CS166 LBA Assignment

I. Introduction

In this LBA assignment, we model the grocery prices in different neighborhoods in Buenos Aires and Taipei. There are three main questions we want to address. First, what is the basic average price for each product? Second, how does these two factors - the brand of grocery store and geographical location of the grocery store - influence and modify the basic price of product? Third, we want to see if the geographical location is correlated with rental prices in Buenos Aires or not.

II. Modeling and Assumption[1]

In the statistical model, the observed price of grocery is a function influenced by three factors - the base price of the product, the store brand where it is sold, and the neighborhood of this store. More specifically, the price is sample from a normal distribution with the equation as f(x) = base_price x store_multiplier x location_multiplier as the mean. The standard deviation of the price is modeled by a hyper-parameter as the error term - sigma, which is further modeled by has a gamma prior - e.g. Gamma (1,1), to reflect the random error/nosies generated from collecting data.

There are several other necessary assumptions and specifications made in building this model:

1. The base price of each product is **independent** from each other, which means e.g. the price of eggs is not influenced by the price of tomatoes.
2. The the **base price parameter** is modeled by a **half-Cauchy** prior to reflect the fact of heavy and broad tails (the price of a type of grocery should be around a certain price, but there are be some far right tail to represent e.g. some high-end organic grocery). It is set to be centered around 1. A lower boundary of 0 for the distribution is set to account for that in reality, these parameters are are positive real numbers.
3. Two **multipliers** (brand and geographical location) are modeled by a **log-normal** distribution based on the fact that they are positive, real, and continuous numbers. These properties can be more easily modeled by the log-normal distribution since we can control the mean to be around 1, and have a broad spread of the distribution by manipulating scales in log-normal distribution.
4. Also, we take effects of the brand and location multiplier are applied towards the price of different types of groceries indifferently. This means that in the same store, we don't have different store and location multipliers for different types of groceries (e.g egg and tomatoes). In reality this is likely to be true: for example, in some high-end supermarkets/ expensive regions, we may find all types of groceries to be inflated at the same time.

---

[1]#modeling: gave a through modeling of the grocery prices based on Bayesian statistics.

However, that may not be true in some cases, as we may also find a store to have cheap eggs but very expensive tomatoes at the same time. But this assumption is necessary to help us to simply the question, so we do not need to run ten models for each of the product to find its corresponding store and location parameters.

5. One good thing about Bayesian inference is that the specific initial values for each distribution parameter matter less and less when the amount of observations are growing, as we are keep updating our priors. Since we have a rather larger dataset - each student have 30 data point to contribute, and we have dozens of students together. So we can be less worry about the initial prior parameter setup.

I make a graph to represent the model: in which the grey circle representing the observed real price of groceries.[2][3]

