Econometrics Final Paper
EC15
Spring 2023
Analysing the Causation of Patents Filed on Homelessness
Tej Chhabra and Jaden Richardson

Introduction:

In our empirical study, we investigated the empirical relationship between the number of patents filed in a state as a measure of innovation and the impact it has on homelessness on a state-by-state basis in the United States. Our hypothesis, which we hope to find evidence for, is that the higher the patent count is in a given region, the higher the relative homeless population will be as well. In essence, as the patent count rises, things such as cost of living, income and other things that can all be encompassed within the issue of homelessness will change. At the end of this paper, we hope to show evidence that there is not only a correlation between these two metrics but also causation.

Data Description:

In searching for our data we focused on finding data sets that were able to group the data on a state-by-state basis to maintain consistency between data sets with large enough sample sizes. This also allowed us to rely on more reliable federal and state data as opposed to the county-by-county or city-by-city data that may or may have been available or structured in the same way. We found our homelessness data from the Department of Housing and Development and their Annual Homelessness Assessment Report. Our Patent Count Data comes from the Federal Reserve Bank of St Louis. The origin of each patent is determined by the primary residence of the first-named inventor on the patent application. The number of patents given includes utility, plant, design, and reissue patents. Our population data came from the Census Bureau. Although the Census Bureau only collects data every 10 years, statistical estimates based on samples are published for years between full Census counts. These include all people residing in the US, including unauthorized immigrants and people living here temporarily for work or school. For state-by-state home prices, we were able to find data from the Federal Housing Finance Agency. Department of Education publishes the percentage of American

students entering public high school who graduate with a standard diploma in four years, which we will use as our high school graduation rate. We were also able to find data from the Bureau of Economic Analysis that specified the amount spent on average per person in each state. Also from the Bureau of Economic Analysis, we were able to find state-by-state GDP data that we used in our regression.

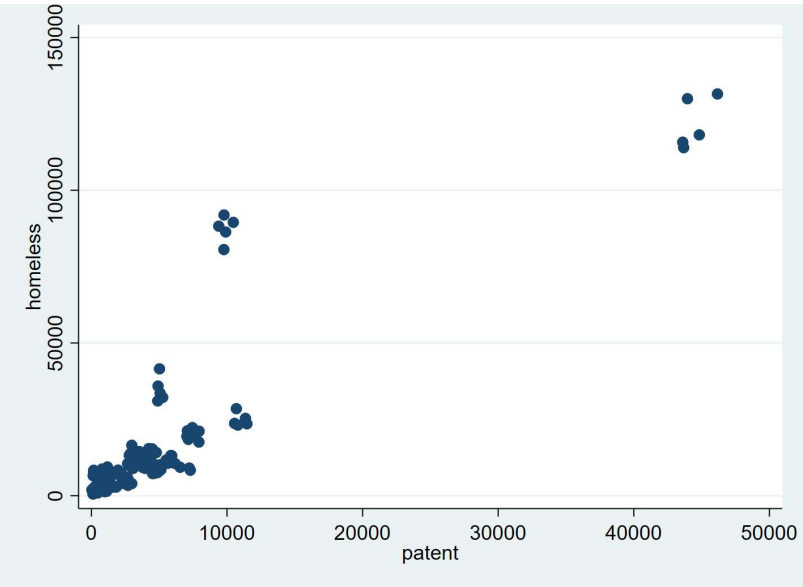
Table 1: Data description

Variable	**Mean**	**Standard deviation**	**Min**	**Max**
Homeless (number of people)	10862.59	20480.08	542	131532
Patents (number filed)	3172.165	6418.326	46	46172
Population (number of people)	6352178	7172772	579054	3.94e+07
House prices (House price index)	239.9316	57.83911	161.0625	495.63
HS Graduation Rate (Percent of class)	.8376457	.0523626	.61	.97
GDP (Millions of dollars)	369,910.7	471,799.6	29,690.7	2,975,083
Personal expenditures (Millions of dollars)	252,517.7	300,746.9	22,655	1,852,976

