Math 189 Homework 6

Hien Bui, Cole Clisby, Cheolmin Hwang, Andrew Li, James Yu

Intro

In this report, we utilize a dataset containing information about babies/mothers to create a logistic regression model that predicts if a mother is a smoker or not. First we explore the data to determine which variables would be useful in our model to help classify an observation. We then split the data up into training and test sets and perform logistic regression.

Body

Data

```
babies <- read.table('babies.dat', header=TRUE)
head(babies)
```

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

```
nrow(babies)
```

```
## [1] 1236
```

Source: The Child Health and Development Studies (CHDS) data are presented in Stat Labs: Mathematical Statistics Through Applications by Deborah Nolan and Terry Speed (Springer).

URL: https://github.com/tuckermcelroy/ma189/blob/main/Data/babies.dat (https://github.com/tuckermcelroy/ma189/blob/main/Data/babies.dat)

Description: This data set consists of multiple data factors of babies at birth, including (but not limited to) weight and duration of pregnancy. It also includes data about the birth mother, including (but not limited to) age at conception, height, and weight. There are 1236 observations in the dataset.

1. Baby’s Birth Weight (bwt): in nearest ounce
2. gestation: duration of pregnancy in days
3. parity: 1 if baby is first born, 0 otherwise
4. Mother’s Age (age): mother’s age at conception in years
5. Mother’s Height (height): in inches
6. Mother’s Weight (weight): prepregnancy weight in pounds

7. Smoking Indicator (smoke): 1 if mother smokes, 0 if not, 9 if unknown

```r
# There are 10 rows that have unknown for smoke (value 9)
nrow(babies[babies$smoke == 9,])
```

```
## [1] 10
```

```r
# Drop the '9' rows from table
babies <- subset(babies, smoke != 9)

# Removing outliers
babies <- subset(babies, age != 99)
babies <- subset(babies, gestation != 999)
babies <- subset(babies, height != 99)
babies <- subset(babies, weight != 999)
nrow(babies)
```

