July 31, 2013

Brown Bag Presentation: Estimation of Race in Non-Mortgage Contexts

[redacted]

Consumer Financial
Protection Bureau

Presentation Outline and Goals

- Briefly discuss why estimating race matters to CFPB
- Concerns regarding current commonly accepted methods of assigning race/ethnicity to individuals
- Our proposed alternative for estimating likelihood of belonging to a given race/ethnicity
- Issues in using these estimates to examine relationship between reported race and outcomes

The Problem: We Need To Know Customers' Race, but Can't Ask

- Outside of mortgage products (covered under HMDA), Reg B generally prohibits financial institutions from asking/keeping track of their customers' race or ethnicity
- Meanwhile, ECOA prohibits racial discrimination in the provision of credit in many forms
- These two laws, both well-meaning, result in an inability to directly test for disparate outcomes based on reported race/ethnicity

Current “State of the Art” Is a Hodgepodge

- For some races, identification for purposes of fair lending analysis is done on basis of place of residence (e.g., African American)
- Others rely on distributions of race based on surname (e.g., Hispanics, Asian Americans)
- Discomfort with a probabilistic world + greater ease in remediation mean current fair lending analysis typically relies on use of “threshold rules”, where individuals are assigned to a given race/ethnicity when the above data sources identify a probability of membership greater than some commonly accepted level
  - Lose quite a bit of statistical power by eliminating sizeable chunk of population
  - Practically, there is a TON of selection going on here

CFPB OR Sought Alternatives to Current Standards

- We wanted to find a solution that met the following goals:
    - Easily implementable process
    - Provides most accurate estimates relative to reported race/ethnicity
    - Provides most accurate estimates of relationship between reported race/ethnicity and financial outcomes
    - Data transparency (i.e., publicly available)

- Any other concepts that should be key?

Data for Proxy comes from the Decennial Census

- Types of information we can typically receive from institutions include some geographic identifier (typically street address) and name of applicants
- We can geocode the address to a latitude/longitude, then use this to identify a geographic unit for which the 2010 Census provides race/ethnicity population counts
  - Currently using census tract, developing information for block group
- For surname, Census Bureau published a race/ethnicity breakdown for all surnames that appear more than 100 times in the 2000 enumeration
  - Provides estimates for 89.8% of population

The CFPB methodology then uses Bayes' Rule to create estimates

- Description of methodology provided in Elliott, et. al. (2008)
- Recall Bayes' Rule:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

- $p(r|s)$ =Pr(race r conditional on having surname s)
- $q(g|r)$ =Pr(living in geography g conditional on being race r)
- CRITICAL ASSUMPTION: $q(g|r) = q(g|r,s)$
  - □ In practical terms, this assumption states that surname provides no <u>meaningful</u> information on geography after accounting for race
  - □ E.g., We assume that Hispanic individuals named Lopez live in areas with similar racial demographics to other Hispanic individuals

The formula

- Given the assumptions listed earlier, the probability of belonging to race r given geography g and surname s is

$$\Pr(r|g,s) = \frac{p(r|s)q(g|r)}{\sum_{r \in R} p * q}$$

- To avoid additional complications, current implementation:
  - Uses only one surname (typically primary applicant, first surname)
  - Excludes observations for which geography or surname distributions cannot be found (workarounds exist)

Future adjustments can be made, pending data/computing power

- Using mixture modeling and machine learning algorithms, can use distributions of common first names found in datasets (e.g., Taro, Jose, Jamal)
- This will become more important as time goes on, since Census has no plans to repeat 2000 Census surname analysis for 2010
- Could attempt to use socioeconomic characteristics found in nationally representative datasets along with race/ethnicity (e.g., income)
  - □ This presents a whole new set of issues, to be discussed later

How do we determine if this proxy is “good”?

