MINERVA SCHOOLS AT K.G.I.
LBA
Quang Tran
CS146 Fall 2019

Meta data

Lidl, Zeughofstraße 23 A

At 3.30 pm 11/09/2019

ALDI, Gerichtstraße 2-3
At 2pm 11/09/2019

Model

The unknowns:

- The base price for each of item types (apples, bananas, etc.) is unknown and is assumed to come from an exponential prior with (lambda=0.0000001).

$$base\ price\ \sim Exponential(\lambda = 1e^{-6})$$

The lambda is that small because the base price is just a scale parameter and we do not know the currency used. Small lambda corresponds to large variance in exponential distribution ($\sigma^2 = 1/\lambda^2$). The exponential also has the support of positive real numbers, which is how base price is constraint.

- The location location scaler is unknown and is assumed to come from a gamma distribution:

$$location\ multiplier\ \sim gamma(8.55712644,\ 7.55717524)$$

Gamma distribution is chosen also because the support is positive real numbers, which is how the location multiplier is constraint. The parameters of this distribution (alpha= $8.55712644,\ beta = 7.55717524$) are because:

- We want the distribution to center around 1, which requires the mode to be around 1.

- I also believe that most of the time, no location has a multiplier lower than 0.5 than the average, and no location has a multiplier higher than 2 than the average, so I chose the parameters (by optimization) such that 95% confidence interval is approximately [0.5; 2].

I am aware that because we have only two unknowns but three equations (one for the mode and two for the confidence interval using the cdf of gamma), the optimization results will surely result in only a very approximate answers that try to match these three criteria (i.e., analytical solutions do not exist.)

- The store brand scaler is unknown and is assumed to come from also a gamma distribution.

$$brand\ multiplier\ \sim gamma(8.55712644,\ 7.55717524)$$

The reason for this distribution and its parameters is the same as the case above.

- Beta is unknown and is assumed to come from a broad ($\lambda = 1e^{-4}$) exponential distribution:

$$\beta \sim Exponential(\lambda = 1e^{-4})$$

Likelihood:

Then the price (the data) comes from a gamma distribution with mean equal the the product of base price, brand multiplier, and location multiplier. I set alpha equals the product times beta.

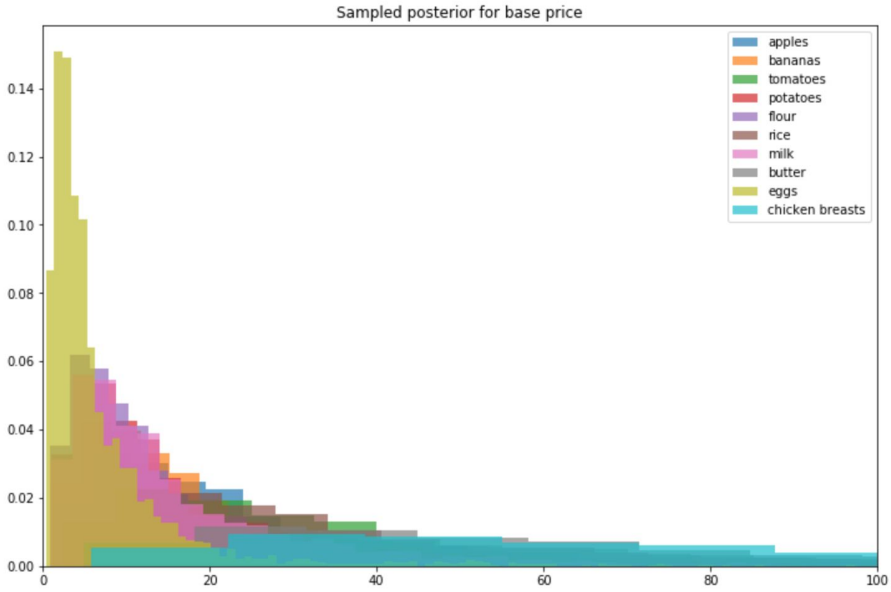
$$price\ \sim Gamma(base\ price\ *\ brand\ multiplier\ *\ location\ multiplier\ *\ \beta,\ \beta)$$

