ECN 140 – Empirical Project Cesar Rodriguez-Munoz

What was the Effect of Leukemia Cases on Churchill County from 2000 to 2002? An Examination Using a Difference-in-Differences Model.

I. Abstract

The goal of this report is to examine the effect that the rise in leukemia cases, in Churchill County, had on real house sale values. To accomplish this, I used a difference-in-differences model. The model takes the difference between average house sale prices between Churchill and Lyon after the rise in cases and subtracts this from the difference between average house sale prices between the counties before the rise in cases. My report found that the rise in cases had a -12.5% effect on average real sale prices in Churchill, but I cannot conclude a causal effect from my model. Further research and data collection is warranted to study the true causal effect of leukemia cases on real house sale prices.

II. Introduction

Churchill, a county in Nevada, experienced a surprising amount of leukemia cases within a short period of time. Statistics from the American Cancer Society suggest that Churchill County should only expect about one case of leukemia every 5 years, based on the population of the county. In 1997, the residents of Churchill experienced their first, and only, case of leukemia that year. This first case was not a cause for concern, but by 2000, Churchill had 8 cases of leukemia, far exceeding the number of cases a county like Churchill should see. This was followed by an additional 4 cases in 2001, and a total of 16 cases in 2002.

While the Nevada Health Department and the CDC were unable to determine the cause of the sudden spike in leukemia cases, there was no indication that Churchill was uniquely inhospitable. Still, house prices dropped as residents left the county in mass, seemingly in fear of the increased number of cases.

The question naturally arises, did house prices drop in Churchill due to the increase in leukemia cases, or would Churchill have experienced this drop in house value regardless of the sudden uptick in leukemia cases? To examine this question, I am using a pooled cross-sectional dataset of 10,204 house sales, from the years of 1990 to 2002. This data set includes house sales from a nearby county, Lyon, which will prove useful as Lyon County did not experience a single case of leukemia over the same period of time, and Lyon has similar characteristics to Churchill.

To answer the question of whether leukemia cases did affect house values in Churchill, I will use a difference in differences model with this data set. If the assumptions of the model hold, we should be able to observe the causal effect of leukemia cases on house prices in Churchill. Additionally, I will give explanations of the model and examine whether we can infer a causal inference in further sections.

III. Data Summary

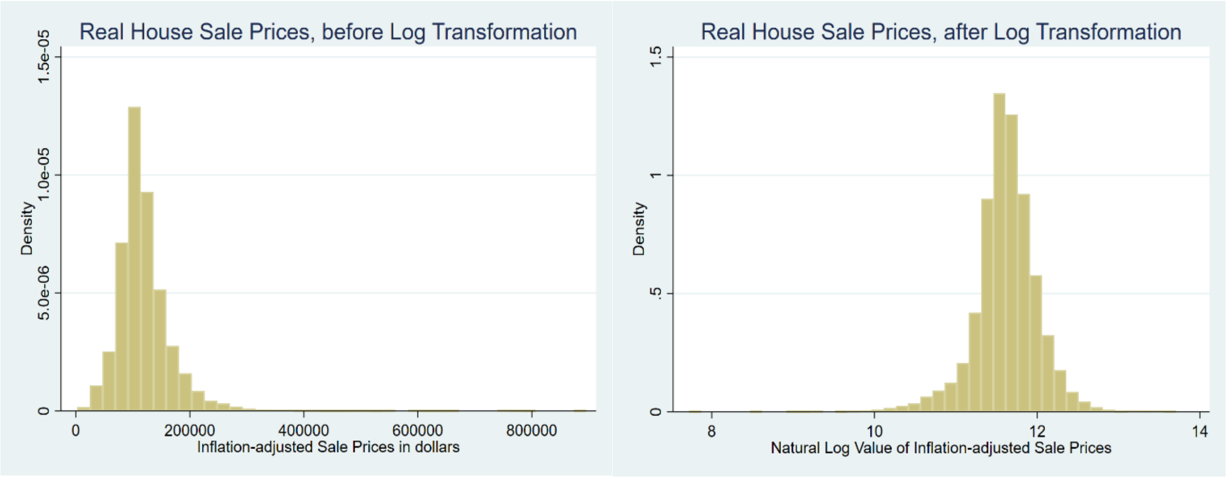
In this section, I will explain general trends in the data and give reasons why I created new variables or transformed variables.