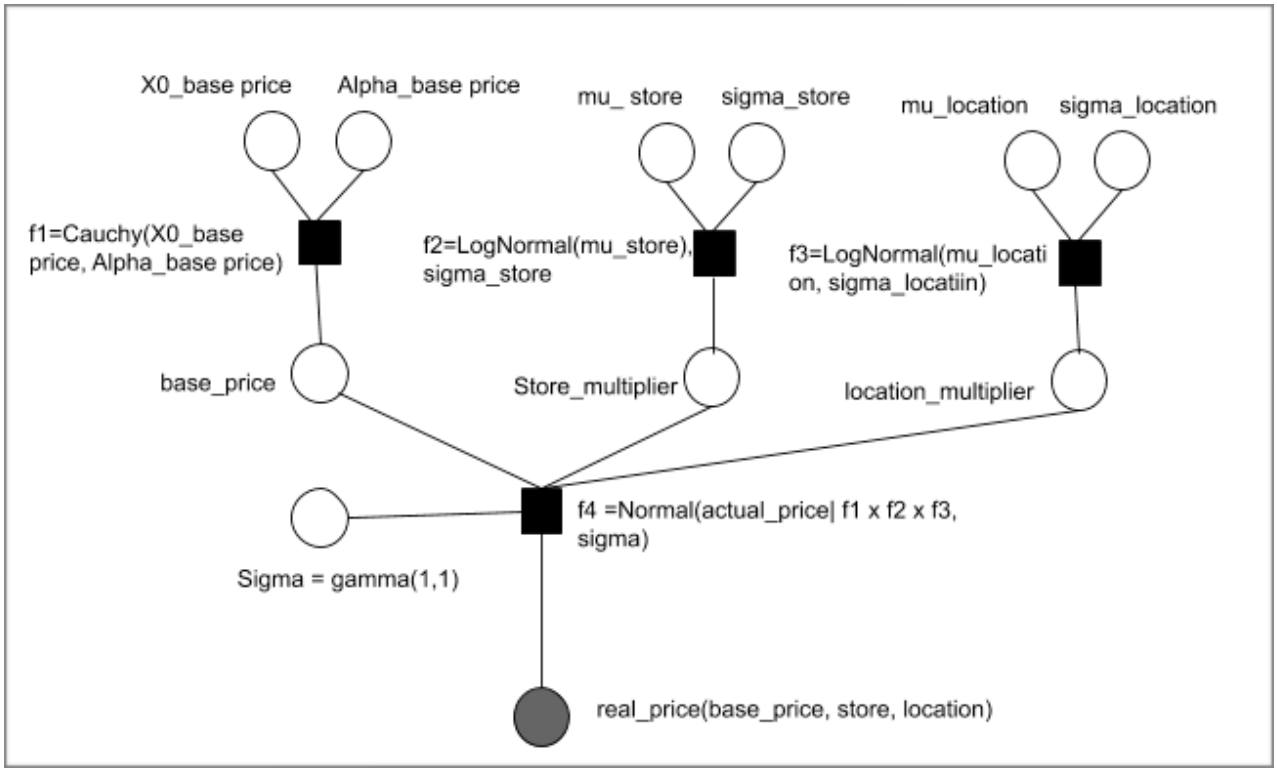

Figure 1. A graph representation of the model.

III. Results and Interpretation

1. The mean of each type of grocery

[2] #distribution: gave a detailed analysis of what types of distribution are appropriate to modeling, and provide justifications.

[3] #probability: used and justified the use a Bayesian probability model to construct the model

