ECON 1190 Problem Set 5: Difference in Differences

Claire Duquennois

Name:

1 Empirical Analysis from Lucas Davis' (2004, American Economic Review)

This exercise uses data from Lucas Davis' paper, "The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster," published in the *American Economic Review* in 2004. This paper studies the effects of the emergence of a child cancer cluster on housing prices to estimate the willingness to pay to avoid this environmental health risk.

The data can be found by following the link on the AER's website which will take you to the ICPSR's data repository.

2 Set Up

2.1 Loading the Packages

Load any R packages you will be using: **Code:**

```
#install.packages("haven",repos = "http://cran.us.r-project.org")
#install.packages("dplyr",repos = "http://cran.us.r-project.org")

library(haven)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(ggplot2)
```

2.2 Cleaning and constructing the data

Thus far in the course the datasets we have been working with were already assembled and cleaned. When doing econometric analysis from scratch, finding, cleaning and compiling the datasets constitutes much of the work. For this project we will do a little bit more of this prior to analysis since the replication files are much more "raw" then for the other papers we have replicated.

The main datasets used in the analysis consist of four files: two listing information on real estate sales in Churchill county and two listing real estate sales in Lyons county. The variables in these four files are not all coded and labeled in the same way so we need synchronize them.

To save you time and busywork, the 3 code chunks below synchronize three of the four raw data files. You will synchronize the last raw data file and merge it in.

File 1:

```r
#Opening the `cc.dta` file which contains home sales records for Churchill County.

temp1<-read_dta("cc.dta")
temp1<-as.data.frame(temp1)

#Rename and keep only the needed variables
temp1<-temp1 %>%
  rename(
    parcel=var1,
    date=var3,
    usecode=var10,
    sales=var16,
    acres=var17,
    sqft=var19,
    constryr=var20
    )

temp1<-temp1[, c("parcel","date","usecode","sales","acres","sqft","constryr")]

# limiting observations to those where
# 1) the sales date is reported
# 2) is in the time period we are interested in (date<=20001300)
# 3) is for the type of property we are interested in, which will have a usecode of 20.

temp1<-temp1[!is.na(temp1$date),]
temp1<-temp1[temp1$usecode==20,]
temp1<-temp1[temp1$date<=20001300,]

# generate two new variables: a Churchill county indicator, cc and a Lyon County indicator, lc.
temp1$cc<-1
temp1$lc<-0
```

File 2:

```
#Opening the `lc.dta` file which contains home sales records for Lyons County.

temp3<-read_dta("lc.dta")
temp3<-as.data.frame(temp3)

#Rename and keep only the needed variables

temp3<-temp3 %>%
  rename(
    parcel=var1,
    date=var2,
    usecode=var3,
    sales=var4,
    acres=var5,
    sqft=var6,
    constryr=var7
    )

temp3<-temp3[, c("parcel","date","usecode","sales","acres","sqft","constryr" )]

# limiting observations to those where
# 1) the sales date is reported
# 2) is in the time period we are interested in (date<=20001300)
# 3) is for the type of property we are interested in, which will have a usecode of 20.

temp3<-temp3[!is.na(temp3$date),]
temp3<-temp3[temp3$usecode==20,]
temp3<-temp3[temp3$date<=20001300,]

# generate two new variables: a Churchill county indicator, cc and a Lyon County indicator, lc.
temp3$cc<-0
temp3$lc<-1
```

File 3:

```r
#Opening the `lc2.dta` file which contains home sales records for Lyons County.

temp4<-read_dta("lc2.dta")
temp4<-as.data.frame(temp4)

#Rename variables
temp4<-temp4 %>%
  rename(
    parcel=var1,
    date=var2,
    sales=var3,
    acres=var4,
    sqft=var5,
    constryr=var6
    )

# generate two new variables: a Churchill county indicator, cc and a Lyon County indicator, lc.
temp4$cc<-0
temp4$lc<-1

#set the usecode for these data to 20 for all observations
temp4$usecode<-20


# limiting observations to those where
# 1) the sales date is reported
# 2) is in the time period we are interested in (date<=20001300)

temp4<-temp4[!is.na(temp4$date),]
temp4<-temp4[temp4$date>=20001300,]

#keep only the needed variables
temp4<-temp4[, c("parcel","date","usecode","sales","acres","sqft","constryr","cc","lc" )]
```

Merging together the three cleaned files.

```r
temp<-rbind(temp1, temp3, temp4)
rm(temp1, temp3, temp4)
```

2.2.1 Question: Let's clean the `cc2.dta` file. We need to make this set of sales records compatible with the other three sets of sales records we just cleaned and merged.

