Race, Writing, Computation

Amanda Su

4/29/2020

Contents

0.1 Abstract 1
0.2 Introduction 1
0.3 Literature Review and Paper Review 2
0.4 Replication 2
0.5 Extension 3
0.6 References 5
Appendix 6

0.1 Abstract

So, Long, and Zhu (2019) determine that novelists marked as "white" versus "black" produce different narratological effects with respect to the interaction of race and religious authority, finding that black writers who cite the Bible are more likely to cite it in a social context compared to white writers who cite the Bible in their novels. I was able to successfully replicate the results of the authors' paper. For my extension, I decided to reconstruct the paper's primary model using a Bayesian approach. I found that the results of the model were largely the same the as that of the original, proving that the original results are even more robust than the authors initially claimed. This corroborates and strengthens the author's conclusions about how race and writing intersect across more than a century of U.S. fiction.

0.2 Introduction

To test their hypothesis, So, Long, and Zhu (2019) drew from a larger corpus that was constructed from a list of the most frequently held novels by American authors published between 1880 and 2000 as catalogued by WorldCat. The authors narrowed down the original 6,000 authors represented in the corpus to only those novels written by authors with marked racial identities, labeling the authors by gender and/or only if the author identified in one particular way or if their identity was documented in the scholarly record. The authors then selected novels written by authors who identified as "black" or "African-American" to represent their "corpus of novels by black authors" and created a parallel corpus of "white" writers, which far outnumber black writers in the larger corpus, by selecting works that similarly skewed canonical. They limited this corpus to canonical writers because otherwise our comparison of "white" and "black" writers would be an anarchic comparison of distinguished black writers against a sea of high and low white writers of all genres. To extract the "canonical" writers, they assembled a list of authors from the Norton Anthology of American Literature (2003) and Harold Bloom's The Western Canon whom they could identify as "white." The authors then use a sequence alignment method to identify quotations of repetitions of specific lines and phrases to determine textual commonality between texts. Throughout this process, the authors acknowledge several biases in their methods. In selecting the corpus, they omit African-American novels which are not traditionally marked as "novelistic" to maintain the corpus's canonical skew. They also recognize that their crude, provisional identification of authors' racial identities is complicated by the particular exigencies of shifting social and historical circumstance and may have not have any implication on novels written under the sign of such identities. To test their theory about novelists of different races producing different narratological

effects in their works with respect to the Bible, the authors constructed a mixed model that explains whether or not a text is "social" as a function of the author's gender, race, whether or not they cited the Bible as a control variable, and the interaction of the race and bible variables, also accounting for the random effect of a single novel. Their results conclude black writers who cite the Bible are more likely to cite it in a social context compared to white writers who cite the Bible in their novels.

I was able to replicate all results found by So, Long, and Zhu (2019). The authors generously made their data available alongside their paper. CITE ORIGINAL DATA SOurCE[1]

For my extension, I decided to reconstruct the paper's primary model using a Bayesian approach. While the authors used the glmer function to fit a generalized mixed-effects model, I instead use stan_glmer to fit a Bayesian generalized linear mixed effects model with group-specific terms. The Bayesian model addsp riors on the regression coefficients and priors on the terms of a decomposition of the covariance matrices of the group-specific parameters. I found that the results of my model were largely the same the as that of the original, proving that the original results are even more robust than the authors initially claimed. This corroborates and strengthens the author's conclusions about how race and writing intersect across more than a century of U.S. fiction.

0.3 Literature Review and Paper Review

Can computational methods tell us anything new and interesting about how racial difference is expressed in literature? Do authors of different racial identifications (for example, "white" versus "black") consistently use different patterns of language, style, and narrative, and if so, what are these patterns? Do they remain stable or change over time?

This scholarly project bridges two scholarly fields historically seen as incompatible: cultural analytics (also known as "computational criticism") and critical race studies. It does so by discovering generative points of contact between data science and critique, two sets of methods typically viewed as antithetical. Cultural analytics is an emerging field wherein humanist scholars leverage the in- creasing availability of large digital materials and the affordances of new computa- tional tools. This allows them to study, for example, semantic and narratological patterns in the English-language novel at the scale of centuries and across tens-of-thousands of texts. While cutural analytics scholars have taken on an expanding array of topics, including genre and cultural prestige, the topic of race and racial difference has remained relatively understudied. Since computational methods demand the quantification of one's objects of study, it's likely easier to accept measuring a novel's popularity by sales figures or classifying its genre by diction than labeling it according to discrete racial identifiers. Such labeling is an affront to critical race studies, the mission of which is the deconstruction of racial categories. As such, recent scholarship on the relationship between computation and race has been critique-oriented. Scholars of science and technology, such as Cathy O'Neil and Safiya A. Noble, have documented how computational algorithms used by banks and online search engines intensify racial stratification and oppression by articulating racial minorities as fixed, quantified types that reinforce existing patterns of social inequality. Tara McPherson has shown that the history of modern computation is deeply intertwined with the history of racial formation in the US since the 1960s. The authors of this paper uses a computational model to study race and literature in order to determine both the model's affordances and its inadequacies. I make use of Jarrett (2007), Stepto (2001), Spillers (2003), and Earhart (2015).

