Math 189: Homework 6

Linking Baby Features to Smoking

In this assignment you will again examine the Baby data set. It is believed that smoking status has a causal impact on birthweight, and perhaps on gestation. Can we flip this around, to see if we can predict smoking status from baby characteristics? Although it does not seem possible that a baby's characteristics would cause smoking status, it could still be useful to develop a predictive model. (Why?)

We will develop a logistic regression model to do the classification. The task is to develop a model to predict whether a given mother is a smoker or not. This data can be found on GitHub.

Metadata for *babies.dat*

- **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.

Tasks

Analyze the dataset according to the following steps:

1. Explore the data graphically in order to investigate the association between the **smoking indicator** and bwt, gestation, or *any of the other variables* that seem appropriate. Which of the other features seem most likely to be useful in predicting smoker status? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.
2. There are 1236 observations. Split the data into a training set and a test set, of sizes that you select (and justify).
3. Perform logistic regression on the training data in order to predict smoking indicator using the variables that seemed most associated.
4. Generate the prediction probabilities for the test data, and discuss the results.
5. In your conclusion, discuss possible applications of such a predictive model. Comment on how it is possible to predict from variables that are not causing a phenomenon.

Remarks

Your R Markdown Notebook report should have a introduction, body, conclusion (and optional appendix). Importantly, your code should run!