

# Reproducing Research; The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster

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

This project 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:**

```
library(haven)
library(dplyr)
library(gridExtra)
library(ggplot2)
library(scales)
library(lfe)
library(stargazer)
library(lubridate)
library(kableExtra)
```

## 2.2 Cleaning and constructing the data

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 to synchronize them.

**File 1:**

```
#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:

```
#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.

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

### 2.2.1 \*\*Let's clean the cc2.dta file.

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:

```
temp5<-read_dta("cc2.dta")
temp5<-as.data.frame(temp5)

temp5<-temp5 %>%
  rename(
    parcel =parcel__,
    date =sale_date,
    usecode =land_use,
    sales =sales_price,
    acres=acreage,
    sqft=sq_ft,
    constryr =yr_blt
  )

temp5$cc<-1
temp5$lc<-0

temp5<-temp5 %>% select(parcel,date,usecode,sales,acres,sqft,constryr,cc,lc )
```

### 2.2.2 How is the date formatted in the temp dataset and how is it formatted in the one we are cleaning?

Answer:

The date in the new file is YYMMDD and the original temp file is YYYYMMDD.

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

Code:

```
# Convert from YYMMDD to Date object and then format it to YYYYMMDD
temp5$date <- format(mdy(temp5$date), "%Y%m%d")
```

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

Code:

```
temp5<-temp5[!is.na(temp5$date),]
temp5<-temp5[temp5$date>=20001300,]

temp<-rbind(temp, temp5)
```

### 2.2.5 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$sale_yr <- as.numeric(substr(temp$date, 1, 4))
temp$sale_mm<- as.numeric(substr(temp$date, 6, 7))
temp$sale_day <- as.numeric(substr(temp$date, 9, 10))
temp$age<- temp$sale_yr - as.numeric(temp$constryr)
temp$log<-log(temp$sales)
temp$date <- ymd(temp$date)
temp$q<- quarter(temp$date, with_year = F)
```

**2.2.6** 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 than 150
- If the square footage is 0.
- If the square footage is greater than 10000.
- If the 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:**

```
temp <- temp %>% filter(sales != 0)
temp <- temp %>% filter(sqft != 0)
temp <- temp %>% filter(sale_mm != 0)
temp <- temp %>% filter(age < 150)
temp <- temp %>% filter(date < as.Date("2002-09-01"))
temp <- temp %>% filter(sqft < 10000)

temp$age <- ifelse(temp$age < 0, 0, temp$age)
temp$sale_day <- ifelse(temp$sale_day == 32, 31, temp$sale_day)

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



2.2.7 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$parcel <- paste0(temp$cc, temp$parcel)
```

2.2.8 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:

```
nvhpi<-read_dta("price.dta")
nvhpi<-as.data.frame(nvhpi)

tempn <- merge(temp, nvhpi[, c("year", "quarter", "nvhpi")],
               by.x = c("sale_yr", "q"),
               by.y = c("year", "quarter"),
               all.x = TRUE)

tempn$adj_index<- (tempn$sales*100)/tempn$nvhpi

tempn %>% filter(nvhpi==100) %>% select(sale_yr,q) %>% head(1) %>% kable(format = "latex") %>%
  kable_styling(latex_options = "HOLD_position","striped")
```

| sale_yr | q |
|---------|---|
| 2000    | 1 |

**Answer:**

Year 2000 quarter 1. This is because for the base year, the adjusted price and actual price will be equal. For these to be equal the index must equal 100 to cancel out the numerator multiplication by 100.

**2.2.9** 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:

```
tempn$Post <- ifelse(tempn$sale_yr > 1999, 1, 0)
```

### 3 Summary Statistics:

3.1 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:

```
d1<- tempn %>% filter(sale_yr<2000) %>% select(sales,acres,sqft,age,cc)

d1 <- d1 %>%
  mutate(cc_label = ifelse(cc == 0, "Lyon County", "Churchill County"))

summary_stats <- d1 %>%
  group_by(cc_label) %>%
  summarise(
    Mean_Sale_Price = mean(sales, na.rm = TRUE),
    SD_Sale_Price = sd(sales, na.rm = TRUE),
    Mean_acres= mean(acres, na.rm = TRUE),
    SD_acres = sd(acres, na.rm = TRUE),
    Mean_sqft = mean(sqft, na.rm = TRUE),
    SD_sqft = sd(sqft, na.rm = TRUE),
    Mean_age = mean(age, na.rm = TRUE),
    SD_age = sd(age, na.rm = TRUE)
  ) %>% t()

options(scipen = 999)

# Create the table with summary statistics
summary_stats %>%
  kable(caption = "Summary Stats Comparing Churchill and Lyon County") %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(2:ncol(summary_stats), width = "3cm") %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

Table 2: Summary Stats Comparing Churchill and Lyon County

| cc_label        | Churchill<br>County | Lyon County |
|-----------------|---------------------|-------------|
| Mean_Sale_Price | 104060.6            | 109700.2    |
| SD_Sale_Price   | 47363.51            | 50116.53    |
| Mean_acres      | 1.358784            | 1.277188    |
| SD_acres        | 3.778717            | 10.619720   |
| Mean_sqft       | 1500.558            | 1487.027    |
| SD_sqft         | 464.1558            | 430.5895    |
| Mean_age        | 16.87541            | 10.51184    |
| SD_age          | 20.67638            | 15.46309    |

Answer:

This shows that between the four characteristics the means and standard deviations are similar, and therefore comparable. This makes a stronger case the the difference in difference approach that the author uses.

## 4 Analysis:

### 4.1 \*\* 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 specification on the log of real home sales and the sales price.\*\*

Note: our 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:**

**Code:**

```
m1<- felm(log~Post+cc+Post*cc, data=tempn)
m2<- felm(sales~Post+cc+Post*cc, data=tempn)
stargazer(m1,m2)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Nov 24, 2024 - 6:57:36 PM

Table 3:

|                                 | <i>Dependent variable:</i> |                              |
|---------------------------------|----------------------------|------------------------------|
|                                 | log                        | sales                        |
|                                 | (1)                        | (2)                          |
| Post                            | 0.231***<br>(0.011)        | 24,732.440***<br>(1,295.891) |
| cc                              | -0.048***<br>(0.010)       | -5,639.533***<br>(1,214.235) |
| Post:cc                         | -0.076***<br>(0.020)       | -7,953.154***<br>(2,400.092) |
| Constant                        | 11.519***<br>(0.006)       | 109,700.200***<br>(761.825)  |
| Observations                    | 9,860                      | 9,860                        |
| R <sup>2</sup>                  | 0.065                      | 0.052                        |
| Adjusted R <sup>2</sup>         | 0.065                      | 0.051                        |
| Residual Std. Error (df = 9856) | 0.401                      | 49,248.300                   |
| <i>Note:</i>                    |                            |                              |
| *p<0.1; **p<0.05; ***p<0.01     |                            |                              |

## 4.2 Interpret each of the coefficients you estimated in the regression using the log real sales.

**Answer:**

$\beta_0$ : Houses in the control group (lyon county) have an average log(house price) of 11.519

$\beta_1$ : after the treatment, the log(house price) increase by approximately 23.1%

$\beta_2$ : Being in the treatment group (churchill county) is associated with a 4.8% decrease in log real sales

$\beta_3$  the combined effect of Treatment and churchill county reduces log real sales by approximately 7.6%.

## 4.3 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 $\leq$ 1999<br>(pre) | 109,700.2   | 104,060.67       |
| Year $>$ 1999<br>(post)   | 134,432.64  | 120,839.95       |

**Answer:**

Case 1:  $post = 0$  &  $churchill = 0$

- 109,700.2

Case 2:  $post = 0$  &  $churchill = 1$

- $109,700.2 - 5,639.53(1) = 104,060.67$

Case 3:  $post = 1$  &  $churchill = 0$

- $109,700.2 + 24,732.44(1) = 134,432.64$

Case 4:  $post = 1$  &  $churchill = 1$

- $109,700.2 + 24,732.44(1) - 5,639.53(1) - 7,953.154(1) = 120,839.95$

**4.4 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:**

Parallel trends assumption, this assumes that without the cancer cluster the housing prices of Churchill county would follow the same pricing pattern over time as Lyon county. That way we can say that the only meaningful reason for the deviation of Churchill county is because of the cancer cluster.

The evidence for this is convincing because we have addition time series data for each county housing prices, and they have roughly followed the same pattern for the past decade. However, having more than just two counties in the sample would make it more convincing.

**4.5 Re-estimate both 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:**