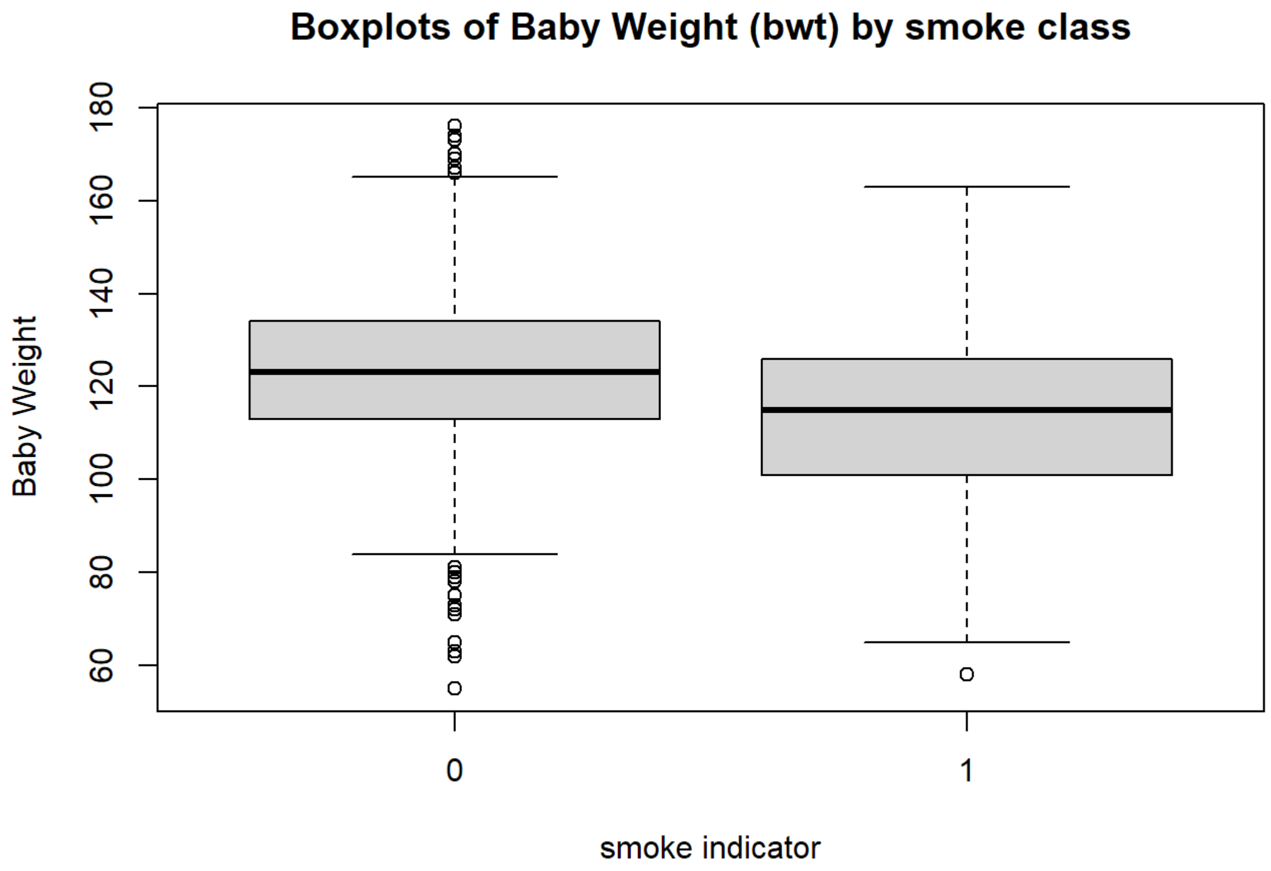
```
## [1] 1174
```

After removing outliers/unknown values we now have 1174 observations. We remove these values since they will impact our visualizations and model. It seems as though any unknown values were filled with 9, 99, or 999, relative to the scale of the variable. The variables are clearly unknown and filled with the value on purpose, like age being 99, weight being 999, and gestation being 999, since those are very unlikely and physically improbable. A mother's height being 99 inches is also improbable, since a quick google search shows that the tallest woman in history was recorded as being roughly 90 inches tall.

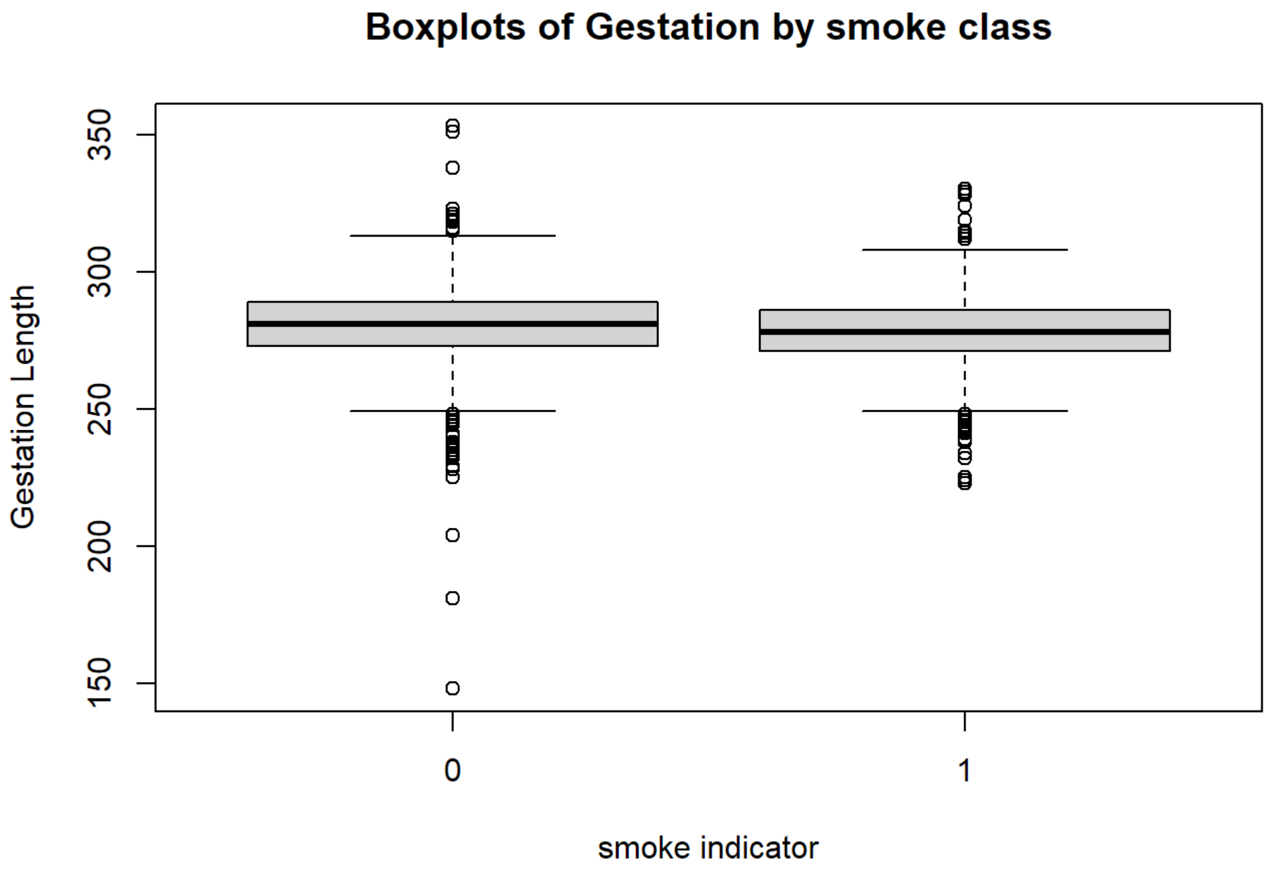
Regardless, we decided to remove these observations since they may impact our logistic regression model due to their values being much greater than 3rd quartile of each variable, which means it could skew the prediction for a specific group. This is the safest choice since we do not know the reason as to why these variables were filled with these values. It could be that it is indicating something important, but we do not have a way of knowing what that is.

