Independent Project: What Variables Drive Homelessness in the United States?

Braden Baseley

12/3/2019

1. Introduction

This paper seeks to understand what economic factors drive homelessness in the United States. According to the most recent data from the United States Department of Housing and Urban Development's (HUD) *Annual Homeless Assessment Report to Congress*, more than half a million Americans experienced homelessness on a given night in 2018, up slightly from 2017[1]. Many factors have been linked to homelessness in the United States, which typically include economic factors (e.g. income inequality and rising cost of living) and psychological factors (e.g. mental health and drug use). Homelessness leads to several socially sub-optimal outcomes, which can include poorer health among homeless individuals[2], lost productivity, and greater incarceration rates[3]. Therefore, if we can better understand what factors lead to homelessness in the United States, policymakers can intervene on at-risk populations in order to mitigate these problems.

To facilitate my analysis, I am collecting state-level data (hence, the unit of analysis is the state) for the year 2017. My dependent variable is homelessness per capita (specifically, the homeless rate per 10,000 people). I have included several independent variables: the Gini index (a measure of income inequality), the unemployment rate, and the median gross rent as a percentage of household income. I hypothesize that, all else equal, the Gini index and the homelessness rate have a positive relationship (i.e. as income inequality increases, so does the rate of homelessness). I also hypothesize that an increase in the unemployment rate leads to an increase in the homeless rate, holding all other variables constant. Moreover, I hypothesize that an uptick in median gross rent as a percentage of household income yields an increase in the homelessness rate, all else equal.

2. Description of Datasets and Variables

Unfortunately, I could not find a single dataset that included all of the variables of interest. As a result, I had to source and collate data from a few disparate sources: the United States Department of Housing and Urban Development (HUD), the U.S. Census Bureau (Census) and the Bureau of Labor Statistics (BLS). A brief summary of each of the dependent and independent variables, including how they were coded and sourced, can be found in the table below. A description for each variable is listed below the table.

Variables

Table 1: Description of Variables

Variable Name	Type of Variable	Coded Name	Data Source(s)
Homelessness Per Capita	Dependent	homeless	HUD & Census
Gini Index	Independent	gini	Census
Unemployment Rate	Independent	unemp	BLS
Median Gross Rent as a Percentage of Household Income	Independent	rent_income	Census

[1] https://www.hud.gov/press/press_releases_media_advisories/HUD_No_18_147
[2] https://www.cdc.gov/phlp/publications/topic/resources/resources-homelessness.html
[3] https://www.prisonpolicy.org/reports/housing.html

(1) **homeless**: This variable measures the number of homeless persons per 10,000 people. I calculated this rate by dividing the number of homeless people (sourced from HUD) in each state by its total population (sourced from Census), and then multiplying the value by 10,000. For a detailed description of how HUD defines the homeless population in the United States, please see part 2A below.

(2) **gini**: This variable measures income inequality. It can take on a value between 0 and 100, with 0 indicating perfect equality (i.e. every member receives an equal share of income) and 100 indicating perfect inequality (i.e. only one recipient or group of recipients receives all the income). Although the Census Bureau calculates the Gini index on a scale from 0 to 1, I recoded the variable by multiplying each value by 100 in order to make it easier to interpret.

(3) **unemp**: This variable measures the number of unemployed people in each state as a percentage of its labor force, which can be used as one proxy for the health of the economy. BLS defines an unemployed person as one who does not have a job, has actively looked for work in the past 4 weeks, and is currently available for work. The labor force is defined as all people age 16 and older who are either employed or unemployed. BLS calculates this every month. Therefore, to get yearly estimates for every state, I averaged each state's monthly unemployment rates over the course of 2017.

(4) **rent_income**: This variable reflects the median percentage of household income spent on gross rent, which can be thought of as a measure of housing affordability. Typically, 30% is believed to be the maximum share of income that households should allocate to rent, otherwise they face a considerable housing burden. Census calculates this statistic directly, so I did not have to recode it.

