Regional demographic and mobility trends in Virginia

Elizabeth Goodwin

12/20/22

Table of contents

1 Introduction

The late 19th to early 20th century was likely the fastest changing time period in world history, especially in America. It moved from a pre-civil war agrarian society to something almost unrecognizable. The macro view can hide things, however. How did this change apply to individual people, and in the context of our local area? That is the primary question and goal of this paper: to help understand the historical and demographic trends in data. The primary dataset, two full count decennial census records, give us the microdata that allows us to zoom in on specific subgroups and populations. The secondary dataset, linked census record IDs, allow us to look at individual people over time and see how they individually were impacted by these changes. One dataset gives us precision, and the other gives us time.

2 Literature Review

The overall field of historical economics on inequality and mobility is quite large, especially regarding historical linked census records, so I will only be able to scratch the surface here. There are some common themes, however. For one, historical census data is both a uniquely good and bad source of data. With the microdata of the full decennial census, you get a level of person by person detail unheard of for any modern economic data. On the other hand, actually using that microdata can be much more difficult than it seems, and it records very little.

Much of the research on similar topics is on later time periods, but using similar datasets. A good example of this is Collins and Wanamaker (2014), where they use linked census datasets to estimate the effects of the great migration. In addition, there are many in this time period, but primarily talking about inter generational mobility specifically, such as Olivetti and Paserman (2015). These use another method, as Census data is already natively linked between households.

One of the most interesting papers I've read on the topic is Dupont and Rosenbloom (2018). First off, they both directly used the same year of of census data I used, and they linked with other Census data. However, that is not the main reason I find it interesting. That time period, the 1860s to be specific, was a time just after the Civil war. This caused huge economic upset in the region, but also lead to higher economic mobility more generally, even compared to the north. Despite this, it still wasn't especially mobile, with large amounts of wealth persistence. While this paper doesn't touch on mobility in specific very often, change in varioius variables over a person's life is similiar, and it would be interesting to dive deeper into other non-income related variables from the time period.

Abramitzky et al. (2021) examines the effectiveness of automatic census linking methods like those used at the Census Linking Project (Abramitzky et al. 2022). It interestingly finds that they are actually quite accurate, even when used in regression models! They do deviate from manually linked results, but not by much. The method used in this paper is one of the more robust ones as well. This will inform the rest of this paper quite a bit, as it may mean the bias is not in the method you use to link but more if they can really be linked accurately in the first place. This may also open up room for a more large scale linking procedure in the future as text and image processing improves. Baker, Blanchette, and Eriksson (2020), while at a later time period, deals pretty heavily with education in an interesting way. It takes 1940 Census records, the latest available at the time, and backlinks them to when they are children. This was then used to estimate the effects of a devastating cotton harvest on education, and actually saw an increase in educational achievement! This is likely because they were not at school, meaning.

Overall this literature review was a bit all over the place without a specific theme asides from 'related to my data'. While I thought I would be able to find more regarding the variables I mostly focused on (literacy, education estimates, etc), I mostly couldn't find it and picked some different but using similar datasets as this paper instead.

3 Data and Methods

This paper is primarily created from a combination of two sources. The first of which is the 1860 and 1910 county decennial census, provided by IPUMS (Ruggles et al. 2022). This is what it sounds like, a full digitized dataset of every recorded american in these time periods. For performance and relevancy reasons, I am only using the Virginia census count in this paper. This may lead to some limitations in linked census records, but not for unlinked records. This dataset is very good on it's own for it's pure size, although unfortunately given it's historical age the number of useful variables is relatively small.

The second main source of data is closer to an addition to the previous one than a unique dataset of it's own, but it is the core of this paper: Abramitzky et al. (2022), or the Census Linking Project. This project attempts to link entries from one census to another census, which I know firsthand is extremely difficult. The most impressive thing about this dataset is simply the pure size of it. While some of the newer released or more niche linking years were relatively small, the 1860 to 1910 linking had over 22,000 links for Virginia alone! This, of course, pales in comparison to the number of people who were in the census to begin with, but it's still very impressive and useful.

