Beth Karp

12/6/2020

Writeup on Data Skills II Final Project

**Background**

This project began as a recreation of a project I did over the summer in Excel, demonstrating that there is no relationship between bar passage rates, Multistate Professional Responsibility Exam cut scores, and professional discipline rates for lawyers. I took on that project because states and the National Conference of Bar Examiners (NCBE), which designs the bar and MPRE materials for most states, claim that the tests are necessary to “protect the public” even though they are extremely expensive and difficult, do not test the skills necessary for legal practice, and tend to exclude the most historically marginalized demographics from the profession. The most telling argument I have seen that the exams test privilege is that people who complete 75% of a major test prep company’s bar prep course have passage rates well north of 90%, which is higher than any state’s overall passage rate. The prep courses cost around $2,000 and require about two months of full-time study.

As I began to gather data, I decided that I would rather explore a new question: whether it is possible to observe a statistical relationship between state characteristics and bar passage rates. If there is a correlation between racial demographics or economic indicators and passage rates, that would support the argument that the bar exam tests privilege. It is a very imperfect way to probe at the question, since not all states administer the same exam or have the same passing scores, the data is state-level rather than individual-level, economic and racial measures are likely to be highly correlated, and people who take the bar exam in a given state do not necessarily live in that state. But the approach is not totally useless. The bar exam was instituted in most states to reserve legal licensure for the “right kind of people,” and if you see lower bar passage rates in states with higher minority populations or higher rates of poverty or unemployment, then the bar exam is likely fulfilling its original, inequitable mission.

**Data**

I. Bar pass rates

I scraped the bar passage rate datasets from tables on the NCBE Statistics website, restricting the data to the years 2015-2018 (the years for which lawyer discipline information is also publicly available) and annual counts. For each jurisdiction, I pulled the annual total of takers and the number of passers, then took the quotient to get a pass rate.

II. Poverty rates

I used US census data to find poverty rates. The data was available in a fairly clean Excel spreadsheet, but it was not quite ready to use straight from the site. There is data from 1980 to 2019, with two sets of data for 2017 (it seems they updated their processing system that year). My code extracts one dataset for each year from 2015-2018 with two columns: state and poverty rate.

III. Layer discipline

This was the most complicated dataset. The ABA only publishes lawyer discipline data in PDF form, and each chart spans multiple pages. After I converted the charts into CSVs using Tabula, I had to fix specific rows and cells, remove the New Hampshire 2015 row (which has erroneous data), and sum the information from multiple New York districts into a single New York row. I had to design a unique function for each page in the 2015 and 2016 data get usable CSVs. The discipline data ultimately yielded three new columns: number of lawyers in the state, number of ethics complaints, and a column taking the quotient as complaints per lawyer.

IV. Unemployment rates and educational attainment

I pulled unemployment and educational attainment data from the USDA’s Economics Research Service website. The data needed virtually no cleaning; it mostly just needed to be filtered to the needed rows and columns.

V. Race

The racial data also came from the US Census. The data did not require much cleaning, but it took a surprising amount of work to replace numerical codes with race names since the data key was a PDF.

**Regressions**

The first regression regresses complaints per attorney on just about every available covariate in my dataset. A few results came out statistically significant, but you would expect some to in any sufficiently large set of covariates and it is hard to make sense of the coefficients. In all likelihood, some covariates are robbing others of explanatory power since they are correlated with one another.

The second regression looks only at the relationship between bar passage rates and complaints per lawyer. There is no statistically significant correlation.

The third regression regress bar passage rates on just about every available covariate. It is just as useless as the first regression.

The fourth and fifth regressions regress bar passage rates on Black and white state populations. Each shows a pretty large and statistically significant correlation, although is quite low.

The sixth and seventh regressions look for a correlation between poverty rates and unemployment rates, respectively, controlling for white state population. Both regressions show a statistically significant correlation between the economic measures and bar passage rates at the 10% level.

The last regression looks at the correlation between educational attainment (measured by percentage of state residents with a bachelor’s degree) and bar passage rates. There is no correlation.

**Plots**

My plots are visualizations of what I thought were the most interesting relationships. I plotted each relationship two ways: first as a scatterplot with point colors representing a variable I thought might be closely related to the X or Y variable, then as a scatterplot with a line of best fit.

The poverty rate vs. pass rate plots show a negative correlation between poverty rates and educational attainment rates and a weak negative correlation between poverty rates and pass rates.

The unemployment rate vs. pass rate plots show a positive correlation between poverty rate and unemployment rate (which is why I ran them separately in the regressions) and a weak negative correlation between unemployment rates and pass rates.

The white population vs. pass rates plots show a weak positive correlation between percent of white people in the population and pass rates.

Finally, the pass rate vs. complaints-per-lawyer plot shows that there is absolutely no correlation between pass rates and disciplinary complaints against lawyers.