Task 1

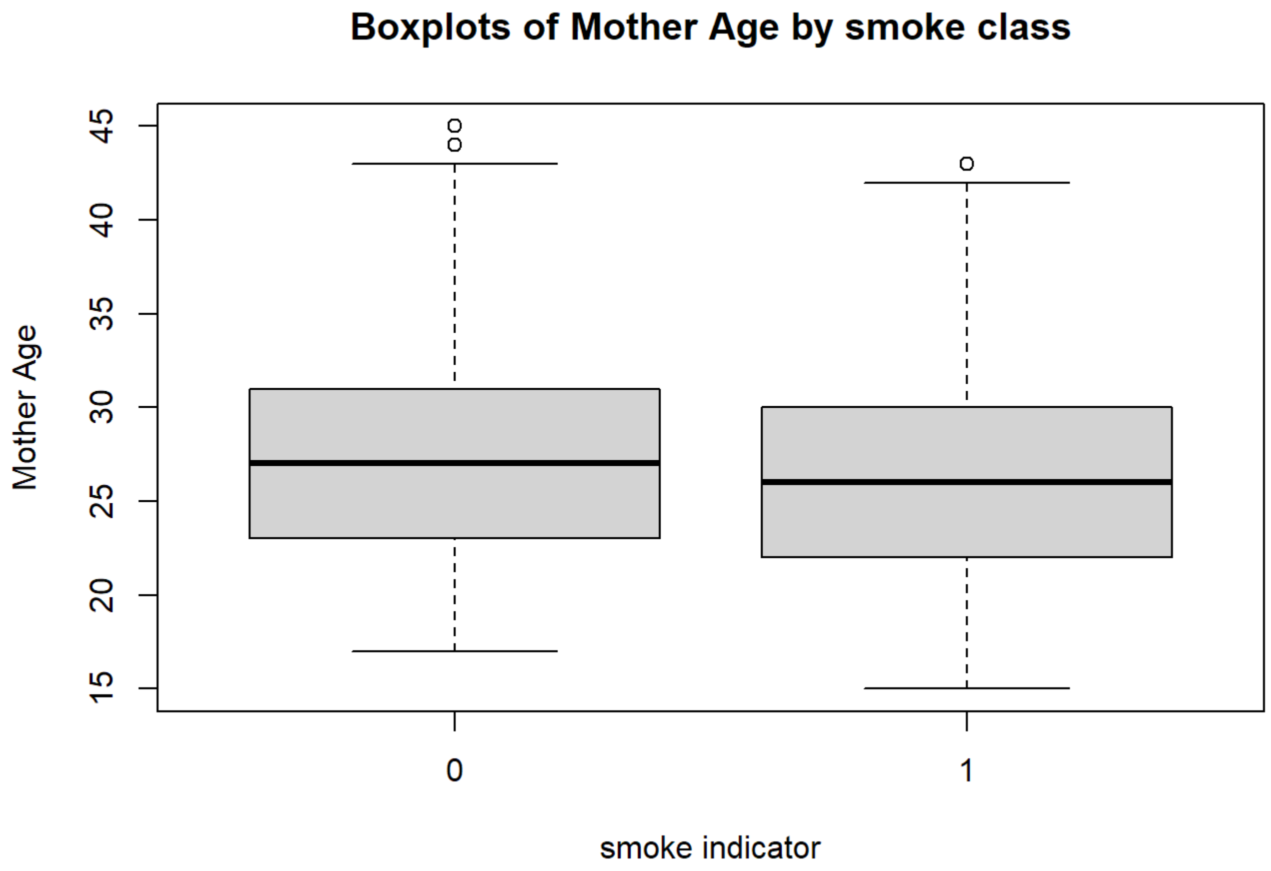
```r
boxplot(babies$bwt ~ babies$smoke, data=babies, xlab='smoke indicator', ylab='Baby Weight')
title('Boxplots of Baby Weight (bwt) by smoke class')
```