3.1 Limitations: Unlinked

The unlinked dataset is excellent data quality wise. It is as close as you can get to perfectly representative data, but with a few exceptions. The most important for this paper is simply the quality of the variables. I end up relying a lot on projected variables derived from 1950 occupational income scores. Given that that is nearly a century after the recording of this census, it should be taken with a serious grain of salt. Even within the variables that were recorded, it is often difficult to know if missing/NA variables are significant or not. For example, the 1940 census appeared to have a lot more N/A variables in general, a good example of this is in the literacy section. The percent that were outright coded as illiterate was dwarfed by the percent blank. If you interpret the blank including many literate people, that would give you drastically different conclusions than if you assumed it was majority illiterate (and therefore not recorded).

3.2 Limitaitons: Linked

Linked census records are very powerful, but can also be very dangerous. The sample size is not actually the problem, it is how they are sampled to begin with. A wealthier person is more likely to be literate, for instance. The result of that is that they are more likely to have their name correctly spelled on the census form, and be linked in the first place. There are countless examples like these that can severely bias any estimated gained from the. Within this study in particular, while this does give a lot of insight, it is absolutely not a perfectly representative sample.

3.3 Summary Statistics

In Table 3 you can see the summary statistics of some of the basic variables of the dataset. The coding of specific variables is somewhat hard to parse, however. The next sections will break down interesting findings on a variable by variable basis.

3.3.1 Literacy

One of the more interesting variables available at our disposal is the literacy variable. It is coded with five primary categories: N/A, Illiterate, read but can't write, write but can't read, and fully literate. In all the datasets the partially literally categories were very small, so I included them in literate to simplify things. As you can see in Table 4, in 1860 26.0% of black Americans in Virginia were literate, compared to 40.7% of white Americans. 56% more white people were literate than black people! In a very positive note, this seems to converge significantly by 1910, with a black rate of 51.6% to the white rate of 69.2%. In absolute terms, they both rose a pretty similar amount, but in relative terms they actually converged!

This story changes somewhat when zooming in more, however. If you zoom into specifically Williamsburg city, James city county, and york county (I will refer to this as the Williamsburg Area from now on), the 1860 difference was actually very small. There are more literate white people, but not by a particularly large amount. By 1910, however, the regional differences are generally much smaller, and it mostly matches the state wide racial trends. This would be interesting to look into at a closer level in the future, especially with regards to specific industrialization and educational history in this area.

3.3.2 Education

Another interesting variable to observe is Education. The specifics of how this works is complicated, however, as it is not a direct measure. It is an implied measure after using 1950 occupational income scoring. Essentially, it is trying to estimate the percent chance someone has >1 year of schooling by using their occupation as a proxy. In Figure 1 you can see this does cause a pretty significant problem: it is heavily driven by outliers. This makes some sense at an intuitive level as well, as only a small amount of jobs during this time period would even require it, and that is assuming that this data is accurate. You can see this in more detail in Table 1, where the median, a stat more resistant to outliers, where large mean deviations simple don't show up in the median. In fact, median results are weirdly identical between many of the rows. As an example, you see zero median location differences despite the Williamsburg area having a much more educated white population than the rest of the state.

However, in both median and mean differences you do see a change on race lines, with every measure reporting a lower result for black Americans, although the differences in mean values is much greater than the median. It seems likely that this is just the result of this trend being driven by educated elites, with nearly all of those being white.

