Homework 1

Christine Chung

April 4, 2016

Introduction

Homework 1 for Machine Learning in Public Policy, CAPP 30254
Spring 2016.

Files can be found at:
`https://github.com/cszc/Machine-Learning-for-Public-Policy/tree/master/HW1`

Problem A

This section lists the summary statistics for the data set ***mock_student_data***, which contains 1000 rows.

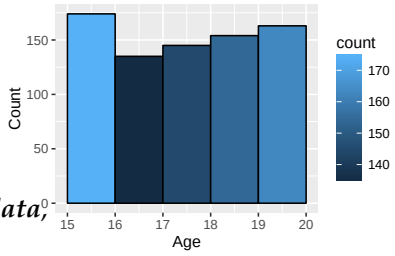

Figure 1: Histgram: Age

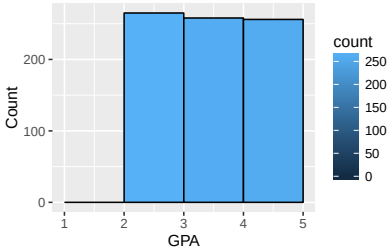

Figure 2: Histgram: GPA

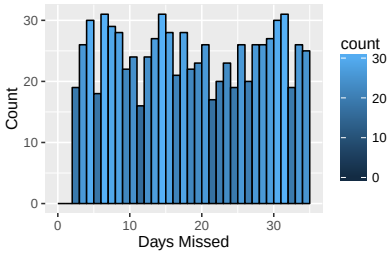

Figure 3: Histogram: Days Missed

A. Statistics

Summary Statistics

Statistic	Age	Days Missed	GPA
count	771.000000	779.000000	808.000000
mean	500.500000	16.996109	2.988447
median	500.5	17.0	3.0
mode	15	2	6
std	1.458067	0.818249	9.629371
min	15.000000	2.000000	2.000000
25%	16.000000	2.000000	9.000000
50%	17.000000	3.000000	18.000000
75%	18.000000	4.000000	27.000000
max	19.000000	4.000000	34.000000

Missing Values

First Name	Last Name	State	Gender	Age	GPA	Days Missed	Graduated
0	0	116	226	229	221	192	0

B. Filling in Missing Genders

The output for the new predicted genders using the genderize API are contained in the document *mock_with_gender.csv*.

C. Filling in Missing Fields with Means and Class Conditional Means

Simple Means

The means for each column can be found in the section above. The output for the new predicted values using means are contained in the document *mock_with_means.csv*.

Class Conditional Means

Note that for the class conditional means, only those that had graduated high school had missing values for gpa and days missed.

Column	Graduated	Not Graduated
age	16.959	17.052
gpa	3.505	2.516
days missed	16.776	19.229

The output for the new predicted values using conditional means are contained in the document *mock_with_cond_means.csv*.

Other Methods

THERE ARE a few other methods that I can think of for filling in missing fields. One way is using a decision tree to predict class labels for each missing value. Another way is to use regression imputation (using simple linear regression or logistic regression) to predict missing values. However, this second method may be difficult to do if some rows are missing multiple values, some of which would be used in the regression model.

For example, the distribution of the predicted values for Age regressed on Gender, GPA, Days_missed, and Graduates, looks like:

summary(predict(fit))

Min.	1st Qu.	Median	Mean 3rd Qu.	Max.
16.47	16.88	16.96	16.97	17.08

Because variation for among the missing values is not very large,

I will employ a much simpler tool for this assignment. In addition to conditioning on "Graduated", I will also condition on "Gender" producing four mean values per missing variable:

Graduated & Gender	Age	GPA	Days Missed
Yes - Female	16.968	3.541	17.793
Yes - Male	16.949	3.466	15.622
No - Female	16.968	2.514	19.298
No - Male	17.136	2.518	19.156

The output for the new predicted values are in the document *mock_partc.csv*.

Problem B

1. Consider 4 Students

Based on the coefficients, it is not possible to tell based on the information provided whether or not Chris or David has a higher probability of graduating. The only information we have is each of their respective incomes. We know that holding all else equal, David would have a lower probability of graduationg since the coefficient for log income is negative. However, we are not able to make this claim since we are not given the individual characterstics of Bob and David and we are not told that they are otherwise the same.

2. Coefficient for AfAm_Male

The probability of graduating is given by: *1/(1+exp(-(intercept + xBeta))* where xBeta is a vector.

Holding all else equal, the coefficient for AfAm_Male means that being African-American and Male decreases the likelihood of graduating. African-American Males are less likely to graduate than African-American Females (coefficient for AfAm) and non-African-American males (coefficient for Male). However, it is impossible to disaggregate the affect of being African American and being Male.

3. Interpretation of Ages

The coefficient for Age is -0.013. This means, holding all else equal, that the older you are the less likely you are to graduate. By using the polynomials age and ageˆ2, we are smoothing the distribution of ages. This allows us to estimate a non-linear function.

4. Dropping Variables

I would drop any variables that are not significant. I would need the p or T value in order to make this decision. In addition, I would drop either Male or Female. They should not both be in the regression. One should be encompassed in the default value, that is, the constant value. In addition, I would include AfAm as its own variable, dropping AfAm-Male, and instead add the interaction variables *AfAm x Male* and *AfAm x Female* (assuming these are significant variables).