To begin, it is important to establish the variables I used to create my model. These variables include the lot size of the house sold in acres, the interior size of the house in square feet, the age of the house sold, and the perceived condition of the home rated on a scale from zero to five. These variables are regressed on the real house sale price. With that clarification, let us dive into the transformations and generation of variables.

The first variable I adjusted was the real sale prices of houses. Intuitively, a variable that measures house prices is prone to skewness. This is because we can observe very large house sale values, with theoretically no limitation, but the smallest sale value we can observe is a house that sells for zero dollars. This inherently leads to skewness.

Skewness makes it more difficult to fit a linear regression to the data, so a transformation on this variable can help us avoid this. To check if the real house sale variable is skewed, I made a histogram of the variable, which is shown below:

Figure 1 (on the right) and Figure 2 (on the left)

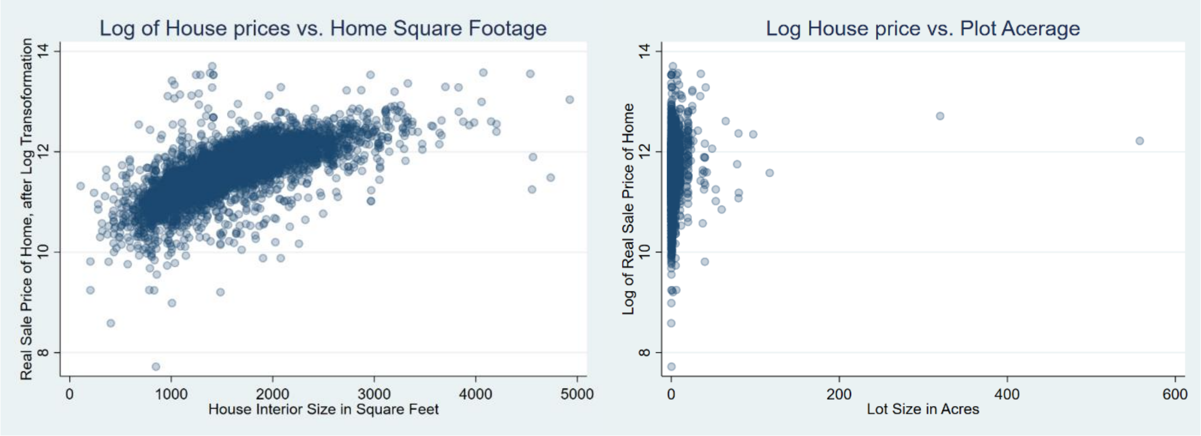


We can see that the right tail of the histogram in Figure 1 continues past $800,000, which greatly contributes to right skewness. Specifically, the skewness value of the histogram in Figure 1 is about 3.67, which value of 0 would be perfectly normal. This means that the data is heavily right-skewed, and a transformation of this variable is justifiable. Taking the natural log of this variable will help to create a more normal distribution, which is shown in Figure 2.

The skewness value of this distribution in Figure 2 is only about -0.52. This is considered slightly left-skewed, but we can still reasonably consider this distribution normal with such a small skewness value. By utilizing this transformation, my model will more easily fit this data, and we will more consistently derive accurate estimators for our parameters of interest.

Next, I examined the variables used in the model that I constructed. There is worthy concern that these variables might not move in a linear fashion with the log of real sales prices of houses. After checking each of the variables used in the model, a notable interaction involved the square footage of the home and the lot size of the home. Using scatter plots shows the non-linear relationship of these variables, shown below:

Figure 3 (shown on the right) and Figure 4 (shown on the left)



If we attempted to fit a linear line through this data, we would be missing several data points. If the factors did move linearly, a larger home or lot would always proportionally be more expensive than a smaller home. What is more likely, and what we are observing, is that a large home and a large lot tend to only be marginally more expensive, and that marginal increase in price decreases at the home and lot size increases. This is known as a diminishing marginal return.

