Predicting Race Using Demographic Information

In this exercise, we return to the problem of predicting the ethnicity of individual voters given their surname and residence location using Bayes' rule. This exercise is based on the following article: Kosuke Imai and Kabir Khanna. (2016). "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24(2): 263-272.

In this exercise, we attempt to improve that prediction by taking into account demographic information such as age and gender. As done earlier, we validate our method by comparing our predictions with the actual race of each voter.

Name	Description
`county`	County census id of voting district.
`VTD`	Voting district census id (only unique within county)
`total.pop`	Total population of voting district

Other variables are labeled in three parts, each separated by a period. See below for each part. Each column contains the proportion of people of that gender, age group, and race in the voting district.

Name	Description
`gender`	Male or female
`age groups`	Age groups as defined by U.S. Census (see table below)
`race`	Different racial categories (see table below)

Below is the table for variables describing racial categories:

Name	Description
`whi`	non-Hispanic whites in the voting district
`bla`	non-Hispanic blacks in the district
`his`	Hispanics
`asi`	non-Hispanic Asian and Pacific Islanders
`oth`	other racial categories
`mix`	non-Hispanic people of two or more races.

Below is the table for age-group variables, as defined by the U.S. Census:

Name	Description
`1`	18–19
`2`	20–24
`3`	25–29
`4`	30–34

Name	Description
5	35–39
6	40–44
7	45–49
8	50–54
9	55–59
10	60–64
11	65–69
12	70–74
13	75–79
14	80–84
15	85+

We use three data sets in this exercises, two of which were already introduced in Section 6.1. The first data set is a random sample of 10,000 registered voters contained in the csv file, `FLVoters.csv`. Table 6.1 presents the names and descriptions of variables for this data set. The second data set is a csv file, `cnames.csv`, containing a modified version of the original data set, `names.csv`, after making appropriate adjustments about a special value as done in Section 6.2. Table 6.3 presents the names and descriptions of variables in this data set. Finally, the third data set, `FLCensusDem`, contains the updated census data with two additional demographic variables – gender and age. Unlike the other census data we analyzed earlier, each observation of this data set consists of one voting district and the proportion of each demographic by age, gender, and race within that district. The tables above present the names and descriptions of variables in this data set of Florida districts. There is also a table that contains the age groupings used in the variable names of the `FLCensusDem.csv` file.

Question 1

Use Bayes' Rule to find a formula for the probability that a voter belongs to a given racial group conditional on their age, gender, surname, and residence location. Given the data sets we have, can we use this formula to predict each voter's race? If the answer is yes, briefly explain how you would make the prediction. If the answer is no, explain why you cannot apply the formula you derived.

Question 2

Assume that, given the person's race, the surname is conditionally independent from residence, age, and gender. Express this assumption mathematically and also substantively interpret. Show that under this assumption, the probability that a voter belongs to a given racial group conditional on their age and gender as well as their surname and residence location is given by the following formula.

$$\frac{P(\text{residence}, \text{age}, \text{gender} \mid \text{race}) P(\text{race} \mid \text{surname})}{P(\text{residence}, \text{age}, \text{gender} \mid \text{surname})}$$

Question 3

Using the formula derived in the previous question, we wish to compute the predicted probability that a voter belongs to a given racial group, conditional on their age and gender as well as their surname and residence location. Provide a step-by-step explanation of how to do this computation using the data. Hint: you will need to modify the formula without invoking an additional assumption such that all quantities can be computed from the data sets we have. The definition of conditional probability and the law of total probability might be useful.

Question 4

Use the procedure described in the previous question, compute the predicted probability for each voter in the `FLVoters.csv` that the voter belongs to a given racial group conditional on their age, gender, surname and residence location. Exclude the voters with missing data from your analysis. Also, note that the csv file `cnames.csv` has been processed from `names.csv` using the code from Section 6.1. Thus, there is no need to re-adjust the values to account for negligibly small race percentages, but the racial proportions by surname are initially expressed as percentages rather than as decimals.

Question 5

Given the results in the previous question, identify the most likely race for each individual in `FLVoters.csv`, given their surname, residence, age, and gender.

Question 6

To validate this race prediction methodology, compare the race predictions you've made in the previous question with the self-reported races of the voters, specifically for white, black, Hispanic, and Asian voters. How often did you correctly predict the race of the individuals? How often did you get false positives? How does your model compare to the predictions made in Section 6.1 based on surname and residence location alone?