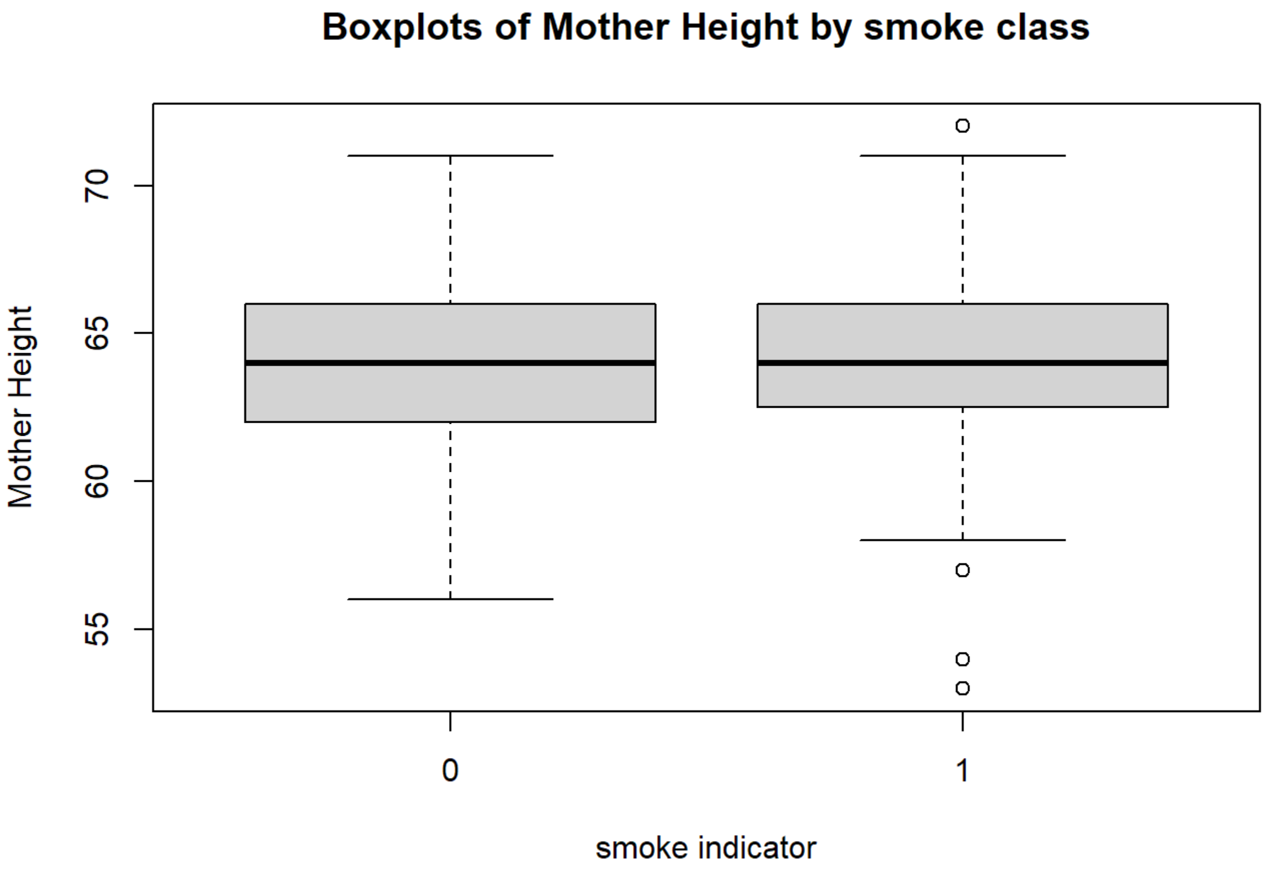
```
boxplot(babies$gestation ~ babies$smoke, data=babies, xlab='smoke indicator', ylab='Gestation Length')
title('Boxplots of Gestation by smoke class')
```