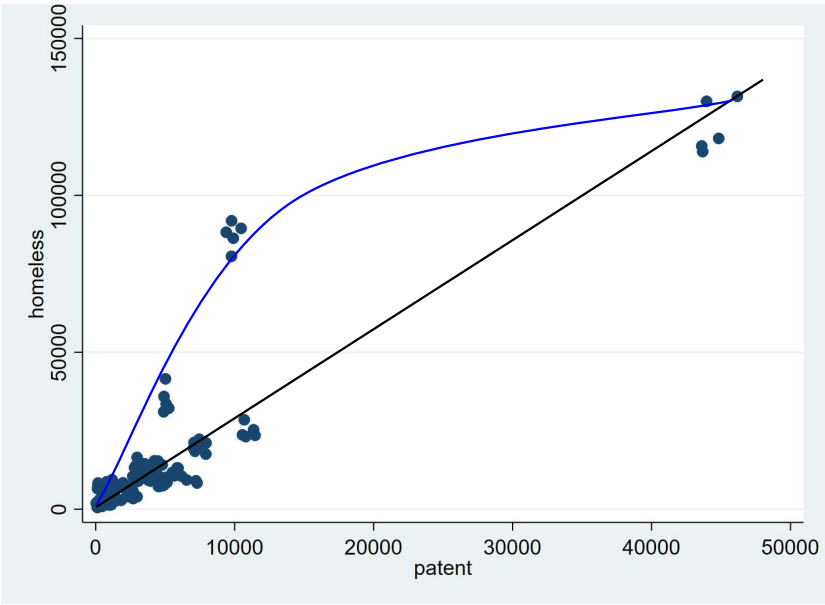
All of the above data is given for observations taken once a year per state from the years 2014 - 2018 including the District of Columbia. This results in an n of 254, as there is no data point for the District of Columbia for the year 2018.

Empirical analysis:

Graph 1: Patents vs Homeless



Graph 2: Different forms - Log vs Ln



The initial graphical results shown in *Graph 1* indicate that there is a clear correlation between the number of patents in a state and the number of homeless people in a state from the years 2014 and 2018. Using the RESET test for functional form, we used the expected values from the regression in *Table 3* to determine whether our functional form was correct. The results from the output yielded a p-value of less than 0.00001, allowing us to reject the null hypothesis that there are omitted interaction, quadratic, and cubic terms at a 1% confidence level.

The following table, *Table 2*, shows the different multivariable regressions run on this data and their results. Additionally, looking at the scatter plot shows a very linear pattern, so we assumed linearity. The comparisons of the different forms are shown in *Graph 2*.

Table 2: Regression results

Regression Analysis

	(1) homeless	(2) homeless	(3) homeless	(4) homeless	(5) homeless	(6) homeless
patent	3.468***	2.03***	2.003***	1.991***	.863**	.916**
	(.271)	*(.23)*	*(.238)*	*(.231)*	*(.369)*	*(.368)*
population		.001***	.001***	.001***	-.003***	-.004***
		(0)	*(0)*	*(0)*	*(.001)*	*(.001)*
home_price			10.148*	3.97	-5.661	-8.338
			(6.114)	*(6.653)*	*(8.056)*	*(7.498)*
perc_hs_grad_rate				-37528.317***	-54411.287***	-56198.053***
				(9551.536)	*(10202.715)*	*(10399.318)*
gdp					.069***	.045**
					(.015)	*(.021)*
pers_exp						.059*
						(.031)
_cons	3450.619***	-1180.549	-3661.118*	29408.982***	46932.9***	48841.076***
	(825.895)	(829.761)	(2028.099)	(8944.294)	(9826.891)	(9825.31)
Observations	619	250	250	250	250	250
R-squared	.315	.82	.821	.828	.896	.901
Adj R^2	.314	.819	.819	.825	.894	.899
F-stat	164.131	765.142	559.595	386.506	267.168	167.638
Robust	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors are in parentheses
*** p<.01, ** p<.05, * p<.1

Variables justification

The following section will be about justifying our variable selection and omitted variable bias, detailing the bias that each variable causes and our speculations about their correlation with both our explained and explanatory variables.

The first variable that we chose to include in our analysis was the population. We expected this variable to be positively correlated with homelessness, and positively correlated with patent count, leading there to be a positive bias. Including this variable yielded a smaller weight on the patent, which means that there was a positive bias, but the actual coefficient is negative. This means that our initial hypothesis that population is positively correlated with patents is wrong and that it is actually negatively correlated with it.

We then chose to include the home price index in our regression, thinking that it would have a positive correlation with homelessness and a negative correlation with patents, leading to a negative overall bias. Findings indicate that this was incorrect and that the omitted variable bias was in a positive direction.

The next variable that we included in our regression analysis was the high school graduation rate. We hypothesized that this is positively correlated with patent count, and negatively correlated with homeless, leading there to be a negative bias. We were incorrect, as including this variable revealed a positive bias, meaning that the graduation rate is negatively correlated with the patent count, which may be because a high graduation rate encourages students to leave their state and go to certain states that have better universities.

After this we included GDP per state on the basis that it is both positively correlated with patent count and homelessness, yielding a positive bias. This was correct, as including GDP cut the emphasis on the patent count by half.

Lastly, we included personal expenditures which we thought would have a negative correlation with patent count and a positive correlation with homelessness. This was also correct, showing a negative bias on the coefficient of the patent count before it was included.

