MSIN0094 Third Assignment

Due Friday 10am, Dec 13, 2024

Candidate number: TBDH9

Word count: 2050

1. Descriptive Analytics (20 pts)

Q1 From `data_full`, generate a new variable, `final_price`, which is the actual retail price for each week (i.e., Recommended Retail Price after discounts). (**8pts**)

- Write your code below to generate final_price from RRP and discount. (**2pts**)

```
# write your codes below

data_full <- data_full %>%
  mutate(final_price = RRP * (1 - discount))

glimpse(data_full)
```

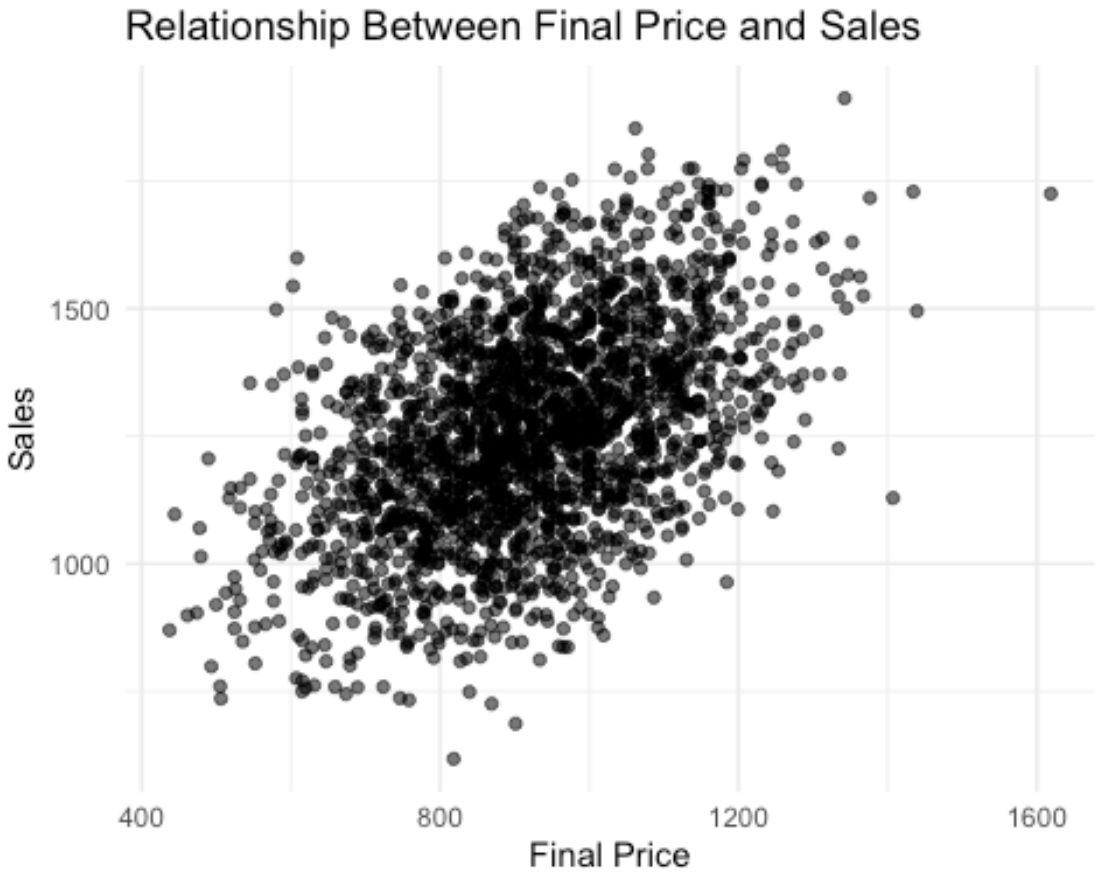
```
Rows: 2,080
Columns: 13
$ product_id        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
1…
$ brand             <chr> "Samsung", "Samsung", "Samsung", "Samsung",
"Samsung…
$ technology        <chr> "OLED", "OLED", "OLED", "QLED", "OLED", "QLED",
"QLE…
$ resolution        <chr> "1080p", "4k", "4k", "1080p", "4k", "4k", "4k",
"4k"…
$ support_HDR       <int> 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0,
1…
$ screensize        <chr> "50-59", "30-39", "30-39", "60+", "50-59", "40-49",
…
$ week_id           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1…
$ sales             <int> 1745, 1444, 1183, 1474, 1647, 1461, 1626, 1530,
1706…
$ RRP               <int> 1399, 1099, 1049, 949, 1299, 999, 1249, 1349, 1349,
…
$ discount          <dbl> 0.12, 0.12, 0.18, 0.13, 0.17, 0.12, 0.07, 0.15,
0.14…
$ marketing_expense <dbl> 3794.739, 3794.739, 3794.739, 3794.739, 3794.739,
37…
$ cost_shifter      <dbl> 618.6324, 571.6035, 644.6133, 593.3050, 646.6439,
60…
```

```
$ final_price       <dbl> 1231.12, 967.12, 860.18, 825.63, 1078.17, 879.12,
11…
```

