MSIN0094 Third Assignment

Due Friday 10am, Dec 13, 2024

Candidate number:VSXB6

Word count:1986

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**)

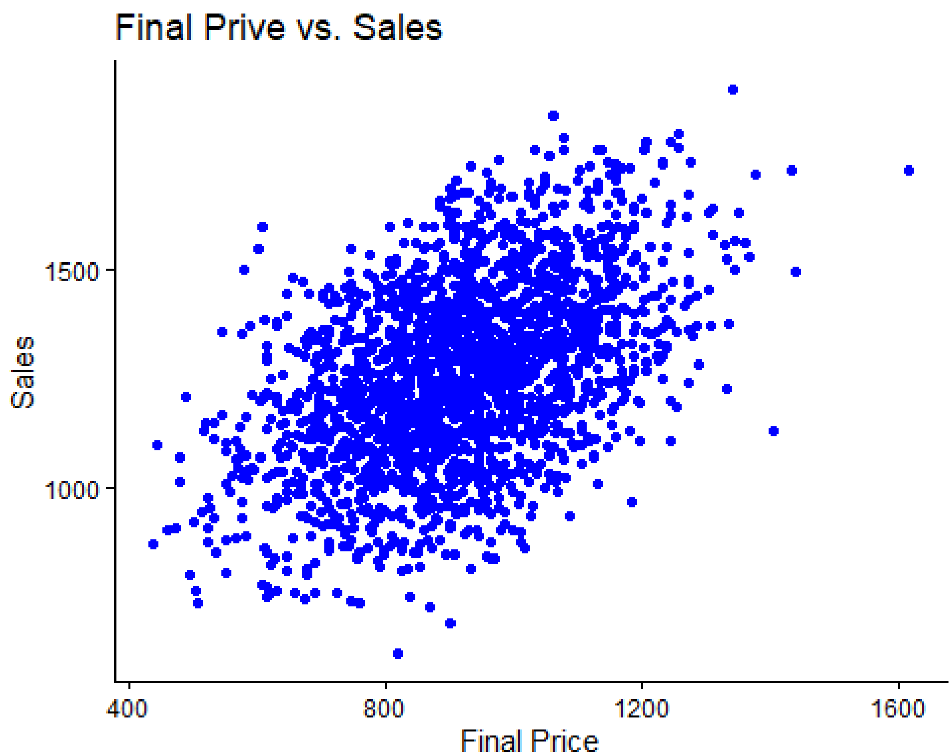
- Write your code below to generate final_price from RRP and discount. (**2pts**)

```
# write your codes below
# mutate a new column called final_price and calculate the values using
 RRP and discount
data_full <- data_full %>%
  mutate(
    final_price = RRP * (1 - discount)
  )
```

- 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**)

```
# write your codes below
# load package
pacman::p_load(ggplot2)

# create scatter plot
ggplot(data_full, aes(x = final_price, y = sales)) +
  geom_point(color = "blue") +
  theme_classic() +
  labs(title = "Final Prive vs. Sales", x = "Final Price", y = "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)

From the scatterplot, I observe a positive relationship between final price and sales. When the final price increases, sales also tend to increase. However, this does not identify as a causal effect. A causal effect refers to estimating the unbiased impact of a business intervention on an outcome. The scatterplot only captures correlation, not causation. The reasons are as follows.

Firstly, correlation is not causation. The scatterplot only shows an association, but it cannot prove that increasing price causes sales to increase. Secondly, some important confounding variables are not controlled, such as marketing promotions or seasonality. Thirdly, the prices are not randomly assigned. Companies usually set higher prices for products with higher demand, which may lead to reverse causality.

In conclusion, although the plot shows a positive association, we cannot claim that raising prices would increase 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
# create a new variable
data_sales_by_brand <- data_full %>%
  # mutate a new column called dollar_sales
  mutate(dollar_sales = final_price * sales) %>%
  # use brand to divide into groups
  group_by(brand) %>%
  # calculate the average of dollar_sales
  summarise(avg_weekly_dollar_sales = mean(dollar_sales, na.rm = TRUE))
 %>%
  # arrange the order by DESC (high to low)
  arrange(desc(avg_weekly_dollar_sales))

# 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.
```

Samsung has the highest average weekly dollar sales, which is 1,277,231, followed by Sony (1,226,644) and LG (1,119,492).

