MSIN0094 Third Assignment

Due Friday 10am, Dec 13, 2024

Candidate number: NZQG9

Word count: 1970

1 Data Merge and Wrangling (20 pts)

Q1 Merge the 3 datasets into `data_final` based on the following instructions (**10pts**):

- Use `left_join()` to join `data_product` and `data_sales` based on appropriate common keys. Assign the joined dataset to `data_full` (**2pts**). Explain the rationale for your codes, including which data frame is chosen as the main data frame, which dataset is put as the first argument of `left_join()`, and the choice of common keys. (**3pt**; 100 words).

```
# Merge data_product and data_sales below

data_full <-  data_sales %>%
  left_join(data_product, by = c("product_id" = "product_id")) #data_sales is our main
set and we left join data_product as data_product has the most granular level of product-
week
```

In this first merge, we keep data_sales as the main data set and left join data_product to create data_full. We choose data_sales for our main set as it is at the most granular level, product-week, while data_product is only at the product level. Because we want to analyse trends of sales over weeks, it makes sense to keep all the information from data_sales and just add in product information from data_product.

- Further use `right_join()` to join `data_marketing` and `data_full` based on appropriate common keys. Assign the joined dataset to `data_full` (**2pts**). Explain the rationale for your codes, including which data frame is chosen as the main data frame, which dataset is put as the first argument of `right_join()`, and the choice of common keys. (**3pts;** 100 words)

```
# Merge data_marketing onto data_full below

data_full <- data_marketing %>%
  right_join(data_full, by = c("brand" = "brand", "week_id" = "week_id")) #Merging by
brand and week level, data_full is our main dataset and joining data_marketing to it.
```

For the second merge, we keep data_full as our main data set, but because it is a right merge the code looks a little bit different. Because data_marketing is at the brand-week level and data_full is at the product-week level, data_full is more granular and should be kept as our main data set. This merge simply adds marketing information for each brand which will be the same across a brand's products but vary across weeks.

Q2. From `data_full`, perform the following descriptive analytics (**10pts**)

- Generate a new variable, `final_price`, which is the actual retail price for each week (i.e., Recommended Retail Price after discounts). (**2pts**)

```
# write you code below to generate final_price from RRP and discount
data_full <- data_full %>%
  mutate(final_price = RRP * (1 - discount))
```

- 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. (**3pts**) Which brand has the highest average weekly dollar sales? (**1pts**).

Sony has the highest average weekly dollar sales at $3,857,785.05 based on the data set provided.

```
#Before calculating average, checking each brand has same number of distinct weeks so
we can use mean in subsequent calculation

data_full %>%
  group_by(brand) %>%
  summarise(total_weeks = n_distinct(week_id))
```

```
# A tibble: 4 × 2
  brand   total_weeks
  <chr>         <int>
1 LG               52
2 Philips          52
3 Samsung          52
4 Sony             52
```

```
data_sales_by_brand <- data_full %>%
  mutate(dollar_sales = final_price * sales) %>% # Calculate dollar sales
  group_by(brand, week_id) %>% # Group by brand and week
  summarise(weekly_dollar_sales = sum(dollar_sales, na.rm = TRUE), .groups = "drop") %>%
# Sum dollar sales by brand-week
  group_by(brand) %>% # Group by brand
  summarise(avg_weekly_sales  = mean(weekly_dollar_sales, na.rm = TRUE)) %>% # Compute
the average weekly sales per brand
  arrange(desc(avg_weekly_sales)) # Rank from highest to lowest

# View the result
print(data_sales_by_brand)
```

```
# A tibble: 4 × 2
  brand   avg_weekly_sales
  <chr>              <dbl>
1 Sony            3857785.
2 Samsung         3805236.
3 LG              2702107.
4 Philips           25310.
```

```
#Checking validity of results, summing rows for each brand

# Count the number of rows for each brand
brand_row_count <- data_full %>%
  group_by(brand) %>%
  summarise(row_count = n(), .groups = "drop") %>% # Count rows for each brand
  arrange(desc(row_count)) # Sort by number of rows

# View the result
print(brand_row_count)
```

```
# A tibble: 4 × 2
  brand   row_count
  <chr>       <int>
1 Sony         9984
2 LG           6656
3 Samsung      4992
4 Philips      2808
```

```
# 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_sales
  <chr>              <dbl>
1 Sony            3857785.
2 Samsung         3805236.
3 LG              2702107.
4 Philips           25310.
```

- 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. Discuss whether the above average sales ranking can causally inform us which brand has the highest brand equity? (**4pts**; 100 words)

The sales ranking above orders brands by a flat average weekly sales amount, offering insight into overall company performance but not brand equity. Variations in marketing expenses and discounts across brands and weeks mean sales were not achieved under equal conditions. Additionally, brand equity cannot be measured accurately if products vary significantly, as is the case with these companies.