- Measuring reported characteristics
  - □ Matching population distribution
  - □ Contingency Tables (currently not in presentation, can discuss)
  - □ Receiver Operating Characteristics
- Measuring relationship between reported characteristics and outcomes
  - □ “Improper Support” (You have a better name?)
  - □ Omitted Variable Bias
  - □ Fair Lending Specific: Disparate Treatment v. Impact

To measure accuracy, we use HMDA + data

- Some institutions (3 total, at this time) provided name and address information in addition to standard HMDA and HMDA+ variables
- HMDA also requires collection of data on reported race and ethnicity when possible
- Provides large dataset on which we can compare estimates outcomes for both reported race and proxy
- Results shown here are for one lender only

Comparison of population estimates

- Bayesian proxy comes closest to matching true distribution
- Typically closest for each race/ethnicity

Race/Ethnicity	HMDA Rep.	Joint Proxy	Name Proxy	Geog. Proxy
Non-Hispanic White	0.73	0.68	0.67	0.66
Black	0.07	0.08	0.10	0.09
Asian/Pacific Islander	0.09	0.09	0.08	0.07
Native	0.003	0.004	0.006	0.005
Hispanic	0.11	0.12	0.12	0.15
Multiracial /Other	0.01	0.02	0.02	0.02

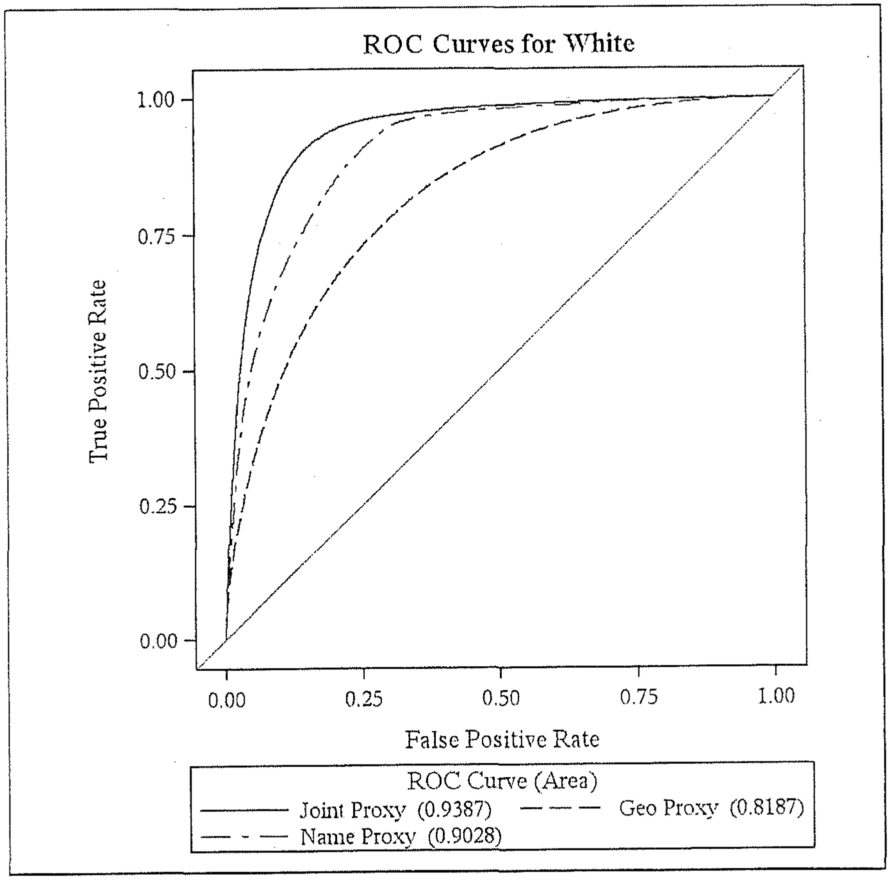
Why aren’t we matching more closely?

