Race, Writing, Computation

Amanda Su

4/29/2020

Contents

0.1 Abstract

So, Long, and Zhu (2019a) 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 (2019a) drew from a larger corpus 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 only if they identified in one particular way or if their identity was documented in the scholarly record. So, Long, and Zhu (2019a) 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. 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 because otherwise their 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. So, Long, and Zhu (2019a) 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 (2019a). The authors generously made their data available alongside their paper at Harvard Dataverse.[1] I used R[2] to complete my replication, which is publically accessible in my Github repository.[3]

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 from the rstanarm package.[4] The Bayesian model adds priors 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 So, Long, and Zhu (2019a) initially claimed. This corroborates and strengthens their conclusions about how race and writing intersect across more than a century of U.S. fiction.

0.3 Literature Review and Paper Review

So, Long, and Zhu (2019a) bridge two scholarly fields historically seen as incompatible: cultural analytics (also known as "computational criticism") and critical race studies. Cultural analytics is an emerging field wherein humanist scholars leverage the increasing availability of large digital materials and the affordances of new computational tools, allowing them to survey semantic and narratological patterns in the English-language novel at the scale of centuries and across tens-of-thousands of texts. Cultural analytics scholars have explored an expanding array of topics, including genre and cultural prestige, but the topic of race and racial difference has remained relatively understudied. While recent scholarship on the relationship between computation and race has been critique-oriented, pointing to computation's role in intensifying racial stratification and reinforcing existing patterns of social inequality[5], So, Long, and Zhu (2019a) seek to determine both computation's affordances and its inadequacies in the study of race and literature. They specifically research whether computational methods can reveal if and how racial difference is expressed in literature — if authors of different racial identifications consistently use different patterns of language, style, and narrative, and if so, what are these patterns? Do they remain stable or change over time?

Generations of traditional literary and religious studies scholars have pointed to the Bible as the basis for the Western cultural imaginary. Traditional scholars of literature and the Bible have long argued that even as the world of the novel has increasingly secularized, its commitment to religious ideas and language has persisted. Canonical literary scholar Northrop Frye conceived the idea of The Great Code to declare the Bible's universal commonality and significance.[6] A more modern inflection of this trope — "virality" — additionally explains the allure of the language of the Bible.[7] Indeed, the Bible possesses a distinct "resonance" that attracts both white and black writers.

So, Long, and Zhu (2019a) seek to unsettle the narrative of the Great Code by excavating differences in how black and white novelists quote the Bible in their works. Their early attempts, however, were unfruitful. First, they assessed whether one group cited the Bible more frequently than the other by randomizing the race labels in their dataset and pulling from them a null distribution of quotation counts. They found that the actual amount of Bible quotation by each group was not significantly different from this null distribution. That is, had we assigned the race labels randomly, we could have expected the same rates of quotation. They

[1] So, Long, and Zhu (2019b)
[2] R Foundation" (2020)
[3] Github repository.
[4] Gabry and Goodrich (2020)
[5] O'Neil (2016) and Noble (2018)
[6] Frye (1982)
[7] Prickett (1996)

then tested whether or not white and black writers cite the Bible at different rates over time, finding that black writers did not explicitly cite the Bible more or less than white writers at any point in time. The authors then examined if novelists were citing different parts of the Bible, noting differences between the ideological orientation of the Old and New Testaments. However, after they analyzed whether chapters from either were being cited at different rates, the results were inconclusive. Then, they looked to the words surrounding the aligned Bible passages but could not conclude that white and black writers, as a whole, used a different vocabulary when invoking the Bible or discussed different topics.

So, Long, and Zhu (2019a) then readjusted their approach by examining whether writers differed in how they cited the bible by looking for moments of sociality, defined as the presence of two or more characters engaged in a dialogue or interaction. While not denying the centrality of the Bible in black communities and its frequency of citation among black writers, black studies scholars argue black writers' inovcation of the Bible usually occurs through a process of "critical modification and revision[8]. This process takes several forms: irony, criticism, and dialogism. The latter refers to the Bible's mention inciting debate or dialogue by a cohort of characters, including potentially the narrator, rather than occuring as a monologic polemic or sermon. Scholars argue that the Bible's appearance and quotation in novels by black authors tend to be very dialogic and interactive all in an effort to question the Bible's normative or hegemonic "white" meaning. This understanding ultimately informed the creation of the final model in So, Long, and Zhu (2019a) explaining the sociality of a text as a function of the novelist's race.

0.4 Replication

To test their theory about novelists of different races producing different narratological effects in their works with respect to the Bible, So, Long, and Zhu (2019a) 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.

