CS146-LBA

Katja Della Libera

October 2020

Report Summary

The following report investigates the effect of country location as well as store brand rating (budget to luxury) on the price of common food staples. An effect of both variables was detected.

In addition, the data showed variation in the distribution of prices for different goods, and some effect of GDP on the coefficient of each country.

A second model fit the rent price rather than the country and yielded worse results, suggesting the country is a better predictor of the price of these goods.

The conclusion is limited by small sample size and further investigation is needed.

Contents

1 The Data[1]

1.1 Collection

The data was collected in October 2020 by undergraduate students in seven countries around the world (see second column in table 1. Each student visited two supermarkets to collect price data on ten goods, taking three samples where possible. In addition, each student assigned the supermarket a brand perception rating from the three categories in the third column of table 1 and researched an estimate of rent prices in the neighborhood of the supermarket.

Product	**countries**	**brand ratings**
Apples	Germany	Budget(cheap)
Bananas	Guatemala	Mid-range
Tomatoes	Morocco	Luxury(expensive)
Potatoes	South Korea	
Flour, white	UK	
Rice, basmati	USA	
Milk, full cream	Vietnam	
Butter		
Eggs		
Chicken breasts		

Table 1: Overview of the possible values for the three categorical variables product, country, and supermarket brand rating.

1.2 Data Processing

1.2.1 Cleaning the data

One of the biggest challenges in the processing of the data set concerned the data for rent prices in the neighborhood of the supermarket. The units varied widely from price per square meter to price for one room or price for three rooms. To make the data set more usable, I only used data from the four less frequent countries in the data set (Vietnam, Guatemala, Morocco, South Korea) and the less frequent cities in the common countries (Konstanz, Riverside, Las Vegas) directly. In this case, I assume the data collected here is a reasonable approximation for the cost of housing in those cities for the typical inhabitant. Where necessary, I took the mean of a range that was given.

For the remaining three cities, Berlin, London, and San Francisco, I standardized the data set by using the following three sources:

1. **San Francisco**: https://www.zumper.com/blog/map-san-francisco-neighborhood-rent-prices-fall-2018/
2. **Berlin**: https://www.immobilienscout24.de/content/dam/is24/ibw/dokumente/mietmap_berlin_2020.jpg
3. **London**: https://maps.london.gov.uk/rents/

[1] #organization: My report is easy to follow because of the division into manageable pieces. The table of contents further allows readers to quickly find the desired section.

1.2.2 Standardizing units

Since the data was collected in different currencies and the items are sold at different quantities, the amount of each good and its currency is standardized.

The currency used throughout the modeling is Euros (EUR), and the unit for each item is in a metric scale, 1kg for everything except for milk (1l) and eggs (10count).

2 Model 1: brand and country

2.1 The Model

2.1.1 Variables

The model approximates the average cost of a good based on three categorical variables: the good itself (e.g. banana or apple), the country in which the supermarket is located (e.g. Germany or Guatemala) and the brand perception of the supermarket brand (one of three categories). Each of these is independent and observed for each of our data points.

The dependent variable of the mathematical relationship described in equation 1 is the price of the good itself, which is observed.

We use that to find the parameters representing the average price of each good across the world $p_{product}$, the coefficient adjusting for the country $c_{country}$, and the coefficient adjusting for the brand perception b_{brand}.

2.1.2 Mathematical expression of the model

The fundamental relationship between the variables we want to arrive at is:

$$price \sim Normal(p_{product} * c_{country} * b_{brand}, \sigma_{product}) \tag{1}$$

where p = the average price of a product, c = the coefficient associated with the country, b = the coefficient associated with the brand perception. The subscripts indicate a separate value for each product/country/brand. So we will have $p_{Germany}$, $p_{Vietnam}$, etc.

Because there are seven different countries and three different possible store brand perceptions, the model will fit for all ten of these parameters plus the average price p for each of our ten goods.

2.1.3 Assumptions

Programmed into the model are several assumptions, in particular the prior distributions for the three sets of coefficients, average prices p, country coefficients c, and brand coefficients b.

As stated above, all of these coefficients should be positive so it makes sense to restrict the value to the positive reals.

In addition, it is desirable to keep the mean of the country and brand coefficients close to 1 so the average price reflects the actual average price.

Therefore, I decided on a log-normal distribution for all three priors, as it has the desired domain and the size of the tail and mean are easily adjusted via constant inputs to my model.

To make sure the mean of the coefficients is around 1, I used the equation for the mean of the log-normal distribution (equation 2) to solve for input constant μ given a value for standard deviation σ (equation 5).[2]

[2] #distributions: I use my knowledge of different distributions to chose an appropriate prior for my parameters. In addition, I use the equations associated with the log-normal distribution to achieve desired traits.

