Data Exploration Project

Course Code: ECON 5300

Introduction

In this project, the Walmart real world dataset has been analyzed to understand since average temperatures have been rising over time if this rising in temperature is impacting their sales. The dataset comprised 45 Walmart stores from February 2010 to October 2012. In short, their research question is to know whether higher temperatures casually affect Walmart sales.

Data Preparation

In Walmart sales dataset comprises of 8 number of features store number that indicates the which treated as a factor for fixed effects using factor(store), date (the week of the sale in day-month-year) which is later parsed using lubridate() library and extract month of year to understand the seasonal pattern, Weekly sales for the given store in the given week, while holiday flag has data if week is a special holiday week or not. Temperature which is the average temperature in the region (in Fahrenheit) this feature is precomputed by taking square of the original column, Fuel Price is the cost of fuel in the region (in dollars per gallon) and CPI is the consumer price index, finally, Unemployment is the unemployment rate in the region.

Research Analysis

Before performing regression analysis, it is important to think through what could create a spurious correlation between temperature and sales. The two primary confounders are seasonality and store-level permanent differences. December brings low temperatures and high sales; summer brings high temperatures, and moderate sales season drives both variables simultaneously. Without controlling this, temperature would appear negatively related to sales simply because holidays happen in winter. Similarly, some stores experience high temperatures year-round and may differ in size. Customer base, or regional wealth compared across stores would conflate climate with store characteristics. I avoid controlling variables caused by temperature, such as product mix, because these are mediators controlling them that would block the very causal effect I want to measure.

By combining store fixed effects and month-of-year fixed effects, I compare each store to itself across weeks that are unusually warm or cold for that calendar month. Any remaining effect on sales is the causal effect of temperature, purged of seasonal and store-level confounding.

Regression Variable Flow Chart

The flow chart is a decision tree that helps us figure out what role each variable plays in the regression model. The goal of this flow chart is not just including every variable in the regression model but to include exactly the right variable so that regression coefficients on the treatment variable provide hones, clean interpretations.

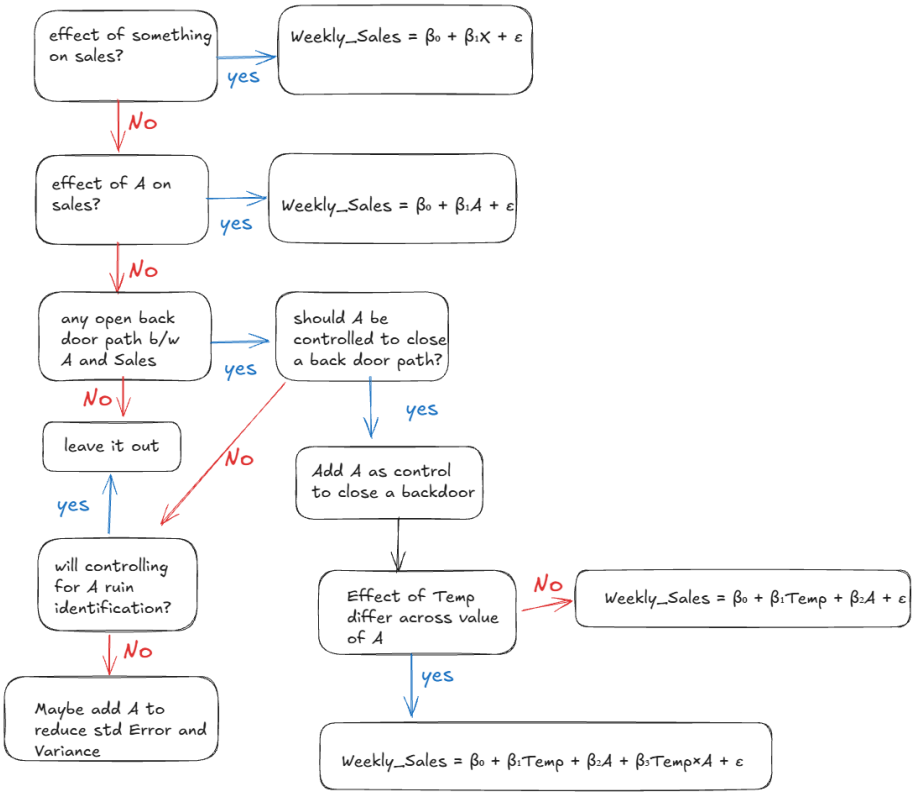

Figure01. Flowchart for Regression Variables

Graphical Analysis/Exploration

Is Relationship Linear?