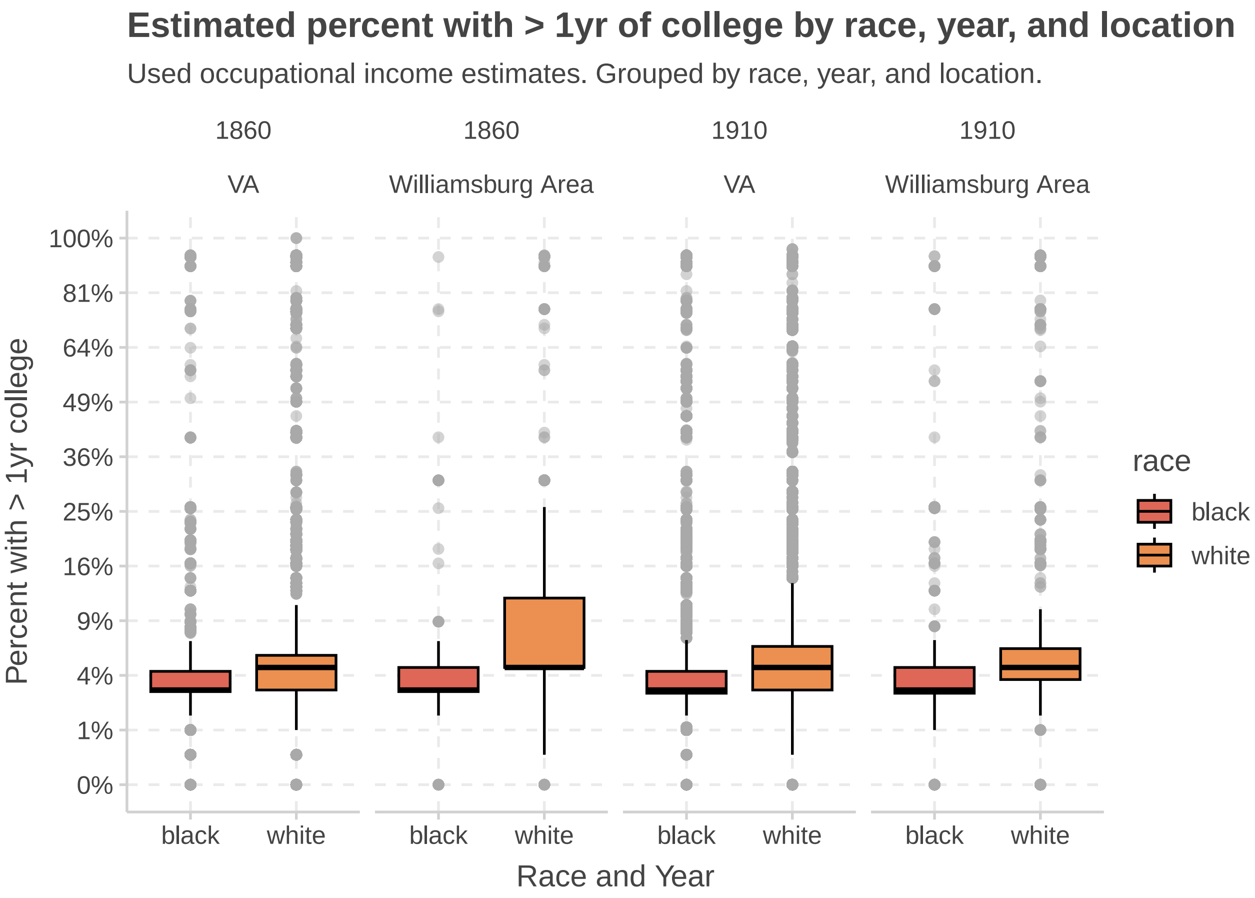

Figure 1: Educational Estimates

3.4 Income

Educational incomes were previously estimated, but those estimations used 1950s occupation categories as a proxy. This measure uses the same data, but estimating pay instead. Again, this is a very crude measure, as many of the 1950 occupations it is based on don't even exist at the time, but it is still a useful measure. As you can see in Table 2, the data becomes extremely skewed when leaving in N/A values. The median result becomes zero, so I had to remove all N/A values entirely. This likely does create problems, however. If it correlates with race or location it could bias these estimates, so take this with a grain of salt.