TO BE EDITED AND COMPLETED

0.4 Replication

To test their theory about novelists of different races producing different narratological effects in their works with respect to the Bible, the authors constructed a model that explains whether or not a text is "social" as a function of the author's gender, race, whether or not they cited the Bible as a control variable, the interaction of the race and bible variables, and the random effect for each novel.

[1] I used R citation() to complete my replication, which is publically accessible in my Github repository.

I was able to successfully replicate every aspect of the paper.

0.5 Extension

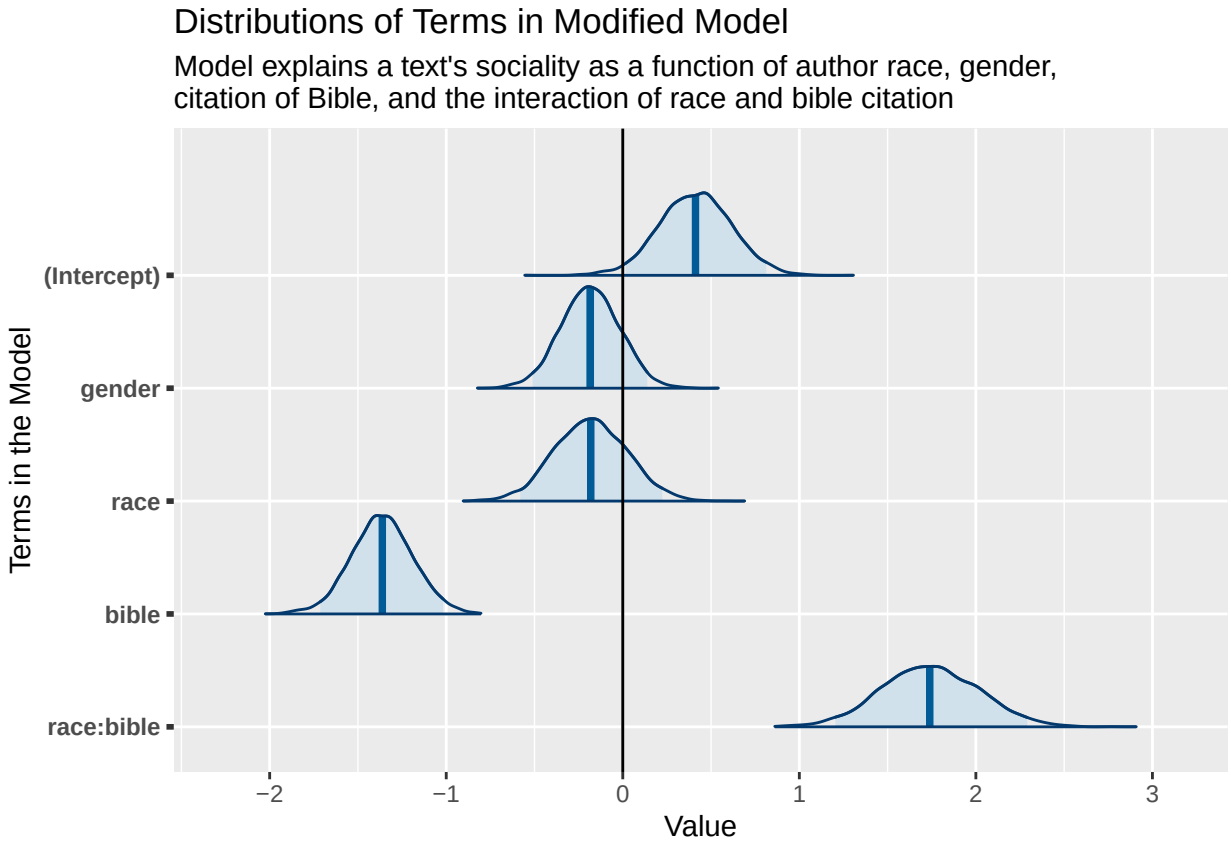
Table 1: Effect of Author Gender, Race, Bible Citation, and Race and Bible's Interaction on the Sociality of a Text

Statistic	Mean	St. Dev.
(Intercept)	0.4121839	0.2037830
gender	-0.1844503	0.1656071
race	-0.1815309	0.2088270
bible	-1.3621587	0.1718259
race:bible	1.7389784	0.2801020

