THE EFFECT OF WEATHER SHOCKS ON HOUSING PRICES IN CALIFORNIA COUNTIES

GIANLUCA CRESCENZO

Abstract. Using California county data from 2010-2019, we investigate how year-to-year weather shocks shape local house prices. We've computed the temperature and precipitation z-scores to measure deviations in each year's weather patterns relative to a county's 1901-2000 climate, then merge them with Census median-price data and population controls found from California's Department of Finance. Two panel regressions are estimated —a baseline model and one which implements county and year fixed effects. We will also consider robust effects on our baseline model. Across all versions, unusually warmer years lead to a small rise in median prices, while precipitation shocks show no consistent effect. This is an example of a citation [Bur20]

1. Introduction

2. Background

We can formally measure temperature and precipitation shocks by computing z-scores of temperature and precipitation data. Let Temp_{ct} denote the average temperature of county c during year t and $\overline{\text{Temp}_c}$ denote the historical average temperature of county c. The *temperature z-score* is given by:

$$z\text{Temp}_{ct} = \frac{\text{Temp}_{ct} - \overline{\text{Temp}_c}}{\sigma_{\text{Temp},c}}, \tag{1}$$

where $\sigma_{\text{Temp},c}$ are the county-specific temperature standard deviations. The precipitation z-scores are defined similarly. Conceptually, Equation 1 measures the number of standard deviations by which a county's average temperature/precipitation differs from its long-run historical mean. For example, if $z\text{Temp}_{ct} = 0$, the average temperature in county c during year t differs by 0 standard deviations from the long-run historical average. Likewise, if

$z\text{Precip}_{ct} = -1$, then the average precipitation in county c during year t differs by -1 standard deviations compared to the long-run historical average; i.e., it was an unusually dryer year.

Given the purpose of this paper, there are many advantages in considering z-scores rather than raw temperature and precipitation values. The conversion of our data into z-scores allows us to express each year's weather data in units of a given county's historical volatility, hence a value of $z\text{Temp}_{ct} = +1$ is equally "unusual" regardless of the specified county. Moreover, the estimated coefficient on a z-score directly shows the effect of a one standard deviation change. As such, the computed estimators in this paper are easily interpretable as well as unit-free.

Because these weather shocks are essentially random from the perspective of buyers and sellers, they give us a source of exogenous variation, that is, year-to-year swings in temperature and precipitation have no effect on local housing markets. This makes $z\text{Temp}_{ct}$ and $z\text{Precip}_{ct}$ valuable estimators for our regression, any predicted effect is less likely to be caused by any extraneous economic effect. In practice, a hot, dry year could raise cooling costs, effect construction schedules, or change the desirability of outdoor amenities. Likewise, volatility in weather patterns can be closely linked with violent crime, another factor which effects housing prices. *Note that our model merely describes a relationship between weather deviations and median housing prices. Based off the few examples given, we have no way of precisely knowing which channel weather shocks take to have an effect on housing prices.*

3. Data

This paper uses panel data on California counties from 2010 to 2019. The main independent variables of this paper are the temperature and precipitation z-scores computed using Equation 1. The temperature and precipitation averages through January-December of a given year were found from the National Centers for Environmental Information. Average temperatures were given in $^\circ F$, whereas precipitation was measured in inches. Included in both data sets was the $1901-2000$ mean, which we used as the long-run historical average for

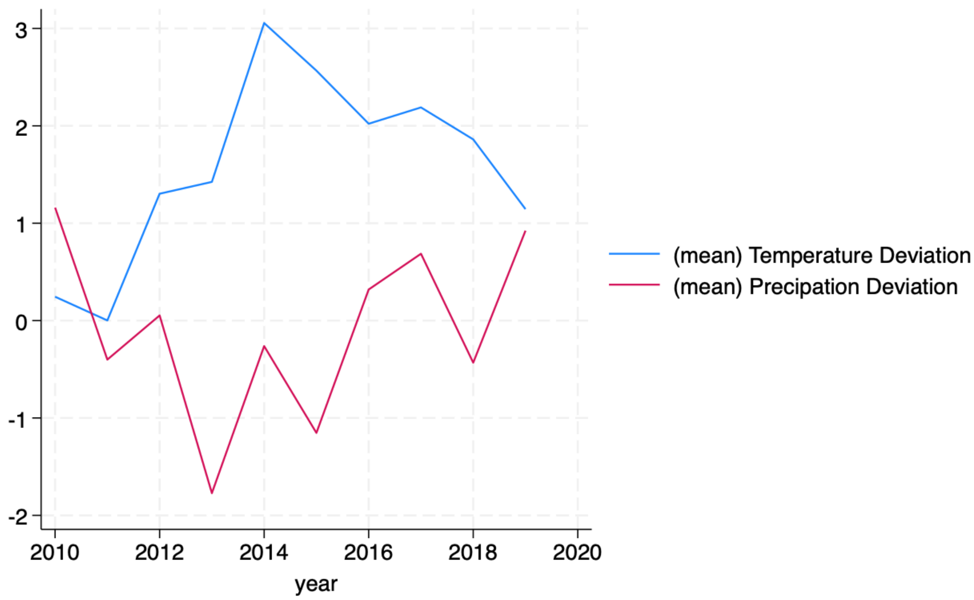
Equation 1. We were able to find data on temperature and precipitation for all 58 counties in California.

