Data Task

Javier Arturo Bardales

2024-03-16

Description of the Task

The following is a coding task I have completed for a position's application process. There are three sections which cover different data sets and research questions pertaining to mobility.

1 Pandemic Economic Trends

In this first section, we attempt to evaluate the level of employment disparity – the difference in the level of employment between low and high-wage workers – in U.S. cities over time since the start of the COVID-19 panedmic. We evaluate two datasets: weekly city-level employment data relative to Jan 4-31 2020 from Paychex, Intuit, Earnin and Kronos and weekly city-level small-business opening and revenue data from Womply.

Before we begin our analysis, we would like to merge the small-business Womply data and employment data. To do so, we merge by date. Since the day of reporting is different, we round to the nearest Sunday (the reporting day for the the Womply data) in the employment data and merge.

The Womply data is already indexed by each Sunday.

Table 1: Distribution of Weekdays in Womply Data

weekday	n
Sunday	5777

However, the employment data is all indexed by each Friday. To merge the two data sets, the dates in both data sets must be equivalent for joining. To index by the closest Sunday, the *date* in the employment data can be offset two days forward. This indexing also makes *date* in the employment data equivalent to *date* in the Womply data (already indexed by Sunday). I perform this transformation with the lubridate package so that the days transition appropriately to the next month for Sundays at the end of the month.

Table 2: Distribution of Weekdays in Employment Data

weekday	n
Friday	6572

After indexing the employment data to the nearest Sunday, the two data sets can be joined using *date* as well as *cityid*, to ensure each observation matches to the appropriate city and time.

Note that *date* in my data is equivalent to the *week* column in the example table from the instructions.

Here is a sample of the data:

City ID	Employment	Employment Above Median	Date	Merchants Opening	···
1	0.002080	0.01020	2020-02-23	-0.004260	···
2	-0.001800	0.00375	2020-02-23	-0.000243	···
3	0.002070	0.01050	2020-02-23	-0.020600	···
4	-0.000427	0.00546	2020-02-23	-0.002630	···

1.

For 1, we would like to find the fraction of city-weeks present in both data sets.

Because the date variables have been converted to date-type, the time period can be filtered with less than or greater than operators to only observe the desired period.

To check the number of shared city-weeks between data sets, I create a variable merging the information in *date* and *cityid*. This *cityweek* variable uniquely identifies all observations in both data sets. Filtering the time for both data sets and comparing *cityweek* should provide an estimate for the number of cityweeks present in both data sets.

Using this comparison, I find that no cityweek is present in one data set and not the other, so 100% of cityweeks in the merged data are present in both data sets.

2.

For 2, we would like to: (1) calculate the difference between employment levels for workers in the top and bottom quartiles of the income distribution, termed employment disparity", (2) categorize cities as above and below the median employment disparity on March 1, 2020, and (3) plot the mean employment disparity for both types of city across time.

Employment disparity can be defined as the difference between emp_incq4 and emp_inc1, which represent employment levels for workers in the top and bottom income quartiles, respectively.

$$emp\_disp = emp\_incq4 - emp\_incq1$$

emp_disp is my variable for employment disparity. Negative values imply employees in the top income quartile have lower employment levels than those in the bottom income quartile, and positive values imply the reverse.

Since March 1, 2020 is a Sunday, applying summary transformations to employment disparity on rows with dates equal to March 1, 2020 should should be sufficient (no other transformation or date checking required).

I take the median of the national distribution of cities (which I presume is the median of all the cities present in the data set) on March 1, 2020 and create a binary variable (*emp_dispabovemed*) checking whether each city's employment disparity strictly exceeds the median.

$$emp\_dispabovemed = 1(emp\_disp \geq median(emp\_disp))$$

Cities with an employment disparity equal to or below the median are coded as zero. The two groups are mutually exclusive. Depending on how many cities had an employment disparity equal to the median, this definition would affect results compared to a strategy which creates two binary variables like $1(emp\_disp \geq median(emp\_disp))$ and $1(emp\_disp \leq median(emp\_disp))$.