```
m1<- felm(log~Post+cc+Post*cc|parcel, data=tempn)
m2<- felm(sales~Post+cc+Post*cc|parcel, data=tempn)
stargazer(m1,m2)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Nov 24, 2024 - 6:57:37 PM

Table 5:

|                                 | <i>Dependent variable:</i>  |                               |
|---------------------------------|-----------------------------|-------------------------------|
|                                 | log                         | sales                         |
|                                 | (1)                         | (2)                           |
| Post                            | 0.164***<br>(0.010)         | 18,329.420***<br>(994.386)    |
| cc                              | -0.349***<br>(0.092)        | -39,750.330***<br>(9,237.041) |
| Post:cc                         | -0.094***<br>(0.017)        | -10,677.550***<br>(1,664.551) |
| Observations                    | 9,860                       | 9,860                         |
| R <sup>2</sup>                  | 0.951                       | 0.967                         |
| Adjusted R <sup>2</sup>         | 0.809                       | 0.871                         |
| Residual Std. Error (df = 2542) | 0.181                       | 18,133.760                    |
| <i>Note:</i>                    | *p<0.1; **p<0.05; ***p<0.01 |                               |

**Answer:**

parcel is a house, by having a fixed effect for each house you are controlling for any unobserved (omitted variables) for that house that could contribute to its housing price.

This could potentially lead to overfitting of the model, by having a large number of fixed effects with few observations for each of them.

Additionally adding a fixed effect for house doesn't allow for changes to the house over time, for example renovations, a fixed effect for parcel holds all house characteristics constant, so it can't account for renovations.

- 4.6 In order to better assess 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.