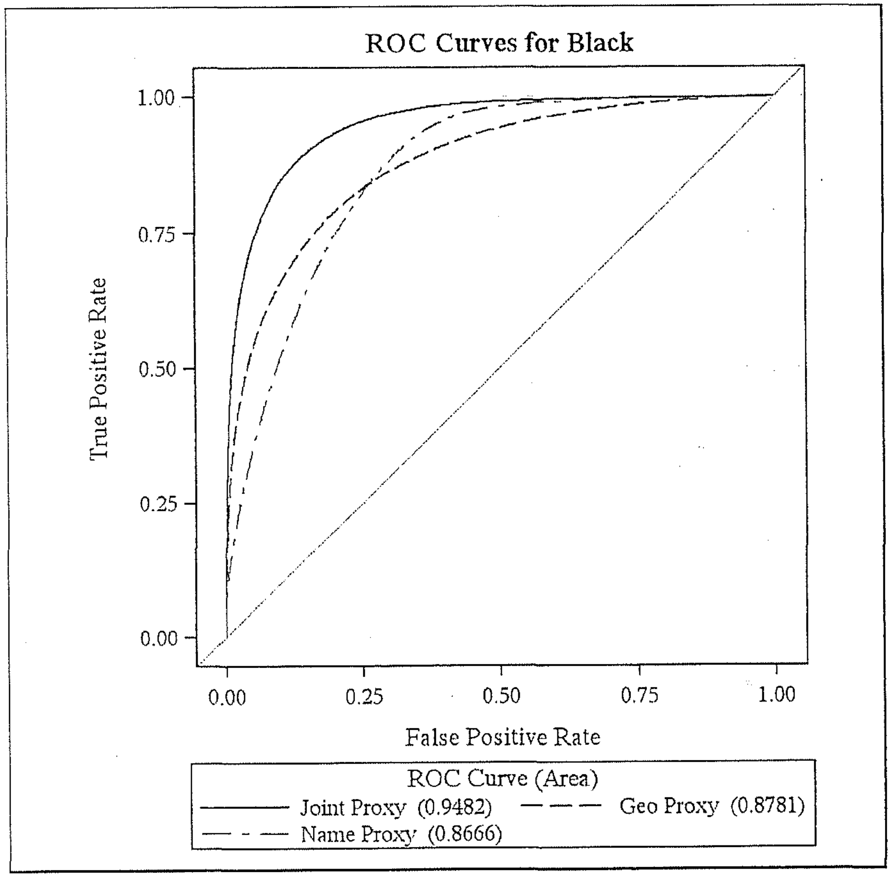
- Proxy is based on estimates drawn from the Census population
- Mortgage owners are likely different from the general population on a variety of covariates also correlated with race
- The impact of this difference in distribution between the general population and who uses a particular financial product will differ for each situation
- In this context, one might naively assume that overestimating membership in the treatment group would result in attenuation IF no selection exists
  - □ It probably does, but still matched pretty closely, especially for a product (mortgage) where we would expect more selection to occur

Comparison of Individual Race Estimates

- The Bayesian estimate most closely matches the population, but how well does it capture individuals' reported races?
- The Receiver Operating Characteristic (ROC) describes the probability that, if two individuals were randomly chosen from the sample, the one with the higher probability is more likely to belong to the treatment group
- A ROC curve represents how the false positive and false negative rates change as a threshold rule is applied from 1 to 0
- The area under this curve is equal to the ROC probability described above

Two notable improvements in accuracy…





AUC Results

BES1

	White	Black	Hispanic	Asian/ Pac. Isl.	Native Am.	Mult./ Other
Joint	0.9387	0.9482	0.9615	0.9723	0.6863	0.6355
	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0077)	(0.0050)
Geo	0.8187	0.8781	0.8376	0.8460	0.6713	0.5834
	(0.0007)	(0.0010)	(0.0009)	(0.0009)	(0.0076)	(0.0054)
Name	0.9028	0.8666	0.9492	0.9617	0.6210	0.6366
	(0.0005)	(0.0008)	(0.0006)	(0.0006)	(0.0073)	(0.0050)
p-value, H_0: Joint = Geo	<.0001	<.0001	<.0001	<.0001	0.0086	<.0001
p-value, H_0: Joint = Name	<.0001	<.0001	<.0001	<.0001	<.0001	0.8198

