
##   Plural Sex MomAge Weeks Gained Smoke BirthWeightGm Low Premie Marital
## 1      1   1     32    40     38     0       3146.85   0      0       0
## 2      1   2     32    37     34     0       3288.60   0      0       0
## 3      1   1     27    39     12     0       3912.30   0      0       0
## 4      1   1     27    39     15     0       3855.60   0      0       0
## 5      1   1     25    39     32     0       3430.35   0      0       0
## 6      1   1     28    43     32     0       3316.95   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).
- **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.

Part 1. I have divided the dataset, which consists of observations from 1409 individuals, into a training and a test data set. I fit a logistic regression model to predict marital status on the training data set.

Using the final model, I predict the outcome on the test data set, plot, and discuss my results. **Part 2 (60 points)**. I created two question that I solved using logistic modeling and liner regression model. First question was to determine if premature babies tend to have lower birthweight than non-premature babies. My second question was whether the weight of the mother gained during pregnancny,the sex of the baby, whether the baby is premature or not and weeks of gestation are statistically significant predictors in determining the weight of a baby.

Project responses should be entered below.

```
# This R code chunk contains the calc_ROC function.
calc_ROC <- function(probabilities, known_truth, model.name=NULL)
  {
  outcome <- as.numeric(factor(known_truth))-1
  pos <- sum(outcome) # total known positives
  neg <- sum(1-outcome) # total known negatives
  pos_probs <- outcome*probabilities # probabilities for known positives
  neg_probs <- (1-outcome)*probabilities # probabilities for known negatives
  true_pos <- sapply(probabilities,
                     function(x) sum(pos_probs>=x)/pos) # true pos. rate
  false_pos <- sapply(probabilities,
                     function(x) sum(neg_probs>=x)/neg)
  if (is.null(model.name))
    result <- data.frame(true_pos, false_pos)
  else
    result <- data.frame(true_pos, false_pos, model.name)
  result %>% arrange(false_pos, true_pos)
  }
```

Part 1

```
# Do a logisitic regression model of the dataset, the significant predictors were mother's age and birth weight in grams
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Plural + Sex + Weeks + Gained + Low + Premie + Smoke , train_data, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Plural + Sex +
##     Weeks + Gained + Low + Premie + Smoke, family = binomial,
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9493  -0.7683  -0.4282   0.8492   2.8552
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.3884139  1.7584645   3.064  0.00218 **
## MomAge        -0.2118561  0.0153721 -13.782  < 2e-16 ***
## BirthWeightGm -0.0002669  0.0001834  -1.455  0.14559
## Plural        -0.2350152  0.4487728  -0.524  0.60050
## Sex           -0.0071716  0.1482256  -0.048  0.96141
## Weeks          0.0113199  0.0426993   0.265  0.79093
## Gained        -0.0030391  0.0053100  -0.572  0.56709
## Low           -0.1574192  0.3641721  -0.432  0.66555
## Premie         0.4467372  0.3089124   1.446  0.14813
## Smoke          0.7980204  0.1989608   4.011 6.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1121.3  on 1117  degrees of freedom
## AIC: 1141.3
##
## Number of Fisher Scoring iterations: 5
```

```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Plural + Weeks + Gained + Low + Premie
 + Smoke , train_data, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Plural + Weeks +
##     Gained + Low + Premie + Smoke, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9478  -0.7683  -0.4290   0.8479   2.8542
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.3802769  1.7504953   3.074  0.00212 **
## MomAge        -0.2118910  0.0153561 -13.798  < 2e-16 ***
## BirthWeightGm -0.0002658  0.0001821  -1.460  0.14426
## Plural        -0.2349377  0.4488932  -0.523  0.60072
## Weeks          0.0111907  0.0426194   0.263  0.79288
## Gained        -0.0030482  0.0053065  -0.574  0.56568
## Low           -0.1564754  0.3636828  -0.430  0.66701
## Premie         0.4466962  0.3089119   1.446  0.14817
## Smoke          0.7980537  0.1989606   4.011 6.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1121.3  on 1118  degrees of freedom
## AIC: 1139.3
##
## Number of Fisher Scoring iterations: 5
```

```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Plural + Gained + Low + Premie + Smoke
 , train_data, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Plural + Gained +
##     Low + Premie + Smoke, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9412  -0.7707  -0.4269   0.8471   2.8490
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.7872617  0.8194483   7.062 1.64e-12 ***
## MomAge        -0.2120732  0.0153422 -13.823  < 2e-16 ***
## BirthWeightGm -0.0002516  0.0001739  -1.447    0.148
## Plural        -0.2433990  0.4491066  -0.542    0.588
## Gained        -0.0030780  0.0053034  -0.580    0.562
## Low           -0.1595719  0.3635008  -0.439    0.661
## Premie         0.3990960  0.2500440   1.596    0.110
## Smoke          0.8000626  0.1988558   4.023 5.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1121.4  on 1119  degrees of freedom
## AIC: 1137.4
##
## Number of Fisher Scoring iterations: 5
```

```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Plural + Gained + Premie + Smoke , tra
in_data, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Plural + Gained +
##     Premie + Smoke, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9189  -0.7655  -0.4323   0.8513   2.8618
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.6826963  0.7849267   7.240 4.49e-13 ***
## MomAge        -0.2121482  0.0153388 -13.831  < 2e-16 ***
## BirthWeightGm -0.0002107  0.0001466  -1.438    0.151
## Plural        -0.2830023  0.4428889  -0.639    0.523
## Gained        -0.0030092  0.0052983  -0.568    0.570
## Premie         0.3776815  0.2453941   1.539    0.124
## Smoke          0.8007287  0.1989279   4.025 5.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1121.6  on 1120  degrees of freedom
## AIC: 1135.6
##
## Number of Fisher Scoring iterations: 5
```

```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Plural + Premie + Smoke , train_data,
family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Plural + Premie +
##     Smoke, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9148  -0.7646  -0.4333   0.8453   2.8859
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.6745226  0.7838778   7.239 4.52e-13 ***
## MomAge        -0.2115878  0.0152949 -13.834  < 2e-16 ***
## BirthWeightGm -0.0002292  0.0001429  -1.604    0.109
## Plural        -0.3190235  0.4375472  -0.729    0.466
## Premie         0.3742500  0.2450667   1.527    0.127
## Smoke          0.7978181  0.1989472   4.010 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1121.9  on 1121  degrees of freedom
## AIC: 1133.9
##
## Number of Fisher Scoring iterations: 5
```

