ECON 1190 Problem Set 5: Difference in Differences

Claire Duquennois

Name: Rohan Krishnan

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 the R packages we will be using:

```
library(haven)
library(dplyr)
```

```
##
## 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 (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(lfe)
```

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

```
library(broom)
```

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:

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

temp1<-read_dta("~/Downloads/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("~/Downloads/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:

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

temp4<-read_dta("~/Downloads/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.

```
temp<-rbind(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('~/Downloads/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<-temp2[temp2$usecode==20,]
# generate two new variables: a Churchill county indicator, cc and a Lyon County indicator, lc.
temp2$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 data is in YYYYMMDD in the temp file whereas it is written as MDDYY for months 1 to 9 and MMDDYY for months 10 to 12 in the data set I'm cleaning.

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

Code:

```
temp2$date <- as.character(temp2$date)
temp2$date <- ifelse(nchar(temp2$date) == 5, paste0('0', temp2$date), temp2$date)
temp2_date <- as.Date(as.character(temp2$date), format = '%m%d%y')

temp2_convert <- format(temp2_date, '%Y%m%d')
temp2_convert <- as.numeric(temp2_convert)

temp2$date <- temp2_convert
```

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.

```
temp2<-temp2[!is.na(temp2$date),]
temp2<-temp2[temp2$date>=20001300,]
temp<-rbind(temp1, temp2, temp3, temp4)
```

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:

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
temp$year <- as.numeric(substr(temp$date, 1, 4))
temp$month <- as.numeric(substr(temp$date, 5, 6))
temp$day <- as.numeric(substr(temp$date, 7, 8))

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

temp$logsales <- log(temp$sales)

quarters <- ymd(temp$date)
```

```
## Warning: 3 failed to parse.
```

```
temp$quarter <- quarter(quarters)
```

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:

```
#Check duplicates
dup.parcel.date <- duplicated(temp[,c('parcel', 'date')])
data <- temp[!dup.parcel.date,]
dup.pdccacres <- duplicated(data[,c('sales','date','cc','acres')])
data <- data[!dup.pdccacres,]

#Check outliers
data <- subset(data, sales != 0)
data <- subset(data, age <= 150)
data <- subset(data, sqft != 0)
data <- subset(data, sqft <= 10000)
data <- subset(data, date <= 20020900)
data <- subset(data, month != 0)

#Re code values
data$age <- ifelse(data$age < 0, 0, data$age)
data$day <- ifelse(data$day == 32, 31, data$day)
```

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:

```
data$parcel <- as.character(data$parcel)
data$parcel <- ifelse(data$cc == 1, paste0(data$parcel, '_church'), paste0(data$parcel, '_lyons'))
```

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:

```
price <- read_dta('~/Downloads/price.dta')

data2 <- merge(data, price, by = c('year', 'quarter'))
data3 <- data2 %>%
  mutate(adj_rsp = ((sales*100)/nvhpi))
data3 <- data3 %>%
  mutate(logrsp = log(adj_rsp))
```

Answer: The base year and quater is 1990 Q1.

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:

```
data3$post <- ifelse(data3$year>1999, 1, 0)
```

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?

Code:

```
baseline <- data3 %>%
  filter(post == 0)

age.mod <- felm(age ~ cc, baseline)
acres.mod <- felm(acres ~ cc, baseline)
sqft.mod <- felm(sqft ~ cc, baseline)
sales.mod <- felm(sqft ~ cc, baseline)

library(stargazer)

stargazer(age.mod, acres.mod, sqft.mod, sales.mod,
          se=list(age.mod$rse, acres.mod$rse, sqft.mod$rse, sales.mod$rse),
          type = 'latex', header = F, no.space = TRUE,
          title = ' Baseline Characteristics Churchill vs Lyons
          County')
```