1) First, load the data and rename the relevant columns so that the names match up and keep the listed variables (see the table below).

2) generated two new variables: `cc` which will be equal to 1 for all observations since this is Churchill county data and `lc` which will equal 0 for all observations

Old Name	New Name	Description
parcel__	parcel	Parcel identification number
sale_date	date	Sale date
land_use	usecode	Land use code
sales_price	sales	Sale price
acreage	acres	Acres
sq_ft	sqft	Square Footage
yr_blt	constryr	Year constructed

Code:

```
temp2<-read_dta("cc2.dta")
temp2<-as.data.frame(temp2)

temp2<-temp2 %>%
  rename(
    parcel=parcel__ ,
    date=sale_date,
    usecode=land_use,
    sales=sales_price,
    acres=acreage,
    sqft=sq_ft,
    constryr=yr_blt
    )


temp2<-temp2[, c("parcel","date","usecode","sales","acres","sqft","constryr" )]

temp2$cc<-1
temp2$lc<-0
```

2.2.2 Question: Compare the formatting of the date variable in the data you are cleaning and the `temp` file you will be merging it with. What do you notice? How is the date formatted in the `temp` dataset and how is it formatted in the one you are cleaning?

Answer: The dates are not formatted in the same way. The first data set uses a YYYYMMDD format while the second appears to be using a MDDYY format.

2.2.3 Question: Convert the dates in the data you are cleaning to the format used in `temp` (YYYYMMDD).

Code:

```
temp2$month=trunc(temp2$date/10000)
temp2$day=trunc(temp2$date/100)-temp2$month*100
temp2$year=2000+temp2$date-temp2$month*10000-temp2$day*100

temp2$date=temp2$year*10000+temp2$month*100+temp2$day
```

2.2.4 Question: Limit your observations to observations where (date>=20001300) and observations where the sales date is reported. Then merge your data to the `temp` file.

```r
temp2<-temp2[!is.na(temp2$date),]
temp2<-temp2[temp2$date>=20001300,]

temp2<-temp2[, c("parcel","date","usecode","sales","acres","sqft","constryr" , "cc", "lc")]

temp<-rbind(temp,temp2)

rm(temp2)
```

2.2.5 Question: Now that we have merged the four files of sales data, we need to create some additional variables and do some further data cleaning. Generate the following seven variables:

- A variable with the sales year
- A variable with the sales month
- A variable with the sales day
- A variable for the age of the home
- The log nominal sales price.
- The quarter (1-4) within the year

Code:

```
temp$year=trunc(temp$date/10000)
temp$month=trunc(temp$date/100)-temp$year*100
temp$day=temp$date-temp$month*100-temp$year*10000

temp$age<-temp$year-temp$constryr

temp$lognomsales<-log(temp$sales)

temp$quarter<-0
temp$quarter[temp$month%in%c(1,2,3)]<-1
temp$quarter[temp$month%in%c(4,5,6)]<-2
temp$quarter[temp$month%in%c(7,8,9)]<-3
temp$quarter[temp$month%in%c(10,11,12)]<-3
```

2.2.6 Question: We now want to check that all the observations in the data make sense and are not extreme outliers and re-code any variables with inexplicable values.

Drop the following observations:

- If the sale price was 0.
- If the home is older then 150
- If the square footage is 0.
- If the square footage is greater than 10000.
- If if date is after Sept. 2002 since that is when the data was collected.
- If the month is 0.

Re-code the following observations:

- If the age of the home is negative, replace with 0.
- If the day is 32 replace with 31.

We also want to make sure there are no duplicate sales records in the data. Drop the duplicate of any observation that shares the same parcel number and sales date, or that shares the same sales price, date, cc, and acres.

Hint: `distinct()` may be useful.