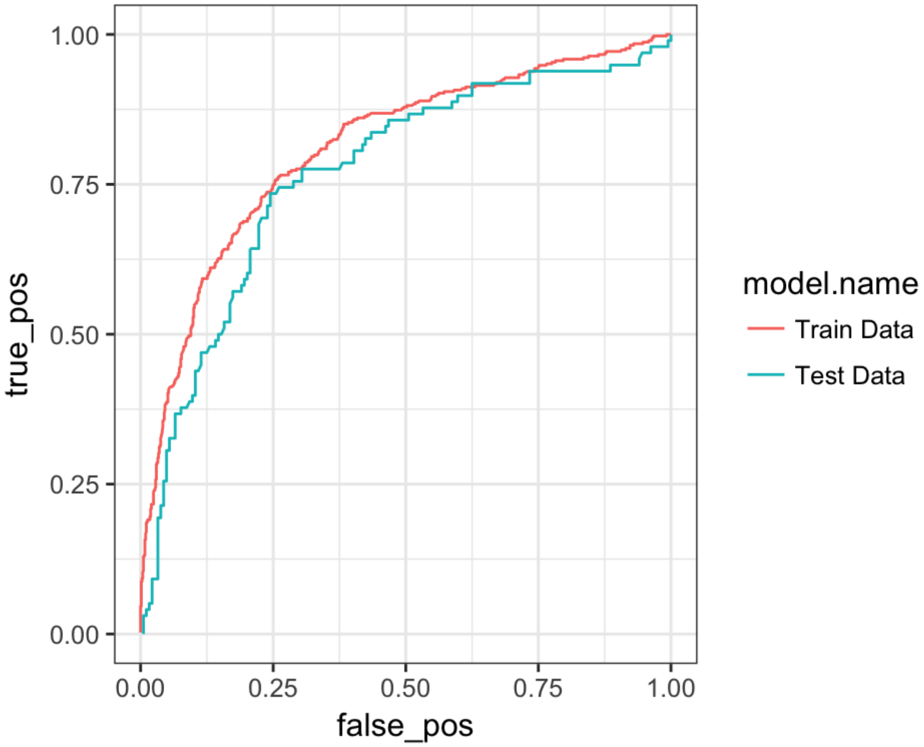
```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Premie + Smoke , train_data, family=binomial)
summary(glm.out)
```

```
## 
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Premie + Smoke, 
##     family = binomial, data = train_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9006  -0.7604  -0.4358   0.8429   2.9019  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    5.2925173  0.5793422   9.135  < 2e-16 ***
## MomAge        -0.2129544  0.0152030 -14.007  < 2e-16 ***
## BirthWeightGm -0.0002020  0.0001379  -1.465    0.143    
## Premie         0.3435171  0.2412112   1.424    0.154    
## Smoke          0.8109828  0.1983385   4.089 4.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1122.5  on 1122  degrees of freedom
## AIC: 1132.5
## 
## Number of Fisher Scoring iterations: 5
```

```
glm.out <- glm(Marital ~ MomAge + BirthWeightGm + Smoke, train_data, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = Marital ~ MomAge + BirthWeightGm + Smoke, family = binomial,
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8805  -0.7614  -0.4364   0.8507   2.8602
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.6492943  0.5251789  10.757  < 2e-16 ***
## MomAge        -0.2121187  0.0151430 -14.008  < 2e-16 ***
## BirthWeightGm -0.0003022  0.0001188  -2.542    0.011 *
## Smoke          0.8014536  0.1982184   4.043 5.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1451.2  on 1126  degrees of freedom
## Residual deviance: 1124.5  on 1123  degrees of freedom
## AIC: 1132.5
##
## Number of Fisher Scoring iterations: 5
```