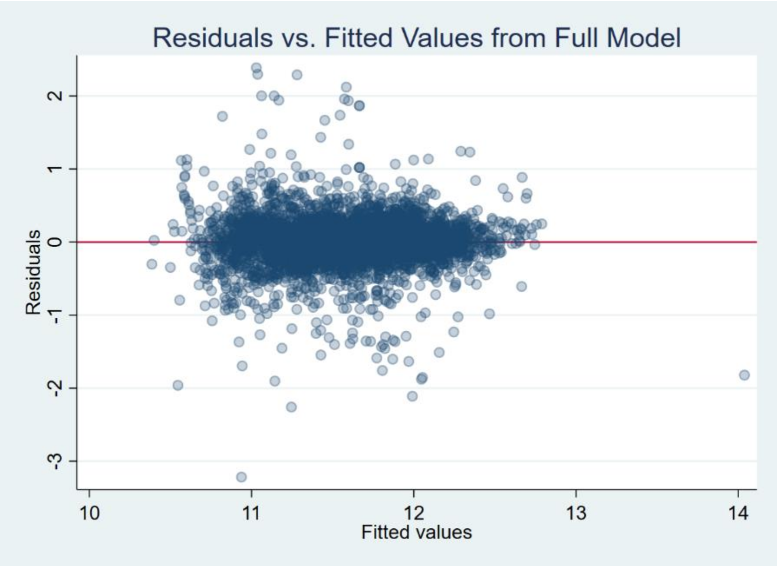
The presence of diminishing marginal return justifies the addition of a term if our goal is to accurately fit the data. To achieve this goal, I will create a squared term for each of these variables. This means we will simply take the squared value of every observation of square footage and acreage in our data set and add them as new variables into the model. Doing this will allow the slope square footage and the acreage to decrease as the values increase and, as a result, capture the diminishing marginal return we observe.

My last variable to focus on is the “condition” variable. In this data set, the condition variable is missing values. A model that does not contain this variable will have 10,204 observations, but a model that includes this variable will only have 9,773 observations. This fact in itself could be a good reason to omit this variable, but I argue this would cause omitted variable bias, meaning that we could not infer a causal relationship from the model. The reason omitting this variable would cause omitted variable bias is because of the correlation between the square footage of a home and the rated condition of that home. It is fair to assume that homes that are larger are also going to be better quality, and as a result, will have a better condition rating.

To do this, I checked the correlation between the square footage of a home and the rated condition of the home. The correlation coefficient between these two variables is 0.5251. This can be considered a moderately high correlation and justifies the idea that the variables are indeed correlated. Therefore, we can conclude that omitting the condition variable would cause the omitted variable bias, and we would not be able to conclude a causal relationship from the difference-in-differences model.

Finally, to end this section, I want to examine the potential for heteroscedasticity in my model. Heteroscedasticity is an issue because when we run regression models in Stata, Stata assumes homoscedasticity in the data. Running a model with heteroskedasticity when assuming homoscedasticity will make our estimators less accurate, therefore we should avoid running this regression when possible (it should be noted that running a heteroskedastic model while assuming homoscedasticity will not make our estimators biased). A simple way to check for heteroskedasticity is with a graph off fitted values plotted against the residuals of the model. A plot of this is shown below:

Figure 5



Our assumption of homoscedasticity is violated if we can observe a trend in the plotted residuals. We can see that the residuals grow larger in magnitude relative to the predicted values, as the predicted values approach 11.5, and they decrease in magnitude after that point. This shows a trend in the plotted residuals, and this justifies a method to adjust for this heteroscedasticity in the model. We do this by using the “robust” command in Stata when running the model.

IV. Regression Analysis

Now, let us examine the full model and its and implications. Here is the full model I constructed and an explanation of the variables:

$$Log\ \widehat{House\ Price} = \hat{B}_0 + \hat{B}_1 year2000 + \hat{B}_2 Churchill + \hat{B}_3 year2000 * Churchill + \hat{B}_4 acres + \hat{B}_5 acres2 + \hat{B}_6 sqft + \hat{B}_7 sqft2 + \hat{B}_8 age + \hat{B}_9 condition$$

The variables “acres,” “sqft,” and “condition” measure the lot size in acres of the house sold, the interior square footage of the house, and the rated condition of the home, respectively. Additionally, “acres2” and “sqft2” are the squared values of “acres” and “sqft” that I created, referenced earlier in this report. Along with continuous variables, there are two dummy variables in this model, which are “year2000” and “Churchill.” These variables only take on the values of one or zero. In the context of the data set, “year2000” will equal one when the house was sold in or after the year, and “Churchill” will equal one if the house was sold in Churchill County.