Code:

```r
temp<-temp[temp$sales!=0,]
temp<-temp[temp$age<150,]
temp<-temp[temp$sqft!=0,]
temp<-temp[temp$sqft<10000,]
temp<-temp[!(temp$month==10 & temp$year==2002),]
temp<-temp[temp$month!=0,]
temp$age[temp$age==-1]<-0
temp$day[temp$day==32]<-31

temp<-temp%>% distinct(parcel,date, .keep_all = TRUE)
temp<-temp%>% distinct(sales,date,cc,acres, .keep_all = TRUE)
```

2.2.7 Question: Lyons and Churchill counties could be using the same parcel numbers for different parcels in each county (ie they may each have a parcel identified as 205 within their separate systems). Modify the parcel variable so parcel numbers are uniquely identified.

Code:

```
temp$parcel<-(2*temp$cc*100000000)+(3*temp$lc*100000000)+temp$parcel
```

2.2.8 Question: We want to adjust the sales price using the Nevada Home Price Index (nvhpi) which is available for each quarter in the price.dta file. Merge the index into your dataset and calculate the index adjusted real sales price ($\frac{salesprice*100}{nvhpi}$) as well as the log of this real sales price. What is the base year and quarter of this index?

Code:

```
index<-read_dta("price.dta")

temp <- left_join(temp, index, by = c("year", "quarter"))

temp$realsales<-temp$sales*100/temp$nvhpi
temp$logrealsales<-log(temp$realsales)
```

Answer: The index is set to 100 for the first quarter 2000, which is thus the reference period.

2.2.9 Question: In the paper, Davis maps the cumulative number of leukemia cases that occur in Churchill county in figure 1. For simplicity, we assume a binary treatment: the cancer cluster did not affect outcomes prior to 2000 and did after. Generate a "Post" indicator for years after 1999.

Code:

```
temp$post<-0
temp$post[temp$year>1999]<-1
```

3 Summary Statistics:

3.1 Question: Create a table comparing baseline characteristics for four variable between Lyon and Churchill prior to 2000. What do these regressions tell you and why they are important?

```r
variables<-c("sales", "acres", "sqft", "age")

balreg1<-felm(sales~cc, temp[temp$post==0,])
balreg2<-felm(acres~cc, temp[temp$post==0,])
balreg3<-felm(sqft~cc, temp[temp$post==0,])
balreg4<-felm(age~cc, temp[temp$post==0,])

stargazer(balreg1, balreg2,balreg3, balreg4, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Feb 20, 2023 - 11:34:43 AM

Table 2:

	Dependent variable:			
	sales	acres	sqft	age
	(1)	(2)	(3)	(4)
cc	−5,800.983***	0.101	13.977	6.377***
	(1,212.379)	(0.208)	(10.850)	(0.431)
Constant	109,839.300***	1.277***	1,486.888***	10.493***
	(758.658)	(0.130)	(6.789)	(0.270)
Observations	7,051	7,051	7,051	7,051
R^2	0.003	0.00003	0.0002	0.030
Adjusted R^2	0.003	−0.0001	0.0001	0.030
Residual Std. Error (df = 7049)	49,690.660	8.532	444.693	17.681
Note:		*p<0.1; **p<0.05; ***p<0.01		

Answer: It is important for us to understand the differences between Lyon and Churchill counties because we are effectively going to be using Lyon as a counter factual for Churchill. We want to be convinced that home prices in the two counties should follow parallel trends. There do appear to be some differences between the housing stock and sales between the two counties though they do not seem to be massive. When looking at these differences, we want to think about whether any of these differences could explain a divergent path in home prices after 2000.

4 Analysis:

4.1 Question: Specify and then estimate the standard difference-in-differences estimator to look at how home sales prices changed between Churchill and Lyons county after the emergence of the cancer cluster. Estimate your specification on the log of real home sales and the sales price. (2 pages)

Note: Your results will not exactly match the values in the paper. His approach is more specific. We model the risk perception of the cancer cluster as a [0, 1] variable: 0 prior to 1999 and 1 after. In the paper, he allows for the perceived risk to increase over the time window in which cases were growing, by using the spline function illustrated in figure 1 which creates more variation and detail in the data.

Answer: We can estimate the following where we are interested in the β_3 coefficient,

$$LogRealSales_i = \beta_0 + \beta_1 ChurchillCo_i + \beta_2 Post_i + \beta_3 ChurchillCo_i * Post_i + \epsilon_i.$$

$$Sales_i = \beta_0 + \beta_1 ChurchillCo_i + \beta_2 Post_i + \beta_3 ChurchillCo_i * Post_i + \epsilon_i.$$

Code:

```
reg1<-felm(logrealsales~cc+post+cc*post, temp)
reg2<-felm(sales~cc+post+cc*post, temp)

stargazer(reg1,reg2, type="latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Feb 20, 2023 - 11:34:44 AM

Table 3:

	Dependent variable:	
	logrealsales	sales
	(1)	(2)
cc	−0.039***	−5,800.983***
	(0.009)	(1,211.269)
post	0.041***	24,692.800***
	(0.010)	(1,287.583)
cc:post	−0.077***	−7,422.297***
	(0.019)	(2,378.767)
Constant	11.630***	109,839.300***
	(0.006)	(757.964)
Observations	10,119	10,119
R^2	0.008	0.051
Adjusted R^2	0.007	0.051
Residual Std. Error (df = 10115)	0.388	49,645.190
Note:	*p<0.1; **p<0.05; ***p<0.01	

4.2 Question: Which table in the paper reports equivalent results?

Answer: Table 2 in the paper reports results that are equivalent to those estimated on the log real sales price above.

4.3 Question: Interpret each of the coefficients you estimated in the regression using the log real sales.

Answer: Since the dependent variable is the log of the real sales price, we can interpret the coefficients as percentages. Thus we see that homes in Churchill county sell for about 3.9% less then in Lyons county. Homes sell for about 4.1% more (in real term) in the years after 1999. But that homes in Churchill county after 1999 experience an additional price penalty of 7.7%, which we attribute to the emergence of the cancer cluster.

4.4 Question: Use the estimated coefficients for the effect on the sales price to report the estimated sales price in each of the situations below. Show your calculations.

	Lyon County	Churchill County
Year<=1999		
Year>1999		

Answer:

	Lyon County	Churchill County
Year<=1999	109,839 USD	109,839-5,800=104,039 USD
Year>1999	109,839+24,692= 134,531 USD	109,839+24,692-5,800-7,412= 121,319 USD

4.5 Question: What assumption must hold for us to be able to attribute the estimated effect as the causal effect of the cancer cluster? Do you find the evidence convincing in this case?

Answer:

For these estimates to be cause, we must believe that absent the cancer cluster, home prices in Churchill would have experienced the same price changes as those in Lyons: ie we must believe in the parallel trends assumption. The evidence in Figure 2 is quite compelling in this regards as it seems that home prices in the two counties closely followed the general pattern for Nevada, prior to the emergence of the cancer cluster. In addition, the summary statistics also these counties are quite similar so that Lyon county is a good counter factual for Churchill.

4.6 Question: Re-estimate both your regressions above but with the addition of parcel fixed effects. What concerns does the addition of parcel fixed effects help address? What is the drawback of using this specification?

