Introduction To The Difference-In-Differences Regression Model

We'll show how to use the DID model to estimate the effect of hurricanes on house prices

In this chapter, we will study the **Difference-In-Differences regression model**. The DID model is a powerful and flexible regression technique that can be used to estimate the differential impact of a 'Treatment' on the treated group of individuals or things.

We will also illustrate the use of the Difference-In-Differences regression model to estimate the effect of hurricanes on property prices in the United States.

Defining the terms: Treatment, treated group, control group

The words 'treatment' and 'treated group' may invoke a picture of a randomized controlled trial to test the efficacy of a drug or medical treatment.

While the DID model can indeed be used very effectively in that setting, in statistics, it is customary to ascribe a much broader interpretation to the word 'Treatment'. '**Treatment**' is any event that selectively affects only some of the individuals or things in a study. Examples of Treatment include an increase in state-mandated minimum wage that affects only restaurants in one state (as analyzed in the well-cited study by Card and Krueger in 1994 (https://econpapers.repec.org/article/aeaaecrev/v_3a84_3ay_3a1994_3ai_3a4_3ap_3a772-93.htm)), or the opening of a new airline route connecting two regions of a large country, or a natural disaster that affects only some parts of a country, or an experimental drug or medical procedure that is administered to only some of the participants in a study. In all these examples, the **unit** of study is respectively a restaurant, a town or a county, a county or a state, or a volunteer.

A study comprises many units (individuals or things) divided into a **treatment group** or a **control group** depending on whether they were or were not subjected to the treatment.

The response variable

In each of such studies, one wants to measure an outcome , a response, and know if it will achieve a mean value that is statistically different within the treatment group than in the control group. For example, the 1994 study by Card and Krueger analyzed whether an increase of minimum wage by New Jersey in 1992 from $4.25 to $5.05 resulted in a statistically significant change in **employment level** amongst fast food restaurant workers in New Jersey from that in neighboring Pennsylvania which did not change its minimum wage. Other examples of a response variable are SAT score of the participant, pollution level in a county, and tree cover in a country.

The Effect of Time

In practice, a complication is introduced by the passage of time. Whatever be the response variable being measured, be it SAT scores, employment level, house price inflation, or blood sugar level of participants, the natural flow of time will change the value of this variable in a potentially significant way as the study progresses from the pre-treatment to the post-treatment phase of the experiment. The experimenter must discount the partial effect of time (and the numerous hidden factors that time acts as a proxy for) on the change in the mean value of the response variable in both the control group and the treatment group. In other words, the experimenter must determine if the treatment itself caused any change in the mean value of the response variable within the treatment group that was *over and above* what was caused by the passage of time, *and*, whether this additional treatment-induced effect was observed much more in the treated group than in the control group.

The Difference-in-Differences (DID) regression model can be used to easily and quite elegantly perform all of the above mentioned analysis.

The fitted DID model will tell us whether there is evidence of a net-additional effect observed in the treated group that is purely treatment induced, the estimated value of this, whether this estimate is statistically significant and if so, the 95% or 99% confidence intervals are around the estimated effect.

Structure of the Difference-In-Differences model

The following equation illustrates the structure of the DID model:

$$y_i = \beta_0 + \beta_1 * Time\_Period_i + \beta_2 * Treated_i + \beta_3 * (Time\_Period_i * Treated_i) + \epsilon_i$$

The Difference-In-Differences regression model (Image by Author)

The first thing we note about this equation is that, it is that of **a linear regression model**.

y_i is the observed response for the *ith* observation. It is the value being measured in each group before and after treatment.

β_0 is the intercept of regression.

Time_Period_i is a dummy variable that takes the value 0 or 1 depending on whether the *ith* measurement refers to the pre or post treatment period respectively.

Treated_i is a dummy variable that takes the value 0 or 1 depending on whether the *ith* measurement refers to an individual in the control group or the treatment group respectively.

*(Time_Period_i*Treated_i)* is an interaction term. It stores the multiplication of the two dummy variable values for the *ith* observation.

ϵ_i is the error term associated with the *ith* observation and it captures the effect of all factors that the model was not able to adequately represent.

The two dummy variables in the model yield the follow 2 X 2 matrix of regression equations:

	$Time\_Period_i = 0$	$Time\_Period_i = 1$
$Treated_i = 0$	$y_i = \beta_0 + \epsilon_i$	$y_i = \beta_0 + \beta_1 + \epsilon_i$
$Treated_i = 1$	$y_i = \beta_0 + \beta_2 + \epsilon_i$	$y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$

The matrix of possible regression equations produced by the two dummy variables (Image by Author)

DID model is trained using the Ordinary Least Squares Regression technique.

For the trained (a.k.a. fitted) model, the corresponding expectations are as follows. The caps (^) above the coefficients indicate that they are the estimated (fitted) values of the corresponding coefficients. Replacing *y_i* with the expected value of *y_i* also allows us to drop the error term *ϵ_i* since in a well-behaved OLS regression model, the expected value of the error term is zero:

$$E(y_i|Time\_Period_i = 0, Treated = 0) = \hat{\beta}_0$$
$$E(y_i|Time\_Period_i = 1, Treated = 0) = \hat{\beta}_0 + \hat{\beta}_1$$
$$E(y_i|Time\_Period_i = 0, Treated = 1) = \hat{\beta}_0 + \hat{\beta}_2$$
$$E(y_i|Time\_Period_i = 1, Treated = 1) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$$

The expected values

(predictions) from the fitted regression model for each of the four scenarios yielded by the two dummy variables (Image by Author)

We wish to calculate the difference in the expected value of *y_i* between the before (pre-)and after (post-)treatment phases of the study.

For the treatment group, the difference in expectations works out as follows:

$$E(y_i|Time\_Period_i = 1, Treated = 1) - E(y_i|Time\_Period_i = 0, Treated = 1)$$
$$= (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) - (\hat{\beta}_0 + \hat{\beta}_2) = \boxed{\hat{\beta}_1 + \hat{\beta}_3}$$

The difference in estimated response within the treatment group between the after-treatment and before-treatment phases of the study (Image by Author)

Similarly, for the control group we have:

$$E(y_i|Time\_Period_i = 1, Treated = 0) - E(y_i|Time\_Period_i = 0, Treated = 0)$$
$$= (\hat{\beta}_0 + \hat{\beta}_1) - (\hat{\beta}_0) = \boxed{\hat{\beta}_1}$$

The difference in estimated response within the control group between the after-treatment and before-treatment phases of the study

(Image by Author)

The difference between the two differences gives us the **net effect of the treatment on the treatment group**:

$$E(DID\ Effect) = \left(\hat{\beta}_1 + \hat{\beta}_3\right) - \left(\hat{\beta}_1\right) = \boxed{\hat{\beta}_3}$$

The expected value of the Difference-In-Difference effect between the treatment and control group (Image by Author)

We see that this Difference-in-differences effect is the coefficient of the interaction term (Time_Period_i*Treatment_Group_i).

It is this result that gives the DID model much of its usefulness.