The main dependent variable of this paper is county median housing prices. The data was sourced from the United States Census Bureau. The data only includes counties with populations greater than or equal to 65000, lowering the amount of observed counties from 58 to 39.

Including in the regression are the county population estimates and population density between the years of 2010 and 2019. The population data was readily available from the State of California's Department of Finance. Rather than finding a data set which contains the population density of each county, we've instead computed it directly using data on California counties land area found from `https://onlinecalifornia.us/countyarea.shtml`. Population and population densities are included in this model because of their correlation with housing prices and independence of weather patterns. Omitting them would bias the estimated effect of temperature and precipitation z-scores, hence their inclusion helps isolate the causal impact of weather shocks on housing prices.

Table 1. Summary Statistics

	Mean	Standard Deviation	Minimum	Maximum
Year	2014.5	2.875971	2010	2019
Temperature	60.44974	4.511999	50.7	76.5
Historical Average Temperature	58.3359	4.1331	51.7	72.2
Precipitation	24.43574	17.71261	2.06	82.59
Historical Average Precipitation	24.90051	14.14373	3.38	63.39
County Population	975545.3	1674471	72348	1.01e+07
Median House Price	415952.6	229061.1	130800	1233600
County Square Miles	2721.245	3391.93	46.69	20052.5
Standard Deviation of Temperature	1.335813	.0912748	1.186779	1.521001
Temperature z-score	1.581177	.9970067	-.7024201	3.511163
Standard Deviation of Precipitation	9.357683	4.940643	1.295387	19.86239
Precipitation z-score	-.0875689	.9638408	-2.207122	2.308191
County Population Density	1025.875	2915.774	24.72207	18880.89
Log of Median House Price	12.80108	.5207476	11.78142	14.02545
N	390			



Within the sample, San Bernadino County is the largest by land area (approximately 20053 square miles), whereas San Francisco County is the smallest (approximately 47 square miles) and most densely populated, at roughly 189000 residents per square mile. At the sparse end of the distribution, Mendocino County has the lowest population density at 22 residents per square mile, with an approximate population size of 87000.

The largest temperature z-score was measured from Santa Barbara County during 2014. In fact, when sorting the 390 samples by temperature deviations, the top fourteen z-scores were *all* measured during 2014. The largest precipitation z-score was measured from Imperial County during 2010. The lowest precipitation z-score was measured from San Mateo County during 2013.

County level data, especially in California, will inevitably hide price differences across neighborhoods in large counties. Los Angeles County, for example, spans Long Beach, Downtown L.A., Santa Clarita, and Lancaster —all of which have drastically different housing markets. Combining these into a single county-wide median will most definitely introduce measurement error, thus weakening the estimated weather effect of our model. A better approach for future work would be to consider ZIP-code level data on median housing prices and weather patterns. Moreover, similar measurement error may be attributed to the land-area data found from [ref]. Although it could be considered negligible, without knowing the accuracy or credibility of the website one can assume that any mis-measurement will be absorbed into the model's error term.

4. Empirical Strategy

This paper will be investigating two different estimating equations. Let Housing_{ct} denote the median housing price in county c during year t, Population_{ct} the population of county c during year t, and PopDensity_{ct} the population density of county c during year t. $z\text{Temp}_{ct}$ and $z\text{Precip}_{ct}$ will be defined as per Equation 1. Our baseline model is the following:

$$\text{Housing}_{ct} = \beta_0 + \beta_1 z\text{Temp}_{ct} + \beta_2 z\text{Precip}_{ct} + \beta_3 \text{Population}_{ct} + \beta_4 + \text{PopDensity}_{ct} + \epsilon_{ct}. \quad (2)$$

As we've briefly mentioned in Section 3, we've included population and population density in the regression to increase the precision of β_1 and β_2, as well as curb any effect of omitted variable bias. Table 2 indicates that $\text{Cov}(\text{Population}_{ct}, z\text{Temp}_{ct})$ is non-zero with some statistical significance, suggesting that the exclusion of Population_{ct} from our regression would have contributed to omitted variable bias.

