An Analysis of the Racial Wage Gap

Zachary Bisson[a,*], Anna Elfstrum[a,*], Shawn Graven[a,*], Jack Kuivanen[a,*], Caleb Olson[a,*], Kate Soukkala[a,*]

[a] *University of Wisconsin-River Falls, 410 South Third Street, River Falls, WI, 54022*

Abstract

"More than half of whites — 55 percent — surveyed say that, generally speaking, they believe there is discrimination against white people in America today" (Gonyea, 2017). What is true and what is real are becoming harder and harder to decern. When it comes to the claims (or negations) of a wage gap existing between dominant and non-dominant races it is no different. Here, using statistical analysis tools on data from the Economic Policy Institute, we have found that a statistically significant gap does exist.

In the following sections you will find a basic introduction, literature review, theoretical analysis, empirical analysis, conclusion, and finally a list of references cited within this work. The literature review contains a broad overview of what we found in our initial research and our theoretical analysis sets up the questions and hypothesis tests that we perform in the empirical analysis. Lastly, we have our conclusion which sums up our findings clearly and concisely.

1. Introduction

Legislation has been passed and social norms challenged to help level the playing field for different races and genders. Yet, according to Williams (1987), "risk-averse employers believe and act as if black workers are on average less productive than their white counterparts; employers thus hire blacks at a wage discount or not at all." Williams (1987) goes on to say that there is a second case that "presumes blacks and white are equally productive on average, but black display a greater variance in ability; hence risk-avers employers' hiring decision could precipitate a racial wage gap." Due to this, it holds that business owners are more likely to make productivity and skill-based decisions based on race rather than incur the cost of acquiring and interpreting statistically significant data. This is witnessed by looking at historical wage data amongst seemingly

*Equal contribution, author order is alphabetical

Email addresses: zachary.bisson@my.uwrf.edu (Zachary Bisson), anna.elfstrum@my.uwrf.edu (Anna Elfstrum), shawn.graven@my.uwrf.edu (Shawn Graven), jack.kuivanen@my.uwrf.edu (Jack Kuivanen), caleb.olson@my.uwrf.edu (Caleb Olson), katelyn.soukkala@my.uwrf.edu (Kate Soukkala)

disparate groups to see how over time, wages have increased but not at the same rate. The wage gap remains.

2. Literature Review

$y = f(x)$ is a common expression of the idea that a given output is a function of all the inputs. This is a deceiving simple concept but an important one that has made research into the wage gap difficult. There are numerous published articles that try to pinpoint the reason a wage gap exists among multiple diversity categories. So many in fact that some factor-combinations yield no gap. Take for example, the work published by Black et al. (2006).

> *We find that these wage differences generally appear to be the consequence of differences in premarket factors: age, the levels and types of education, and English fluency and/or assimilation. In particular, among college-educated men who speak English at home, our estimated wage gaps are very close to 0 for Hispanic and Asian men. Similarly, the unexplained wage gap is approximately 0 for black men with college-educated parents not born in the South. We provide fragmentary evidence that the unexplained gap for other black men - Southern-born men and those born elsewhere to poorly educated parents - is related to the generally poor quality of education afforded these men at the precollege and college levels.*

Which is in direct contrast to Matt Huffman's work (2004) where he found evidence of increasing tendencies toward racial discrimination as the job stakes are raised (high-status jobs).

The majority of published works we reviewed found evidence to support the claim that a wage gap exists between underrepresented populations and the dominate population. Studies ranged from scopes as broad as the work of Oliver and Shapiro in 2006 that looked at total debt-to-asset ratios to scopes as narrow as the work of Broyles and Fenner (2010) looking specifically at the field of STEM. What it comes down to is while there is a lot of information published, there are not many works reproducing or confirming those results. We set out to find the answer ourselves. We found that despite all the disparate studies, the overwhelming find is the gap exists. Our findings also confirm this.