After the DID model is trained, the fitted coefficient of the *interaction term (Time_Period_i*Treatment_Group_i)* will give us the the estimated difference-in-differences effect that we are seeking. The coefficient's t-score and corresponding p value will tell us whether the effect is significant and if so, we can construct the 95% or 99% confidence interval (https://timeseriesreasoning.com/contents/interval-estimation/) around the estimated coefficient using the coefficient's standard error reported by the model.

Let's illustrate the procedure for building and training a Difference-In-Differences regression model using an interesting real world example.

How to build a Difference-In-Differences model to estimate the effect of coastal weather events on house prices

We'll use the DID model to estimate the effect of coastal weather events on house prices in the United States. Specifically, we'll analyze the effect of the the 2005 Atlantic hurricane season (https://en.wikipedia.org/wiki/2005_Atlantic_hurricane_season) which was *the* most active hurricane season in recorded history up until 2020.

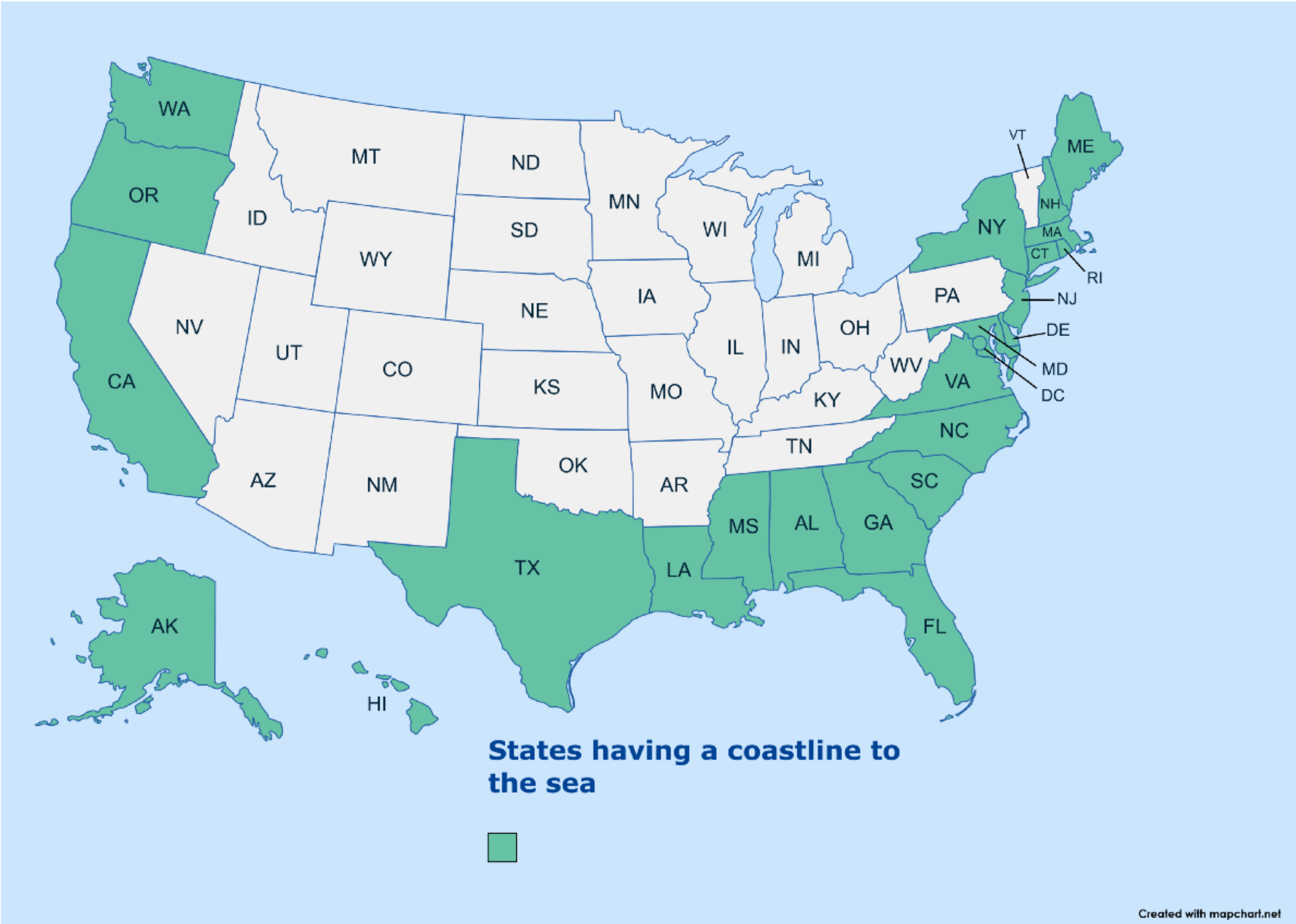
Incidentally, this topic has been extensively researched using a variety of methods. Some researchers have focused on the effect of a single storm or many storms on the house prices in a single city (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3074762) or a singe state (https://www.nber.org/papers/w27542) while others have zoomed out their attention to a regional or national level (https://link.springer.com/article/10.1057/s11369-021-00212-9). There are hyper-local studies of the effect of severe weather events on the house prices in a single US county (https://onlinelibrary.wiley.com/doi/full/10.1111/jfr3.12626), while others have studied the effect of several years worth of severe weather events (https://mpra.ub.uni-muenchen.de/19360/) on the house prices of several coastal cities. There is also an interesting recent study on estimating the impact of distant but approaching (https://link.springer.com/article/10.1007/s11146-021-09843-3) hurricane on property prices.

Several of these studies have used the Difference-In-Differences regression model (or some variation or enhancement thereof). Interestingly, although perhaps unsurprisingly, the findings from these studies are diverse and contradictory depending on the methodology used by the researchers and extent of the spatial and temporal scope of the study.

Our approach to the problem

In the rest of this chapter, we will build a rather simple Difference-In-Differences regression model to study the effect of the 2005 hurricane season on the change in the House Price Index a.k.a. house price inflation in the coastal states that were heavily impacted by the hurricane season versus the ones that weren't. Our model will be a simple one compared to the ones employed in the previous work in this area. Nevertheless, as we will soon see, we will arrive at the same sorts of results as obtained in the research literature in this area.

In our little experiment, the '**treatment**' will mean being subjected to the full brunt of 2005 hurricane season. The '**unit**' being subjected to (or not subjected to) the treatment is a US state having a coastline to the sea. There are 24 of such states in the United States:



States with a coastline to the sea (Source: MapChart (https://www.mapchart.net/feedback.html) under CC BY-SA 4.0 (https://www.mapchart.net/feedback.html))

Defining the criteria for being included in the Treatment group

We'll decide whether a state falls in the treatment group by examining the actions taken by the US Federal Emergency Management Agency (https://www.fema.gov/) (FEMA) in that state during the 2005 Atlantic hurricane season.