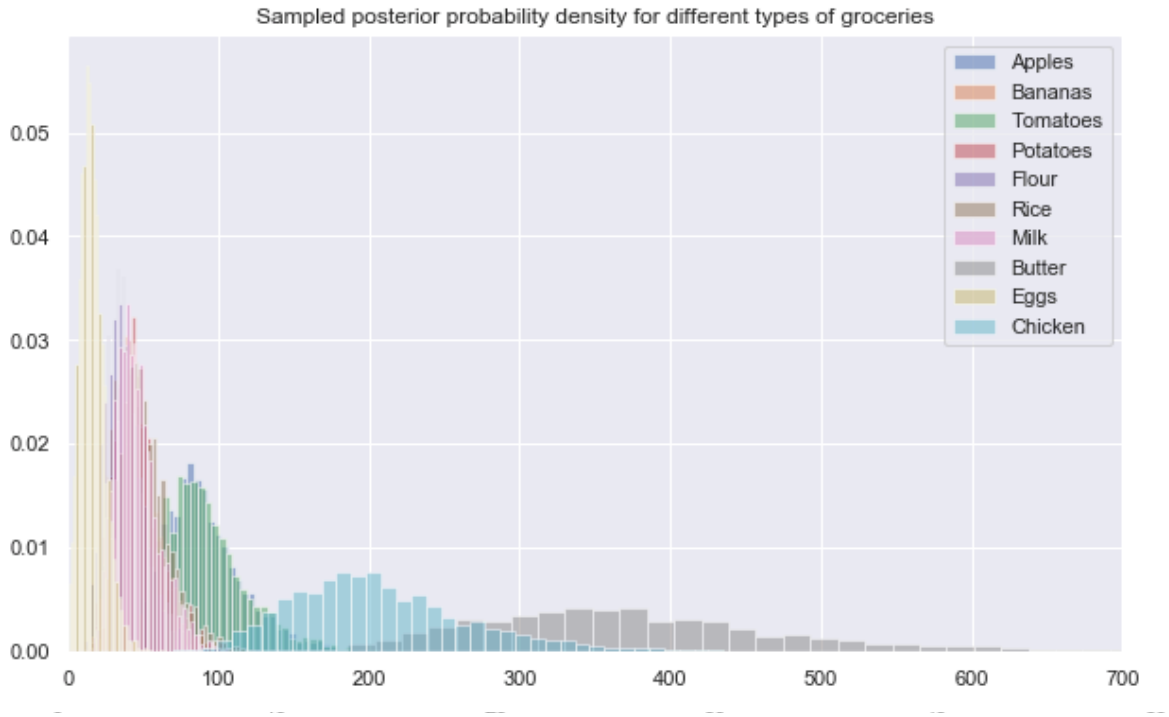

Table 1: Sampled mean of each grocery

Apple	Mean price
Banana	89.58
Tomato	44.04
Potato	90.83
Flour	46.42
Flour	37.81
Rice	51.58
Milk	45.96
Butter	379.91
Eggs	16.24
Chicken	209.18

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
base_price[1]	89.58	1.19	27.87	46.96	85.42	157.15	549	1.0
base_price[2]	44.04	0.6	15.6	20.61	41.74	81.84	686	1.0
base_price[3]	90.83	1.21	28.41	47.5	86.55	161.56	554	1.0
base_price[4]	46.42	0.64	15.77	22.41	43.77	84.58	616	1.01
base_price[5]	37.81	0.53	13.23	17.46	35.91	68.91	628	1.0
base_price[6]	51.58	0.7	17.11	25.86	48.7	92.94	594	1.0
base_price[7]	45.96	0.62	14.98	23.48	43.65	81.62	585	1.0
base_price[8]	379.91	5.06	116.05	203.99	362.8	661.23	526	1.0
base_price[9]	16.24	0.28	8.14	3.38	15.17	34.9	819	1.0
base_price[10]	209.18	2.79	64.06	111.3	199.54	364.55	528	1.0

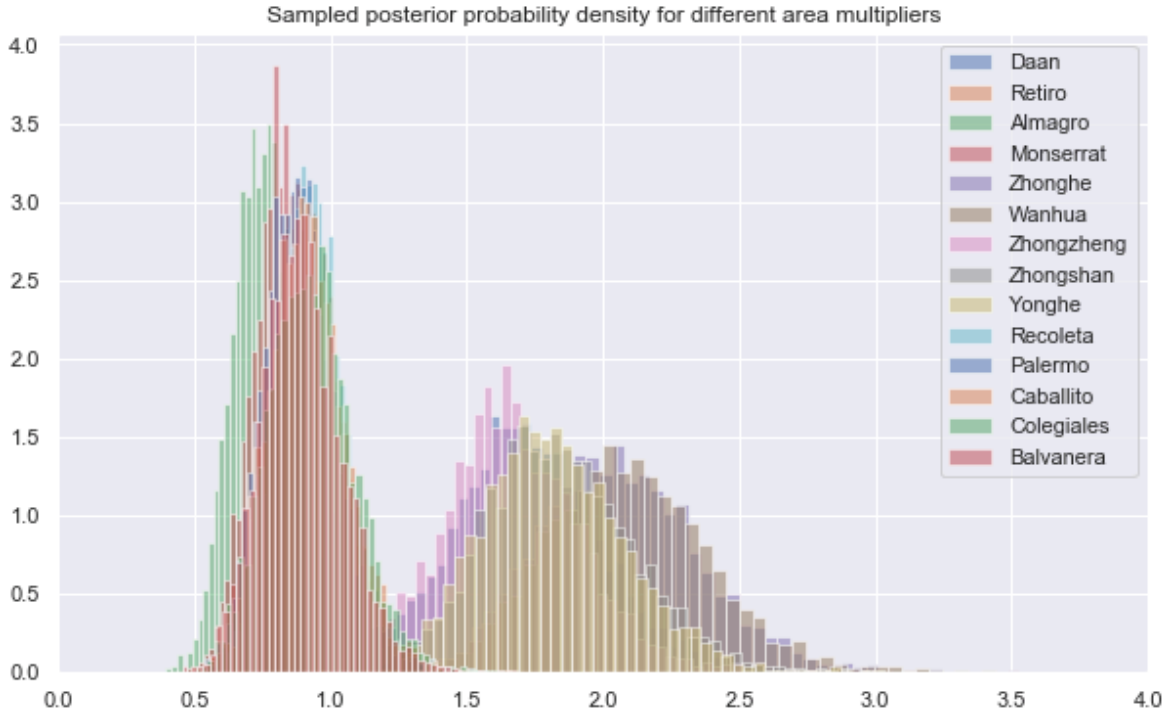

Figure 2. A histogram representing the sampled posterior probability density for different types of groceries.

