Modeling and Representing Data Logisitic Regession

Derek Wales

27SEP19

Summary:

For this Homework we were analyzing a data set that had a myriad of different mothers all of whom had different characteristics and developing a model to answer several research questions. This was accomplished through Exploratory Data Analysis, Critical Thinking, and Logistic Regression. Which ultimately lead to the model and coefficients below.

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Black + \beta_2 Asian + \beta_3 MotherWeight + \beta_3 Smoking$$

where $y_i|x_i \sim$ Bernouilli(π_i)

Using our final model, we were able to answer these key research questions (code in the appendix).

Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the odds ratio of pre-term birth for smokers and non-smokers?

Yes, based upon our model the confidence interval for mothers who do smoke increase the odds of a premature within a confidence interval (5%/95%) 0.011320313/0.714031603, log scale.

Is there any evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race? If so, characterize those differences.

Yes, the odds of being premautre go up if you are a smoker, and for Black and Asian mothers as well (confidence interval in the model section).

Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

Mothers weight had a larger effect than anticipated.

Intro:

For this homework students were provided a simplified data from a Yerushalmy Study which collected the details of more than 15,000 births at the Kaiser Foundation Hospital in Oakland, Ca. The first step in analysis was getting a clean form of it into R. I did this using the read.csv function and putting the results into a dataframe (referred to as Smoking). Upon inspection I realized that the races in the dataframe were similar. Therefore, I also grouped the first six categories of the race variable into one because they were all variations of Caucasian, and as well as education categories 6 and 7 (trade school/uncertain if they finished high school) because of their similarities. Additionally, I added a new factor variable to the dataframe, premature_f (Note: Varun Prasad, MIDS 2021 assisted me in getting the premature weights as factors). This variable is valued at 1 or 0 (1 for if the baby is premautre, 0 if not), and treated it as a categorical variable. This made it much easier to see if when are response variable was true. Instead of trying to find a numerical value with a specific cut off that would have been required if I used gestation time.

Data and Exploratory Data Analysis:

With the new dataframe we could begin exploratory data analysis. This is the process of looking at our data in a series of formats to see which factors will be the most influential when we develop our model. (Note: BirthID and Date were not explored because they have no effect on baby weight).

I also did not put birthweight in the model, because birthweight is a response to various things about the mother through the pregnancy. I focused all of my analysis on traits of the mother that could potentially determine if the baby was premature.

Since I had a categorical variable as my response variable, I looked a few of the graphs with numerical values first. Starting with the age of the mother (plot in appendix). After that I explored the other numerical variables, like height and mother weight, each with a boxplot. From that analysis it was hard to draw any conclusions on what the relationship may be, none of the plots had a distinct difference that I could see if the baby would be premature or not.

Next, I began to look to at the other categorical variables in the data set, but instead of using graphs, I used a Chi-Squared test to determine if each variable, when compared with my premature as factor variable, was strong enough to reject the null hypothesis (using .05 as a cut off).

Going through the dataset there was some evidence that Race, Education, Parity and Smoking may affect whether the baby was premature (based upon p value being less than .05 or near it). Additionally, based upon its high p value for income (0.9087) I knew to be skeptical of any final model that included that as a term.

Of note, the p value for smoking was 0.0694, which is close to the threshold of 0.05. So I decided to pay close attention to it when I was building the model, especially since it was a specific research question.

After completing Exploratory Data Analysis, I had an idea of what the terms in my model should look like and what was likely to be in the model I went through the selection process.

Model Selection Process and Assessment:

Moving into model selection, I decided to keep all the variables (except for the birth weight, date and baby ID) and work forward using the Akaike information criterion (abbreviated as AIC) to determine the best model. I chose the AIC over the Bayesian Information Criterion (or BIC) because we had a complete data set and I did not want to penalize the model for including more information that was readily available.

