Effect on the Employment Level by the 2021 VA Minimum Wage Hike with Variable Selection

Jacob Langley

2022-11-30

Abstract

This paper examines what effect Virginia's new minimum wage may have had on the employment level among food and beverage retailers in the state. On May 1st 2021, Virginia raised its minimum wage from the federal minimum of $7.25 to $9.50 and again to $11.00 on January 1st 2022. In the same timeframe Virginia's neighbor, North Carolina, kept the federal minimum of $7.25. This paper shows that thus far, neither minimum wage hike has had a statistically significant effect on the employment level among food and beverage retailers. I use a Difference In Differences (DID) regression with North Carolina as a control, and used logistic and lasso regressions for variable selection guidance.

Data

I have constructed a dataset of county level demographic data for Virginia and North Carolina from multiple sources. The employment level data for food and beverage retailers comes from the Bureau of Labor Statistics. The county college graduation rate is the total number of college graduates in the county from the US Census 2020 American Community Survey on Educational Attainment, divided by the county's total population. The rest of the demographic data including total population, average age, gender proportion, race/ethnicity percentages, and the population density are all calculated from the County Population by Characteristics section from the US Census Bureau's website.

Table 1 shows the descriptive statistics of the data used. North Carolina has 100 county equivalent areas and Virginia has 133, which gives 233 distinct counties. The dataset is comprised of two cross sectional snapshots representing Time = 0 and Time = 1 for each linear regression. The main regression consists of Employment snapshots from 2019 and 2021. The secondary regression, which is used to check for a lag effect, consists of snapshots from 2019 and 2022. All demographic data is also from 2019 and 2021, except for the county college graduation rate which is from 2020.

Table 1: Descriptive Statistics: Virgina & North Carolina

($number of counties = 233$)

Measure	# of distinct counties	Time = 0	Time = 1	Total	Mean	Standard Deviation	Min	Max
Food and Beverage Employment (2019 Q1 and 2021 Q4)	196	189	180	369	919.66	1670.43	11.00	12746.30
Food and Beverage Employment (2019 Q1 and 2022 Q1)	201	189	191	380	894.85	1654.87	10.00	12336.30
Average age (grouped)	233	233	233	466	8.92	0.69	6.93	10.86
Gender (1 = 100% women)	233	233	233	466	0.51	0.02	0.36	0.54
Percent White	233	233	233	466	76.72	17.04	20.40	98.90
Percent Black	233	233	233	466	21.04	16.55	0.90	79.10
Percent Hisp/Latino	233	233	233	466	6.88	5.42	0.90	43.00
Population Density	233	233	233	466	579.73	1368.08	3.20	10365.90
College Graduation Rate	233	233	233	466	18.64	8.49	7.30	60.90

Since there are two sets of cross sectional data of 233 counties, there is a total of 466 possible observations. For the demographic data all 233 counties are accounted for in both 2019 and 2021. 2022, however, has not been posted yet so any 2022 calculations uses 2021 demographic data. As for the employment data, 44 counties are missing from 2019, 53 are missing in 2021, and 42 are missing from 2022. The employment data is an average of the three months in the given quarter.

The dataset is then set up for a DID regression by adding three dummy variables: Control, Time, and Treatment. The Control variable is used to denote the difference between the subject (Virginia) and the control (North Carolina). The Time variable is used to denote the snapshot of time before, and after the treatment (minimum wage hike(s)). The Treatment variable then an interaction term between Control and Time to denote the subject after the treatment. In this way, DID Regression attempts to isolate the effect of the treatment from the differences in the subject versus the control, as well as the differences in different time frames.

Because of the large number of missing counties in the employment level data, selection bias may influence the results. Counties that do not have the resources to collect employment level data may be more rural with a lower average income and may be more exposed to a minimum wage hike. Table 2 shows the descriptive statistics of the counties not represented in the Employment Data, and Table 3 shows those that are.

Table 2: Descriptive Statistics of Counties Missing from Employment Level Dataset ($TotalMissingObservations = 97/466$)

Measure	Counties	2019	2021	Total	Mean	SD	Min	Max
Average age (grouped)	60	44	53	97	9.16	0.61	7.27	10.31
Gender (1 = 100% women)	60	44	53	97	0.50	0.03	0.43	0.54
Percent White	60	44	53	97	76.41	17.27	32.50	98.90
Percent Black	60	44	53	97	22.38	17.07	1.00	66.90
Percent Hisp/Latino	60	44	53	97	5.06	4.82	0.90	41.00
Population Density	60	44	53	97	300.08	1067.23	5.40	9550.80
College Graduation Rate	60	44	53	97	15.76	5.01	7.30	31.60

Table 3: Descriptive Statistics of Counties Observed in Employment Level Dataset ($TotalObservedObservations = 369/466$)

Measure	Counties	2019	2021	Total	Mean	SD	Min	Max
Average age (grouped)	196	189	180	369	8.86	0.70	6.93	10.86
Gender (1 = 100% women)	196	189	180	369	0.51	0.02	0.36	0.54
Percent White	196	189	180	369	76.80	17.01	20.40	98.90
Percent Black	196	189	180	369	20.69	16.42	0.90	79.10
Percent Hisp/Latino	196	189	180	369	7.36	5.48	1.00	43.00
Population Density	196	189	180	369	653.24	1428.93	3.20	10365.90
College Graduation Rate	196	189	180	369	19.39	9.04	7.30	60.90

Methodology

I laid out the original research design with a Directional Acyclic Graph (DAG, figure 1) in which I lay out my research assumptions for the major roots of the employment level in a typical western capitalist economy. Most of these causes are hard to quantify, so there are not publicly available datasets to evaluate them. In grey are roots whose data is not available to me, and in blue are datasets that I was able to get. Demographic data is readily available, and as the graph suggests it acts as a proxy for the demand level in the model.

