Econ 388 Data Project 3

Mike King

17 April 2023

1 Introduction

The research question we are trying to answer is straightforward: what effect did county-level per-capita Covid-19 mortality rates in 2020 have on housing prices in 2021? This is an interesting and controversial topic. Covid-19 was a very hectic time for the country as a whole, and many perople were affected in different ways. The purpose of this memo is to address specifically how the mortality rates affected house prices, and to use casual inference to do so. This will be done through the use of econometrics as taught in Econ 388. This paper contains three sections: first a section discussing data sources, next a section discussing the technical details of the econometric analysis, and lastly a section discussing the conclusions of our analysis. All relevant tables and figures are included at the end of the file for reference.

2 Data

The data used by the analysis team was collected from various sources and contains the data required for several variables relevant to our analysis. We will discuss them one by one in detail.

The first data set contains the total number of Covid-19 cases and the total number of deaths related to Covid-19 broken down by county for all of the united states on a daily basis. This data set was obtained from the New York Times, who kept detailed numbers during all of 2020 and for some time following as well. This data set was filtered so that it contained only the data from the year 2020 and then sorted so that it contained total numbers of cases and deaths for each county for the whole year. These were used later on to form our right-hand-side variable of county-level per-capita Covid-19 mortality rates. This variable was formed as it is our main RHS variable of interest.

The second data set contains information about the age demographics of each county in the United States. This data set was used to form a variable of percent of population age 65 or older. Age 65 was used as a proxy to represent the age of retirement. Thus, this variable represents the percent of the population that is retired. This is important because it helps capture the different age characteristics of each county, which helps control for the effects of age demographics on housing prices.

The third data set contains information on the racial breakdown of each of the counties in the United States. This data was used to form RHS variables of the percentage of the population that identify as white, black, islander, Asian, and Native American/First Nation person. These variables help control for the effects of different racial demographics on on housing prices e.g. help represent the fact that parts of the country have different racial demographics, which may be correlated with differences in housing prices.

This data set also contains estimates of the population of each of the counties, which were used to calculate the previously mentioned percentage variables. These percentages were calculated simply by dividing the raw age/race numbers by these population estimates. Both this data set and the previously mentioned one were obtained from U.S. Census data, and therefore, hopefully have very accurate results.

The next data set contains data on the House Pricing Index. This data was obtained from the Federal Housing Finance Agency, and was used to find the values of our dependent variable as well as a lagged variable for HPI to compensate for serial correlation (more discussion on this in the Empirical Analysis section). This data was merged in with the other data using the "fips" code that was contained in the data set. The data containing HPI with the base year of 2000 was used.

The last data set contains data that classifies each state based on region. This data was used to for a categorical variable of region in our regression. The data breaks the United States into 9 total regions with the East North Central region being the base group. These categorical variables were used to control for the fact that different regions may have inherent differences in housing prices. This data was obtained from a csv found on GitHub that is linked in the README file. Summary statistics of each of the RHS variables can be found in figure 1.

3 Empirical Analysis

3.1 Serial Correlation

Initially, a regression was run that did not contain the lagged variable of HPI. The results of this regression can be seen in figure 2. In this regression, the t-stat for the coefficient on the county-level per-capita Covid-19 mortality rate variable is -.975, which has a p-value of .33. This means that our variable for the county-level per-capita Covid-19 mortality rate is not statistcally significant. The coefficient is also very large, and is coupled with a very large standard error. This is a very interesting result. However, the Durbin-Watson statistic is only 0.578 meaning that there is autocorrelation.

After this regression, another regression was run using the lagged HPI variable. This improved the Durbin-Watson score to 1.503, within the normal range of 1.5 to 2.5. In this new regression, the t-stat of our coefficient on the county-level per-capita Covid-19 mortality rate variable is 1.78, which has a p-value of .074. This means that we can reject the null-hypothesis that the coefficient is zero at the 10 percent confidence level! More details about this regression can be found in figure 3.

3.2 Testing for Regional Effect

As a part of the analysis, the handsome coding team also ran a test on the joint effect of the regional variables. Some of these variables have coefficients with t-stats that are very low and statistically insignificant, while others have t-stats that are large and are clearly very statistically significant. Thus, our wonderful coding team ran an f-test to find the joint significance of all of these 8 categorical variables. This code for this is included at the end of the python file. The f-stat was found to be approximately 11.1, showing that these categorical variables are clearly jointly significant!