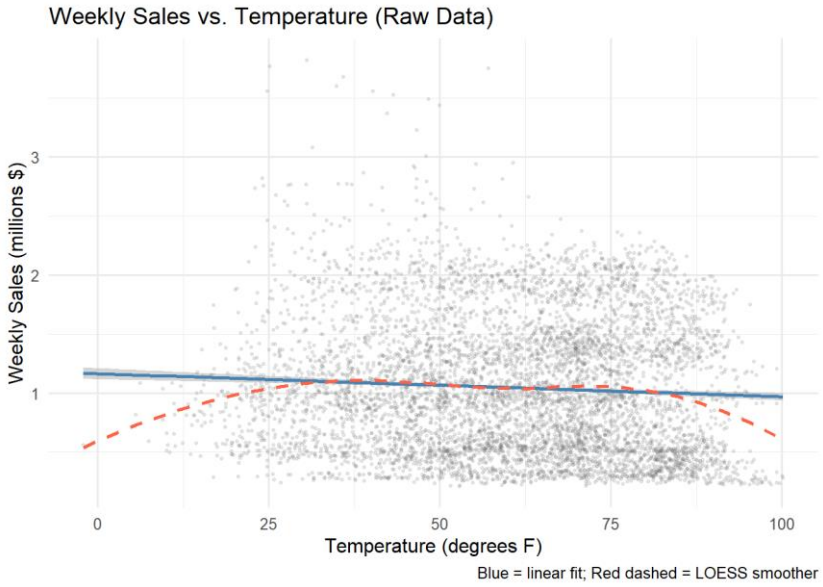

Fig02. Raw scatter of weekly sales vs. temperature across all stores and weeks. The LOESS smoother (red dashed) reveals an inverted-U shape, motivating a quadratic temperature term.

The red dashed LOESS curve shows a clear, inverted U-shape. Sales seem to be highest when temperatures are moderate and lower when it's either very cold or very hot. That feels intuitive, since people may avoid going out in extreme weather. If we only used a linear model, we would miss this curved pattern. The blue linear fit is almost flat with a slight negative slope, which suggests that a simple straight-line relationship doesn't tell the full story. This graph makes a strong case for adding a squared temperature term to better capture the nonlinear effect.

Seasonal Sales Movement

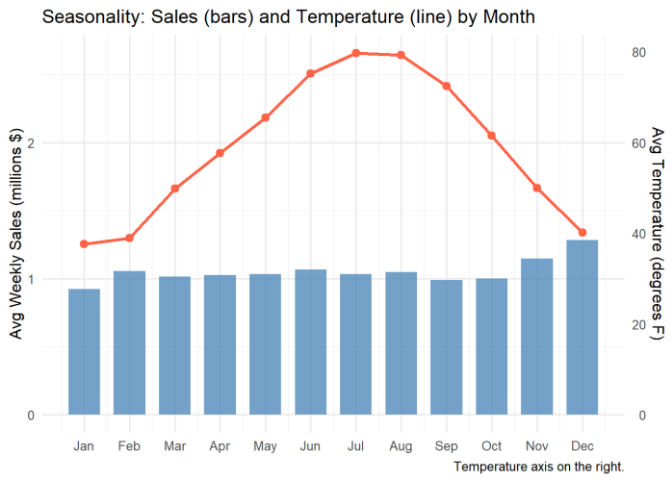

Fig03. Average weekly sales (bars) and average temperature (line) by calendar month. Both follow strong seasonal patterns – confirming that month fixed effects are necessary to avoid spurious correlation.

This graph makes the confounding issue much clearer. December has the highest sales, which is likely driven by holiday shopping, but it also has relatively low temperatures. On the other hand, the summer months have the highest temperatures, yet sales during those months are average. If we ran a regression without controlling for month effects, temperature might appear negatively related to sales. But that would be misleading. What we would really be capturing is the holiday spike in December, not a true causal effect of temperature. This is exactly why including month fixed effects is important; it helps separate seasonal demand patterns from the actual impact of weather.

