Data

Comment [BES1]: Focus is on race/ethnicity. Would be good to have sex added at some point.

Surname data used to calculate the probability of belonging to a specific race/ethnicity comes from data derived from Census 2000 and released by the US Census Bureau in 2007. This release lists each name held by at least 100 enumerated individuals along with a breakdown of the proportion of individuals with that name belonging to each of the six OMB defined race and ethnicity categories, defined as: White, Black, Asian/Pacific Islander, American Indian/Alaskan Native, Hispanic, and Multiracial/Some Other Race. This classification holds Hispanic as mutually exclusive from the other categories, with any individuals identified as Hispanic belonging only to that category, regardless of racial background. Similarly to the process for racial identification used in the HMDA loan application registry, the Census relies on self-identification of both race and ethnicity when determining identity for these individuals, with an exception made for classification to the "Multiracial/Some Other Race" category. In many cases individuals identifying as "Some Other Race" then specified a Hispanic nationality (e.g., Salvadoran, Puerto Rican); in these cases the Census identified the respondent as Hispanic. Additionally, the US Census did not provide exact counts/percentages for surname/race or ethnicity combinations with nonzero values less than five. In those instances we apportioned the remaining non-assigned total population for the surname evenly across the omitted categories.[1] In total, the list provides 151,-671 surnames, covering approximately 90 percent of the population. Word (2008) provides a more detailed description of how the Bureau cleaned and developed the final list from initial responses. Information on racial/ethnic identity by geography uses census tract-level data from the more recent 2010 US Census Summary File 1.[2]

To compare the constructed race/ethnicity proxies against a reported truth, we use data from the HMDA Loan/Application Register of one of the HMDA+ reporters, with additional information on credit, loan characteristics, name, and address provided. Under rules associated with collection of this data, the provider of the loan application should offer the applicant an opportunity to voluntarily self-identify their racial or ethnic background; if the applicant does not do so the lender must provide information on the applicant to the best of their ability, with the exception of applications completed by mail, over the phone, or online. As the proportion of

Comment [BES2]: Let's move away from this word choice.

[1] Might want to include simple example as appendix
[2] http://www.census.gov/genealogy/www/data/2000surnames/surnames.pdf

online applications provided increases, the proportion of applications with race omitted will increase as well; as a result imputation of race will become more important in this area when evaluating fair lending outcomes, even though current reporting rules allow for accurate measurement.

Comment [BES3]: Don't know if this is a relevant point for this write-up. I would, however, make the point that there may be "error" in both self-reporting or reporting done by lenders.

Construction of Proxy

For race and ethnicity, the name and geography information can be combined to form a joint probability using the methodology described in Elliott, et al. [3] This methodology requires the following steps:

Comment [BES4]: As a technical note, do we consider this a "joint probability" in the technical sense?

- Generate probabilities $p(r|s)$, the probability of belonging to race or ethnicity r given surname s, and $q(g|r)$, the proportion of the population of individuals in race or ethnicity r who live in geographic area g.
- Through application of Bayes' Theorem and the law of total probability, the likelihood that an individual with surname s living in geographic area g belongs to race or ethnicity r is described by

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

- For applicants with compound surnames or a co-applicant with a different surname, the probability generated for this paper uses the first surname found on the application.[4] While this sidesteps issues surrounding the presence of multiple individuals on the application, including relationships between applicants, using only one surname results in a simpler implementation with little, if any, loss of accuracy.[5]
- For applicants with multiple addresses, only the address of the primary applicant is used for purposes of geocoding, for similar reasons.

Comment [BES5]: Might want to use the term "primary applicant" as used in subsequent bullet.

[3] Elliott et. al., "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, Sept. 2009.

[4] Might want to add a footnote mentioning that in practical terms, this means using the first name that appears in the applicant last name field in the data...since we do not actually see the contracts.