FEMA provides direct assistance to individuals in counties that have suffered wide-spread damage due to disasters. This type of assistance is called Individual Assistance (https://www.fema.gov/assistance/individual) and differs from the other type of assistance that FEMA offers called Community Assistance (https://www.fema.gov/floodplain-management/community-assistance-program). We will count the number of counties in each coastal state which qualified for receiving individual assistance from FEMA at anytime during the 2005 Atlantic hurricane season. Here are those those state-wise counts:

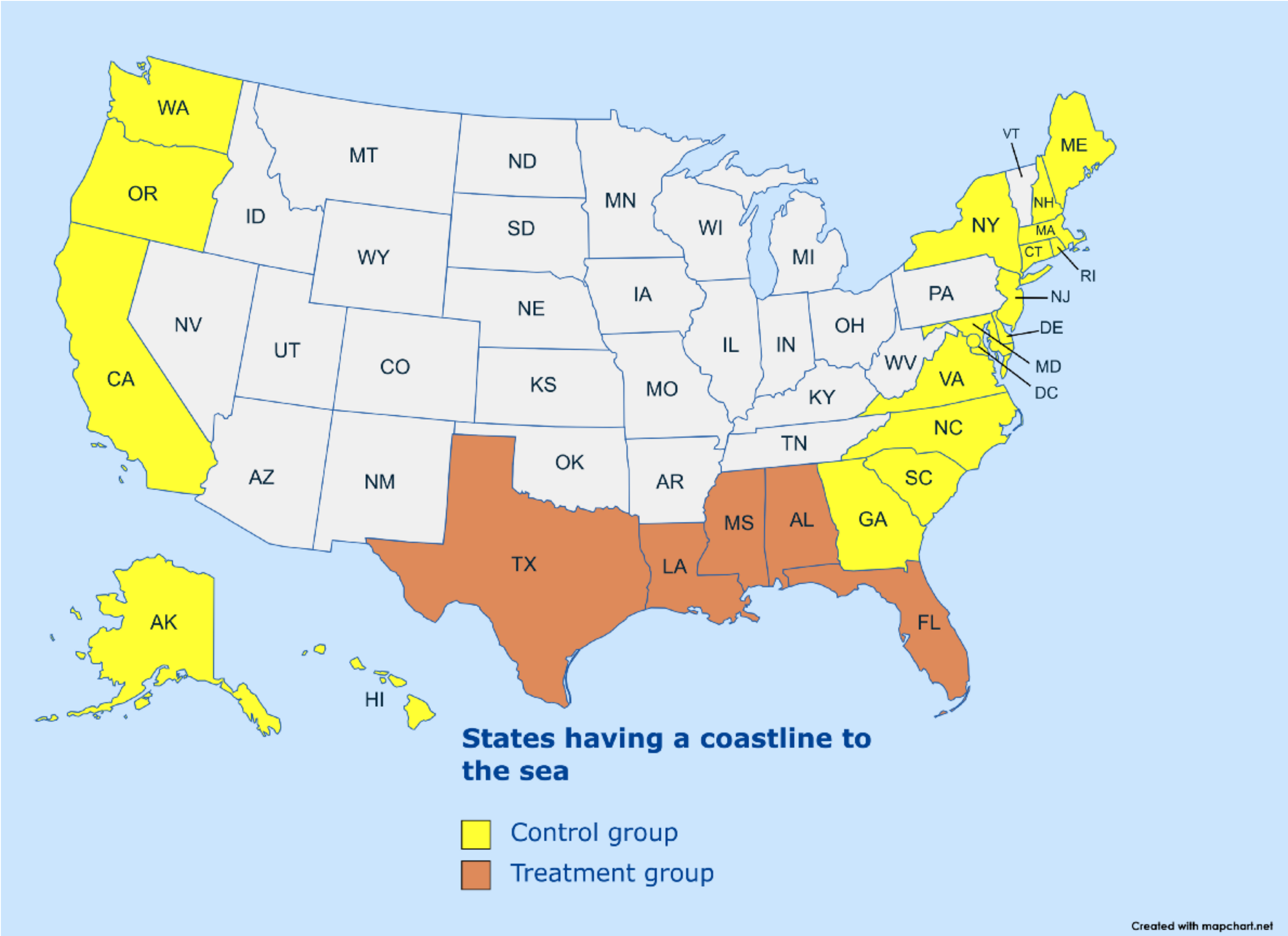
State	Number of counties receiving IA
Georgia	0
North Carolina	0
Texas	22
Massachusetts	9
Alabama	14
Mississippi	49
South Carolina	0
New Hampshire	6
Louisiana	55
Connecticut	0
Maine	0
Rhode Island	0
New York	11
California	8
Alaska	0
New Jersey	9
Delaware	0
Florida	23
Washington	0
Oregon	0
Virginia	0
Maryland	0
District of Columbia	0
Hawaii	0

State-wise counts of counties qualifying for IA during the 2005 Atlantic hurricane season. Data source: List of disasters declared by FEMA in 2005 (https://www.fema.gov/disaster/declarations?field_dv2_state_territory_tribal_value=All&field_year_value%5B%5D=2005&field_dv2_declaration_type_value=All&field_dv2_incident_type_target_id_sele (Image by Author)

If a county qualified for IA more than once, we will count it multiple times. The rationale behind the double counting is that during each disaster, some of the damaged property may have been different than the property damaged during the previous disaster. Similarly, some of the rebuilt or repaired property may also have gotten damaged again in a subsequent incident. Both cases can impact the resale value of the property. Additionally, multiple disaster events in the same county may, in theory and at least temporarily, make properties in that county less attractive to potential home buyers thereby depressing the prices or reducing the growth in prices. On the other hand, a reduction in transaction-worthy housing inventory in the county may (temporarily) increase house price inflation. Our regression model should help us determine which of these effects are dominant.

The table shown above contains a wide variability in counts and we are faced with the question of how to determine if a state was affected 'enough' to be considered a Treatment state. Should we consider New Hampshire with 9 affected counties as a Treatment state? What about California with 8 affected counties, or New York state with 11 affected counties? At the other end of the counts scale are the gulf states of Louisiana, Alabama and Mississippi which were by all accounts greatly affected and are clearly 'Treatment' group states.

We'll try to resolve this question by drawing the line at the **median of counts**. Any state with a count greater or equal to the median (14) will fall into the treatment group. The rest will be part of the control group. Here is the how the group-wise map looks like:



Treatment and control groups amongst the sea-facing states (Source: MapChart (https://www.mapchart.net/feedback.html) under CC BY-SA 4.0 (https://www.mapchart.net/feedback.html))

As we can see from the map, we would be dealing with a **highly unbalanced** data set with the treatment group being far smaller than the control. This will almost certainly influence in the quality of the estimates produced by our DID model.

Setting up the Treatment column

Using the treatment group selection criteria outlined above, we'll add a column called ***Disaster_Affected*** and set its value to 1 for states with a count ≥14, and to 0 for the rest:

State	Number of counties receiving IA	Disaster_Afected
Georgia	0	0
North Carolina	0	0
Texas	22	1
Massachusetts	9	0
Alabama	14	1
Mississippi	49	1
South Carolina	0	0
New Hampshire	6	0
Louisiana	55	1
Connecticut	0	0
Maine	0	0
Rhode Island	0	0
New York	11	0
California	8	0
Alaska	0	0
New Jersey	9	0
Delaware	0	0
Florida	23	1
Washington	0	0
Oregon	0	0
Virginia	0	0
Maryland	0	0
District of Columbia	0	0
Hawaii	0	0

(Image by Author)

Setting up the Time Period column

Next, we will add a ***Time_Period*** column which we will set to 0 to indicate the period before the start of the 2005 hurricane season, and to 1 to indicate the period after the end of the hurricane season. Notice below that we have duplicated the rows so that each state has a row with *Time_Period=0* and a row with *Time_Period=1*.