DID Regression takes a snapshot of the target and control before and after the treatment, and then compares them. It evaluates the dependent variable to see if the target of the study significantly diverged from the control accounting

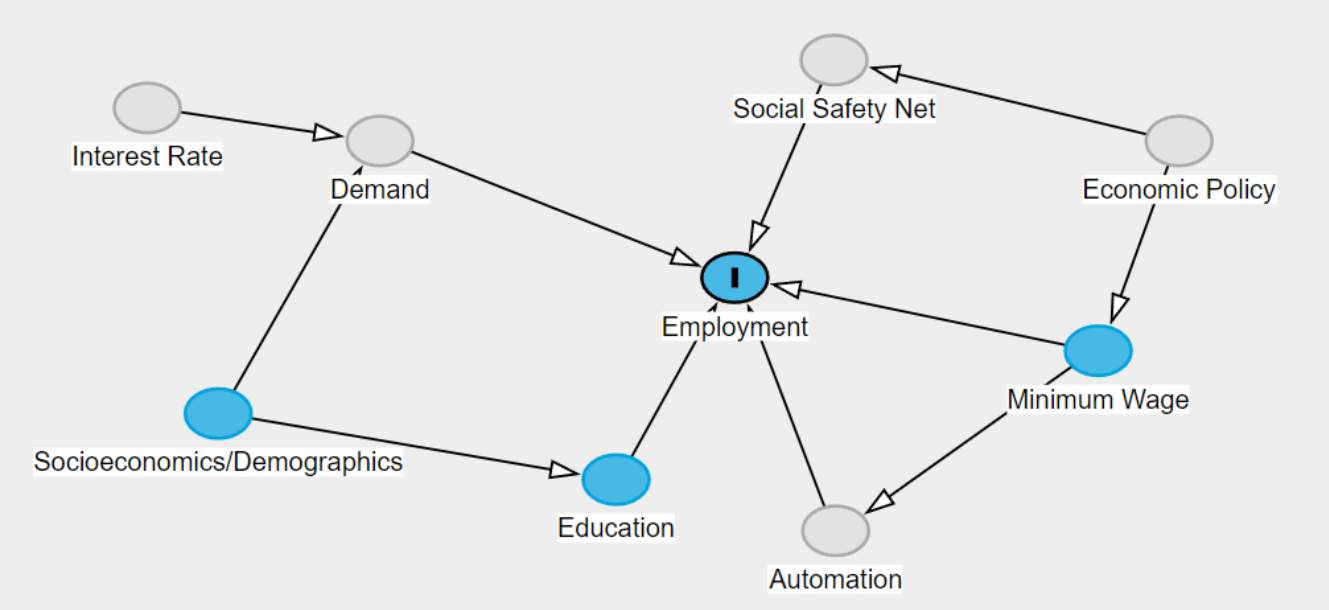


Figure 1: Unemployment DAG

for the timeframe, and the inherent differences between target and control. If the treatment variable is statistically significant then it means the treatment caused a measurable divergence in the subject of interest.

The first regressions include all demographic data without variable selection. Logistic and Lasso Regressions are then used to select variables that have the most direct effect on the employment level. The goal of the variable selection process I use is to deselect variables that are acting as proxies for the differences between the two states (i.e. the control variable).

The correlation between percent white and percent black was too high (-96%) to include both without violating the multicollinearity assumption, so a secondary goal of the variable selection is to determine which of the two is a better predictor of the food and beverage employment level.

Models

Using either 2021 or 2022 data, the regression immediately shows the Minimum Wage Hike to be non-significant. Which means, according to this initial model, the Minimum Wage Hike had no measurable effect on the employment level among food and beverage retailers.

A potential issue with the original DID model is that it is unclear how good of a control North Carolina is for Virginia. If they are very dissimilar then the predictors could just be telling the differences in employment between the two states. If that is the case then North Carolina is a poor control and the model would be invalid.