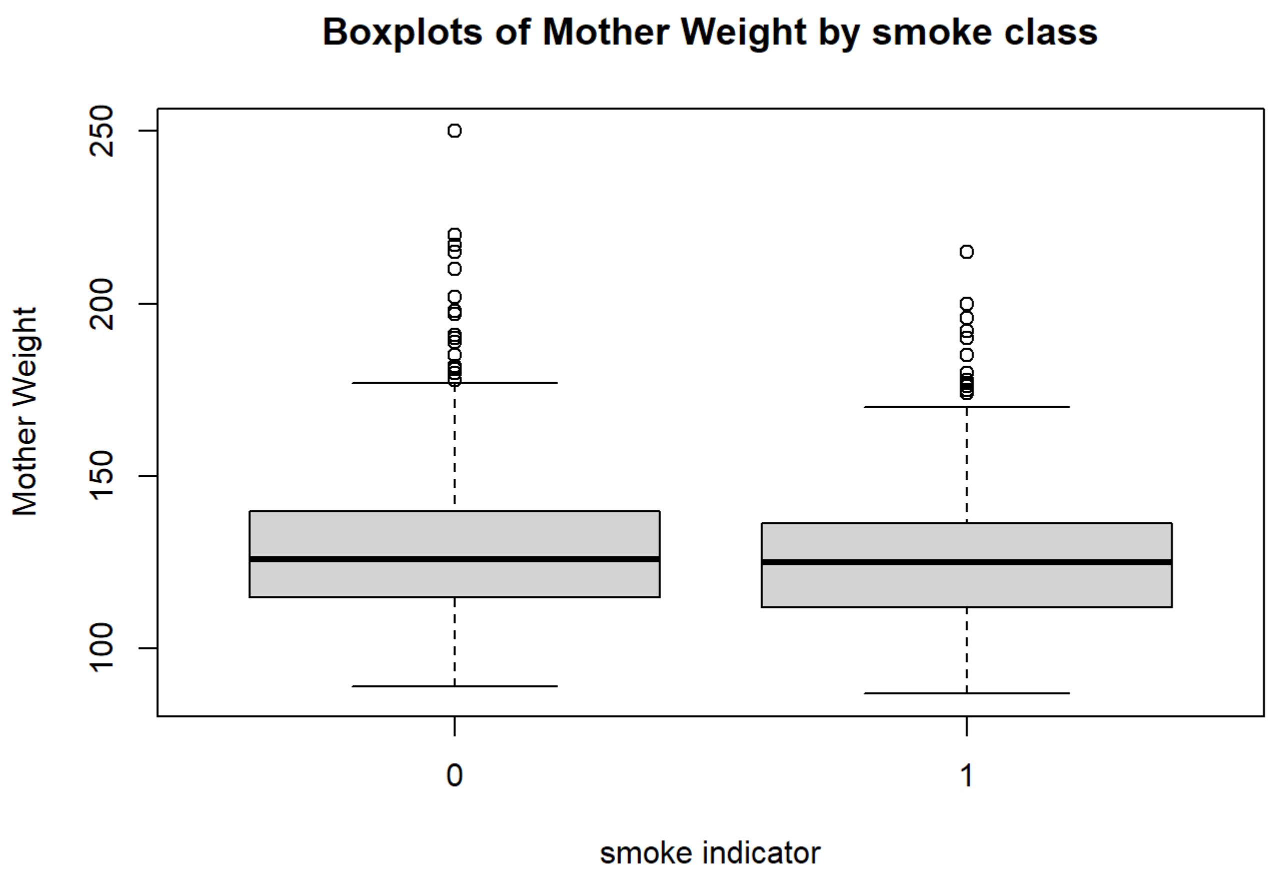
```
boxplot(babies$age ~ babies$smoke, data=babies, xlab='smoke indicator', ylab='Mother Age')
title('Boxplots of Mother Age by smoke class')
```



```
boxplot(babies$height ~ babies$smoke, data=babies, xlab='smoke indicator', ylab='Mother Height')
title('Boxplots of Mother Height by smoke class')
```

