Race Proxy Method-Two applicants
July 23, 2012

I. Problem

All fair lending analysis requires data on the race, ethnicity, and/or gender of loan applicants. In the case of mortgage lending, the Home Mortgage Disclosure Act requires lenders to collect and report these data. This allows for direct and accurate comparisons of outcomes across race, ethnicity, or gender.

Unfortunately, other financial products such as auto loans may not offer direct data on these prohibited basis groups (PBGs). In such cases, PBG status must be imputed from other attributes of the applicants. Surname and street address may hint toward an applicant's race or ethnicity, while a first name may indicate gender. To give a few hypothetical examples, [redacted] Miller of Bangor, Maine is most likely female, white, and not Hispanic. [redacted] Gonzales of Miami is most likely male and Hispanic. The proxy method will assign a vector of probabilities to every loan applicant, indicating our degree of confidence that the individual is of any given PBG.

Some applicants will have ambiguous combinations of name and location. In this case, a threshold rule can be used to exclude loan applicants whose race cannot be reliably estimated. For instance, we could classify an applicant as Hispanic only if name and geographic data show at least an 80% probability that the applicant is Hispanic.

Another challenge arises when applications have multiple applicants. If individuals applying together may have different races or genders, some rule must be chosen to categorize applications by PBG. For our purposes, an application is assigned a race or ethnicity whenever at least one of its applicants is of that race or ethnicity. A different rule is used for gender analysis: an application is classified as female only if all of its applicants are female.

II. General Method

The current method generates one set of race probabilities based on surnames and another based on local demographics. Surname-based race probabilities are created using data from the 2000 Census on the frequency of common surnames by race.[1] Geography-based race probabilities are generated using tract-level racial demographics from the 2010 Census. The two sets of probabilities are then combined using Bayesian updating, as discussed in Section II.C.

A. Race/Gender data

The proxy defines five to six mutually exclusive and collectively exhaustive ethnic categories. In the six-category case, each individual is either Hispanic, non-Hispanic (NH) white, NH black, NH Asian/Pacific Islander (API), NH American Indian (AI), or NH multiple/other. In the five-category case, the NH multiple/other option is removed. The justification for removing the NH multiple/other category was the inconsistent treatment of "Some other race" responses between

[1] http://www.census.gov/genealogy/www/data/2000surnames/index.html

the census tract file and the census surname file. In the surname file (but not in the tract file), "Some other race" responses were assigned to the various standard races.

Due to this inconsistency, the six-category model did not include "Some other race" individuals in its model of surname. This could lead the number of people assigned to the "Some other race" category to be biased downward. The five-category model corrects this issue by modifying the tract and surname files to remove the problematic category. "Some other race" or "Two or more races" individuals are assigned a standard race, leaving only the five standard categories. Specifically, each of the four standard NH categories absorbs a share of "Some other race" individuals proportional to its share of the tract population.

Applications using the six-category model: Ally, [redacted]
Applications using the five-category model: [redacted]
([redacted] and [redacted] also used the five-category model, but had major differences in method from the auto exams.)

Gender probabilities are created separately using a list of common first names by gender, provided by the Social Security Administration. The list is organized by birth year. Thus, a typical record might show that a Tracy born in 1985 had an 88% probability of being female. Our current method aggregates occurrences of each name across all birth years. This may entail some loss of accuracy, since some names have gradually changed gender over the years (e.g. Leslie). However, aggregating across years offers the countervailing advantage of increasing sample size, which may be useful for rare names.

A. Surnames

The current method for name parsing allows each applicant to have up to two surnames, to account for hyphenation. An application with an applicant and a coapplicant could thus have as many as four surnames.

Joint probabilities are computed by name rather than by individual. For example, consider an application with last names Ramirez-Washington and Washington. Let H(R) be the probability that a person named Ramirez is Hispanic, and let H(W) be the corresponding probability for Washington. Then the Ramirez-Washington/Washington application is classified as Hispanic with probability:

$$H(R)H(W) + (1-H(R))H(W) + H(R)(1-H(W))$$

$$\approx(.94)(.01) + (.06)(.99) + (.94)(.99)$$

$$= .9994$$

In words, the probability that at least one applicant is Hispanic is the probability that at least one name is Hispanic. This simple approach avoids the use of conditional probabilities by assuming that the name of one applicant is not a predictor of the race of the other.

However, this assumption of independence is most likely false. Individuals tend to make loan applications jointly with relatives or spouses, generally of the same reported race. Thus the racially ambiguous name Jackson (54% black, 43% white) would likely designate a black applicant when used alongside the (generally black) name Washington. It might instead designate a white applicant when used alongside the (generally white) name Epstein.

The independence assumption may thus lead us to overestimate diversity within applicant pairs. Coapplicants Jackson and Epstein are likely both white, but our method assigns them a 54.22% probability of including at least one black applicant. Fortunately, this falls far short of any threshold we might apply (e.g. 70%, 80%). Applying a sufficiently high threshold should help minimize the bias from our independence assumption.

B. Census tract demographics

Another indicator of race comes from local demographics near the address(es) of the applicant(s). The current method uses ArcGIS software to match each applicant address to a census tract. The demographic composition of the tract in 2010 then becomes a vector of probabilities for the applicant's race. For instance, an applicant in a census tract that is 60% white has a 60% probability of being white.

To simplify the analysis, the address of the coapplicant is not used in this process. This does not entail much loss of accuracy. Of 5.5 million observations for one institution, only 350,000 showed variation in address between applicant and coapplicant. Furthermore, applicants and coapplicants are likely to be of the same race, allowing us to use the address of just one to estimate the race of both.

Unfortunately, the local demographics analysis is set up to classify a single individual, not a pair. As a result, it will tend to misclassify some different-race applicant pairs as white only or minority only. It thus errs in the opposite direction of the surname analysis. Again, a sufficiently high threshold for classification should minimize the bias.

C. Combining surnames and local demographics

Surname and local demographic probabilities are combined by Bayesian updating to generate our authoritative set of probabilities. The method is described in Elliott et al., "Census Data for Proxies", Health Serv Outcomes Res Method (2009).