Fortunately, there is only one observation with a employment disparity equal to the median (no bunching of values at the median).

With the new binary variable, the mean employment disparity can be calculated for each week in each type of city, which is graphed below.

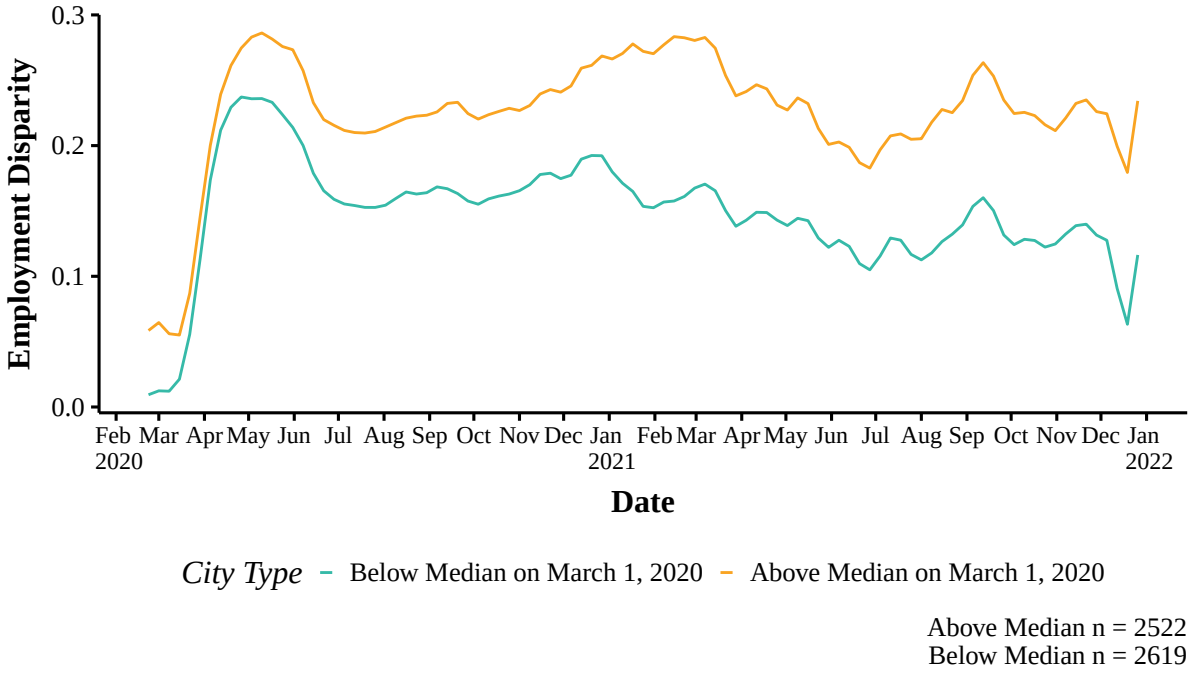

Figure 1: Mean Employment Disparity

On average, cities with higher early March 2020 employment disparities indeed had higher employment disparities throughout the pandemic than cities that experienced lower early March 2020 employment disparities

3.

For 3, we would like to evaluate the feasibility of small business permit requirements as an instrumental variable to find the LATE effect of small business openings on employment disparity.

I assume "small business permit requirements" refers to the minimum number of small business permit requirements needed to open a small business in a particular city, regardless of the type of business. The true number of permit requirements a business needs may vary because of occupation and business type.

(a) Independence/exogeneity of the instrument

If small business permit requirements satisfied independence, then cities would essentially receive their minimum permit requirements at random or approximately randomly. This scenario seems extremely unlikely – the number of permit requirements would likely be influenced by a city's political leaning or policy agenda, which would affect other small business and labor policies and hence employment disparity. For example, a city with high permit requirements may have more PPP money to give businesses (that hire low wage labor) during the pandemic, which they may use to retain more employees. If these other factors were known and could be controlled for, then the IV could have more validity.

