Data Modeling Exercise

NIQ

October 2023

Preface

This take home assignment is designed to assess the candidate's proficiency in executing tasks that are commonly performed in data-related roles at NIQ. The following exercise is split as follows:

1. Introduction
    - Helps to get familiar with the main dataset
2. Basic Data Preparation
    - Tests standard knowledge of data queries in python
3. Data Description and Reprocessing
    - Tests standard knowledge of simple data analysis procedures
4. Modeling Exercise
    - Tests standard knowledge of simple modeling techniques

Introduction

Start by loading the data into memory. This is a standard store level data-set for a group of Stock Keeping Units (SKUs). In the data we should have the following columns:

- ITEM_ID - Unique code associated with each item
- ITEM_DESCRIPTION - Item name with its associated number
- CHANNEL - Market that is covered by the database
- STORE_ID - Unique store id. Id values do not have any meaning
- PERIOD_ID - Unique week id with meaningful values. The lowest value represent the first week in the data and the largest value represents the last. Each subsequent iteration from the first week adds a single week
- SALES_UNITS_EXPANDED - Number of units sold
- SALES_VOLUME_EXPANDED - Volume sold
- SALES_VALUE_EXPANDED - Value of units sold

ITEM_ID	ITEM_DESCRIPTION	STORE_ID	PERIOD_ID	SALES_UNITS_EXPANDED
1	001_Item	31905	70	1
1	001_Item	31905	71	3
1	001_Item	31905	72	1
1	001_Item	31905	73	4
1	001_Item	31905	77	1

Each row is interpreted as data for a single SKU at a given time period at a given store. So for example, the first row tells us that 1 unit of 001_Item was sold at store 31905 and week 70.

1 Basic Data Preparation

The following set of tasks will test familiarity with basic data manipulation techniques.

1. First lets start by making sure that our data is the correct format. ITEM_ID, STORE_ID, PERIOD_ID should be integers, all sales columns should be floats and the rest should be objects

2. Convert PERIOD_ID in proper date format
   Assume that the first period lowest value of PERIOD_ID is associated with week number 19 and year 2019. Standard procedure at NIQ is to record the Sunday of a given week as the date for this PERIOD_ID.

   - Hint: Use week calendar

3. Calculate numeric distribution of each item by month
   Numeric distribution for an item is defined as

$$\frac{\sum \text{Shops in which item was present}}{\sum \text{Shops in universe}}$$

   Assume that these 8 items represent the whole market so the union of their shops should be equal to the universe. What is the distribution value of 005_Item? Which month is it the highest?

2 Data Description and Pre-processing

In this section we want to get familiar with the data set and get a feel for working with sales price data

1. Draw a histogram of SALES_UNITS_EXPANDED for each item

   (a) What distribution does it follow?

   (b) What are the parameters of this distribution?

   (c) Does the answer change if we applied a log transformation?

2. Generate a variable called ACTUAL_PRICE
   This will be the associated price per unit for a particular SKU at a given store week. Use the available columns to obtain this value

3. Draw SALES_UNITS_EXPANDED and ACTUAL_PRICE on a single graph with PERIOD_ID on the x-axis

   (a) Do you notice any seasonal patters? Is it similar for all items?

   (b) Can we deduce that price reductions lead to sales increase from this graph? Why/Why not?

   Now plot a second graph with SALES_UNITS_EXPANDED and number of shops with PERIOD_ID on the x-axis

   (c) Does this support your conclusion in (b)?

4. Discuss how would you approach generating REGULAR_PRICE and PROMO variables
   - REGULAR_PRICE - price per unit in absence of any promo
   - PROMO - a binary variable that indicates for each observation whether price promo was carried out
   Context: When we work with store week sales price data we want to know at the most granular level when price promo is executed. We are given no indicators by stores when they are offering discounts and our job is to figure it out by ourselves using this data.