Using R to "work forward," I continually put each one of the variables and comparing it against the "null model" (a model which states that the coefficients have no effect). From this I was able to find which model gave me the lowest AIC using the "working forward" methodology. (Note Nathan Warren, MIDS 2021 assisted me in using the stepwise function).

I then repeated the same process, except "working backward" using R to iteratively take values away from the model. Once I had a result, I was then able to compare the results between working forward and backward and see which most accurately represented the data.

From this I was able to see that working forward resulted in the lowest (best) AIC value and that there were only a few statistically significant variables, race (category 7 and 8 Black and Asian respectively) and the pregnant weight of the mother. I also specifically added smoking back into the model because it was one of the main research questions.

I then when building the Logistic Regression modeled, I used these four factors which I ultimately resulted in the model below. Interestingly, adding smoking back into the model did result in it generating a signification p value (below 0.05).

Final Model

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Black + \beta_2 Asian - \beta_3 MotherWeight + \beta_3 Smoking$$

where $y_i|x_i \sim$ Bernouilli(π_i)

Coefficients in table below

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4634	0.6149	-0.75	0.4511
Black	0.7576	0.2116	3.58	0.0003
Asain	0.8124	0.3954	2.05	0.0399
Mother Weight	-0.0109	0.0047	-2.31	0.0211
Smoker	0.3614	0.1790	2.02	0.0435

Conclusion/Remarks:

The final model had the terms Black, Asian, Smoking (all factor variables), and Mother Weight (a numerical variable) as coefficients. The exponentiated values for the factor variables were Black - 0.5738, Asian - 0.6599, and Smoking - 0.1306. This means that with everything else being held constant, if a mother is Black it increases the odds that the baby will be born premature by 0.5738. Additionally, the weight of the mother was also a significant factor in determining if the baby would be premature, with the log odds being -0.0109, with the exponentiated value being 0.0001, this means as the mother gets heavier the odds drop of the baby being premature.

From here we had our final model the next step was to do Model Assessment. The primary way of doing this was by plotting the residuals against the fitted values. With the goal of seeing if anything fell outside of the confidence interval, depicted by the red lines in the graph in the appendix. From this we can infer how effectively our model is selecting against the actual data by seeing if many of our model points fall outside of the confidence interval.

Additionally I had to validate the sensitivitiy (or the true positive rate), the specificity (the true negative rate), and accuracy (how often the model predicts correctly). These values were 0.5731, 0.6226, and 0.6133 respectively. This means that holistically the model is prone to errors when predicting whether or not a mother will have a baby prematurely.

The next step was to plot the residuals with sensitivity and specificity, also known as the Area Under the Curve (or AOC Graph). A perfect fit for a model would have the AOC = 1 and know effect would be 0.5 (the same results as a coin toss). Based upon this our model with an AOC 0.618 was not a particularly strong fit.

Although our model does help predict whether a baby will be premature, it is hard to make lifestyle inferences from the data. Specifically, if the mother is heavier, it is less likely that the baby will be premature. However, weight in and of itself is not necessarily the best measure because sometimes people can gain weight and not necessarily become healthier. We also may not have a large enough sample size for some of the minority races.

This means that according to our model, the odds of a baby being premature are 0.6291 if the mother is Black or Asian, the chances increase by 0.5738/0.6599 respectively, and 0.1306 if the mother is a smoker. Those odds decrease as the mother gets heavier.