Table 3: Regression without home price index

Regression Analysis

homeless	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
patent	.902	.372	2.42	.016	.169	1.635	**
population	-.004	.001	-4.44	0	-.005	-.002	***
perc_hs_grad_rate	-54568.318	9710.038	-5.62	0	-73694.511	-35442.126	***
gdp	.046	.021	2.18	.03	.004	.087	**
pers_exp	.058	.031	1.88	.061	-.003	.118	*
Constant	45432.38	8101.009	5.61	0	29475.548	61389.213	***

Mean dependent var	10912.912	SD dependent var	20639.955
R-squared	0.901	Number of obs	250
F-test	192.445	Prob > F	0.000
Akaike crit. (AIC)	5110.404	Bayesian crit. (BIC)	5131.533

*** $p<.01$, ** $p<.05$, * $p<.1$

Statistical significance

The first issue of statistical significance is in looking at the coefficients of house prices in the regressions that it is included in *Table 2*. Performing a t-test on the coefficient of house prices in the third regression gives us a p-value of 1.66 which is statistically significant at a 10 percent confidence level. Performing a t-test on the coefficient of house prices in regression 4-6 yields p-values of 0.607, 0.702, and 1.112 respectively, all showing a non-significant coefficient of home_price. Because of these results, we can conclude that house price has no statistical significance in the regression on the number of homeless people in the state. Intuition suggests that the higher the price of homes, the higher the number of homeless people in the state, so it is an interesting result that it is not statistically significant.

In regards to the rest of the variables, when performing t-tests, all of the coefficients yielded a less than 0.01 p-value, indicating that each of them is statistically significant.

Because there are no other individually insignificant variables, we did not find it necessary to conduct any F tests to see if there was any joint significance between variables.

Briefly discussing R^2 values, we found that each of our R^2 achieved a higher than .80, indicating that the independent variables explained over 80% of the variation in the number of homeless people. Statistically speaking, the explanatory power and statistical significance of personal expenditures on homelessness were less than the other variables, but due to the economic significance of its meaning, we chose to keep it in the regression.

Ultimately, due to the statistical insignificance of home prices, we chose to omit it in our final regression which is shown in *Table 3.*

Economic Significance

Looking at the regression shown in *Table 3*, we found that the coefficient of patents is .902, meaning that for each patent filed, the amount of homeless people increases by .902. This is an interesting observation that is consistent with our hypothesis: The more patents filed in a state, the more the homeless population there will be.

The coefficient of the population in this regression is -.004, meaning for every additional person in a state, the amount of homeless people decreases. This is an odd result, and has an interesting interpretation: the number of homeless people starts with a constant of 45432 people, and 0 people with a home. Once there are 1000 people in a state, the homeless population does not reduce by much, only four people. This puts the effect of population on homelessness into context, showing that it does not have a large impact in comparison to other explanatory variables.

The coefficient of the high school graduation rate is -54568.318. This is also a fairly intuitive result: as the amount of high school graduates increases, the fewer homeless people in a

state. The interpretation of this coefficient is a little different to the others as the value of the graduation rate is reported as a percentage divided by 100. So, as the high school graduation rate increases by one percentage point, the homeless population decreases by 545 people. This is a statistic which can be used in policy implementations when trying to reduce the homeless population in a state.

We next added GDP, another measure of economic activity, which yielded the statistically significant result that as GDP increases, so does the amount of homeless people in a state. This is a result that is in line with our hypothesis; As the GDP increases by 1 million dollars, the amount of homeless people increases by 0.046 people, an expected correlation. This is what we were expecting, as we are arguing that as these measures of standards of living increase, so does the amount of homeless people.

The last variable that we added to our regression was pers_exp, or the total personal expenditures in a state. The coefficient of this variable was .058, indicating that for every additional 1 million dollars spent by people in that state, the number of homeless people increases by 0.058. The rationale behind this is very similar to the one for GDP, and is an expected result.

Shortcomings:

Randomness

The assumption of randomness in OLS was violated. Because we observed the same state multiple times we cannot make a claim of randomness. In order to be able to do this there would have had to be an equal probability of one observation being drawn into the sample and the probability of one being drawn would have to be independent of another one being drawn.

Homoskedasticity

Conducting a test for homoskedasticity, We found that with an H_0 that the results are homogeneous, we would reject the null with a p-value of less than 0.0001, and so our regression did not pass the assumption of homoskedasticity. Thus, we used robust regression to conduct our analysis to correct this.

Sample Size

Another shortcoming in our regression analysis is in the sample size. Due to the fact that there was a limited number of available data sets fitting our needs, we were working with data from an already limited number of years. There are also only 50 states which further limited the number of data points we were able to use. Additionally, we were reliant on the already limited number of existing data sets available because we were not the ones collecting our own data.