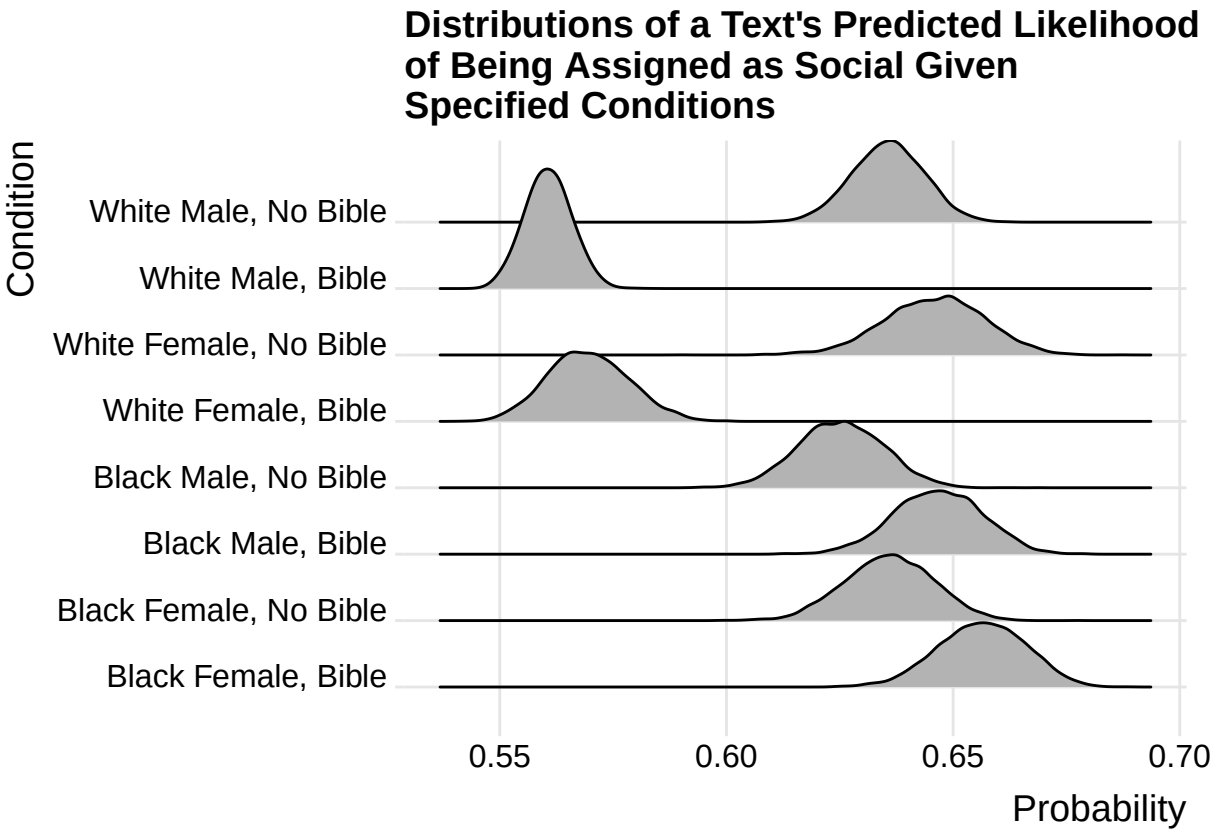
Table 1: Effect of Author Gender, Race, Bible Citation, Race and Bible's Interaction on the Sociality of a Text and the Random Effect of Single Novels

Statistic	Mean	St. Dev.
(Intercept)	0.6200658	0.3123184
gender	-0.3992170	0.2715278
race	-0.2416955	0.2908725
bible	-1.5138484	0.2522027
race:bible	1.9552393	0.3904437

Graphic 1: Distribution of Coefficients on Author Gender, Race, Bible Citation, and the Interaction of the Latter Two



Graphic 2: Distribution of Predicted Likelihoods of a Text Being Social Given An Author's Race, Gender, and Citation of the Bible in Their Work



Whereas So, Long, and Zhu (2019) decided to perform a maximum likelihood estimation of generalized linear models to determine the predicted values of model coefficients, I perform a full Bayesian estimation full Bayesian estimation to find the average expected values for coefficients. King, Tomz, and Wittenberg (2000) wrote that expected value averages are preferrable to predicted values because the latter contains both fundamental and estimation uncertainty, whereas the former only has to account for the estimation uncertainty caused by not having an infinite number of observations. As a result, predicted values have a larger variance than expected values. I ultimately found that the primary results of the authors' paper are largely unchanged even when using a Bayesian approach to create the model.

0.6 References

Earhart, Amy. 2015. *Traces of the Old, Uses of the New.* Ann Arbor, MI: University of Michigan Press.

Jarrett, Gene Andrew. 2007. "Addition by Subtraction: Toward a Literary History of Racial Representation." Legacy.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" 44 (2). American Journal of Political Science: 347–61.

So, Richard Jean, Hoyt Long, and Yuancheng Zhu. 2019. "Race, Writing, and Computation: Racial Difference and the Us Novel, 1880-2000." Journal of Cultural Analytics.

Spillers, Hortense. 2003. *Black, White, and in Color: Essays on American Literature and Culture.* Chicago: University of Chicago Press.

Stepto, Robert. 2001. *From Behind the Veil: A Study of Afro-American Narrative.* Urbana, IL: University of Illinois Press.

Appendix

Results from So, Long, and Zhu (2019) were successfully replicated.[2]

include_graphics("original-paper/Figures/Fig3.png")

[2]All analysis for this paper is available at my Github repository.