Appendix (Note: EDA at the end for readability)

```
# Model Selection Code
full_model <- glm(premature_f ~ as.factor(smoke) + as.factor(race) + parity
                  + as.factor(Edu) + as.factor(inc) + mht + mpregwt,
                  family = binomial, Smoking)

null_model <- glm(premature_f~1 ,family = binomial, Smoking)

BIC_forward <- step(full_model, trace = 0, direction = "both")
#AIC_forward$call
summary(BIC_forward)
```

```
##
## Call:
## glm(formula = premature_f ~ as.factor(smoke) + as.factor(race) +
##     as.factor(Edu) + mpregwt, family = binomial, data = Smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7356  -0.6722  -0.5576  -0.4106   2.4417
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.684033   1.124515   0.608 0.542994
## as.factor(smoke)1  0.288894   0.184335   1.567 0.117062
## as.factor(race)6   0.154865   0.516398   0.300 0.764258
## as.factor(race)7   0.770692   0.222714   3.460 0.000539 ***
## as.factor(race)8   0.905970   0.407690   2.222 0.026269 *
## as.factor(race)9  -0.752839   1.051521  -0.716 0.474020
## as.factor(Edu)1   -0.541029   0.948966  -0.570 0.568593
## as.factor(Edu)2   -0.887619   0.940660  -0.944 0.345368
## as.factor(Edu)3   -0.706669   0.993071  -0.712 0.476713
## as.factor(Edu)4   -1.547889   0.955839  -1.619 0.105360
## as.factor(Edu)5   -1.064617   0.958531  -1.111 0.266708
## as.factor(Edu)7    1.825677   1.484105   1.230 0.218640
## mpregwt           -0.012147   0.004844  -2.508 0.012147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 795.91  on 856  degrees of freedom
## AIC: 821.91
##
## Number of Fisher Scoring iterations: 5
```

```
BIC(BIC_forward)
```

```
## [1] 883.8879
```

```
AIC_backward <- step(null_model, scope = formula(full_model), trace = 0, direction = 'backward')
AIC_backward$call
```

```
## glm(formula = premature_f ~ 1, family = binomial, data = Smoking)
```

```r
summary(AIC_backward)
```

```
##
## Call:
## glm(formula = premature_f ~ 1, family = binomial, data = Smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6468  -0.6468  -0.6468  -0.6468   1.8262
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45833    0.08669  -16.82   <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: 841.83  on 868  degrees of freedom
## Residual deviance: 841.83  on 868  degrees of freedom
## AIC: 843.83
##
## Number of Fisher Scoring iterations: 4
```

```r
#Building the final model because it was part of the question
final_model_premature <- glm(premature_f ~ as.factor(race) + mpregwt +as.factor(smoke),
                             family = binomial, Smoking)
summary(final_model_premature)
```

```
##
## Call:
## glm(formula = premature_f ~ as.factor(race) + mpregwt + as.factor(smoke),
##     family = binomial, data = Smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0253  -0.6707  -0.5806  -0.4810   2.1957
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.463443   0.614941  -0.754 0.451067
## as.factor(race)6   0.505334   0.487853   1.036 0.300281
## as.factor(race)7   0.757554   0.211588   3.580 0.000343 ***
## as.factor(race)8   0.812376   0.395397   2.055 0.039919 *
## as.factor(race)9  -0.871072   1.044921  -0.834 0.404492
## mpregwt           -0.010900   0.004726  -2.307 0.021077 *
## as.factor(smoke)1  0.361417   0.179025   2.019 0.043507 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 816.34  on 862  degrees of freedom
```

```
## AIC: 830.34
##
## Number of Fisher Scoring iterations: 5
```

```
exp(confint(final_model_premature, level = 0.95))
```