Table 4: Original OLS Estimates of the Effect of Minimum Wage on the Employment Level in VA

	Dependent variable:			
	Employment among Food and Beverage Retailers in VA and NC			
	2019-2021	2019-2021	2019-2022	2019-2022
	(1)	(2)	(3)	(4)
Time	103.932	105.244	100.688	103.335
	(212.090)	(212.868)	(208.852)	(209.625)
Control	−465.194**	−484.558**	−448.790**	−467.481**
	(211.483)	(212.001)	(209.056)	(209.615)
Minimum Wage Hike	−83.746	−85.009	−98.450	−101.498
	(285.939)	(286.984)	(279.417)	(280.453)
Average age (grouped)	−477.285****	−516.424****	−464.212****	−500.939****
	(131.521)	(129.825)	(126.973)	(125.537)
Gender	−1,833.666	−1,850.932	−937.176	−933.759
	(3,583.782)	(3,598.168)	(3,428.171)	(3,442.005)
Percent White	−13.036***		−12.956***	
	(4.683)		(4.510)	
Percent Black		10.878**		10.842**
		(4.825)		(4.645)
Percent Hisp/Latino	30.304*	30.580*	32.960*	33.356*
	(17.222)	(17.360)	(16.838)	(16.969)
Population Density	−0.240***	−0.233***	−0.248****	−0.243***
	(0.075)	(0.075)	(0.073)	(0.074)
College Graduation Rate	105.830****	106.196****	105.191****	105.573****
	(10.351)	(10.606)	(10.134)	(10.377)
Constant	5,190.012**	4,316.541*	4,601.541**	3,701.801*
	(2,156.160)	(2,212.769)	(2,049.545)	(2,099.113)
Observations	369	369	380	380
R^2	0.349	0.344	0.349	0.344
Adjusted R^2	0.332	0.328	0.333	0.328
Residual Std. Error	1,364.857 (df = 359)	1,369.843 (df = 359)	1,351.731 (df = 370)	1,356.771 (df = 370)
F Statistic	21.358**** (df = 9; 359)	20.913**** (df = 9; 359)	22.005**** (df = 9; 370)	21.537**** (df = 9; 370)
Note:				* p<0.1; ** p<0.05; *** p<0.01; ****p<0.001

The next step is to use logistic regression to identify which variables are most useful to distinguish Virginia from North Carolina. Those variables that do a good job at categorizing the two states and a poor job of predicting the employment level can be dropped from the model. Their effect would be better captured in the Control variable. As Table 5 shows, all the listed variables for both years were found to be significant (p<.05) in predicting the difference between the two states, except average age, percent black, and percent white. The Logistic regression had a 23% error rate in its prediction for all four versions.

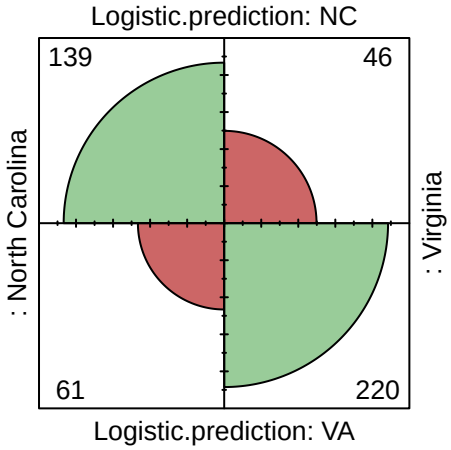
Once I determined which variables were good at parsing the difference between the states, I then ran a lasso regression to find the strongest determiners of the employment level. Lasso regressions shrink the model and force less important variables to zero relative to the lambda tuning parameter. This method is especially useful when working with small datasets because it is less sensitive to outliers.

Generally in Lasso Regression the weaker predictors will be pushed to 0 first, while the stronger ones will be pushed to 0 later. These results suggest that Percentage Black is a weaker predictor than Percentage White of the employment level among food and beverage retailers in North Carolina and Virginia. This creates a natural lambda selection at log(lambda) = 5 meaning we would drop every variable left of the double bar at log lambda = 5.27. This solves our colinearity problem between Percent White and Percent Black. Out of the other four variables we are dropping; Treatment, Gender, Time, and Population Density, only Population Density was significant in the original model. The DID variables will need to stay to maintain the assumptions of DID regression, but these results do allow us to drop Gender and Population Density from the final model.

Table 5: Logistic Regression to Evaluate the Classification Power of the Variables

