Rita Kurban

Professor Scheffler

CS146

09 November 2018

The Cost of Basic Goods

Metadata:

I went to two EDEKA stores. I visited the first one on November 6 at around 12 pm. The address is Fischerinsel 12, 10179 Berlin. I visited the second store at around 5 pm on the same day. The address is Heidelberger Str. 90, 12435 Berlin:

Pic. 1 and 2. Two EDEKA stores. Fischerinsel is on the left, EDEKA Lawrenz is on the right.

Data Cleaning:

Before building the model, I cleaned the data first in google sheets and then in Python. In google sheets, I removed all the umlaut letters and substituted them with o, a, u letters. I also checked if people used combinations of oe, ae, ue instead of umlauts and also substituted them with single letters. After that, I looked at the brands, found some discrepancies in the writing, and used "replace" command to use the same writing across brands. For example, some people used "Honey crunch" while others wrote "honeycrunch." Finally, I added neighborhoods by transferring them from the initial store assignment sheet.

In Python, I used "lower()" to convert all uppercase characters into lowercase characters to ensure that Python doesn't see them as different instances. I also transferred all the brand names and prices to the bottom of the dataset and added the products as a separate column, so that the remaining columns were 'store,' 'neighborhood,' 'brand,' 'price,' and 'product.' I removed rows that don't have product prices because I couldn't use them in the model. Finally, I used "LabelEncoder()" to convert categorical variables to integers so that stan can work with them freely. I also added 1 to all the numbers since Python counts from 0, while stan counts from 1.

Model:

In my model, each type of product has a base price, with multipliers depending on product brand, store brand, and geographical location. I decided to use the Cauchy distribution as a prior for my base prices since this distribution has heavy tales which makes it possible to sample different base prices. This distribution is informed by two parameters: location and scale. Since most of the prices were around 2 euros, I decided that the location parameter should be 2,

while the scale parameter should be broad enough to make it possible to get diverse values, so I set it to 5. I also restricted the values to positive numbers in the parameters section because the prices cannot go into negatives or be 0. A change in the parameters can make it possible to sample prices not in euros but in other currencies.

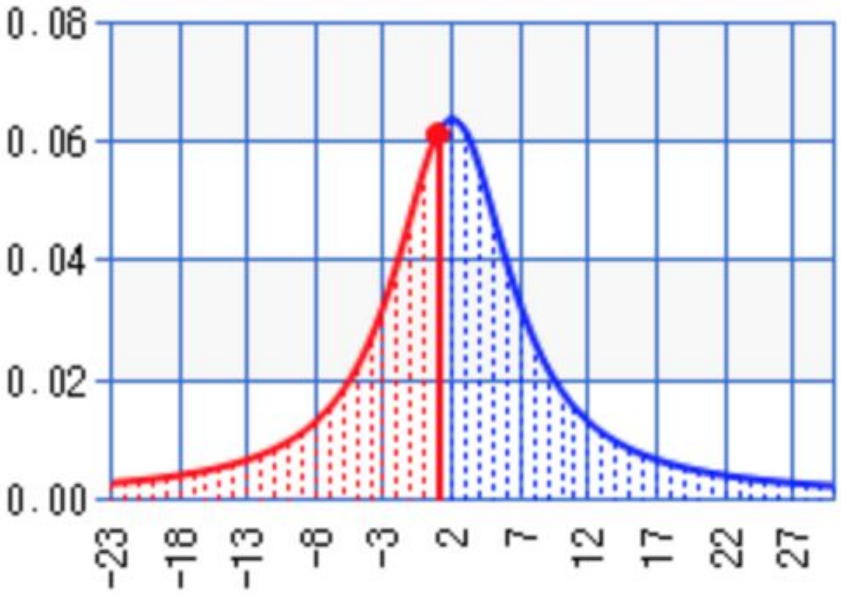

Fig.1. Cauchy prior distribution from which different base prices can be generated.