Table 1: Educational Estimates

year	Location	race	n	median	mean	sd
1860	VA	black	52940	3.0	4.4	7.4
1860	VA	white	690854	4.6	9.9	17.7
1860	Williamsburg Area	black	1172	3.0	4.7	7.8
1860	Williamsburg Area	white	3007	4.6	15.2	23.2
1910	VA	black	672309	3.0	4.9	9.7
1910	VA	white	1377256	4.6	10.3	17.0
1910	Williamsburg Area	black	6837	3.0	4.6	8.6
1910	Williamsburg Area	white	7258	4.6	11.3	19.3

Assuming that, however, there is plenty of interesting information to discover. First off, the mean and median are once again very skewed, but this skew seems to be correlated with race. While not universal, it does appear that the mean values are greater than median for white Americans, but less than the median for black Americans. This makes it seem likely that these are highly dependent upon the tails of the distribution. This also lead me to, for Figure 2, log the occupational income score before putting it on the figure. I did this for a pretty simple reason, as income distributions in real life follow a pretty similar trend where logged values make more sense. White occscores are higher in every single example, but particularly interesting results can be seen when looking at the time trends. I am not sure what to make of this, but the time trends on the state wide results decrease far more over time for black americans than white americans. This trend does not really seem to be the case in Williamsburg, however. This could be because it was a relatively rural and poorer area generally, so increasing racial differences were not as prevalent. This requires further study.

Table 2: Occupational Income Estimates

(a) Including N/A as 0

year	Location	race	n	median	mean	sd
1860	VA	black	52940	0	4.2	8.2
1860	VA	white	690854	0	5.0	10.4
1860	Williamsburg Area	black	1172	0	5.8	8.6
1860	Williamsburg Area	white	3007	0	5.0	11.5
1910	VA	black	672309	0	6.3	8.6
1910	VA	white	1377256	0	6.8	11.3
1910	Williamsburg Area	black	6837	0	5.6	7.9
1910	Williamsburg Area	white	7258	0	5.8	10.8

(b) Excluding N/A as 0

year	Location	race	n	median	mean	sd
1860	VA	black	13525	20	16.6	7.7
1860	VA	white	173996	14	19.9	11.5
1860	Williamsburg Area	black	438	14	15.5	6.8
1860	Williamsburg Area	white	676	16	22.2	14.2
1910	VA	black	307070	14	13.9	7.6
1910	VA	white	479755	14	19.6	10.9
1910	Williamsburg Area	black	2757	14	13.8	6.4
1910	Williamsburg Area	white	2110	16	20.1	10.7