Table 2. Correlation Matrix

	Temperature z-score	Precipitation z-score	County Population	County Population Density
Temperature z-score	1			
Precipitation z-score	-0.292***	1		
County Population	0.115*	-0.0707	1	
County Population Density	0.0222	0.00637	0.163**	1

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

Equation 2 will give rise to concerns of omitted variable bias, the most drastic being the quality of education in a county. Good schools push local housing prices up, and hotter temperatures —an outcome of weather deviations —are known to be negatively correlated with education quality. Assuming $\text{EducationQuality}_{ct}$ is omitted from both Equation 2 and Equation **??**, and since $\text{Cov}(\text{EducationQuality}_{ct}, \text{Housing}_{ct})$ and $\text{Cov}(\text{EducationQuality}_{ct}, z\text{Temp}_{ct})$

are both non-zero, omitted variable bias will occur. Section **??** contains the effects and direction of this bias in more detail.

The most important result of this paper will be one which considers fixed effects. Let α_c denote the covariates which control for time-invariant effects in each county, and let δ_t denote the covariates which control for county-invariant effects each year between 2010 and 2019. Our regression in question is:

$$\text{Housing}_{ct} = \beta_0 + \beta_1 z\text{Temp}_{ct} + \beta_2 z\text{Precip}_{ct} + \alpha_c + \delta_t + \epsilon_{ct}. \tag{3}$$

Claim: β_1 and β_2 are *unbiased* estimators. Suppose X_{ct} is an omitted variable from Equation 3. It must be the case that it is neither a time nor county fixed effect; i.e., it is something that changes within a county over time, as well as differs across counties in a given year. It cannot be $z\text{Temp}_{ct}$ or $z\text{Precip}_{ct}$ as they are both included in the regression, and similarly it must be true that $\text{Cov}(\text{Housing}_{ct}, X_{ct}) \neq 0$, $\text{Cov}(X_{ct}, z\text{Temp}_{ct}) \neq 0$, and $\text{Cov}(X_{ct}, z\text{Precip}_{ct}) \neq 0$. But how can any covariate which changes within a county over time have an effect on weather deviations? How can any covariate which differs across counties in a given year have an effect on weather deviations? It must be the case that $\text{Cov}(X_{ct}, z\text{Temp}_{ct}) = \text{Cov}(X_{ct}, z\text{Precip}_{ct}) = 0$, giving that X_{ct} is not an omitted variable.

Note that the inclusion of county and year fixed effects in our regression will likely yield a high R^2 —we've effectively captured through $z\text{Temp}_{ct}$ and $z\text{Precip}_{ct}$ all the variance for our dependent variable.

5. Results

Our null-hypothesis for each of our regression equations is that temperature and precipitation do not affect median housing prices, which we denote as:

$$H_0 : \beta_1, \beta_2 = 0,$$

A failure to reject this hypothesis would indicate that temperature and precipitation deviations do not have any effect on our outcome variable. Rejecting the null-hypothesis in favour of our alternate hypothesis:

$$H_a : \beta_1, \beta_2 \neq 0,$$

indicates that there is a relationship between weather deviations and median housing prices.

Table 3. Baseline Panel Results

Dependent variable is the log of median house prices	(1)
Temperature z-score	0.107***
	(0.024)
Precipitation z-score	0.091***
	(0.025)
County Population	0.000***
	(0.000)
County Population Density	0.000***
	(0.000)
Constant	12.530***
	(0.042)
Observations	390
R^2	0.237

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

In our baseline a model, a one standard deviation change in average temperatures raises county median house prices by 10.7%, while a one standard deviation change in average precipitation results in a 9.1% change in median house prices. Both coefficients are statistically significant at the $p < 0.001$ level, and as such we can reject the null-hypothesis that temperature and precipitation deviations do not have an effect on median housing prices. [talk about economic significance here]. Although statistically significant at the $p < 0.001$ level, county population and population has little to no effect on median housing prices. Excluding them results in a $\widehat{\beta_1} = 0.11894$ and $\widehat{\beta_2} = .0905279$, indicating that omitted variable bias regarding population and population density is not a concern.