(b) Exclusion restriction

If small business permit requirements satisfied the exclusion restriction, then the minimum number of permit requirements in a city would not affect employment disparity except through its effect on the number of small business openings. Small business permit requirements plausibly satisfies this assumption as businesses with low-wage workers likely would not hire differently based on these minimum permit requirements compared to businesses hiring high-wage workers.

(c) Monotonicity

Small business permit requirements plausibly satisfies monotonicity since a larger minimum number of permit requirements for a city is likely associated with a higher barrier for entry and fewer small business openings, and vice-versa.

2 Binned Scatter Plots

In this section, we would like to analyze the relationship between the percent of single-parent households and future earnings at the county-level across the U.S. The data was constructed

from 2000 and 2010 decennial U.S. census data linked to federal income tax returns and data from 2005-2015 American Community surveys.

Our outcomes and covariates are separated in different files. We must merge them together.

To merge the two data sets and only keep counties in present in both, I use an inner_join.

1.

In 1, we would like to create an unweighted binned scatter plot between single-parent share in a county and average future family income for white and black kids in the county with parents at the 25th percentile of the national income distribution. We would like to create this scatterplot without a package that does most of the work.

I create a vigintile (cut up into 20) variable for both X and Y variables.

I report the Pearson correlation coefficient for each graph between the two underlying variables.

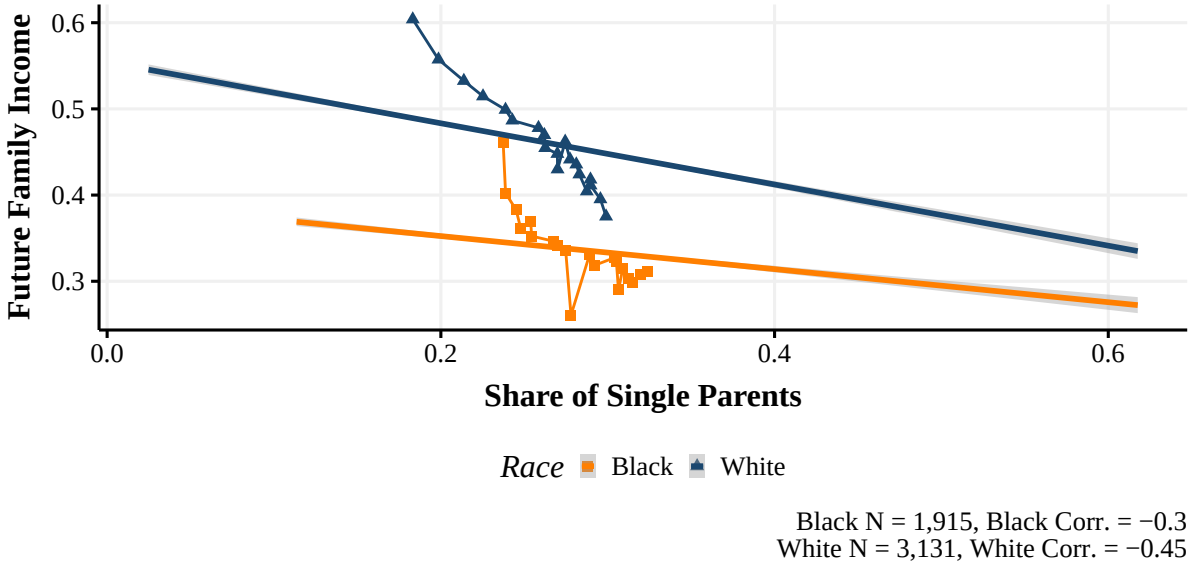

Figure 2: Unweighted Mean of Future Family Income for Children of Parents at the 25th National Income Percentile