Table 2: Baseline Characteristics Churchill vs Lyons County

	Dependent variable:			
	age	acres	sqft	
	(1)	(2)	(3)	(4)
cc	6.391***	0.101	13.904	13.904
	(0.459)	(0.176)	(11.022)	(11.022)
Constant	10.486***	1.278***	1,486.922***	1,486.922***
	(0.236)	(0.160)	(6.597)	(6.597)
Observations	7,047	7,047	7,047	7,047
R^2	0.030	0.00003	0.0002	0.0002
Adjusted R^2	0.030	−0.0001	0.0001	0.0001
Residual Std. Error (df = 7045)	17.681	8.535	444.729	444.729
Note:				*p<0.1; **p<0.05; ***p<0.01

Answer: These regression indicate the difference in average sales age of homes, acres, sqft, and sales price are between Churchill and Lyons county. $/beta_0$ gives the Lyons county average and beta_0 + beta_1 gives the Churchill county average. These regressions are important because it allows to establish a baseline difference between the two groups before examining the post period.

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: The DID estimator examines the effects of the treated group interacted with the treatment period. If the estimator of the interaction is significant, it indicates that Churchill county experienced a statistically significant change after 2000.

In the log sales model, the estimator was -0.078 with a p-value close to 0. Similarly, in the sales model, the estimator was -7784.6 with a p-value close to 0. Since both estimators were highly statistically significant, we can reject the null hypothesis and concluded there is some relationship between the interaction of being post 2000 and being in Churchill county with log sales and sales.

Code:

```
data3$diffindiff <- data3$post * data3$cc
did.sales.mod <- felm(sales ~ cc + post + diffindiff, data = data3)
did.logrsp.mod <- felm(logrsp ~ cc + post + diffindiff, data = data3)

stargazer(did.logrsp.mod, did.sales.mod,
          se = list(did.logrsp.mod$rse, did.sales.mod$rse),
          type = 'latex', header = F, no.space = TRUE,
          title = 'DID for Sales Price')
```

Table 3: DID for Sales Price

	Dependent variable:	
	logrsp	sales
	(1)	(2)
cc	−0.039***	−5,794.951***
	(0.010)	(1,192.116)
post	0.039***	24,541.750***
	(0.009)	(1,322.663)
diffindiff	−0.078***	−7,784.605***
	(0.019)	(2,316.072)
Constant	11.628***	109,826.900***
	(0.006)	(782.265)
Observations	10,062	10,062
R^2	0.008	0.050
Adjusted R^2	0.007	0.050
Residual Std. Error (df = 10058)	0.389	49,657.010
Note:	*p<0.1; **p<0.05; ***p<0.01	

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

Answer: Table 2 reports equivalent results, though they are displayed differently.

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

Answer: beta_0 = 11.628: Properties not in Churchill and before 2000, the average log real sales price is $11.628.

beta_1 = -0.039: Properties in Churchill are associated with a statistically significant decrease of 3.9% in the log sales price.

beta_2 = 0.039: Properties after 1999 are associated with a statistically significant increase of 3.9% in log real sales price.

beta_3 = -0.078: Properties in Churchill sold after 1999 are associated with a statistically significant decrease of 7.8% in log real sales price.

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	109826.9	104031.9
Year>1999	134368	120789.1

Answer: Lyon <=1999: 109826.9 + 0 + 0 + 0 = 109826.9 Lyon >1999: 109826.9 + 0 + 24541.8 + 0 = 134368.7 Churchill <=1999: 109826.9 -5795 + 0 + 0 = 104031.9 Churchill >1999: 109826.90 + 24541.8 - 5794.95 - 7784.61 = 120789.1

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: The parallel trends assumption must hold. We must believe that the sales price in Churchill and Lyons changed parallel (approx.) to each other during the pre period. Since the DID estimator is significant and the counties are in similar locations, I think the evidence is sufficient to contest that PTA holds. It is important to note, however, that each county may have had localized economic phenomena effecting their sales prices that are not captured in the analysis.