Assumptions and how they are reflected in the model

1. We don't know which currency used. This is reflected in the broad prior over base price.
2. We treat base prices independently, i.e., knowing data for the base price of one product does not tell us anything about base prices of any other products. This means if a new product comes, we need to sample from the predictive fixed prior again and cannot make use of the data for old products.
3. We make the similar assumption about independence about the brand and the location multipliers but the assumption on the location multiplier is less realistic: usually knowing information about some locations tell us something about that about other locations. After all, all locations are in Berlin and geographically connected/ close to each other! All of these (assumption 2) and 3)) are reflected by the fixed priors.
4. The beta of the gamma likelihood means that we assume that there is a same common variance underlying for all combinations of 1) item types (bananas, apples, etc.), 2) neighborhoods, and 3) store brands.

Results

- **Base prices**



Item	Mean base price	STD	98% confidence interval
Apples	39.78	99.11	3.16; 264.8
Bananas	30.22	76.08	2.35; 203.01
Tomatoes	61.97	157.32	4.92; 413.37
Potatoes	22.76	57.66	1.8; 150.22
Flour	20.7	52.42	1.61; 140.26
Rice	56.63	142.47	4.47; 379.15
Milk	22.77	58.08	1.79; 150.22
Butter	113.3	285.68	8.86; 760.36
Eggs	8.43	20.85	0.66; 57.09

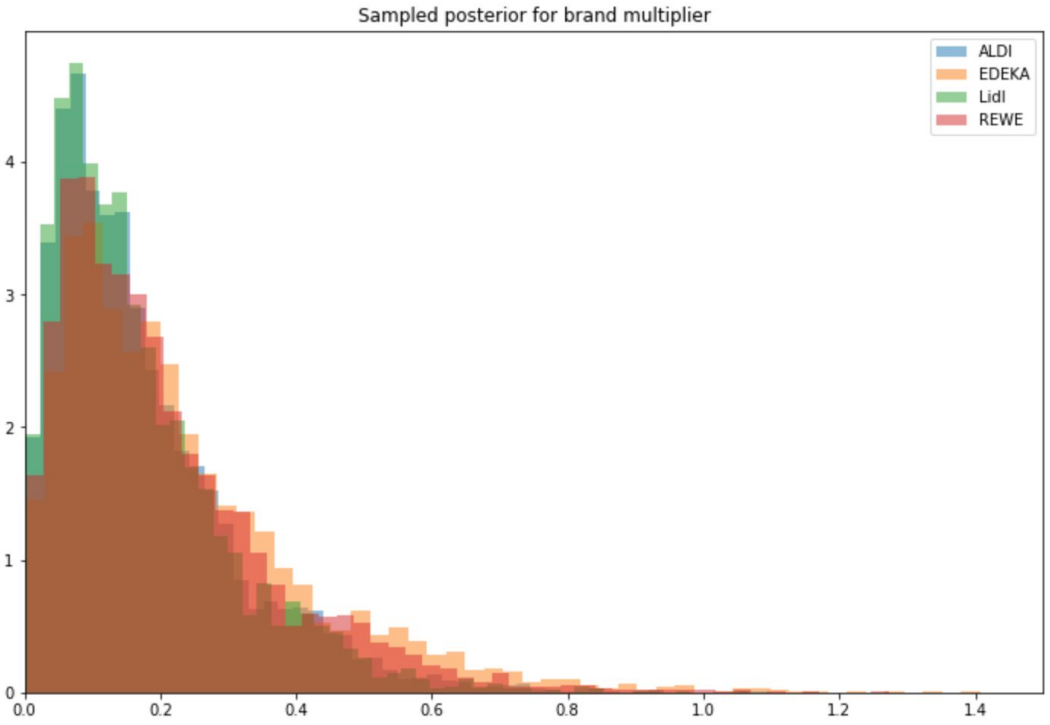
Chicken breasts	138.86	353.0	10.83; 903.16

We see that the lowest average base price is for eggs (8.43) and the highest is for chicken breasts (138.86), which makes sense. There is the most variance around the average for chicken breasts with the widest confidence interval, which is reasonable given how little data for this item is and the large variance within the original data points for this category.

The numerical value for the average base prices does not match what we would usually observe in the original data in Euros because the currency used is unknown, and the base price is just a scaling parameter and should be detached from any currency used when interpreted. Therefore, this parameter is mostly for comparisons among different item types (apples, bananas, etc.) rather for getting a sense of how much that item is in any particular currency.