4 Conclusion

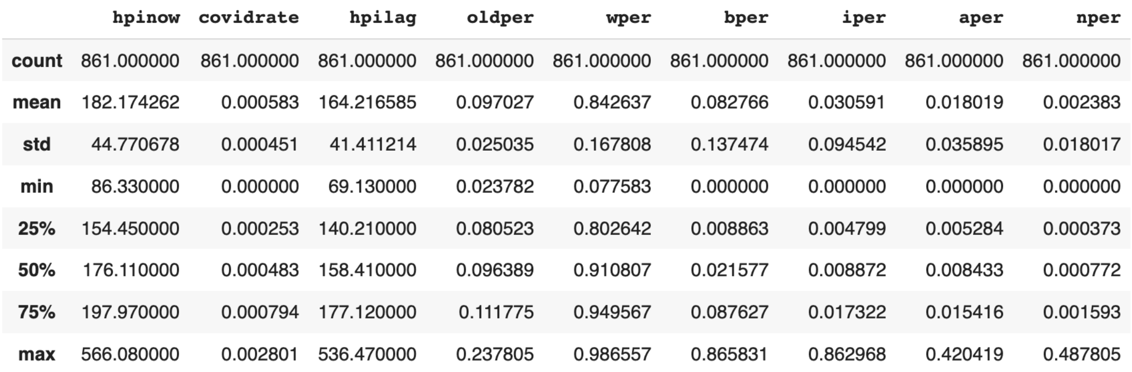
The most important results can be found in figure 3, the figure with the results from the regression with autocorrelation. This regression shows that the effects of the Covid mortality rate are statistically significant at the 10 percent confidence level. The coefficient on the Covid mortality rate says that for a 1 percent increase in the Covid mortality rate leads to a 9.78 percent increase in HPI. We also can say with recent certainly that this is causal; we have adjusted our model for serial correlation and used categorical variables to account for regional differences.

However, it is impossible to capture the effects of everything in our model. It is possible that the model misses some important variables that might change the result some. This is especially relevant to this model because the data on the Covid mortality rate coefficient seems so sensitive to changes in the model. This is because the variance in the rate is so high. This leads to very high standard error. It may be beneficial to add another level of serial correlation and or explore more effects, like income or otherwise, if the data can be procured. However, with an R-squared value of .981 in our regression model, adding more variables increase our risk of over-fitting.

Another interesting thing to mention is that the condition number of the matrices used to calculate our OLS estimators are very large. This is noted in a warning in both figures 2 and 3. These condition numbers may show multicolinearity, but more likely are a result of noise in the data. The condition numbers are not big enough to be of large concern either way.

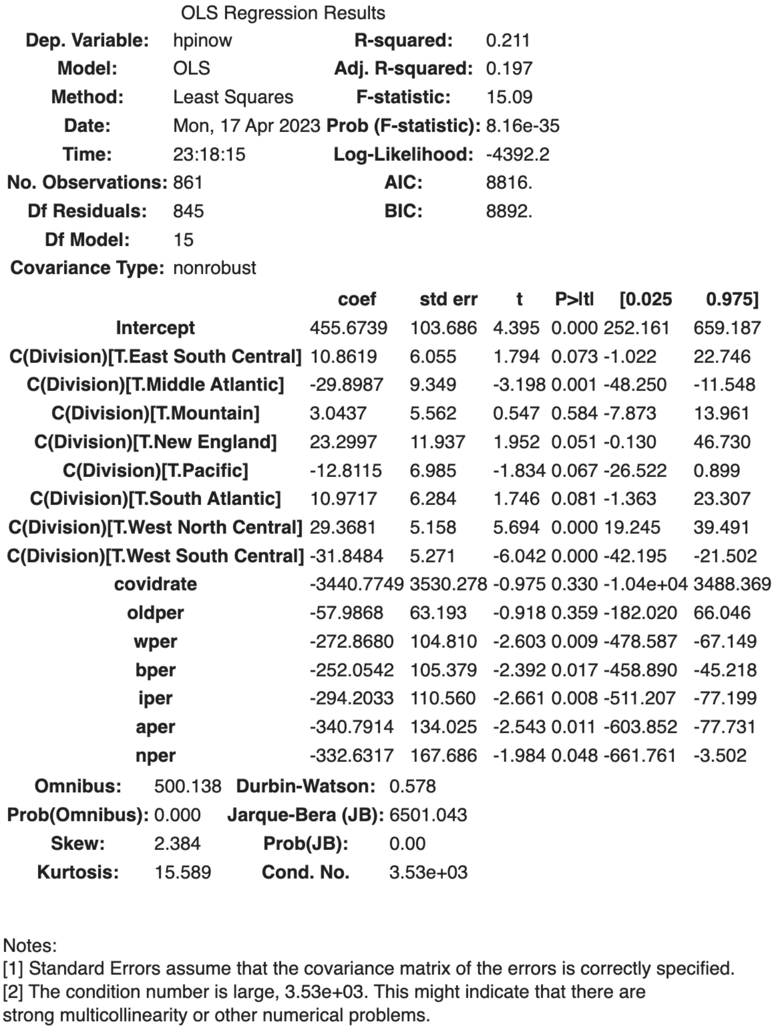
Please let us know what further steps you would like us to take with this model bossman.