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:

```
did.sales.modfe <- felm(sales ~ cc + post + diffindiff | parcel, data = data3)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite
```

```
did.logrsp.modfe <- felm(logrsp ~ cc + post + diffindiff | parcel, data = data3)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or not positive definite
```

```
stargazer(did.sales.modfe, did.logrsp.modfe,
          se = list(did.sales.modfe$rse, did.logrsp.modfe$rse),
          type = 'latex', header = F, no.space = TRUE,
          title = 'DID Sales Price with Fixed Effects')
```

Table 5: DID Sales Price with Fixed Effects

	Dependent variable:	
	sales	logrsp
	(1)	(2)
cc		
	(0.000)	(0.000)
post	19,331.010***	−0.013
	(1,731.851)	(0.011)
diffindiff	−12,030.140***	−0.105***
	(2,206.747)	(0.019)
Observations	10,062	10,062
R^2	0.965	0.955
Adjusted R^2	0.865	0.826
Residual Std. Error (df = 2636)	18,742.630	0.163
Note:	*p<0.1; **p<0.05; ***p<0.01	

Answer: Parcel fixed effects essentially gives each house its own coefficient. This removes omitted variables that may be different between houses like school district. There is a risk of magnifying individual house factors that contribute sale price and also losing some of the country wide effects on sales price.

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:

```
library(ggplot2)
data.binned <- data3 %>%
  mutate(year_bins = cut(year, breaks = 7))
bins <- data.binned %>%
  group_by(year_bins, cc) %>%
  summarize(avgprice = mean(sales), sd = sd(sales))
```

```
## `summarise()` has grouped output by 'year_bins'. You can override using the
## `.groups` argument.
```

```
ggplot(bins, aes(x=year_bins, y = avgprice, group = factor(cc), color = factor(cc))) +
  geom_line() +
  geom_point() +
  geom_errorbar(aes(ymin = avgprice-1.96*sd, ymax = avgprice+1.96*sd),
                width = .2, position = position_dodge(.05)) +
  labs(title = 'Change in Mean Price over Time by County',
       x = 'Year', y = 'Average House Price', color = 'County') +
  scale_color_discrete(labels = c('Churchill', 'Lyons')) +
  scale_y_continuous(breaks = 11, labels = 2) +
  theme_bw()
```


Change in Mean Price over Time by County
Average House Price
2
(1990,1992]
(1992,1993]
(1993,1995]
(1995,1997]
(1997,1999]
(1999,2000]
(2000,2002]
Year
County
Churchill
Lyons

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}^7beta_{b}Bin_b times ChurchillCo_c+lambda_b+gamma_c+u_{it}.

Code:

```
data.binned$year_ind <- factor(relevel(data.binned$year_bins, ref = "(1997,1999]"))

binned.mod <- felm(logrsp ~ year_ind*cc + year_ind + cc, data = data.binned)

stargazer(binned.mod, se = list(binned.mod$rse), type = 'latex',
          header = F, no.space = TRUE,
          title = 'Binned Years Regression')
```