Slide 16

BES1	WHat is this? Stephens, Bryce (CFPB), 7/23/2013

Use of proxy in estimation of outcomes

- Ideally, we want to know the coefficient on the regression

$$y_{isg} = \beta r_{isg} + \varepsilon_{isg}$$

- Unfortunately, we don't know r, but believe we know some value $\bar{r}_{sg}$ s.t. $r_{isg} = \bar{r}_{sg} + v_{isg}$
- Regressing y on $\bar{r}$ results in coefficient estimate of

$$\tilde{\beta} = \frac{cov\left(y_{isg}, \bar{r}_{sg}\right)}{var\left(\bar{r}_{sg}\right)} = \frac{cov\left(\beta\left(\bar{r}_{sg} + v_{isg}\right) + \varepsilon_{isg}, \bar{r}_{sg}\right)}{var\left(\bar{r}_{sg}\right)} = \frac{\beta \sigma_{\bar{r}_{sg}}{}^2}{\sigma_{\bar{r}_{sg}}{}^2} = \beta$$

IF we assume $cov(v, \bar{r}) = cov(\varepsilon, \bar{r}) = 0$

Consumer Financial
Protection Bureau

What can go wrong with this estimation

1. The previous slide assumes a coefficient of 1 on the regression of the proxy on the reported truth
   - If this is not true it will bias our result upward or downward, depending on the coefficient magnitude
2. The previous slide assumes no omitted variable bias
   - Just as critically, for our purposes, it assumes that the bias of the reported race/ethnicity is equal to that of the proxy

1. Does the magnitude of the proxy match reality?

- To check this, perform a seemingly unrelated regression, with the outcome variables equal to a binary indicator for each of the race/ethnicity categories, and the proxy as the left hand side variable
  - No constant term in regression, as this is how proxy will be used in practice
  - SUR format allows for correlation in errors across categories
  - Calculate generalized goodness of fit measure equal to

$$R^2 = 1 - \frac{\sum_{r=1}^{R} SSR_{\bar{r}}}{\sum_{r=1}^{R} SSR_{tot}}$$

Bayesian proxy creates reasonable results, with large improvement in fit/reduction in residual error

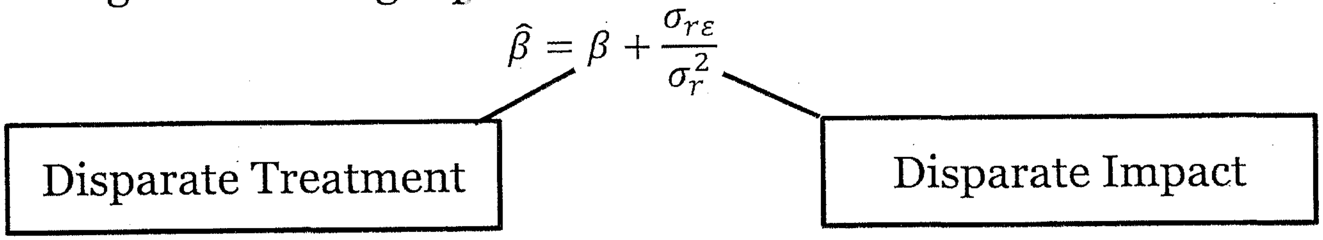
	Joint Proxy	Name Proxy	Geographic Proxy
Non-Hispanic White	1.01	1.011	1.014
	(0.0001)	(0.0001)	(0.0002)
Black	0.995	0.994	0.999
	(0.0004)	(0.0006)	(0.0005)
Native	0.632	0.612	0.631
	(0.0028)	(0.0039)	(0.0038)
Asian/Pacific Islander	1.003	1.016	1.077
	(0.0003)	(0.0003)	(0.0009)
Hispanic	0.991	0.995	0.993
	(0.0003)	(0.0003)	(0.0004)
Multiracial/Other	0.603	0.598	0.452
	(0.0018)	(0.0027)	(0.0029)
Gen. R^2	0.62	0.526	0.25

2. Omitted Variable Bias and FL Analysis

- What happens if we don't believe the covariances with the error terms are equal to zero?
  - □ Regression using reported truth:
  $$\hat{\beta} = \beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2}$$
  - □ Regression using proxy:
  $$\tilde{\beta} = \beta + \frac{\sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2}$$

2. Omitted Variable Bias and FL Analysis

- What happens if we don't believe the covariances with the error terms are equal to zero?
  - □ Regression using reported truth:



$$\hat{\beta} = \beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2}$$

Disparate Treatment

Disparate Impact

  - □ Regression using proxy:

$$\tilde{\beta} = \beta + \frac{\sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2}$$

- Though the disparate impact terms above converge to the same value as $\bar{r} \rightarrow r$, they will differ

How different are these two values?