State	Number of counties receiving IA	Disaster_Afected	Time_Period
Georgia	0	0	0
North Carolina	0	0	0
Texas	22	1	0
Massachusetts	9	0	0
Alabama	14	1	0
Mississippi	49	1	0
South Carolina	0	0	0
New Hampshire	6	0	0
Louisiana	55	1	0
Connecticut	0	0	0
Maine	0	0	0
Rhode Island	0	0	0
New York	11	0	0
California	8	0	0
Alaska	0	0	0
New Jersey	9	0	0
Delaware	0	0	0
Florida	23	1	0
Washington	0	0	0
Oregon	0	0	0
Virginia	0	0	0
Maryland	0	0	0
District of Columbia	0	0	0
Hawaii	0	0	0
Georgia	0	0	1
North Carolina	0	0	1
Texas	22	1	1
Massachusetts	9	0	1
Alabama	14	1	1
Mississippi	49	1	1
South Carolina	0	0	1
New Hampshire	6	0	1
Louisiana	55	1	1
Connecticut	0	0	1
Maine	0	0	1
Rhode Island	0	0	1

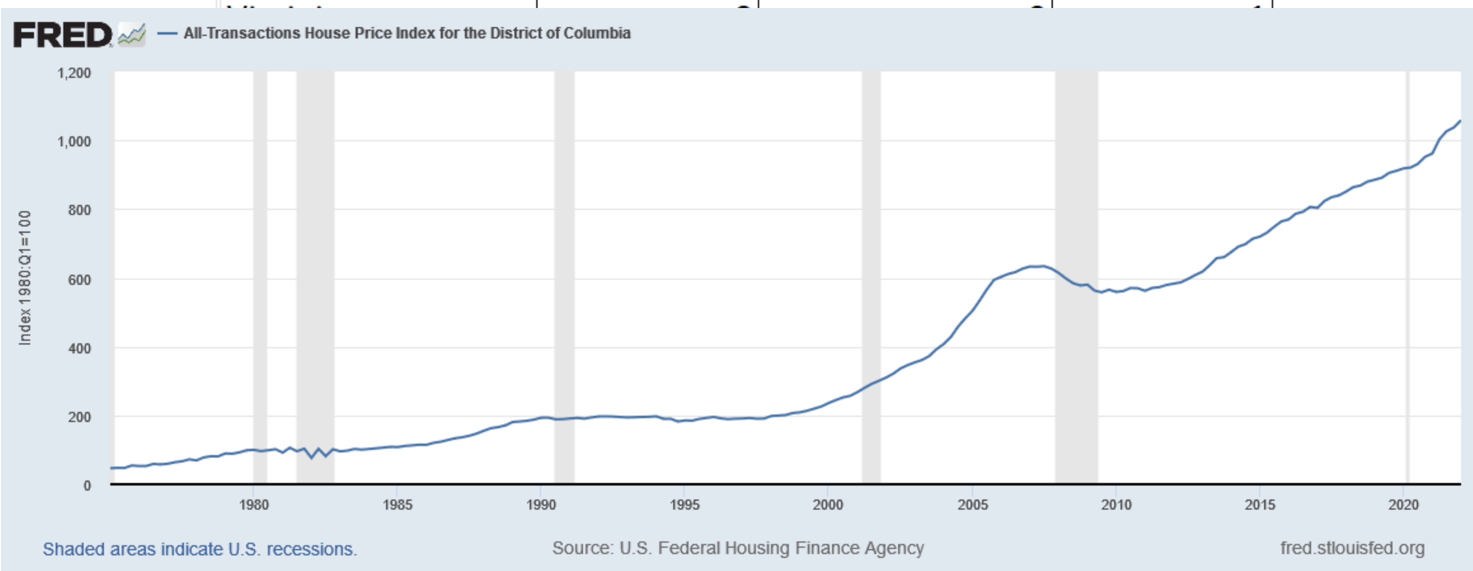
New York	11	0	1
California	8	0	1
Alaska	0	0	1
New Jersey	9	0	1
Delaware	0	0	1
Florida	23	1	1
Washington	0	0	1
Oregon	0	0	1

(Image by Author)

The methodology for calculating the value of the response variable

This section described the procedure for calculating the values of the response variable *y_i*.

Our goal is to study the effect of the 2005 hurricane season on house prices in the coastal states. To that end, we'll use the **state-wise *All Transactions House Price Index* published by the US Federal Reserve**, and available for download under the public domain license (https://fred.stlouisfed.org/categories/27290?t=public%20domain%3A%20citation%20requested%3Bquarterly&ob=pv&od=desc) from US FRED (https://fred.stlouisfed.org/). Here's how the index looks like for the District of Columbia:



U.S. Federal Housing Finance Agency, All-Transactions House Price Index for the District of Columbia (https://fred.stlouisfed.org/series/DCSTHPI) [DCSTHPI], retrieved from FRED, Federal Reserve Bank of St. Louis;, June 12, 2022 (public domain (https://fred.stlouisfed.org/categories/27290?t=public%20domain%3A%20citation%20requested%3Bquarterly&ob=pv&od=desc))

We will access 24 of these time series data sets for the 24 states of interest and we'll knock them together into a 24-state data panel as follows:

	DATE	GASTHPI	NCSTHPI	TXSTHPI	MASTHPI	ALSTHPI	MSSTHPI	SCSTHPI	NHSTHPI	...	NJSTHPI	DESTHPI	FLSTHPI	WASTHPI	ORSTHPI	VASTHPI	MDSTHPI	DCSTHPI	HISTHPI
0	01-01-75	73.88	65.99	55.58	67.89	75.03	75.21	75.39	77.94	...	64.02	77.19	65.80	46.22	51.12	69.87	61.86	46.55	66.77
1	01-04-75	71.95	67.07	58.53	66.04	72.23	72.14	69.82	55.42	...	59.25	89.82	83.42	47.19	51.49	66.70	62.31	47.60	53.85
2	01-07-75	74.27	68.83	56.11	67.29	74.54	72.56	72.26	59.46	...	60.09	115.94	66.84	49.70	54.17	67.38	65.06	47.33	57.84
3	01-10-75	73.65	70.49	59.36	70.26	71.73	73.11	69.98	52.83	...	62.81	71.78	68.26	48.02	53.23	67.77	66.91	54.55	55.94
4	01-01-76	71.29	68.55	58.66	67.41	77.11	73.06	71.04	81.90	...	62.63	79.68	67.98	50.33	57.13	69.39	67.23	53.03	53.82
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
184	01-01-21	424.35	441.51	374.29	891.48	366.78	298.41	445.77	563.07	...	586.06	521.41	531.04	741.63	648.23	520.50	523.69	961.53	664.94
185	01-04-21	448.85	467.45	396.31	939.29	384.40	309.70	470.31	596.88	...	614.01	542.94	565.29	798.23	691.69	544.82	548.31	1003.22	699.55
186	01-07-21	475.90	498.68	421.49	988.02	405.76	326.08	496.85	632.46	...	644.11	570.16	607.29	844.40	733.89	569.69	570.22	1026.32	727.43
187	01-10-21	499.85	521.18	439.06	1010.73	421.80	336.11	519.90	651.12	...	664.58	589.42	641.32	872.01	753.41	583.51	582.80	1037.33	762.28
188	01-01-22	523.17	546.14	458.99	1035.47	433.73	345.65	544.18	670.56	...	685.23	604.87	676.58	910.83	778.00	601.04	599.59	1058.77	799.98