3. Theoretical Analysis

3.1. Hyppothesises

3.1.1. White

$$H_0 : WhiteIncome \propto AllIncome$$

Our null hypothesis is that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of white individuals who are 16 and older in the United States.

$$H_A : WhiteIncome \not\propto AllIncome$$

3.1.2. Black

Our alternate hypothesis is that the median income of white individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older.

$$H_0 : BlackIncome \propto AllIncome$$

Our null hypothesis is that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of black individuals who are 16 and older in the United States.

$$H_A : BlackIncome \not\propto AllIncome$$

3.1.3. Hispanic

Our alternate hypothesis is that the median income of hispanic individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older.

$$H_0 : HispanicIncome \propto AllIncome$$

Our null hypothesis is that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of hispanic individuals who are 16 and older in the United States.

$$H_A : HispanicIncome \not\propto AllIncome$$

Our alternate hypothesis is that the median income of black individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older.

3.2. Models

Our model is simple, `mod1` predicting `Median` only depends on the `Date`. `mod2` preforms the $\log(Median)$ as it makes the data more linear as monetary values tends to fit log regressions much better than linear ones. These will represent *AllIncome*. `modBlack` includes the `Black` factor and will represent the Black population and non-black's. The same is done with `modWhite` and `modHispanic` with the white's and the hispanics respectively.

$$mod1 : Median = \beta_1 Date + \beta_0 + e$$
$$mod2 : \log(Median) = \beta_1 Date + \beta_0 + e$$
$$modWhite : \log(Median) = \beta_1 Date * White + \beta_0 + e$$
$$modBlack : \log(Median) = \beta_1 Date * Black + \beta_0 + e$$
$$modBlack : \log(Median) = \beta_1 Date * Hispanic + \beta_0 + e$$

4. Empirical Analysis

4.1. Data

4.2. About the Data

We have two sources of data, one from U.S. Bureau of Labor Statistics (BLS) and the majority of data from Economic Policy Institute (EPI).

BLS maintains a data set called `cpsaat`, this data summaries the wage earnings per type of job, based on race and gender. To access the data in R we use a `curl_download` to retrieve the `.xlsx` file off the internet. To read the file we use the function `readxl::read_excel`.

EPI hosts a lot of data on wage statistics including, minimum wage, the participation, and earnings of each *race*, *gender*, *education* level, and much more. Due to the way EPI presents the data, it cannot be downloaded with `curl`. Instead, I have accessed the data with the package `epidata`, this simple package interfaces with EPI so that you don't have to manually download the data. EPI does not contain individual observations for wage, instead it provides 2 summarizations of the data grouped by *race*, *age*, *gender*, and *education*. This is the *Median*, 50% of people make more and 50% of people make less than this value. The other one is mean called *Average*, this is the sum of wages added up and divided by the amount.

$$\bar{x} = \frac{\sum_{i=0}^{n-1} x_i}{n}$$

To reduce the effect of the highest earners we will be using the median, like they use in the housing market as a high outlier will only add one rather than a lot more.

4.3. Clean Data

As with most data, it will have to be cleaned. This includes pivoting the tibble into a longer tibble, as it will work better for `ggplot2`. This current format is called wide format as it has many columns. To fix this we can convert it into long format, as there are many rows, with `pivot_longer`. When we do this sometimes the new column we create contains more than one value, to remedy this issue we can use `seperate` and mutate if necessary to get the values in the right column. Another inconsistency we should be aware of is that the currency values are in different years, not a large difference, but something that should be corrected.

`Minimum_wage` has data in terms of 2018, the other data is in 2019 USD. As it will be easiest and the latest data, we will be using 2019. Although small, there will be a difference and we need to adjust for inflation. The package `priceR` allows us to convert those monetary values into other ones using online inflation data.

As the data was imported with `epidata`, the column names have been changed from what the csv has. So we need to fix that to conform to consistency. For this project the names will be captained.