Description of Datasets

2A: HUD

HUD provides a point-in-time count of the number of homeless people in about 3,000 cities and counties in the United States. The point-in-time data collection process occurs every year on a single night in January, wherein state and local planning agencies (known as "Continuums of Care") work alongside volunteers to identify individuals and families living in emergency shelters, transitional housing programs and unsheltered settings[4]. Currently, HUD defines homelessness in four ways. Therefore, if any individual or family falls into one of these four categories, they are considered homeless[5]:

1. "Individuals and families who lack a fixed, regular, and adequate nighttime residence and includes a subset for an individual who is exiting an institution where he or she resided for 90 days or less and who resided in an emergency shelter or a place not meant for human habitation immediately before entering that institution."
2. "Individuals and families who will imminently lose their primary nighttime residence."
3. "Unaccompanied youth and families with children and youth who are defined as homeless under other federal statutes who do not otherwise qualify as homeless under this definition."
4. "Individuals and families who are fleeing, or are attempting to flee, domestic violence, dating violence, sexual assault, stalking, or other dangerous or life-threatening conditions that relate to violence against the individual or a family member."

One potential downside to the point-in-time estimates is that there is variation in count methodology within and across communities each year. However, this should not be much of a problem for this particular project because I am only looking at one year of data as opposed to a time series. According to the National Alliance to End Homelessness, "the annual point-in-time counts result in the most reliable estimate of people experiencing homelessness in the United States from which progress can be measured."[6] This serves as my rationale for using HUD's point-in-time estimates for measuring homelessness as opposed to other sources.

2B: U.S. Census Bureau

[4]https://www.hud.gov/press/press_releases_media_advisories/HUD_No_18_147
[5]https://www.hud.gov/sites/documents/PIH2013-15HOMELESSQAS.PDF
[6]https://endhomelessness.org/resource/what-is-a-point-in-time-count/

My second dataset is the American Community Survey (ACS), which is conducted by the Census Bureau. The ACS surveys roughly 3.5 million households in the United States every year. Samples are drawn from the Census Bureau's Master Address File (MAF), which is a list of all known living quarters and selected nonresidential units in the country. The ACS includes estimates of social characteristics, housing characteristics, economic characteristics and demographic characteristics for myriad geographic areas, such as states (including the District of Columbia and Puerto Rico), counties, cities, school districts, congressional districts, census tracts and block groups. I am pulling state populations, Gini indexes and median percentage of household income spent on gross rent estimates from the ACS.

The ACS can be broken down into 1-year releases and 5-year releases. ACS 1-year estimates are compiled annually for geographic areas that have at least 65,000 people. For areas with populations smaller than 65,000 people, the Census Bureau pools 5 consecutive years of ACS data to come up with the 5-year release. As such, the 5-year ACS program offers statistically more reliable results compared to the 1-year ACS program. At the same time, the 1-year ACS program is the most current data, whereas the 5-year data is less current since it includes years-old information. This means that there is a trade off between the currency of the data versus its statistical reliability when comparing the 1-year program against the 5-year program. For the sake of this project, I shall use the 5-year estimates due to their statistical reliability.

2C: Bureau of Labor Statistics

Lastly, I shall use the BLS Current Employment Statistics (CES) program to collect unemployment data. The CES program is a stratified, simple random sample of worksites. According to the BLS, the CES program surveys about 142,000 businesses and government agencies (or 689,000 individual worksites) across the country in order to gather data on nonfarm employment levels, hours of work and earnings of workers on payrolls. The worksites, which are clustered by Unemployment Insurance (UI) account number, is a major identifier on the Longitudinal Database (LDB) of employer records. This serves as both the sampling frame and the benchmark source for the BLS' estimates. The sample strata are defined by three characteristics: state, industry and employment size. Optimum allocation, which distributes a fixed number of sample units across a set of strata to minimize the sampling error, determines the sampling rates for each stratum. The CES program reports data at varying levels of granularity, including national, state (including the District of Columbia, Puerto Rico and the Virgin Islands) and local (about 450 metropolitan areas).