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)

No, the above average sales ranking does not causally identify which brand has the highest brand equity. Average sales only show which brand sells more, but they do not represent brand equity. There are many factors that can affect sales, such as price, product quality, marketing expense, seasonality. If these factors are different across brands, then the difference in average sales cannot be interpreted as difference in brand equity.

Therefore, without controlling for other confounding variables, average sales cannot represent which brand has the highest brand equity.

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
# Marketing Mix Model using linear regression model
ols_1 <- feols(
  sales ~ final_price + marketing_expense,
  data = data_full
)

# summary
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'))
```

(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 coefficient of final_price is 0.618. When marketing_expense is controlled, one-unit increase in final_price can lead to 0.618 units increase in sales. In other words, higher prices are correlated with higher sales.

The p-value for final_price is lower than 0.001, which is far below the 0.05 threshold. This means that final_price is statistically significant, and it is strongly associated with sales.

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 refers to an econometric issue with OLS linear regression, in which a focal explanatory variable is correlated with the error term, violating the Conditional Independence Assumption (CIA). There are two main causes.

1. Omitted Variable Bias (OVB): A determinant of the outcome variable correlated with the focal explanatory variable, but is not included in the regression, either due to data unavailability or ignorance of data scientists.

2. Reverse Causality: The phenomenon that the independent variable affects the dependent variable, and the dependent variable also affects the independent variable at the same time.

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

1. Omitted Variable Bias (OVB): In Amazon's context, OVB arises when an important factor which influences sales is not considered in the above linear regression. For example, seasonality is a relevant factor that may affect sales. Customer demand for TVs during major shopping events such as Black Friday and Christmas increase dramatically, leading to higher sales. At the same time, Amazon may change the pricing decision such as raising prices or offering temporary discounts, causing sales to increase. Since seasonality affects both sales and final_price, failing to include seasonal dummy variables makes final_price correlated with the error term, resulting in OVB.

2. Reverse Causality: Amazon adjusts final_price based on product demand. If a TV is selling very well in a certain period, Amazon may increase the price

because high demand shows that customers are willing to buy more. In this case, not only does sales affect price, but also price influence sales. Since price is partly determined by sales, final_price becomes correlated with the error term, resulting in reverse causality.

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 the discount each week is randomized, then Equation 1 can give the causal effect of price on sales.

In Amazon's case, randomization means that the discount and final_prize (which is calculated after discount) are no longer chosen based on demand, popularity, or seasonality, which usually cause endogeneity. This means randomization removes Omitted Variable Bias (OVB) because price is no longer correlated with unobserved factors such as seasonality in the error term. It also removes reverse causality since price is not adjusted by sales anymore.

With random price changes, final_price becomes independent of the error term, so the Conditional Independence Assumption (CIA) holds. In this case, the coefficient of final_price in Equation 1 can be interpreted as the causal effect of price on sales when marketing_expense is under control.

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)

Based on the regression result, customers do not always prefer larger screens. The regression model controls for brand, technology, resolution, support_HDR, and marketing_expense, so the coefficients of screen size show how screen size affects sales when all other factors stay the same. Compared to the baseline, which is the screen size under 40 inches, the coefficient for 50-59 inches is the largest and strongly positive, indicating the highest sales. The 40-49 inch and 60+ inch screen size are also positive and statistically significant, but the coefficient for the 60+ inches is smaller than the 50-59 inch one. This means larger screens do not always generate higher sales.

In conclusion, mid-size screen (50-59 inches) have higher sales than all the other sizes, the but it doesn't mean that larger screen always gain more popularity.

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`

A valid instrumental variable (IV) is a set of variables that satisfies the following requirements:

(1) Exogeneity: the IV is unrelated to the error term

(2) Exclusion Restriction: the IV affects outcome (sales) only through explanatory variable (final_price)

(3) Relevance: the IV is correlated with explanatory variable (final_price)

I choose cost_shifter and competitor price as the instrumental variables because they satisfy the three IV requirements:

(a) cost_shifter:

- Exogeneity: cost_shifter is not related to the error term. Supplier cost changes, such as shipping cost or manufacturing cost, are external to customers' demand and are not affected by unobserved factors that drive sales.

- Exclusion Restriction: cost_shifter affects sales only through final_price. Higher supplier cost may cause Amazon to increase its price, but customers do not buy more or less because the supplier's cost changed. Therefore, cost_shifter does not directly influence sales.

- Relevance: cost_shifter is correlated with final_price. When supplier cost increases, Amazon may adjust the final_price.

(b) competitor price:

- Exogeneity: competitor price is external to Amazon's demand. The pricing strategies are influenced by competitors such as Walmart, rather than by Amazon's unobserved factors, so it is not related to the error term.

- Exclusion Restriction: competitor price affect Amazon's sales only through Amazon's own final_price. Customers buy on Amazon based on Amazon's prices, not because the change of competitor's price. Therefore, competitor price does not directly influence Amazon's sales.

- Relevance: competitor prices are strongly related to Amazon's pricing decisions. When competitors offer discounts, Amazon often adjusts its final_price to stay competitive.

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

No, the VAT tax rate can not be used as an instrumental variable for final_price. Although VAT affects final_price, but it can also directly affect sales. That is, a higher VAT rate increases the total purchase cost for consumers and may reduce the demand, which violates the Exclusion Restriction requirement. In addition, VAT may also relate to broader economic conditions, such as inflation, which influence demand, so it is not fully exogenous. In conclusion, since VAT can directly affect sales and may correlate with the error term, it is not a valid IV for final_price.

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
# Instrument variable cost_shifter predicts final_price
ols_stage1 <- feols(
  final_price ~ cost_shifter + marketing_expense,
  data = data_full
)

# summary
summary(ols_stage1)

OLS estimation, Dep. Var.: final_price
Observations: 2,080
Standard-errors: IID
                    Estimate Std. Error   t value   Pr(>|t|)
(Intercept)       256.793013  40.789853  6.295512 3.7275e-10 ***
cost_shifter        1.098999   0.065810 16.699696  < 2.2e-16 ***
marketing_expense   0.000051   0.002938  0.017277 9.8622e-01
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 152.4   Adj. R2: 0.117532

### Stage 2: write the second-stage regression
# mutate a new variable called final_price_hat -> generated predicted p
rice using ols_stage1
data_full$final_price_hat <- predict(ols_stage1)

# Causal effect of predicted price on sales
ols_stage2 <- feols(
  sales ~ final_price_hat + marketing_expense,
  data = data_full
)

# summary
summary(ols_stage2)

OLS estimation, Dep. Var.: sales
Observations: 2,080
Standard-errors: IID
                     Estimate Std. Error  t value  Pr(>|t|)
(Intercept)       1665.622296  63.066790  26.4104 < 2.2e-16 ***
final_price_hat     -0.733567   0.067583 -10.8543 < 2.2e-16 ***
marketing_expense    0.093342   0.003316  28.1501 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 172.0   Adj. R2: 0.305004

# 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'))
```