The term of interest is the interaction between “year2000” and “Churchill.” The slope coefficient of this term captures the difference-in-difference estimator, and, assuming our assumptions of the model hold, gives us the causal effect of the cases of leukemia on real house prices in Churchill County. Below is a table with the values of the slope coefficients for each term and the associated standard error.

Difference-in-Difference Models for The Effect of Cases on Average Real House Prices in Churchill

	(1) Simple Model, no Controls	(2) Model with Control Variables	(3) Robust model with Controls
Sale on or after 2000	0.0395	-0.0205	-0.0205
	(0.0101)	(0.00615)	(0.00537)
Sale Churchill	-0.0399	0.0840	0.0840
	(0.00951)	(0.00659)	(0.00733)

Interaction Term	-0.0774	-0.125	-0.125
	(0.0187)	(0.0114)	(0.0117)
Lot Size in Acres		0.0139	0.0139
		(0.000670)	(0.00177)
Lot Size Squared		-0.0000236	-0.0000236
		(0.00000150)	(0.00000345)
House Square Feet		0.000796	0.000796
		(0.0000241)	(0.0000467)
Square Feet Sqaured		-0.000000101	-0.000000101
		(6.26e-09)	(1.33e-08)
Building Age in Years		-0.00421	-0.00421
		(0.000164)	(0.000248)
Condition of Home		0.193	0.193
		(0.00646)	(0.00767)
Constant	11.63	10.26	10.26
	(0.00596)	(0.0236)	(0.0432)
Observations	10204	9773	9773
Adjusted R^2	0.007	0.641	0.641

Standard errors in parentheses
All $p < 0.001$

In the third column of the table, the slope coefficient of the interaction term is -0.125. In the context of this model, this means leukemia cases decreased average house sale values by 12.5% in Churchill County, if the assumptions of the difference-in-differences model hold.

We can conclude a causal effect because of the implications of the difference-in-differences model. The model allows us to take the difference in average real sale prices between Churchill and Lyon after the spike in cases and subtract this difference from the difference between average real sale prices between Churchill and Lyon before the spike in cases. The difference between average sale prices after the spike in cases contains two types of information. First, it contains information that includes factors that do not vary with time. Second, it includes information that explains the difference in average sale prices caused by the rise in cases. To extract the information that explains the effect of cases on average house sale prices, we can take the difference in house sale prices after the rise in cases and before the rise in house sale prices. This is because the difference between the counties before the rise in cases only contains information that explains the difference in average house sale prices from factors that do not vary with time. Therefore, the differences of these differences leave us with only the information that explains the change in average house prices that is explained by the rise in cases, which is the slope coefficient of the interaction term.

With this being said, I do not believe the assumptions of the difference-in-differences model holds, as zero conditional mean is violated by the omitted variable bias. A variable that was not included in this data set was the household income of the individuals who lived in the house when the house was sold, and therefore it could not be controlled for in this model. The issue being that household income would almost certainly correlate with the condition of the home. As a result, household income being omitted would cause omitted variable bias. One solution would be to omit the condition variable as well, but as stated earlier in this report, the condition of the home correlates with the size of the home, so this variable needs to be controlled for.

Considering this, I cannot conclude that the slope coefficient of the interaction term gives us the causal effect of leukemia cases on average home sale prices in Churchill County.

V. Conclusion

My model does give an association between leukemia cases in Churchill and average house sale prices but not a causal effect. To further study this causal effect, I recommend a new data set be constructed that includes household income. A model that includes household income would more likely estimate the causal effect of cases on average house sale prices.