- Though the disparate impact terms above converge to the same value as $\bar{r} \rightarrow r$, they will differ
- We can decompose the true disparate impact term, and compare the two:

$$\frac{\sigma_{\bar{r}\varepsilon}+\sigma_{v\varepsilon}}{\sigma_{\bar{r}}^2+2\sigma_{\bar{r}v}+\sigma_v^2} \quad \text{v.} \quad \frac{\sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2}$$

- $\sigma_{\bar{r}}^2 + 2\sigma_{\bar{r}v} + \sigma_v^2 > \sigma_{\bar{r}}^2$ biases proxy coefficient upward
- $\sigma_{v\varepsilon}$ means that if there are missing variables that impact both the measure of the proxy and the impact on outcomes, this will bias the coefficient of the proxy differently depending on the nature of the relationship
  - E.g., Income level in an examination of race and house price (low income correlated positively with minority status, negatively with house price)

Testing the Role of OVB

- Can use available variables from HMDA+ mortgage data to look at the role of omitted variables in this empirical context
- Want to compare results in environment where outcome is correlated with omitted variables in both proxy and outcome estimation to results where the two should be uncorrelated
  - Finding the former is easy: House price, APR, etc.
  - Finding the latter is more difficult: One potential option includes testing whether, conditional on qualifying for GSE-backing, a loan is picked up by either Fannie Mae or Freddie Mac

OVB: Example 1 – Basic Simulation

- Two Data Generating Processes:

$$y_1 = .1(Hispanic) + .3(Black) + .2(API) + .4(Native) + .6(Other) + \varepsilon$$

$$y_2 = .1(Hispanic) + .3(Black) + .2(API) + .4(Native) + .6(Other) + .0005(Income) + \varepsilon$$

- Results shown are just from one run (has been done over large range of simulations, results are similar)

Example 1 Results – no OVB

	HMDA Reporting	Joint Proxy	Name Proxy	Geography Proxy
Hispanic (=.1)	0.101***	0.0923***	0.0972***	0.0852***
	(0.00130)	(0.00146)	(0.00158)	(0.00227)
Black (=.3)	0.300***	0.264***	0.226***	0.269***
	(0.00161)	(0.00218)	(0.00325)	(0.00276)
Asian/Pacific Islander (=.2)	0.201***	0.196***	0.203***	0.244***
	(0.00142)	(0.00163)	(0.00187)	(0.00448)
Native (=.4)	0.398***	0.125***	0.160***	0.0788**
	(0.00740)	(0.0207)	(0.0302)	(0.0284)
Multiracial/Other	0.597***	0.0873***	0.134***	0.00793
(=.6)	(0.00531)	(0.0116)	(0.0196)	(0.0245)
Observations	519372	519372	519372	519372
Adjusted R-squared	0.111	0.049	0.031	0.026

Example 1 with OVB

	HMDA Reporting	Joint Proxy	Name Proxy	Geography Proxy
Hispanic (=.1)	-0.0704***	-0.113***	-0.104***	-0.206***
	(0.00333)	(0.00363)	(0.00388)	(0.00554)
Black (=.3)	0.134***	-0.0231***	-0.0306***	-0.105***
	(0.00413)	(0.00541)	(0.00798)	(0.00675)
Asian/Pacific Islander (=.2)	0.258***	0.268***	0.258***	0.771***
	(0.00363)	(0.00404)	(0.00460)	(0.0110)
Native (=.4)	0.302***	-0.765***	-0.184*	-1.200***
	(0.0189)	(0.0512)	(0.0743)	(0.0696)
Multiracial/Other (=.6)	0.516***	0.338***	0.521***	-0.179**
	(0.0136)	(0.0289)	(0.0482)	(0.0598)
Observations	519372	519372	519372	519372
Adjusted R-squared	0.016	0.013	0.011	0.015