Code:

```
tempn$year_bin <- cut(tempn$sale_yr,
                      breaks = c(1990, 1991, 1993, 1995, 1997, 1999, 2001, 2003),
                      labels = c("1990-91", "1992-93", "1994-95", "1996-97", "1998-99", "2000-01",
                                "2002-03"),
                      right = TRUE)
```

```
results <- tempn %>%
  group_by(cc, year_bin) %>%
  summarise(
    mean_price = mean(sales, na.rm = TRUE),
    sd_price = sd(sales, na.rm = TRUE),
    n = n(),
    se_price = sd_price / sqrt(n), # standard error
    ci_lower = mean_price - qt(0.975, df = n-1) * se_price, # lower bound of 95% CI
    ci_upper = mean_price + qt(0.975, df = n-1) * se_price # upper bound of 95% CI
  )
```

```
Churchill<- head(results,7)
Lyon<- results[9:15, ]
```

```
p <- ggplot() +
  geom_line(data = Churchill, aes(x = year_bin, y = mean_price, group = 1, color = "Lyon"),
            linewidth = 1, alpha = 0.7) +
  geom_point(data = Churchill, aes(x = year_bin, y = mean_price, color = "Lyon"),
             size = 3, shape = 16, alpha = 0.7) +
  geom_errorbar(data = Churchill, aes(x = year_bin, ymin = ci_lower, ymax = ci_upper,
                                     color = "Lyon"), width = 0.2, size = 0.75, alpha = 0.7) +

  geom_line(data = Lyon, aes(x = year_bin, y = mean_price, group = 1, color = "Churchill"),
            linewidth = 1, alpha = 0.7) +
  geom_point(data = Lyon, aes(x = year_bin, y = mean_price, color = "Churchill"),
             size = 3, shape = 16, alpha = 0.7) +
  geom_errorbar(data = Lyon, aes(x = year_bin, ymin = ci_lower, ymax = ci_upper,
                                 color = "Churchill"), width = 0.2, size = 0.75, alpha = 0.7) +

  labs(
    title = "Average House Prices Over Time by County",
    subtitle = "Comparing Churchill and Lyon counties",
    x = "Years",
    y = "Average House Price"
  ) +
  theme_minimal(base_size = 15) +
```



```

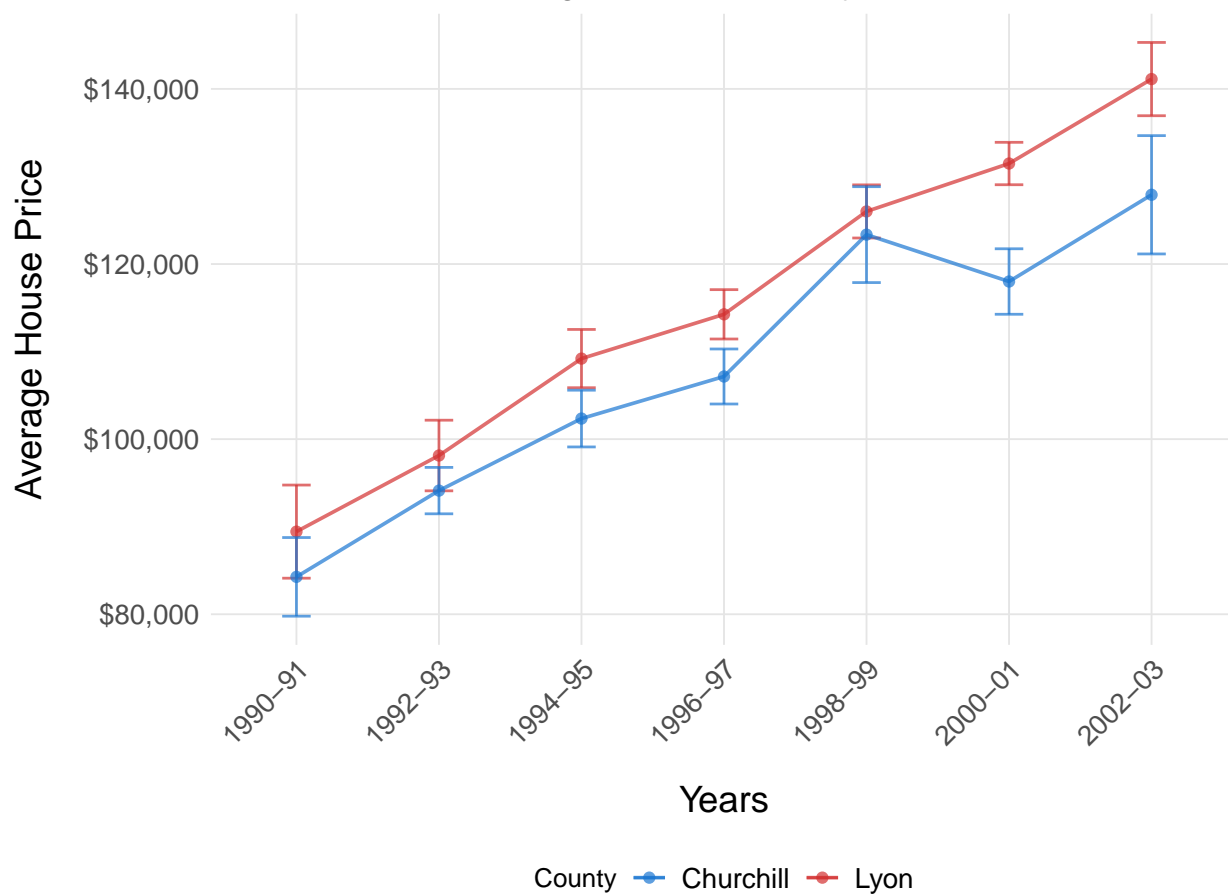
theme(
  axis.text.x = element_text(angle = 45, hjust = 1, size = 16),
  axis.text.y = element_text(size = 16),
  axis.title = element_text(size = 21),
  axis.title.x = element_text(size = 21, margin = margin(t = 20)),
  axis.title.y = element_text(size = 21, margin = margin(r = 20)),
  plot.title = element_text(size = 28, face = "bold", hjust = 0.5),
  plot.subtitle = element_text(size = 22, face = "italic", hjust = 0.5),
  plot.caption = element_text(size = 10, face = "italic", color = "gray"),
  panel.grid.major = element_line(color = "gray90", size = 0.5),
  panel.grid.minor = element_blank(),
  legend.key.size = unit(0.8, "cm"),
  legend.text = element_text(size = 16),
  legend.position = "bottom",
  legend.direction = "horizontal",
  legend.box = "horizontal",
  legend.box.spacing = unit(0.5, "cm")
) +
scale_color_manual(
  values = c("Churchill" = "#1976D2", "Lyon" = "#D32F2F"),
  name = "County",
  labels = c("Churchill", "Lyon")
) +
scale_y_continuous(labels = scales::label_dollar())

grid.arrange(p, ncol = 1, nrow = 1, widths = unit(1, "npc"), heights = unit(0.75, "npc"))