[189 rows x 25 columns]

The House Price Index data for all seacoast states from Q1 1975 to Q1 2022 (Image by Author)

For our study, the time periods of interest to us are the 4 quarters immediately prior to the 2005 hurricane season and the 4 quarters immediately following the season. The hurricane season itself ran from 8 June 2005 to 6 Jan 2006. Hence, we are interested in house price index change across the quarters starting from 1 July 2004, 1 October 2004, 1 January 2005 and 1 April 2005, and then again across the 4 quarters following the 2005 season namely, 1 April 2006, 1 July 2006, 1 October 2006 and 1 January 2007. Let's zoom into this region of interest to see how it looks like:

DATE	GASTHPI	NCSTHPI	TXSTHPI	MASTHPI	ALSTHPI	MSSTHPI	SCSTHPI	NHSTHPI	LASTHPI	CTSTHPI	MESTH
01-04-04	283.99	274.38	189.31	632.4	244.79	211.39	277.61	399.62	195.01	383.82	418
01-07-04	287.88	277.28	190.84	662.87	249.32	214.21	282.88	417.3	198.3	403.32	438
01-10-04	292.44	282.01	193.04	673.94	253	217.16	288.17	425.46	201	411.06	447
01-01-05	296.26	286.49	194.03	688.44	256.38	218.47	293.12	436.77	203.23	421.53	460
01-04-05	300.24	290.32	197.14	704.79	262.46	222.81	298.13	447.11	207.44	435.69	470
01-07-05	305.19	296.91	200.31	715.82	267.91	227.34	305.69	457.87	211.4	447.71	482
01-10-05	309.19	303.18	202.68	720.96	273.42	232.51	312.46	463.14	218.55	456.74	490
01-01-06	312.39	308.18	205.07	721.53	278.92	236.32	316.83	466.7	225.14	462.73	494
01-04-06	314.26	312.39	208.29	712.73	283.51	243.22	322.1	464.09	231.03	464.98	492
01-07-06	316.81	317.67	211.39	707.41	288.07	247.96	326.82	462.13	235.87	465.44	495
01-10-06	322.74	323.44	214.29	709.36	293.6	252.45	335.01	465.22	239.69	468.13	502
01-01-07	325.34	328.62	217.08	703.93	295.52	256.26	336.98	463.88	242.26	471.21	506
01-04-07	326.44	331.71	220.72	693.92	299.64	256.33	339.46	460.75	244.71	467.58	502

4 quarters before the hurricane season

Quarters overlapping the hurricane season

4 quarters after the hurricane season

The four quarters of interest immediately preceding and immediately following the 2005 hurricane season. (Image by Author)

For each state, we will calculate the average quarter-over-quarter fractional change in the house price index over the two sets of quarters. Doing so will give us the value of the response variable, namely, the average Q-o-Q change in HPI in the pre-treatment and the post-treatment phases of the study for each state.

The Q-o-Q fractional change in house price index across any two consecutive quarters *i* and *(i-1)* can be calculated using the following formula:

HPI Fractional Change = [HPI_i — HPI_(i-1)]/HPI_(i-1)

Here are the Q-o-Q fractional change values for the 4 quarters of interest before and after the 2005 hurricane season. The highlighted cells illustrate the calculation for one of the quarters:

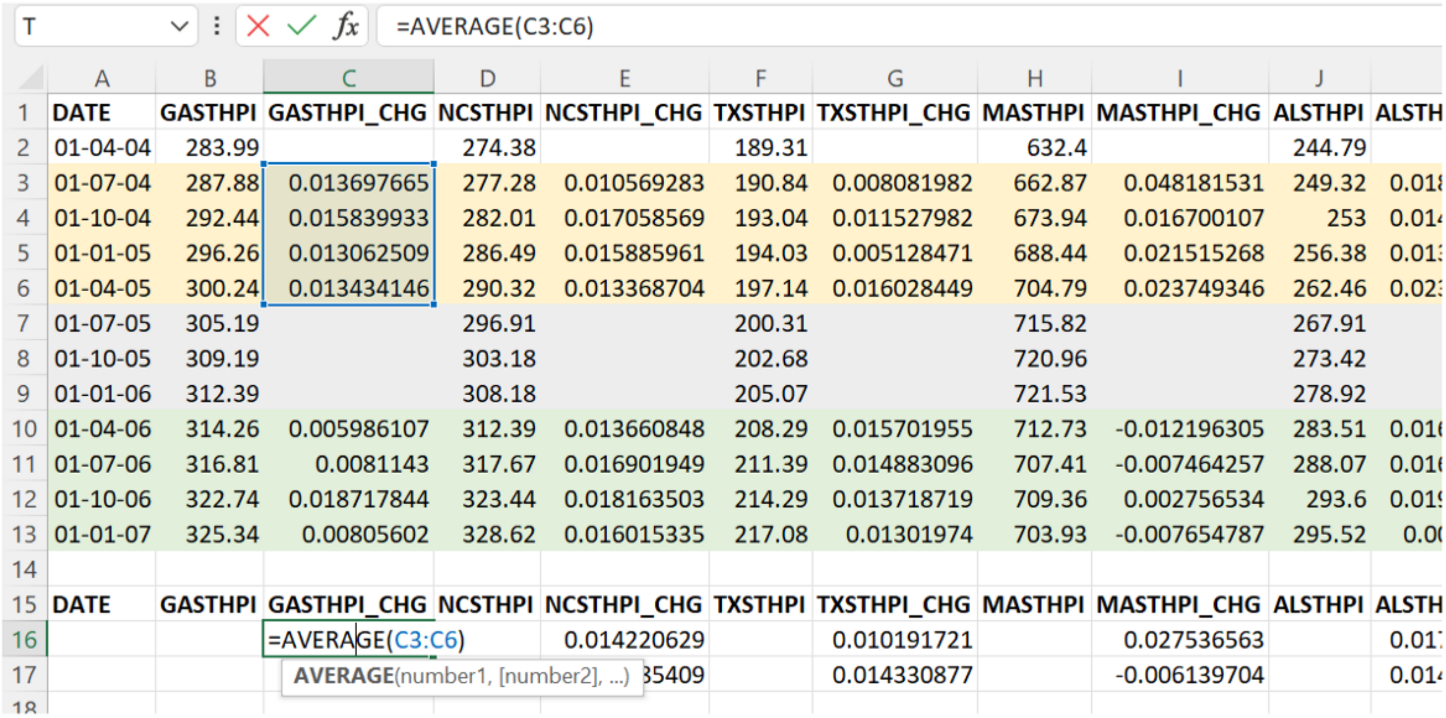
T =(B3-B2)/B2

	A	B	C	D	E	F	G	H
1	DATE	GASTHPI	GASTHPI_CHG	NCSTHPI	NCSTHPI_CHG	TXSTHPI	TXSTHPI_CHG	MAS
2	01-04-04	283.99		274.38		189.31		6
3	01-07-04	287.88	=(B3-B2)/B2	277.28	0.010569283	190.84	0.008081982	66
4	01-10-04	292.44	0.015839933	282.01	0.017058569	193.04	0.011527982	67
5	01-01-05	296.26	0.013062509	286.49	0.015885961	194.03	0.005128471	68
6	01-04-05	300.24	0.013434146	290.32	0.013368704	197.14	0.016028449	70
7	01-07-05	305.19		296.91		200.31		71
8	01-10-05	309.19		303.18		202.68		72
9	01-01-06	312.39		308.18		205.07		72
10	01-04-06	314.26	0.005986107	312.39	0.013660848	208.29	0.015701955	71
11	01-07-06	316.81	0.0081143	317.67	0.016901949	211.39	0.014883096	70
12	01-10-06	322.74	0.018717844	323.44	0.018163503	214.29	0.013718719	70
13	01-01-07	325.34	0.00805602	328.62	0.016015335	217.08	0.01301974	70

Calculation of the Q-o-Q fractional change in HPI for the quarters of interest

(Image by Author)

Next, we take the vertical average of each block of 4 quarters to arrive at the average fractional change in HPI across 4 quarters both before and after the 2005 hurricane season. We repeat this calculation for each state to get the value of the response variable HPI_CHG for the pre-treatment and post-treatment phases.



T =AVERAGE(C3:C6)

	A	B	C	D	E	F	G	H	I	J	K
1	DATE	GASTHPI	GASTHPI_CHG	NCSTHPI	NCSTHPI_CHG	TXSTHPI	TXSTHPI_CHG	MASTHPI	MASTHPI_CHG	ALSTHPI	ALSTH
2	01-04-04	283.99		274.38		189.31		632.4		244.79	
3	01-07-04	287.88	0.013697665	277.28	0.010569283	190.84	0.008081982	662.87	0.048181531	249.32	0.01
4	01-10-04	292.44	0.015839933	282.01	0.017058569	193.04	0.011527982	673.94	0.016700107	253	0.01
5	01-01-05	296.26	0.013062509	286.49	0.015885961	194.03	0.005128471	688.44	0.021515268	256.38	0.01
6	01-04-05	300.24	0.013434146	290.32	0.013368704	197.14	0.016028449	704.79	0.023749346	262.46	0.02
7	01-07-05	305.19		296.91		200.31		715.82		267.91	
8	01-10-05	309.19		303.18		202.68		720.96		273.42	
9	01-01-06	312.39		308.18		205.07		721.53		278.92	
10	01-04-06	314.26	0.005986107	312.39	0.013660848	208.29	0.015701955	712.73	-0.012196305	283.51	0.01
11	01-07-06	316.81	0.0081143	317.67	0.016901949	211.39	0.014883096	707.41	-0.007464257	288.07	0.01
12	01-10-06	322.74	0.018717844	323.44	0.018163503	214.29	0.013718719	709.36	0.002756534	293.6	0.01
13	01-01-07	325.34	0.00805602	328.62	0.016015335	217.08	0.01301974	703.93	-0.007654787	295.52	0.0
14											
15	DATE	GASTHPI	GASTHPI_CHG	NCSTHPI	NCSTHPI_CHG	TXSTHPI	TXSTHPI_CHG	MASTHPI	MASTHPI_CHG	ALSTHPI	ALSTH
16			=AVERAGE(C3:C6)		0.014220629		0.010191721		0.027536563		0.01
17			AVERAGE(number1, [number2], ...)		35409		0.014330877		-0.006139704		0.01

Calculation of the average Q-o-Q fractional change in HPI across 4 quarters preceding and following the hurricane season (Image by Author)

Note that for each state, we have calculated two response values: the top value is the pre-treatment value and the bottom one is the post-treatment value. Thus, there is one value corresponding to *Time_Period=0* and another one corresponding to *Time_Period=1*. Let's include these average values in the data set we will use to train the DID model:

State	Number of counties receiving IA	Disaster_Afected	Time_Period	HPI_CPG
Georgia	0	0	0	0.0140086
North Carolina	0	0	0	0.0142206
Texas	22	1	0	0.0101917
Massachusetts	9	0	0	0.0275366
Alabama	14	1	0	0.0175851
Mississippi	49	1	0	0.0132524
South Carolina	0	0	0	0.0179883
New Hampshire	6	0	0	0.0285133
Louisiana	55	1	0	0.0155742
Connecticut	0	0	0	0.0322646
Maine	0	0	0	0.0300314
Rhode Island	0	0	0	0.0393889
New York	11	0	0	0.0334331
California	8	0	0	0.0606154
Alaska	0	0	0	0.0311449
New Jersey	9	0	0	0.041487
Delaware	0	0	0	0.0402581
Florida	23	1	0	0.0591272
Washington	0	0	0	0.0375659
Oregon	0	0	0	0.037789
Virginia	0	0	0	0.0487666
Maryland	0	0	0	0.0532513
District of Columbia	0	0	0	0.0568036
Hawaii	0	0	0	0.059643
Georgia	0	0	1	0.0102186
North Carolina	0	0	1	0.0161854
Texas	22	1	1	0.0143309
Massachusetts	9	0	1	-0.00614
Alabama	14	1	1	0.0145692
Mississippi	49	1	1	0.0204715
South Carolina	0	0	1	0.0155569
New Hampshire	6	0	1	-0.001502
Louisiana	55	1	1	0.0185072
Connecticut	0	0	1	0.0045526
Maine	0	0	1	0.0056873
Rhode Island	0	0	1	-0.00087

New York	11	0	1	0.0049155
California	8	0	1	-0.00174
Alaska	0	0	1	0.0157814
New Jersey	9	0	1	0.0060337
Delaware	0	0	1	0.0119038
Florida	23	1	1	0.007313
Washington	0	0	1	0.0259023
Oregon	0	0	1	0.0242574
Virginia	0	0	1	0.0108559
Maryland	0	0	1	0.0125506
District of Columbia	0	0	1	0.0120927
Hawaii	0	0	1	0.0093707

The data set to be used for training the Difference-In-Differences model (Image by Author)

The last column of the above data set set HPI_CPG is our response variable *y_i*.

The data set is available for download **from here** (https://gist.github.com/sachinsdate/1fc451683137398e11c75b2e47031cf1).

Now that our data set is built, we can get back to the task of building and training the DID model.

Building the Difference-In-Differences model for house price inflation

Let's start by stating the equation for our DID model:

$$\boxed{\begin{aligned} HPI\_CHG_i = \beta_0 + \beta_1 * Time\_Period_i + \beta_2 * Disaster\_Affected_i + \\ \beta_3 * (Time\_Period_i * Disaster\_Affected_i) + \epsilon_i \end{aligned}}$$

The equation of the DID model used for estimating the effect of hurricane disasters on house price changes (Image by Author)

To build and train the model, we'll use Python and Python based libraries Pandas (https://pandas.pydata.org/getting_started.html) and statsmodels (https://www.statsmodels.org/stable/gettingstarted.html).