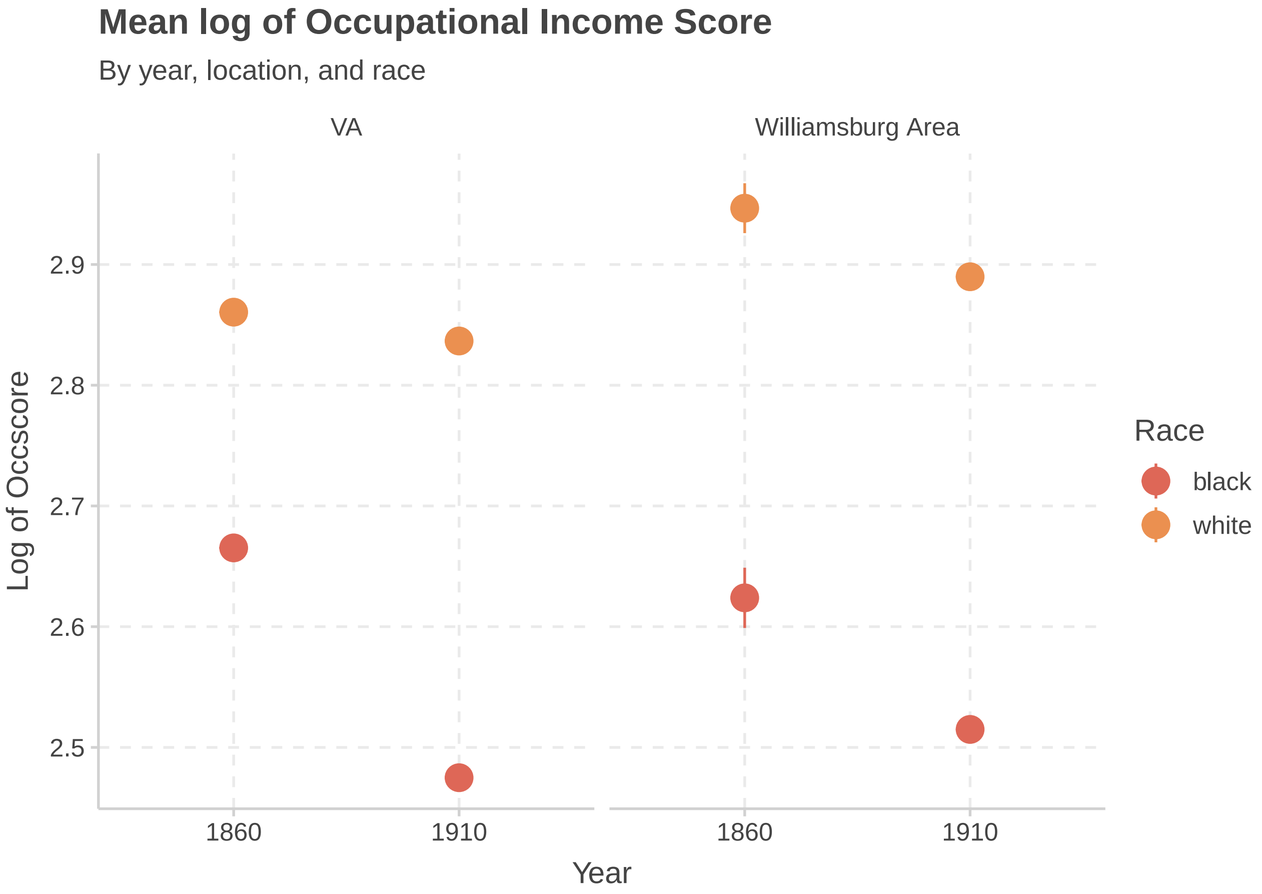

Figure 2: Occupational Income Figure

Table 3: Summary Statistics of main dataset

(a) 1860

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Size of Family	25	0	6.0	2.9	1.0	6.0	33.0
school attendance	3	0	1.1	0.4	1.0	1.0	9.0
In the labor force	3	43	0.4	0.5	0.0	0.0	1.0
number of mothers in household	9	0	1.0	0.6	0.0	1.0	8.0
number of fathers in household	6	0	0.8	0.5	0.0	1.0	5.0
Occupational Income Score	48	0	5.0	10.3	0.0	0.0	80.0
Predicted Years of College	131	74	9.6	17.2	0.0	4.6	100.0
literacy	4	0	1.7	1.9	0.0	0.0	4.0

(b) 1910

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Size of Family	27	0	5.6	2.8	1.0	5.0	47.0
school attendance	2	0	1.2	0.4	1.0	1.0	2.0
In the labor force	3	39	0.6	0.5	0.0	1.0	1.0
number of mothers in household	10	0	1.0	0.6	0.0	1.0	9.0
number of fathers in household	10	0	0.9	0.5	0.0	1.0	9.0
Occupational Income Score	51	0	6.7	10.5	0.0	0.0	80.0
Predicted Years of College	171	62	8.2	14.8	0.0	3.3	96.0
literacy	4	0	2.6	1.8	0.0	4.0	4.0

Table 4: Literacy Crosstabulations by Race and Year

		VA			Williamsburg/James/York		
year	race	Illiterate	Literate	N/A	Illiterate	Literate	N/A
1860	black	21%	25.7%	53.4%	7.7%	39.4%	52.9%
1860	white	7.7%	40.6%	51.6%	5.4%	47.3%	47.3%
1910	black	22.3%	51.6%	26.1%	19.9%	54.1%	26%
1910	white	5.5%	69.2%	25.3%	3.9%	74.7%	21.4%

3.5 Summary of Dataset

3.6 Methods

The methods of the paper are overall relatively limited because of severe historical data limitations as well as the scope of this paper. While there is not a lot specifically novel about the data or methods used in this paper, the regional focusing and use of linkings make this significantly more versatile.