3. Descriptive Statistics

Table 2: Summary Statistics

	mean	sd	median	min	max	range	skew	kurtosis
homeless	14.58068	9.7589761	10.97785	4.929309	50.78577	45.85646	2.0103709	3.7009016
gini	46.35500	1.8846114	46.37000	41.800000	51.29000	9.49000	-0.0708842	-0.0920540
rent_income	29.40000	1.7826831	29.50000	24.900000	33.60000	8.70000	0.0230244	0.3439386
unemp	4.15000	0.8981273	4.25000	2.400000	7.00000	4.60000	0.3390669	0.4759403

I shall begin by looking at the dependent variable. The average rate of homelessness is about 14.58068 people per 10,000, which is greater than the median rate of homelessness (10.97785 people per 10,000). A mean greater than its median implies a right-skewed distribution, which I confirm by looking at the calculated value of skewness (2.0103709). Skewness values that are positive are indicative of right-skewed distributions, whereas negative values are indicative of left-skewed distributions. The spread of the *homeless* distribution also appears to be quite large, with a range of 45.85646 people per 10,000. The standard deviation of the distribution, which represents how far on average the values of the data are spread from the mean, is 9.7589761 people per 10,000. Ultimately, based on these descriptive statistics, I infer that the distribution is not normal. To confirm my intuition, I ran a Shapiro-Wilk Normality test. This test posits that, under the null hypothesis, the distribution is normal. My p-value for this test is less than 0.05, meaning that I can safely reject the null hypothesis and conclude the distribution is not normal.

Next, I will look at the independent variables. The average Gini index value is 46.355, which is roughly identical to its median of 46.37. Additionally, the skewness of the distribution is nearly 0, which implies its distribution is roughly symmetrical. The range is quite small (9.49), as is the standard deviation (1.8846114), which suggests that most of the values are very close to the mean. Similarly, the mean for *rent_income* (29.4%) is very close to its median (29.5%), with a skewness also very close to 0. As such, the distribution for *rent_income* is also roughly symmetrical. The range of *rent_income* is also quite small (8.7%); moreover, the standard deviation (1.7826831%) is minuscule, so most of the values are close to the mean. Lastly, the average unemployment rate is 4.15% for the dataset, which is close to its median of 4.25%. Additionally, the range appears to be quite low (4.6%), and the standard deviation is very low (0.8981273%). As such, most of the values are close to the mean. I ran a Shapiro-Wilk Normality test for each of the independent variables individually, and all of them have a p-value greater than 0.05. As a result, I cannot reject the null hypothesis, so I conclude that the independent variables are all normal.

4. Initial Models

Model 1

I begin my running a multivariate linear regression model. The equation for Model 1 is given by:

$$homeless = \alpha + \beta_1 gini + \beta_2 rent\_income + \beta_3 unemp$$

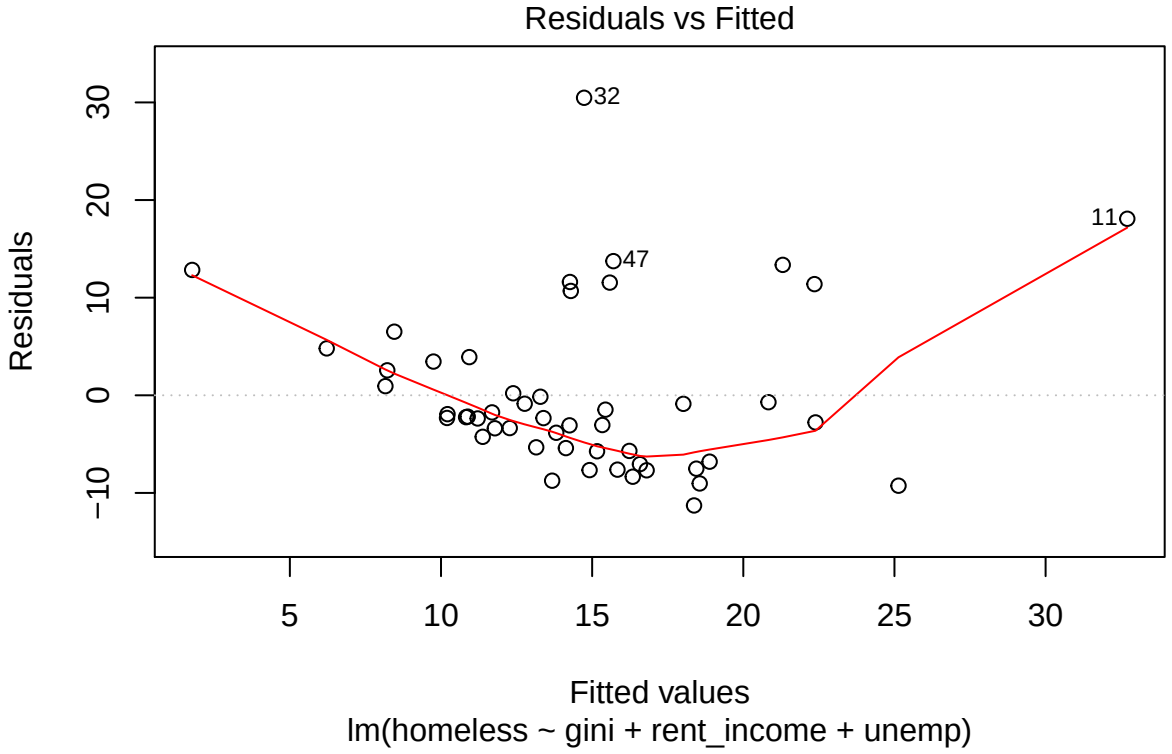
The results for Model 1 are given in Table 3 below.