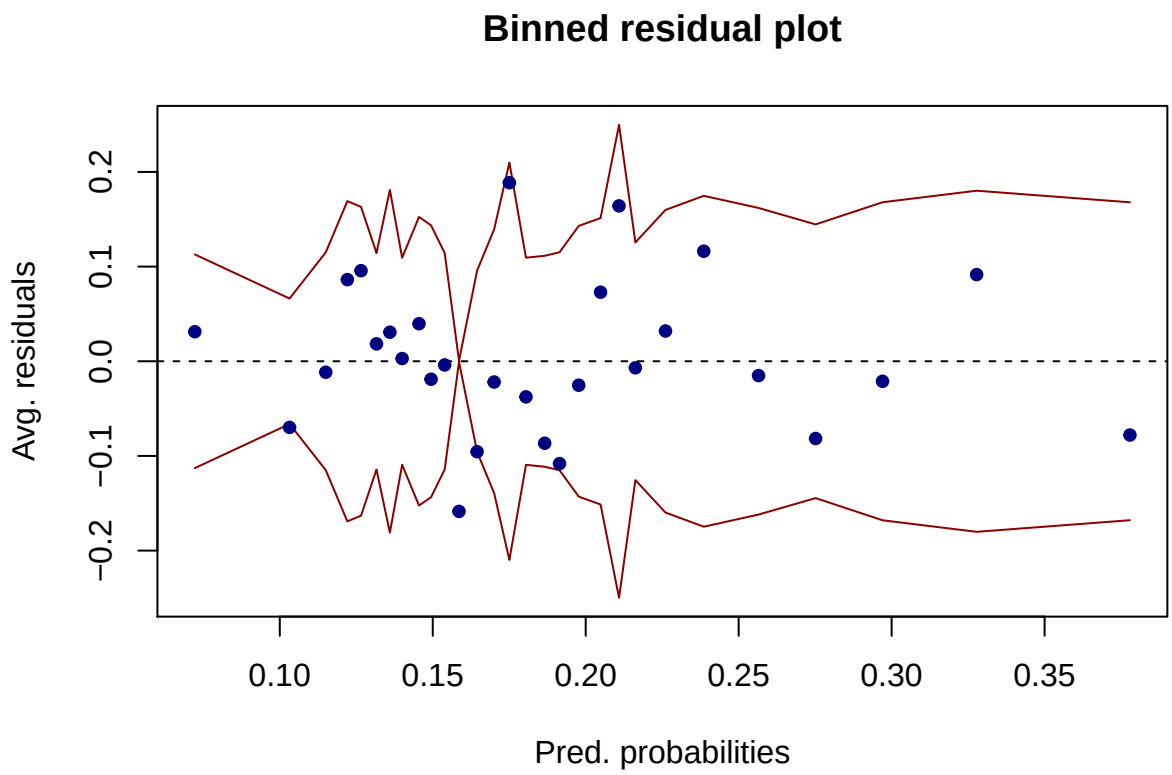
```
## Waiting for profiling to be done...
```

```
##                         2.5 %    97.5 %
## (Intercept)       0.19107244 2.1336497
## as.factor(race)6  0.58499993 4.0944394
## as.factor(race)7  1.40217975 3.2191251
## as.factor(race)8  1.00467474 4.7994271
## as.factor(race)9  0.02287415 2.1436291
## mpregwt           0.97982445 0.9981617
## as.factor(smoke)1 1.01138463 2.0422081
```

```
# Model Validation Code
# Testing to see if we have any residuals
rawresid1 <- residuals(final_model_premature,"resp")

#binned residual plot
binnedplot(x=fitted(final_model_premature),y=rawresid1,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")

#binned residual plot
binnedplot(x=fitted(final_model_premature),y=rawresid1,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



```
#Accuracy/Sensitivity/Specificity
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(final_model_premature) >= mean(Smoking$premature),
                            as.factor(Smoking$premature_f),positive = "1")
Conf_mat$table
```

```
##           Reference
## Prediction   0   1
##          0 439  70
##          1 266  94
```

