Math 189 Homework 6: Linking Baby Features to Smoking

Yunchun Pan, Siqing Lyu, Nathan Ng, Pudan Xu

2/23/2021

Introduction

In this assignment, we examine the relationship between characteristics of babies and their mothers and their smoking status, using the Baby data set **babies.dat**[1]. We develop a logistic regression model to predict whether the mother is a smoker or a non-smoker based on the baby's and the mother's characteristics that are most likely to be associated with the mother's smoking status. The Baby dataset has 1236 observations of babies and each observation has the following variables:

- **bwt**: Baby's weight at birth, to the nearest ounce
- **gestation**: Duration of the pregnancy in days, calculated from the first day of the last normal menstrual period
- **parity**: Indicator for whether the baby is the first born (1) or not (0)
- **age**: Mother's age at the time of conception, in years
- **height**: Height of the mother, in inches
- **weight**: Mother's prepregnancy weight, in pounds
- **smoke**: Smoking indicator for whether the mother smokes (1) or not (0); (9) denotes unknown

In order to determine whether a mother is smoker or not, we first explore the data using boxplots to visually investigate the association between the mothers' smoking status and the other characteristic variables in the baby dataset. Using the variables that are strongly associated with the smoking indicator, we split the data into a training and a testing set and fit a logistic regression on the training set. Afterwards, we estimate the coefficients of fitted model and use them to predict the probabilities of babies having smoking mothers for the testing set. In the end, we discuss the results based on the computed probabilities and some possible applications of our model.

Our Work

We first load the Baby data set into R to get the data we use throughout this assignment and save it into **baby**. There are 1236 observations in **baby**, but 10 observations whose mothers' smoking status are unknown. Because we want to build a model that only predicts whether the mother is a smoker or not, we remove those 10 observations and display the top 6 rows of the updated **baby**.

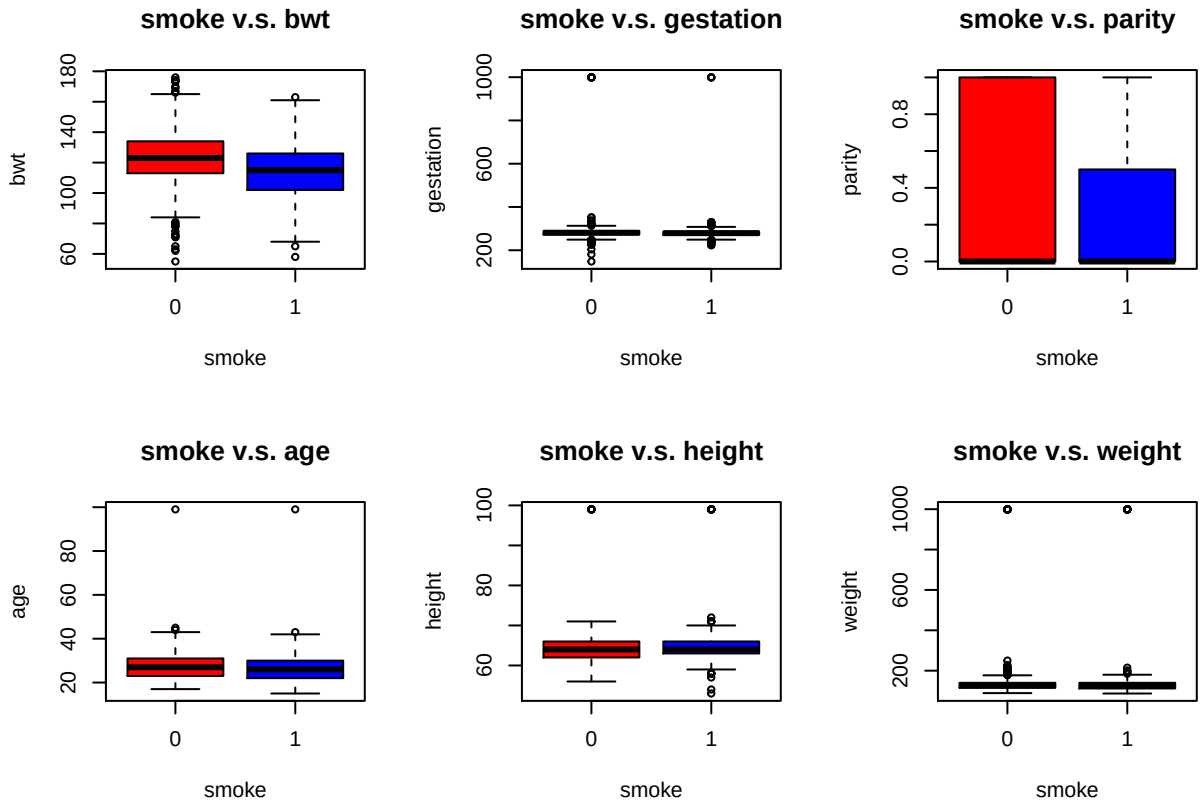
```
baby <- read.csv("../Data/babies.dat", sep="")
baby <- baby[baby$smoke!=9,]
head(baby)
```