External Validity

In terms of External Validity it is hard to know how our results would translate to other contexts such as other locales or other time periods. First, each locale whether it be a state or country has different rules and regulations with regard to patents. For example, some require patents to be renewed after different time periods than others or have different regulations for what is needed to have a patent issued etc. This issue also applies to housing where each locale has varying baselines for the amount of housing, cost of housing, and process of buying or renting a home. Given these differences, our results would likely be different depending on the locations included.

Measurement error

Another issue that arose was measurement error. There was no way for us to fact-check the data we had collected. We were unable to confirm or control whether people were lying about

smaller things like their personal expenditures. The larger issue was with data provided by the different government agencies. There was no way to tell if states were collecting their data using different methods or who they were counting for categories such as homelessness that are very fluid. For example, one locale could include those in temporary housing while another may only include those who filed for federal housing assistance. In essence, there is no guarantee that the numbers reported by one locale would be the same if we used the methods of a different locale to measure the same variable.

Extensions:

Simultaneous Causality

There is an arguable case for simultaneous causality within our variables. It may be that the effects of the two on each other fight, with an increase in patent count resulting in an increase in homelessness due to the increase of innovation and productivity driving costs of living upwards, but an increase in homelessness decreasing the patent counts due to the perception that an area is not very nice, so it is not a good place to set up a business that involves a patent filing. Unfortunately due to the scope of this paper and the resources we have, we were unable to explore potential solutions to this problem, but it is something that should be explored in an extended version of this paper.

External Validity

In the shortcomings section we discussed arguments against our results being externally valid, but in this section, we are going to briefly highlight some situations in which external validity should theoretically hold. Inside the United States, using the same metrics and states but with different years should stay relatively accurate, as it controls for most structural differences and only changes one variable, the year. However, we do not possess the data to test our theory

of external validity, so the validity of this hypothesis is unknown. In any case, the scope of this paper does not require external validity, particularly as it has little to no forecasting purposes.

Time Trend

In this section, we explored adding some additional variables to account for growth over time, including a new variable t, which represents the number of years since 2014. The results of this change are shown below in *Table 4.*

Table 4: Regression with Time Trend

Linear regression

homeless	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
patent	.859	.359	2.39	.017	.152	1.567	**
population	-.004	.001	-4.82	0	-.006	-.002	***
perc_hs_grad_rate	-49232.643	8732.167	-5.64	0	-66433.043	-32032.244	***
gdp	.043	.02	2.18	.031	.004	.083	**
pers_exp	.072	.031	2.31	.022	.011	.133	**
t	-1039.549	313.46	-3.32	.001	-1656.995	-422.102	***
Constant	43006.726	7548.5	5.70	0	28137.885	57875.568	***

Mean dependent var	10912.912	SD dependent var	20639.955
R-squared	0.905	Number of obs	250
F-test	168.150	Prob > F	0.000
Akaike crit. (AIC)	5101.019	Bayesian crit. (BIC)	5125.669

*** $p<.01$, ** $p<.05$, * $p<.1$

Accounting for a possible time trend indicates a statistically significant coefficient of time that results in a smaller homeless population each year after 2014 across all states. This could be for a number of reasons, but is most likely due to a federal program to help reduce homelessness that had varying degrees of success in each state, but did have an overall impact. The inclusion of this time trend did not have a large effect on the coefficient of patents, so our conclusion of the causality of patent count on homelessness still has evidence in support of it.

Beta coefficient

For the last part of this section, we chose to examine the regression if we used beta, or standardized coefficients instead of normal coefficients. We did this in order to compare the

actual effects of each variable and see which explanatory variable was actually the most important to look at. The results of this regression are shown in *Table 5*.

Table 5: Beta coefficients with regression

homeless	Coefficient	Robust std. err.	t	P>|t|	Beta
patent	.8593721	.3590826	2.39	0.017	.2689031
population	-.004041	.0008389	-4.82	0.000	-1.408421
perc_hs_grad_rate	-49232.64	8732.167	-5.64	0.000	-.1156259
gdp	.0434604	.0199731	2.18	0.031	.9992504
pers_exp	.0717867	.031046	2.31	0.022	1.050132
t	-1039.549	313.4604	-3.32	0.001	-.0713709
_cons	43006.73	7548.5	5.70	0.000	.

The results of this regression show that relative to other variables such as personal expenditure, high school graduation rate, and population, the effect of patent count on the homeless is significantly less. This is also important for policy decisions as it informs policymakers that while encouraging patents and innovation is important, there are other influencing factors that are perhaps more important.

Conclusion:

In conclusion, all of our original assumptions going into our analysis were met. When regressed using a linear model, our data supported our hypothesis that the higher the patent count in a given region, the higher the relative homeless population will be in that same region. However, while the weight on the patent count decreased significantly, and is now ~.88 it can still be used as a strong indicator of homelessness in a region. As the Patent Count increased, the variables that are encompassed within homelessness changed as well, along with the homelessness rate itself.