Table 3: Model 1 Results

	Dependent variable:
	homeless
gini	-154.659^{*}
	(80.594)
rent_income	3.344^{***}
	(0.819)
unemp	-1.047
	(1.465)
Constant	-7.710
	(31.158)
Observations	50
R^2	0.269
Adjusted R^2	0.221
Residual Std. Error	8.611 (df = 46)
F Statistic	5.643^{***} (df = 3; 46)
Note:	*p<0.1; **p<0.05; ***p<0.01

Looking at the results, *gini* is statistically significant only at the 10% level, whereas *rent_income* is statistically significant at the 1% level. I will begin by giving an overview of statistical significance by looking at *gini* as an example. The t-statistic for *gini* is -1.919 (-154.659/80.594). Therefore, given that the p-value is p<0.1, there is less than a 10% chance that we would get a t-statistic as large as +1.919 or -1.919 simply due to chance. As such, we have evidence against the null hypothesis that the coefficient is equal to 0.

The coefficient for *gini* surprised me given that it is negative when I hypothesized it would be positive. In particular, all else equal, a 1 point increase in the Gini index leads to an average drop in the homelessness rate of about 154.659 people per 10,000. I suspect that there is actually a non-linear relationship between the Gini index and the homelessness rate, which would explain the questionable estimates. Meanwhile, the sign of the *rent_income* variable matched my expectations. Holding all other variables constant, given a 1 percentage point increase in median gross rent as a percentage of household income, the homelessness rate increases by 3.344 people per 10,000 on average. The coefficient for *unemp* is negative (-1.047), although it is not statistically significant. This would imply that, given a 1 percentage point increase in the unemployment rate, the homelessness rate would decline on average by 1.047 people per 10,000, all else equal. The y-intercept of Model 1 is -7.71, which means that if all of the independent variables were set equal to 0, the expected homelessness rate would be -7.71 people per 10,000, which of course does not make sense in practical terms. The adjusted R^2 for Model 1 is 0.221, which means that about 22.1% of the variation in the homelessness rate is explained by the independent variables in the model. Nevertheless, the plot of the residuals against fitted values shows a clear U-shape, which suggest that there is non-linearity in the data. As such, I will try to account for this in the next model.



Model 2

I shall make some adjustments to the model given my results from Model 1. As mentioned earlier, I suspect there may be some non-linear relationships between the homelessness rate and some of the independent variables, especially after looking at some scatterplots. To test whether or not it makes sense to add quadratic terms to the model, I used the `resettest` function on 3 simple linear models (homeless ~ gini, homeless ~ unemp, and homeless ~ rent_income). The null hypothesis of the RESET test says that no higher power of the independent variables would fit the data better. P-values below 5% indicate that we can reject the null hypothesis, which means the functional form is misspecified.

For *gini*, the p-value for the RESET test is 0.001289, which means that I can reject the null hypothesis that no higher powers of *gini* would fit the data better. As such, I will add a squared term for *gini*. Likewise, the p-value of the RESET for *rent_income* was also less than 0.05, so I will add a squared term for this variable as well. However, the RESET test for *unemp* had a p-value greater than 0.05, so I will not be including a quadratic term for this variable.