[1]Source: This dataset is found from http://www.stat.berkeley.edu/users/statlabs/labs.html. It accompanies the excellent text Stat Labs: Mathematical Statistics through Applications Springer-Verlag (2001) by Deborah Nolan and Terry Speed.

```
##   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
```

Here, we create boxplots for six variables, which are **bwt**, **gestation**, **parity**, **age**, **height**, and **weight**, versus the binary variable **smoke** to visually investigate the association between **smoke** and the other features.

```
par(mfrow=c(2, 3))
boxplot(bwt ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. bwt')
boxplot(gestation ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. gestation')
boxplot(parity ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. parity')
boxplot(age ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. age')
boxplot(height ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. height')
boxplot(weight ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. weight')
```

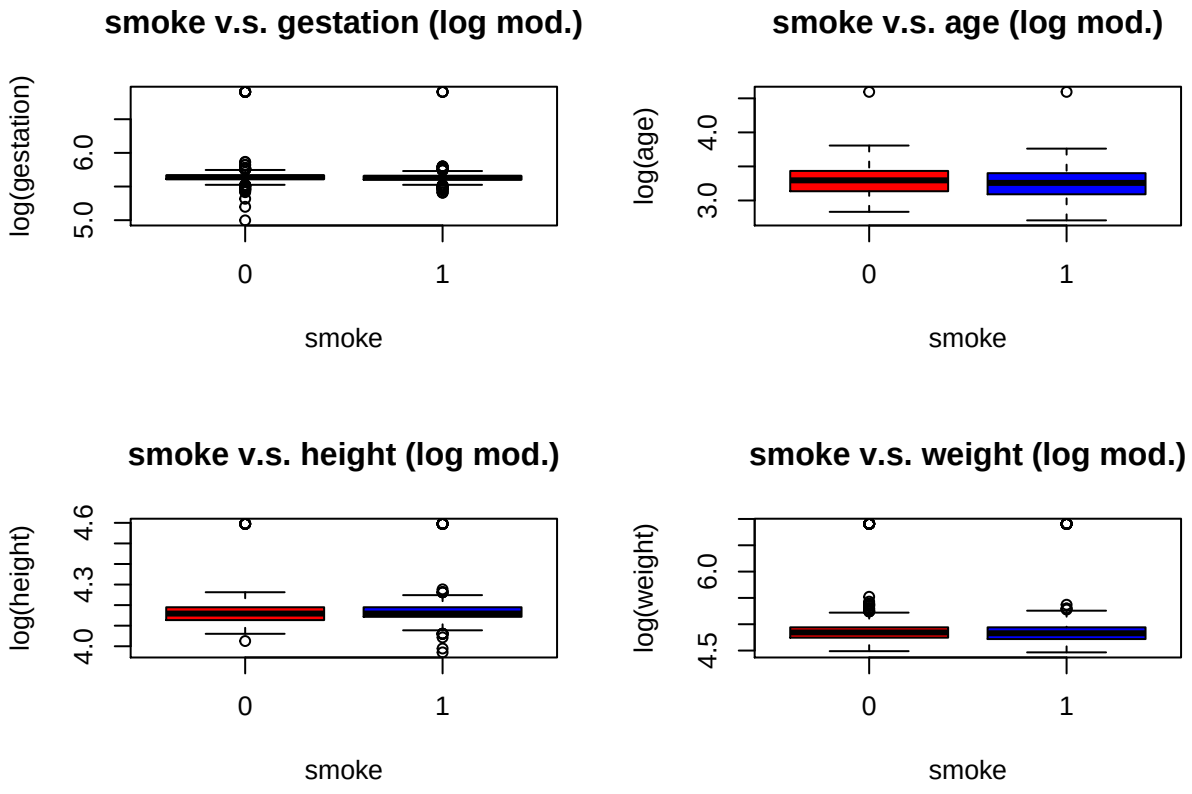


By observing the above six boxplots, we find that the distributions of birth weights of babies given birth by

