

# Questions for the case study: Optimizing Promotions for Supermarkets using Data Analytics

This assignment deals with estimating sales from transactions data and promotion optimization.<sup>1</sup> The goal is twofold: (i) understanding what factors affect the demand for items in supermarkets as well as exploring how to estimate demand models from data and (ii) studying one way to formulate the promotion optimization problem solved by category managers in supermarkets. Please submit your answers to the following questions.

1. What factors do you think one should incorporate in the demand function? Please, discuss.

## Estimation – Single Item Model

**Load the csv file “singleitemSKU88.csv”.** The data consists of three years (156 weeks) of weekly transactions for one particular item (labeled as SKU88). Each row corresponds to a given week for which we report: the week number (from 1 to 156, where 1 corresponds to the first week of January 2011), the price of the item (normalized to 1), and the aggregate sales during that week (note that because of the data aggregation, sales are not integer numbers). Using R, run the file “Studentspart1.R” to load the data and add the seasonality (explained below). This will allow you to prepare the data for the next question.

2. In this question, you are required to estimate the demand function for SKU88 using R. As we discussed, the demand at time  $t$  depends on several observable features, such as: seasonality, trend and price. We next consider two specific models.

In each model, we assume that the demand has a log-log form (i.e., the logarithm of the demand is linear in the logarithm of the price).

We also assume that the seasonality is monthly (i.e., constant for 4 consecutive weeks and repeated from year to year). In addition, you are required to split the dataset into two parts: training set (consists of the first 2 years) and testing set (consists of the last year). Estimate the parameters of the models below using the training set.

- (a) Estimate the parameters for the following model:

$$\log(d_t) = a_0 + a_t + \text{Trend} \cdot t + \beta_0 \log(p_t), \quad (1)$$

where  $d_t$  and  $p_t$  are the demand and the price during week  $t$  respectively. Our goal is to estimate the parameters  $a_0$  (the intercept or SKU effect),  $a_t$  (the seasonality – a total of 12 different parameters),<sup>2</sup> Trend (a single parameter that captures the trend effect) and  $\beta_0$  (the price effect). In this model, we have a total of 15 parameters to estimate.

- (b) Now, consider and estimate the parameters for the following model:

$$\log(d_t) = a_0 + a_t + \text{Trend} \cdot t + \beta_0 \log(p_t) + \beta_1 \log(p_{t-1}) + \beta_2 \log(p_{t-2}), \quad (2)$$

where  $p_{t-1}$  and  $p_{t-2}$  are the prices during weeks  $t-1$  and  $t-2$  respectively. Our goal is to estimate the parameters  $a_0$ ,  $a_t$ , Trend,  $\beta_0$ ,  $\beta_1$  (effect of the last price on current demand) and  $\beta_2$  (effect of the second last price on current demand). In this model, we have a total of 17 parameters to estimate.

In both (a) and (b), you are required to submit (i) the summary of the regressions; (ii) the values of  $R^2$  and MAPE on the test set.

---

<sup>1</sup>The data used in this case is synthetic, and was generated for educational purposes only.

<sup>2</sup>We have 13 different  $a_t$  values but we estimate 12 dummy variables, as one of the parameters is normalized.

3. For the model in equation (2), what can you say on the magnitudes of the estimated parameters  $\beta_0$ ,  $\beta_1$  and  $\beta_2$ ? Does this make sense and why?
4. Assume that  $p_7 = 1$ ,  $p_6 = 0.8$  and  $p_5 = 0.8$ . Compute the values of the predicted demand  $d_7$  using models (1) and (2) above.
5. Discuss the differences between the two models (1) and (2) above. For which type of products do you think that model (2) is better? Please discuss.

### Estimation – Multiple Items

Consider a small category of items (e.g., fresh orange juice) with 5 different brands. Our goal is to simultaneously estimate the demand models for each brand.<sup>3</sup> **Load the csv file “multipleitem-part1.csv”**. The data consists of three years (156 weeks) of weekly transactions for five different brands (labeled as  $B1, \dots, B5$ ). Each row corresponds to a given week for which we report: the week number, the prices of each brand (normalized to 1) and the aggregate sales during that week. Using R, run the file “Studentspart2.R”. This will allow you to prepare the data for the next question.