2 Marketing Mix Modeling (40pts)

Q3. 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
ols_1 <- feols(sales ~ final_price + marketing_expense, data = data_full)
```

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

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

```
Error: The `regex` supplied `lines_insert()` did not match a unique line.
```

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

	(1)
(Intercept)	-2.509***
	(0.157)
final_price	0.002***
	(0.000)
marketing_expense	0.049***
	(0.001)
Num.Obs.	24440
R2	0.209
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

The coefficient of final_price in our linear regression model is 0.002 with a p value less than 0.001. The coefficient of marketing expense was 0.049 with a p value less than 0.001. This means that increasing the final price by 1 dollar increases final sales by 0.002 units at a statistically significant level. Similarly, increasing marketing expense by 1 dollar increases sales by 0.049 units also at a statistically significant level.

> ***Q4.*** 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. (**14pts**)

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

Endogeneity is an issue which can occur in regression models when one or more of our independent variables are correlated with the error term of our model. We can categorise these issues as falling into one of three groups, Omitted Variable Bias (OVB), Simultaneity, and Measurement Error (ME).

OVB can occur in a regression model when a confounding variable, which affects both our independent and dependent variables, is not included in our model. As a result, this variable's effect is not measured and may be incorrectly captured by the models we do include. Simultaneity occurs when instead of our independent variable(s) affecting our dependent variable, the relationship goes both ways to some extent and our dependent variable may have a causal effect on one or more of our independent ones. Finally, ME occurs when the metrics used to measure our variables are incorrect and contribute to the error term, meaning the coefficient is not reflective of reality and biases our model.

- In Amazon's context, concrete examples to illustrate each endogeneity cause (**8pts**, 200 words)

In the context of our analysis, we can note that there may be OVB related to the lack of a metric for brand equity or brand perception. In the analysis of the final_price coefficient, we noted that an increase in price increases sales of the product. However, this is counter-intuitive, and may be because a company with strong brand perception and equity are able to charge higher prices while retaining a larger customer base.

Regarding Simultaneity, high/low sales for a product may impact how amazon or the companies behind each product choose to determine it's final_price or marketing expenses. For example if sales are low for a specific TV, there might be more of a push with marketing or discounts on prices. This represents a potential causal relationship that flows both ways in our model and could lead to endogeneity.

Finally, ME could occur in any of the variables in our dataset. If the final_price variable applies the wrong discount, or if marketing expenses are not calculated correctly, noise could potentially be introduced into our model. In a practical sense, it means the model could mistakenly attribute increases or decreases in sales to our explanatory variables which are not reflective of the true information.

> ***Q5.*** 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**; 150 words)

In the previous section I noted the issues of possible unobserved confounders such as product/brand quality as well as simultaneity. If the discounts are randomized, this solves some of the endogeneity problems in our model as changes in the final_price variable would now be independent of our omitted confounders. If randomization is implemented correctly in our dataset, the effect of final_price would be exogenous and thus Equation 1 would give the causal effect of price on sales. However, this is also contingent on any measurement errors being eliminated to ensure our data is reflective of real world information. If these issues, unobserved confounders, simultaneity, and measurement errors, are resolved through randomization and robust data collection, we can confidently assert that Equation 1 establishes a causal relationship between price and sales.

Q6. The following is how we typically measure brand equity using Marketing Mix Modeling: We can further add brand factor variables to Equation 1, so that the coefficients of the brand factors represent the brand equity of each brand relative to the baseline brand after controlling for other confounding variables. (**8pts**)

- Factorize brand into a factor variable, `brand_factor`, using *Philips* as the baseline group and run the above regression. Interpret the coefficients of `brand_factor`. (**4pts**)

  In this regression, Phillips is the baseline group so other brand coefficients will show how their sales compare to Phillips. Firstly, the coefficient for Samsung was −1.416 statistically significant at the 5% level. This means that if we hold final_price and marketing_expense as constants, sales for Samsung will be on average 1.416 units less than Phillips. Secondly, the coefficient for LG was −4.489 highly statistically significant with a p value less than 0.001. Similarly, sales for LG will be −4.489 units less than Phillips on average all else held constant. Finally, the Sony brand coefficient was −4.356 also with a p value less than 0.001 implying high statistical significance. Thus, sales for Sony are 4.356 units less than Phillips with other variables held constant.

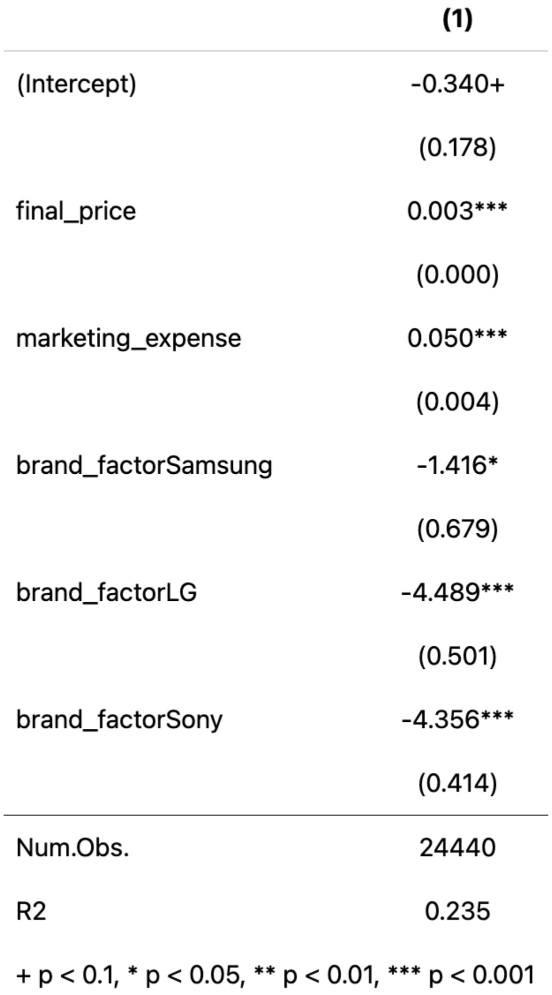
```
# mutate brand_factor below

