Problem Set 5: Difference-in-Differences

Claire Duquennois

Group Member 1:

Group Member 2:

Group Member 3:

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.

2 Set Up

2.1 Loading the Packages

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

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

library(haven)
library(dplyr)
library(stargazer)
library(lfe)
library(ggplot2)
```

2.2 Finding the data

The data can be found by following the link on the AER's website which will take you to the ICPSR's data repository. For this assignment we will be using the `cc.dta`, `cc2.dta`, `lc.dta`, `lc2.dta` and `price.dta` which can be found online. In addition you will want to download the `allpriceindex.dta` file from the course canvas page.

2.3 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.

2.3.1 Question: Open the `cc.dta` file. This file contains home sales records for Churchill County. You will need to rename and keep only the following variables:

Old Name	New Name	Description
var1	parcel	Parcel identification number
var3	date	Sale date
var10	usecode	Land use code
var16	sales	Sale price
var17	acres	Acres
var19	sqft	Square Footage
var20	constryr	Year constructed
var23	class	

Code:

```
temp1<-read_dta("../../../PSdata/PS5_DIDdata/data/cc.dta")
temp1<-as.data.frame(temp1)

temp1<-temp1 %>%
  rename(
    parcel=var1,
    date=var3,
    usecode=var10,
    sales=var16,
    acres=var17,
    sqft=var19,
    constryr=var20,
    class=var23
    )


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

2.3.2 Question: Next we want to limit our observations to observations where the sales date is reported and that are in the time period we are interested in (date<=20001300) and the type of property we are interested in, which will have a usecode of 20.

Code:

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

2.3.3 Question: Finally we need to generated two new variables: a Churchill county indicator, cc, and a Lyon County indicator, lc'. Setccequal to 1 for all observations andlc‘ which will equal 0 for all observations.

Code:

```
temp1$cc<-1
temp1$lc<-0
```

2.3.4 Question: Next open the `cc2.dta` file. We need to make this set of sales records compatible with the set of sales records we just cleaned. The way the variables are coded in this data however are different so we need to rename the relevant columns so that the names match up.

Old Name	New Name	Description
parcel__	parcel	(same as above)
sale_date	date	
land_use	usecode	
sales_price	sales	
acreage	acres	
sq_ft	sqft	
yr_blt	constryr	
class	class	

```
temp2<-read_dta("../../../PSdata/PS5_DIDdata/data/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,
    class=class
    )


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

2.3.5 Question: Here too we need to generated two new variables: `cc` which will be equal to 1 for all observations and `lc` which will equal 0 for all observations.

Code:

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

2.3.6 Question: Compare the formatting of the date variable in the two datasets you are working with. What do you notice? How is the date formatted in the first dataset you loaded and how is it formatted in the second?

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.

2.3.7 Question: Convert the dates reported in the second dataset to the format used in the first (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
```

2.3.8 Question: For this dataset we limit our observations to observations where (date>=20001300) and observations where the sales date is reported.

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

2.3.9 Question: Keep the same variables as in the first data set and merge the two data sets so that the observations from second datasets are added as new rows to the first dataset.

Code:

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

temp<-rbind(temp1,temp2)

rm(temp1, temp2)
```

2.3.10 Question: Next open the `lc.dta` file which has sales data for Lyons county. We need to make this set of sales records compatible as well. Rename the variables as follows.

Old Name	New Name	Description
var1	parcel	(same as above)
var2	date	
var3	usecode	
var4	sales	
var5	acres	
var6	sqft	
var7	constryr	
var11	class	

Code:

```
temp3<-read_dta("../../../PSdata/PS5_DIDdata/data/lc.dta")
temp3<-as.data.frame(temp3)

temp3<-temp3 %>%
  rename(
    parcel=var1,
    date=var2,
    usecode=var3,
    sales=var4,
    acres=var5,
    sqft=var6,
    constryr=var7,
    class=var11
    )


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

2.3.11 Question: Here too we need to generated two new variables but this time set cc equal to 0 for all observations and lc equal 1 for all observations.

Code:

```
temp3$cc<-0
temp3$lc<-1
```

2.3.12 Question: Keep observations where the sales date is reported and that are in the time period we are interested in (date<=20001300) and the type of property we are interested in, which will have a usecode of 20.