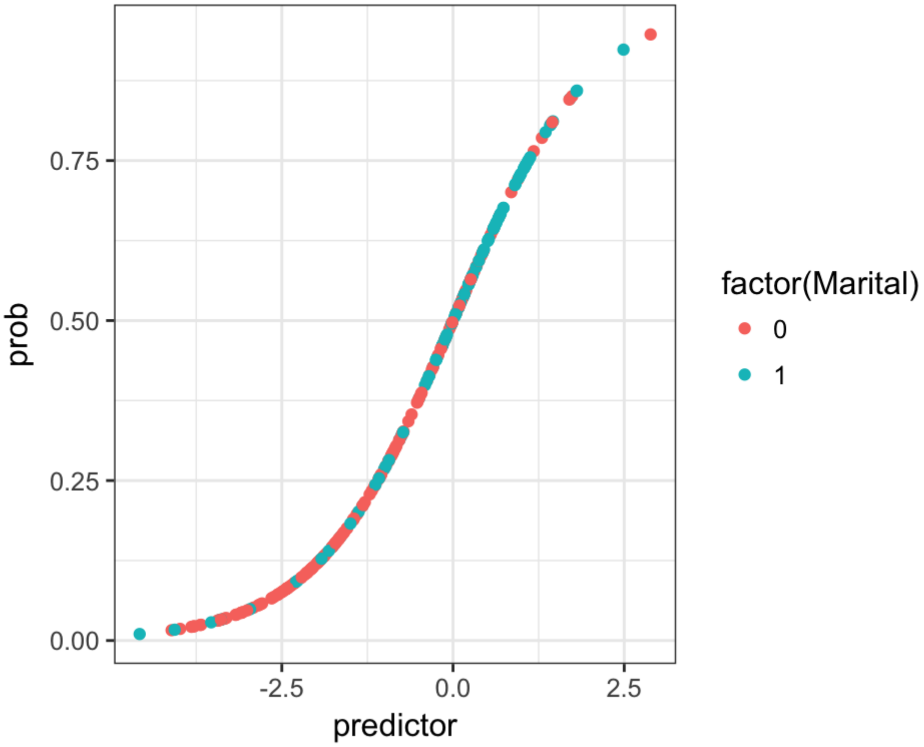
```
#Using the final model predict the outcome on the test data set
#Predict the outcome on the Train Data Set using the calc_ROC function
test_prob <- predict(glm.out, test_data, type="response")
ROC1 <- calc_ROC(probabilities=glm.out$fitted.values,
                 known_truth=train_data$Marital,
                 model.name="Train Data")
#Predict the outcome on the Test Data Set using the calc_Roc function
ROC2 <- calc_ROC(probabilities=test_prob,
                 known_truth=test_data$Marital,
                 model.name="Test Data")
ROC_all <-rbind(ROC1, ROC2)
#Perform the ggPlot
ggplot(ROC_all, aes(x=false_pos, y=true_pos, color=model.name)) + geom_line()
```



```
#Find the AUC of both Test Data and Train Data
ROC_all %>% group_by(model.name) %>%
  mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T)) %>%
  arrange(desc(AUC))
```

```
##        AUC
## 1 1.580293
```

```
#Plot of the fitted probability of marital status as a function of the predictors, color
ed by marital status, on the test data
test_pred <- predict(glm.out, test_data )
log_data <- data.frame(predictor=test_pred, prob=test_prob, Marital=test_data$Marital)
ggplot(log_data, aes(x=predictor, y=prob,color= factor(Marital))) + geom_point()
```