All brands, neighborhoods, and stores have their multipliers. The multipliers come from a gamma distribution with parameters alpha = 9 and beta = 8. I choose this numbers because the distribution's density mass is centered around 1:

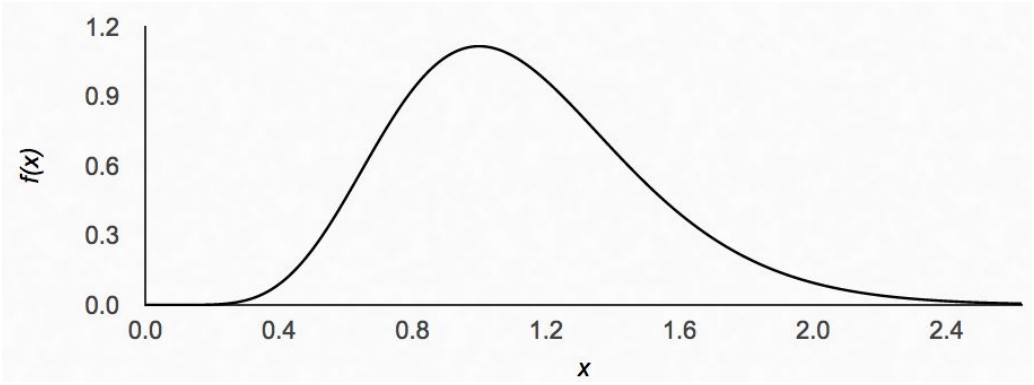

Fig.2. Gamma prior distribution from which different multipliers can be generated.

Final prices are sampled from a normal distribution with sigma = 0.1 to ensure the variance is not too high. The mean is calculated as the product of the base price and three multipliers. Below, you can find a graphical representation of the model:

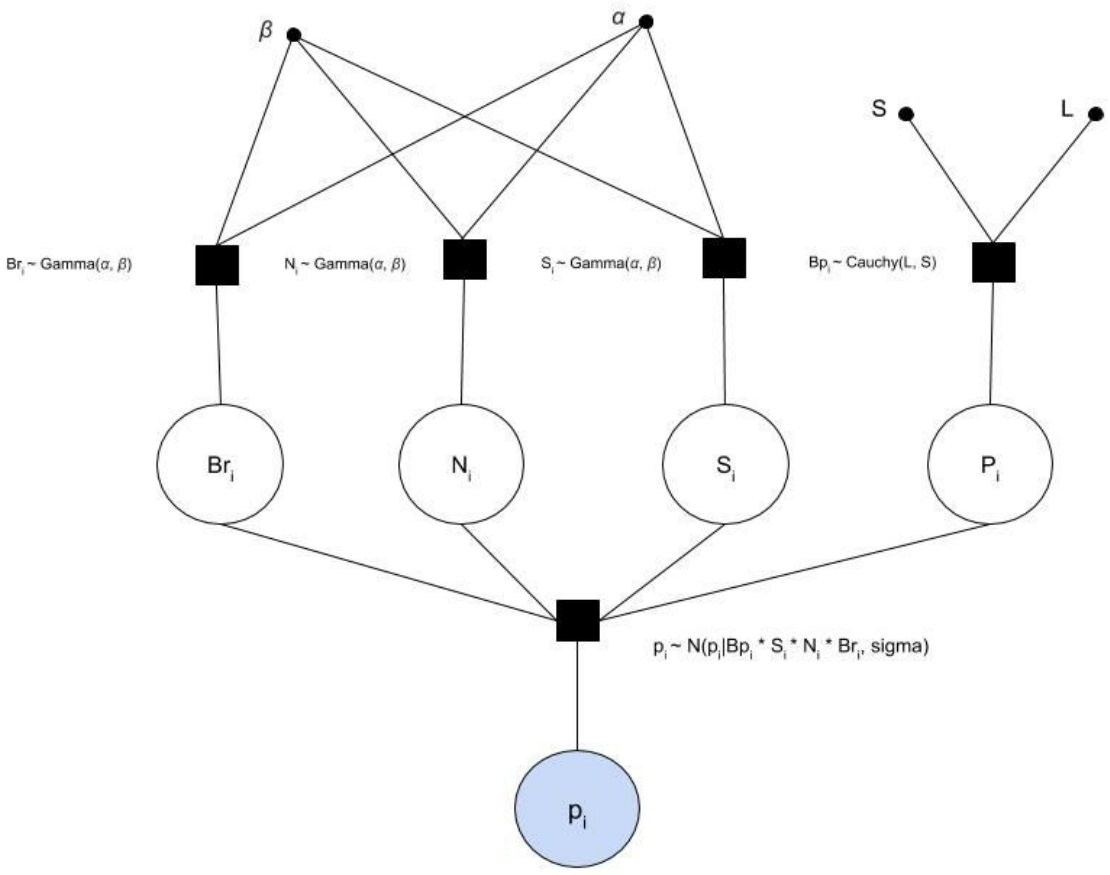

Fig.3. Graphical representation of the model. Br, N, S, P, p stand for "brand", "neighborhood", "store", "product", and "price" respectively.