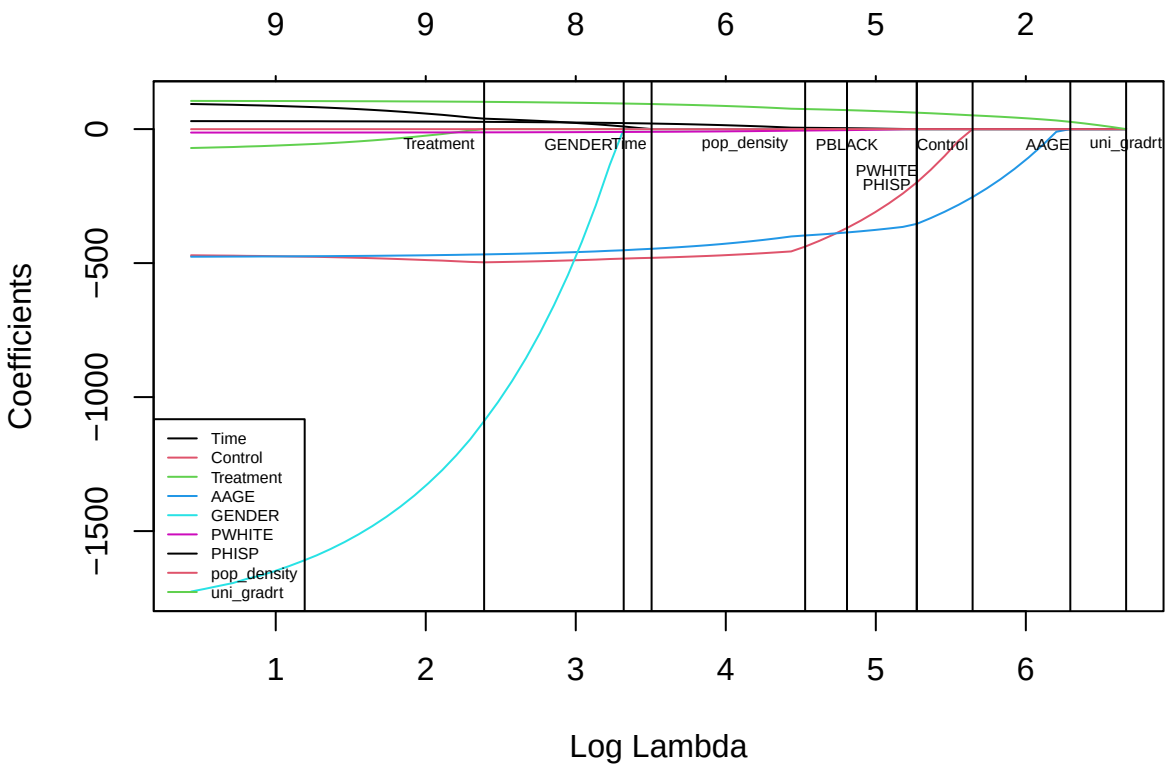
	Dependent variable:			
	Control: North Carolina = 0, Virginia = 1			
	2019-2021	2019-2021	2019-2022	2019-2022
	(1)	(2)	(3)	(4)
Average age (grouped)	−0.024	0.024	−0.024	0.024
	(0.222)	(0.219)	(0.222)	(0.219)
Gender	−30.221****	−30.889****	−30.221****	−30.889****
	(7.025)	(7.079)	(7.025)	(7.079)
Percent White	0.012*		0.012*	
	(0.007)		(0.007)	
Percent Black		−0.008		−0.008
		(0.008)		(0.008)
Percent Hisp/Latino	−0.341****	−0.340****	−0.341****	−0.340****
	(0.043)	(0.043)	(0.043)	(0.043)
Population Density	0.003****	0.003****	0.003****	0.003****
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
College Graduation Rate	0.037**	0.039**	0.037**	0.039**
	(0.018)	(0.019)	(0.018)	(0.019)
Constant	15.383****	16.376****	15.383****	16.376****
	(3.941)	(3.964)	(3.941)	(3.964)
Observations	466	466	466	466
Log Likelihood	−229.935	−230.847	−229.935	−230.847
Akaike Inf. Crit.	473.870	475.695	473.870	475.695
Note:	* p<0.1; ** p<0.05; *** p<0.01; ****p<0.001			