- **Store brand multiplier**

	mean	se_mean	sd	1%	99%	n_eff	Rhat
brand_multip[1]	0.18	5.9e-3	0.14	0.01	0.64	542	1.01
brand_multip[2]	0.23	7.8e-3	0.18	0.01	0.84	544	1.01
brand_multip[3]	0.17	5.8e-3	0.13	0.01	0.63	540	1.01
brand_multip[4]	0.21	6.9e-3	0.16	0.01	0.75	543	1.01



0.01 0.64 542 1.01

brand_multip[2] 0.23 7.8e-3 0.18 0.01 0.84 544 1.01

brand_multip[3] 0.17 5.8e-3 0.13 0.01 0.63 540 1.01

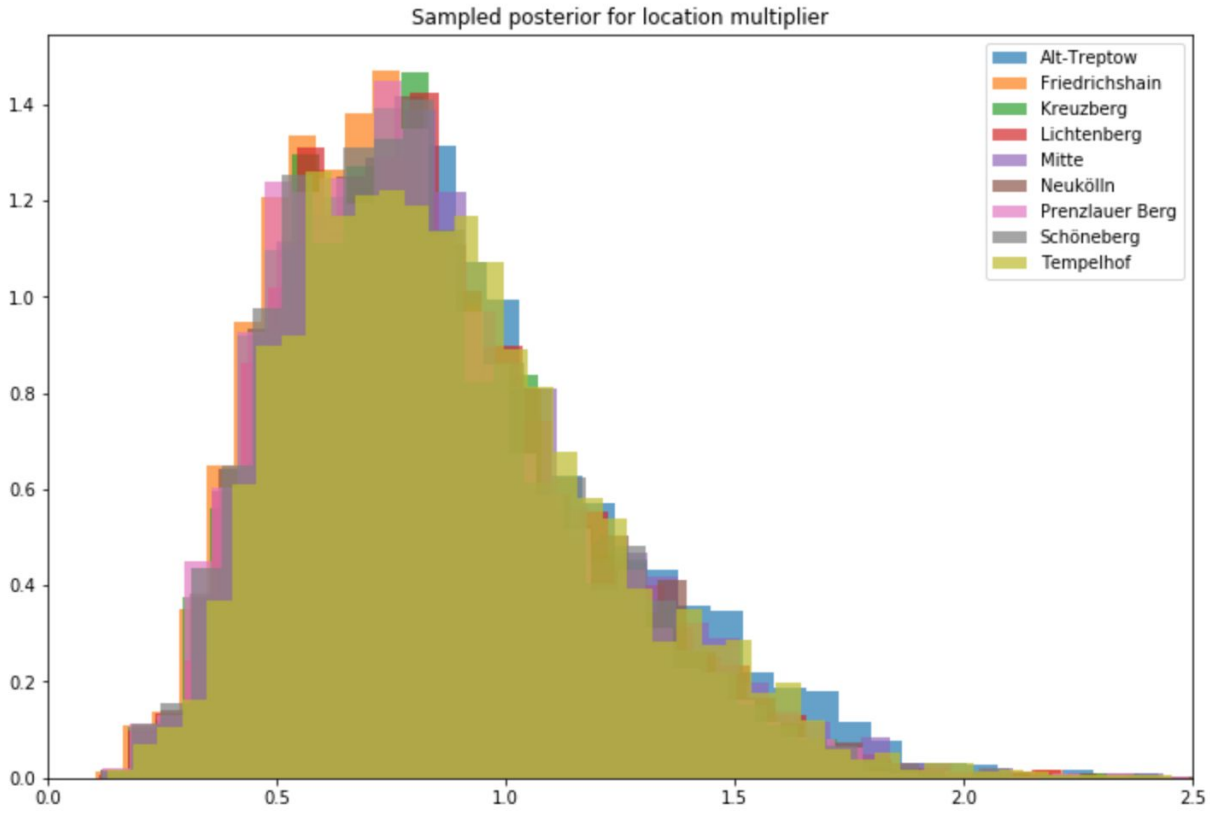
brand_multip[4] 0.21 6.9e-3 0.16 0.01 0.75

Store	Mean	STD	98% Confidence interval
Aldi	0.18	5.9e-3	0.01; 0.64
EDEKA	0.23	7.8e-3	0.01; 0.84
Lidl	0.17	5.8e-3	0.01; 0.63
REWE	0.21	6.9e-3	0.01 0.75

EDEKA has the highest average multiplier, meaning that it's the most expensive store. The least expensive store is Lidl, which makes sense for a discounted supermarket. There is largest variance for EDEKA, which corresponds to least consistency around prices in different EDEKA store.