The formula for Model 2 is given by:

$$homeless = \alpha + \beta_1 gini + \beta_2 gini^2 + \beta_3 rent\_income + \beta_4 rent\_income^2 + \beta_5 unemp$$

The results of Model 2 can be found in Table 4 below.

Table 4: Model 2 Results

	Dependent variable:
	homeless
gini	$-5{,}133.678^{**}$
	(2,240.892)
I(giniˆ2)	$5{,}398.955^{**}$
	(2,433.206)
rent_income	-25.441^{*}
	(14.505)
I(rent_incomeˆ2)	0.484^{*}
	(0.246)
unemp	-0.960
	(1.452)
Constant	$1{,}564.626^{***}$
	(494.748)
Observations	50
R^2	0.435
Adjusted R^2	0.371
Residual Std. Error	7.739 (df = 44)
F Statistic	6.785^{***} (df = 5; 44)
Note:	*p<0.1; **p<0.05; ***p<0.01

The adjusted R^2 for Model 2 is 0.371, meaning that the covariates explain about 37.1% of the variation in the homelessness rate. As such, Model 2 is an improvement compared to Model 1 with respect to the adjusted R^2. The coefficients on both *gini* and $gini^2$ are statistically significant at the 5% level. The coefficient on *gini* is negative while the coefficient on $gini^2$ is positive, which implies that *gini* has an increasing effect on the homelessness rate. In other words, for low values of *gini*, a 1 point increase in the Gini index leads to a decline in the homelessness rate. After a certain point (specifically, when the Gini index exceeds 0.475, which is the absolute value of $\beta_1/(2\beta_2)$), a 1 point increase in the Gini index leads to an increase in the homeless rate. Because the cutoff point is so low (and no states in the dataset have a Gini index lower than 41.8), this largely supports my initial hypothesis that there is a positive relationship between the Gini index and the homelessness rate.

The coefficients on both *rent_income* and $rent\_income^2$ are both statistically significant at the 10% level. The coefficient on *rent_income* is negative while the coefficient on $rent\_income^2$ is positive, which also implies that *rent_income* has an increasing effect on the homelessness rate. In particular, for low values of *rent_income*, a 1 percentage point increase in *rent_income* leads to a decline in the homeless rate. However, after a certain threshold (i.e. when *rent_income* exceeds 26.282), a 1 point increase in *rent_income* leads to an increase in the homeless rate. In fact, only 3 states in the dataset (6% of the sample) have a *rent_income* value less than 26.28, so I can conclude that the relationship is mostly a positive one. I hypothesized that

there is a positive relationship between *rent_income* and the homelessness rate because as housing becomes less affordable, more people are crowded out of the rental market. The increasing effect observed here largely supports this hypothesis.

The coefficient for *unemp* is still negative and statistically insignificant. I believe there may be some omitted variable bias that could be causing this odd relationship, which I will attempt to reconcile in Model 3. Lastly, the y-intercept of the model is 1,564.626. This means that, given a value of 0 for all of the independent variables, the average homelessness rate is 1,564.626 people per 10,000. Unlike Model 1, this value is positive albeit extremely large.

Model 3

I believe there may be other unobserved factors that are affecting the parameter estimates. In particular, I suspect there are many other demographic features that correlate with the homelessness rate (e.g. racial makeup of the state, proportion of veterans in the state, age demographics of the state, etc.). Additionally, I hypothesize that less harsh climates are more prone to issues of homelessness because homeless people would seemingly rather sleep outside where it is warm rather than someplace cold. To account for this, I will add a *region* variable to the model, which is a categorical variable that can take on four values: Northeast, South, North Central and West. These are just the default regions included in R Studio, so I use them for simplicity. I posit that homelessness varies by region, which likely stems from variation in some the aforementioned factors. A summary of the proportion of states in each region can be found in the table below.

Table 5: Proportion of States in Each Region

Var1	Freq
Northeast	0.18
South	0.32
North Central	0.24
West	0.26

The equation for Model 3 is given by:

$$homeless = \alpha + \beta_1 gini + \beta_2 gini^2 + \beta_3 rent\_income + \beta_4 rent\_income^2 + \beta_5 unemp + \beta_6 region$$

The results for Model 3 are given in Table 6 below.