smoking and non-smoking mothers are different, so it's reasonable to believe **bwt** is associated with **smoke**. Since **parity** is a categorical variable, we decide to perform chi-square test of independence to evaluate if there is an association between the categories of **parity** and **smoke**. Moreover, because there are certain amount of outliers in boxplots of all other variables, we decide to apply some transformations to their data and plot new boxplots to further explore their associations with **smoke**.

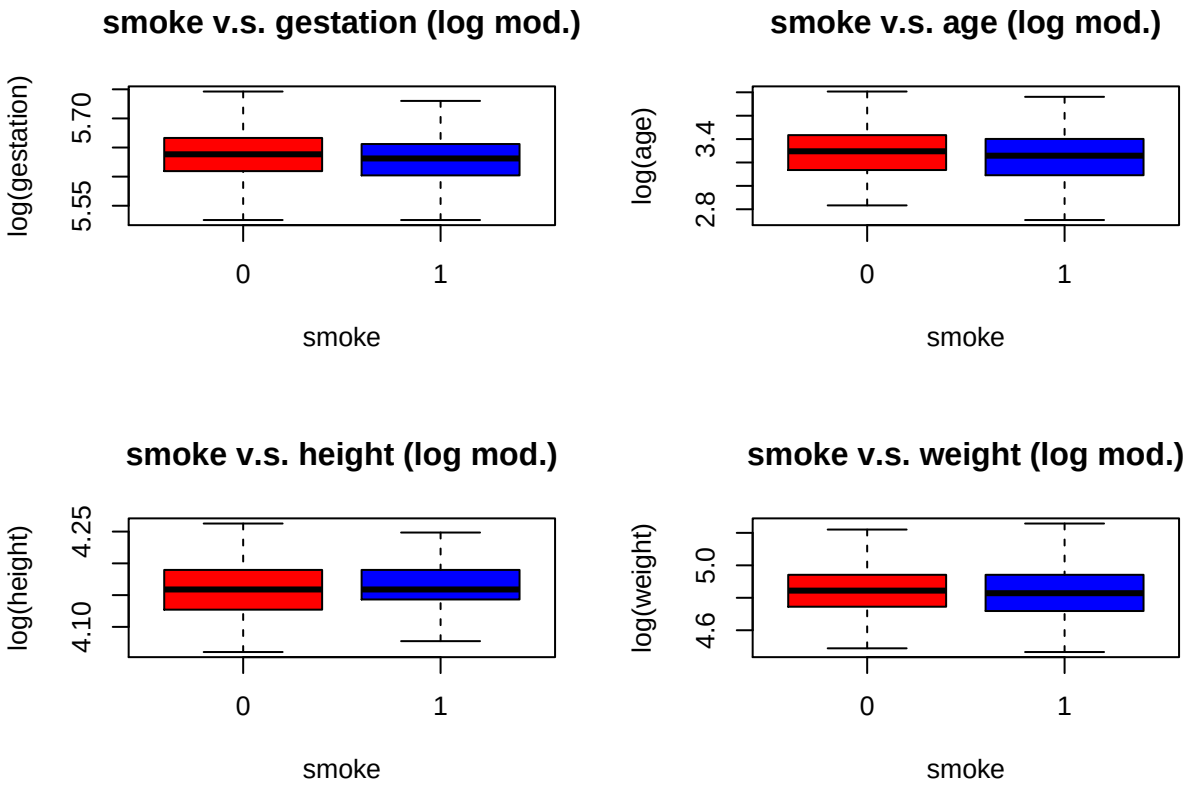
Here, we apply log transformation to those variables whose data have some outliers, which are **gestation**, **age**, **height**, and **weight**. Then, we make new boxplots for these transformed variables versus the binary variable **smoke**.

```
par(mfrow=c(2, 2))
boxplot(log(gestation) ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. gestation (log mod.)')
boxplot(log(age) ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. age (log mod.)')
boxplot(log(height) ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. height (log mod.)')
boxplot(log(weight) ~ smoke, data=baby,
        col=c("red","blue"), main='smoke v.s. weight (log mod.)')
```