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

0.5 Extension

Table 1: Mixed Model Explaining the Fixed Effects 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)	0.6291125	0.3060727
gender	-0.3897267	0.2738443
race	-0.2365954	0.2876978
bible	-1.5220742	0.2439998
race:bible	1.9580150	0.4026135

I fit a mixed effects model explaining the fixed effects of an author's gender, race, their citation of the Bible, the interaction between race and Bible, and the random effects of single novels on the sociality of a text. I create a mixed model with "title" added as a random effects variable as So, Long, and Zhu (2019a) did to ensure that no single novel might be contributing a disproportionate amount of Bible quotations and become the source of any specific effect. Whereas So, Long, and Zhu (2019a) 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 to find the average expected values for coefficients. 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

[8] Valkeakari (2007)

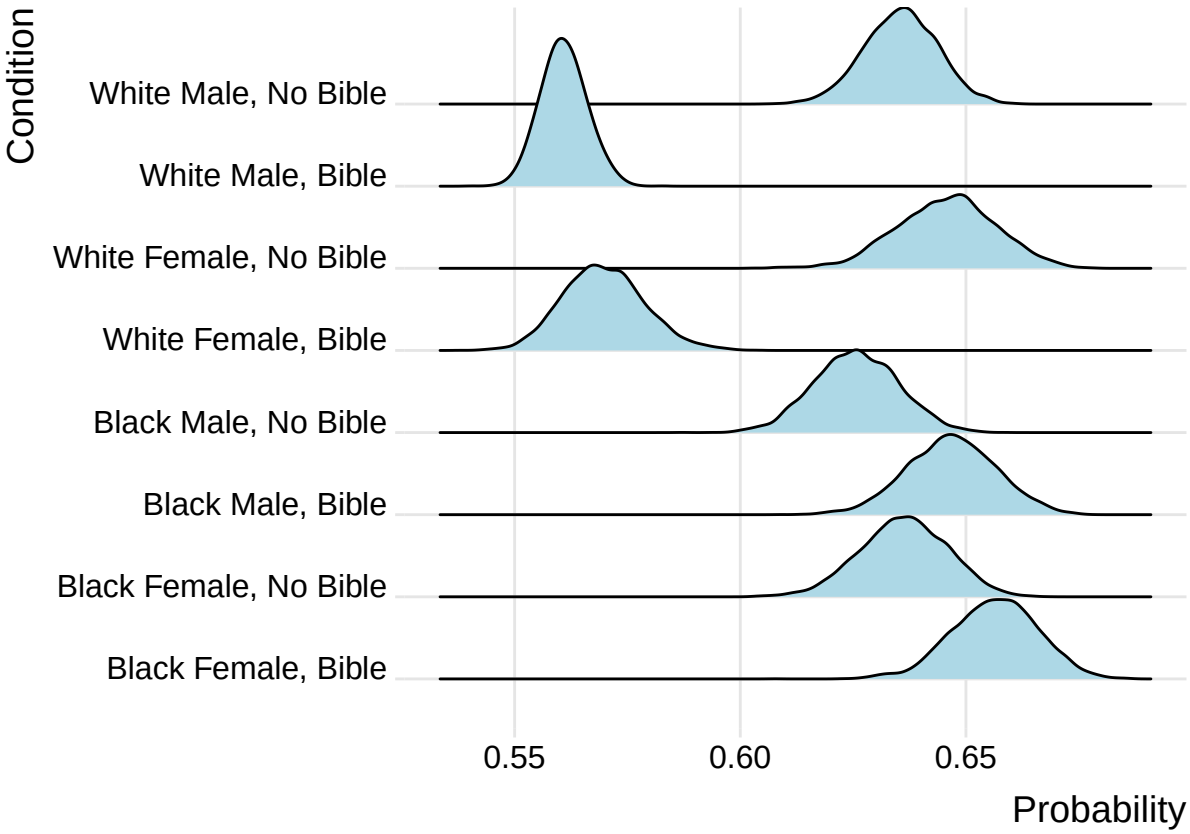
infinite number of observations.ˆ[King, Tomz, and Wittenberg (2000)} As a result, predicted values have a larger variance than expected values. I ultimately found that the primary results of the original paper are largely unchanged even when using a Bayesian approach to create the model. Both models show that an author being male, the author being black, and the author citing the Bible negatively affect the sociality of a text while the main coefficient of interest — the interaction of race and bible or the author both being black and citing the bible while controlling for gender — has a positive effect on sociality.

Table 2: Fixed Model Explaining the Effect of Author Gender, Race, Bible Citation, and Race and Bible's Interaction on the Sociality of a Text

Statistic	Mean	St. Dev.
(Intercept)	0.4119915	0.1983781
gender	-0.1830980	0.1636683
race	-0.1742306	0.2012682
bible	-1.3559118	0.1744646
race:bible	1.7272689	0.2748273

I fit an additional fixed effects model explaining the effects author's gender, race, their citation of the Bible, the interaction between race and Bible on the sociality of a text without the random effects of single novels. This allows me to construct the resulting posterior distributions using the model and find the predicted likelihoods of a text being marked as "social" given the following conditions: a white female author not citing the Bible, a white female author citing the Bible, a white male author not citing the Bible, a white male author citing the Bible, a black female author not citing the Bible, a black female author citing the Bible, a black male author not citing the Bible, and a black male author citing the Bible.