Logistic Confusion Matrix

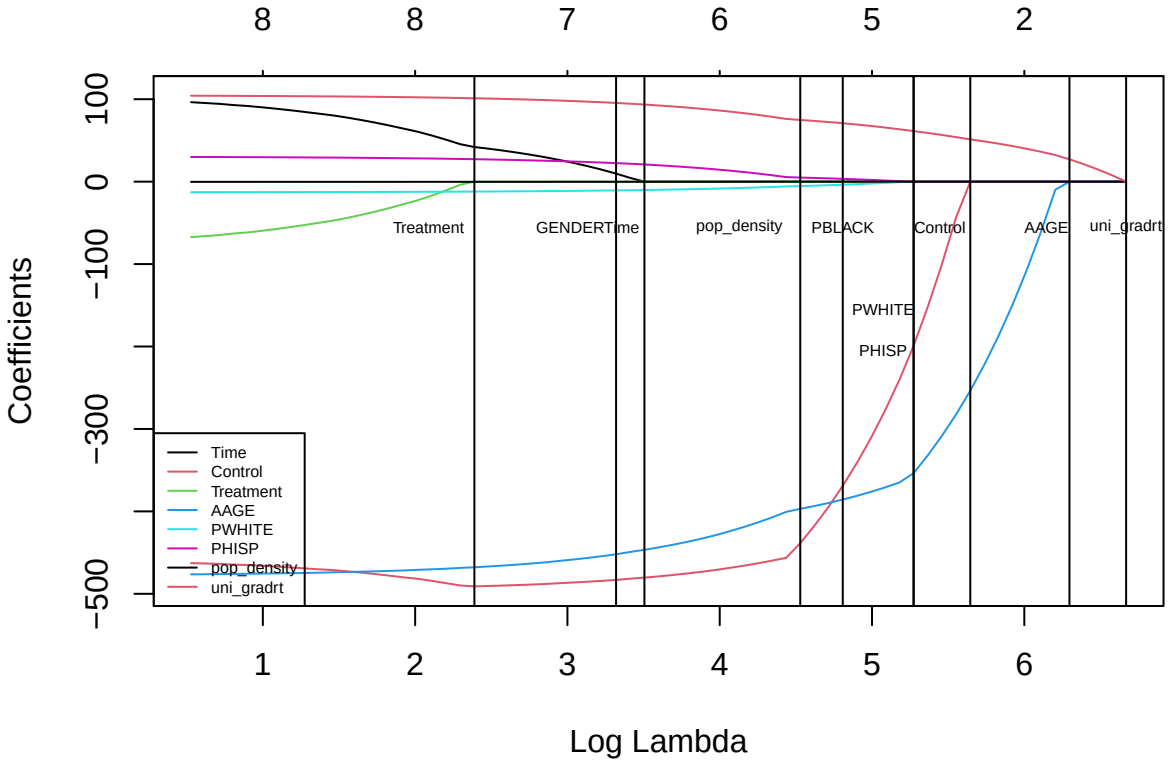


	lambdas
Treatment	2.389
GENDER	3.319
Time	3.506
pop_density	4.529
PBLACK	4.808
PWHITE	5.273
PHISP	5.273
Control	5.645
AAGE	6.297
uni_gradrt	6.669

Lasso Regression Plot



Lasso Plot with Percentage White (zoomed in: dropped gender)



Lasso Plot with Percentage Black (zoomed in: dropped gender)

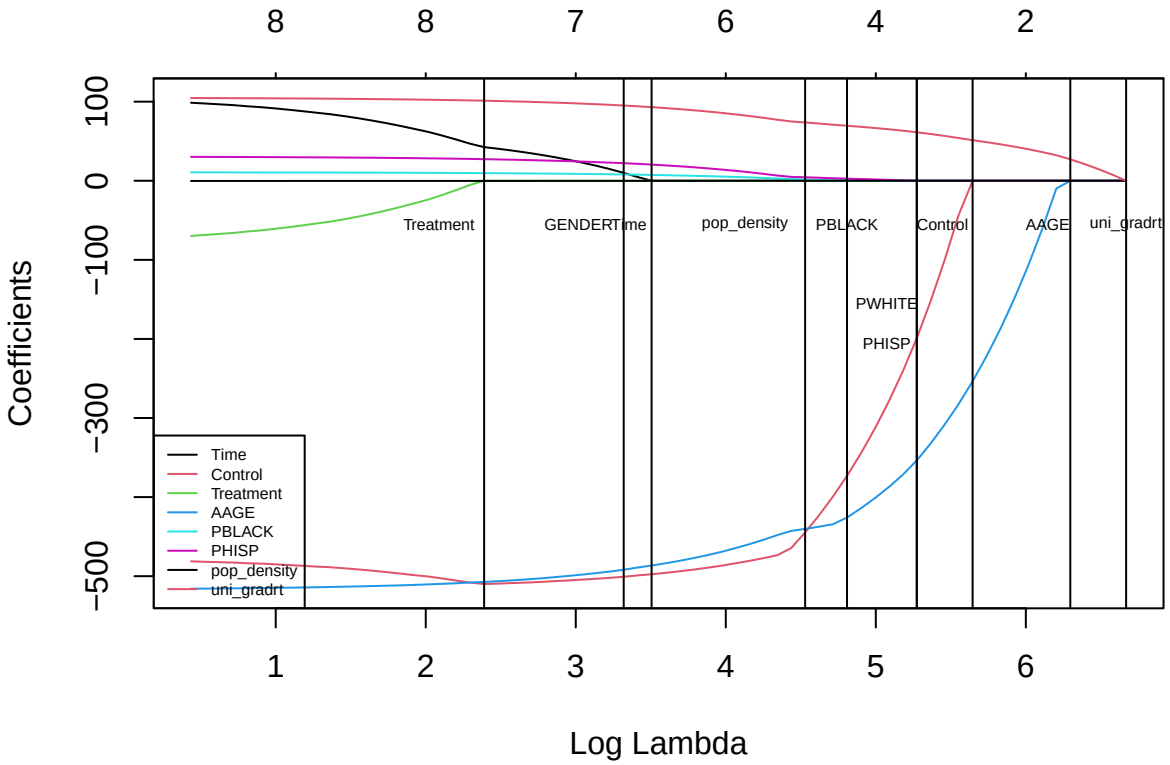


Table 6: Comparing the Original Regression to the Lassoed Regression

	Dependent variable:		
	Employment among Food and Beverage Retailers		
	Original	Lassoed	Control Test
	(1)	(2)	(3)
Time	103.932		62.737
	(212.090)		(148.002)
Control	-465.194^{**}	-654.318^{****}	
	(211.483)	(151.078)	
Minimum Wage Hike	−83.746		
	(285.939)		
Average age (grouped)	-477.285^{****}	-431.529^{***}	-367.700^{***}
	(131.521)	(131.965)	(134.668)
Gender	−1,833.666		6.962
	(3,583.782)		(3,682.644)
Percent White	-13.036^{***}	-10.526^{**}	-11.244^{**}
	(4.683)	(4.663)	(4.789)
Percent Hisp/Latino	30.304^{*}	9.746	25.324
	(17.222)	(16.036)	(16.055)
Population Density	-0.240^{***}		
	(0.075)		
College Graduation Rate	105.830^{****}	87.167^{****}	79.194^{****}
	(10.351)	(8.630)	(8.849)
Constant	$5{,}190.012^{**}$	$4{,}150.095^{****}$	3,285.383
	(2,156.160)	(1,189.485)	(2,190.122)
Observations	369	369	369
R^2	0.349	0.329	0.295
Adjusted R^2	0.332	0.320	0.283
Residual Std. Error	1,364.857 (df = 359)	1,377.728 (df = 363)	1,414.474 (df = 362)
F Statistic	21.358^{****} (df = 9; 359)	35.595^{****} (df = 5; 363)	25.205^{****} (df = 6; 362)
Note:			* p<0.1; ** p<0.05; *** p<0.01; ****p<0.001