Code:

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

2.3.13 Question: Check that everything is compatible and add these observations to your dataset.

Code:

```
temp<-rbind(temp,temp3)

rm(temp3)
```

2.3.14 Question:Repeat these steps with ‘lc2.dta’ where

Old Name	New Name	Description
var1	parcel	(same as above)
var2	date	
var3	sales	
var4	acres	
var5	sqft	
var6	constryr	
var8	class	

Code:

```
temp4<-read_dta("../../../PSdata/PS5_DIDdata/data/lc2.dta")
temp4<-as.data.frame(temp4)

temp4<-temp4 %>%
  rename(
    parcel=var1,
    date=var2,
    sales=var3,
    acres=var4,
    sqft=var5,
    constryr=var6,
    class=var8
    )
```

2.3.15 Question: Generate three new variables: `cc` equal to 0 for all observations; `lc` equal 1 for all observations and `usecode` equal to 20 for all observations.

Code:

```r
temp4$cc<-0
temp4$lc<-1
temp4$usecode<-20
```

2.3.16 Question: Keep observations where the sales date is reported and that are in the time period we are interested in (date>20001300). Check that everything is compatible and add these observations to our dataset.

Code:

```
temp4<-temp4[!is.na(temp4$date),]
temp4<-temp4[temp4$date>=20001300,]
temp4<-temp4[, c("parcel","date","usecode","sales","acres","sqft","constryr","class","cc","lc" )]

temp<-rbind(temp,temp4)

rm(temp4)
```

2.3.17 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
- A variable of the age of the home squared
- A variable of the property acreage squared
- 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$month[temp$month==0]<-1

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

temp$age2<-temp$age*temp$age
temp$acres2<-temp$acres*temp$acres
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
```

2.3.18 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.

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)
```

2.3.19 Question: Modify the class variable so that it is discreet: round the value up to the nearest 0.5 increment between 0 and 4.5. Set any values greater than 5 to 0.

Code:

```
temp$class[temp$class>0 & temp$class<0.5]<-0.5
temp$class[temp$class>0.5 & temp$class<1]<-1
temp$class[temp$class>1 & temp$class<1.5]<-1.5
temp$class[temp$class>1.5 & temp$class<2]<-2
temp$class[temp$class>2 & temp$class<2.5]<-2.5
temp$class[temp$class>2.5 & temp$class<3]<-3
temp$class[temp$class>3 & temp$class<3.5]<-3.5
temp$class[temp$class>3.5 & temp$class<4]<-4
temp$class[temp$class>5]<-0
```

2.3.20 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
```

2.3.21 Question: Create a identifying variable that identifies (ie will be the same for) all home sales that occurred within a particular month in a specific county.

Code:

```
temp$group<-temp$cc*1000000+temp$year*100+temp$month
```

2.3.22 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("../../../PSdata/PS5_DIDdata/data/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.

2.3.23 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:

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

3 Summary Statistics:

3.1 Question: Create a table comparing baseline characteristics between Lyon and Churchill prior to 2000. To do this, USE LOOPING to run several models where a characteristic of interest is regressed on the Churchill county indicator. Store each regression model and report the results. what do they tell you and why they are important?

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

allModelsList <- lapply(paste(variables,"~cc"), as.formula)
allModelsResults <- lapply(allModelsList, function(x) lm(x, temp[temp$post==0,]))


stargazer(allModelsResults[[1]],allModelsResults[[2]],allModelsResults[[3]],allModelsResults[[4]], allM
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 3:03:17 PM

Table 5:

	Dependent variable:				
	sales	acres	sqft	age	class
	(1)	(2)	(3)	(4)	(5)
cc	−5,800.983***	0.101	13.977	6.377***	−0.342***
	(1,212.379)	(0.208)	(10.850)	(0.431)	(0.019)
Constant	109,839.300***	1.277***	1,486.888***	10.493***	2.076***
	(758.658)	(0.130)	(6.789)	(0.270)	(0.012)
Observations	7,051	7,051	7,051	7,051	7,051
R^2	0.003	0.00003	0.0002	0.030	0.043
Adjusted R^2	0.003	−0.0001	0.0001	0.030	0.043
Residual Std. Error (df = 7049)	49,690.660	8.532	444.693	17.681	0.791
F Statistic (df = 1; 7049)	22.894***	0.237	1.659	218.546***	314.803***
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.

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

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

Code:

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


stargazer(reg1,reg2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 3:03:17 PM

Table 6:

	Dependent variable:	
	logrealsales	sales
	(1)	(2)
cc	-0.039^{***}	$-5{,}800.983^{***}$
	(0.009)	(1,211.211)
post	0.041^{***}	$24{,}692.800^{***}$
	(0.010)	(1,287.521)
cc:post	-0.077^{***}	$-7{,}412.232^{***}$
	(0.019)	(2,377.830)
Constant	11.630^{***}	$109{,}839.300^{***}$
	(0.006)	(757.927)
Observations	10,120	10,120
R^2	0.008	0.051
Adjusted R^2	0.007	0.051
Residual Std. Error (df = 10116)	0.388	49,642.800
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: (2 pages) Estimate three new regressions by adjusting your main difference-in-difference specification with logrealsales as the outcome by adding the same controls and fixed effects as those used by Davis in table 3. Cluster your standard errors as he does. How do your estimates compare to his? What is the main difference between this approach and the one that he uses?

Code:

```
regmain1<-felm(logrealsales~cc+post+cc*post+acres+acres2+sqft+age+age2|0|0|group,temp)
regmain2<-felm(logrealsales~cc+post+cc*post+acres+acres2+sqft+age+age2|class+year+month|0|group,temp)
```

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

```
regmain3<-felm(logrealsales~cc+post+cc*post|parcel+year+month|0|group,temp)
```

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

```
stargazer(regmain1,regmain2,regmain3, se = list(regmain1$rse, regmain2$rse, regmain3$rse))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 3:03:18 PM

Answer: These estimates still find a negative impact on home prices but the magnitude of the coefficient is small by about 5%. The two approaches are fairly similar, though his is a bit 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's regression, 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 a little more variation and detail in the data.

Table 9:

	Dependent variable:		
	logrealsales		
	(1)	(2)	(3)
cc	-0.004	0.066^{***}	
	(0.006)	(0.008)	(0.000)
post	-0.003		
	(0.006)	(0.000)	(0.000)
acres	0.012^{***}	0.012^{***}	
	(0.002)	(0.002)	
acres2	-0.00002^{***}	-0.00002^{***}	
	(0.00000)	(0.00000)	
sqft	0.001^{***}	0.0004^{***}	
	(0.00001)	(0.00001)	
age	-0.009^{***}	-0.007^{***}	
	(0.0005)	(0.0005)	
age2	0.00004^{***}	0.00002^{*}	
	(0.00001)	(0.00001)	
cc:post	-0.075^{***}	-0.108^{***}	-0.106^{***}
	(0.011)	(0.011)	(0.019)
Constant	10.893^{***}		
	(0.015)		
Observations	10,120	10,120	10,120
R^2	0.608	0.645	0.955
Adjusted R^2	0.608	0.643	0.829
Residual Std. Error	0.244 (df = 10111)	0.233 (df = 10080)	0.161 (df = 2646)
Note:			*p<0.1; **p<0.05; ***p<0.01

4.7 Question: We would like to check for parallel trends in the pre-period. Using only the data prior to the emergence of the cancer cluster (1990-1998), create an indicator set to 1 for 1990-1994 and set to 0 for 1995-1998. Use your basic specification to test for parallel trends and discuss your results.