data_full <- data_full %>%
   mutate(brand_factor = factor(brand, levels = c("Phillips", setdiff(unique(brand),
"Phillips")))) #Converting brand to factor, setting Phillips as the baseline group

# run the regression below
ols_brandeffect <- feols(sales ~ final_price + marketing_expense + brand_factor, data
= data_full)
```

```
The variable 'brand_factorPhilips' has been removed because of collinearity (see
$collin.var).
```

```
# please do not modify.
modelsummary(ols_brandeffect,
             stars = TRUE,
             gof_map = c('nobs','r.squared'))
```

```
Error: The `regex` supplied `lines_insert()` did not match a unique line.
```

- Based on the regression results, please rank brands based on the order of brand equity from the highest to the lowest. Explain how you get the ranking. (**4pts**; 100 words)



	(1)
(Intercept)	-0.340+
	(0.178)
final_price	0.003***
	(0.000)
marketing_expense	0.050***
	(0.004)
brand_factorSamsung	-1.416*
	(0.679)
brand_factorLG	-4.489***
	(0.501)
brand_factorSony	-4.356***
	(0.414)
Num.Obs.	24440
R2	0.235

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Given the coefficients of each brand_factor variable represent sales compared to our baseline brand of Phillips, we can rank the brands based on the magnitude and direction of this difference. At the top, Phillips has the highest brand equity as all else held equal, the other brands all sell less units on average. Samsung is second as their coefficient was −1.416 meaning they sell only 1.416 units less than Phillips. Sony and LG take the 3rd and 4th places respectively as their coefficients were −4.356 and −4.489, significantly less than both Phillips and Samsung.

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)? (**6pts**; 150 words)

Based on the regression output, we can assume that the baseline screen size is the 30-39 inch category, as this is omitted from the output to avoid multicollinearity. The regression testing the effect of screen size on sales shows that larger screens in the 50-59 inch range increase sales by 2.574 units statistically significant at the 0.05 level implying customers prefer screens in this range compared to the smaller baseline. The only other statistically significant screen size factor was the up to 29 inch variable, which showed sales for products in this category were −0.133 units less all else held constant to the larger baseline. This implies that generally, customers prefer larger screens to smaller ones all else held constant.

However, given that the coefficients for the 40-49 inch and 60 inch and above categories are not statistically significant, we cannot determine whether size reaches a point of diminishing returns or whether 40-49 inch screens are preferred to smaller sizes.

3 Instrumental Variables and Natural Experiments (40pts)

> ***Q8.*** One way to obtain causal effects of price on sales from secondary data is to use the instrumental variable method. (**16pts**)

- List two variables you you would collect as instrumental variables for `final_price`; Justify why they qualify as valid instruments (**12pts**; 6pt for each correct instrument; up to 300 words in total). Tips: first discuss the general requirements of a valid instrument and how your proposed variables can satisfy these requirements.

Instrumental variables should meet 3 requirements of relevance, exogeneity, and exclusion restriction to be good candidates for our model. Regarding relevance, our IV must be correlated with the endogenous variable, in our case final_price. Secondly, the IV should be uncorrelated with the regression's error term and thus exogenous. Finally, the exclusion restriction requirement means our IV should only affect the dependent variable, sales, through the endogenous variable, final_price and not a direct causal pathway.

In our regression model, two possible instrumental variables could be raw material prices and shipping costs. Firstly, raw material prices are relevant as they are tied to the production costs which affect final_price. It is also exogenous as they are higher up in the product manufacturing process and so likely do not affect our error term containing noted unobserved factors like brand equity. Finally, raw material prices affect sales only through changes in final_price, so the variable does not directly cause changes in sales and meets the exclusion restriction requirement.

Shipping costs are also a good IV for similar reasons to raw material prices. Shipping costs affect final prices so they are relevant, and only affect sales through changes in final_price meeting exclusion restriction. They also have no relation to brand equity or product quality as they are not inherent characteristics of the product and therefore meet the exogeneity requirement as well.

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

  VAT tax rate should not be used as an instrumental variable for final_price for two main reasons. Firstly, tax rates are often constant over long periods of time, meaning the variable does not have the necessary variation for it to be a useful IV. Secondly, there may be endogeneity issues with this variable. Notably, changes to tax rates are often in response to underlying macroeconomic conditions, meaning that an increase in VAT may influence sales through factors outside of its link to final_price. This would violate the exclusion restriction requirement and make VAT a bad IV.

Q9. Assume you have collected two instrument variables `Z1` and `Z2` into `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` assuming `marketing_expense` as the only confounding variable (**8pts;** this question is just to illustrate the code, no need to really use R to estimate a regression)

- Correct first stage codes and explanation of the code (**4pts**)
- Correct second stage codes and explanation of the codes (**4pts**)