There is an apparent glaring contradiction between the Lasso regression and Table 6. After dropping the weak variables from the model, Percent Hisp/Latino actually gets significance shifted away from it. However, in the context of table 5 it makes sense. Of the variables listed as highly significant ($p < .001$) in table 5, Percent Hisp/Latino is the only one still in our model. This implies that it is only good at predicting employment insofar as it is good at differentiating between North Carolina and Virginia. We should expect to see a large difference in concentrations of Hispanic populations in the two states.

In the control test, Control and Treatment are dropped and significance is shifted back onto Percent Hisp/Latino confirming that it is acting as a proxy for control. A similar description can apply to Population Density, though it had a lower lambda value. To show this in Table 7 I've run a logistic regression on Percent Hisp/Latino and Population Density compared to all the other demographics. Hisp and Pop got an error rate of 27% and all the others got an error rate of 38%. Since Hisp/Latino appears to be acting as a proxy for control, it will also be dropped in the final model.

In the following density graphs we can see that Average Age, Gender, Percent White, and Percent Black are all very similar between the two states. Whereas Percent Hisp/Latino, Population Density, College Graduation Rate, and the Employment Level are noticeably different.

Results

The variables that were weak predictors of employment, or that were acting as proxies for Control have been dropped; leaving us with a more predictive model that only holds the necessary variables shown in Table 8. In this final model, the Minimum Wage Hike is still shown to be non-significant, meaning it has not effected the employment level in the state. Updating the employment data to 2022 shows the same result. Variable selection in this model was important because treatment is similar enough to control that if there are proxies taking significance away from control, it would also be shifting significance away from the Minimum Wage Hike. This way we can be more confident in our conclusion. The p value for Treatment in the final model was 0.76, and 0.72 for final 2022.

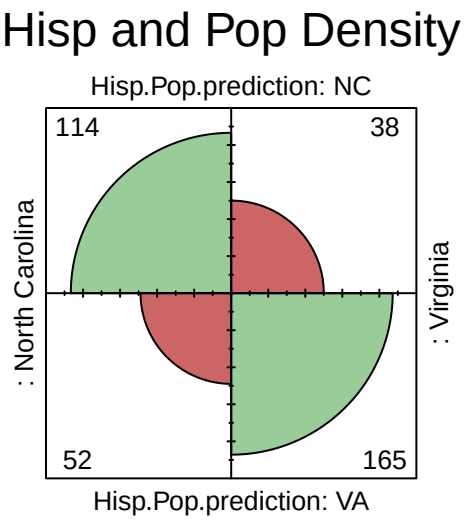
The remaining significant variables can be interpreted as follows:

Average Age: for every 5 years the average age of the population is older, the employment level among Food and beverage retailers decreases by 471 positions, working out to be 94 positions per average year older.

Percent White: for each percentage point higher a county is whiter, the employment among food and beverage retailers level decreases by 10 positions

College Graduation Rate: for every percentage point higher of people with a college degree, food and beverage retailer employment in the county increases by 88 jobs.

State difference: employment among Food and beverage retailers in VA lowers by 664 when compared to NC.



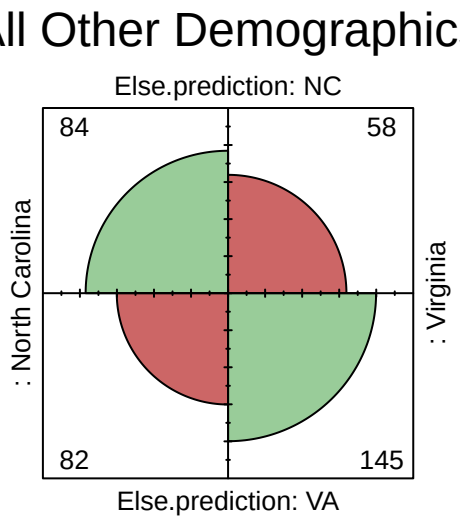
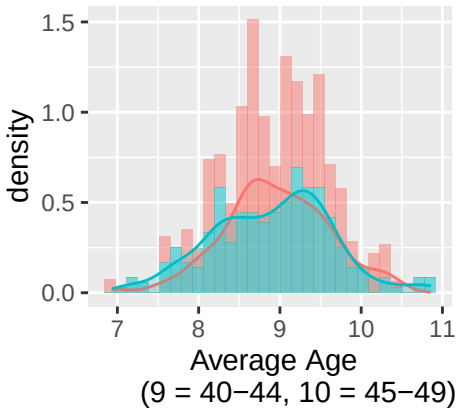
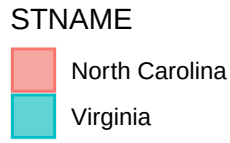
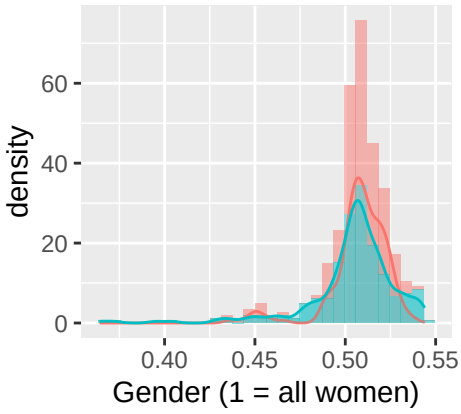


