Math 189: Homework 6

Alan Lui, Derek So, Xiangyu Wei

February 28, 2021

Introduction

This study looks into a classification task focusing on the `babies` dataset. Using a logistic regression model, we will incorporate a child's birth features to predict whether the mother smoked during their pregnancy. While using effects/correlations to predict causation is not necessarily sound, it may reveal important links between smoking and the features of a child's birth. Such a model, for example, could be used to check whether mothers were truthful about their smoking habits, should their child's birth weight be anomalous (and a predictive model is valid).

Tasks & Analysis

Dataset

The dataset that is used in this report is the *babies.dat* dataset, which contains the following features:

- **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
- **smoking Indicator**: for whether the mother smokes (1) or not (0); (9) denotes unknown.

The dataset is collected for each new mother in a *Child and Health Development Study*. It is found http://www.stat.berkeley.edu/users/statlabs/labs.html, and it is presented by *Stat Labs*: Mathematical Statistics through Applications Springer-Verlag (2001) by Deborah Nolan and Terry Speed. We extracted the dataset from https://github.com/tuckermcelroy/ma189/blob/main/Data/babies.dat at 2021-01-12 20:07:02 PST.

The following code is written to load and clean (remove missing/unknown values) the dataset.

```
#import dataset
babies <- read.csv("Data/babies.dat", sep="")

#removing missing/unknown values
babies <- babies[babies$smoke != 9 & babies$height != 99 & babies$age != 99 &
                 babies$weight != 999 & babies$gestation != 999,]
```

Methods

After exploring the potential factors that could be used to predict mother smoking status using scatter plots and box plots, we separate the data into training set and testing set, and train a logistic regression model using the selected features.