```
# do not modify the above #| eval: false, it's to ask R not to run this code block.

# show the estimation code below and describe the steps

### Stage 1: write the first-stage regression you would run below

IV_1ststage <- feols(final_price ~ Z1 + Z2 + marketing_expense, data = data_full) #Firstly, we need to predict the variable final_price using our instrumental variables Z1 and Z2, as well as the only confounding variable marketing expense. We assign this regression to IV_1ststage

data_full <- data_full %>%
  mutate(predicted_final_price = predict(IV_1ststage)) #Next, we mutate a new variable predicted_final_price that takes the values from our first stage regression.

### Stage 2: write the second-stage regression you would run below
IV_2ndstage <- feols(sales ~ predicted_final_price + marketing_expense, data = data_full) #Finally, we run a regression to predict sales based on our IV adjusted predicted_final_price variable, accounting for the confounding variable marketing expense as well.
```

Q10. 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 stars on sales than 4 stars; so on and so forth. Based on the causal inference methods we learned in class, solve the causal question. (**16pts**)

- name of the method, and the intuition of why the chosen method can give causal effects (4pts;100 words)

The method I will use is Regression Discontinuity Design (RDD). This method can prove causality through the comparison of observations just above and just below a set threshold. In our case, this threshold would be the star rating which we could set as 4.5, 4, 3.5 stars. This method assumes that products just below/above the threshold are highly similar and so the difference in product sales can be attributed to the star rating.

- data you are going to collect for the chosen causal inference method (4pts; 100 words)

The main data we would have to collect for our RDD is product ratings, which is our continuous running variable, as well as sales data for those products. Additionally, we would want to collect data on possible control variables such as the marketing expenses or price of the product. We would be especially interested in collecting ratings data on products that are near our set threshold of 4.5, such as products with 4.49, 4.51 ratings. We could expand this regression to include other thresholds such as 4, 3.5 etc. This would increase the amount of data we would need to include.

- after data collection, the procedures to execute the method using linear regressions (8pts; 200 words)

Firstly, we have determined that our continuous running variable for our RDD is amazon product ratings. We then want to define the thresholds that cutoff our treated and untreated groups. For this RDD, I will use products with a rating up to 0.1 below the threshold as the untreated group, and products with rating up to 0.1 above the threshold as the treatment group. This range, also called a bandwidth, brings with it inherent trade offs. Notably, a smaller bandwidth means experimental units will be more similar which improves our internal validity. However, this also means a smaller sample size which could decrease external validity.

The next step is to examine the continuity of our observed characteristics. Given we are dealing with observations just below and above a cutoff and RDD relies on these products being highly similar, we need to include this step to ensure that characteristics like price, marketing, or brand equity do not change drastically around our cutoff. This can be done through performing a balance test, in which we run regressions of our observed characteristics on our treatment indicator and test for statistically significant differences above and below the cutoff.

Next, we run our RDD regression to estimate the causal effect of crossing our determined cutoff on sales. This regression includes both a binary variable Treated indicating whether the product is above or below the cutoff, as well as our rating running variable and the error term. We then interpret the coefficient of treated in our regression output, which should represent the causal effect of being above the rating threshold.