Code:

```
reg1fe<-felm(logrealsales~cc+post+cc*post|parcel, temp)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
reg2fe<-felm(sales~cc+post+cc*post|parcel, temp)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
stargazer(reg1,reg1fe,reg2, reg2fe, type="latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Feb 20, 2023 - 11:34:44 AM

Table 6:

	Dependent variable:			
	logrealsales		sales	
	(1)	(2)	(3)	(4)
cc	−0.039***		−5,800.983***	
	(0.009)		(1,211.269)	
post	0.041***	−0.013	24,692.800***	19,404.230***
	(0.010)	(0.009)	(1,287.583)	(1,002.443)
cc:post	−0.077***	−0.106***	−7,422.297***	−11,938.330***
	(0.019)	(0.014)	(2,378.767)	(1,658.334)
Constant	11.630***		109,839.300***	
	(0.006)		(757.964)	
Observations	10,119	10,119	10,119	10,119
R^2	0.008	0.954	0.051	0.965
Adjusted R^2	0.007	0.827	0.051	0.865
Residual Std. Error	0.388 (df = 10115)	0.162 (df = 2668)	49,645.190 (df = 10115)	18,694.960 (df = 2668)
Note:				*p<0.1; **p<0.05; ***p<0.01

Answer: Adding parcel fixed effects means that we are using changes in prices for the same house to identify treatment effects. This helps control for many unobservable characteristics about homes that are in the data that we could not otherwise control for. The disadvantage is that our estimates will be driven only by homes that observed being sold more tan once in this period of time, which could be a selected group that is not representative of the typical home. Nonetheless, estimates using this method return similar results which is reassuring.

4.7 Question: In order to better asses how home prices in Churchill and Lyon counties compare to each other over time, calculate the average price of sold homes in each county for 7 two year bins of the data (bin the years 90 and 91 together, 92 and 93 together, . . .). Plot the evolution of this average for the two counties on the same graph. Include bars to indicate the confidence interval of the calculated means. (2 pages)

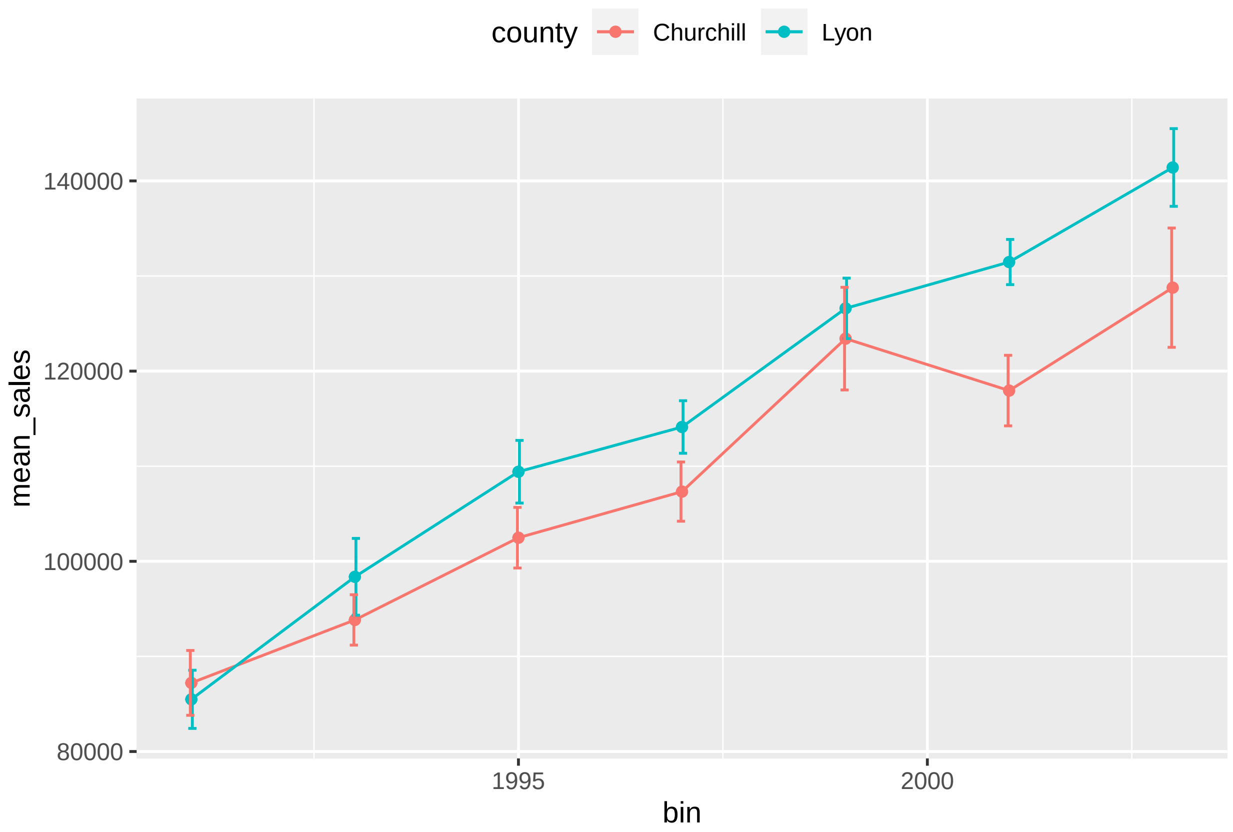
Hint: You want a plot that looks something like the third set of graphs on the following page: http://www.sthda.com/english/wiki/ggplot2-error-bars-quick-start-guide-r-software-and-data-visualization

Code:

```
temp$bin[temp$year%in%c(1990,1991)]<-1991
temp$bin[temp$year%in%c(1992,1993)]<-1993
temp$bin[temp$year%in%c(1994,1995)]<-1995
temp$bin[temp$year%in%c(1996,1997)]<-1997
temp$bin[temp$year%in%c(1998,1999)]<-1999
temp$bin[temp$year%in%c(2000,2001)]<-2001
temp$bin[temp$year%in%c(2002,2003)]<-2003


means<-temp %>% group_by(cc, bin) %>%
  summarize(mean_sales = mean(sales, na.rm = TRUE),  n=n(), sd=sd(sales))
```

```
## ‘summarise()‘ has grouped output by ’cc’. You can override using the ‘.groups‘ argument.
```

```
means$se<-means$sd/sqrt(means$n)

means$county<-"Churchill"
means$county[means$cc==0]<-"Lyon"

plotnew<-ggplot(means, aes(x=bin, y=mean_sales, group=county,color=county)) +
  geom_line()+
  geom_point()+
  geom_errorbar(aes(ymin=mean_sales-1.96*se, ymax=mean_sales+1.96*se), width=.2, position=position_dodge
  theme(legend.position="top")
plotnew
```


county
Churchill
Lyon
mean_sales
140000
120000
100000
80000
1995
2000
bin

4.8 Question: Using the bins of two years constructed above, estimate an event study specification using the 98-99 bin as your omitted category. That is estimate the specification below and present your results in a table. (2 pages)

$$logrealsales_{icb} = \sum_{b=-98/99}^{7} \beta_b Bin_b \times ChurchillCo_c + \lambda_b + \gamma_c + u_{it}.$$

```
temp$cc_1991<-ifelse(temp$bin == 1991 & temp$cc==1, 1, 0)
temp$cc_1993<-ifelse(temp$bin == 1993 & temp$cc==1, 1, 0)
temp$cc_1995<-ifelse(temp$bin == 1995 & temp$cc==1, 1, 0)
temp$cc_1997<-ifelse(temp$bin == 1997 & temp$cc==1, 1, 0)
temp$cc_1999<-ifelse(temp$bin == 1999 & temp$cc==1, 1, 0)
temp$cc_2001<-ifelse(temp$bin == 2001 & temp$cc==1, 1, 0)
temp$cc_2003<-ifelse(temp$bin == 2003 & temp$cc==1, 1, 0)


regev<-felm(logrealsales~cc_1991+cc_1993+cc_1995+cc_1997+cc_2001+cc_2003|cc+year, temp)