OVB: Example 2 – HMDA + Data

- Use HMDA+ data to look at differences in estimates between reported race/ethnicity and proxy for processes with large amounts of potential omitted variables
- Look at regression of loan amount on race/ethnicity, along with DTI, FICO, LTV, and income

Example 2: Regression of Loan Amount

	HMDA Reporting	Joint Proxy	Name Proxy	Geography Proxy
Hispanic	-5.425***	-7.966***	-3.944***	-35.95***
	(0.791)	(0.848)	(0.908)	(1.242)
Black	-17.66***	-63.40***	-37.78***	-118.1***
	(1.021)	(1.324)	(1.818)	(1.618)
Asian/Pacific Islander	68.16***	78.82***	70.60***	324.8***
	(0.805)	(0.881)	(1.015)	(2.342)
Native	-12.17**	-313.4***	-59.06***	-568.0***
	(4.446)	(12.40)	(17.52)	(16.28)
Multiracial/Other	6.177	279.4***	213.8***	505.7***
	(3.238)	(6.730)	(11.16)	(13.26)
Observations	314880	314880	314880	314880
Adjusted R-squared	0.274	0.289	0.275	0.337

OVB: Example 3

- Can also use HMDA data to look at how the proxy performs when measuring outcomes uncorrelated with any omitted variables
- One example is, conditional on a loan qualifying as GSE-conforming, whether a loan is purchased by Fannie Mae or Freddie Mac
  - While for the most part both entities purchase loans meeting the same standards, there are some esoteric differences in loan quality and purchasing decisions across the two
- This dataset includes two sets of loans
  - One set of loans come from an underwriting program with a shaky reputation, meaning those idiosyncratic differences will matter more
  - The other comes from a program with a better reputation, meaning loans should have better overall quality both GSEs consider investment-worthy

Example 3: GSE Purchaser, Total Portfolio

	HMDA Reporting	Joint Proxy	Name Proxy	Geography Proxy
Black	-0.0164***	-0.0343***	-0.00761	-0.0489***
	(0.00490)	(0.00618)	(0.00770)	(0.00779)
Hispanic	-0.0138***	-0.0184***	-0.0125**	-0.0414***
	(0.00351)	(0.00382)	(0.00400)	(0.00567)
Native	-0.0538**	-0.0825	0.0532	-0.0883
	(0.0186)	(0.0550)	(0.0774)	(0.0712)
Asian/Pacific Islander	0.00717*	0.00506	0.00647	0.0365***
	(0.00317)	(0.00352)	(0.00397)	(0.00995)
Multiracial/Other	-0.0104	0.0192	-0.00212	0.0668
	(0.0129)	(0.0277)	(0.0449)	(0.0550)
Observations	200993	200993	200993	200993
Adjusted R-squared	0.053	0.053	0.053	0.053

Example 3: GSE Purchaser, “Quality Loans”

	HMDA Reporting	Joint Proxy	Name Proxy	Geography Proxy
Black	-0.0126	-0.0168	-0.00587	-0.0207
	(0.0107)	(0.0130)	(0.0131)	(0.0166)
Hispanic	-0.0133*	-0.0129*	-0.0109	-0.0129
	(0.00580)	(0.00621)	(0.00648)	(0.00930)
Native	-0.159***	-0.186	0.0194	-0.262
	(0.0433)	(0.123)	(0.127)	(0.160)
Asian/Pacific Islander	0.0180***	0.0167***	0.0187***	0.0365**
	(0.00435)	(0.00471)	(0.00536)	(0.0130)
Multiracial/Other	0.0157	0.0355	-0.0364	0.217*
	(0.0259)	(0.0411)	(0.0637)	(0.0891)
Observations	84136	84136	84136	84136
Adjusted R-squared	0.059	0.059	0.059	0.059

Conclusions

- Joint proxy appears to successfully improve accuracy of measuring reported race/ethnicity
- Potential pitfalls include weak statistical power for smallest groups and potential for omitted variable bias
    - This means the more "hands-on" products in terms of underwriting and pricing will likely be more difficult
    - When products have more cleanly defined processes and rules, the proxy can more accurately measure potential disparities
- Any questions/comments/suggestions greatly welcomed