I decided not to omit the multiplier for "No brands" because products that have no brand tend to be cheaper than those produced by companies that spend money on marketing. Therefore, I expect this multiplier to be pretty small.

My model also assumes that brand multipliers stay the same independent of the products (same for apples and chicken) which might not be 100% true. For example, if a vegetable brand is famous for its apples, the multiplier might be a bit higher for apples than for potatoes. Moreover, the model assumes that all the neighborhoods are independent of each other which is also not true. However, I don't think these small inaccuracies can have a dramatic effect on the final prices.

Results:

Product	Average Price	Base Price
Apples	2.282711	2.381226
Bananas	1.431739	1.505796
Butter	4.077870	4.655343
Chicken	9.845563	16.318597
Eggs	2.574876	3.381883
Flour	1.064358	1.069067
Milk	1.045189	1.048828
Potatoes	1.352609	1.420275
Rice	2.338808	2.412683
Tomatoes	3.436967	3.457224

Table 1. Basic price for each product as estimated by the model compared to the simple average.

As expected, the base prices on each product seem to be similar to average prices because some multipliers have a positive effect, while others affect the price negatively. Chicken is the only exception because its base price is different from the average price. However, this is still the most expensive product. I believe that the number is so high because multiple chicken brands are

really expensive, so they skew the distribution towards larger numbers which means that a simple average might be a less reliable metric.

	Store	Multiplier
0	aldi	1.103302
1	edeka	0.850347
2	lidl	0.811144
3	rewe	0.929646

	Neighborhood	Multiplier
0	alt-treptow	0.964826
1	friedrichshain	1.080026
2	kreuzberg	1.065152
3	lichtenberg	0.954068
4	mitte	1.063220
5	neukoelln	0.847655
6	prenzlauer berg	0.915700
7	schoeneberg	1.094441
8	tempelhof	1.332327

Table 2. Multipliers for stores and neighborhoods in alphabetical order. The higher the multiplier, the bigger its positive effect on price is, multipliers <1 decrease the price.

The tables indicate that the magnitude of the effect of neighborhood on final prices is similar to the effect of the store. Surprisingly, Aldi's multiplier is the largest even though this store is one of the cheapest in Berlin. The reason for that might be that the dataset is biased if, for example, students who went to Aldi recorded the prices of the most expensive products. Alternatively, this store might be more expensive than other stores, but this is unlikely. The rest of the multipliers are distributed as expected with REWE being the most expensive store, followed by EDEKA and Lidl. To further explore why Lidl has the highest coefficient, I plotted the samples for different stores:

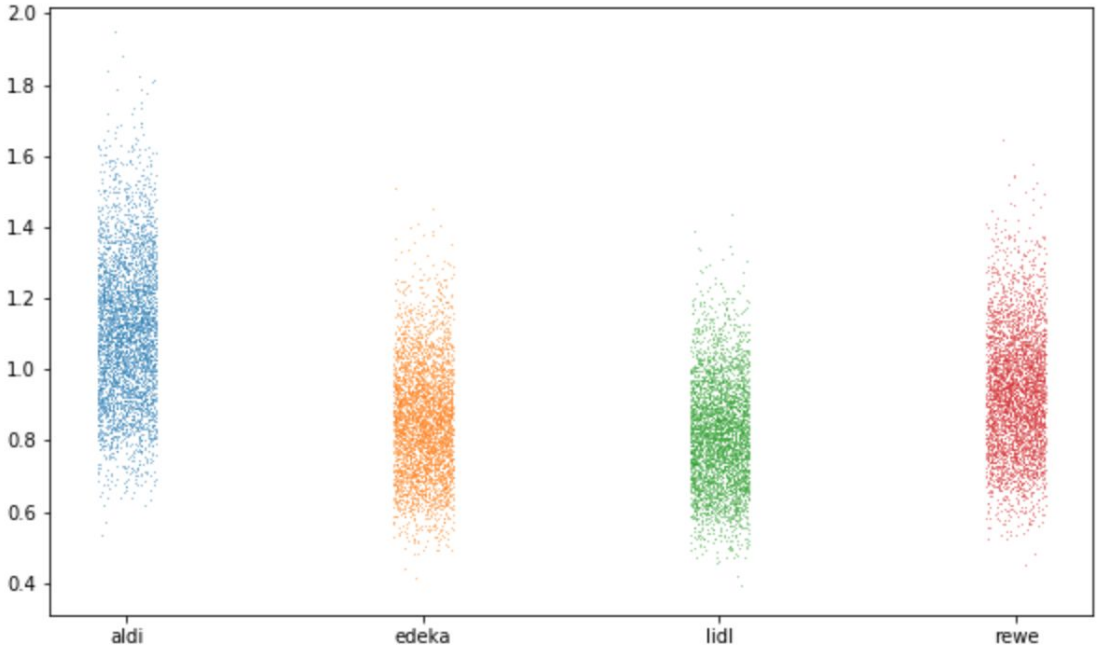

Fig.4. Samples for different stores

The graph corresponds to the multipliers and shows that most of the Aldi samples are indeed in the higher numbers than the samples for other stores.