[5] I think you should site to an appendix or table that explores this sensitivity (i.e., this is an analysis that should be performed)

To maintain statistical validity of this process, only one assumption is required: That the direct relationship between race and geography does not depend on surname. At a practical level, violation of this assumption requires that individuals with specific surname/race combinations tend to live in certain geographic areas that have a different racial/ethnic distribution than other individuals of the same race. For example, if African Americans with the last name Jones preferred to live in a certain neighborhood more than both African Americans in general and all people with the last name Jones, this would skew the results. The difficulty in coming up with an example that would meet both conditions demonstrates the specificity of this assumption.

Refinements to Proxy

In the census surname data, the bureau suppressed exact counts for racial/ethnic categories with 2-5 occurrences for a given name. Similarly to Elliott et. al., in these cases we distribute the sum of the suppressed counts for each surname evenly across all categories with missing nonzero counts. For applicants with multiple surnames, only one name was used to calculate a probability, sacrificing precision and paying the cost of measurement error in order to avoid the potentially contentious issue of how to infer the relationship across last names. In those multiple surname cases, the first surname to appear was matched against the surname list; if this name appeared in the list its distribution was used; otherwise if it appeared the second surname provided the race/ethnicity distribution used in estimation.

Comment [BES6]: You mention this in the opening section.

Comment [BES7]: Consider rewording this.

Measurement Concerns

When evaluating the impact of use of proxies in place of reported race/ethnicity for purposes of estimating impacts on outcomes, ~~among other considerations the roles of~~ four general concerns arise: classification error, omitted variable bias, non-mean-preserving distributions of proxies, and lack of a one-to-one relation between reported race and the proxy.[6]

Classification Error

[6] I would footnote other considerations as necessary.

Given general concerns across the economics and ~~statistical~~ statistics literature regarding the role of classical, mean-zero measurement error, one might worry that the use of a constructed proxy would result in error that would mismeasure the true disparity. However, under the basic assumptions used to construct the Bayesian proxy $\hat{r}_{isg}$, reconsider it as the mean probability of belonging to race r for individuals with last name s in geography g, $\hat{r}_{isg} = \bar{r}_{sg}$. In expectation, we should expect the reported race to have the form

$$r_{isg} = \bar{r}_{sg} + \nu_{isg} \ (1)$$

, with $(\nu, \bar{r}) = 0$. This assumption states that the measurement error in estimating race has no systematic relationship with the proxy itself.

Comment [BES8]: Again, let's be conscious of the use of this word. In a theoretical framework, we care about comparison to the truth. In practice, the reported values we observe may, themselves, be impacted by measurement error.

Comment [BES9]: Not sure what this is, an orthogonality condition? I guess it is the notation.

In terms of using the proxy to evaluate outcomes, we are interested in estimating

$$y_{isg} = \beta r_{isg} + \varepsilon_{isg} \ (2)$$

but since we do not know the race, must instead evaluate

$$y_{isg} = \tilde{\beta}\bar{r}_{sg} + u_{isg} \ (3)$$

. This will yield the coefficient estimate

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

under the additional assumption that $cov(\varepsilon, \bar{r}) = 0$.

Omitted Variables

The consistent estimate shown above relies on a lack of omitted variables across three estimating equations; relaxation of these assumptions will have consequences. As an example, consider equation (2) in the presence of omitted variables; e.g., $cov(r, \varepsilon) \neq 0$. In this case, regressing y on r results in a parameter estimate of

$$\hat{\beta} = \beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2} \ (4)$$

. Here, β represents the true direct impact of race on outcomes, what would be thought of in the legal literature as disparate treatment. $\hat{\beta}$, meanwhile, represents the true direct impact of race on outcomes, as well as the indirect relationship race has with outcomes based on its correlation with other factors that also influence the outcome. The additional amount beyond the direct

relationship is also known as disparate impact. The impact term in the above equation can also be expressed in terms of the race proxy, as

$$\hat{\beta}_r = \beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2} = \beta + \frac{cov(\bar{r} + \nu, \varepsilon)}{var(\bar{r} + \nu, \bar{r} + \nu)} = \beta_r + \frac{\sigma_{\bar{r}\varepsilon} + \sigma_{\nu\varepsilon}}{\sigma_{\bar{r}}^2 + 2\sigma_{\bar{r}\nu} + \sigma_\nu^2} \quad (5)$$