It is also important that we construct a new data set because we need the household income associated with each observed home sale for every observation. If we only used the average household income of Churchill residents for each year, our results would be inaccurate, and we would be no closer to determining the causal effect of leukemia cases on average house sale prices.

```

cap log close
log using Cesar_RM_Fall22_project, replace

cap log close
log using myproject, replace

use "D:\1A.SchoolHW\2022 Fall\ECN 140\Emperical project\EPDATA_F2022.dta", clear

//installing package
ssc install estout, replace

//variables generated
gen lnrealsales = log(realsales)
gen sqft2 = sqft^2
gen year2000 = year >= 2000
gen y99 = year >= 1999
gen casesPresent = 0
replace casesPresent = 1 if cases > 0
gen acres2 = acres^2

sort year

// Code for data summary

// get a look at how the data is distributed
histogram realsales, frequency  title("Real House Sale Prices, before Log Transformation") // does appeare to be right skewed
histogram lnrealsales, frequency title("Real House Sale Prices, after Log Transformation") xtitle("Natural Log Value of Inflation-adjusted Sale Prices")

//sactter plot of each variable, indiviually
scatter lnrealsales sqft, title("Log House price vs. Home Square Footage") ytitle("Log of Real Sale Price of Home") mcolor(%25)

scatter lnrealsales acres, title("Log House price vs. Plot Acerage") ytitle("Log of Real Sale Price of Home") mcolor(%25)

scatter lnrealsales age, title("Log House price vs. Age of Home") ytitle("Log of Real Sale Price of Home") mcolor(%25)

scatter lnrealsales condition, title("Log House price vs. Perceived condition of Home") ytitle("Log of Real Sale Price of Home") mcolor(%25)

// get correlation coefficient of two variables
correlate sqft condition

//plot residuals to check for heteroskedasticity
reg lnrealsales year2000##county acres sqft sqft2 age condition
rvfplot, yline(0) title("Residuals vs. Fitted Values from Full Model") mcolor(%25)

//

// Code for Regression Analysis

//full model
reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition, robust

reg lnrealsales year##county acres sqft sqft2 age condition, robust
margins, at(year=(1990(1)2002) county=(0(1)1))
marginsplot, title("Average House Price per Year, by County") ytitle("Inflation-Adjusted Sale Price of House") xtitle("Year")