4 Empirical Analysis

This section focuses on the empirical analysis and results of the linked datasets. It is broken down in a similar way as the summary statistics section, on a variable by variable basis.

4.1 Literacy: Linked

Extending previous analysis, I wanted to expand on the topic of literacy by connecting it to changes in individual people over time. Of course, this is limited by the fact that literacy has one of the most direct connections to linking in the first place. The model(s) used can be seen in in Table 5 with a selection of various explanatory variables. All models are logistic models, as it is estimating a outcome I transformed into a binary dependent variable earlier in the paper.

All the models are trying explain the linked person's literacy status in the year 1910. To begin with, model first uses the respect 1910 literacy scores using 1860 literacy scores of the same person. There is, of course, a significant result. What gets interesting is that by just adding race alone it becomesmoderately smaller, but the interaction of race and the 1860 variable is gigantic, and enough to offset the 1860 score itself. In other words, being literate in 1860 moderately increases your chances of being literate in 1910, but only if you are white.

This result is fascinating for a particular reason though. If you were literate in an earlier year, you probably didn't stop being literate in a later one! However, the fact that even just by itself, it doesn't explain all that much is interesting. Overall these regressions are thought provoking but in need of more research. This regression is also available in graph form at Figure 3. In addition, I tested various Williamsburg area specific variables but they were not significant.

Table 5: Literacy Models

	Only 1860	+race	+race*1860	1860 + loc	+loc*1860	both	all
(Intercept)	1.811***	−0.471***	−0.362***	1.203***	1.240***	0.357***	0.571***
	(0.021)	(0.051)	(0.053)	(0.315)	(0.342)	(0.045)	(0.119)
lit1860	0.265***	0.155*	−1.011***	0.266***	0.013	−0.142	−0.571*
	(0.058)	(0.062)	(0.194)	(0.058)	(0.872)	(0.116)	(0.252)
white		2.680***	2.545***			0.488***	0.238+
		(0.056)	(0.058)			(0.008)	(0.128)
lit1860 × white			1.354***			0.236***	0.762**
			(0.207)			(0.026)	(0.282)
VA				0.610+	0.574+	0.054	−0.162
				(0.316)	(0.343)	(0.045)	(0.119)
lit1860 × VA					0.254	−0.066	0.366
					(0.874)	(0.114)	(0.253)
VA × white							0.251+
							(0.128)
lit1860 × VA × white							−0.530+
							(0.284)
Num.Obs.	22 035	22 035	22 035	22 035	22 035	22 035	22 035
R2						0.162	0.162
AIC	17 499.3	15 191.3	15 143.3	17 498.0	17 499.9	11 468.5	11 467.4
BIC	17 515.3	15 215.3	15 175.3	17 522.0	17 531.9	11 524.5	11 539.4
Log.Lik.	−8747.670	−7592.644	−7567.655	−8746.008	−8745.968	−5727.235	−5724.715
F	20.877	1172.590	772.723	12.300	8.216	849.254	607.413
RMSE	0.34	0.31	0.31	0.34	0.34	0.31	0.31