- Visualise using scatter plot the relationship between final price and sales. Tips: you can use `ggplot2` and `geom_point` to create the scatter plot. Write your code below to create the scatter plot. (**2pts**)

```r
# write your codes below
library(ggplot2)

ggplot(data_full, aes(x = final_price, y = sales)) +
  geom_point(alpha = 0.6) +
  theme_minimal() +
  labs(
    x = "Final Price",
    y = "Sales",
    title = "Relationship Between Final Price and Sales"
  )
```



- Do you observe a positive or negative relationship between final price and sales? Is this relationship causal? Why or why not? (**4pts**, 150 words)

There is a positive correlation between final price and sales in the scatter plot, but this association arises in observational data rather than a randomised setting. This pattern reflects product-mix heterogeneity and omitted-variable bias: higher-priced models differ on key features such as having better gaming specs, HDR support or larger screens, which are confounders jointly affecting both price and sales. Therefore, such positive correlation does not imply a causal price elasticity.

The relationship is not causal: the dataset is observational, non-experimental, with no random assignment of price, and mixes different products and sizes. It violates the independence condition required for causal inference, and cannot satisfy the potential outcome framework. Therefore, there is no proof higher price leads to bigger sales.

> ***Q2.*** Use `dplyr` to compute the average weekly dollar sales (final price * unit sales) for each brand across all weeks (i.e., the result should be 1 average per brand). Rank the brands from the highest average dollar sales to the lowest average dollar sales. (**6pts**) Which brand has the highest average weekly dollar sales? (**2pts**).

```
# write you code below
library(dplyr)

data_sales_by_brand <- data_full %>%
  mutate(dollar_sales = final_price * sales) %>%
  group_by(brand) %>%
  summarise(avg_weekly_dollar_sales = mean(dollar_sales, na.rm = TRUE)) %>%
  arrange(desc(avg_weekly_dollar_sales))

glimpse(data_full)
```

```
Rows: 2,080
Columns: 13
$ product_id        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
1…
$ brand             <chr> "Samsung", "Samsung", "Samsung", "Samsung",
"Samsung…
$ technology        <chr> "OLED", "OLED", "OLED", "QLED", "OLED", "QLED",
"QLE…
$ resolution        <chr> "1080p", "4k", "4k", "1080p", "4k", "4k", "4k",
"4k"…
$ support_HDR       <int> 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0,
1…
$ screensize        <chr> "50-59", "30-39", "30-39", "60+", "50-59", "40-49",
…
$ week_id           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1…
$ sales             <int> 1745, 1444, 1183, 1474, 1647, 1461, 1626, 1530,
1706…
$ RRP               <int> 1399, 1099, 1049, 949, 1299, 999, 1249, 1349, 1349,
…
```

```
$ discount          <dbl> 0.12, 0.12, 0.18, 0.13, 0.17, 0.12, 0.07, 0.15,
0.14…
$ marketing_expense <dbl> 3794.739, 3794.739, 3794.739, 3794.739, 3794.739,
37…
$ cost_shifter      <dbl> 618.6324, 571.6035, 644.6133, 593.3050, 646.6439,
60…
$ final_price       <dbl> 1231.12, 967.12, 860.18, 825.63, 1078.17, 879.12,
11…
```

```
# please do not modify.
# print out the ranking of brands based on average weekly dollar sales
data_sales_by_brand
```

```
# A tibble: 4 × 2
  brand   avg_weekly_dollar_sales
  <chr>                     <dbl>
1 Samsung                1277231.
2 Sony                   1226644.
3 LG                     1119492.
4 Philips                1086011.
```

The highest weekly dollar sales Samsung has, with $1 277 231 of sales.