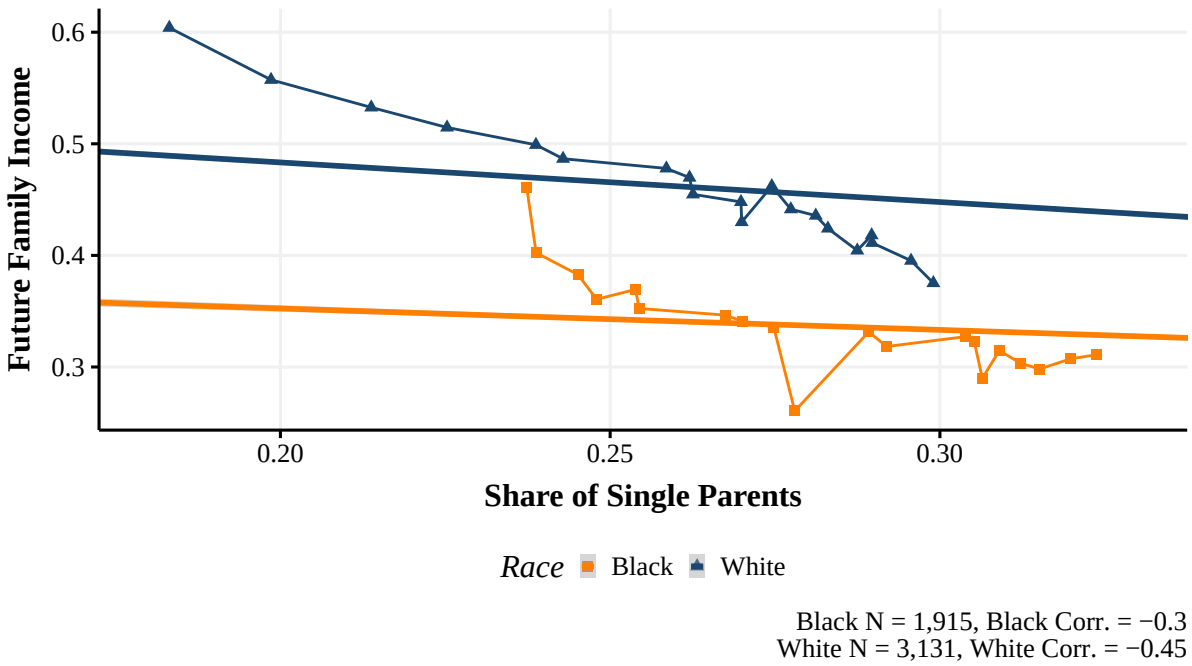

Figure 3: Unweighted Mean of Future Family Income for Children of Parents at the 25th National Income Percentile Zoomed

2.

In 2, we would like to replicate the plots in 1 but using the number of each children in a county as weights.

Refer to part 1 of this question for details on the functions used to create these figures. Note that missing weights are treated as zero. I weighted both the averages of future family income variables and the averages of the shares of single parents.