As we can see from the above four boxplots, it's not very obvious that the distribution of these variables are different for **smoke == 0** and **smoke == 1** due to outliers, so we decide to remove outliers from them and check if boxplots with outliers removed would better help us examine the association between these variables and **smoke**.

```
par(mfrow=c(2, 2))
boxplot(log(gestation) ~ smoke, data=baby, outline =  FALSE,
        col=c("red","blue"), main='smoke v.s. gestation (log mod.)')
boxplot(log(age) ~ smoke, data=baby, outline =  FALSE,
        col=c("red","blue"), main='smoke v.s. age (log mod.)')
boxplot(log(height) ~ smoke, data=baby, outline =  FALSE,
        col=c("red","blue"), main='smoke v.s. height (log mod.)')
boxplot(log(weight) ~ smoke, data=baby, outline =  FALSE,
        col=c("red","blue"), main='smoke v.s. weight (log mod.)')
```



As we can see from the above boxplots with outliers removed, it is still not evident that the distribution of these variables are different for **smoke** == **0** and **smoke** == **1**, so we decide to include all of them as predictors when building our first logistic regression, and then check if there is an association between them and **smoke** by comparing their p-values computed by logistic regression model with the significance level 0.05.

Here, we perform a chi-square test of independence to test whether there is a an association between the categories of **parity** and **smoke**. The null hypothesis is that the categories of **parity** and **smoke** are independent, while the alternative hypothesis is that there is an association between the categories of **parity** and **smoke**. At the significance level of 0.05, if the p value is larger than or equal to 0.05, we fail to reject the null hypothesis that these two categorical variables are independent and hence **parity** is not useful for predicting **smoke**. Otherwise, we reject the null hypothesis and **parity** is helpful for predicting **smoke**.

```
cat("The significance level:", 0.05)
```

```
## The significance level: 0.05
```

```r
chisq.test(baby$smoke, baby$parity)['p.value']
```

```
## $p.value
## [1] 0.7025691
```

Since the p value of this test is 0.7025691 and it is much larger than 0.05, we fail to reject the null hypothesis that the categorical variables **parity** and **smoke** are independent, and there are clear evidences for us to conclude that there is no association between the categories of **parity** and **smoke**. Hence, **parity** is not helpful for predicting **smoke**.

To predict **smoke** using associated variables, we split the Baby dataset into a training and testing set and build a logistic regression model afterwards. The training set contains 80% observations of baby_0 (**smoke == 0**) and 80% data in baby_1 (**smoke == 1**), and the remaining 20% of data in baby_0 and baby_1 are combined into a testing set. We set the training size to be 80% because we want to provide sufficient information for the logistic regression model to estimate related coefficients before we predict the probability of a baby having a smoking mother.

```r
baby_1 <- baby[baby$smoke == 1,]
baby_0 <- baby[baby$smoke == 0,]
paste("Number of mothers that smoke:", dim(baby_1)[1])
```

```
## [1] "Number of mothers that smoke: 484"
```

```r
paste("Number of mothers that do not smoke:", dim(baby_0)[1])
```

```
## [1] "Number of mothers that do not smoke: 742"
```

```r
train_size <- 0.8

baby_1_train_size <-floor(dim(baby_1)[1]*train_size)
baby_0_train_size <-floor(dim(baby_0)[1]*train_size)
baby_train <- rbind(baby_1[1:baby_1_train_size,],
                    baby_0[1:baby_1_train_size,])
baby_test <- rbind(baby_1[(baby_1_train_size+1):dim(baby_1)[1],],
                   baby_0[(baby_0_train_size+1):dim(baby_0)[1],])
```