Let's begin by importing all the required packages:

```
import pandas as pd
from patsy import dmatrices
import statsmodels.api as sm
```

Next, we'll load the data set into a Pandas DataFrame as follows:

```
df = pd.read_csv('us_fred_coastal_us_states_avg_hpi_before_after_2005.csv', header=0)
```

Form the regression expression in Patsy (https://patsy.readthedocs.io/en/latest/quickstart.html) syntax. The intercept is assumed to be present and will be included in the data set automatically:

```
reg_exp = 'HPI_CHG ~ Time_Period + Disaster_Affected + Time_Period*Disaster_Affected'
```

Using Patsy, carve out the training matrices:

```
y_train, X_train = dmatrices(reg_exp, df, return_type='dataframe')
```

Build the DID model:

```
did_model = sm.OLS(endog=y_train, exog=X_train)
```

Train the model:

```
did_model_results = did_model.fit()
```

Print the training summary:

```
did_model_results.summary()
```

We see the following output (I have highlighted the interesting parts):

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                HPI_CHG   R-squared:                       0.536
Model:                            OLS   Adj. R-squared:                  0.504
Method:                 Least Squares   F-statistic:                     16.92
Date:                Mon, 13 Jun 2022   Prob (F-statistic):           1.88e-07
Time:                        16:59:46   Log-Likelihood:                 145.14
No. Observations:                  48   AIC:                            -282.3
Df Residuals:                      44   BIC:                            -274.8
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.0371      0.003     13.157      0.000       0.031       0.043
Time_Period                     -0.0278      0.004     -6.985      0.000      -0.036      -0.020
Disaster_Affected               -0.0139      0.006     -2.258      0.029      -0.026      -0.001
Time_Period:Disaster_Affected    0.0197      0.009      2.260      0.029       0.002       0.037
==============================================================================
Omnibus:                        5.463   Durbin-Watson:                   1.165
Prob(Omnibus):                  0.065   Jarque-Bera (JB):                4.279
Skew:                           0.623   Prob(JB):                        0.118
Kurtosis:                       3.767   Cond. No.                         6.78
==============================================================================
```

Training output of the Difference-In-Differences regression model (Image by Author)

How to interpret the training output of the DID model

We see that the adjusted R-squared is 0.504. The model has been able to explain more than 50% of the variance in the response variable HPI_CHG. That is a great result. The p value of the F-statistic is 1.88e-07 which is statistically speaking, highly significant, leading us to conclude that the model's variables are jointly significant and they are together doing a much better job of explain the variance in HPI_CHG than a simple mean model.

We also note is that all coefficients are statistically significant as indicated by their p values which are all smaller than 0.05.

The equation of the fitted model is as follows:

$$HPI\_CHG_i = 0.0371 - 0.0278 * Time\_Period_i - 0.0139 * Disaster\_Affected_i + 0.0197 * (Time\_Period_i * Disaster\_Affected_i) + e_i$$

The equation of the fitted Difference-In-Differences model (Image by Author)

Time_Period and Disaster_Affected are 0/1 dummy variables. The four possible combinations are:

Let's see how to interpret each combination of the two dummy variables: *Time_Period* and *Disaster_Affected*. We'll also switch to working with expected values of *HPI_CHG*, which results in dropping of the subscript *i* as also the residual error term *e_i*.

Time_Period_i=0 and Disaster_Affected_i=0

We get the following equation:

$$E(HPI\_CHG) = 0.0371$$

Expected Q-o-Q change in house price index in the control group states during the pre-hurricane period (Image by Author)

This equation gives us the estimated mean inflation in house prices in the **control group** during the four quarters immediately preceding the 2005 hurricane season. The value of the estimated mean inflation is simply the intercept of regression: 0.0371, or 3.71%.

Time_Period_i=1 and Disaster_Affected_i=0

$$E(HPI\_CHG) = 0.0371 - 0.0278$$

Expected Q-o-Q change in house price index in the control group states during the post-hurricane period (Image by Author)

This equation give us the estimated mean inflation in house prices in the **control group** states in the post-treatment period, i.e. during the four quarters following the hurricane season. The value of the estimated mean inflation is 0.0371—0.0278=0.0093, or 0.93%.

Time_Period_i=0 and Disaster_Affected_i=1

$$E(HPI\_CHG) = 0.0371 - 0.0139$$

Expected Q-o-Q change in house price index in the treatment group states during the pre-hurricane period (Image by Author)

This equation gives us the estimated mean house price inflation in the **treatment group** states during the four quarters prior to the start of the hurricane season. The value of this inflation is 0.0371—0.0139=0.0232, or 2.32%.

Time_Period_i=1 and Disaster_Affected_i=1

$$E(HPI\_CHG) = 0.0371 - 0.0278 - 0.0139 + 0.0197$$

Expected Q-o-Q change in house price index in the treatment group states during the post-hurricane period (Image by Author)

This equation gives us the estimated mean house price inflation in the treatment group during the four quarters following the end of the hurricane season. The value of this inflation is 0.0371—0.0278—0.0139+0.0197=0.0151 or 1.51%.

Let’s tabulate our findings:

	Treatment Group	Control Group
Time_Period	E(HPI_CHG|DisasterAffected=1)	E(HPI_CHG|DisasterAffected=0)
0	2.32%	3.71%
1	1.51%	0.93%
δE(HPI_CPG)	**-0.81%**	**-2.78%**

Estimated change in House Price Index in the Treatment and Control groups before and after the Treatment (Image by Author)

The third row of the table mentions the vertical differences (post-season—pre-season) in the estimated values.

We see that for those in the Disaster Affected group, the inflation in house prices in the four quarters following the hurricane season were lower by 0.81% as compared to the house price inflation experienced in the four quarters prior to the start of the hurricane season.

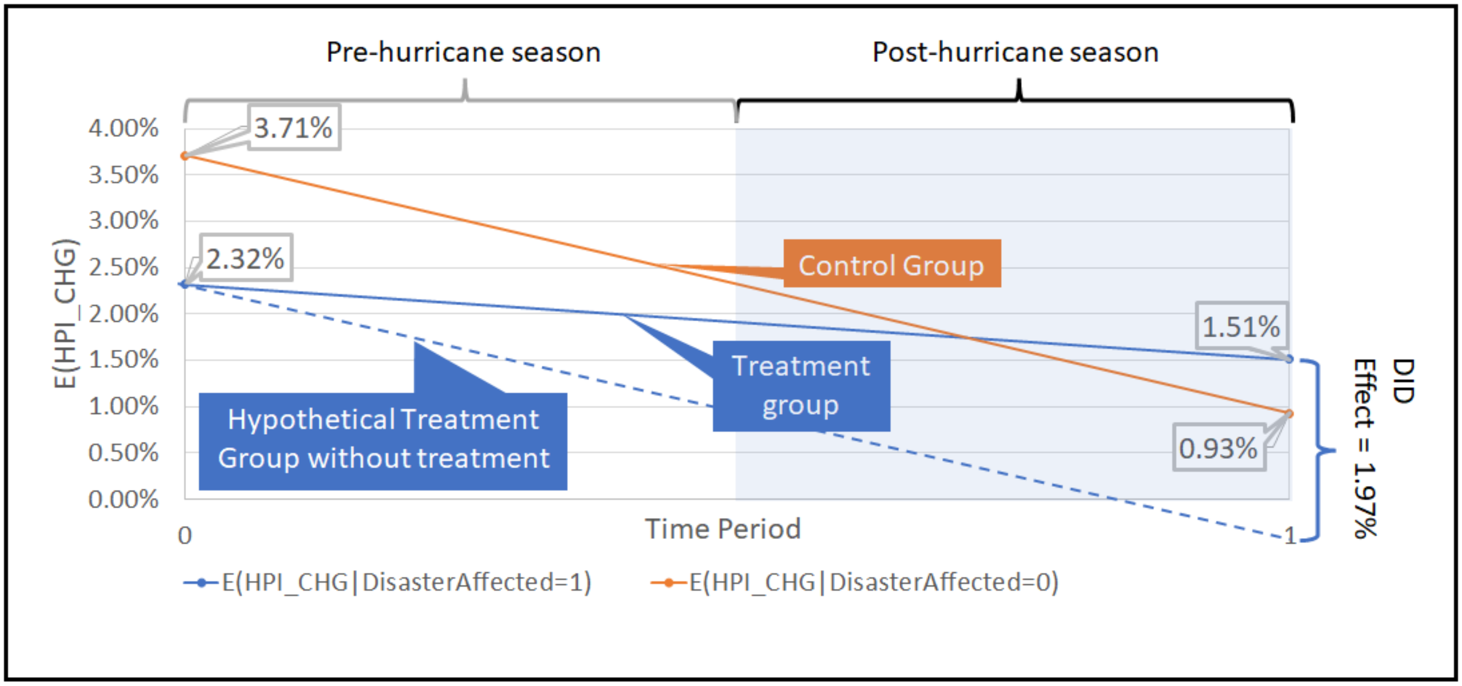
For those in the non Disaster Affected group, the inflation in house prices in the four quarters following the hurricane season were lower by 2.78% as compared to the house price inflation experienced in the four quarters prior to the start of the hurricane season.

The difference-in-difference effect between the two groups is:

$$\delta E(HPI\_CHG|Disaster\_Affected = 1) - \delta E(HPI\_CHG|Disaster\_Affected = 0) = (-0.81\%) - (-2.78\%) = \boxed{1.97\%}$$

The estimated Difference-In-Differences effect (Image by Author)

The following graphic may help in visualizing the various estimated values:



Estimated change in House Price Index in the Treatment and Control groups before and after the Hurricane season a.k.a. Treatment (Image by Author)

The value of 1.97% is exactly the value of the coefficient of *Time_Period*Disaster_Affected* interaction term reported by the trained DID regression model:

$$HPI\_CHG_i = 0.0371 - 0.0278 * Time\_Period_i - 0.0139 * Disaster\_Affected_i + \boxed{0.0197} * (Time\_Period_i * Disaster\_Affected_i) + e_i$$

The fitted DID model (Image by Author)

The estimated difference-in-differences of 1.97% suggests that the house price inflation in the states that were especially affected by the 2005 hurricane season cooled down less than in the rest of the coastal states after the season ended. One way to explain this effect is by noting that inflation is often inversely proportional to supply. Due to extensive property damage suffered by the treatment group states, the resulting reduction in house inventory may have temporarily fed house price inflation in those states during the four quarters immediately following the end of the hurricane season.

Here's the source code used in this chapter:

```
import pandas as pd
from patsy import dmatrices
import statsmodels.api as sm
```

```


#Load the data set into a Pandas Dataframe
df = pd.read_csv('us_fred_coastal_us_states_avg_hpi_before_after_2005.csv', header=0)

#Print it
print(df)

#Form the regression expression in Patsy syntax. The intercept is assumed to be present and will be
# included in the data set automatically
reg_exp = 'HPI_CHG ~ Time_Period + Disaster_Affected + Time_Period*Disaster_Affected'

#Carve out the training matrices
y_train, X_train = dmatrices(reg_exp, df, return_type='dataframe')

#Build the DID model
did_model = sm.OLS(endog=y_train, exog=X_train)

#Train the model
did_model_results = did_model.fit()

#Print out the training results
did_model_results.summary()
```

view raw difference_in_differences_regression.py hosted with ❤ by **GitHub**

Citations and Copyrights

Data set

All-Transactions House Price Index for various US states, courtesy of U.S. Federal Housing Finance Agency, retrieved from FRED, Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/searchresults?st=All-transactions+House+Price+Index);, June 12, 2022 (available in public domain (https://fred.stlouisfed.org/categories/27290?t=public%20domain%3A%20citation%20requested%3Bquarterly&ob=pv&od=desc)). **The curated version of the data set used in this chapter is available for download from here** (https://gist.github.com/sachinsdate/1fc451683137398e11c75b2e47031cf1).

Paper and Book Links

Card, David and Krueger, Alan, (1994), Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania (https://EconPapers.repec.org/RePEc:aea:aecrev:v:84:y:1994:i:4:p:772-93), *American Economic Review*, 84, issue 4, p. 772–93.

Ortega, Francesc and Taspinar, Suleyman, Rising Sea Levels and Sinking Property Values: The Effects of Hurricane Sandy on New York's Housing Market (March 29, 2018). Available at SSRN (https://ssrn.com/abstract=3074762) or http://dx.doi.org/10.2139/ssrn.3074762 (https://dx.doi.org/10.2139/ssrn.3074762)

Liao, Yanjun and Graff Zivin, Joshua and Panassie, Yann, How Hurricanes Sweep Up Housing Markets: Evidence from Florida. Available at SSRN (https://ssrn.com/abstract=4103049) or http://dx.doi.org/10.2139/ssrn.4103049 (https://dx.doi.org/10.2139/ssrn.4103049)

Fisher, J.D., Rutledge, S.R. The impact of Hurricanes on the value of commercial real estate. *Bus Econ* 56**,** 129–145 (2021). https://doi.org/10.1057/s11369-021-00212-9 (https://doi.org/10.1057/s11369-021-00212-9)

Seung Kyum Kim, Richard B. Peiser, The implication of the increase in storm frequency andintensity to coastal housing markets, Journal of Flood Risk Management, 26 May 2020, https://doi.org/10.1111/jfr3.12626 (https://doi.org/10.1111/jfr3.12626)

Anthony Murphy & Eric Strobl, 2010. The impact of hurricanes on housing prices: evidence from U.S. coastal cities (https://ideas.repec.org/p/fip/feddwp/1009.html), Working Papers (https://ideas.repec.org/s/fip/feddwp.html) 1009, Federal Reserve Bank of Dallas.

Fang, L., Li, L. & Yavas, A. The Impact of Distant Hurricane on Local Housing Markets. *J Real Estate Finan Econ* (2021). https://doi.org/10.1007/s11146-021-09843-3 (https://doi.org/10.1007/s11146-021-09843-3)

Images

PREVIOUS: What Are Dummy Variables And How To Use Them In A Regression Model (https://timeseriesreasoning.com/contents/dummy-variables-in-a-regression-model/)

NEXT: A Guide To Building Linear Models For Discontinuous Data (https://timeseriesreasoning.com/contents/linear-regression-models-for-discontinuous-data/)

UP: Table of Contents (https://timeseriesreasoning.com/)