Heteroskedasticity Check (non-constant variance in regression errors)

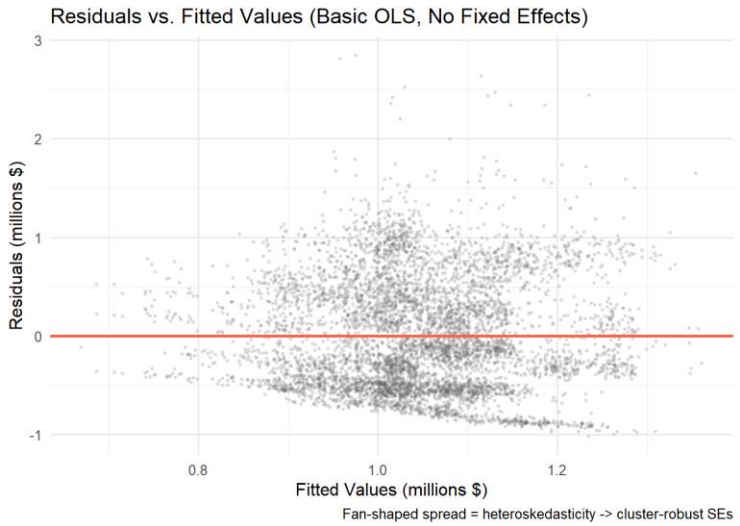

Fig 04. Residuals vs. fitted values from a basic OLS model. The fan-shaped spread indicates heteroskedasticity, motivating cluster-robust standard errors.

The residual plot shows a clear fan shape as the fitted values increase, which suggests heteroskedasticity. In other words, stores with higher predicted sales tend to have more variability in their residuals. On top of that, the data include repeated observations for the same stores over time, so those observations are not independent. Since each store appears many times, errors within a store are likely correlated. Because of both the changing variance and the within-store correlation, I use standard errors clustered at the store level. This approach adjusts for heteroskedasticity and for serial correlation within stores over time.

Regression Analysis

Effect of Temperature on Weekly Walmart Sales

	(1) Linear Temp	(2) Quadratic Temp
Temperature (F)	695.600*	3413.487**
	(317.868)	(1191.867)
Holiday Week	32288.759***	34530.091***
	(5577.607)	(5935.226)
Fuel Price ($/gal)	-25336.711+	-24477.898+
	(12909.272)	(12779.186)
Consumer Price Index	907.019	542.021
	(2390.606)	(2351.879)
Unemployment Rate	-31738.884**	-32958.879**
	(11427.153)	(11242.304)
Temperature-squared (F sq)		-26.672*
		(11.148)
Num.Obs.	6435	6435
R2	0.937	0.938
R2 Adj.	0.937	0.937

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Store and month-of-year fixed effects in all models. SEs clustered by store.

Table01. Regression Estimates of the Effect of Temperature on Weekly Walmart Sales

This table reports the results of your regression on the effect of temperature on weekly Walmart sales, comparing a simple linear model with a quadratic (nonlinear) one. The first column assumes temperature affects sales in a straight line. The second column allows the relationship to curve, which better matches what we saw in the graphs. In the quadratic model, the coefficient on temperature is positive and statistically significant. This means that when temperatures are relatively low, a 1°F increase is associated with about $3,400 more in weekly sales. However, the temperature-squared term is negative and significant, which confirms the inverted U-shape. As temperatures keep rising, the positive effect weakens and eventually turns negative. So, warming helps sales when it's cold, but very high temperatures start to reduce sales. That's exactly why the quadratic term was necessary. Holiday weeks have the largest effect on the model, increasing sales by roughly $34,500 on average. That makes an intuitive sense, since major shopping periods like Thanksgiving and Christmas drive much higher demand.

Fuel prices have a negative and marginally significant effect, suggesting that higher gas prices may slightly reduce sales, possibly because consumers travel less. The unemployment rate has a strong and statistically significant negative effect, meaning higher unemployment is associated with lower

sales. The Consumer Price Index is not statistically significant, so it doesn't appear to add much explanatory power to this setting.

The R^2 is 0.938, which means the model explains about 94% of the variation in weekly sales. With 6,435 observations and fixed effects for store and month included, this is a very strong fit.

Overall, the results show that temperature has a meaningful and statistically significant nonlinear effect on Walmart sales. Moderate warming boosts sales when it's cold, but extreme heat eventually reduces them.