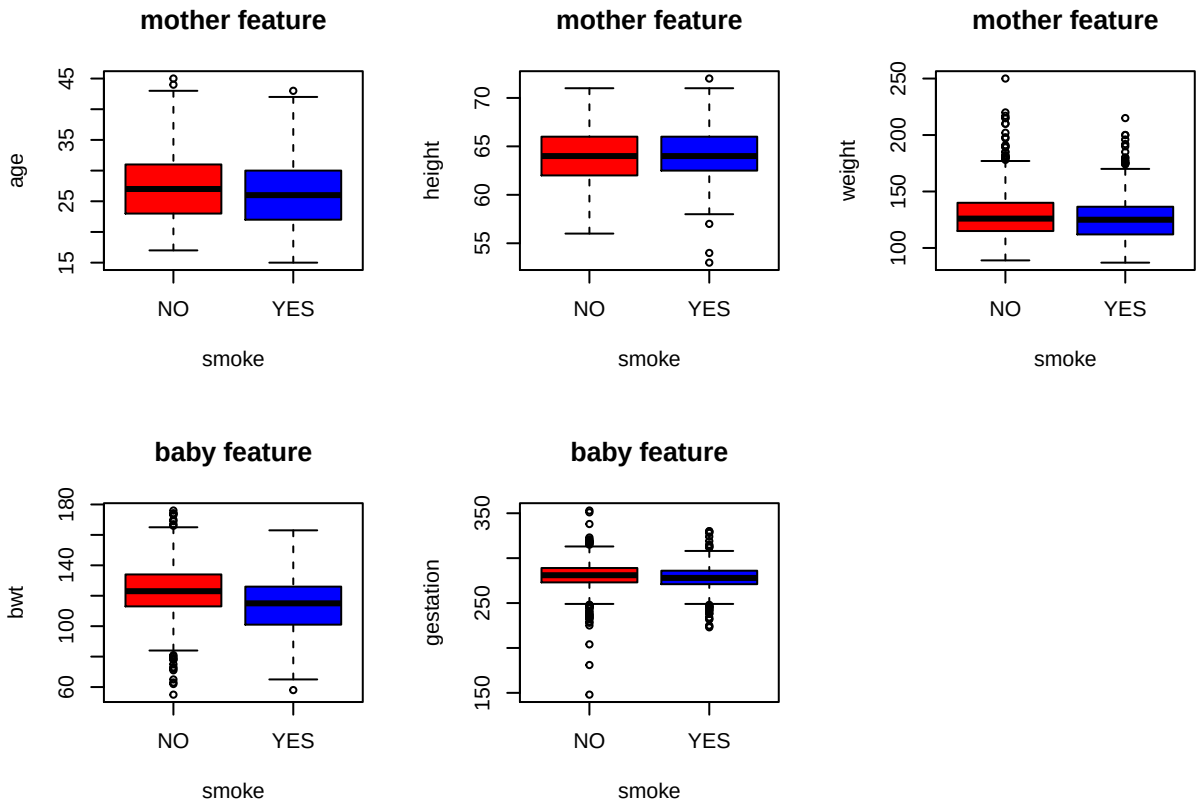
Analysis

1. Explore the data graphically. Describe your findings.

We first use box plots to visualize the relationship (red represents NO smoking, blue represents YES smoking).

```
par(mfrow = c(2, 3)) # format the graphs
args <- list(col = c("red", "blue"), names = c("NO","YES"))
#these features are of the mother and not the baby
boxplot(data = babies, age ~ smoke, col = args[[1]], names = args[[2]], main = "mother feature")
boxplot(data = babies, height ~ smoke, col = args[[1]], names = args[[2]], main = "mother feature")
boxplot(data = babies, weight ~ smoke, col = args[[1]], names = args[[2]], main = "mother feature")

#baby features
boxplot(data = babies, bwt ~ smoke, col = args[[1]], names = args[[2]], main = "baby feature")
boxplot(data = babies, gestation ~ smoke, col = args[[1]], names = args[[2]], main = "baby feature")
```



Boxplots are an effective representation of differences in distribution, based on the mother's smoking status. Overall, the distributions were fairly similar to each other and no feature seems to stand out as a key prediction factor.

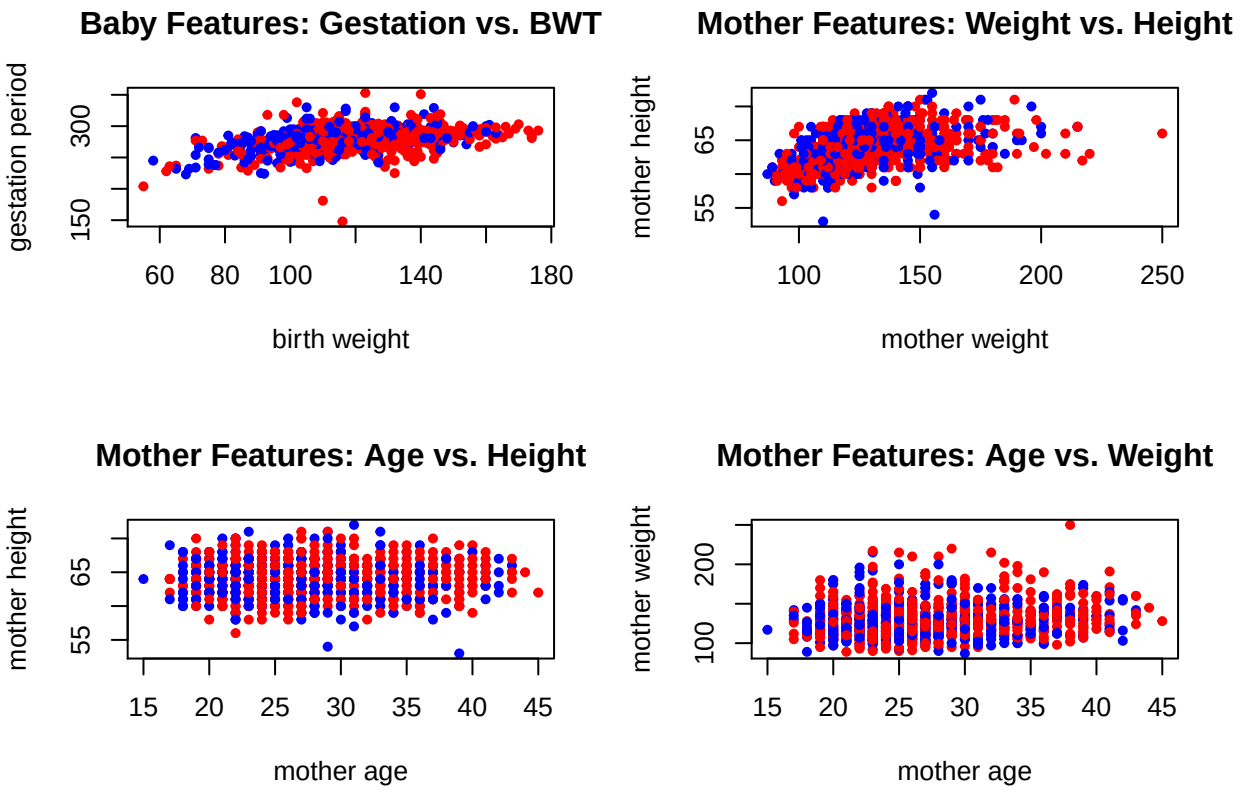
Then we use scatter plots to visualize the relationship between variables, separated by colors (red represents NO smoking, blue represents YES smoking).

```
par(mfrow = c(2, 2))
col_scale = ifelse(babies$smoke == 1, "blue", "red")
plot(babies$bwt, babies$gestation, pch = 20, col = col_scale,
     xlab = "birth weight", ylab = "gestation period",
     main = "Baby Features: Gestation vs. BWT")

plot(babies$weight, babies$height, pch = 20, col = col_scale,
     xlab = "mother weight", ylab = "mother height",
     main = "Mother Features: Weight vs. Height")

plot(babies$age, babies$height, pch = 20, col = col_scale,
     xlab = "mother age", ylab = "mother height",
     main = "Mother Features: Age vs. Height")

plot(babies$age, babies$weight, pch = 20, col = col_scale,
     xlab = "mother age", ylab = "mother weight",
     main = "Mother Features: Age vs. Weight")
```