> ***Q3.*** In Marketing, we refer to brand equity as the additional sales a brand can obtain when everything else is equal, i.e., the causal effect of brands on sales. Does the above average sales ranking causally identify which brand has the highest brand equity? Why or why not? (**4pts**; 150 words)

The weekly sales ranking does not represent causal brand equity, which is defined as ceteris paribus causal effect of the brand on sales when all product attributes are held constant. Although Samsung has the highest average, the average sales confound brand with screensize, technology, and marketing expense, producing product-mix confounding typical of non-experimental data. Therefore, the difference in sales is driven by product mix and marketing conditions, not by causal effect of brand equity effect itself.

Brand equity requires controlled comparison analogous to an RCT and isolation, where products are matched on observable characteristics so the brand indicator functions as the treatment; then, a regression model could have then be made controlling for price, size, HDR support, same promotion.

2. Marketing Mix Modeling and Endogeneity (28pts)

> ***Q4.*** Run a Marketing Mix Modeling linear regression as follows (**6pts**):

- Run the linear regression below using `fixest` package (Equation 1 hereinafter) (**2pts**).

```
# write you codes for the regression below
library(fixest)
library(modelsummary)
```

```
# the formula
# sales = a + b * final_price + c * marketing_expense + ϵ

ols_1 <- feols(
  sales ~ final_price + marketing_expense,
  data = data_full
)

summary(ols_1)
```

```
OLS estimation, Dep. Var.: sales
Observations: 2,080
Standard-errors: IID
                   Estimate Std. Error t value  Pr(>|t|)
(Intercept)       421.271308  20.126281 20.9314 < 2.2e-16 ***
final_price         0.618258   0.019680 31.4152 < 2.2e-16 ***
marketing_expense   0.093787   0.002806 33.4189 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 145.6   Adj. R2: 0.502144
```

```
# do not modify the code below; this is to print out the results

modelsummary(ols_1,
             stars = T,
             gof_map = c('nobs','r.squared'))
```

```
Warning in gzfile(file, "rb"): cannot open compressed file
'/Library/Frameworks/R.framework/Versions/4.5-
arm64/Resources/library/parameters/R/parameters.rdx',
probable reason 'No such file or directory'

Error in gzfile(file, "rb") : cannot open the connection
```

(Intercept)	421.271***
	(20.126)
final_price	0.618***
	(0.020)
marketing_expense	0.094***
	(0.003)
Num.Obs.	2080
R2	0.503

- p < 0.1, * p < 0.05, ** p < 0.01, ***

$p < 0.001$

- Interpret the coefficients of `final_price`, including coefficients and statistical significance (**4pts**).

The estimated coefficient on final_price is 0.618, with a standard error of 0.020, and it is statistically significant at $p < 0.001$. It represents the partial effect of price on sales, and conditional estimated ceteris paribus.

This means:

- Holding marketing_expense constant, a one-dollar increase in price is associated with 0.618 additional units sold, but endogeneity prevents causal interpretation
- The effect is precisely estimated and highly statistically significant, allowing us to reject the null hypothesis that price has no effect on sales.
- Although price usually reduces demand, the positive and significant coefficient reflects product-mix and brand-mix confounding in the raw dataset. Because product-specific factors are not controlled for in this OLS model and CIA is violated, the regression attributes part of that demand to price itself.

Thus, the model finds a positive and statistically significant association, but this should not be interpreted as a true causal price elasticity.

> ***Q5.*** Based on the regression coefficients reported above, discuss the endogeneity issues with `final_price` in Equation 1. For each endogeneity cause, explain the general definitions and then give concrete examples in Amazon's context. (**12pts**)

- General definition of each endogeneity cause (**6pts**, 200 words)

Endogeneity arises when an explanatory variable is correlated with the regression error term, causing OLS estimates to be biased and inconsistent. The first major source is omitted-variable bias, which occurs when relevant determinants of the dependent variable are excluded from the model. If these omitted factors also influence the explanatory variable, OLS incorrectly attributes their effects to the included regressor, distorting the estimated coefficient. The second source is reverse causality, where the direction of causation runs both ways: the dependent variable affects the explanatory variable. When this happens, price changes reflect demand conditions rather than causing them, meaning the regressor is jointly determined with the outcome. A third common source is measurement error, where the observed explanatory variable differs from the true underlying construct. When measurement error correlates with unobserved demand drivers or is systematically patterned, OLS no longer isolates the causal relationship. A fourth related mechanism is simultaneity, where firms choose price jointly with other strategic variables that directly affect sales, making price an outcome of the same

optimisation problem as demand. Across all these cases, the regressor absorbs the influence of unobserved forces embedded in the error term, violating the exogeneity assumption.