4.4. Methodology

4.4.1. Using the Data

After acquiring all the data, the next step was cleaning all the data. Once the data is cleaned and reorganized the next was filtering it for all the different hypothesis. Once the data was separated into the different races and age groups the data is then represented in the form of graphs demonstrating the how the different races compare to each other on a median income basis. The graphs however do not demonstrate our hypothesis well enough.

Following the graphs is chow testing. After preforming the first chow test we can reject our null hypothesis that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of white individuals who are 16 and older in the United States. This is due to the p-value being less than `.01`.

The second chow test also results in a rejection of our null hypothesis demonstrating there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of black individuals who are 16 and older in the United States. Since the p-value is less than 0.01, we can conclude that the median income of black individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older.

The final chow test results again in a rejection of the null hypothesis. This claims that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of Hispanic individuals who are 16 and older in the United States. We can then conclude that the median income of Hispanic individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older. Following the chow testing concludes our hypothesis on the data that whites make more on a median income basis than average United States residents and that black people and Hispanics also make less median income than the average United States resident.

4.4.2. Open Sorce

This document is created open source, meaning anyone can view the code, comment on it, and suggest modifications to the documents. As such, the RMarkdown used to create these documents, both HTML and PDF are available at GitHub zekrom-vale IncomeOnRace as well as all modifications in the repository. Both versions are are created with the same multi-file RMarkdown using `knitr` to knit statically with PDF or interactively with HTML. The most up to date version of the PDF and HTML versions will be available there. The GitHub repository also archives the data we used after it was cleaned in csv format.

4.4.3. HTML Version

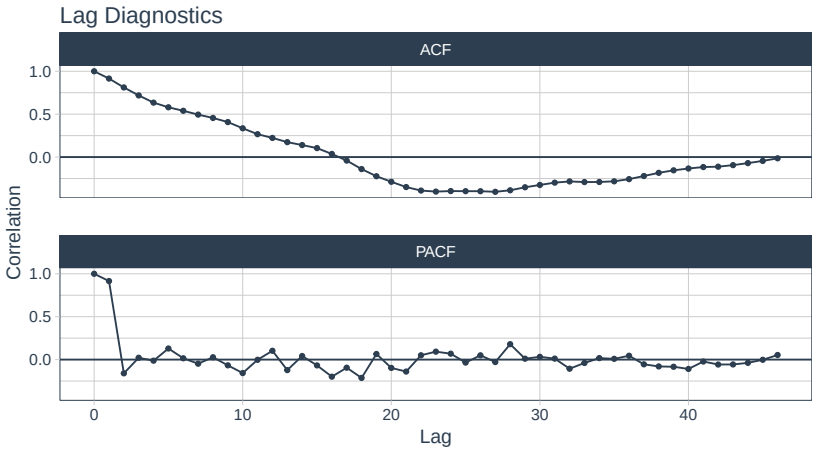
An interactive HTML document / web site is available at GitHub zekrom-vale IncomeOnRace and is recommended over the static PDF version. The

HTML version uses the package `plotly` that allows users to zoom, filter, play animations, and inspect data in graphs. The static PDF version uses `ggplot2` that does not support interactivity.

4.5. Results

4.5.1. Detect Autocorrelation

Autocorrelation is a major issue in time series as it breaks the independent observations that OLS (Ordinary Least Squares) expects. So, autocorrelation must be removed before fitting the model. This can be resolved by adding lags to the regression.