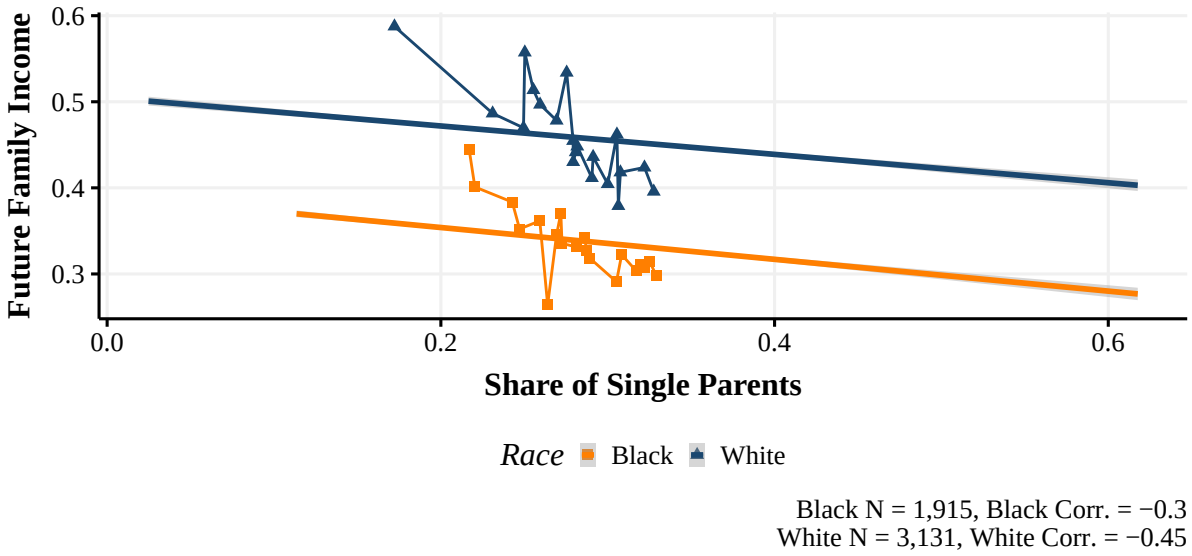

Figure 4: Weighted Mean of Future Family Income for Children of Parents at the 25th National Income Percentile

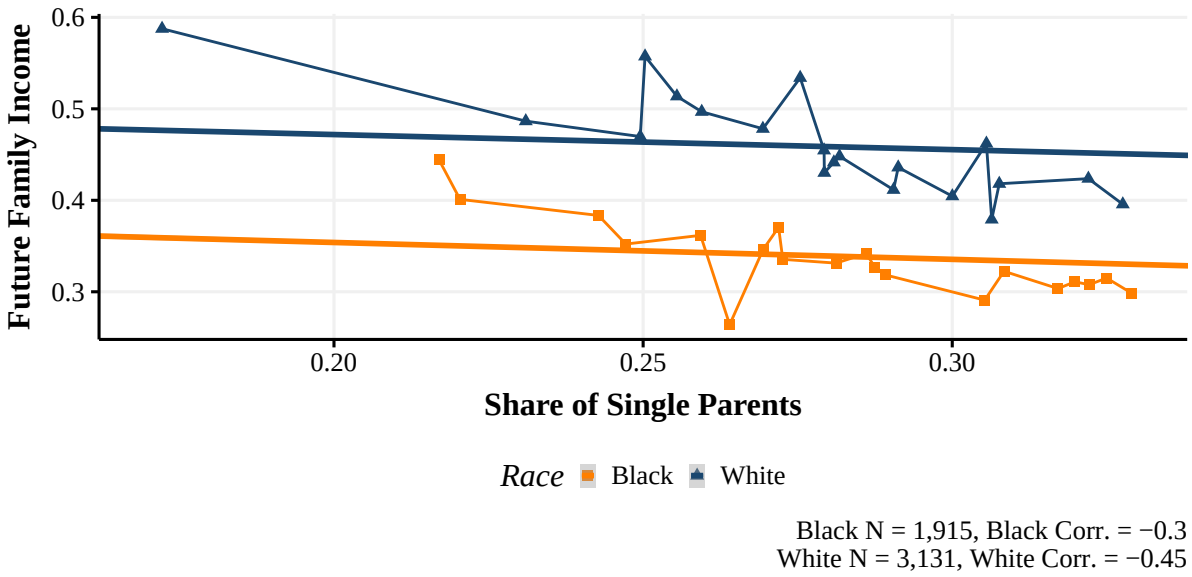

Figure 5: Weighted Mean of Future Family Income for Children of Parents at the 25th National Income Percentile Zoomed

3.

Let's discuss the value of weighing the scatterplot in 2 with the number of children in each county.

Weighing the binscatter with the number of children could provide a more accurate and relevant analysis of the relationship between counties' single-parent share and future family income. The weighted average is likely a less biased approximation of the true population average for future family income and county single-parent shares, allowing for more justifiable extrapolation from the graph to other populations outside the sample.

For example, suppose no weights are used. If one county had an extremely low average future family income, but this county also contained the most number of children living there, the average future family income would likely be overestimated, as well as the quality of life and opportunities available for children of the county in that bin (and similar counties outside the sample). The unweighted average can conceal potential differences in children's ability or future ability to generate earnings, by under- and over-stating different types of counties' mobility.

Specifically in the data, it does appears that more white children are better off than what the unweighted graphs showed, and racial disparities in future family income are larger than what an unweighted estimate would suggest. For black children in counties with relatively higher rates of single parents, future family income seems to be more depressed than without the weights. Hence, the weights better reflect the equitable circumstances and well-being of the population.