- In Amazon's context, concrete examples to illustrate each endogeneity cause (**6pts**, 200 words)

In Amazon's TV marketplace, omitted-variable bias arises because the regression does not control for core product attributes—screen size, resolution, OLED/QLED panel type, gaming performance, brand reputation and review ratings. Premium TVs are expensive and also inherently popular, so the model wrongly attributes quality-driven demand to price. Reverse causality is present because Amazon uses dynamic pricing. When a TV experiences unexpectedly strong sales—often due to viral reviews, stock scarcity or algorithmic recommendations—Amazon frequently raises the price in response. Thus high sales lead to higher prices, not the other way around. Measurement error further affects final_price: the dataset does not fully capture coupons, lightning deals, warehouse-deal adjustments, third-party seller price changes, or temporary on-site rebates. If highly demanded TVs receive unrecorded discounts, recorded prices overstate the actual transaction price, biasing the estimated relationship. Simultaneity also matters: price and marketing expenditure are jointly chosen during events like Black Friday or Prime Day. Amazon often increases list prices before applying headline discounts or coordinates price positioning with paid placement, making price part of a multi-variable promotional strategy. Each mechanism creates correlation between final_price and unobserved demand shocks in the error term, generating endogeneity.

> ***Q6.*** If the discount each week in our dataset is randomized by Amazon each week, will Equation 1 give the causal effect of price on sales? Give your reasoning. (**6pts**; 200 words)

If weekly discounts were truly randomised by Amazon, then Equation 1 would move closer to identifying the causal effect of price on sales, but only under additional conditions that are rarely met in practice. Randomisation breaks the correlation between price and many demand-related factors such as product quality, stock levels, or brand-specific demand shocks, because each week's discount no longer reflects managerial optimisation or real-time demand. In that narrow sense, randomized discounts make final_price exogenous with respect to unobserved demand determinants, satisfying the core requirement for causal identification in OLS.

However, randomisation alone is not sufficient. First, randomisation must occur within product, not only across products or across weeks; otherwise, underlying differences between brands, screen sizes, and specifications still confound the relationship. Second, price must be the only channel through which the randomized discount affects sales. If discounts simultaneously influence visibility (e.g., through Amazon's ranking algorithms, badge placement, or promotion slots), then the exclusion restriction is violated because sales respond to both price and non-price effects of the discount. Third, the randomisation must not change consumer composition or trigger inventory-related allocation rules.

Therefore, while randomized discounts greatly reduce endogeneity, Equation 1 recovers a clean causal price effect only if randomisation is both product-level and affects no other sales drivers beyond price.

> ***Q7.*** From the below regression designed by another data scientist, discuss whether customers always prefer larger screens (i.e., everything else being equal, a larger screen always leads to higher sales)? (**4pts**; 150 words)

No, this regression does not support the claim that "bigger is always better". Screen size enters as a factor, so each coefficient is the incremental sales relative to the omitted base category (likely <40 inches), holding price and marketing constant. Moving from small TVs to 40–49 inches raises weekly sales by about 49 units, and 50–59 inch models add roughly 144 units, so demand clearly increases for mid-to-large screens. However, the coefficient for 60+ inches (≈82) is much lower than for 50–59 inches and only moderately higher than for 40–49 inches. Given the small standard errors, this drop is substantively and statistically meaningful.

So the pattern is non-monotonic: customers prefer larger screens up to around 50–59 inches, but the very largest TVs do not generate the highest sales once we control for price and marketing. Space constraints and niche demand for 60+ inch models are plausible reasons.

3. Instrumental Variables (20pts)

> ***Q8.*** One way to obtain causal effects of price on sales from secondary data is to use the instrumental variable method. (**12pts**)

- List two variables you you would collect as instrumental variables for `final_price`

[placeholder for discussion answers]

- Can one use the VAT tax rate of TV products as an instrument variable for `final_price`? (**4pts**; 100 words)

No. The VAT rate on TVs cannot serve as a valid instrument for final_price in this context. For an instrument to work, it must vary in a way that shifts price while remaining unrelated to underlying demand shocks. VAT on TVs in the UK is a **national, fixed, product-category tax** that does not change week-to-week or across brands, so it lacks the **relevance** needed for identification: a variable with no meaningful variation cannot explain variation in final_price. It also risks violating **exogeneity**, because category-specific VAT changes typically coincide with policy responses to market conditions, not clean, demand-independent shocks.