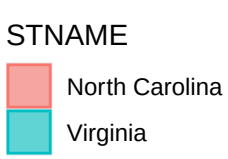
Table 7: Hisp/Latino and Population Density are Good Classifiers

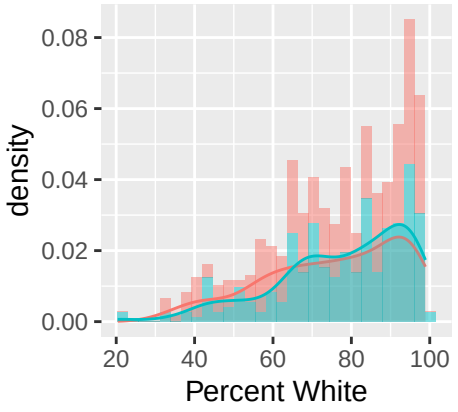
	Dependent variable:	
	Control: North Carolina = 0, Virginia = 1	
	Hisp and Population	Other
	(1)	(2)
Percent Hisp/Latino	-0.298^{****}	
	(0.042)	
Population Density	0.003^{****}	
	(0.0004)	
Average age (grouped)		0.005
		(0.172)
Gender		-13.921^{**}
		(5.833)
Percent White		0.001
		(0.007)
College Graduation Rate		0.057^{****}
		(0.015)
Constant	1.266^{****}	6.031^{*}
	(0.234)	(3.267)
Observations	369	369
Log Likelihood	−192.885	−243.257
Akaike Inf. Crit.	391.770	496.513
Note:	* p<0.1; ** p<0.05; *** p<0.01; ****p<0.001	

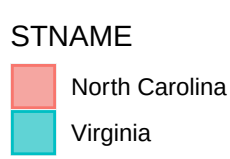

density
1.5
1.0
0.5
0.0
7
8
9
10
11
Average Age
(9 = 40−44, 10 = 45−49)

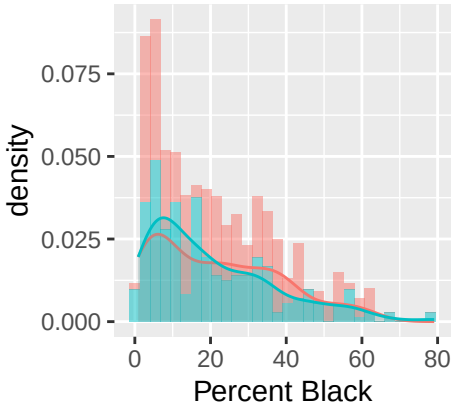

STNAME
North Carolina
Virginia

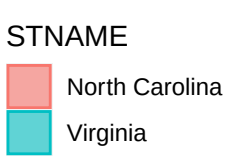

density
60
40
20
0
0.40
0.45
0.50
0.55
Gender (1 = all women)