Marginal Effects plot

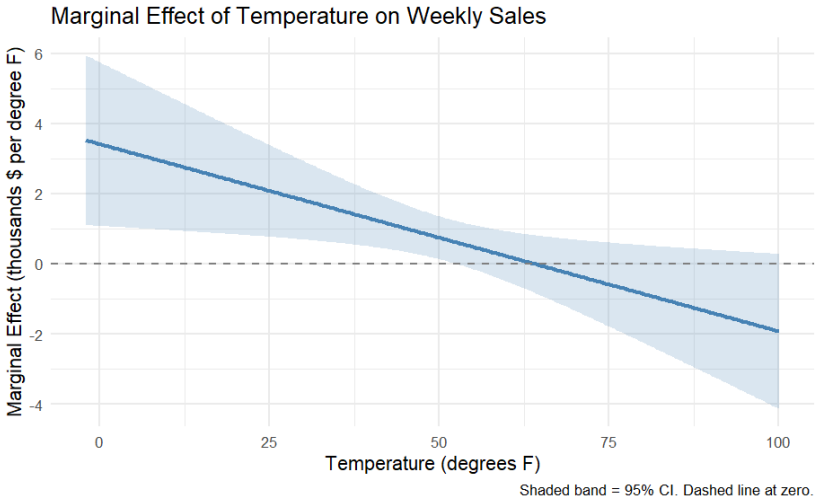

Fig05. Estimated marginal effect of a 1-degree-F temperature increase on weekly sales, as a function of the current temperature. The effect transitions from positive (warming helps) to negative (warming hurts) around the turning point.

This graph shows how the effect of temperature on sales changes depending on how warm it already is. When it's cold, a small increase in temperature helps sales. For example, around 20°F, each additional degree is linked to about $2,300 more in weekly sales. Even at 40°F, the effect is still positive, though smaller. So, warming up from very cold weather seems to encourage more shopping. As temperatures move into the mid-60s, the benefits fade. Around 64°F, the marginal effect is basically zero. That's the point where sales are at their peak, warming further doesn't help anymore.

Once temperatures get higher than that, the effect turns negative. At 75°F and above, each additional degree reduces sales, and by 100°F the drop becomes noticeable. In other words, extreme heat seems to discourage shopping. The shaded area shows uncertainty, but the overall pattern is clear. Sales increase as it warms up from cold temperatures, level off around the mid-60s, and then decline when it becomes too hot.

```
Sales-maximizing temperature: 64.0 degrees F

Marginal effects at select temperatures:
   20.0 F:  +2.3 thousand dollars per degree F
   40.0 F:  +1.3 thousand dollars per degree F
   64.0 F:  +0.0 thousand dollars per degree F
   75.0 F:  -0.6 thousand dollars per degree F
   90.0 F:  -1.4 thousand dollars per degree F
  100.0 F:  -1.9 thousand dollars per degree F
```

Table.02 Estimated Marginal Effects of Temperature on Weekly Sales

This table shows how the effect of temperature on sales changes depending on how warm it already is. Sales are highest at around 64°F, which is the sales-maximizing temperature. At that point, increasing temperatures further doesn't increase sales. When it's cold, warming helps. For example, at 20°F, each extra degree is linked to about $2,300 more in weekly sales, and at 40°F the effect is still positive, though smaller. Once temperatures rise above the mid-60s, the effect turns negative. At 75°F and above, each additional degree starts to reduce sales, and by 100°F the drop is noticeable. In short, moderate weather boosts sales, but extreme heat begins to hurt them.

Conclusion

The relationship between temperature and sales is clearly nonlinear. The initial graph, the quadratic regression, and the marginal effects plot all tell the same story. When its very cold, warmer temperatures tend to increase sales. But once temperatures rise past the mid-60s, further heat begins to reduce sales. In other words, moderate weather helps, while extreme heat hurts. Importantly, this pattern remains even after controlling fixed effects, store fixed effects, holidays, fuel prices, inflation, and unemployment.

From a business perspective, this matters. As average temperatures rise, more weeks may fall into the "too hot" range where sales begin to decline. The effect per degree is not huge in any single week, but across 45 stores and many weeks, even modest temperature impacts translate into meaningful revenue differences. Stores in already warm climates are likely the most exposed. Walmart could respond by investing in better in-store cooling, running targeted summer promotions, or leaning more heavily on online ordering during heat waves to reduce the impact of extreme heat. The causal interpretation is fairly strong because of how the effect is identified. We are not comparing hot states to cold states. Instead, we compare each store to itself within the same calendar month. For example, we ask whether Store 12 sells more in a January week that is warmer than usual compared to a January week that is colder than usual. That within-store, within-month comparison isolates unusual temperature variation and makes the result much more convincing than a simple cross-sectional correlation.