> ***Q9.*** Assume you have identified one instrument variable `cost_shifter` in `data_full`. In the code blocks below, write down the two regressions you would need to run in order to estimate the causal effects of `final_price` on `sales`, including `marketing_expense` as the only control variable (**8pts**)

- Correct first stage codes and explanation of the code (**3pts**)
- Correct second stage codes and explanation of the codes (**3pts**)

```
# show the estimation code below and describe the steps

### Stage 1: write the first-stage regression

ols_stage1 <- feols(
  # final_price is endogenous → we explain it using the instrument
  final_price ~ cost_shifter + marketing_expense,
  data = data_full
)

# Extract the predicted (instrumented) component of price
# This captures only the variation in final_price explained by the instrument

data_full <- data_full %>%
  mutate(
    final_price_hat = predict(ols_stage1, data_full)
  )


### Stage 2: write the second-stage regression

ols_stage2 <- feols(
  # Regress sales on the instrumented price and the control
  # final_price_hat replaces the original endogenous price
  sales ~ predicted_TV + marketing_expense,
  data = data_full
)
```

```
Error in feols(sales ~ predicted_TV + marketing_expense, data = data_full):
The variable 'predicted_TV' is in the RHS of the formula but not in the data
set.
Maybe you meant: `product_id` or `final_price_hat`?
```

```
# The coefficient on final_price_hat is the causal price effect

# do not modify the code below; this is to print out the results

modelsummary(list(
  "First Stage" = ols_stage1,
  "Second Stage" = ols_stage2
),
stars = T, gof_map = c('nobs','r.squared'))
```

```
Error: object 'ols_stage2' not found
```

- Based on the results of the two regressions, discuss the causal effect of `final_price` on `sales` (**2pts**)

The IV second stage shows a causal coefficient on final_price_hat of –0.734* (s.e. 0.068). Holding marketing expense constant, a $1 increase in final price causally reduces weekly sales by about 0.73 units. The effect is highly statistically significant ($p < 0.001$) and now has the economically sensible negative sign, in contrast to the biased positive OLS estimate we saw earlier, indicating that the IV procedure has corrected the upward endogeneity bias in the price coefficient.

Q10. Design the A/B/N testing (**20 pts**)

Step 1: Unit of Randomisation

The unit must satisfy the **Stable Unit Treatment Value Assumption (SUTVA)** and minimise **spillover** and **crossover**. For an interface feature like "AI Virtual Try-On", **customer-level randomisation** is the most appropriate.

Session-level or **page-level** assignment risks showing different versions to the same user across visits, violating SUTVA by creating treatment inconsistency. **Device-level** assignment also creates crossover because Amazon shoppers commonly switch between mobile, tablet, and desktop. **Geographic (cluster) randomisation** has low statistical power and high spillover risk because try-on usage (screenshots, shared links, outfit inspiration) can influence users in other regions.

Randomising at the **customer account** ensures each user receives a stable treatment across all devices and sessions, prevents contamination, and mirrors the Instagram A/B/N case where user-level assignment preserved consistent exposure.

3.1 Step 2: Randomisation Scheme (3 pts)

Given 100,000 available users, I use **simple randomisation** with a **1:1:1 allocation** across Control, Treatment A (real-photo try-on), and Treatment B (avatar try-on). This provides high power and ensures symmetry across treatment arms. Following the lecture, I apply **stratified randomisation** on pre-treatment variables such as historical spend, Prime status, and device type to improve covariate balance.

```
#illustrative code
set.seed(123)
data_users$treat <- sample(c("control","A","B"),
                           size = nrow(data_users),
                           replace = TRUE,
                           prob = c(1/3,1/3,1/3))

Error: object 'data_users' not found
```