5. (Advanced) Generate REGULAR_PRICE and PROMO parameters for each item-store-week

   - Hint: Draw an SKU-store-week graphs and use ACTUAL_PRICE to help you deduce whether any price promo was carried out
   - Hint: Assume that promo should not last longer than 17 weeks and the minimum price drop at a given store-week required to be considered as a promo is 5%

Load the additional data set and compare your answers in 4 and 5
To understand REGULAR_PRICE and PROMO review the definition in 4
PROMOTION_PRICE_INDEX - is a ratio of ACTUAL_PRICE and REGULAR_PRICE

6. Discuss the difference between your approach/answers and the new data

7. List and apply the standard data pre-processing techniques before proceeding to modeling stage

   - Hint: Start by identifying and dealing with outliers and then proceed to use appropriate tools to correct possible issues that can arise with sales price data

3 Modeling Exercise

In this section we will try to estimate the impact of our independent variables on our dependent variable

Assume: Data generating process (DGP) for each item is given by

$$y_{it} = x'_{it}\beta + \epsilon_{it} \tag{1}$$

- y_{it} is a scalar value of dependent (aka target) variable in group i at time t
- x_{it} is vector of independent variables in group i at time t with β being a vector of associated coefficients
- ϵ_{it} is a scalar value of idiosyncratic error in group i at time t

Let x_{it} be REGULAR_PRICE, PROMO and PROMOTION_PRICE_INDEX
y_{it} be SALES_UNITS_EXPANDED.

1. List necessary and sufficient assumption to get consistent and asymptotically normal estimate of the marginal impact of price parameters on sales using OLS?
   - Do we require normality of errors to obtain consistent estimate?
   - How does violating spherical errors impact consistency of the estimate?
2. Run OLS to obtain coefficient for **each item**
   (a) Interpret coefficient estimates. Which sign would we expect? What values did we get?
   (b) Perform t-test and F-test. What can we infer from this? Are these test values accurate?
   - Hint: Recall how t and F statistics are calculated

Now change the DGP assumption to a multiplicative model and use this specification from now on

$$y_{it} = x_{it}^{\beta} * \epsilon_{it} \tag{2}$$

3. Estimate β within this new model for **each item**
   - Hint: Apply an appropriate monotone transformation to convert this model to something more familiar

   (a) How does the coefficient interpretation change from the model in (1)
   (b) Perform t-test and F-test. What can we infer from this? Are these test values accurate?
   (c) Which model do you think is more realistic? Explain

Let $\epsilon_{it} = \alpha_i * \omega_t * \upsilon_{it}$ and use this definition from now on

- α_i is a shop specific effect which varies by i (e.g. turnover) with with $E(\alpha_i|x_{it}) = 0$
- ω_t is a week specific effect which varies by t (e.g. seasonality) with $E(\omega_t|x_{it}) = 0$
- υ_{it} is the remaining idiosyncratic term with $E(\upsilon_{it}|x_{it}) = 0$

4. How does this change the estimation procedure in 3? What adjustments, if any, are needed to ensure the consistency of coefficients? Explain

Now assume $E(\alpha_i|x_{it}) \neq 0$ and $E(\omega_t|x_{it}) \neq 0$ and proceed with this assumption to all future tasks

5. Perform OLS under these assumptions and obtain the coefficients for **each item**. How does the estimate change from the before? What does it say about our data?

Lastly, we want to test the impact of each item on each other. To do this, in addition to own independent variables from a given item, we need to add to our x_{it} values from competitor items. So for example, a model for 001_Item would include its own REGULAR_PRICE, PROMO and PROMOTION_PRICE_INDEX and then same values for each competitor. Also, we want to transform REGULAR_PRICE for competitors only into ratios such that

$$RPR_{1j} = \frac{\text{001\_Item REGULAR\_PRICE}}{\text{00j\_Item REGULAR\_PRICE } \forall j}$$

6. (Advanced) Estimate β within this final model for **each item**. How would you interpret these values? What do they imply?

- Hint: Try to think of a way to fill NA values for competitor items store-week with some value
- Hint: Generate a variable that would track the presence of each competitor in each store-week