The level of significance that I choose for the regression model 95%. The final model uses two predictors MomAge and BirthWeightGm. Then I performed a plot for the two ROC curves of both the train data and test data. Best on the area under the curve (AUC), the train data has .81 compared to test data which is .77. Therefore my final predictors of MomAge and BirthWeightGm both provided a good fit in the train data and test data set. Then a fitted probability plot was also made to be able to predict the outcome of marital status based on the predictors of mom’s age and birth weight. Based on the plot, there was not a seperation of marital status based on the predictor value. Therefore based on the data it would be hard to infer, based on the age of the mother and birthweight of the baby, if the mother is married or not.

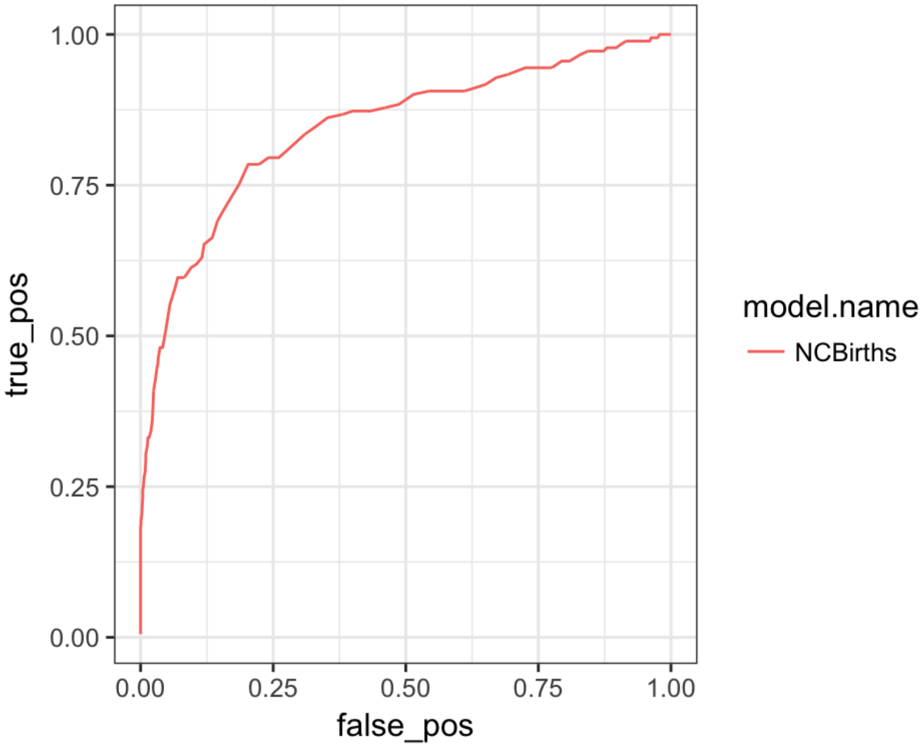
Part 2

Are premature babies tend to have lower birthweight than non-premature babies?

```
#Do a Logistic Regression Model, BirthWeightGm as the independent variable and Premie as
 the dependent variable
glm3.out <- glm(Premie ~ BirthWeightGm , data= NCbirths, family=binomial)
summary(glm3.out)
```

```
##
## Call:
## glm(formula = Premie ~ BirthWeightGm, family = binomial, data = NCbirths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7656  -0.4414  -0.2961  -0.1699   3.3868
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.1255508  0.5782306   10.59   <2e-16 ***
## BirthWeightGm -0.0026472  0.0001963  -13.49   <2e-16 ***
## ---
## 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:  754.35  on 1407  degrees of freedom
## AIC: 758.35
##
## Number of Fisher Scoring iterations: 6
```

```
#Plot an ROC Curve
Birth_ROC <- calc_ROC(probabilities=glm3.out$fitted.values,
                 known_truth=NCbirths$Premie,
                 model.name="NCBirths")
ggplot(Birth_ROC, aes(x=false_pos, y=true_pos)) +
  geom_line(data=Birth_ROC, aes(color=model.name))
```



```
##Find the AUC of the ROC Curve
Birth_ROC %>% group_by(model.name) %>%
  mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T)) %>%
  arrange(desc(AUC))
```

```
##         AUC
## 1 0.8483047
```