The tables also indicate that Tempelhof is the most expensive district and Neukoelln is the cheapest. Here are the samples for ten different neighborhoods:

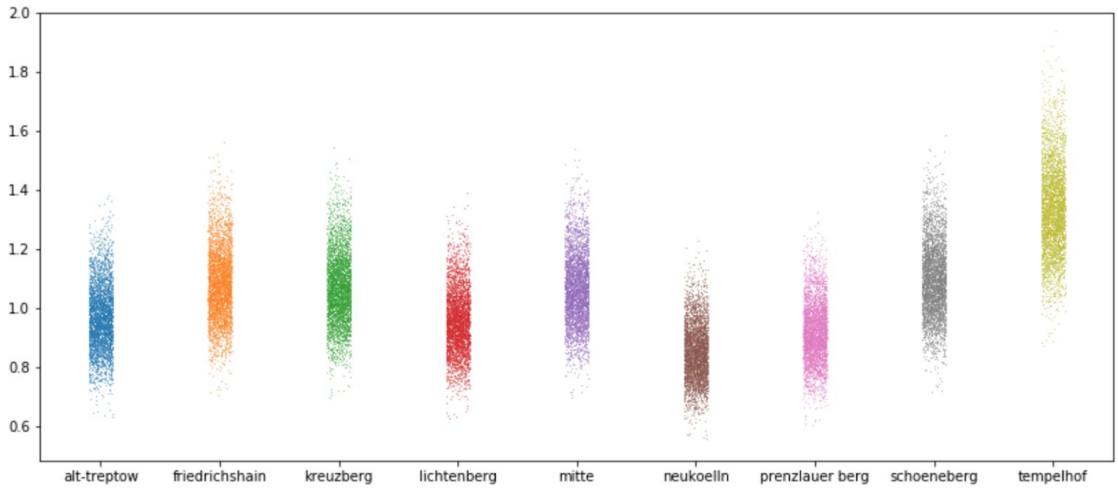

Fig.5. Samples for different neighborhoods

I checked the data and realized that there is only one observation for Tempelhof, and the store that corresponds to this neighborhood is Aldi. Therefore, this one observation might be unreliable and skew the multipliers to the higher numbers.

I also explored whether a price variation by geographical location correlates with variation in rental prices in Berlin. To do so, I calculated average rent prices for districts from the model. After that, I created a scatter plot to find out whether the data points are correlated. The plot indicates that there is some sort of correlation except Schoenberg that seems to be pretty far off:

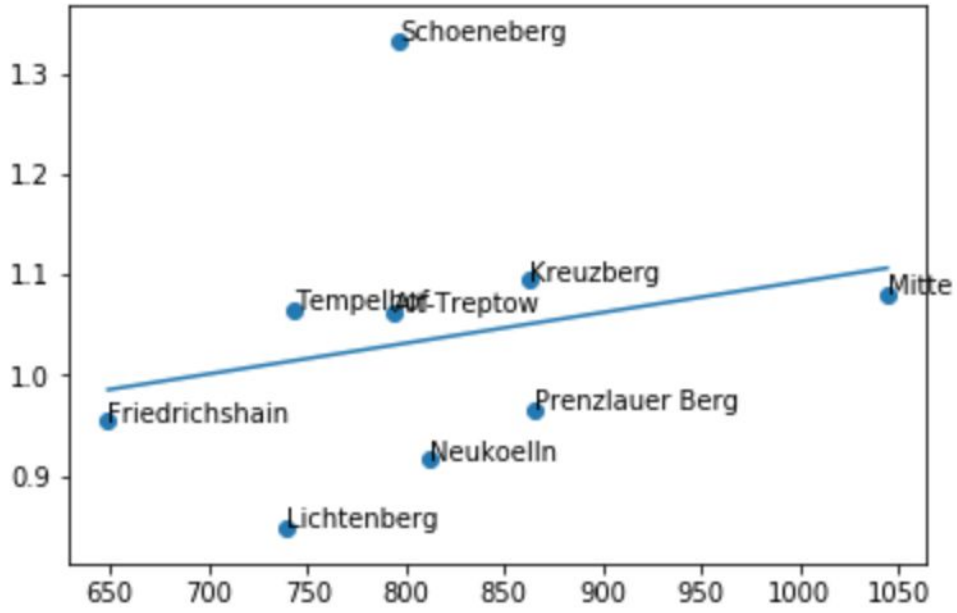

Fig.5. Correlation of price variation by geographical location with variation in rental prices in Berlin.