Lags in years

```
Breusch-Godfrey test for serial correlation of order up to 1

data:  mod0
LM test = 10.624, df1 = 1, df2 = 560, p-value = 0.001184
```

As you can see there is a lot of autocorrelation as indicated by the ACF graph and agreed by the `bgtest`, the larger the value the more correlated the data is to the previous value. To fix it we update the models to include a lag of the dependent variable. According to the PACF, it appears that there is only one lag that is required to fix the issue. $+\beta Median_{t-x} \forall x \in \mathbb{Z}$

These models now become:

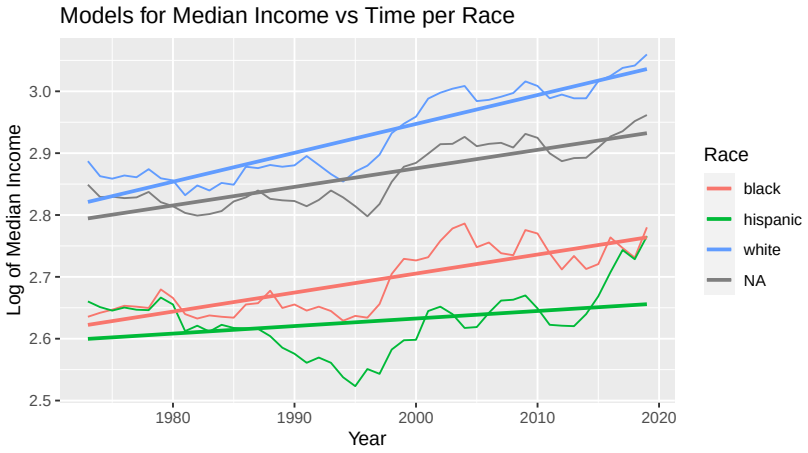
$$mod0 : Median_t = \beta_3 Date + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$$mod1 : Median = \beta_3 Date + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$$mod2 : \log(Median) = \beta_3 Date + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$$modWhite : \log(Median) = \beta_3 Date * White + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$modBlack : \log(Median) = \beta_3 Date{*}Black{+}\beta_2 Median_{t-2}{+}\beta_1 Median_{t-1}{+}\beta_0{+}e$
$modBlack : \log(Median) = \beta_3 Date{*}Hispanic{+}\beta_2 Median_{t-2}{+}\beta_1 Median_{t-1}{+}\beta_0{+}e$
Where $White$, $Black$, and $Hispanic$ are binary features based on $Race$.



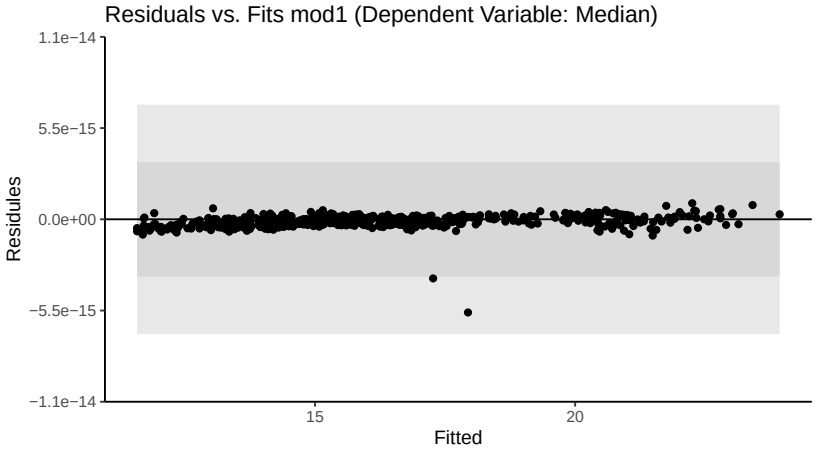
Does not incorporate lags into the visualization