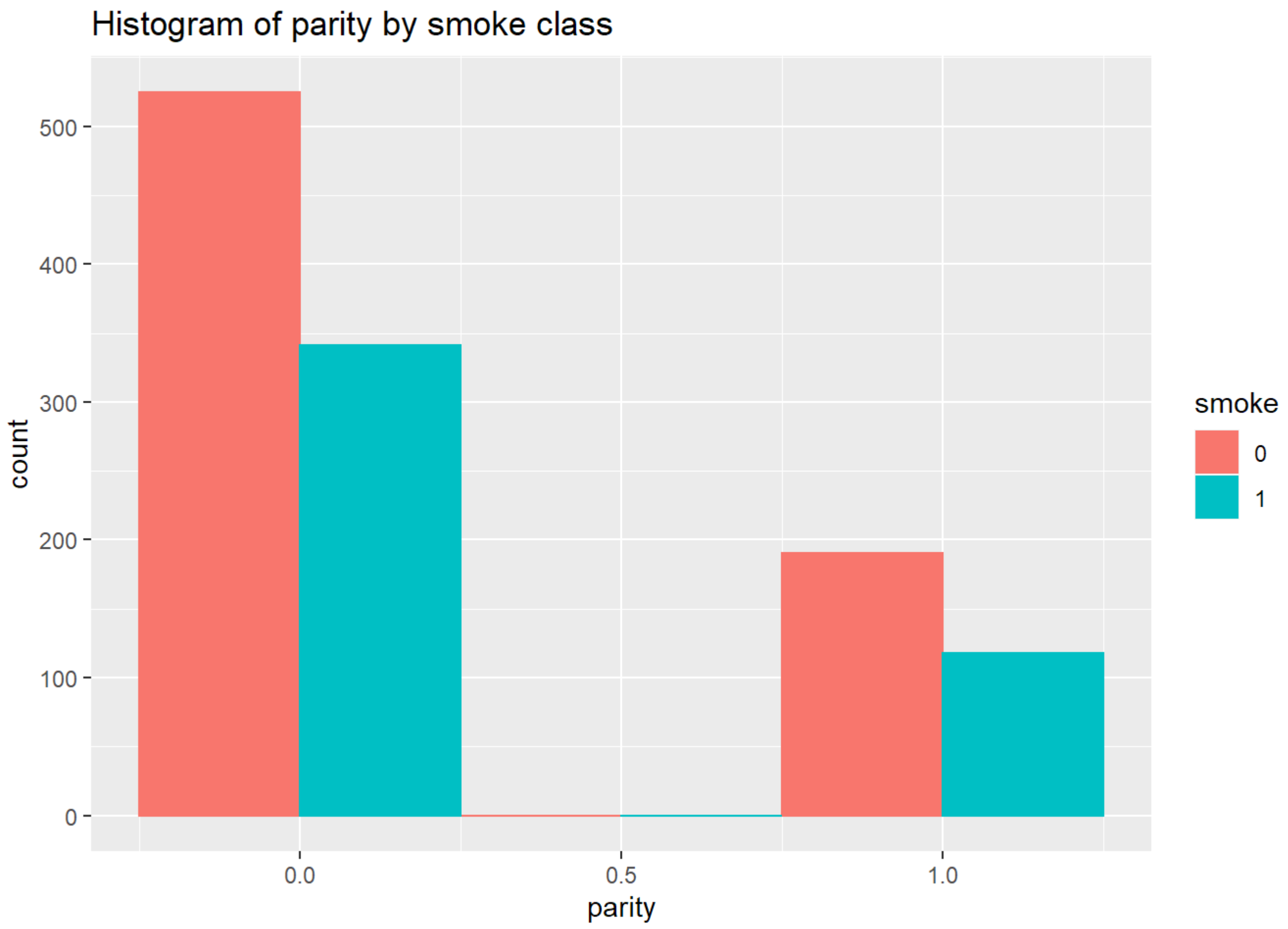


```
boxplot(babies$weight ~ babies$smoke, data=babies, xlab='smoke indicator', ylab='Mother Weight')
title('Boxplots of Mother Weight by smoke class')
```