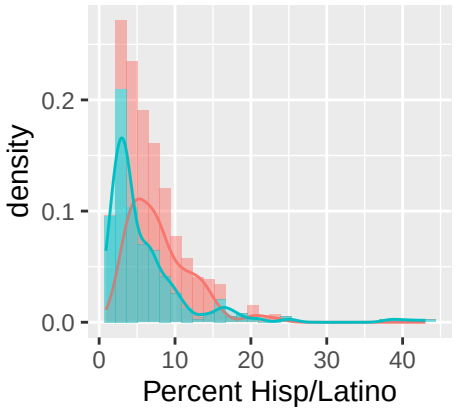

STNAME
North Carolina
Virginia

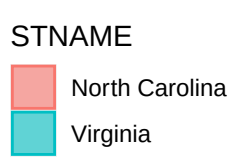

density
0.08
0.06
0.04
0.02
0.00
20
40
60
80
100
Percent White

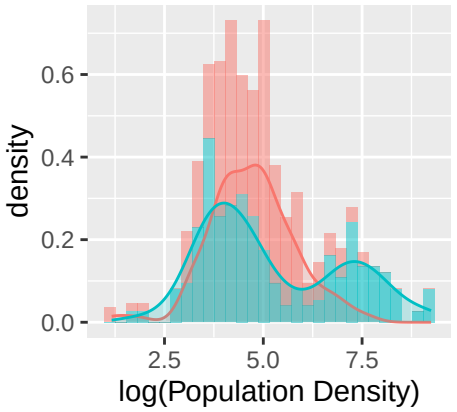

STNAME
North Carolina
Virginia

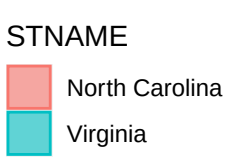

density
0.075
0.050
0.025
0.000
0
20
40
60
80
Percent Black

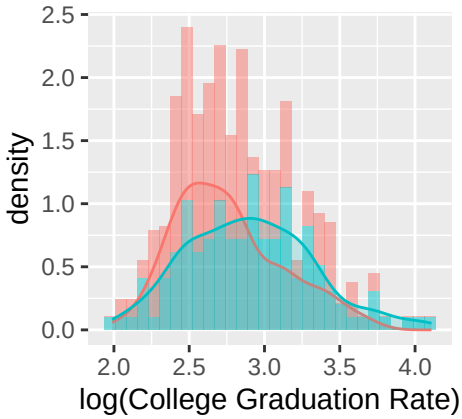

STNAME
North Carolina
Virginia

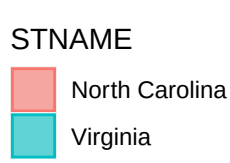

density
0.2
0.1
0.0
0
10
20
30
40
Percent Hisp/Latino

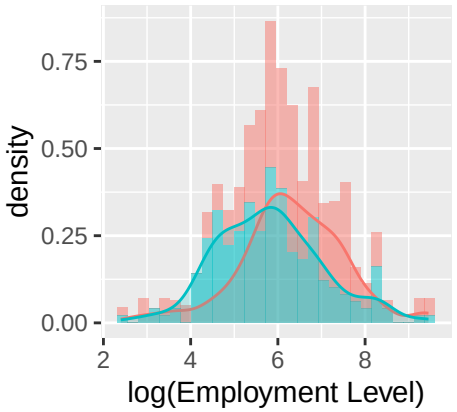

STNAME
North Carolina
Virginia


density
0.6
0.4
0.2
0.0
2.5
5.0
7.5
log(Population Density)


STNAME
North Carolina
Virginia


density
2.5
2.0
1.5
1.0
0.5
0.0
2.0
2.5
3.0
3.5
4.0
log(College Graduation Rate)


STNAME
North Carolina
Virginia


density
0.75
0.50
0.25
0.00
2
4
6
8
log(Employment Level)

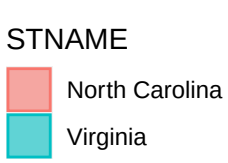

STNAME
North Carolina
Virginia

Table 8: Final Model

	Dependent variable:		
	Employment among Food and Beverage Retailers		
	Original	Final Model	Final with 2022 employment
	(1)	(2)	(3)
Time	103.932	115.606	106.746
	(212.090)	(214.208)	(211.138)
Control	−465.194**	−631.897***	−630.617***
	(211.483)	(203.816)	(202.076)
Minimum Wage Hike	−83.746	−87.194	−100.218
	(285.939)	(288.896)	(282.709)
Average age (grouped)	−477.285****	−471.180****	−456.023****
	(131.521)	(115.022)	(111.188)
Gender	−1,833.666		
	(3,583.782)		
Percent White	−13.036***	−10.379**	−10.397**
	(4.683)	(4.662)	(4.502)
Percent Hisp/Latino	30.304*		
	(17.222)		
Population Density	−0.240***		
	(0.075)		
College Graduation Rate	105.830****	87.974****	87.734****
	(10.351)	(8.528)	(8.385)
Constant	5,190.012**	4,500.775****	4,371.644****
	(2,156.160)	(992.585)	(970.073)
Observations	369	369	380
R^2	0.349	0.329	0.327
Adjusted R^2	0.332	0.318	0.316
Residual Std. Error	1,364.857 (df = 359)	1,379.735 (df = 362)	1,368.186 (df = 373)
F Statistic	21.358**** (df = 9; 359)	29.567**** (df = 6; 362)	30.245**** (df = 6; 373)
Note:			* p<0.1; ** p<0.05; *** p<0.01; ****p<0.001

Discussion

The most fundamental problem with this study is the possibility of a major selection bias. It's possible that by selecting counties that have the ability to report employment data, I am eliminating the only counties that have had adverse effects on their employment levels from the minimum wage hikes. Another open question is whether North Carolina is similar enough to Virginia for it to be a valid control, even after using variable selection. I did check into a very short lag effect (Q4 2021 to Q1 2022), however its possible it would take longer. Virginia is planning a third hike from $11 an hour to $12 on January 1, 2023 and then two more: $13.50 in 2025 and $15 in 2026. A future study could collect the missing data manually and do another analysis since it appears there will be ample opportunity.

Sources

U.S. Bureau of Labor Statistics. (n.d.). QCEW data files. U.S. Bureau of Labor Statistics. Retrieved December 7, 2022, from https://www.bls.gov/cew/downloadable-data-files.htm

U.S. Census Bureau. (n.d.). Educational Attainment. Explore census data. Retrieved December 7, 2022, from https://data.census.gov/table?t=Educational%2BAttainment&g=0400000US37%240500000&y=2020&tid=ACSST5Y2020.S1501

US Census Bureau. (2021, October 8). County population by characteristics: 2010-2019. Census.gov. Retrieved December 7, 2022, from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html

US Census Bureau. (2022, June 30). County population by characteristics: 2020-2021. Census.gov. Retrieved December 7, 2022, from https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-detail.html

Virginia Open Data Portal. Tyler Data & Insights. (n.d.). Retrieved December 7, 2022, from https://data.virginia.gov/