5 Appendix: Figures



	hpinow	covidrate	hpilag	oldper	wper	bper	iper	aper	nper
count	861.000000	861.000000	861.000000	861.000000	861.000000	861.000000	861.000000	861.000000	861.000000
mean	182.174262	0.000583	164.216585	0.097027	0.842637	0.082766	0.030591	0.018019	0.002383
std	44.770678	0.000451	41.411214	0.025035	0.167808	0.137474	0.094542	0.035895	0.018017
min	86.330000	0.000000	69.130000	0.023782	0.077583	0.000000	0.000000	0.000000	0.000000
25%	154.450000	0.000253	140.210000	0.080523	0.802642	0.008863	0.004799	0.005284	0.000373
50%	176.110000	0.000483	158.410000	0.096389	0.910807	0.021577	0.008872	0.008433	0.000772
75%	197.970000	0.000794	177.120000	0.111775	0.949567	0.087627	0.017322	0.015416	0.001593
max	566.080000	0.002801	536.470000	0.237805	0.986557	0.865831	0.862968	0.420419	0.487805

Figure 1: This table shows summary statistics for each of the variables in the regression. This table was generated using the Python summary function. The variables are defined as follows: ”hpinow” is the HPI for 2021, ”covidrate” is the county-level per-capita Covid-19 mortality rate, ”hpilag” is the HPI from 2020, ”oldper” is the percentage of the population that is over the age of 65, ”wper” is the percentage of the population that identifies as white, ”bper” is the percentage of the population that identifies as black, ”iper” is the percentage of the population that identifies as pacific islander, ”aper” is the percentage of the population that identifies as Asian, ”nper” is the percentage of the population that identifies as Native American/First Nation person. It is assumed that these percentages are not colinear. This is a big assumption, but it is corroborated by the fact that condition numbers of the matrices used in OLS when all of the percentages are included are actually lower than the ones when one of the percentages is ommitted.



OLS Regression Results

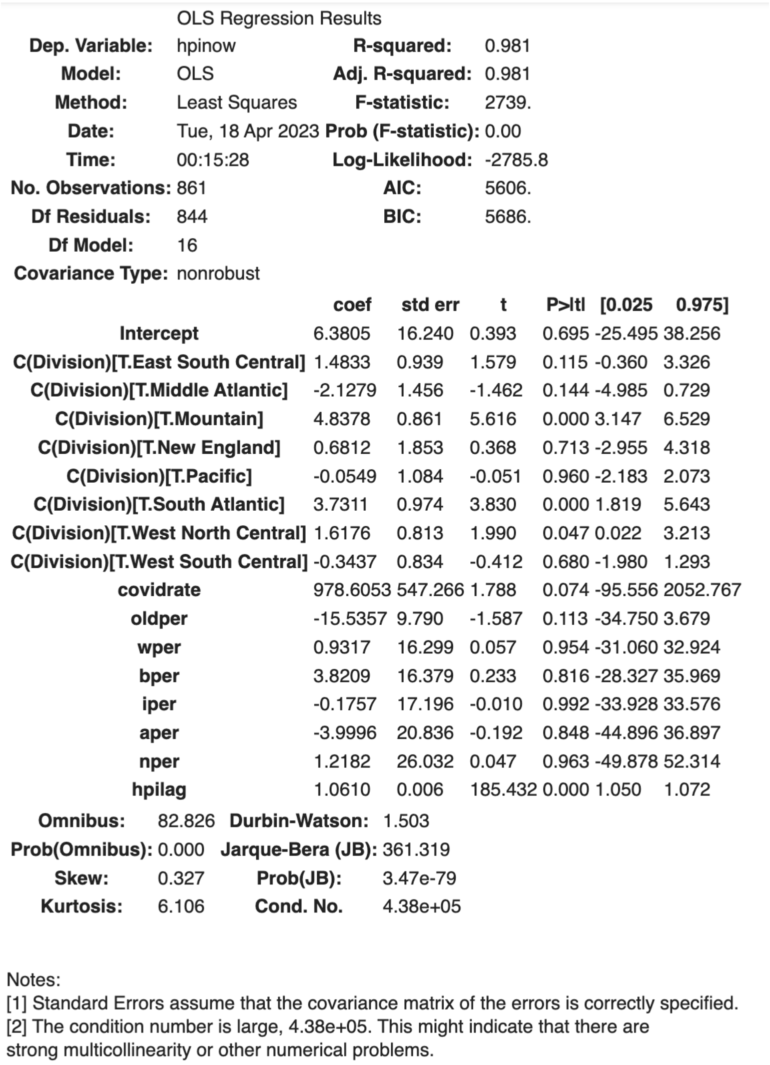
Dep. Variable:	hpinow	**R-squared:**	0.211
Model:	OLS	**Adj. R-squared:**	0.197
Method:	Least Squares	**F-statistic:**	15.09
Date:	Mon, 17 Apr 2023	**Prob (F-statistic):**	8.16e-35
Time:	23:18:15	**Log-Likelihood:**	-4392.2
No. Observations:	861	**AIC:**	8816.
Df Residuals:	845	**BIC:**	8892.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>ltl**	**[0.025**	**0.975]**
Intercept	455.6739	103.686	4.395	0.000	252.161	659.187
C(Division)[T.East South Central]	10.8619	6.055	1.794	0.073	-1.022	22.746
C(Division)[T.Middle Atlantic]	-29.8987	9.349	-3.198	0.001	-48.250	-11.548
C(Division)[T.Mountain]	3.0437	5.562	0.547	0.584	-7.873	13.961
C(Division)[T.New England]	23.2997	11.937	1.952	0.051	-0.130	46.730
C(Division)[T.Pacific]	-12.8115	6.985	-1.834	0.067	-26.522	0.899
C(Division)[T.South Atlantic]	10.9717	6.284	1.746	0.081	-1.363	23.307
C(Division)[T.West North Central]	29.3681	5.158	5.694	0.000	19.245	39.491
C(Division)[T.West South Central]	-31.8484	5.271	-6.042	0.000	-42.195	-21.502
covidrate	-3440.7749	3530.278	-0.975	0.330	-1.04e+04	3488.369
oldper	-57.9868	63.193	-0.918	0.359	-182.020	66.046
wper	-272.8680	104.810	-2.603	0.009	-478.587	-67.149
bper	-252.0542	105.379	-2.392	0.017	-458.890	-45.218
iper	-294.2033	110.560	-2.661	0.008	-511.207	-77.199
aper	-340.7914	134.025	-2.543	0.011	-603.852	-77.731
nper	-332.6317	167.686	-1.984	0.048	-661.761	-3.502