The boxplots for the two groups are very similar across the 5 numeric variables, with baby weight (bwt) looking to have the most difference between groups, where smoke=0 has a higher median than smoke=1.

```
library(ggplot2)
temp_babies <- babies
temp_babies$smoke <- as.factor(temp_babies$smoke)
ggplot(temp_babies, aes(x=parity, fill=smoke, color=smoke)) + geom_histogram(position='dodge', b
ins=3) + ggtitle('Histogram of parity by smoke class')
```



```
parity_0 <- c()
parity_0 <- append(parity_0, dim(babies[babies$parity == 0 & babies$smoke ==0,])[1])
parity_0 <- append(parity_0, dim(babies[babies$parity == 0 & babies$smoke ==1,])[1])

parity_1 <- c()
parity_1 <- append(parity_1, dim(babies[babies$parity == 1 & babies$smoke ==0,])[1])
parity_1 <- append(parity_1, dim(babies[babies$parity == 1 & babies$smoke ==1,])[1])

parity_smoke <- rbind(parity_0, parity_1)
colnames(parity_smoke) <- c('Smoke=0', 'Smoke=1')
rownames(parity_smoke) <- c('Parity=0', 'Parity=1')
parity_smoke
```

```
##          Smoke=0 Smoke=1
## Parity=0     525     341
## Parity=1     190     118
```

```
smoke_0_parity_0_prop <- 525 / (525 + 190)
smoke_1_parity_0_prop <- 341 / (341 + 118)
smoke_0_parity_0_prop
```

```
## [1] 0.7342657
```

```
smoke_1_parity_0_prop
```

```
## [1] 0.7429194
```

Plotting a histogram of parity by groups and calculating the proportions directly shows that the two groups have similar distributions for parity.

Ultimately, we see that many of the variables have similar distributions across the two groups of smoke. So the visualizations were not very helpful in determining which features would be useful to predict smoker status.

Task 2

```
babies.train <- babies[sample(nrow(babies), 880), ]
babies.test <- babies[sample(nrow(babies), 294), ]
```

We chose to split the data into about 75% (880 entries) training and 25% (294 entries) test set using random sampling to give the model enough data to effectively predict the test set. We use a 75-25 split because we want to use more of the data to help train the model, and use a smaller proportion to test our well our model performs, after we train it. We use random sampling since it should give us a representative sample of the dataset, since each observation has an equal chance of being included in the sample.

Task 3

```
# perform logistic regression on training
y_train <- babies.train$smoke
X_weight <- babies.train$weight
X_height <- babies.train$height
X_parity <- babies.train$parity
X_bwt <- babies.train$bwt
X_gestation <- babies.train$gestation
X_age <- babies.train$age
all.fit <- glm(y_train ~ X_weight+X_height+X_parity+X_bwt+X_gestation+X_age, data=babies.train,
 family=binomial)
summary(all.fit)
```

```
##
## Call:
## glm(formula = y_train ~ X_weight + X_height + X_parity + X_bwt +
##     X_gestation + X_age, family = binomial, data = babies.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7052  -0.9823  -0.7654   1.1987   2.2919
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.373619   2.253972  -1.053  0.29230
## X_weight    -0.008200   0.004074  -2.013  0.04413 *
## X_height     0.092288   0.033139   2.785  0.00535 **
## X_parity    -0.216954   0.175420  -1.237  0.21617
## X_bwt       -0.031321   0.004673  -6.703 2.04e-11 ***
## X_gestation  0.005628   0.004879   1.153  0.24873
## X_age       -0.027542   0.013403  -2.055  0.03990 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1177.7  on 879  degrees of freedom
## Residual deviance: 1114.2  on 873  degrees of freedom
## AIC: 1128.2
##
## Number of Fisher Scoring iterations: 4
```

