Awesome Paper

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

This is the abstract.

It consists of two paragraphs.

---

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

2. Literature Review

(**Keating2019?**)

Here are two sample references: (Pais, 2011; **bob?**).

3. Theoretical Analysis

3.1. Hyppothesises

---

*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)

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. Import Data

4.2.1. 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, or they call 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.

4.2.2. Import cpsaat Data

Make sure we have internet and if not abort if not

`cpsaat` data is provided online at bls.gov. As it is a direct link we can download it and save it to a temporary file and process the data with `readxl::read_excel()`

4.2.3. Import EPI Data

Get the data at EPI. As there is no direct link avalable we cannot use `curl`, instead there is a package that we can use to access the data, `epidata`. This will download data in the background.

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 inconsistancy we should be aware of is that the currency values are in different years, not a large difference, but something that should be corrected.

4.3.1. Clean cpsaat11

Looks fine.

4.3.2. Clean Labor_force_participation

4.3.3. Clean Medianaverage_hourly_wages

4.3.4. Clean Minimum_wage

This data 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.

4.3.5. Fix inconsistant case

As the data was imported with `epidata`, the colum names have been changed from what the csv has. So we need to fix that to conform to consistency.

4.4. Export the data as `.csv` files

To backup our data we will export the cleaned tibbles.

```
if(!dir.exists("../data"))dir.create("../data")

cpsaat11%>%
    write_csv("../data/cpsaat11.csv")

Minimum_wage%>%
    write_csv("../data/Minimum_wage.csv")

Participation%>%
    write_csv("../data/Participation.csv")

Wages%>%
    write_csv("../data/Wages.csv")
```

4.5. Methodology

4.6. Results

4.7. Estimators

```
WagesAll=Wages%>%
    filter(is.na(Race),is.na(Gender))
WageTs=ts(WagesAll, start = min(WagesAll$Date), end = max(WagesAll$Date), frequency = 1)

acf(WageTs[, "Median"])
```

There is a lot of autocorelation so we update the models to include a lag of the dependent variable. $\beta Median_{t-1}$

These models now become:

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

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

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

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

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

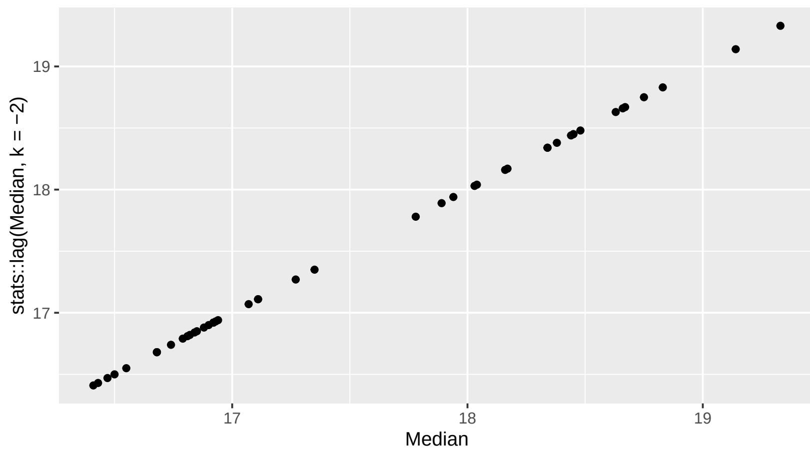
$$modBlack : \log(Median) = \beta_2 Date * Hispanic + \beta_1 Median_{t-1} + \beta_0 + e$$

Were $White$, $Black$, and $Hispanic$ are binary features based on $Race$.

```
mod0=dynlm(Median~Date+stats::lag(Median, -1), data = Wages)

bgtest(mod0, order = 1, type = "F", fill = NA)
```