stargazer(regev, type="latex")
```

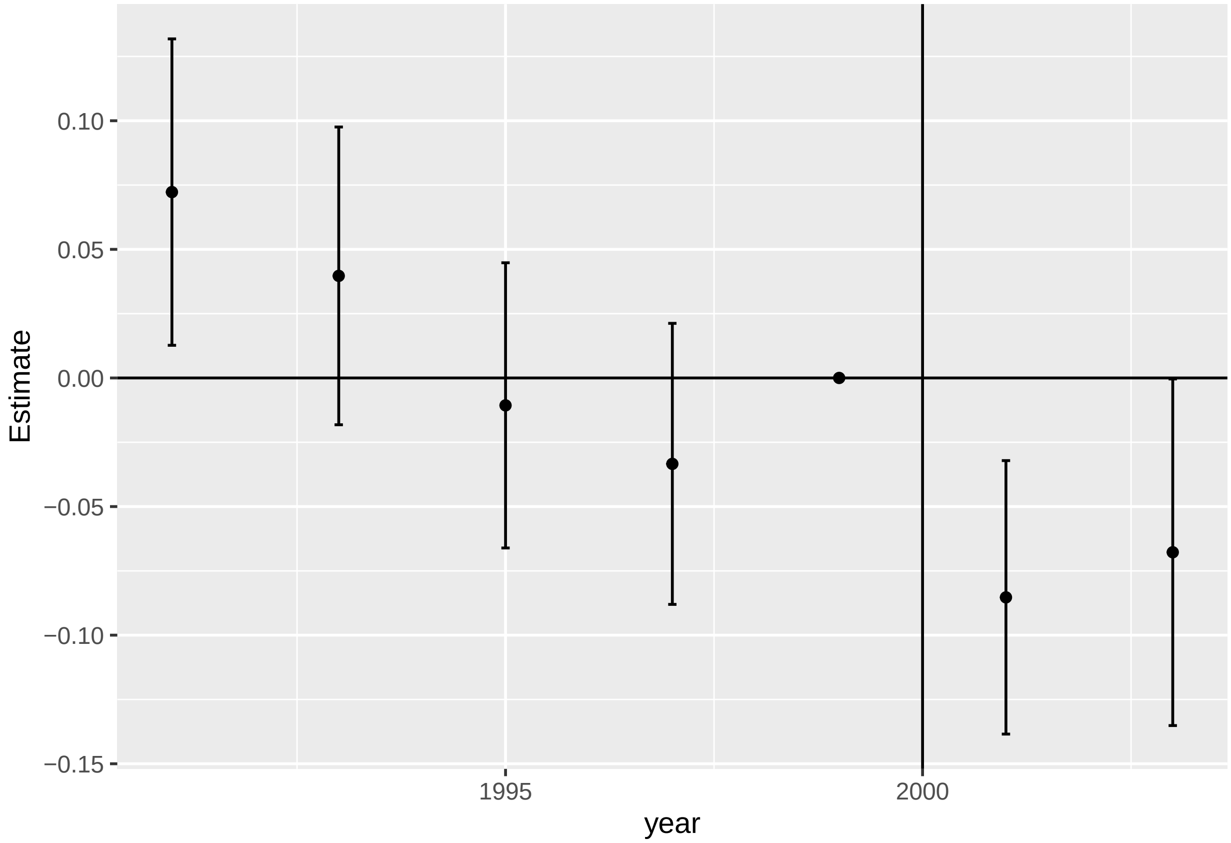
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Feb 20, 2023 - 11:34:45 AM

Table 7:

	Dependent variable:
	logrealsales
cc_1991	0.072**
	(0.030)
cc_1993	0.040
	(0.030)
cc_1995	$-$0.011
	(0.028)
cc_1997	$-$0.033
	(0.028)
cc_2001	$-$0.085***
	(0.027)
cc_2003	$-$0.068**
	(0.034)
Observations	10,119
R^2	0.026
Adjusted R^2	0.024
Residual Std. Error	0.385 (df = 10099)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.9 Question: Use your results to plot an event study figure of your estimates showing your estimated coefficients and 95% confidence level intervals around them.

```
#plot of differences coefficients

res<-coef(summary(regev))
res<-as.data.frame(res)

a<-c(0,0,0,0)

res<-rbind(res,a)

year<-c(1991,1993,1995,1997,2001,2003,1999)
res<-cbind(res,year)
res$ci<-1.96*res$`Std. Error`

names(res)<-c("Estimate","se", "t",  "p", "year", "ci")
# Use 95% confidence interval instead of SEM
didplot2<-ggplot(res, aes(x=year, y=Estimate)) +
    geom_errorbar(aes(ymin=Estimate-ci, ymax=Estimate+ci),width=.1) +
    geom_vline(xintercept = 2000)+
      geom_hline(yintercept = 0)+
     geom_point()
didplot2
```



4.10 Question: What patterns are we looking for in the two graph you just produced?

Answer: We want to see a pattern of parallel trends prior to “treatment” and a break in the pattern of parallel trends after “treatment”. In this case, the cases of pediatric leukemia started gaining notice around 2000. We can see that the average home sale price index of Churchill county follows a similar patterns to that in Lyons prior to this point in time and experiences a break in this pattern after this point in time. However there does seem to be a slight difference in the slope: home prices look like they may be growing more slowly in Churchill county which suggests some problems with making a clear causal claim. This issue is highlighted in the event study estimates.

5 Submission instructions:

1) Knit your assignment in PDF (It should be 28 pages long).
2) Make sure you have ONE question and answer per page (this allows gradescope to easily find your answers).
3) Upload your assignment PDF to gradescope.