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

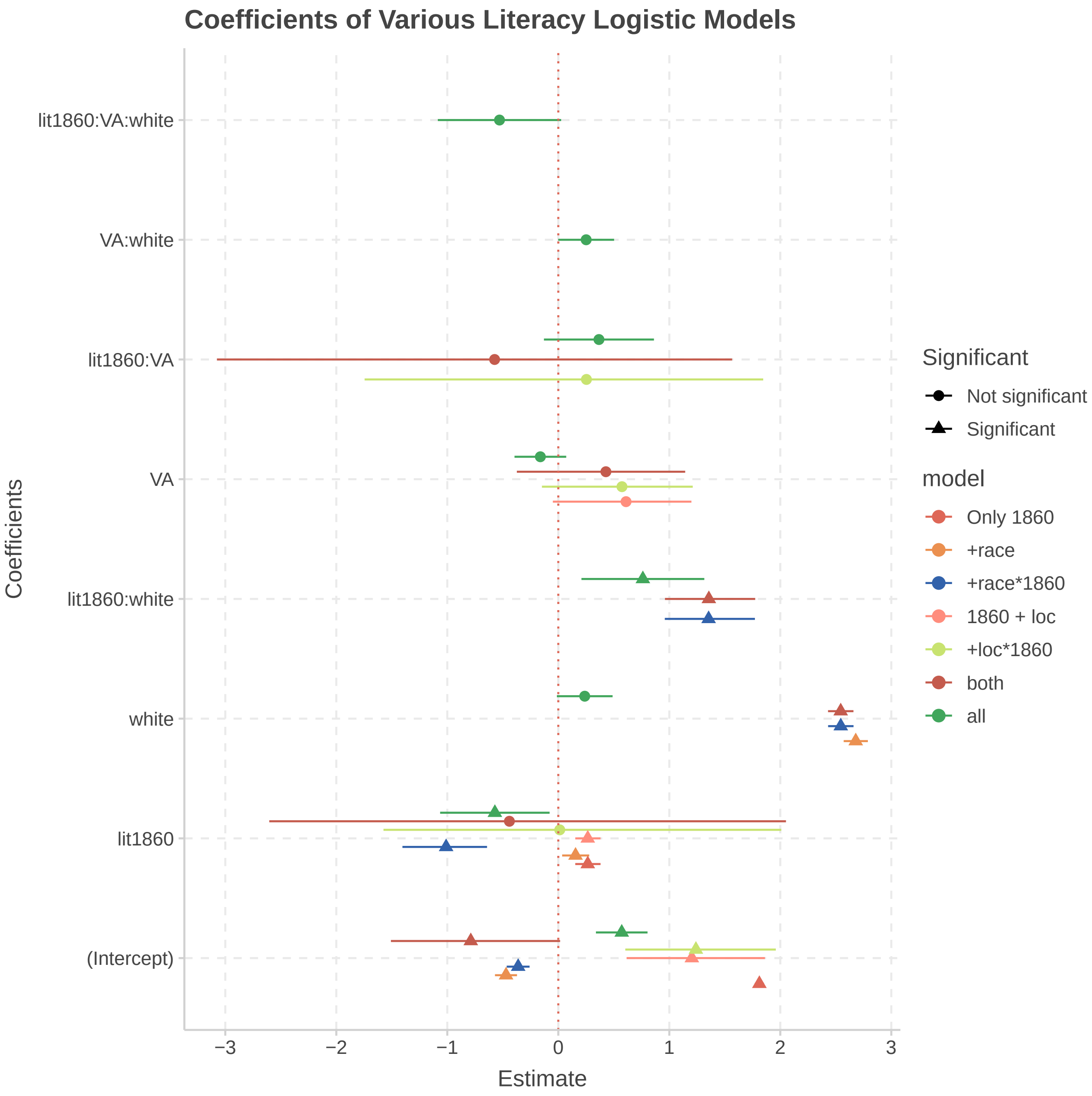

Figure 3: Literacy Model Coefficients

4.2 Education: Linked

The next variable of interest I investigated was the education score. It being a continuous estimation made it easier to estimate in a regression format, and allowed me to easily calculate pairwise differences between specific people. The dependent variable in all of these regressions is the difference between the percentage estimate that they had >1 year of col-