```

# Average House Prices Over Time by County

*Comparing Churchill and Lyon counties*



- 4.7 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.

$$\log \text{realsales}_{icb} = \sum_{b=-98/99}^7 \beta_b \text{Bin}_b \times \text{ChurchillCo}_c + \lambda_b + \gamma_c + u_{it}.$$

```
tempn$year_bin <- relevel(tempn$year_bin, ref = "1998-99")
es<- felm(log~factor(year_bin)+cc+factor(year_bin)*cc, data=tempn)

stargazer(es)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Nov 24, 2024 - 6:57:40 PM

Table 6:

|                            | <i>Dependent variable:</i>  |
|----------------------------|-----------------------------|
|                            | log                         |
| factor(year_bin)1990-91    | −0.376***<br>(0.026)        |
| factor(year_bin)1992-93    | −0.288***<br>(0.018)        |
| factor(year_bin)1994-95    | −0.164***<br>(0.018)        |
| factor(year_bin)1996-97    | −0.100***<br>(0.017)        |
| factor(year_bin)2000-01    | 0.050***<br>(0.015)         |
| factor(year_bin)2002-03    | 0.110***<br>(0.018)         |
| cc                         | −0.042**<br>(0.019)         |
| factor(year_bin)1990-91:cc | −0.016<br>(0.040)           |
| factor(year_bin)1992-93:cc | 0.047<br>(0.030)            |
| factor(year_bin)1994-95:cc | −0.012<br>(0.028)           |
| factor(year_bin)1996-97:cc | −0.034<br>(0.028)           |
| factor(year_bin)2000-01:cc | −0.085***<br>(0.027)        |
| factor(year_bin)2002-03:cc | −0.069**<br>(0.035)         |
| Constant                   | 11.681***<br>(0.011)        |
| Observations               | 9,263                       |
| R <sup>2</sup>             | 0.118                       |
| Adjusted R <sup>2</sup>    | 0.116                       |
| Residual Std. Error        | 0.382 (df = 9249)           |
| <i>Note:</i>               | *p<0.1; **p<0.05; ***p<0.01 |

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

```
res<-coef(summary(es))
res<-as.data.frame(res)

res<-tail(res,6)