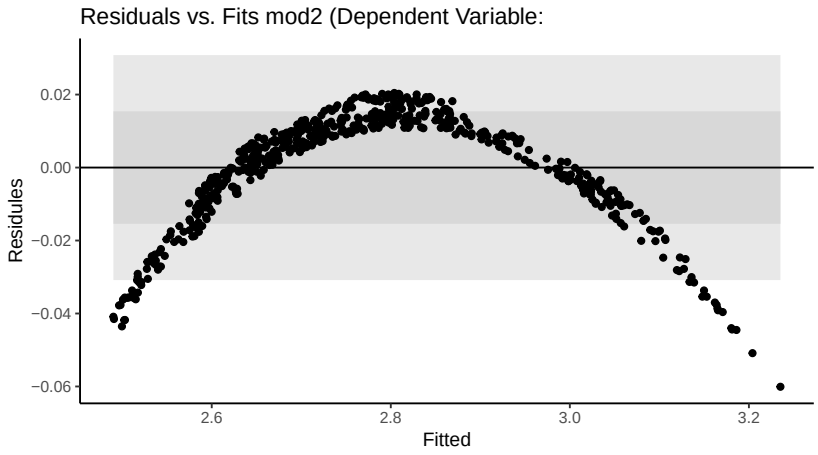
4.5.2. Model Accuracy

mod	.metric	.estimator	.estimate
mod1	rmse	standard	0.00000
mod2	rmse	standard	14.60402

According to the testing data, `mod1` is the best model at predicting as it has a lower `RMSE`.

```
Warning: Removed 1 rows containing missing values (geom_point).
```





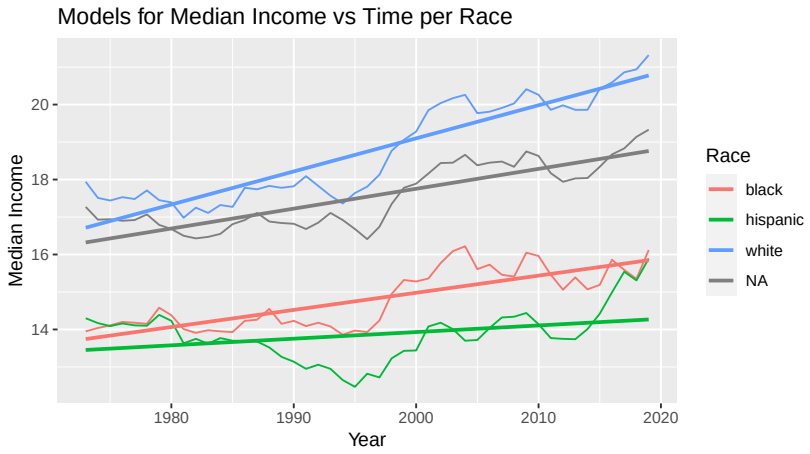
Even the Residuals vs Fitted graph agrees with `RMSE` as most of the data is gathered around 0 in the gray aria and spread out. Where `mod2` shows a curve going outside that aria. So the models will be based on `mod1`.

$$modWhite : Median = \beta_3 Date * White + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$$modBlack : Median = \beta_3 Date * Black + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

$$modBlack : Median = \beta_3 Date * Hispanic + \beta_2 Median_{t-2} + \beta_1 Median_{t-1} + \beta_0 + e$$

So, the updated model looks like this:



4.5.3. Chow Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
138	0	NA	NA	NA	NA
136	0	2	0	163.6678	0

Table 1: Model comparison, 'wage' equation

	Dependent variable:	
	Median	
	(1)	(2)
Constant	0.00*** (0.00)	0.00*** (0.00)
Date	−0.00*** (0.00)	−0.00*** (0.00)
R		−0.00** (0.00)
stats::lag(Median, -1)	1.00*** (0.00)	1.00*** (0.00)
stats::lag(Median, -2)		
Date:R		0.00** (0.00)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

After performing a chow test we can reject our null hypothesis that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of white individuals who are 16 and older in the United States, since our p-value is less than 0.01. We conclude that the median income of white individuals aged 16 and older in the United States is significantly higher than the median income of individuals aged 16 and older.

Table 2: Model comparison, 'wage' equation

	Dependent variable:	
	Median	
	(1)	(2)
Constant	0.00*** (0.00)	0.00*** (0.00)
Date	−0.00*** (0.00)	−0.00*** (0.00)
R		−0.00 (0.00)
stats::lag(Median, -1)	1.00*** (0.00)	1.00*** (0.00)
stats::lag(Median, -2)		
Date:R		0.00 (0.00)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
138	0	NA	NA	NA	NA
136	0	2	0	NA	NA

After performing another chow test we can reject our null hypothesis that

there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of black individuals who are 16 and older in the United States, since our p-value is less than 0.01. We conclude that the median income of black individuals aged 16 and older in the United States is significantly lower than the median income of individuals aged 16 and older.