Table 6: Model 3 Results

	Dependent variable:
	homeless
gini	−3,846.650*
	(2,118.657)
I(giniˆ2)	4,170.305*
	(2,300.578)
rent_income	−20.804
	(13.546)
I(rent_incomeˆ2)	0.390*
	(0.229)
unemp	−1.772
	(1.374)
regionSouth	−4.918
	(3.101)
regionNorth Central	−1.956
	(3.651)
regionWest	7.149**
	(3.463)
Constant	1,181.398**
	(452.618)
Observations	50
R^2	0.603
Adjusted R^2	0.526
Residual Std. Error	6.721 (df = 41)
F Statistic	7.789*** (df = 8; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

Among all of the *region* variables added to Model 3, only the region *West* is statistically significant (at the 5% level). In order to test if the *region* variables are *collectively* significant, I ran a partial F-test comparing Model 3 against Model 2, the results of which can be seen in Table 7 below. A partial F-test is used to test whether or not adding additional variables improves the explanatory power of the model. This is done by comparing a model with the additional variables (called the unrestricted model) to a model that excludes said variables (called the restricted model). If the p-value from the partial F-test is low enough, it implies that the unrestricted model has a better fit compared to the restricted model since the additional variables are all jointly not equal to zero.

The p-value is 0.0021642, which means that there's a very low chance that the *region* variables are jointly equal to 0. Although two of the regions are statistically insignificant, collectively, the *region* variables are a statistically significant predictor of the homelessness rate. According to the output, the *West* region has a homelessness rate that is greater than the homelessness rate in the *Northeast* (the base region) by 7.149 people per 10,000 on average, net of other factors. Meanwhile, the *South* region has a homelessness rate that is lower than the homelessness rate in the *Northeast* by about 4.918 people per 10,000 on average, all else equal. Lastly, the *North Central* region has a homelessness rate that is lower than the homelessness rate in the *Northeast* by 1.956 people per 10,000 on average, net of other factors.

Table 7: Partial F-Test Results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
44	2634.989	NA	NA	NA	NA
41	1851.949	3	783.0398	5.778529	0.0021642

The coefficient for *gini* and $gini^2$ are both statistically significant at the 5% level. Like Model 2, the coefficient for *gini* is negative and the coefficient for $gini^2$ is positive, which implies that *gini*'s increasing effect on the homelessness rate still holds for this model. The cutoff point is still quite low: When the Gini index exceeds 0.461, a 1 point increase in the index leads to growth in the homelessness rate, all other variables held constant. The coefficient for $rent\_income^2$ is statistically significant only at the 10% level. However, the coefficient for *rent_income* is no longer statistically significant. The signs of these coefficients remain unchanged from Model 2, with *rent_income* being negative and *rent_income*2 being positive. Thus, after a certain threshold (i.e. when *rent_income* exceeds 26.672), a 1 point increase in *rent_income* leads to an increase in the homeless rate, all else equal. Meanwhile, the coefficient for *unemp* still remains negative and statistically insignificant. The adjusted R^2 is 0.526, which means that 52.6% of the variation in the homelessness rate is explained by the model. This is a considerable increase compared to Model 2.

Model 4

My last model will make a slight adjustment to Model 3. When looking at the descriptive statistics earlier, I concluded that the distribution for the homelessness rate is not normal. To make it more normal, I will take the natural log of the dependent variable while leaving all of the other independent variables the same. Hence, the formula is given by:

$$log(homeless) = \alpha + \beta_1 gini + \beta_2 gini^2 + \beta_3 rent\_income + \beta_4 rent\_income^2 + \beta_5 unemp + \beta_6 region$$

The results are given in Table 8 below. The adjusted R^2 is 0.529, which means that 52.9% of the variation in the homelessness rate is explained by the model. This figure is only marginally higher than the R^2 for Model 3 (52.6%). Interestingly, although the signs of the coefficients for *gini* and $gini^2$ remain unchanged from Model 3, they are no longer statistically significant in this model. Meanwhile, both coefficients for *rent_income* and $rent\_income^2$ are statistically significant at the 10% level. Like Model 3, the coefficients for *rent_income* and $rent\_income^2$ are negative and positive, respectively. The coefficient on *unemp* is still negative (-0.071) and statistically insignificant. This means that, given a 1 percentage point increase in the unemployment rate, the homelessness rate would fall by 7.1% on average, all else equal.