2.The effect of area & store multipliers on prices

Figure 3. A histogram representing sampled posterior probability density for area multipliers.

As we can see, all neighborhoods in Taipei have distributions on the right side of neighborhoods in Buenos Aires, meaning grocery prices in Taipei are higher than Buenos Aires due to geographical location reasons.This can also be seen from the mean values for each area multiplier: for Taipei, most multipliers are above 1, meaning that they drives the prices up; while for BA, many multipliers are below 1, meaning that geographical factors bring grocery prices down for BA. This conclusion also matches with the reality as shown in the figure below. (Numbeo, n.d.).

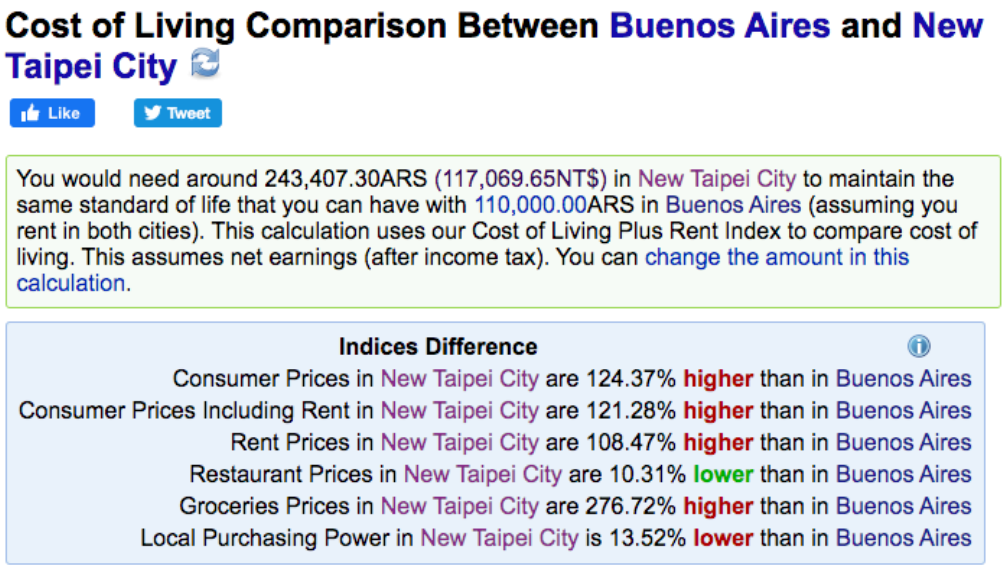

Figure 4. Real Price Index Comparison between Taipei and BA.

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
multiplier_location[1]	1.73	9.8e-3	0.27	1.26	1.71	2.31	737	1.0
multiplier_location[2]	0.91	5.1e-3	0.14	0.66	0.9	1.21	729	1.0
multiplier_location[3]	0.76	4.6e-3	0.12	0.55	0.76	1.03	726	1.0
multiplier_location[4]	0.84	4.9e-3	0.13	0.62	0.84	1.12	676	1.0
multiplier_location[5]	2.06	0.01	0.3	1.53	2.05	2.7	664	1.0
multiplier_location[6]	2.09	0.01	0.31	1.54	2.08	2.75	710	1.0
multiplier_location[7]	1.67	9.5e-3	0.25	1.23	1.65	2.21	681	1.0
multiplier_location[8]	1.84	0.01	0.28	1.33	1.82	2.42	713	1.0
multiplier_location[9]	1.82	0.01	0.27	1.35	1.8	2.4	699	1.0
multiplier_location[10]	0.92	5.2e-3	0.13	0.69	0.92	1.21	643	1.0
multiplier_location[11]	0.88	5.0e-3	0.13	0.65	0.88	1.17	693	1.0
multiplier_location[12]	0.92	5.4e-3	0.14	0.67	0.91	1.22	698	1.0
multiplier_location[13]	0.93	5.6e-3	0.16	0.65	0.92	1.26	788	1.0
multiplier_location[14]	0.9	5.0e-3	0.15	0.63	0.89	1.22	892	1.0