	First Stage	Second Stage
(Intercept)	256.793***	1665.622***
	(40.790)	(63.067)
cost_shifter	1.099***	
	(0.066)	
marketing_expense	0.000	0.093***
	(0.003)	(0.003)
final_price_hat		-0.734***
		(0.068)
Num.Obs.	2080	2080

	First Stage	Second Stage
R2	0.118	0.306

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
- Based on the results of the two regressions, discuss the causal effect of `final_price` on `sales` (**2pts**)

The second-stage result shows a negative causal effect of price on sales. The coefficient of final_price_hat is -0.734 and statistically significant (p-value<0.001), meaning that every one unit of price increase reduces sales by around 0.734 units. This result is different from first-stage OLS result because OLC suffered from endogeneity. Therefore, after removing bias through instrumental variables, the causal effect is negative.

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

Step 1: Decide on the Unit of Randomisation

The unit of randomization should match the nature of the treatment. The possible units include individual, household, device, store, or city level. For the AI Virtual Try-On experiment, the most appropriate unit is the individual customer.

The feature directly affects each customer's personal shopping journey, so randomizing at the household or city level would be too coarse and may mix users with very different behaviors.

Randomizing at the device level is also not ideal because customers may use multiple devices, which would introduce contamination if the treatment is not consistent across devices. Randomizing at the store or city level would create very large clusters and reduce statistical power.

Therefore, assigning the treatment at the individual level creates a clean between-subject design, which fits the digital A/B/N experiments.

Step 2: Decide on Randomisation Allocation Scheme

Since A/B/N testing can be costly and risky, I will randomly assign 100,000 customers into three groups with equal probability.

- Group 1: Control

- Group 2: Treatment A – Real Photo

- Group 3: Treatment B - Virtual Cartoon Figure

```
if (FALSE){
```

```
# Randomly assign customers to Control, Treatment A, or Treatment B
  # data
  experiment_df <- data.frame(
  customer_id = 1:100000
)
  # assign to 3 groups
  set.seed(123)
experiment_df$treatment <- sample(
  c("Control", "A_real_photo", "B_cartoon_avatar"),
  size = nrow(experiment_df),
  replace = TRUE
)

# Check assigned group sizes
table(experiment_df$treatment)
}
```

Step 3: Decide on Sample Selection and Treatment Duration

Use two-sample t-test power analysis.