For robust results we've considered dropping Los Angeles County in Equation 2 as well as only considering coastline counties. For coastline counties, a +1 change in temperature

Table 4. Robust Results

Dependent variable is the log of median house prices	(1) Coastline Counties	(2) Exclude Los Angeles	(3) County & Year Fixed Effects
Temperature z-score	0.114***	0.101***	0.023
	(0.024)	(0.024)	(0.014)
Precipitation z-score	0.077***	0.090***	0.016**
	(0.025)	(0.026)	(0.007)
County Population	0.000***	0.000***	
	(0.000)	(0.000)	
County Population Density	0.000***	0.000***	
	(0.000)	(0.000)	
Year=2010			0.000
			(.)
Year=2011			-0.045**
			(0.017)
Year=2012			-0.109***
			(0.016)
Year=2013			-0.022
			(0.018)
Year=2014			0.010
			(0.034)
Year=2015			0.123***
			(0.029)
Year=2016			0.176***
			(0.030)
Year=2017			0.240***
			(0.031)
Year=2018			0.326***
			(0.030)
Year=2019			0.366***
			(0.026)
Constant	12.946***	12.498***	12.113***
	(0.049)	(0.044)	(0.225)
Observations	190	380	390
R^2	0.274	0.243	0.990

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

z-scores raised median house prices by 11.4%, a 0.7% increase compared to our baseline regression. Similarly, a unit change in precipitation z-scores lead to a 7.7% change in median house prices, a 1.3% decrease compared to our baseline regression. Excluding Los Angeles from our regression resulted in a negligible change in our $\widehat{\beta}_1$ and $\widehat{\beta}_2$ coefficients, indicating that our results are not driven by that outlier county. Excluding coastline counties and/or dropping Los Angeles County from the regression resulting in statistically significant estimators at the $p < 0.001$ level; i.e., we can reject the null-hypothesis that weather deviations do not affect median housing prices.

*We went into depth on the implications of county and year fixed effects on our model in Section 4, the most important being that, when included, results in our estimators β_1 and β_2 to be *unbiased*. Any county-level covariate which is time-invariant —[examples] —would have been picked up by the fixed effects, meaning that our estimators for Equation 2 are over-biased. The same is true for any covariate which gets picked up by year fixed effects. Our $\widehat{\beta}_1$ and $\widehat{\beta}_2$ estimators for Equaton 3 can be considered negligible —a one standard deviation change in average temperatures leads to a 2.3% increase in median housing prices, whilst a one standard deviation change in average precipitation leads to a 1.6% increase in median housing prices. We are unable to reject the null hypothesis for $\widehat{\beta}_1$ at the $p < 0.05$ level, while we can reject the null hypothesis for $\widehat{\beta}_2$ at the $p < 0.01$ level. Every year, excluding 2014, was statistically significant at the $p < 0.001$ level. This is quite surprising, as we mentioned in Section 3 that 2014 was an unusually warmer and dryer year compared to it's long-run historical average.

There are many reasons which indicate why our statistical significant disappears when including county and year fixed effects. Note that county fixed effects remove every time-invarait county trait, so all that is left for $\widehat{\beta}_1$ to estimate are year-to-year weather swings inside each county, which are much smaller

There are many reasons as to why our statistical significance dropped when including county and year fixed effects. Note that county fixed effect remove every time-invariant county trait, so all that is left for $\widehat{\beta}_1$ to estimate are year-to-year weather changes inside each county, which are much smaller. Moreover, our z-score is defined against the 1901-2000 mean, but our data only ranges from 2010-2019. Within such a small range, the inclusion of fixef effects leaves out too little independent variation for the coefficient to remain statistically significant.

Ultimately, the inclusion of fixed effects was able to remove bias from omitted variables at the cost of statistical significance. With less variation in our regression equation, the estimated effect of $\widehat{\beta}_1$ is both smaller and —despite the standard error falling from 0.024 to 0.014 —no longer large enough relative to its standard error to reject at the $p < 0.05$ level.

6. Conclusion

References

[Bur20] Benjamin A. Burton, *The next 350 million knots*, 36th International Symposium on Computational Geometry, LIPIcs. Leibniz Int. Proc. Inform., vol. 164, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2020, pp. Art. No. 25, 17. MR 4117738