Code:

```r
temp$pre<-0
temp$pre[temp$year<1994]<-1

reg1par<-felm(logrealsales~cc+pre+cc*pre, temp[temp$year<1999,])
reg2par<-felm(sales~cc+pre+cc*pre, temp[temp$year<1999,])

stargazer(reg1par,reg2par)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 18, 2023 - 3:03:18 PM

Table 10:

	Dependent variable:	
	logrealsales	sales
	(1)	(2)
cc	−0.049***	−4,552.672***
	(0.013)	(1,566.127)
pre	−0.114***	−22,188.860***
	(0.014)	(1,618.545)
cc:pre	0.063***	3,191.838
	(0.022)	(2,532.026)
Constant	11.655***	114,281.800***
	(0.008)	(980.554)
Observations	6,107	6,107
R^2	0.014	0.046
Adjusted R^2	0.013	0.045
Residual Std. Error (df = 6103)	0.404	47,138.010
Note:	*p<0.1; **p<0.05; ***p<0.01	

Answer: Prior to the emergence of the cancer cluster, the evidence that Lyon and Churchill county homes followed the same price trends is not entirely convincing using this approach. In the regression using log real sales, the results suggest that home prices in Churchill county where on a decreasing relative trend prior to the emergence of the cancer cluster. When using simply the sales price, the coefficient is also positive but not statistically significant at conventional levels.

4.8 Question: (2 pages) 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 each 6 month period. of the data. 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.

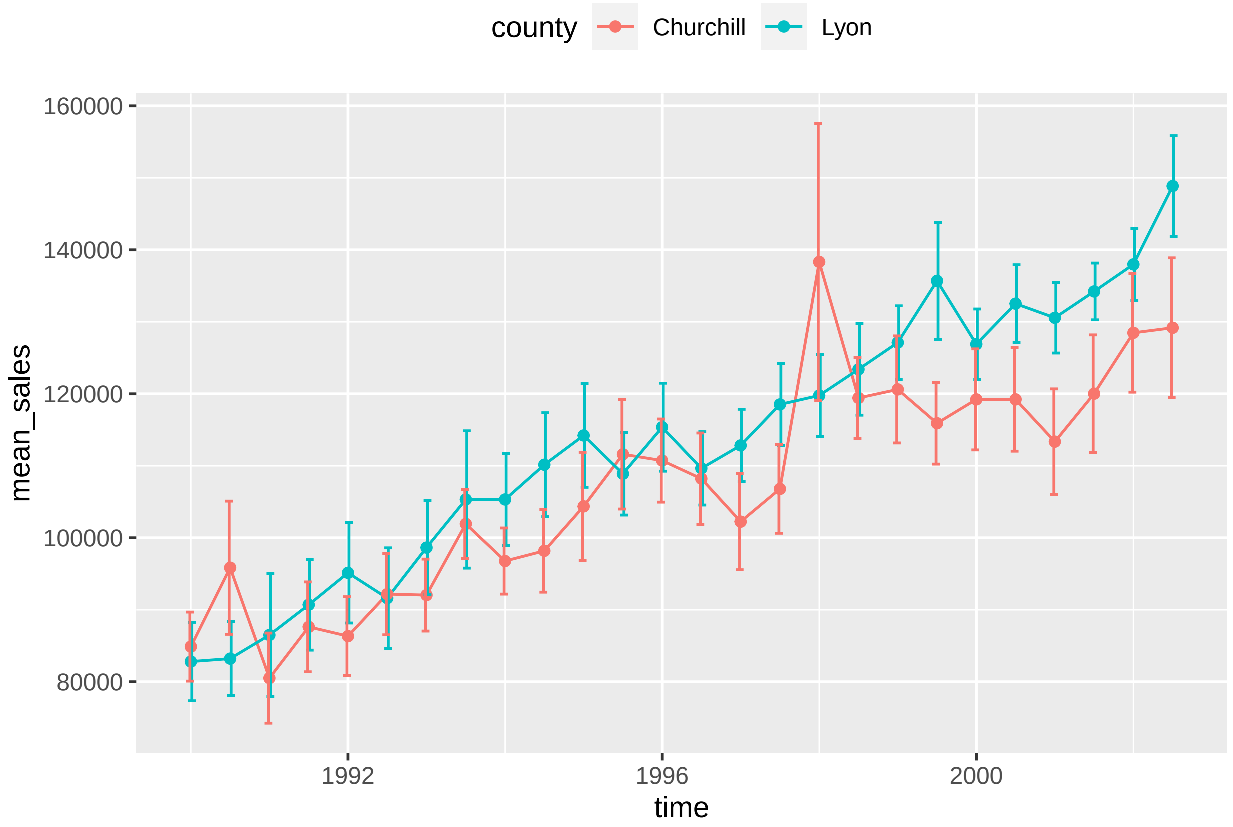
Hint: You want a plot that looks something like the third set of graphs on the following page: http://www.sthda.com/english/wiki/ggplot2-error-bars-quick-start-guide-r-software-and-data-visualization

Code:

```
temp$semester<-0
temp$semester[temp$month%in%c(7,8,9,10,11,12)]<-0.5

temp$time<-temp$year+temp$semester



means<-temp %>% group_by(cc, time) %>%
  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=time, 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
160000
140000
120000
100000
80000
mean_sales
1992
1996
2000
time

4.9 Question: What patterns are we looking for in the two graphs 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 the 1999-2000. We can see that the home price index of Churchill county follows the same patterns as those in Lyons and in Nevada prior to this point in time and experiences a break in this pattern after this point in time.