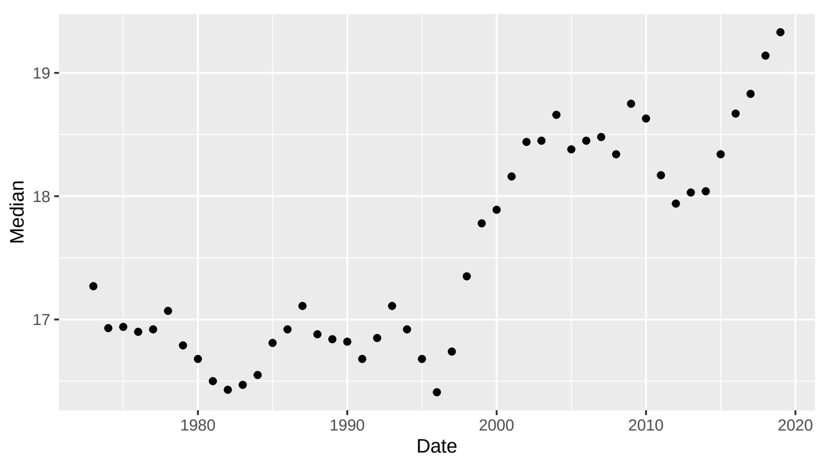
```
	Breusch-Godfrey test for serial correlation of order up to 1

data:  mod0
LM test = 2.4598, df1 = 1, df2 = 559, p-value = 0.1174
```

```
WagesAll%>%
    ggplot(aes(x=Median, y=stats::lag(Median, k=-2)))+
    geom_point()
```




```
mod1=lm(Median~Date+stats::lag(Median, -1), data = Wages)
mod2=lm(log(Median)~Date+stats::lag(Median, -1), data = Wages)
```

```
WagesAll%>%
    ggplot(aes(x=Date, y=Median))+
    geom_point()
```



```r
chow=function(racestr){
    WagesRace=Wages%>%
        mutate(R=if_else(Race==racestr, 1, 0))%>%
        filter(!is.na(Race),is.na(Gender))

    mod2=lm(log(Median)~Date,
                    data=WagesRace
    )


    modRace=lm(log(Median)~Date*R,
                    data=WagesRace
    )

    stargazer(mod2, modRace,
        header=FALSE,
        type=knittype,
      title="Model comparison, 'wage' equation",
      keep.stat="n",digits=2, single.row=TRUE,
      intercept.bottom=FALSE
    )

    anova(mod2, modRace)%>%
        kable()
}
chow("white")
```

Table 1: Model comparison, 'wage' equation

	Dependent variable:	
	log(Median)	
	(1)	(2)
Constant	-3.21^{*} (1.69)	-1.62^{**} (0.71)
Date	0.003^{***} (0.001)	0.002^{***} (0.0004)
R		-4.77^{***} (1.23)
Date:R		0.003^{***} (0.001)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
139	2.5887721	NA	NA	NA	NA
137	0.2989179	2	2.289854	524.7428	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 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.

```
chow("black")
```

Table 2: Model comparison, 'wage' equation

	Dependent variable:	
	log(Median)	
	(1)	(2)
Constant	-3.21^{*} (1.69)	-3.10 (1.99)
Date	0.003^{***} (0.001)	0.003^{***} (0.001)
R		-0.34 (3.45)
Date:R		0.0001 (0.002)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
139	2.588772	NA	NA	NA	NA
137	2.361682	2	0.2270902	6.586694	0.0018567

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.

```
chow("hispanic")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
139	2.588772	NA	NA	NA	NA
137	1.498359	2	1.090413	49.85007	0

After performing our final 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.

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.

Table 3: Model comparison, 'wage' equation

	Dependent variable:	
	log(Median)	
	(1)	(2)
Constant	-3.21^{*} (1.69)	-4.92^{***} (1.59)
Date	0.003^{***} (0.001)	0.004^{***} (0.001)
R		5.11^{*} (2.75)
Date:R		-0.003^{*} (0.001)
Observations	141	141
Note:		*p<0.1; **p<0.05; ***p<0.01

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

6. References

Pais, J., 2011. Socioeconomic background and racial earnings inequality: A propensity score analysis. Social Science Research 40, 37–49. doi:10.1016/j.ssresearch.2010.06.016

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