After splitting the data into a training set and testing set, we build a logistic regression using variables whose relationships with **smoke** can not be confirmed due to the lack of statistical evidence. They are baby's birth weight, gestation, mother's weight, age, and height. To determine whether these variables have relationship with the smoking indicator statistically, we fit the model on the training set and observe the coefficients and the corresponding p-value of each variable. At the significance level of 0.05, a p-value being greater than or equal to 0.05 suggests that there is no relationship between the variable and **smoke**, whereas a p-value being smaller than 0.05 suggests that a relationship does exist between them.

```r
all.fit <- glm(smoke~bwt+gestation+weight+age+height,
               data=baby_train,family=binomial)
summary(all.fit)
```

```
##
## Call:
## glm(formula = smoke ~ bwt + gestation + weight + age + height,
```

```
##     family = binomial, data = baby_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.77867  -1.12966   0.00267   1.12555   1.75619
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.210e+00  1.142e+00   1.935    0.053 .
## bwt         -2.968e-02  4.598e-03  -6.455 1.08e-10 ***
## gestation   -1.148e-04  9.923e-04  -0.116    0.908
## weight       7.798e-05  5.576e-04   0.140    0.889
## age         -1.416e-02  1.139e-02  -1.243    0.214
## height       2.662e-02  1.670e-02   1.594    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1073.0  on 773  degrees of freedom
## Residual deviance: 1024.2  on 768  degrees of freedom
## AIC: 1036.2
##
## Number of Fisher Scoring iterations: 4
```

The information above is the result of the logistic regression model that is fitted on training set and built on those five variables. As we can see, only **bwt**, or the baby's birth weight, has a p-value of 1.08e-10 that is significantly less than 0.05, while all other variables included have p-values that are greater than our significance level of 0.05. This suggests that the smoking status of a mother is not dependent on variables other than **bwt**. Thus, we only include birth weight as the predictor of the probability of a baby having a smoking mother.

```
all.fit <- glm(smoke~bwt,
               data=baby_train,family=binomial)
summary(all.fit)
```

```
##
## Call:
## glm(formula = smoke ~ bwt, family = binomial, data = baby_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.78541  -1.13058  -0.01597   1.12848   1.73228
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.384721   0.544649   6.214 5.15e-10 ***
## bwt         -0.028422   0.004529  -6.275 3.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1073.0  on 773  degrees of freedom
## Residual deviance: 1030.1  on 772  degrees of freedom
## AIC: 1034.1
##
## Number of Fisher Scoring iterations: 4
```

The information above is the result of the logistic regression model that is fitted on training set and built on **bwt**. The coefficient corresponding to the birth weight is equal to -0.028422, and the intercept is 3.384721. Its p-value is 3.50e-10 and less than 0.05, so we reject the hypothesis that there is no relationship between **bwt** and **smoke**.

After fitting the model, we create the following function that uses the intercept and coefficient calculated above to compute the probabilities of babies having smoking mothers for our testing dataset. Again, we only use the baby's birth weight as the predictor because it is the only variable that has association with the smoking status of the mother.