Again, we should not interpret the average values for the four store, which are all less than 1 (0.18, 0.23, 0.17, and 0.21), as having the effect of lowering the price for items. In fact, if we assume these are the only four store brands in Berlin, the mean of the averages is $(0.18 + 0.23 + 0.17 + 0.21)/4 = 0.1975,$ so ALDI and Lidl, whose means are lower than 0.1975, have the effect of making the items less expensive, and EDEKA and REWE generally elevate the price.

- **Location multiplier**

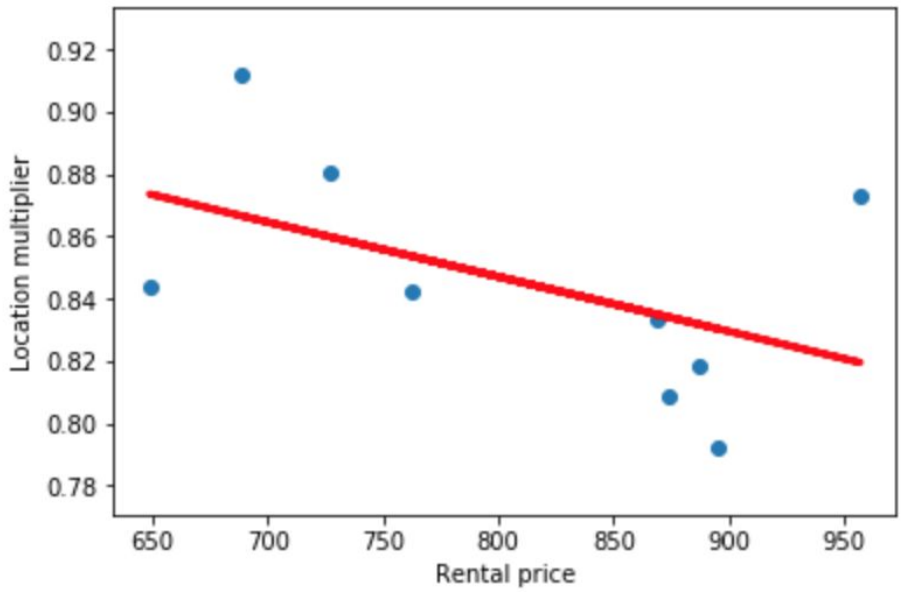


Neighborhood	Mean	STD	98% confidence interval
Alt-Treptow	0.91	0.36	0.29 1.92
Friedrichstein	0.79	0.32	0.25 1.68
Kreuzberg	0.82	0.33	0.25 1.71
Lichtenberg	0.84	0.33	0.26 1.8
Mitte	0.87	0.35	0.27 1.83
Neukollin	0.84	0.34	0.26 1.76
Prenzlauer	0.81	0.32	0.25 1.71
Schonerberg	0.83	0.33	0.26 1.77
Tempelhof	0.88	0.34	0.29 1.84

Friedrichstein has the greatest influence over the price in terms of making the price less expensive because its average value is the lowest (0.79). Alt-Treptow has the effect of highering the price the most with mean 0.91.

- **Correlation between location multipliers and rental prices**

For this exercise, I eyeballed the Berlin neighborhood map and the rental price map given in the assignment and try to visually map the rental prices to the corresponding neighborhoods. For each neighborhood, I computed the average rental price and use that in my scatter plot with the mean of the posteriors over location multipliers.



We see a negative slope of the line of best fit with a negative coefficient (-0.00017553) meaning that generally the more expensive the rentals in a neighborhood are, the lower the multiplier for that location. This is unintuitve. I suspect the problem lies within the error-prone method I used to map the rental prices with the neighborhoods.