Table 8: Model 4 Results

	Dependent variable:
	log(homeless)
gini	−113.599
	(110.998)
I(giniˆ2)	122.579
	(120.529)
rent_income	−1.321*
	(0.710)
I(rent_incomeˆ2)	0.024*
	(0.012)
unemp	−0.071
	(0.072)
regionSouth	−0.419**
	(0.162)
regionNorth Central	−0.276
	(0.191)
regionWest	0.317*
	(0.181)
Constant	47.423*
	(23.713)
Observations	50
R^2	0.606
Adjusted R^2	0.529
Residual Std. Error	0.352 (df = 41)
F Statistic	7.878*** (df = 8; 41)
Note:	*p<0.1; **p<0.05; ***p<0.01

5. Model Summary

Table 9: Summary of Models

	Dependent variable:			
	homeless			log(homeless)
	(1)	(2)	(3)	(4)
gini	−154.659*	−5,133.678**	−3,846.650*	−113.599
	(80.594)	(2,240.892)	(2,118.657)	(110.998)
I(giniˆ2)		5,398.955**	4,170.305*	122.579
		(2,433.206)	(2,300.578)	(120.529)
rent_income	3.344***	−25.441*	−20.804	−1.321*
	(0.819)	(14.505)	(13.546)	(0.710)
I(rent_incomeˆ2)		0.484*	0.390*	0.024*
		(0.246)	(0.229)	(0.012)
unemp	−1.047	−0.960	−1.772	−0.071
	(1.465)	(1.452)	(1.374)	(0.072)
regionSouth			−4.918	−0.419**
			(3.101)	(0.162)
regionNorth Central			−1.956	−0.276
			(3.651)	(0.191)
regionWest			7.149**	0.317*
			(3.463)	(0.181)
Constant	−7.710	1,564.626***	1,181.398**	47.423*
	(31.158)	(494.748)	(452.618)	(23.713)
Observations	50	50	50	50
R^2	0.269	0.435	0.603	0.606
Adjusted R^2	0.221	0.371	0.526	0.529
Residual Std. Error	8.611 (df = 46)	7.739 (df = 44)	6.721 (df = 41)	0.352 (df = 41)
F Statistic	5.643*** (df = 3; 46)	6.785*** (df = 5; 44)	7.789*** (df = 8; 41)	7.878*** (df = 8; 41)
Note:				*p<0.1; **p<0.05; ***p<0.01

In summary, Model 4 has the highest adjusted R^2 among all of the models I ran. I started with a simple linear model and gradually added complexity by transforming the variables (e.g. logarithms, polynomials) to account for nonlinearities. I diagnosed these problems in various ways, such as by looking at plots of residuals vs. fitted values and using the *resettest* function. Additionally, I added a *region* variable to control for regional differences in demographics and climate, which improved the adjusted R^2 considerably.

6. Conclusion

I learned quite a lot during this project. To begin, I sharpened my ability to collect and clean data from disparate sources in order to generate a dataset (this took much longer than I imagined). The project gave me

considerable insight into the "trial and error" process of doing social science research. It's a highly iterative process wherein you run a model, analyze the results, run some diagnostics, make some adjustments to the model, and rinse and repeat. Throughout the process, I found that I was able to confirm some of my initial hypotheses. In particular, I was able to find largely positive relationships between the homelessness rate and the Gini index/median gross rent as a share of household income variables. Across all of my models, the coefficient for unemployment only ever took on a negative values, which challenged my initial hypothesis. At the same time, the coefficient was statistically insignificant across the models, which makes me believe that I still have more work to do. I believe I would do well to add additional controls to the model, perhaps touching on covariates associated with mental health, drug use and/or funding for homelessness initiatives. Ideally, I would like to collect a time series dataset and try some more complex models if I were to continue researching this problem.