Omnibus:	500.138	**Durbin-Watson:**	0.578
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6501.043
Skew:	2.384	**Prob(JB):**	0.00
Kurtosis:	15.589	**Cond. No.**	3.53e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.53e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 2: The regression table for the regression without serial correlation. the variable "covidrate" is the variable for the county-level per-capita Covid-19 mortality rate. Note the "Durbin-Watson" statistic in at the bottom of the table, as referenced in the paper. Table was generated using python OLS summary packages.



OLS Regression Results

Dep. Variable:	hpinow	**R-squared:**	0.981
Model:	OLS	**Adj. R-squared:**	0.981
Method:	Least Squares	**F-statistic:**	2739.
Date:	Tue, 18 Apr 2023	**Prob (F-statistic):**	0.00
Time:	00:15:28	**Log-Likelihood:**	-2785.8
No. Observations:	861	**AIC:**	5606.
Df Residuals:	844	**BIC:**	5686.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.3805	16.240	0.393	0.695	-25.495	38.256
C(Division)[T.East South Central]	1.4833	0.939	1.579	0.115	-0.360	3.326
C(Division)[T.Middle Atlantic]	-2.1279	1.456	-1.462	0.144	-4.985	0.729
C(Division)[T.Mountain]	4.8378	0.861	5.616	0.000	3.147	6.529
C(Division)[T.New England]	0.6812	1.853	0.368	0.713	-2.955	4.318
C(Division)[T.Pacific]	-0.0549	1.084	-0.051	0.960	-2.183	2.073
C(Division)[T.South Atlantic]	3.7311	0.974	3.830	0.000	1.819	5.643
C(Division)[T.West North Central]	1.6176	0.813	1.990	0.047	0.022	3.213
C(Division)[T.West South Central]	-0.3437	0.834	-0.412	0.680	-1.980	1.293
covidrate	978.6053	547.266	1.788	0.074	-95.556	2052.767
oldper	-15.5357	9.790	-1.587	0.113	-34.750	3.679
wper	0.9317	16.299	0.057	0.954	-31.060	32.924
bper	3.8209	16.379	0.233	0.816	-28.327	35.969
iper	-0.1757	17.196	-0.010	0.992	-33.928	33.576
aper	-3.9996	20.836	-0.192	0.848	-44.896	36.897
nper	1.2182	26.032	0.047	0.963	-49.878	52.314
hpilag	1.0610	0.006	185.432	0.000	1.050	1.072

Omnibus:	82.826	**Durbin-Watson:**	1.503
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	361.319
Skew:	0.327	**Prob(JB):**	3.47e-79
Kurtosis:	6.106	**Cond. No.**	4.38e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.38e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 3: The regression table for the regression with serial correlation. the variable ”covidrate” is the variable for the county-level per-capita Covid-19 mortality rate. Note the ”Durbin-Watson” statistic in at the bottom of the table, as referenced in the paper. Table was generated using python OLS summary packages.