```
Conf_mat$overall["Accuracy"];
```

```
##  Accuracy
## 0.6133487
```

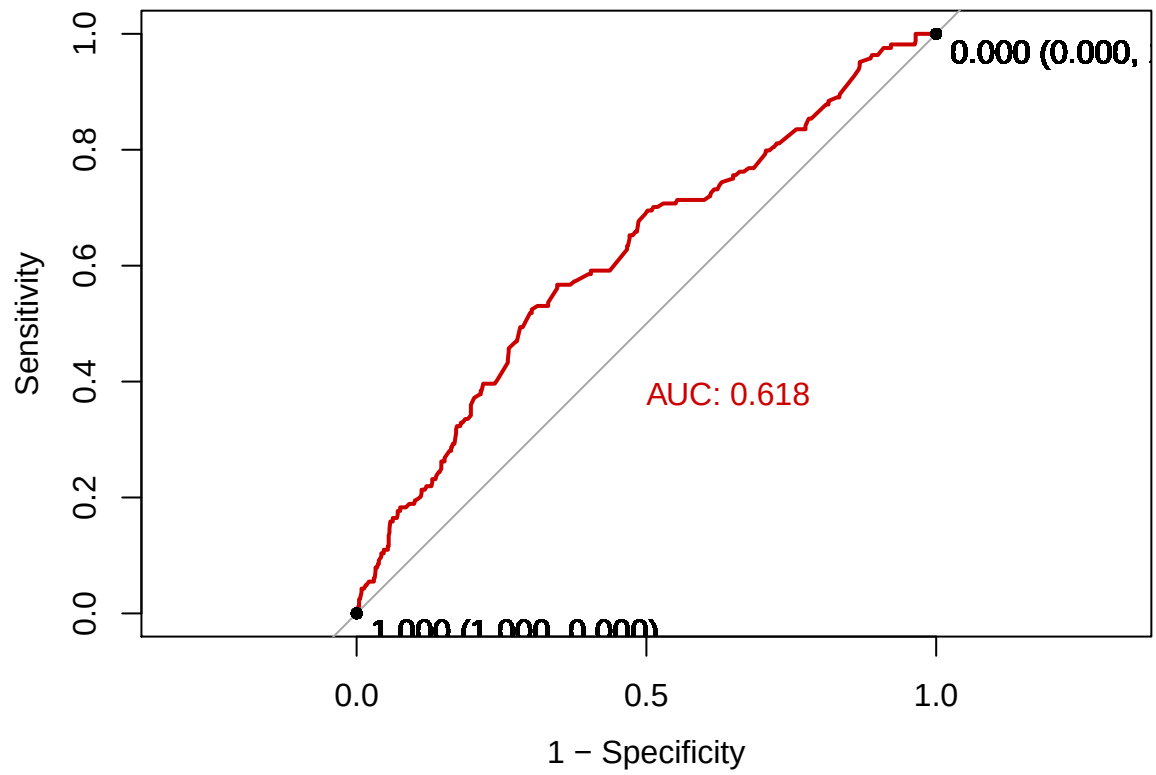
```
Conf_mat$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.5731707   0.6226950
```

```
# Area Under the Curve
roc(Smoking$premature_f,fitted(final_model_premature),plot=T,print.thres=Smoking$premature,
    legacy.axes=T, print.auc = T, print.auc.y = .4, col="red3")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