Still, the average future earnings was not calculated among the current number of children, so the actual impact of these weights on the asymptotic bias is not necessarily clear.

3 College Characteristics

In this section, we would like to explore the relationship between mobility and higher income achievement and university major and tuition characteristics. Specifically, we would like to analyze the determinants of, among those who have parents with income in the bottom income quintile, the percent of students who eventually earn an income level at the top quintile.

The data is from the Department of Education's IPEDS database in 200 and 2013, the College Scorecard, and various estimates of parent and child income by college from different sources. The data is at the college-level.

1.

Let's define the percent of students with parents in the bottom income quintile who reach the top income quintile as the mobility rate.

In 1, we would like to the following regression: mobility rate regressed on tuition in 2000, tuition in 2013, share of students with major in the arts and humanities, share of students with a business major, share of students with a STEM major, share of students with a social science major, Colleg Scorecard median earnings, and the percent of sutdents graduating within 150% of normal time in 2002. We would like to run this regression twice, separately modeling the effect for private and public universities. Reported standard errors are not robust to heteroskedasticity or serial correlation.

Table 3: Determinants of Mobility Rate in Private School

	(1)
(Intercept)	0.0218***
	(0.0024)
Tuition Sticker Price 2000	0.0000
	(0.0000)
Tuition Sticker Price 2013	0.0000**
	(0.0000)
Arts and Humanities Major Share 2000	−0.0001***
	(0.0000)
Business Major Share 2000	−0.0001+
	(0.0000)
STEM Major Share 2000	0.0001**
	(0.0000)
Social Science Major Share 2000	−0.0002***
	(0.0000)
Scorecard Median Earnings 2011	0.0000***
	(0.0000)
Percent of Students Graduating Within 150% of Normal Time in 2002	−0.0227***
	(0.0032)
Num.Obs.	804
R2	0.200
R2 Adj.	0.192
AIC	−4633.0
BIC	−4586.1
Log.Lik.	2326.487
F	24.921
RMSE	0.01

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

Table 4: Determinants of Mobility Rate in Public School

	(1)
(Intercept)	0.0031
	(0.0021)
Tuition Sticker Price 2000	0.0000***
	(0.0000)
Tuition Sticker Price 2013	0.0000
	(0.0000)
Arts and Humanities Major Share 2000	0.0000
	(0.0001)
Business Major Share 2000	−0.0001**
	(0.0000)
STEM Major Share 2000	−0.0001
	(0.0000)
Social Science Major Share 2000	0.0000
	(0.0000)
Scorecard Median Earnings 2011	0.0000***
	(0.0000)
Percent of Students Graduating Within 150% of Normal Time in 2002	−0.0092***
	(0.0028)
Num.Obs.	1038
R2	0.155
R2 Adj.	0.149
AIC	−6001.6
BIC	−5952.1
Log.Lik.	3010.793
F	23.637
RMSE	0.01

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

2.

Let's analyze the regression tables we generated in 1.

In the public school regression, four terms are significant at the 0.001 significance level. In the private school regression, six terms are significant at the 0.01 significance level.

For private institutions, the share of students who are majors in STEM fields has a small, positive and strongly significant relationship with mobility rate. A 1% increase in the share of STEM majors results in a 0.0001 increase in the likelihood that students, whose parents belong to the bottom income quintile, will reach the top income quintile. This seems reasonable: STEM majors typically earn more money than most other majors, and so an increase in the likelihood of having a major that earns more money increases the likelihood that a

random student, whose parents are in the bottom income quintile, achieves an income high enough which gets them to the top 20%.

For public institutions, the share of students who are majors in STEM fields has no statistically significant relationship with mobility rate. If the term were statistically significant, a 1% increase in the share of STEM majors would result in a 0.0001 decrease in the likelihood that students, whose parents belong to the bottom income quintile, will reach the top income quintile. It seems likely that, due to the competitive job market, the private school effects of STEM major share may take away from the relationship in public institutions. Public school graduates may have a tougher time competing with private school graduates and reaping the same benefits from having a STEM major.