Graph 1: Distribution of Predicted Likelihoods of a Text Being Social Given An Author's Race, Gender, and Citation of the Bible in Their Work



This ridges plot visualizes the resulting posterior distributions generated from my fixed effects model. A comparison between the Black Female, Bible and Black Male, Bible ridges and the White Female, Bible, and White Male, Bible ridges reveals that the predicted likelihood of an invocation of the Bible in a text occurring in a social context is greater when it is a black author citing the Bible than when a white author is citing the Bible, as concluded by So, Long, and Zhu (2019a) and my own models.

0.6 Conclusion

The first paragraph is a review of the paper you are replicating.

The second paragraph provides more details on your replications.

The third and fourth paragraphs are more flexible. Indeed, they might be only one paragraph or they might be several. What did you do? What did you find? The final paragraph is different between the Introduction and the Conclusion. In the Introduction, it may not even exist! (We don't want to be overly didactic here. There are many ways to write a great paper.) Or it may just provide a roadmap to the rest of the paper. In the Conclusion, the last paragraph is where you get to speculate: What does it all mean? What should we research next?

0.7 References

Frye, Northrop. 1982. *The Great Code: The Bible and Literature*. Harcourt Brace Jovanovich: New York.

Gabry, Jonah, and Ben Goodrich. 2020. *Rstanarm*. https://mc-stan.org/rstanarm/index.html.

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.

Noble, Safiya. 2018. *Algorithms of Oppression*. New York: NYU Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. New York: Crown.

Prickett, Stephen. 1996. *Origins of Narrative: The Romantic Appropriation of the Bible*. Cambridge University Press: University of Florida Press.

R Foundation". 2020. *The R Project for Statistical Computing*. https://www.r-project.org/.

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

———. 2019b. "Replication Data for: Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000." Harvard Dataverse. https://doi.org/10.7910/DVN/6ANTB8.

Valkeakari, Tuire. 2007. *Religious Idiom and the African American Novel, 1952-1998*. Gainsville, FL: University of Florida Press.

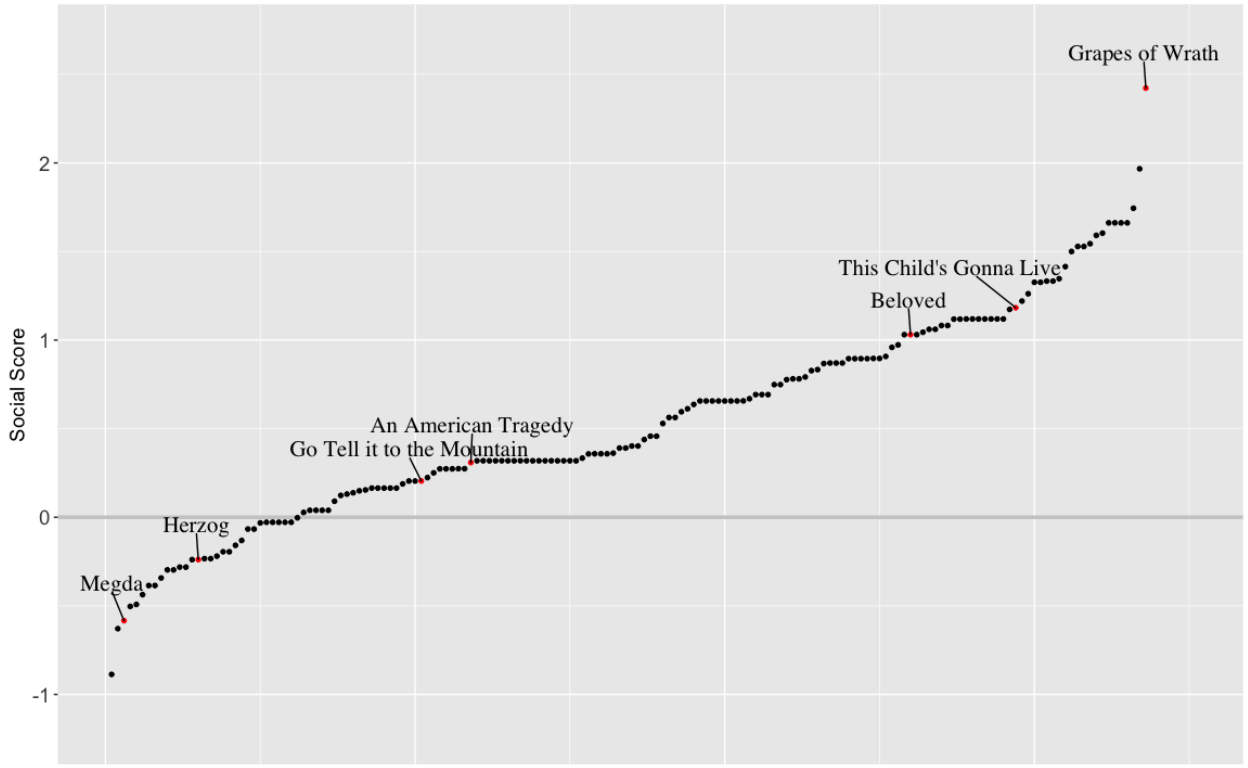


Figure 1: Original Graphic

Appendix

Results from So, Long, and Zhu (2019a) were successfully replicated. All analysis for this paper is available at my Github repository[9]

[9]Github repository.