- Treatment A: +£10 vs Control

- Treatment B: +£5 vs Control - Standard deviation = £100

The required sample size per group can be calculated as follows:

```
if FALSE{

# Power analysis for Treatment A (effect = 10)
power.t.test(delta = 10, sd = 100, sig.level = 0.05,
             power = 0.80, type = "two.sample")

# Power analysis for Treatment B (effect = 5)
power.t.test(delta = 5, sd = 100, sig.level = 0.05,
             power = 0.80, type = "two.sample")
}

Error in parse(text = input): <text>:1:4: unexpected numeric constant
1: if FALSE
       ^
```

Step 4: Collect Data

I need to collect the following two types of data. The data serve two purposes: randomisation check and estimation of treatment effects.

Demographic data — to conduct the randomisation check to confirm that the three groups are balanced.

Behavioural data — to estimate the treatment effects.

Step 5: Interpreting Results from a Field Experiment

First, we need to conduct a randomisation check to ensure that the treatment and control groups have similar characteristics. We compare demographic variables across groups using t-tests. For any significant differences, we include these variables as controls in the regression model.

Next, we analyse the treatment effects by comparing the key outcome metrics between the groups. Since this is an A/B/N experiment, we use linear regression models to estimate the differences in spending across conditions. We can also run a regression to estimate the average treatment effects while controlling for demographic variables.

The regression specification is:

$$spend_i = \alpha + \beta_a.T(A,i) + \beta_b.T(B,i) + \gamma.X_i + \varepsilon_i$$

where T(A,i)and T(B,i)are treatment indicators, and X_i includes demographic controls. Based on the regression estimates, I can determine whether Treatment A or Treatment B leads to higher spending and evaluate which feature provides more value for Amazon.

```
if FALSE{
# 1. Randomisation check using t-tests
t.test(control_A ~ treatment, data = final_df)
t.test(control_B ~ treatment, data = final_df)
# Repeat for other demographic variables if needed

# 2. Create treatment indicators
final_df$treat_A <- ifelse(final_df$treatment == "A_real_photo", 1, 0)
final_df$treat_B <- ifelse(final_df$treatment == "B_cartoon_avatar", 1,
 0)

# 3. Regression model for A/B/N experiment
model <- lm(
  spend ~ treat_A + treat_B +
    control_A + control_B + control_C + control_D,
  data = final_df
)

# 4. View regression results
summary(model)
}

Error in parse(text = input): <text>:1:4: unexpected numeric constant
1: if FALSE
       ^
```

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**)

A natural experiment is an event in which individuals are exposed to the quasi-experimental conditions that are determined by nature or exogenous factors beyond individuals' control. The process governing the exposures arguably resembles randomised experiments.

(1) I choose Regression Discontinuity Design (RDD) method because the key variation comes from a rating cutoff, not from a policy change over time. Amazon may show different products based on the ratings. However, products with ratings just above and just below 4.5 are likely to be very similar in quality and demand, but they receive different treatment in terms of displayed rating. Around this cutoff, comparing sales just above vs just below the threshold can identify the causal effect of a higher rating.

(2) To implement RDD, I would collect the following weekly data from Amazon.

   (a) Average User Rating: this is essential because RDD relies on comparing products just above and below the rating cutoff.
   (b) Displayed Rating (the ratings shown to customers): check how the change of user ratings affects the displayed rating (treatment)
   (c) Sales: the outcome variable
   (d) Price, promotion, advertising, product category, time: these are the factors that affect sales, so it is important to control these variables in regression models

(3) After data collection, I would use a Regression Discontinuity Design (RDD) and run a regression model to estimate the causal effect of ratings on sales. Firstly, I would keep only products with ratings close to the cutoff, for example, between 4.2 and 4.8. Then, I create a treatment variable "highrating", which equals to 1 if the rating is above 4.5, and 0 otherwise. I also will create a running variable, which is the rating minus 4.5.

The regression I would run is as follows:

$$sales = \alpha + \beta . highrating + \gamma . (rating - 4.5) + \delta X + \varepsilon$$

where X includes price, promotion, advertising, product category, and time controls. The coefficient β measures the causal effect of moving from below 4.5 stars to above 4.5 stars on sales.