To quantify the results, I also calculated the Pearson correlation. It indicated a mild correlation of 24%. After excluding Schoeneberg, the coefficient increased to 47% which suggests a higher but also not very strong relationship.

The last multiplier was calculated for different brands. Since the number of brands is much larger than the number of stores and neighborhoods, I didn't print the entire table, only some major statistics. Even though most of the values are centered around 1, the range of this multiplier is much higher than that of the other ones (from 0 to more than 8.) The distribution of mean values is illustrated by the following graph:

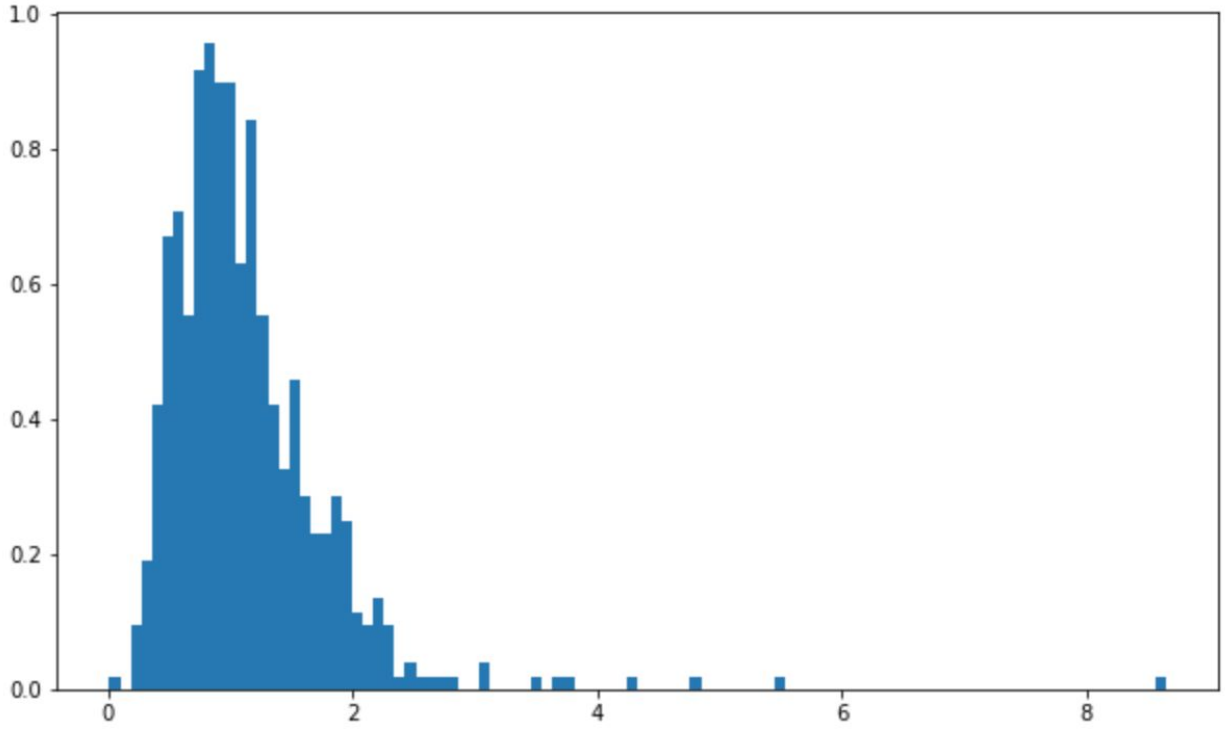

Fig. 6. Sampled Posterior Probability Density for Brands

```
Minimum Value of the Multiplier =  0.01432202233766675
Mean Value of the Multiplier =  1.1243190437063384
Maximum Value of the Multiplier =  8.641482011793387
```

To conclude, brands have a higher effect on prices than stores and neighborhoods. Stores and neighborhoods, in turn, have similar mild effects on the end price of the products. I also indicated a mild correlation between rent prices and grocery prices in different neighborhoods.