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

Then, we use the function defined above to calculate the probability that the mother is a smoker for each observation in the testing dataset. If the probability that the mother is a smoker is greater than 0.5, then we predict that the mother is a smoker. However, if the probability is less than or equal to 0.5, then we predict that the mother is not a smoker. To sum up, our prediction is based on choosing which event, whether the mother is a smoker or a non-smoker, has a higher probability of happening given the prediction results.

```
comp = c(1:dim(baby_test)[1])

for (i in 1:dim(baby_test)[1]) {
  comp[i] = pred_all(baby_test[i, c('bwt')])
}
result <- data.frame("predict prob" = comp,
                     "predict result" = as.numeric(comp > 0.5),
                     "true result" = baby_test$smoke)
result
```

```
##     predict.prob predict.result true.result
## 1      0.4864234              0           1
## 2      0.4580968              0           1
## 3      0.6324028              1           1
## 4      0.5148375              1           1
## 5      0.4793264              0           1
## 6      0.5077358              1           1
## 7      0.3754118              0           1
## 8      0.4935259              0           1
## 9      0.5781452              1           1
## 10     0.4161647              0           1
## 11     0.5987908              1           1
## 12     0.7627505              1           1
## 13     0.5572228              1           1
## 14     0.7778333              1           1
```

```
## 15     0.6710791              1           1
## 16     0.3237823              0           1
## 17     0.5711984              1           1
## 18     0.5290202              1           1
## 19     0.4935259              0           1
## 20     0.6584121              1           1
## 21     0.6190917              1           1
## 22     0.6896182              1           1
## 23     0.2648985              0           1
## 24     0.5711984              1           1
## 25     0.7523108              1           1
## 26     0.6710791              1           1
## 27     0.3054010              0           1
## 28     0.4722377              0           1
## 29     0.8230722              1           1
## 30     0.5781452              1           1
## 31     0.6123671              1           1
## 32     0.5987908              1           1
## 33     0.6324028              1           1
## 34     0.5290202              1           1
## 35     0.6896182              1           1
## 36     0.5501995              1           1
## 37     0.5006310              1           1
## 38     0.4510501              0           1
## 39     0.6324028              1           1
## 40     0.6584121              1           1
## 41     0.4161647              0           1
## 42     0.5781452              1           1
## 43     0.5431561              1           1
## 44     0.4024225              0           1
## 45     0.4440230              0           1
## 46     0.6190917              1           1
## 47     0.4935259              0           1
## 48     0.6455150              1           1
## 49     0.6055993              1           1
## 50     0.4793264              0           1
## 51     0.6389848              1           1
## 52     0.5148375              1           1
## 53     0.5919440              1           1
## 54     0.2594011              0           1
## 55     0.7922184              1           1
## 56     0.7778333              1           1
## 57     0.6519914              1           1
## 58     0.6896182              1           1
## 59     0.5148375              1           1
## 60     0.4230864              0           1
## 61     0.5360954              1           1
## 62     0.2381642              0           1
## 63     0.7305250              1           1
## 64     0.4230864              0           1
## 65     0.7968580              1           1
## 66     0.7415675              1           1
## 67     0.4651603              0           1
## 68     0.3687715              0           1
```

```
## 69     0.7075678             1           1
## 70     0.7415675             1           1
## 71     0.6519914             1           1
## 72     0.5219333             1           1
## 73     0.4092759             0           1
## 74     0.5006310             1           1
## 75     0.5919440             1           1
## 76     0.6123671             1           1
## 77     0.5501995             1           1
## 78     0.6584121             1           1
## 79     0.6190917             1           1
## 80     0.4935259             0           1
## 81     0.6519914             1           1
## 82     0.6519914             1           1
## 83     0.4440230             0           1
## 84     0.7134138             1           1
## 85     0.3491549             0           1
## 86     0.5219333             1           1
## 87     0.7778333             1           1
## 88     0.2818312             0           1
## 89     0.7469769             1           1
## 90     0.5360954             1           1
## 91     0.4651603             0           1
## 92     0.5290202             1           1
## 93     0.3363506             0           1
## 94     0.5431561             1           1
## 95     0.5711984             1           1
## 96     0.6123671             1           1
## 97     0.4230864             0           1
## 98     0.5148375             1           0
## 99     0.5006310             1           0
## 100    0.5987908             1           0
## 101    0.5006310             1           0
## 102    0.6257710             1           0
## 103    0.5987908             1           0
## 104    0.5642233             1           0
## 105    0.6455150             1           0
## 106    0.4440230             0           0
## 107    0.4793264             0           0
## 108    0.3363506             0           0
## 109    0.5642233             1           0
## 110    0.4024225             0           0
## 111    0.4580968             0           0
## 112    0.5360954             1           0
## 113    0.6123671             1           0
## 114    0.5360954             1           0
## 115    0.1948863             0           0
## 116    0.2934778             0           0
## 117    0.3300361             0           0
## 118    0.3300361             0           0
## 119    0.3363506             0           0
## 120    0.4864234             0           0
## 121    0.5987908             1           0
## 122    0.3956070             0           0
```

