Causal Exploration of Retail Volume Drivers

Overview

This document presents an initial causal analysis of macroeconomic indicators and their influence on retail volume. A binary simplification approach has been used instead of immediately deploying multivariate Bayesian modeling, allowing for more interpretable comparisons across discretised predictor-outcome relationships.

This structure aids in early DAG formation and enables examination of questions such as:

- Whether retail volume is influenced by prior values of itself, price, or income
- Whether sentiment or price can be predicted by upstream drivers
- Which confounders must be adjusted for in causal modeling

The aim is to determine foundational causal pathways prior to modeling interventions, pricing strategy, or sentiment shocks.

Step 1: Lag Generation

First- and second-order lags have been computed for relevant predictors to allow for temporal comparison.

```
combined_q_lags <- combined_q_trimmed %>%
  group_by(geo) %>%
  arrange(time) %>%
  mutate(
    volume_lag1       = lag(volume_index, 1),
    volume_lag2       = lag(volume_index, 2),
    price_index_lag1  = lag(price_index, 1),
    sentiment_lag1    = lag(sentiment, 1),
    sentiment_lag2    = lag(sentiment, 2),
    unemployment_lag1 = lag(unemployment_rate, 1),
    unemployment_lag2 = lag(unemployment_rate, 2),
    income_lag1       = lag(household_income, 1),
    income_lag2       = lag(household_income, 2)
  ) %>%
  ungroup()
```

Step 2: Variable Binarisation

Each numeric variable has been binarised relative to its median value. This facilitates stratified conditional probability analysis.

```r
binned_df <- combined_q_trimmed %>%
  mutate(across(where(is.numeric), ~ ifelse(. > median(., na.rm = TRUE), 1, 0))) %>%
  mutate(
    sentiment_lag1      = lag(sentiment, 1),
    sentiment_lag2      = lag(sentiment, 2),
    income_lag1         = lag(household_income, 1),
    income_lag2         = lag(household_income, 2),
    unemployment_lag1   = lag(unemployment_rate, 1),
    unemployment_lag2   = lag(unemployment_rate, 2),
    price_index_lag1    = lag(price_index, 1),
    volume_index_lag1   = lag(volume_index, 1),
    volume_index_lag2   = lag(volume_index, 2)
  )
```

Step 3: Conditional Probability Comparison

Conditional probabilities have been estimated for each predictor relative to binary sales volume.

```r
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
library(dplyr)

results_tbl <- tibble::tibble(
  Hypothesis = c(
    "$P(V = 1 \\mid S_t = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid S_{t-1} = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid I_t = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid I_{t-1} = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid U_t = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid U_{t-1} = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid P_t = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid P_{t-1} = 1) \\ne P(V = 1)$",
    "$P(V = 1 \\mid V_{t-1} = 1) \\ne P(V = 1)$"
  ),
  `P(V = 1 | X = 1)` = c(
    mean(binned_df$volume_index[binned_df$sentiment == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$sentiment_lag1 == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$household_income == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$income_lag1 == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$unemployment_rate == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$unemployment_lag1 == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$price_index == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$price_index_lag1 == 1], na.rm = TRUE),
    mean(binned_df$volume_index[binned_df$volume_index_lag1 == 1], na.rm = TRUE)
```

```
  ) %>% round(3)
)

kbl(results_tbl, booktabs = TRUE, escape = FALSE,
    col.names = c("Hypothesis", "P(V = 1 | X = 1)")) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  column_spec(2, width = "3cm")
```

Hypothesis	P(V = 1 | X = 1)
$P(V = 1 \mid S_t = 1) \neq P(V = 1)$	0.591
$P(V = 1 \mid S_{t-1} = 1) \neq P(V = 1)$	0.593
$P(V = 1 \mid I_t = 1) \neq P(V = 1)$	0.501
$P(V = 1 \mid I_{t-1} = 1) \neq P(V = 1)$	0.501
$P(V = 1 \mid U_t = 1) \neq P(V = 1)$	0.443
$P(V = 1 \mid U_{t-1} = 1) \neq P(V = 1)$	0.444
$P(V = 1 \mid P_t = 1) \neq P(V = 1)$	0.512
$P(V = 1 \mid P_{t-1} = 1) \neq P(V = 1)$	0.513
$P(V = 1 \mid V_{t-1} = 1) \neq P(V = 1)$	0.985

Step 4: Upstream Prediction of Sentiment and Price

The ability of macroeconomic indicators to predict current sentiment or pricing has been assessed.

```
results_tbl <- tibble::tibble(
  Hypothesis = c(
    "$P(S = 1 \\mid U_{t-1} = 1) \\ne P(V = 1)$",
    "$P(S = 1 \\mid I_{t-1} = 1) \\ne P(V = 1)$",
    "$P(P = 1 \\mid U_{t-1} = 1) \\ne P(V = 1)$",
    "$P(P = 1 \\mid I_{t-1} = 1) \\ne P(V = 1)$",
    "$P(P = 1 \\mid S_{t-1} = 1) \\ne P(V = 1)$",
    "$P(P = 1 \\mid V_{t-1} = 1) \\ne P(V = 1)$"
  ),
  `P(V = 1 | X = 1)` = c(
    mean(binned_df$sentiment[binned_df$unemployment_lag1 == 1], na.rm = TRUE),
    mean(binned_df$sentiment[binned_df$income_lag1 == 1], na.rm = TRUE),
    mean(binned_df$price_index[binned_df$unemployment_lag1 == 1], na.rm = TRUE),
    mean(binned_df$price_index[binned_df$income_lag1 == 1], na.rm = TRUE),
    mean(binned_df$price_index[binned_df$sentiment == 1], na.rm = TRUE),
    mean(binned_df$price_index[binned_df$volume_index_lag1 == 1], na.rm = TRUE)
  ) %>% round(3)
)

kbl(results_tbl, booktabs = TRUE, escape = FALSE,
    col.names = c("Hypothesis", "P(V = 1 | X = 1)")) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  column_spec(2, width = "3cm")
```