Table 6: Education Models

	race	location	race + location	+ interaction
(Intercept)	3.731***	4.267*	2.119	3.889
	(0.378)	(2.033)	(2.059)	(5.158)
white	2.547***		2.542***	0.448
	(0.394)		(0.394)	(5.611)
VA		1.823	1.621	−0.158
		(2.036)	(2.035)	(5.171)
white × VA				2.105
				(5.625)
Num.Obs.	22 035	22 035	22 035	22 035
R2	0.002	0.000 04	0.002	0.002
R2 Adj.	0.002	−0.000 009	0.002	0.002
AIC	183 246.1	183 287.1	183 247.4	183 249.3
BIC	183 270.1	183 311.1	183 279.4	183 289.3
Log.Lik.	−91 620.039	−91 640.549	−91 619.722	−91 619.652
F	41.857	0.802	21.246	14.210
RMSE	15.47	15.49	15.47	15.47

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

lege in 1910, subtracted by the same numbers in 1860. When just accounting for race, the intercept difference is significant. In this case, that means that the estimate went up by that many percentage points, roughly. Being white did had a positive increase in the pairwise differences, but both being positive is a good sign. A possible conclusion from this is that a tide raises all boats, but not equally. I would be careful interpreting too much from this data either way, however.

In addition to the results about race, I also included results about location data. Once again this was mostly a disappointment, with the term by itself not being significant, and no terms at all being significant in the interaction model. Given the sample sizes we are dealing with there, though, makes this not particularly surprising.

5 Conclusions

Overall, this paper is limited but was quite thought provoking to write. This topic is huge in general terms, but not that big in specific regional terms, so it definitely could be expanded upon significantly in the future. The biggest conclusion here is that historical statistics are complicated, and that the entire time I was well aware that I was in above my head. I do have some ideas for the future, however. For one, I would like to try replacing OCCSCORE with a newer variable. I know my advisor, Tate Twinam, has told me about a replacement for OCCSCORE he co-created using a regularized logistic regression and some interesting early Iowa census datasets. Lack of good variables in historical census data is one of if not the largest non-linking related issue this topic faces. This even includes simple factor variables. I could barely find anything else to group by besides race and location related variables! I would also like to compare the results of this, specifically about the Williamsburg area, to the results we linked in class. This does use an automated method, and I would be interested to see how far apart they are.

In terms of results of this paper, it has some real insights in understanding the data, but less in distinct conclusions. One of the things that stood out to me was simply the speed of change things happen in. An example of this is how outside of just race, the literacy rates in general skyrocketed over this relatively short period. I was disappointed in the lack of significant or specifically novel results about the Williamsburg area in particular. With more historical background knowledge it may have produced some more interesting results but on a surface level it does not appear to. Learning about the past in an economic and data related context is a very unique way, and this paper helps contribute to that.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59 (3): 865–918. https://doi.org/10.1257/jel.20201599.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Myera Rashid, and Santiago Pérez. 2022. "Census Linking Project: 1850-1940 Crosswalk." Harvard Dataverse. https://

doi.org/10.7910/DVN/GSMUTZ.

Baker, Richard B., John Blanchette, and Katherine Eriksson. 2020. "Long-Run Impacts of Agricultural Shocks on Educational Attainment: Evidence from the Boll Weevil." *The Journal of Economic History* 80 (1): 136–74. https://doi.org/10.1017/S0022050719000779.

Collins, William J., and Marianne H. Wanamaker. 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics* 6 (1): 220–52. https://doi.org/10.1257/app.6.1.220.

Dupont, Brandon, and Joshua L. Rosenbloom. 2018. "The Economic Origins of the Postwar Southern Elite." *Explorations in Economic History* 68 (April): 119–31. https://doi.org/10.1016/j.eeh.2017.09.002.

Olivetti, Claudia, and M. Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850 –1940." *American Economic Review* 105 (8): 2695–2724. https://doi.org/10.1257/aer.20130821.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Megan Schouweiler, and Matthew Sobek. 2022. "IPUMS USA: Version 12.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V12.0.