```
## 123    0.4300384             0           0
## 124    0.5360954             1           0
## 125    0.6519914             1           0
## 126    0.2381642             0           0
## 127    0.8230722             1           0
## 128    0.3237823             0           0
## 129    0.6647752             1           0
## 130    0.3621805             0           0
## 131    0.4722377             0           0
## 132    0.5642233             1           0
## 133    0.4793264             0           0
## 134    0.5148375             1           0
## 135    0.5781452             1           0
## 136    0.4161647             0           0
## 137    0.3820993             0           0
## 138    0.4580968             0           0
## 139    0.6190917             1           0
## 140    0.5501995             1           0
## 141    0.3888318             0           0
## 142    0.4300384             0           0
## 143    0.4580968             0           0
## 144    0.7875011             1           0
## 145    0.5077358             1           0
## 146    0.6190917             1           0
## 147    0.4935259             0           0
## 148    0.4793264             0           0
## 149    0.4510501             0           0
## 150    0.3237823             0           0
## 151    0.3621805             0           0
## 152    0.4651603             0           0
## 153    0.4864234             0           0
## 154    0.4510501             0           0
## 155    0.5006310             1           0
## 156    0.5360954             1           0
## 157    0.5077358             1           0
## 158    0.4440230             0           0
## 159    0.5148375             1           0
## 160    0.3754118             0           0
## 161    0.4024225             0           0
## 162    0.5850613             1           0
## 163    0.4580968             0           0
## 164    0.2539784             0           0
## 165    0.5781452             1           0
## 166    0.4230864             0           0
## 167    0.3888318             0           0
## 168    0.4722377             0           0
## 169    0.6324028             1           0
## 170    0.4651603             0           0
## 171    0.1735496             0           0
## 172    0.4300384             0           0
## 173    0.4370182             0           0
## 174    0.5219333             1           0
## 175    0.5642233             1           0
## 176    0.4300384             0           0
```

```
## 177    0.3491549             0           0
## 178    0.5642233             1           0
## 179    0.5077358             1           0
## 180    0.5572228             1           0
## 181    0.2381642             0           0
## 182    0.4935259             0           0
## 183    0.4864234             0           0
## 184    0.5148375             1           0
## 185    0.2486310             0           0
## 186    0.4370182             0           0
## 187    0.2486310             0           0
## 188    0.2230403             0           0
## 189    0.4370182             0           0
## 190    0.4510501             0           0
## 191    0.3956070             0           0
## 192    0.3556410             0           0
## 193    0.6324028             1           0
## 194    0.4230864             0           0
## 195    0.3754118             0           0
## 196    0.5360954             1           0
## 197    0.8311986             1           0
## 198    0.6455150             1           0
## 199    0.7016523             1           0
## 200    0.3363506             0           0
## 201    0.6389848             1           0
## 202    0.5919440             1           0
## 203    0.5148375             1           0
## 204    0.5501995             1           0
## 205    0.5501995             1           0
## 206    0.3491549             0           0
## 207    0.4161647             0           0
## 208    0.4092759             0           0
## 209    0.5919440             1           0
## 210    0.5501995             1           0
## 211    0.3621805             0           0
## 212    0.6055993             1           0
## 213    0.5006310             1           0
## 214    0.4722377             0           0
## 215    0.5290202             1           0
## 216    0.3175909             0           0
## 217    0.4370182             0           0
## 218    0.6324028             1           0
## 219    0.5360954             1           0
## 220    0.4510501             0           0
## 221    0.4793264             0           0
## 222    0.2818312             0           0
## 223    0.7305250             1           0
## 224    0.6055993             1           0
## 225    0.4651603             0           0
## 226    0.3621805             0           0
## 227    0.6190917             1           0
## 228    0.5431561             1           0
## 229    0.4230864             0           0
## 230    0.5219333             1           0
```

```
## 231    0.5360954              1           0
## 232    0.3300361              0           0
## 233    0.3687715              0           0
## 234    0.6389848              1           0
## 235    0.5077358              1           0
## 236    0.6519914              1           0
## 237    0.3175909              0           0
## 238    0.5642233              1           0
## 239    0.3888318              0           0
## 240    0.5077358              1           0
## 241    0.4440230              0           0
## 242    0.4092759              0           0
## 243    0.5431561              1           0
## 244    0.4370182              0           0
## 245    0.4580968              0           0
## 246    0.5148375              1           0
```

```
cat("Accuracy score:", sum(as.numeric(comp > 0.5) == baby_test$smoke)/length(baby_test$smoke))
```

```
## Accuracy score: 0.597561
```