When $cov(r, \varepsilon) \neq 0$ running the regression in (3) results in an estimate of

$$\tilde{\beta}_{\bar{r}} = \frac{cov\left(\beta\left(\bar{r}_{sg} + \nu_{isg}\right) + \varepsilon_{isg}, \bar{r}_{sg}\right)}{var\left(\bar{r}_{sg}\right)} = \frac{\beta\sigma_{\bar{r}}^2 + \beta\sigma_{\bar{r}\nu} + \sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2} = \beta_r + \frac{\beta_r\sigma_{\bar{r}\nu} + \sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2} \quad (6)$$

Comparing equations (4) through (6), while not straightforward, allows us to draw conclusions about the consequences of violating the initial assumptions about the presence of omitted variables both in the true estimation and in the proxy equation. First, examination of the denominators of (5) and (6) and the dichotomous nature of r versus the continuous nature of its proxy provides evidence that the denominator of the bias for the reported race should be larger than that of the proxy, increasing the size of the proxy's estimated bias/disparate impact. Second, differences between covariances of the omitted variables and the reported race when estimating outcomes and those same variables and the proxy (i.e., $\sigma_{r\varepsilon} \neq \sigma_{\bar{r}\varepsilon}$) will impact the difference between the estimates of outcomes generated by regressions using these variables. Third, the amount of estimate bias in outcomes from the proxy relative to the use of reported race will be impacted by the amount of omitted variable bias present in the estimation of actual race. For example, if income predicts race and is correlated with the proxy, but the information it provides regarding race is not included in the proxy, then this will bias the estimate relative to the result in equation (4). With that said, the size of that bias is mitigated by the variance of the proxy, as well as any disparate treatment. If no disparate treatment exists, then this term would cancel out and not create any difference in disparate treatment estimates between the reported race and the proxy. With that said, as the precision of the proxy estimate increases, this will decrease the covariance terms associated with ν and the variance of the proxy, minimizing the issues listed above and resulting in outcome estimates generated from the proxy approaching those generated directly using reported race.

Lack of One-to-One Relation between Reported Race and Proxy

The above two sections, via equation (1), made the implicit assumption of a coefficient of 1 on the term for the proxy when estimating reported race. If this assumption does not hold, estimation bias results. Specifically, if we reframe equation (1) as

$$r_{isg} = \gamma \bar{r}_{sg} + \nu_{isg} \ (1a)$$

equation (6) becomes

$$\tilde{\beta}_{\bar{r}} = \gamma \beta_r + \frac{\beta \sigma_{\bar{r}\nu} + \sigma_{\bar{r}\varepsilon}}{\sigma^2_{\bar{r}}} \ (6a)$$

. Here we see the impact of the coefficient from (1a) on measuring outcomes depends on its magnitude. While in the estimation results section we will attempt to infer the impact of omitted variables, we can directly estimate $\hat{\gamma}$ and show how this source of bias will impact the results.

Results

Table 1 displays the reported race/ethnicity distribution of our sample data, along with the comparable distributions generated by the joint, name, and geography proxy, respectively. Due to the large sample size of our data set, all differences in percentages of race/ethnicity across distributions are statistically significant at the 1 percent level. In the context of the notation in the previous section, this means that error in our measurement of race is currently nonrandom, and that other covariates could help identify race with further precision. Ignoring the potential for selection, however, overestimating the size of the treatment groups in general should lead to underestimates of the true disparity size, since some individuals in the control group will instead count toward the estimated treatment coefficient. Additionally, a noteworthy pattern emerges from examination of the table; namely, for five of the six classifications the joint proxy comes closest to matching the reported data, with the only exception being the catch-all "Multiracial/Other" category.