```

```
//create table for different models
reg lnrealsales year2000##county
est store m1

reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition
est store m2


reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition, robust
est store m3

esttab m1 m2 m3 using "regTableDiD.rtf", se label ar2 replace ///
    title("Difference in Difference Models for effect of Cases on House Prices") ///
    mtitles("Simple DiD with no Controls" "DiD with Control Variables" "DiD with Robust SEs and Controls")


log close
```

```
                                                                  ___  ____  ____  ____  ____(R)
                                                                 /__    /   ____/   /   ____/
                                                                ___/   /   /___/   /   /___/
                                                                  Statistics/Data analysis

       county │
      name:  <unnamed>
       log:  D:\1A.SchoolHW\2022 Fall\ECN 140\Emperical project\Cesar_RM_Fall22_project.smcl
  log type:  smcl
 opened on:   2 Dec 2022, 21:17:45

. 
. use "D:\1A.SchoolHW\2022 Fall\ECN 140\Emperical project\EPDATA_F2022.dta", clear

. 
. //installing package
. ssc install estout, replace
checking estout consistency and verifying not already installed...
all files already exist and are up to date.

. 
. //variables generated
. gen lnrealsales = log(realsales)

. gen sqft2 = sqft^2

. gen year2000 = year >= 2000

. gen y99 = year >= 1999

. gen casesPresent = 0

. replace casesPresent = 1 if cases > 0
(1,582 real changes made)

. gen acres2 = acres^2

. 
. sort year

. 
. // Code for data summary
. 
. // get a look at how the data is distributed
. histogram realsales, frequency  title("Real House Sale Prices, before Log Transformation") // does
>  appeare to be right skewed
(bin=40, start=2252.5337, width=22337.679)

. histogram lnrealsales, frequency title("Real House Sale Prices, after Log Transformation") xtitle(
> "Natural Log Value of Inflation-adjusted Sale Prices")
(bin=40, start=7.719811, width=.14964041)

. 
. //sactter plot of each variable, indiviually
. scatter lnrealsales sqft, title("Log House price vs. Home Square Footage") ytitle("Log of Real Sal
> e Price of Home") mcolor(%25)

. 
. scatter lnrealsales acres, title("Log House price vs. Plot Acerage") ytitle("Log of Real Sale Pric
> e of Home") mcolor(%25)
```

```
. 
. scatter lnrealsales age, title("Log House price vs. Age of Home") ytitle("Log of Real Sale Price o
> f Home") mcolor(%25)

. 
. scatter lnrealsales condition, title("Log House price vs. Perceived condition of Home") ytitle("Lo
> g of Real Sale Price of Home") mcolor(%25)

. 
. // get correlation coefficient of two variables
. correlate sqft condition
(obs=9,773)

             |     sqft condit~n
-------------+------------------
        sqft |   1.0000
   condition |   0.5251   1.0000


. 
. //plot residuals to check for heteroskedasticity
. reg lnrealsales year2000##county acres sqft sqft2 age condition

      Source |       SS           df       MS      Number of obs   =     9,773
-------------+----------------------------------   F(8, 9764)      =   2094.10
       Model |  924.601583         8  115.575198   Prob > F        =    0.0000
    Residual |  538.883055     9,764  .055190809   R-squared       =    0.6318
-------------+----------------------------------   Adj R-squared   =    0.6315
       Total |  1463.48464     9,772  .149763062   Root MSE        =    .23493

------------------------------------------------------------------------------------
       lnrealsales | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------------+----------------------------------------------------------------
        1.year2000 |  -.0222743   .0062275    -3.58   0.000    -.0344815    -.010067
                   |
            county |
        Churchill  |   .0833612   .0066713    12.50   0.000     .0702841    .0964382
                   |
   year2000#county |
      1#Churchill  |  -.1210268   .0115758   -10.46   0.000    -.1437177   -.0983359
                   |
             acres |   .0046473   .0003203    14.51   0.000     .0040196    .0052751
              sqft |   .0007923   .0000244    32.53   0.000     .0007446      .00084
             sqft2 |  -9.58e-08   6.33e-09   -15.14   0.000    -1.08e-07   -8.34e-08
               age |  -.0040411   .0001656   -24.41   0.000    -.0043657   -.0037166
         condition |   .1882663   .0065355    28.81   0.000     .1754554    .2010772
             _cons |   10.27468   .0238683   430.47   0.000     10.22789    10.32147
------------------------------------------------------------------------------------

. rvfplot, yline(0) title("Residuals vs. Fitted Values from Full Model") mcolor(%25)

. 
. //
. 
. // Code for Regression Analysis
```

```
43 .
44 . //full model
45 . reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition, robust

   Linear regression                               Number of obs     =      9,773
                                                   F(9, 9763)        =    1021.37
                                                   Prob > F          =     0.0000
                                                   R-squared         =     0.6409
                                                   Root MSE          =     .23202

   ------------------------------------------------------------------------------
                |               Robust
    lnrealsales | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
   -------------+----------------------------------------------------------------
     1.year2000 |  -.0205192   .0053686    -3.82   0.000    -.0310427   -.0099958
                |
         county |
     Churchill  |   .0840392   .0073315    11.46   0.000      .069668    .0984104
                |
 year2000#county |
   1#Churchill  |  -.1251008   .0117209   -10.67   0.000    -.1480761   -.1021254
