Causal Analysis: Temperature and Walmart Sales

Brenda Estefania Hernandez Barragan

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.2.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(fixest)
library(lubridate)
library(ggplot2)
df <- read.csv("Walmart_Sales.csv")
```

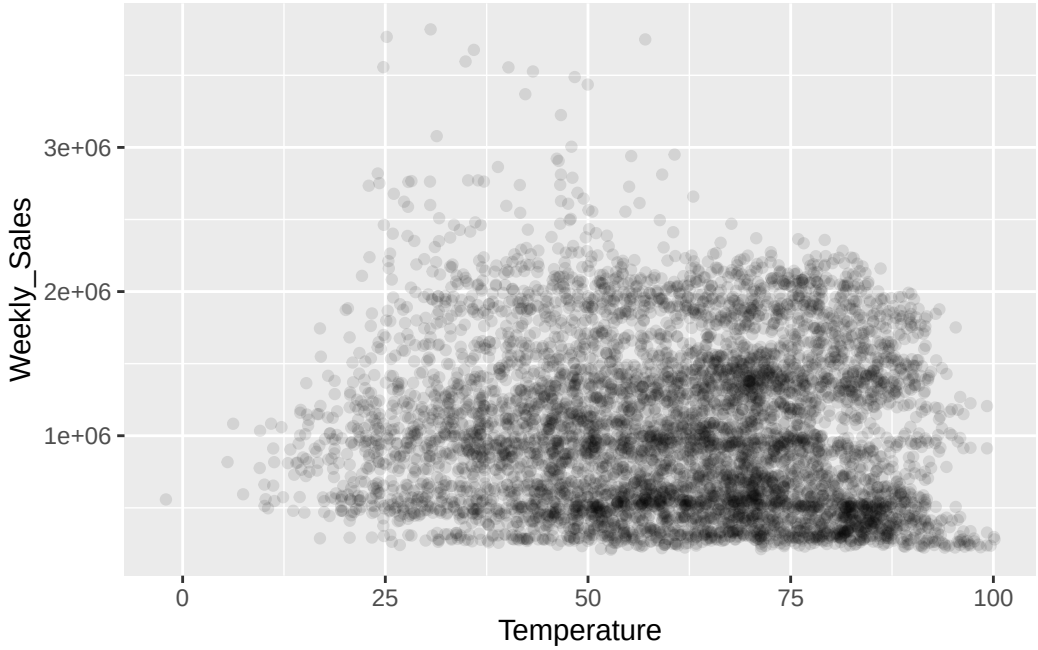
1. Research question

Does higher temperatures casually affect sales?

Answer is: Yes, but not in a linear way. On average, for every additional 1°F, sales are associated with an increase of 0.37% until 68.51 °F. Beyond this point, the same effect becomes negative.

2. Justification