6. In this question, you are required to estimate the demand function for the 5 different brands using R. As before, the demand of each brand  $i$  ( $i = 1, 2, \dots, 5$ ) at time  $t$  depends on several observable features, such as: seasonality, trend and price. In this case, the demand of each brand depends not only on its own price, but also on the prices of the other brands. We next consider a specific model.

As before, we assume that the seasonality is monthly, and that we split the dataset into two parts: training set (consists of the first 2 years) and testing set (consists of the last year).

Estimate the parameters for the following model (using the training set):

$$\log(d_t^i) = a_0^i + a_t + \text{Trend} \cdot t + \beta_0^i \log(p_t^i) + \beta_1^i \log(p_{t-1}^i) + \beta_2^i \log(p_{t-2}^i) + \sum_{j \neq i} \delta_{ji} p_t^j, \quad (3)$$

where  $d_t^i$  and  $p_t^i$  are the demand and the price of brand  $i$  during week  $t$  respectively. Our goal is to estimate the parameters  $a_0^i$  (the intercept for each brand),  $a_t$  (the seasonality, a total of 12 different parameters that are the same for all the brands), Trend (a single parameter that captures the trend),  $\beta_0^i$  (the price effect for each brand),  $\beta_1^i$  (effect of the last price on current demand for each brand) and  $\beta_2^i$  (effect of the second last price on current demand for each brand). We also need to estimate the parameters  $\delta_{ji}$  for  $j \neq i$  (how the price of brand  $j$  affects the demand of brand  $i$ ). In this model, we have a total of 53 parameters to estimate.

Please, submit (i) the summary of the regressions; (ii) the values of  $R^2$  and MAPE on the test set.

7. Repeat the same procedure as in Question 6 using the data from the csv file “multipleitem-part2.csv”. As before, submit (i) the summary of the regressions; (ii) the values of  $R^2$  and MAPE on the test set.
8. Assume that  $p_6^i = 1$  and  $p_5^i = 1$  for all  $i$ , and  $p_7^2 = 1, p_7^3 = 1, p_7^4 = 1, p_7^5 = 1$ . Compute the values of the predicted demands  $d_7^1, d_7^2, d_7^3, d_7^4, d_7^5$  using model (3) calibrated with the data from Question 6, when  $p_7^1 = 1$  and when  $p_7^1 = 0.7$  (i.e., a total of 10 different values). What can you say about the effect of promoting Brand 1?
9. Compare the estimated parameters from Questions 6 and 7. What is the difference between the 5 brands from Question 6 and the 5 brands from Question 7? Please, discuss.

---

<sup>3</sup>In several settings, we cannot estimate the demand at the SKU level and instead, we estimate it at the brand level.

## Optimization Formulation

10. Once the demand models are estimated from data, the next step is to formulate the optimization problem, i.e., maximizing an objective function subject to some constraints. Consider deciding promotions for 5 items throughout the next quarter (composed of 13 weeks). What types of objectives would you consider? Please discuss.
11. Very often, when formulating an optimization problem, it is convenient to use binary decision variables (variables that can take only the values 0 or 1). Denote the price of a particular SKU during week  $t$  by  $p_t$ . In addition, the (normalized) price of the SKU can only take of the following four values: 1 (i.e., full price), 0.9 (10% promotion), 0.8 (20% promotion) and 0.7 (30% promotion). How can one represent such a restriction using linear constraints and binary variables? (Hint: You may use more than one constraint.)
12. As we discussed in the case study, promotions should satisfy several business rules. One such rule is to limit the frequency of promotions for a particular SKU. For example, a retailer may insist that a particular product is promoted at most 4 times during the quarter. How can you model this business rule as a linear constraint with binary variables?
13. Consider the problem of maximizing the total profit throughout the next quarter for a particular SKU (e.g., SKU88). Please, formulate the optimization problem with the constraint on the limitation of the number of promotions to be at most 4.
14. Assume that we solved the optimization problem and obtained the suggested prices for SKU88 for the next 13 weeks. The suggested prices are presented in Exhibit 5 of the case study. What can you say on this output? What tests would you conduct in order to check that the suggested prices are robust? Please discuss.
15. Consider the optimization problem for several items simultaneously. Should the promotions of the different items be at the same time, or at different times? Please discuss the various factors that may affect the answer to this question. (This question is qualitative and you are not expected to solve any optimization problem.)