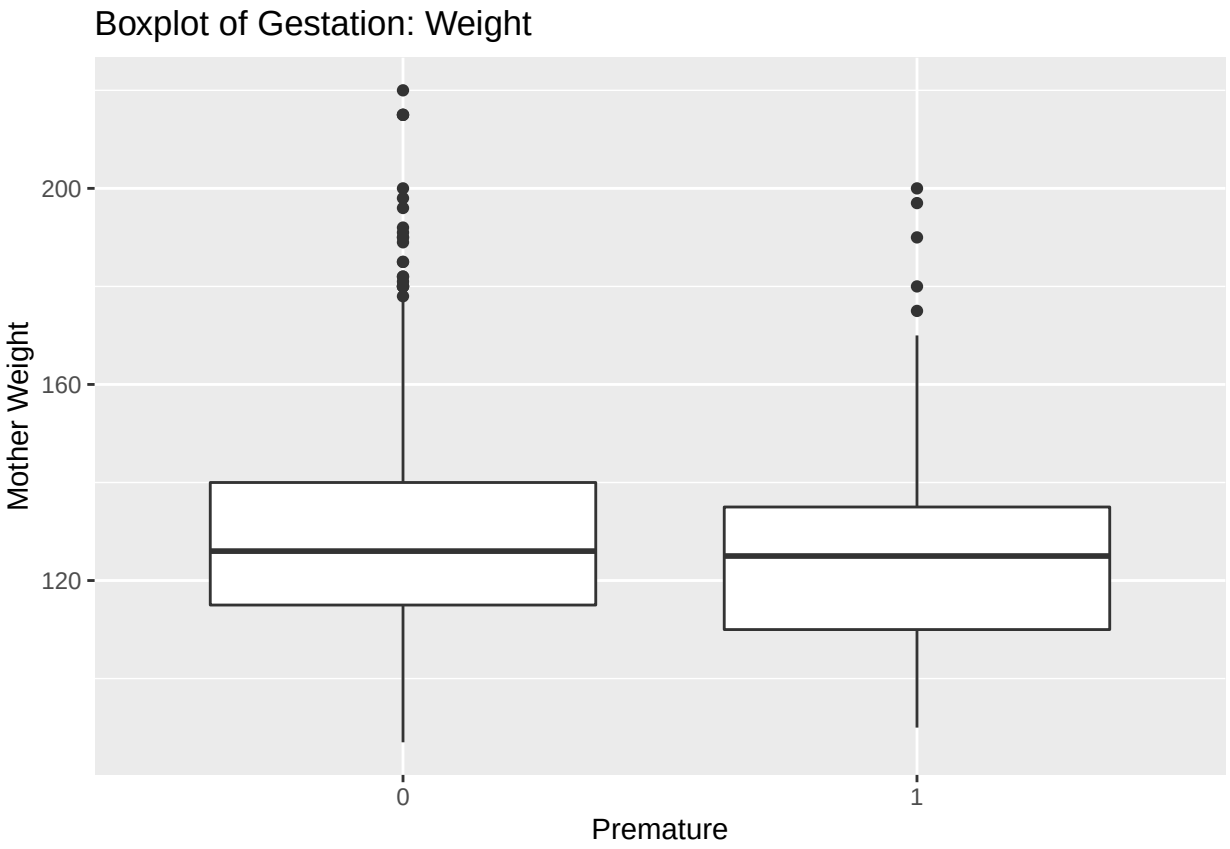

```
##
```

```
## Call:
## roc.default(response = Smoking$premature_f, predictor = fitted(final_model_premature),     plot = T,
##
## Data: fitted(final_model_premature) in 705 controls (Smoking$premature_f 0) < 164 cases (Smoking$prem
## Area under the curve: 0.6182
```