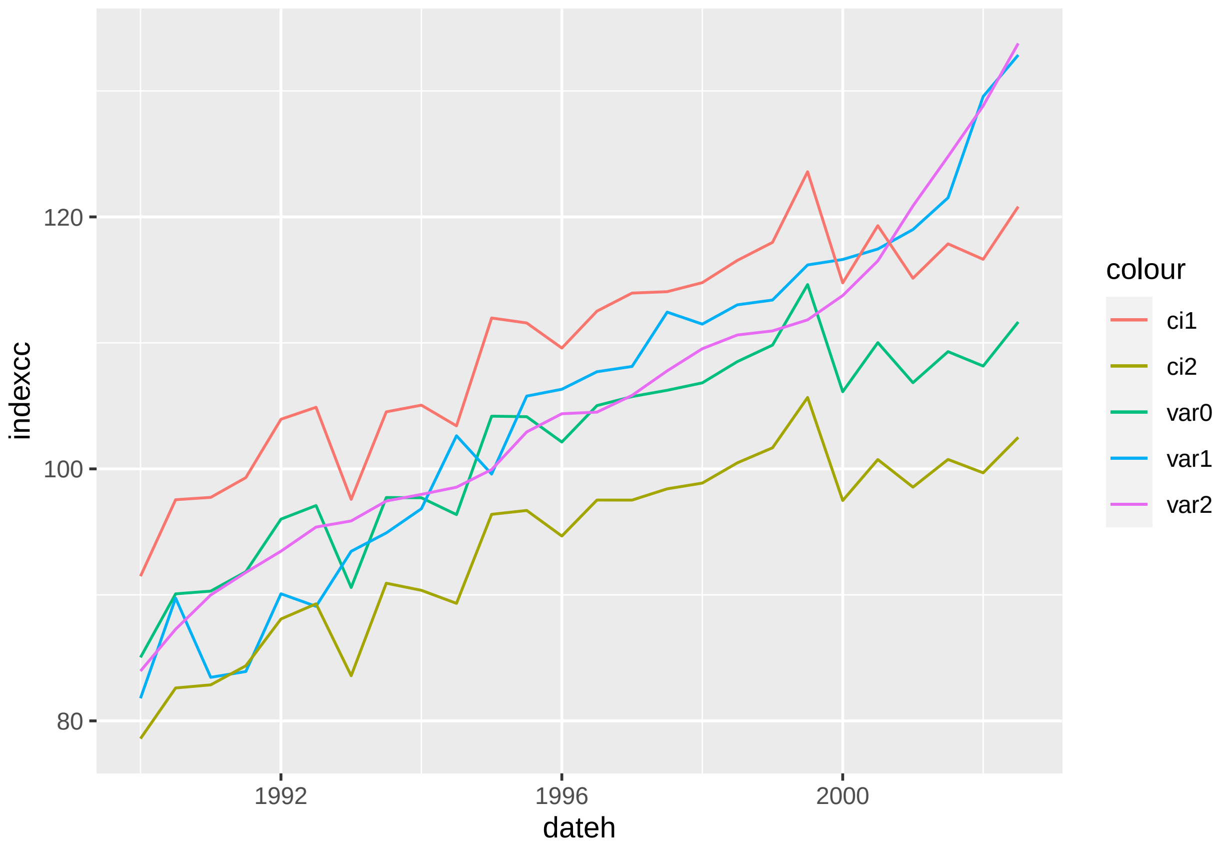
Davis generates a graph similar to the one you just produces but he uses a calculated housing price index for both Lyon and Churchill counties which he combines with the Nevada price index. We will not replicate all these calculations here. I have already replicated the calculations and compiled this data for you. You can find them on the course canvas page. Download the `allpriceindex.dta` file and keep the following variables:

Name	Description
dateh	Year and semester
indexcc	Housing price index in Churchill county
indexlc	Housing price index in Lyon county
indexnv	Housing price index in Nevada
vcc1	Upper confidence interval for indexcc
vcc2	Lower confidence interval for indexcc

4.10 Question: (2 pages) Replicate figures 2 from the paper. Make your figure as visually appealing and informative as possible.

Code:

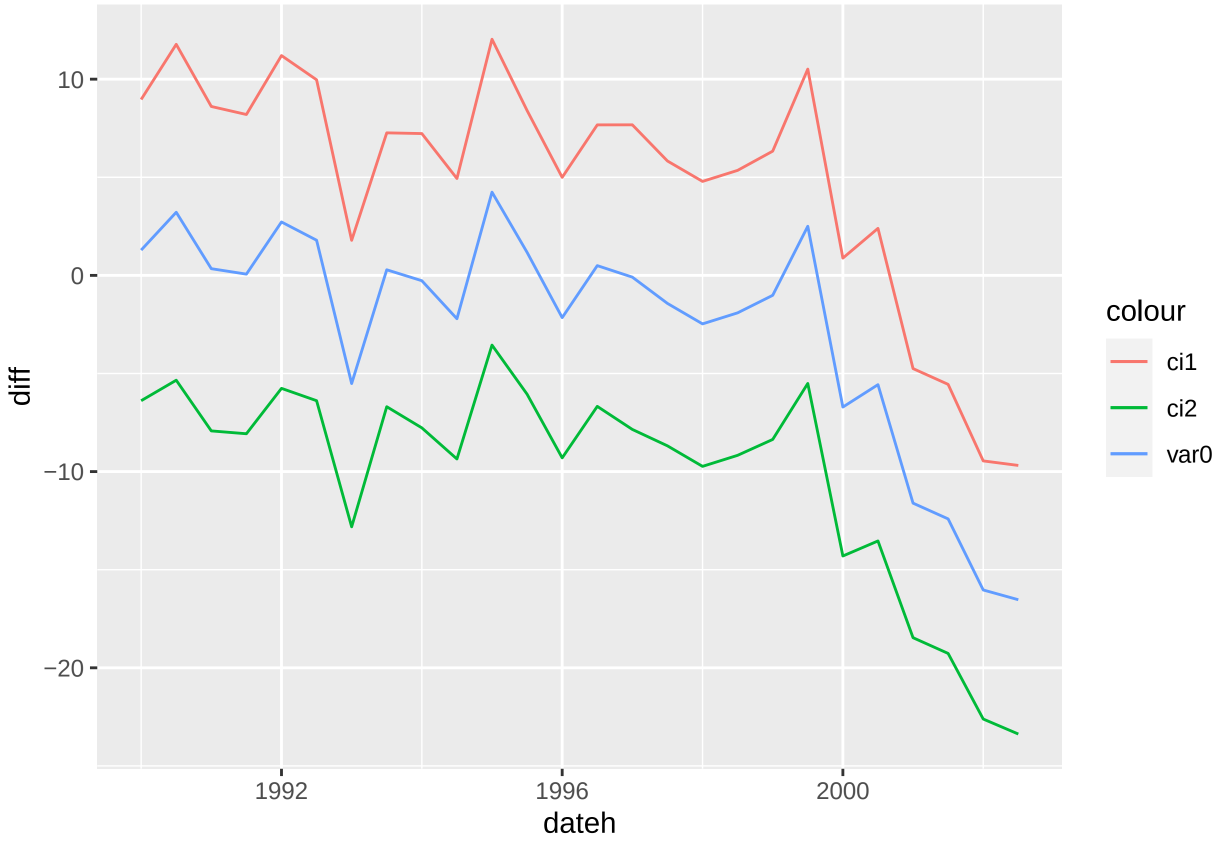
```
figdat<-read_dta("../../../PSdata/PS5_DIDdata/data/allpriceindex.dta")
figdat<-as.data.frame(figdat)
figdat<-figdat[, c("dateh","indexcc","indexlc","indexnv","vcc1","vcc2" )]


plot1<-ggplot(figdat, aes(dateh)) +
    geom_line(aes(y = indexcc, colour = "var0")) +
  geom_line(aes(y = indexlc, colour = "var1"))+
  geom_line(aes(y = indexnv, colour = "var2"))+
  geom_line(aes(y = vcc1, colour = "ci1"))+
  geom_line(aes(y = vcc2, colour = "ci2"))

plot1
```


indexcc
120
100
80
1992
1996
2000
dateh
colour
ci1
ci2
var0
var1
var2

4.11 Question: Calculate the percentage difference in HPI between Churchill county and the state of Nevada. Replicate figures 3 from the paper. Make your figure as visually appealing and informative as possible.

Code:

```
figdat$diff=(figdat$indexcc-figdat$indexnv)*100/figdat$indexnv
figdat$vcc1a=(figdat$vcc1-figdat$indexnv)*100/figdat$indexnv
figdat$vcc2a=(figdat$vcc2-figdat$indexnv)*100/figdat$indexnv

plot2<-ggplot(figdat, aes(dateh)) +
    geom_line(aes(y = diff, colour = "var0")) +
  geom_line(aes(y = vcc1a, colour = "ci1"))+
  geom_line(aes(y = vcc2a, colour = "ci2"))
plot2
```



5 Submission instructions:

- Since this is a group assignment only one member of the group will upload it to gradescope.
- Make sure the final version of your assignment is knit in pdf format and uploaded to gradescope. Make sure you have one question response per page (unless otherwise indicated) so that question positions align with the template in gradescope.The final PDF should be 40 pages long.