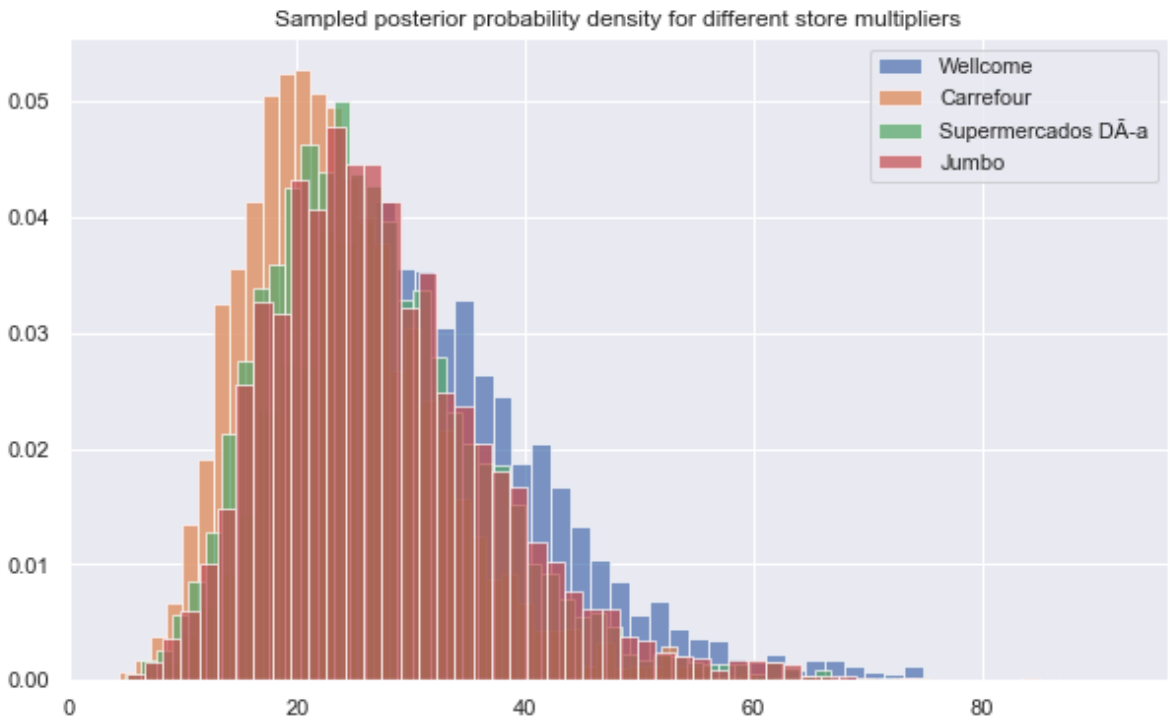

Figure 5. A histogram representing sampled posterior probability density for store multipliers.

Similarly, we also have the distribution for the store multipliers. As we can see, most distributions are overlapped with each other, and it means means that most stores have similar effect on the price (the brand effect is less in both BA and Taipei).

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
multiplier_store[1]	2.27	0.03	0.62	1.3	2.19	3.69	544	1.0
multiplier_store[2]	1.73	0.02	0.46	0.99	1.67	2.79	510	1.0
multiplier_store[3]	1.91	0.02	0.51	1.09	1.86	3.11	523	1.0
multiplier_store[4]	1.98	0.02	0.53	1.13	1.91	3.25	513	1.0

IV. Rental Price and Area Multiplier

As we can observe from the graph above, the rental price and area multiplier are positively correlated, which is understandable, as when the rental prices are higher, store may transfer some of the rental cost to customers by increasing grocery prices.

Reference

Numbeo. (n.d.). *Cost of Living Comparison Between Buenos Aires and New Taipei City.* Retrieved from: https://www.numbeo.com/cost-of-living/compare_cities.jsp?country1=Argentina&country2=Taiwan&city1=Buenos+Aires&city2=New+Taipei+City&tracking=getDispatchComparison