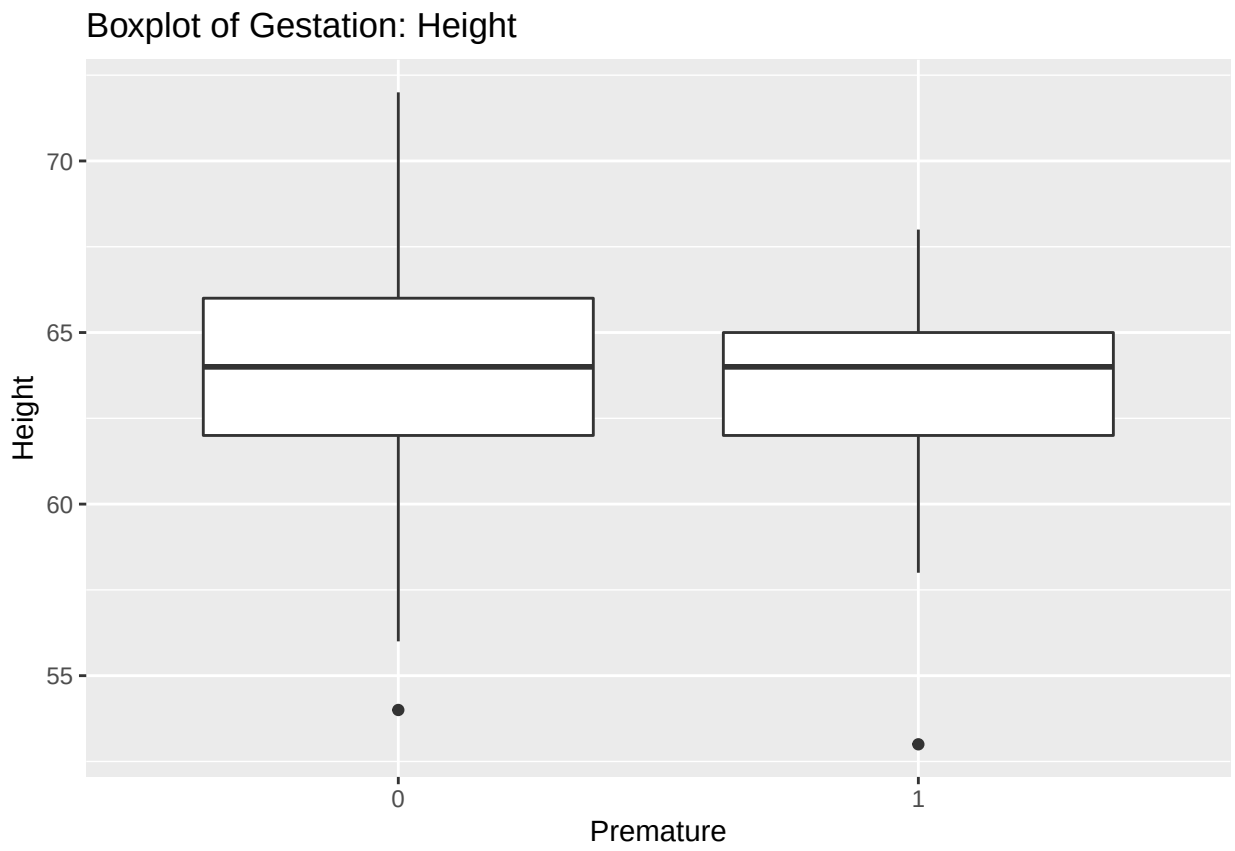
```
#Abhishek Baral, MIDS 2021 Assisted me with adjusting the mean for the AUC curve.

# EDA Tables and Plots
#Age
ggplot(Smoking, aes(x=premature_f, y=mage))+
  geom_boxplot() +
  ylab('Age') +
  xlab('Premature') + ggtitle("Boxplot of Gestation: Age")
```