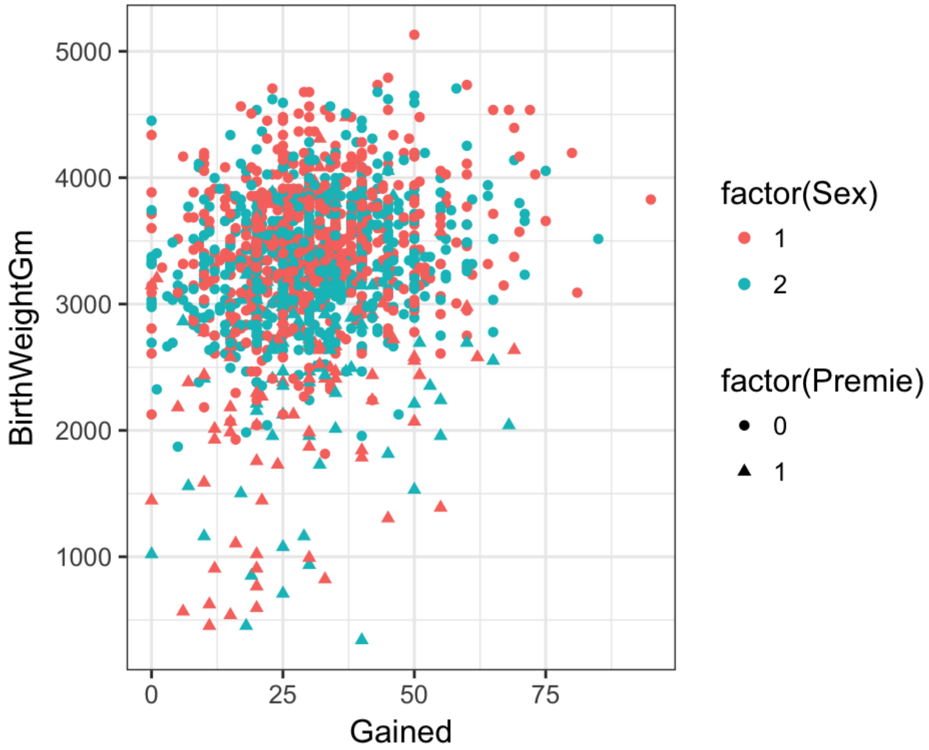
Based on the summary of the logistic model, an increase of one gram of birthweight results in the log odds of a premature baby decreases by .002 with a p-value less than 0.05. An ROC curve was then preformed to determine how well birthweight is able to predict if a baby is premature or not. The ROC curve showed a plot like the model’s predictions were good with an Area Under The Curve equal to .848. From the model as well as the ROC curve I conclude that based on the data premature babies tend to have lower birthweight than non-premature babies.

Is the weight of the mother gained during pregnancny,the sex of the baby, whether the baby is premature or not and weeks of gestation statistically significant predictors in determining the weight of a baby?