                |
          acres |   .0139477   .0017738     7.86   0.000     .0104707    .0174247
         acres2 |  -.0000236   3.45e-06    -6.83   0.000    -.0000304   -.0000168
           sqft |   .0007962   .0000467    17.03   0.000     .0007045    .0008878
          sqft2 |  -1.01e-07   1.33e-08    -7.60   0.000    -1.27e-07   -7.49e-08
            age |  -.0042133   .0002478   -17.01   0.000     -.004699   -.0037277
      condition |   .1928592   .0076735    25.13   0.000     .1778176    .2079008
          _cons |   10.26296   .0432453   237.32   0.000     10.17819    10.34773
   ------------------------------------------------------------------------------

46 .
47 . //create table for different models
48 . reg lnrealsales year2000##county

         Source |       SS           df       MS      Number of obs   =    10,204
   -------------+----------------------------------   F(3, 10200)     =     26.40
          Model |  12.1636567         3  4.05455225   Prob > F        =    0.0000
       Residual |   1566.3638    10,200  .153565078   R-squared       =    0.0077
   -------------+----------------------------------   Adj R-squared   =    0.0074
          Total |  1578.52745    10,203   .15471209   Root MSE        =    .39187

   ------------------------------------------------------------------------------
    lnrealsales | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
   -------------+----------------------------------------------------------------
     1.year2000 |   .0394579   .0101355     3.89   0.000     .0195903    .0593254
                |
         county |
     Churchill  |  -.0399441   .0095062    -4.20   0.000    -.0585781   -.0213101
                |
 year2000#county |
   1#Churchill  |  -.0773647   .0187215    -4.13   0.000    -.1140625   -.0406669
                |
          _cons |    11.6274   .0059601  1950.87   0.000     11.61571    11.63908
   ------------------------------------------------------------------------------
```

```
49 . est store m1

50 .
51 . reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition

      Source |       SS           df       MS      Number of obs   =     9,773
-------------+----------------------------------   F(9, 9763)      =   1935.96
       Model |  937.932911         9  104.214768   Prob > F        =    0.0000
    Residual |  525.551727     9,763  .053830967   R-squared       =    0.6409
-------------+----------------------------------   Adj R-squared   =    0.6406
       Total |  1463.48464     9,772  .149763062   Root MSE        =    .23202

------------------------------------------------------------------------------
 lnrealsales | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  1.year2000 |  -.0205192   .0061513    -3.34   0.001    -.0325771   -.0084613
             |
      county |
  Churchill  |   .0840392   .0065887    12.76   0.000      .071124    .0969544
             |
year2000#county |
1#Churchill  |  -.1251008   .0114352   -10.94   0.000    -.1475161   -.1026854
             |
       acres |   .0139477   .0006703    20.81   0.000     .0126337    .0152616
      acres2 |  -.0000236   1.50e-06   -15.74   0.000    -.0000265   -.0000207
        sqft |   .0007962   .0000241    33.10   0.000      .000749    .0008433
       sqft2 |  -1.01e-07   6.26e-09   -16.13   0.000    -1.13e-07   -8.86e-08
         age |  -.0042133   .0001639   -25.71   0.000    -.0045345   -.0038921
   condition |   .1928592   .0064611    29.85   0.000     .1801942    .2055242
       _cons |   10.26296   .0235842   435.16   0.000     10.21673    10.30919
------------------------------------------------------------------------------

52 . est store m2

53 .
54 .
55 . reg lnrealsales year2000##county acres acres2 sqft sqft2 age condition, robust

Linear regression                               Number of obs     =      9,773
                                                F(9, 9763)        =    1021.37
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6409
                                                Root MSE          =     .23202

------------------------------------------------------------------------------
             |               Robust
 lnrealsales | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  1.year2000 |  -.0205192   .0053686    -3.82   0.000    -.0310427   -.0099958
             |
      county |
  Churchill  |   .0840392   .0073315    11.46   0.000      .069668    .0984104
             |
year2000#county |
1#Churchill  |  -.1251008   .0117209   -10.67   0.000    -.1480761   -.1021254
             |
       acres |   .0139477   .0017738     7.86   0.000     .0104707    .0174247
      acres2 |  -.0000236   3.45e-06    -6.83   0.000    -.0000304   -.0000168
        sqft |   .0007962   .0000467    17.03   0.000     .0007045    .0008878
       sqft2 |  -1.01e-07   1.33e-08    -7.60   0.000    -1.27e-07   -7.49e-08
         age |  -.0042133   .0002478   -17.01   0.000     -.004699   -.0037277
   condition |   .1928592   .0076735    25.13   0.000     .1778176    .2079008
       _cons |   10.26296   .0432453   237.32   0.000     10.17819    10.34773
------------------------------------------------------------------------------
```

```
56 . est store m3

57 .
58 . esttab m1 m2 m3 using "regTableDiD.rtf", se label ar2 replace ///
   >         title("Difference in Difference Models for effect of Cases on House Prices") ///
   >         mtitles("Simple DiD with no Controls" "DiD with Control Variables" "DiD with Robust SEs an
   > d Controls")
   (output written to regTableDiD.rtf)

59 .
60 .
61 . log close
         name:  <unnamed>
          log:  D:\1A.SchoolHW\2022 Fall\ECN 140\Emperical project\Cesar_RM_Fall22_project.smcl
     log type:  smcl
    closed on:   2 Dec 2022, 21:17:55
   ─────────────────────────────────────────────────────────────────────────────────────────────────────
```