Using scatterplots to explore the data again shows that the spread for the variables depending on smoking status is fairly insignificant. The blue and red dots are mostly intertwined with each other, meaning there's no signficant difference.

Thus, between both explorations (box plots and scatter plots), no variables showed significant relationships with the other.

A 2×2 table is used to show the count of `parity` by `smoke`.

```
table(ifelse(babies$smoke == 1, "Smoke YES", "Smoke No"),
      ifelse(babies$parity == 1, "First Born", "Not First"))
```

```
##
##             First Born Not First
##   Smoke No         190       525
##   Smoke YES        118       341
```

Using the 2×2 table, we can see the count of `parity` by `smoke` are different for each condition.

2. There are 1236 observations. Split the data into a training set and a test set, of sizes that you select (and justify).

```
prop_split = .7 # proportion of training set
training_size = as.integer(nrow(babies) * prop_split)
train_ind = sample(1:nrow(babies), training_size)
train = babies[train_ind,] # training set
test = babies[-train_ind,] # testing set
```

We did a 70/30 train/test split, which is a sufficient training amount given the large size of the dataset to prevent overfitting/underfitting.

3. Perform logistic regression on the training data in order to predict smoking indicator using the variables that seemed most associated.

Before conducting any analysis, it is important to identify if the dataset is fit for logistic regression. Looking at the box plots above, the distributions of `age` & `height` seem to be overall symmetric with no apparent outliers, while the distributions of `weight`, `bwt`, & `gestation` seem to have some outliers. Thus, we would need to take the violation of normal distribution into account.

```
cor(babies)
```

```
##                    bwt   gestation       parity          age       height
## bwt         1.00000000  0.40754279 -0.043908173  0.026982911  0.203704177
## gestation   0.40754279  1.00000000  0.080916029 -0.053424774  0.070469902
## parity     -0.04390817  0.08091603  1.000000000 -0.351040648  0.043543487
## age         0.02698291 -0.05342477 -0.351040648  1.000000000 -0.006452846
## height      0.20370418  0.07046990  0.043543487 -0.006452846  1.000000000
## weight      0.15592327  0.02365494 -0.096362092  0.147322111  0.435287428
## smoke      -0.24679951 -0.06026684 -0.009598971 -0.067771942  0.017506595
##                 weight        smoke
## bwt         0.15592327 -0.246799515
## gestation   0.02365494 -0.060266842
## parity     -0.09636209 -0.009598971
## age         0.14732211 -0.067771942
## height      0.43528743  0.017506595
## weight      1.00000000 -0.060281396
## smoke      -0.06028140  1.000000000
```

Looking at the correlation matrix between all features, there seems to be no significant collinearity between the features.

Then, it is simply the matter of performing logistic regression on the training data to predict the smoking indicator. We decide to fit three models that use different features. The first model uses ALL the features. But we feel like since we are predicting the smoking status of mom, we should not be using mother's features. Otherwise, the prediction task is not what we set out to perform. Thus, we also include the second model that uses ONLY significant features from the mode1 1 (which are `bwt`, `age`, & `height`); and the third model that uses ONLY baby features (which are `bwt`, `gestation`, & `parity`).

```
# ALL features
mdl1 <- glm(formula = smoke ~ bwt + gestation + parity + age + height + weight,
            family = binomial, data = train)
summary(mdl1)$coefficients
```

```
##                 Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -2.795549250 2.308772571 -1.210838 2.259576e-01
## bwt         -0.034300595 0.004913025 -6.981563 2.919143e-12
## gestation    0.007557802 0.005166825  1.462756 1.435343e-01
## parity      -0.378265074 0.184274238 -2.052729 4.009887e-02
## age         -0.029332780 0.014323305 -2.047906 4.056923e-02
## height       0.098906989 0.033774100  2.928486 3.406169e-03
## weight      -0.008759451 0.004344218 -2.016347 4.376369e-02
```

```
# significant features
mdl2 <- glm(formula = smoke ~ bwt + age + height,
            family = binomial, data = train)
# ONLY baby features
mdl3 <- glm(formula = smoke ~ bwt + gestation + parity,
            family = binomial, data = train)
```