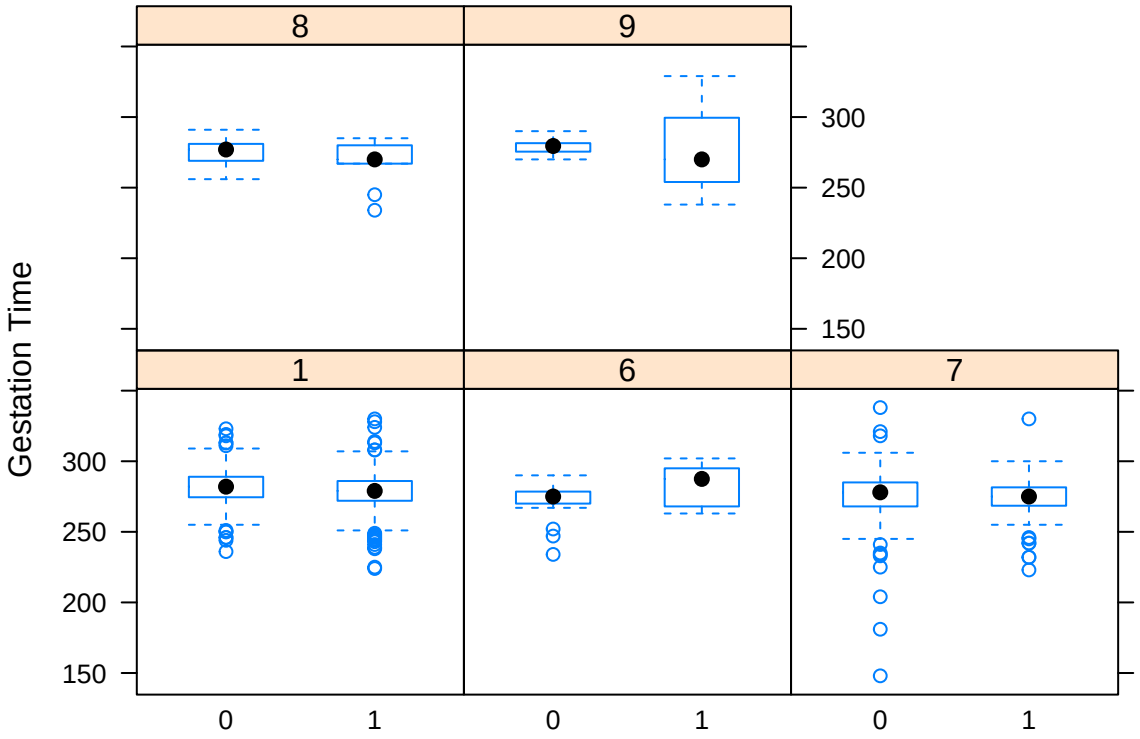

```
#Mother Weight
ggplot(Smoking, aes(x=premature_f, y=mpregwt))+
  geom_boxplot() +
  ylab('Mother Weight') +
  xlab('Premature') + ggtitle("Boxplot of Gestation: Weight")
```




```
#Mother Height
ggplot(Smoking, aes(x=premature_f, y=mht))+
  geom_boxplot() +
  ylab('Height') +
  xlab('Premature') + ggtitle("Boxplot of Gestation: Height")
```




```
#Lets do some more EDA!!!!
#Race vs Gestation (All whites are 1)
bwplot(gestation ~ as.factor(smoke)|as.factor(race), data = Smoking, ylab = 'Gestation Time')
```




```
#Smoking
chisq.test(table(Smoking[,c("smoke","premature_f")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Smoking[, c("smoke", "premature_f")])
## X-squared = 3.2971, df = 1, p-value = 0.0694
```

```
# P-value = .0694

#Race
chisq.test(table(Smoking[,c("race","premature_f")]))
```

```
## Warning in chisq.test(table(Smoking[, c("race", "premature_f")])): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Smoking[, c("race", "premature_f")])
## X-squared = 15.628, df = 4, p-value = 0.003561
```

```
# P-value = .003561

#Education
chisq.test(table(Smoking[,c("Edu","premature_f")]))
```

```
## Warning in chisq.test(table(Smoking[, c("Edu", "premature_f")])): Chi-
```

```
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Smoking[, c("Edu", "premature_f")])
## X-squared = 23.888, df = 6, p-value = 0.0005476
```

```
# P-value = 0.0005476

#Income
chisq.test(table(Smoking[,c("inc","premature_f")]))
```

```
## Warning in chisq.test(table(Smoking[, c("inc", "premature_f")])): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Smoking[, c("inc", "premature_f")])
## X-squared = 4.0413, df = 9, p-value = 0.9087
```

```
# P-value = 0.9087

#Pairity
chisq.test(table(Smoking[,c("parity","premature_f")]))
```

```
## Warning in chisq.test(table(Smoking[, c("parity", "premature_f")])): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Smoking[, c("parity", "premature_f")])
## X-squared = 24.372, df = 11, p-value = 0.01125
```

```
# P-value = 0.01125
```