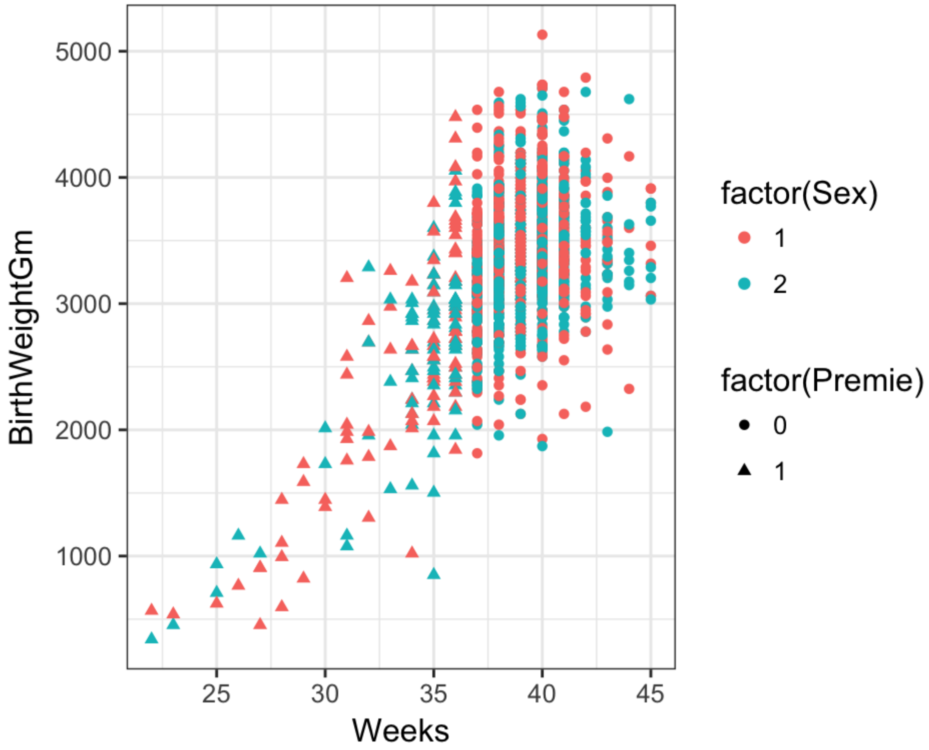
```
# Perform a linear model to find if the predictors are statistically significant in determining the amount of the weight of the baby.

NCbirths_lm <- lm(BirthWeightGm ~ Gained + Sex + Premie + Weeks , NCbirths)
summary(NCbirths_lm)
```

```
## 
## Call:
## lm(formula = BirthWeightGm ~ Gained + Sex + Premie + Weeks, data = NCbirths)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1694.2  -320.1    -9.0   334.2  1643.5 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -965.3189   279.5201  -3.453 0.000570 ***
## Gained         7.1123     0.9473   7.508 1.06e-13 ***
## Sex          -94.3066    26.2730  -3.589 0.000343 ***
## Premie      -306.5226    55.8856  -5.485 4.90e-08 ***
## Weeks        109.4105     7.0633  15.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 492.3 on 1404 degrees of freedom
## Multiple R-squared:  0.3859, Adjusted R-squared:  0.3841 
## F-statistic: 220.5 on 4 and 1404 DF,  p-value: < 2.2e-16
```

```
#Perform The Plots of Each of the Variable
ggplot(NCbirths, aes(x=Gained, y=BirthWeightGm, color=factor(Sex), shape=
factor(Premie))) + geom_point()
```



```
ggplot(NCbirths, aes(x=Weeks, y=BirthWeightGm, color=factor(Sex), shape=
factor(Premie))) + geom_point()
```



To answer this question a multiple linear regression was performed to see what predictors significantly determine the weight of the baby. From the model the amount of weight that a mother has is a significant predictor of the birth weight of the baby with p<.05. The sex of the baby was a significant predictor in determining the birth weight of the baby with p<.05. Whether the baby was premature was a statistically significant predictor in determining the weight of the baby with a p<.05.The amount of weeks of gestation was a significant predictor in determining the weight of the baby with a p <.05.

I performed two graphs: the linear relationship between the amount of weight a mother gains and the birthweight of a baby with categorical predictors of sex and premature. Based on the first graph there seems to be a moderate positive linear relationship between the gain weight of mother and the birthweight of the baby. From the graph the sex of the baby did not have an significant impact on the weight of the baby if the mother gained weight, but premature babies tend to weigh less and the mothers tend to gained little weight during pregnancy.

On the second graph there seemed to be a strong positive linear relationship between the weeks of gestation and the birthweight of the baby. Looking at the graph sex does not seem to effect the weight of the baby when the amount of weeks of gestation increases, but premature babies tend to have lower weeks of gestation and lower birth weight.

From the regression model and the graph the variables gained, sex, premie, and weeks are signficant variables for determining the birthweight of the baby. I did not look at the interaction effect on categorical vairables of sex and premie on quantitative variables gained and weeks in determining the birthweight of the baby. Based on the graphs I would hypothesize that sex did not affect the variables of gained and weeks in determing the birthweight of a baby but I do believe that premie does intereact with gained and weeks in determiing the birthweight of a baby.