Table 3: Model comparison, 'wage' equation

	Dependent variable:	
	Median	
	(1)	(2)
Constant	0.00^{***} (0.00)	-0.00^{***} (0.00)
Date	-0.00^{***} (0.00)	0.00^{***} (0.00)
R		0.00 (0.00)
stats::lag(Median, -1)	1.00^{***} (0.00)	1.00^{***} (0.00)
stats::lag(Median, -2)		
Date:R		−0.00 (0.00)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
138	0	NA	NA	NA	NA
136	0	2	0	59.52971	0

After performing a chow test we can reject our null hypothesis that there is no significant difference between the median income of individuals aged 16 and older in the United States and the median income of hispanic individuals who are 16 and older in the United States, since our p-value is less than 0.01. We conclude that the median income of white individuals aged 16 and older in the United States is significantly higher than the median income of individuals aged 16 and older.

Some limitations to the experiment are the data collection. This is because we are unable to collect everyone's income in the united states to test this. However, the data we do have gives a good representation of the income of people as we currently know it in the United States. Another major issue would be the voluntary data used. People who volunteer to give out this data may not participate due to their current financial status. This would skew the data and ultimately change the outcome.

5. Conclusion

Racism and discrimination have been an enormous issue in the United States, since the foundations of the country. The Emancipation Proclamation, the

13th, 14th, and 15th Amendments were supposed to give minority citizens equal rights, however racist lawmakers passed laws to discriminate against minority citizens. This meant that citizens of this country who were not white were not given an equal opportunity to succeed. One hundred years later, The Civil Rights Act of 1964 and the Voting Rights Act of 1965 were designed to give minority citizens equal protection under the law. Although the United States has made significant progress towards giving minority citizens an equal opportunity as white citizens, wage gap discrimination still exists and is a significant problem.

After performing a significant test, we found that white workers aged 16 and older have a significantly higher wage than the average worker in the United States. We performed two more significant tests and found that black workers and Hispanic workers aged 16 and older have a significantly lower wage than the average worker in the United States. Even with companies striving to promote racial diversity, minority workers typically do not have the level of social capital that white workers do. Minority workers tend to not have the same access to networks of higher-earning individuals as white workers do. This makes upward economic mobility incredibly difficult. Having connections to higher-earning individuals make it much easier to land a job, because the higher-earning individual can be used as a high-quality reference to an employer. If employers gave an equal opportunity to potential candidates, minority workers would be able to enter higher-earning occupations. To lower the racial wage gap, employers need to stop putting such a great importance on who the potential candidate knows and more emphasis on how the potential candidate can contribute to the future success of the company.

6. References

Black, D., Haviland, A., Sanders, S., Taylor, L., 2006. Why Do Minority Men Earn Less? A Study of Wage Differentials among the Highly Educated. Review of Economics and Statistics 88, 300–313. doi:10.1162/rest.88.2.300

Broyles, P., Fenner, W., 2010. Race, human capital, and wage discrimination in STEM professions in the United States. International Journal of Sociology and Social Policy 30, 251–266. doi:10.1108/01443331011054226

Gonyea, D., 2017. Majority of white americans say they believe whites face discrimination. NPR, last modified October 24.

Huffman, M.L., 2004. More pay, more inequality? The influence of average wage levels and the racial composition of jobs on the Black–White wage gap. Social Science Research 33, 498–520. doi:10.1016/j.ssresearch.2003.06.004

Williams, R.M., 1987. Capital, competition, and discrimination: A reconsideration of racial earnings inequality. Review of Radical Political Economics 19, 1–15.