The table above displays predicted probabilities, predictions results made by comparing those probabilities with 0.5, and the true smoking status of mothers. We also calculate the accuracy of our logistic regression model by computing the proportion of the smoking status that are correctly predicted. Our model has an accuracy of 0.597, which means that our model predicts smoking status of 59.7% mothers in the testing set correctly. This score also suggests that our model has a better accuracy than randomly guessing and it is a fairly good predictor for whether the mother is a smoker or not.

Conclusion

To determine whether a mother is smoker or not, we first explore the data using boxplots for variables **bwt**, **gestation**, **parity**, **age**, **height**, and **weight**, versus the binary variable **smoke** to visually investigate the association between smoke and the other features. Based on the results, we find that the distributions of birth weights of babies given birth by smoking and non-smoking mothers appear visually to be different, so it's reasonable to believe **bwt** is associated with **smoke**. Next, we apply log transformation to those variables whose data have some outliers, which are **gestation**, **age**, **height**, and **weight**. Then, we make new boxplots with outliers and without outliers respectively for these transformed variables versus the binary variable smoke. The difference in distributions of these variables in smoking and non-smoking mother groups shown in the boxplots are not evident, so we decide to include all of them as predictors when building our first logistic regression, and then check if there is an association between them and smoke by comparing their p-values computed by logistic regression model with the significance level 0.05. Then we perform a chi-square test of independence to test whether there is an association between the categories of **parity** and **smoke**. The null hypothesis is that the categories of parity and smoke are independent, while the alternative hypothesis is that there is an association between the categories of parity and smoke. Since the p value of this test is 0.7025691 and larger than 0.05, we fail to reject the null hypothesis that the categorical variables parity and smoke are independent, and there are clear evidences for us to conclude that there is no association between the categories of parity and smoke. Hence, parity is not helpful for predicting smoke.

There are 1236 observations in the dataset, but 10 observations whose mothers' smoking status are unknown. Because we want to build a model that only predicts whether the mother is a smoker or not, we remove those 10 observations and then split data into a training set and a test set. The training set contains 80% observations of each of smoking and non-smoking mother groups, and the remaining 20% of data are combined into a testing set. Afterwards, we build a logistic regression using all variables whose relationships with smoke

can not be confirmed due to the lack of statistical evidence. They are baby's birth weight, gestation, mother's weight, age, and height. To determine whether these variables have relationship with the smoking indicator statistically, we fit the model on the training set and observe the coefficients and the corresponding p-value of each variable. The results show that only baby's birth weight, has a p-value significantly less than 0.05, while all other variables included have p-values that are greater than our significance level of 0.05. This suggests that the smoking status of a mother is not dependent on variables other than **bwt**. Thus, we only include birth weight as the predictor of the probability of a baby having a smoking mother. What's more, according to the result of the logistic regression model that is fitted on training set and built on bwt, the coefficient corresponding to the birth weight is equal to -0.028422, the intercept is 3.384721, and p-value is less than 0.05, so we reject the hypothesis that there is no relationship between bwt and smoke.

After fitting this new model on the training set, we create a function that uses the intercept and coefficient calculated using predictor **bwt** to compute the probabilities of babies having smoking mothers for our testing dataset. If the probability that the mother is a smoker is greater than 0.5, then we predict that the mother is a smoker, and if the probability is less than or equal to 0.5, then we predict that the mother is not a smoker. In the ned, our model has an accuracy of 0.597, which means that our model predicts smoking status of 59.7% mothers in the testing set correctly. This score also suggests that our model has a better accuracy than randomly guessing and it is a fairly good predictor for whether the mother is a smoker or not.

As for other application of this predictive model, we can investigate the association between diabetes and people's weight, alcohol consumption, height, and age. When it comes to diabetes, the public usually believe that a relationship exists between sugar intake and biabetes. Hence, we can apply this predictive model to analyze the relationships between sugar intake as well as other features whose association with diabetes can not be confirmed due to the lack of statistical evidence.