4. Generate the prediction probabilities for the test data, and discuss the results.

```
# model 1 results
mdl1_results = ifelse(predict(mdl1, test, type = "response") > .5, 1, 0)
mdl1_acc = sum(mdl1_results == test$smoke)/length(mdl1_results)
# model 2 results
mdl2_results = ifelse(predict(mdl2, test, type = "response") > .5, 1, 0)
mdl2_acc = sum(mdl2_results == test$smoke)/length(mdl2_results)
# model 3 results
mdl3_results = ifelse(predict(mdl3, test, type = "response") > .5,1, 0)
mdl3_acc = sum(mdl3_results == test$smoke)/length(mdl3_results)
# display results
c("Model 1" = mdl1_acc, "Model 2" = mdl2_acc, "Model 3" = mdl3_acc)
```

```
##   Model 1   Model 2   Model 3
## 0.6600567 0.6515581 0.6487252
```

The first model that includes all features has the highest accuracy around 65-67% depending on the train/test split. The second model (ONLY significant features) and the third model (ONLY baby features) have similar accuracies that are usually (but not always) lower than the first full model depending on the train/test split

Conclusion

All of the three models' accuracy hovered just below 70%. Even though their performances (measured by accuracy rate) varies, it seems that the three model all perform relatively similar to each other. This may be a result of the violation of the logistic regression assumption, or a result of the small sample size since a larger one is usually required for this kind of prediction model. The accuracies of the models will be discussed in applications in the discussion.

Discussion

As stated in the introduction, it is possible to use the model to check whether the mother is truthful about their smoking habits. Other applications of such a model may include finding out the smoking proportion of birth mothers given a baby's birth features, at a more analytical standpoint, rather than taking the word of self-disclosure. Of course, both these ideas depend on the robustness of the predictive model produced.

However, given the fairly low accuracy of the model, such a task is likely not effective. Given that a completely random model would have an accuracy of about 50%, a roughly 70% accurate model, while significant, is not much to speak of.

Strangely, gestation period did not seem to be a major factor in the earlier model and was thus removed from the final model. Anecdotal information suggests that early births are an indicator of a mother's lesser wellbeing. However, as with the substandard accuracy of the model, many factors besides smoking could lead to early births and anomalous birthweight, so the fact that gestation is not a major factor should not be surprising. It is important to keep in mind that the logistic regression regression model does not suggest causality but correlations. Though the model does not confirm empirical observations of child birthweight and mother smoking status, it does not necessarily mean that the birth weight of a child is not affected by the smoking status of the mother.

Appendix (adapted from lectures 15 & 16)

The Logistic Regression Model

- Let Yes be 1 and No be 0.
- Consider the conditional probability

$$p(x) = \mathbb{P}[Y = 1|X = x]$$

- We want to model $p(x)$ using a function that gives outputs between 0 and 1 for all values of x. Many functions meet this criterion.
- In logistic regression, we use the *logistic function*.

$$p(x) = \frac{\exp\{\beta_0 + \beta_1 x\}}{1 + \exp\{\beta_0 + \beta_1 x\}}$$

- To fit this model, we use a method called *maximum likelihood estimation*.
- Note that

$$\frac{p(x)}{1 - p(x)} = \exp\{\beta_0 + \beta_1 x\}$$

  The left-hand side is called the *odds*.
- Values of the odds close to 0 indicate very low probabilities of default.
- Values of the odds close to ∞ indicate very high probabilities of default.

How to Estimate the Regression Coefficient

- Suppose we have data $(y_i, x_i)_{i=1}^n$.
- Likelihood function: by model assumption, $\mathbf{P}[y_i = 1|x_i] = p(x_i)$ and $\mathbf{P}[y_i = 0|x_i] = 1 - p(x_i)$. So the conditional "density function" of y_i given x_i can be written compactly as

$$p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

- The joint conditional density of $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$ is

$$L_n(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

- This is a function of β_0 and β_1, because our model assumption is that

$$p(x_i) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$$

- The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen to maximize this likelihood function.
- The logistic regression model can be fitted using a statistical software package, such as `glm` in R.

Multiple Logistic Regression

- Goal: predict a binary response using multiple predictors.
- We generalize the previous model as follows:

$$\log\left(\frac{p(\underline{X})}{1 - p(\underline{X})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

  where is $\underline{X}$ is a p-vector of covariates.
- Equivalently,

$$p(\underline{X}) = \frac{\exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\}}{1 + \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\}}$$

- Let $\underline{\beta} = [\beta_1, \ldots, \beta_p]'$ be the vector of regression coefficients; β_0 is referred to as the intercept.