ML hw 1

Michael Fosco

1)

Here is a description of the data that includes the count, mean, standard deviation, and percentiles. Note that the median is the 50th percentile.

ID Age GPA Days\_missed

count 1000.000000 771.000000 779.000000 808.000000

mean 500.500000 16.996109 2.988447 18.011139

std 288.819436 1.458067 0.818249 9.629371

min 1.000000 15.000000 2.000000 2.000000

25% 250.750000 16.000000 2.000000 9.000000

50% 500.500000 17.000000 3.000000 18.000000

75% 750.250000 18.000000 4.000000 27.000000

max 1000.000000 19.000000 4.000000 34.000000

Below is the list of missing values for each variable.

ID 0

First\_name 0

Last\_name 0

State 116

Gender 226

Age 229

GPA 221

Days\_missed 192

Graduated 0

dtype: int64

And here is the mode for each variable in the data:

First\_name mode: Amy

Last\_name mode: Ross

State mode: Texas

Gender mode: Female

Age mode: 15.0

GPA mode: 2.0

Days\_missed mode: 6.0

Graduated mode: Yes

It is interesting to note that gpa is discrete, only taking on the values of 2, 3, or 4, potentially indicative of measurement error since it is unlikely the only GPAs people ever get at this school are integers. It should also be noted that since the mean age is 17 we are likely dealing with high school students.

The plots for variables are included in other pdf files by variable name and are reproduced below.



Graduated:



Gender:



States:



For part C, instead of just taking the means of the values (conditional on just graduation or non conditional) it should be better to condition on even more so that when we take means we have a more homogenous group. To improve the prediction I conditioned on whether or not someone graduated, gender, and state and then took the mean to fill in the missing value. The result is in wayC.csv

2)

2.A (the first one)

Question:

Based on the coefficients above, who would you think has a higher probability of graduating?

* Chris
* David
* They have the same probability
* Cannot tell based on the information provided

What is your reasoning?  (you need not calculate an exact probability to answer this question. Just explain your reasoning in general terms.)

I expect Chris to have a higher probability of graduating since the difference in logs between 50,000 and 40,000 is a larger jump than between 200,000 and 190,000. Thus, since we have a negative coefficient on the log of the family income, Chris has less of a negative effect due to his family’s money than David. Since Chris has similar characteristics to Adam and David has similar characteristics to Bob, and both Bob and Adam have 50% probability, then I expect Chris to have a higher probability of graduating.

2.A (the second one)

1. Question: The coefficient for AfAm\_Male is negative. How do you interpret this? Does this mean that African-American Males are more likely to not graduate than African-American Females? What about relative to non African American males?

The effect of being an African Male is the sum of the male, African American, and African American male coefficients, which is greater than zero. The effect of being an African American female is the sum of being female and African American, which is less than zero. Thus, African American males are actually more likely to graduate than African American females. Similarly, when we sum coefficients, African American males are more likely to graduate than males. The fact the coefficient on African American males is negative just means that they are less likely than some other group to graduate.

2.B)

For the ages that concern us, there is a negative, quadratic effect with age. The squared effect of age is positive while that on age is negative, indicative of a relationship that decreases at a decreasing rate. The variables estimate the probability by altering the odds ratio through the model.

2.C)

It appears as though age and age squared are collinear. I would consider dropping one of them. However, before I did I would have to examine the theoretical reasoning for why they were included in the model and if the theoretical reasoning were strong enough then I would leave them. I would also look at the residuals and based on the shape of the residuals I would consider changing the shape of some of the variables used (such as changing logs) and potentially dropping some variables.