The p-values can vary from run-to-run since our training set is a random sample, but with a significance level of 0.05, we saw that the p values for height, bwt, and age were generally the most significant predictor variables for the logistic regression model. The p value represents the outcome of the test that the given variable is associated with the outcome variable of smoke. The null hypothesis claims that the given variable is not correlated with the outcome variable, and the alternative claims that the given variable is correlated with it. So if the p-value is less than the significance level, we can reject the null and say that the variable is a significant predictor variable.

```
sig.fit <- glm(y_train ~ X_weight+X_height+X_bwt+X_age, data=babies.train, family=binomial)
summary(sig.fit)
```

```
## 
## Call:
## glm(formula = y_train ~ X_weight + X_height + X_bwt + X_age, 
##     family = binomial, data = babies.train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7365  -0.9848  -0.7654   1.2038   2.2213  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.119948   1.922279  -0.583  0.56015    
## X_weight    -0.007840   0.004056  -1.933  0.05322 .  
## X_height     0.089575   0.032970   2.717  0.00659 ** 
## X_bwt       -0.029058   0.004273  -6.800 1.05e-11 ***
## X_age       -0.023105   0.012579  -1.837  0.06624 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1177.7  on 879  degrees of freedom
## Residual deviance: 1116.8  on 875  degrees of freedom
## AIC: 1126.8
## 
## Number of Fisher Scoring iterations: 4
```

We can use the coefficients from our model to make a prediction function that will return the probablity that a given observation is the group of smoke=1. If the probability is greater than or equal to 0.5, we will classify the observation as smoke=1. Otherwise, we will classify it as smoke=0.

```
# Prediction function
pred_all <- function(obs){
  x <- c(1, obs)
  pred <- as.numeric(as.numeric(x %*% sig.fit$coefficients))
  pred <- 1 / (1 + exp(-pred))
  return (pred)
}
```

Task 4

```r
prediction <- c()

# use model (prediction function?) on test data
for (i in 1:nrow(babies.test)) {
  obs <- (babies.test[i,])
  pred <- pred_all(c(obs$weight,obs$height,obs$bwt,obs$age))
  if (pred >= 0.5) {
    prediction <- append(prediction, 1)
  } else {
    prediction <- append(prediction, 0)
  }
}

babies.test$prediction <- prediction
group0_true <- sum((babies.test$smoke == 0) & (babies.test$prediction == 0))
group1_true <- sum((babies.test$smoke == 1) & (babies.test$prediction == 1))

class_tab <- c(sum(babies.test$smoke==0), sum(babies.test$smoke==1))
class_tab <- rbind(class_tab, c(group0_true, group1_true))
class_tab <- rbind(class_tab, class_tab[1,] - class_tab[2,])
colnames(class_tab) <- c('Smoke 0', 'Smoke 1')
rownames(class_tab) <- c('Num Observations', 'Num Correct', 'Num Wrong')
class_tab
```

```
##                  Smoke 0 Smoke 1
## Num Observations     190     104
## Num Correct          171      37
## Num Wrong             19      67
```

```r
acc <- (group0_true + group1_true) / nrow(babies.test)
acc
```

```
## [1] 0.707483
```

Accuracy is around 0.65-0.7 (variance due to training and test sets being random samples) with our Logistic Regression model, using weight, height, and bwt as predictor variables. If we were to randomly guess the class each time, we would get roughly 0.50 accuracy, so our model is better than chance, although it definitely could be better.

Our model performs much worse for class 1 compared to class 0. This could be because the dataset was slightly skewed towards class 0, with more observations of smoke=0 (715 vs 459). So when we created our training set, our model would have more information on class 0 and may have been biased towards it. This is more of an issue with the dataset as if we took an equal number of samples from each class, it might be unrepresentative of the general population, and the model may not generalize.

Task 5 / Conclusion

Based on our analysis and predictive model, there is a strong correlation between mothers smoking and a decreased birth weight in their babies. However, this analysis cannot be used to claim that smoking causes reduced birth weight. This predictive model can only be used on the surface level to predict whether a mother smokes, and cannot be used as evidence to support causation.

Our predictive model takes in information on a mother’s weight, height, and age, along with the baby’s birth weight to predict if the mother is a smoker or not. This could be useful if we had information about the mother and her baby, but not if she was a smoker or not. A model like this could be used to determine if the mother was a smoker and potentially be used for future analysis with other data. For instance, we could use it to analyze the impacts of smoking on a baby’s future height and weight, relative to the mother.

It is possible to predict from variables that are not causing a phenomenon if the variables in question follow a consistent, representative trend or pattern in association with the phenomenon. If analysis of the variable and the phenomenon is conducted while controlling all other variables, and if the analysis results in a distinct pattern, we can use the variable to predict a phenomenon. However, it is important to note that this model is merely predictive, and not definitive.