Hypothesis	P(V = 1 | X = 1)
$P(S = 1 \mid U_{t-1} = 1) \neq P(V = 1)$	0.319
$P(S = 1 \mid I_{t-1} = 1) \neq P(V = 1)$	0.567
$P(P = 1 \mid U_{t-1} = 1) \neq P(V = 1)$	0.472
$P(P = 1 \mid I_{t-1} = 1) \neq P(V = 1)$	0.552
$P(P = 1 \mid S_{t-1} = 1) \neq P(V = 1)$	0.484
$P(P = 1 \mid V_{t-1} = 1) \neq P(V = 1)$	0.508

Step 5: Confounding Checks

The above results indicate that not only does the set of variables household income, sentiment, or unemployment predict volume of sales, but they also predict price. This makes them clear candidates for counfounders of price and a possible explanation for me counter intuitive behaviour of price. To assess whether the observed relationship between higher prices (price_index) and higher sales volume (volume_index) may be confounded by other variables—specifically, household income, sentiment, or unemployment.

A key concern is that factors such as sales volume at $t-1$ through innertia and unemployment may influence both prices and sales together either contemporaneously or through lagged effects—and, in turn, impact sales volume. This introduces a backdoor path from sales volume at $t-1$ or unemployment to price to volume, potentially biasing our estimates of the causal effect of price.

As a starting point, we will use the benchmark of total effect of price on sales volume by measuring, suggesting that price being high increases the probability of sales volume being high (again, counter intuitive) by about 12% points.

$$P(V = 1 \mid P = 1)) - P(V = 1 \mid P = 0))$$

```
mean(binned_df$volume_index[binned_df$price_index == 1], na.rm = TRUE) - mean(binned_df$volume_index[bi
```

```
## [1] 0.02422261
```

To estimate the causal impact of price on volume while accounting for such confounding, we compute the Average Causal Effect (ACE) using the adjustment formula:

$$ACE = P(V = 1 \mid do(P = 1)) - P(V = 1 \mid do(P = 0))$$

Here, P is the variable we intervene on (price), and confounders such as sales volume at $t-1$ sentiment, income, or unemployment are denoted as Z. To block backdoor paths from Z to P, we apply stratification or adjustment:

$$P(V = 1 \mid do(P = 1)) = \sum_z P(V = 1 \mid P = 1, Z = z) \cdot P(Z = z)$$

This formula adjusts the observed relationship between price and volume by conditioning on the confounding variable Z, effectively simulating an intervention on price while holding Z constant.

```
# --- Compute base difference ---
base_line <- mean(binned_df$volume_index[binned_df$price_index == 1], na.rm = TRUE) -
             mean(binned_df$volume_index[binned_df$price_index == 0], na.rm = TRUE)

# --- ACE: Price effect conditioned on household income ---
```

```r
high_price_high_income_income <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$hous
high_price_low_income_income  <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$hous
low_price_high_income_income  <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$hous
low_price_low_income_income   <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$hous

ace_price_on_sales_income <- (high_price_high_income_income + high_price_low_income_income) -
                             (low_price_high_income_income + low_price_low_income_income)

# --- ACE: Price effect conditioned on unemployment ---
high_price_high_unemp <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$unemployment
high_price_low_unemp  <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$unemployment
low_price_high_unemp  <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$unemployment
low_price_low_unemp   <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$unemployment

ace_price_on_sales_unemp <- (high_price_high_unemp + high_price_low_unemp) -
                             (low_price_high_unemp + low_price_low_unemp)

# --- ACE: Price effect conditioned on lagged volume ---
high_price_high_lag <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$volume_index_l
high_price_low_lag  <- mean(binned_df$volume_index[binned_df$price_index == 1 & binned_df$volume_index_l
low_price_high_lag  <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$volume_index_l
low_price_low_lag   <- mean(binned_df$volume_index[binned_df$price_index == 0 & binned_df$volume_index_l

ace_price_on_sales_lag <- (high_price_high_lag + high_price_low_lag) -
                          (low_price_high_lag + low_price_low_lag)

# --- Build Results Table ---
results_tbl <- tibble::tibble(
  Hypothesis = c(
    "$P(V = 1 \\mid do(P), I_t)$",
    "$P(V = 1 \\mid do(P), U_t)$",
    "$P(V = 1 \\mid do(P), V_{t-1})$"
  ),
  `P(V = 1 | X = 1)` = c(
    ace_price_on_sales_income,
    ace_price_on_sales_unemp,
    ace_price_on_sales_lag
  ) %>% round(3),
  Delta = c(
     ace_price_on_sales_income - base_line,
     ace_price_on_sales_unemp  - base_line,
     ace_price_on_sales_lag    - base_line
  ) %>% round(3)
  )

# --- Render Table ---
kbl(results_tbl, booktabs = TRUE, escape = FALSE,
    col.names = c("Hypothesis", "P(V = 1 | X = 1)", "Delta")) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  column_spec(2, width = "3cm")
```

Hypothesis	P(V = 1 | X = 1)	Delta
$P(V = 1 \mid do(P), I_t)$	0.024	0.000
$P(V = 1 \mid do(P), U_t)$	0.017	-0.007
$P(V = 1 \mid do(P), V_{t-1})$	0.008	-0.016