3.2 Step 3: Sample Size Determination

Amazon expects Treatment A to increase spending by **£10** and Treatment B by **£5**, with **σ = £100**. Using the standard two-sample power formula:

```
n = ((z_0.975 + z_0.8) / (Δ / σ))^2

Error: object 'z_0.975' not found

# Treatment A (Δ = 10):
n = (2.8 / 0.10)^2 = 784

Error in (2.8/0.1)^2 = 784: target of assignment expands to non-language
object

# Treatment B (Δ = 5):
n = (2.8 / 0.05)^2 = 3136

Error in (2.8/0.05)^2 = 3136: target of assignment expands to non-language
object
```

Thus the experiment requires **≈ 3,200 users per arm (≈ 9,600 total)**. Since we have 100,000 customers, no adjustments to allocation are necessary; using all users yields tighter confidence intervals and enables heterogeneity analysis, consistent with the instructor's comments that Step 3 may require rethinking Step 2.

3.3 Step 4: Data Collection

Following the Instagram A/B/N design, I collect data for two purposes:

1. **Randomisation check**
2. **Estimation of the treatment effect**

Pre-treatment covariates: historical spend, device type, Prime membership, browsing history.

Behavioural outcomes: total spending (primary outcome), conversion rate, number of try-on interactions (upload vs avatar), add-to-cart rate, session duration, and return rate.

These metrics mirror the lecture's emphasis on both **engagement measures** and **purchase behaviour**.

3.4 Step 5: Statistical Analysis (4 pts)

Analysis proceeds under an **intention-to-treat (ITT)** framework.

```
# Randomisation checks
t.test(age ~ treat, data = data_users)

Error in eval(m$data, parent.frame()): object 'data_users' not found
```

```r
chisq.test(table(data_users$prime_status, data_users$treat))
```

```
Error: object 'data_users' not found
```

```r
# Estimating treatment effects
Spend_i = α + β_A * A_i + β_B * B_i + ε_i
```

```
Error: object 'α' not found
```

```r
#Where β_A and β_B are the causal effects of real-photo try-on and avatar try-on relative to control.
data_users$treat <- relevel(as.factor(data_users$treat), ref = "control")
```

```
Error: object 'data_users' not found
```

```r
model <- lm(Spend ~ treat, data = data_users)
```

```
Error in eval(mf, parent.frame()): object 'data_users' not found
```

```r
summary(model)
```

```
Error: object 'model' not found
```

Secondary analysis includes comparing A vs B directly and adding covariates for precision. This follows the lecture's recommendation to use regression-based ITT estimation for multi-arm experiments.

> ***Q11.*** Finally, Tom would like to study the causal effect of Amazon rating on product sales. For instance, what is the causal effect of a 4.5-star rating on sales compared to a 4-star rating. Propose **one** natural experiment method to study this causal question. (**12pts**)

3.5 **Method name and intuition**

A suitable natural experiment is a **Regression Discontinuity Design (RDD)** exploiting Amazon's rating-rounding rule. Amazon maps continuous underlying averages (e.g., 4.24 vs 4.26) to discrete displayed ratings (4.0 vs 4.5). Products narrowly above the rounding threshold (≈4.25) show 4.5, while nearly identical products just below show 4.0★. Near this cutoff, assignment to 4.0 vs 4.5 stars is effectively arbitrary, allowing products on either side to be treated as comparable. Any discontinuous jump in sales at the threshold identifies the **local causal effect** of displaying 4.5 rather than 4.0.

3.6 **Data required**

Implementation requires product-level panel data including:

- **Underlying average rating** (running variable).

  Displayed rating (4.0 vs 4.5 indicator).

  Outcome: sales or revenue.

Controls: price, discounting, brand, search rank, Prime eligibility, and week fixed effects.

Data should be restricted to a narrow window around the rounding threshold (e.g., 4.15–4.35) to satisfy RDD comparability conditions. These variables allow measurement of the rating discontinuity's effect while holding other demand drivers constant.

Statistical analysis

Analysis begins with **RDD validity checks**: (1) continuity of the running variable around 4.25 (no manipulation), and (2) continuity of covariates such as price and advertising. The causal effect is then estimated using a **local linear RDD**:

```
sales_i = α + τ * 1{rating_i ≥ 4.25} + f(rating_i) + ε_i

Error in parse(text = input): <text>:1:20: unexpected '{'
1: sales_i = α + τ * 1{
                      ^
```

with flexible functions of the running variable and bandwidth selection. The coefficient τ captures the **local average treatment effect** of receiving a 4.5★ display. Robustness checks include alternative bandwidths and polynomial specifications.