Table 6: Binned Years Regression

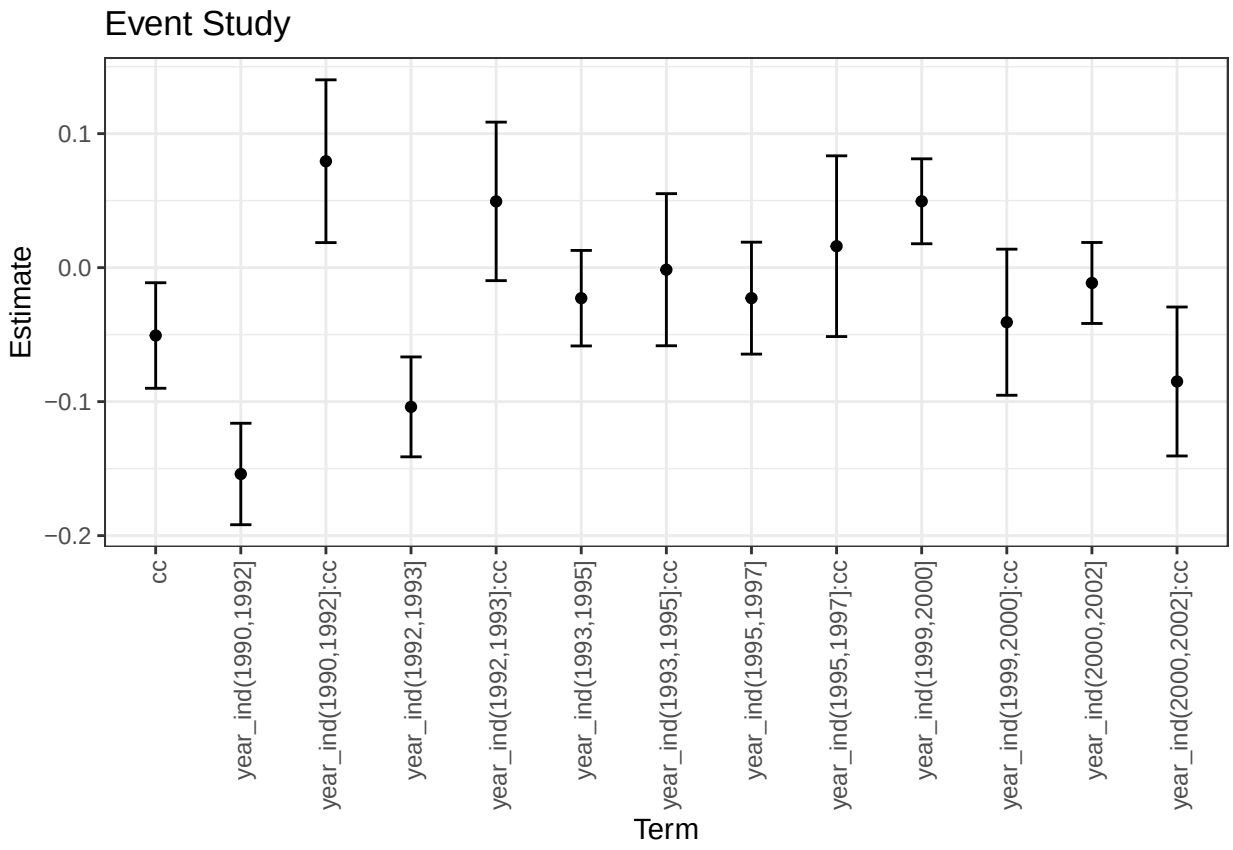
	Dependent variable:
	logrsp
year_ind(1990,1992]	-0.154^{***}
	(0.021)
year_ind(1992,1993]	-0.104^{***}
	(0.020)
year_ind(1993,1995]	-0.023
	(0.018)
year_ind(1995,1997]	-0.023
	(0.020)
year_ind(1999,2000]	0.050^{***}
	(0.015)
year_ind(2000,2002]	-0.011
	(0.014)
cc	-0.051^{**}
	(0.021)
year_ind(1990,1992]:cc	0.079^{**}
	(0.035)
year_ind(1992,1993]:cc	0.049
	(0.031)
year_ind(1993,1995]:cc	-0.002
	(0.030)
year_ind(1995,1997]:cc	0.016
	(0.036)
year_ind(1999,2000]:cc	-0.041
	(0.027)
year_ind(2000,2002]:cc	-0.085^{***}
	(0.030)
Constant	11.666^{***}
	(0.011)
Observations	10,062
R^2	0.024
Adjusted R^2	0.023
Residual Std. Error	0.386 (df = 10048)
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.

Code:

```
binned.mod.coef <- tidy(binned.mod)
binned.mod.coef <- binned.mod.coef[!binned.mod.coef$term == "(Intercept)",]

ggplot(binned.mod.coef, aes(x = term, y = estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin = estimate - 1.96 * std.error,
                    ymax = estimate + 1.96 * std.error), width = 0.25) +
  labs(title = "Event Study", x = "Term", y = "Estimate") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.60, hjust = 1))
```



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

Answer: For the first graph, we are looking at hte difference between Lyons and Churchill before and after the treatment period to see if the path they follow significantly changes post treatment. The intervals allow us to better see if there is a significant difference post 1999.

For the second graph, We want to see if the confidence intervals show a significant difference for the interaction term. The interaction terms illustrate the expected effect on sales price for being in Churchill at a specific binned year. If the interval avoids 0, we can say with 95% confidence that there was some significant effect. By comparing the interaction to the year variable, we can see if being in Churchill was significantly affecting sales price of houses.

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.