```
ggplot(df, aes(x = Temperature, y = Weekly_Sales)) +
  geom_point(alpha = 0.1)
```



Looking at the raw data, I identified 2 main issues. The first one is the distribution of sales being skewed with a few values that are really large (heteroskedasticity). To fix this issue, I used of the **logarithmic form of the Sales** variable. Another reason to use the logarithmic term is that interpretation in percentage is easier and more meaningful. The second issue I identified is that there wasn't a clear linear relationship. Sales seem to reach a plateau or decrease, to fix this issue I decided to use the **quadratic term for the variable Temperature.**

Also, considering all the variables in the data set, the inclusion decision to the regression equation rational is the following.

- Included
  - **Week:** frames the week of the year. This also allows us to see the variation of a week that has a holiday versus a normal week. For example, we might see that the week of Black Friday behaves differently than the week prior.
  - **Temperature:** this is our main question.

  - **Holiday:** if this is not controlled, we might make mistaken assumptions about weather and sales. For example, we might conclude mistakenly that winter drives more sales when in reality, there are several holidays that are the ones driving the higher sales.
  - **Store:** some stores might have a better location and more traffic, enabling higher sales.

- Excluded: I am considering Fuel price, CPI and unemployment as unrelated because they might have an effect in sales but not in temperature. Not including them won't cause a bias for omitted variable.

3. Methodology

Additionally, I realized that is a data panel set, all incidences are observations that belong to the set of 45 stores. If we don't **adjust for Standard Errors** and do clustering, the model will read every week as different new instance. This will make the model "inflate" artificially the size of the sample and will yield artificially optimistic confidence levels. **Clustering by Store** will help the model correlate observations to a specific store, ensuring robust Standard Errors and realistic p-values.

Moreover, I also acknowledge that there might be some innate effects of the store and the time that might affect sales. If not controlled, we will create a backdoor path. To prevent this, **I used Store Fixed Effects that will washout the effects that are constant in the store l**ike the size of the store and its location, while Time Fixed Effects isolate global aspects like inflation or global warming.

4. Results

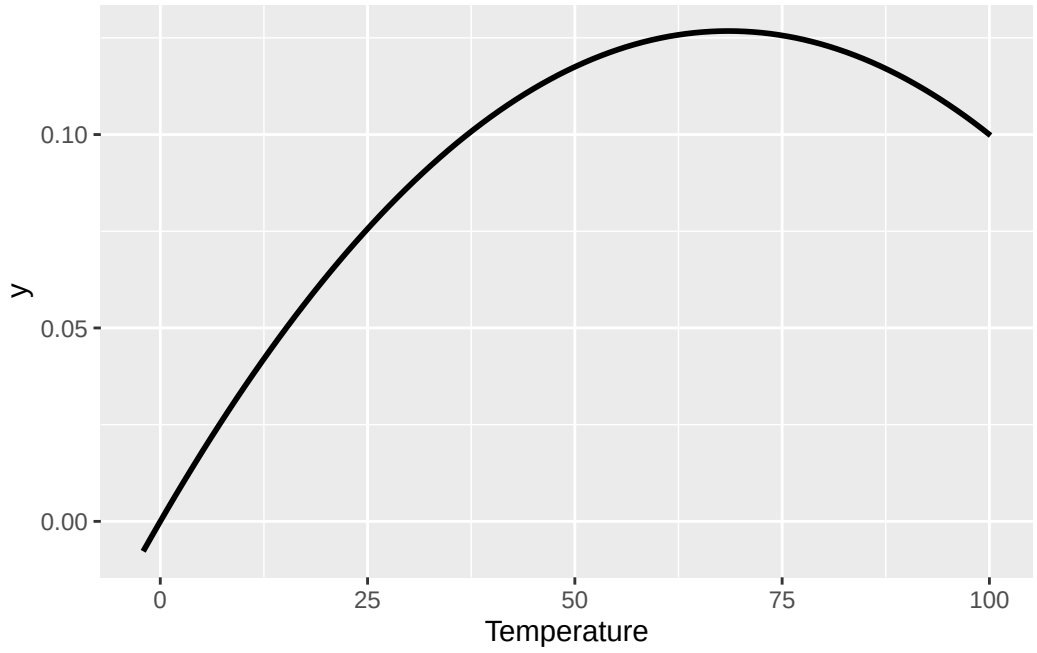
```
df$Date <- as.Date(df$Date, format = "%d-%m-%Y")
df$Month <- as.factor(format(df$Date, "%m"))
df$Year <- as.factor(format(df$Date, "%Y"))
df$log_sales <- log(df$Weekly_Sales)
df$temp_sq <- df$Temperature^2

modelo <- feols(log_sales ~ Temperature + temp_sq + Holiday_Flag | Store + Month + Year, data
etable(modelo)
```

```
                           modelo
Dependent Var.:         log_sales
```

```
Temperature       0.0037*** (0.0008)
temp_sq         -2.7e-5*** (7.18e-6)
Holiday_Flag      0.0321*** (0.0055)
Fixed-Effects:  --------------------
Store                            Yes
Month                            Yes
Year                             Yes
--------------- --------------------
S.E.: Clustered             by: Store
Observations                    6,435
R2                            0.96725
Within R2                     0.01267
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
b1 <- 0.0037
b2 <- -0.000027
ggplot(df, aes(x = Temperature)) +
  stat_function(fun = function(x) b1*x + b2*x^2, size = 1)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



These results show that for every additional 1ºF increase in temperature, sales are associated with an average increase of 0.37%. However, this growth follows a inverted-U shape and is sustained until the temperature reaches 68.51ºF. Beyond this temperature, the effect becomes negative and it leads to a decline in sales as temperatures rise. To get this peak point, I used the derivative of my regression equation (Temp = - 1/(2 2)). These results are significant at 0.1% level.

The Within R2 is one of the most valuable metrics for our research, it tells us that the extra sales driven by the temperature are of 1.26%. In this case, R2 is not as useful for our question since this metric can be artificially inflated by making comparisons of innate differences in stores like size and location.

Additionally, holding everything constant, the presence of a holiday on a given week is associated with an increase of about 3.2% in sales.