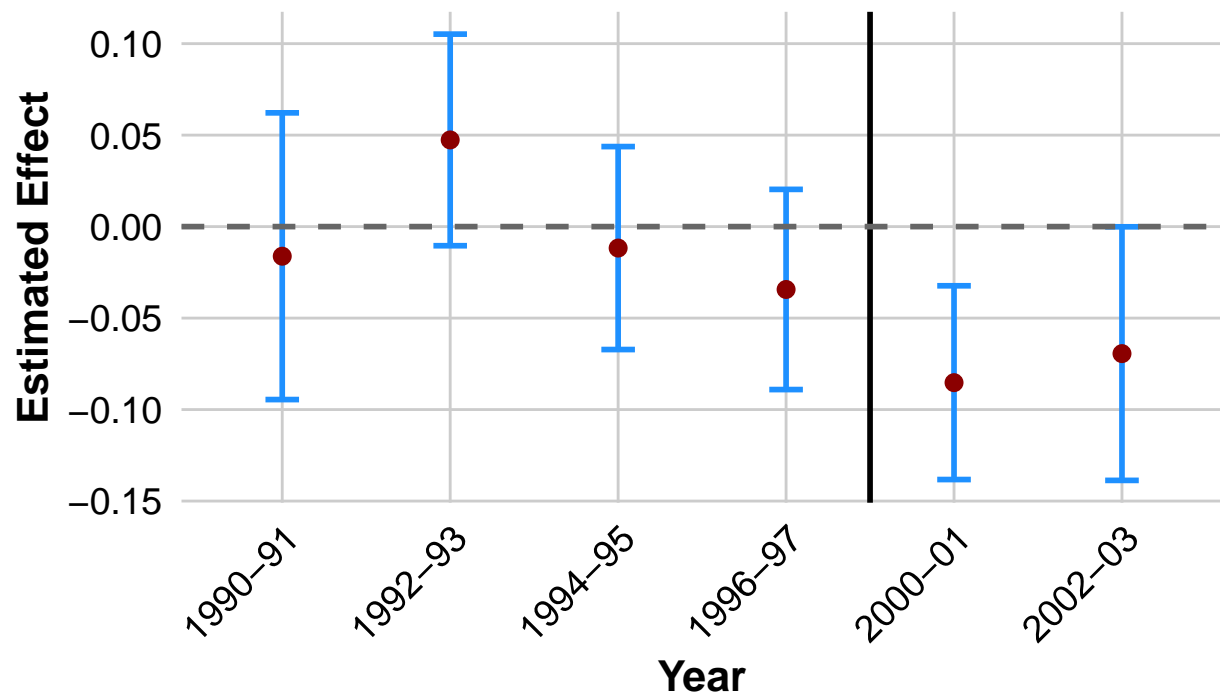
a<-c(0,0,0,0)
res<-rbind(res,a)

year<-c("1990-91", "1992-93", "1994-95", "1996-97", "2000-01", "2002-03")
res<-cbind(res,year)
res$ci<-1.96*res$`Std..Error`
names(res)<-c("Estimate","se", "t", "p","year", "ci")
#Use95%confidenceintervalinsteadofSEM

ggplot(res, aes(x = year, y = Estimate)) +
  geom_errorbar(aes(ymin = Estimate - ci, ymax = Estimate + ci), width = 0.2, color = "dodgerblue",
  size = 1) +
  geom_vline(xintercept = 4.5,color = "black", size = 1) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray40", size = 1) +
  geom_point(size = 3, color = "darkred", shape = 16) +
  labs(
    title = "Event Study Estimates by Year Bin",
    subtitle = "With 95% Confidence Intervals",
    x = "Year", # X-axis label
    y = "Estimated Effect" # Y-axis label
  ) +
  theme_minimal(base_size = 16) +
  theme(
    axis.text.x = element_text(size = 14, color = "black", angle = 45, hjust = 1),
    axis.text.y = element_text(size = 14, color = "black"),
    axis.title = element_text(size = 16, face = "bold"),
    plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 16, face = "italic", hjust = 0.5),
    plot.caption = element_text(size = 12, face = "italic", color = "gray"),
    panel.grid.major = element_line(color = "gray80", size = 0.5),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  )
)
```

# Event Study Estimates by Year Bin

*With 95% Confidence Intervals*



#### 4.9 What patterns are we looking for in the two graphs just produced?

**Answer:**

We are looking to see if after the cancer cluster appears (post 1999) if the effect estimate on house prices in Churchill county decreases (significantly) in the second graph. In the first graph we look to see if this pattern of decreasing housing prices is also present in the control county, checking to see if the pattern only occurs in the treatment and not control, so we attribute the decrease in housing prices to the cancer cluster.