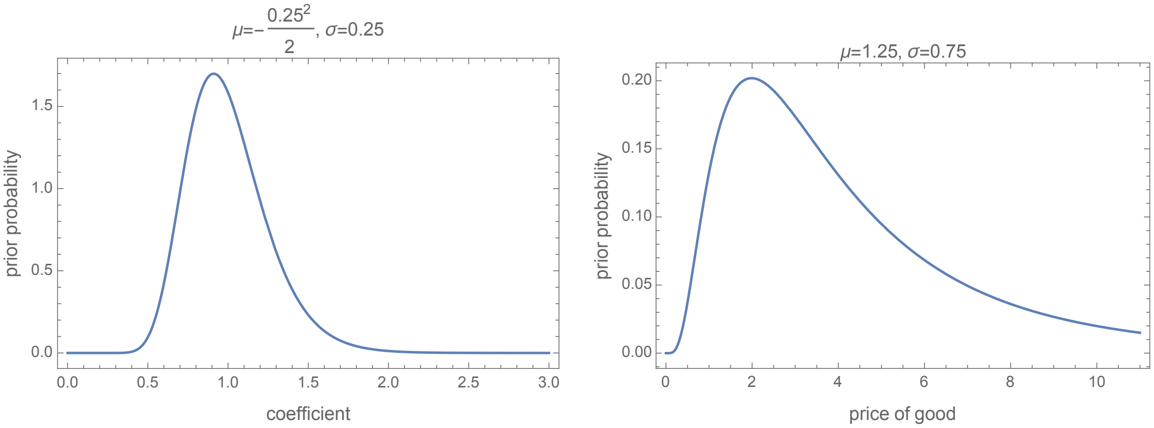
$$mean = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{2}$$

$$1 = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{3}$$

$$\log(1) = \mu + \frac{\sigma^2}{2} \tag{4}$$

$$\mu = -\frac{\sigma^2}{2} \tag{5}$$

Figure 1 shows the two prior distributions, both log-normals with different initialization parameters. the coefficient prior is relatively narrow and with mean 1, the price of goods has a much longer tail.



(a) Prior log-normal distribution for coefficients

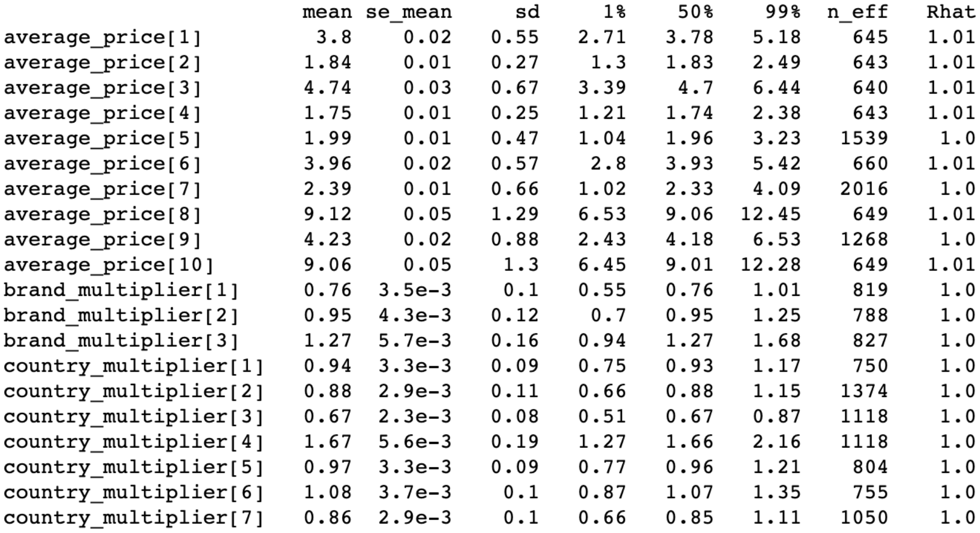
(b) prior log-normal distribution for prices

Figure 1: The log-normal prior for the two coefficients and the average price of the goods

In addition to the priors, we also have to assume any errors in data collection are negligible.

2.2 Results

After fitting for the variables using Bayesian inference in the PyStan environment, we get a table of results as seen in figure 2. For each product there is a price estimate, for each brand a multiplier estimate and for each country a multiplier estimate. Rhat is close to or equal to 1, meaning the model converged for all parameters.



```
                       mean  se_mean      sd      1%     50%     99%   n_eff    Rhat
average_price[1]        3.8     0.02    0.55    2.71    3.78    5.18     645    1.01
average_price[2]       1.84     0.01    0.27     1.3    1.83    2.49     643    1.01
average_price[3]       4.74     0.03    0.67    3.39     4.7    6.44     640    1.01
average_price[4]       1.75     0.01    0.25    1.21    1.74    2.38     643    1.01
average_price[5]       1.99     0.01    0.47    1.04    1.96    3.23    1539     1.0
average_price[6]       3.96     0.02    0.57     2.8    3.93    5.42     660    1.01
average_price[7]       2.39     0.01    0.66    1.02    2.33    4.09    2016     1.0
average_price[8]       9.12     0.05    1.29    6.53    9.06   12.45     649    1.01
average_price[9]       4.23     0.02    0.88    2.43    4.18    6.53    1268     1.0
average_price[10]      9.06     0.05     1.3    6.45    9.01   12.28     649    1.01
brand_multiplier[1]    0.76   3.5e-3     0.1    0.55    0.76    1.01     819     1.0
brand_multiplier[2]    0.95   4.3e-3    0.12     0.7    0.95    1.25     788     1.0
brand_multiplier[3]    1.27   5.7e-3    0.16    0.94    1.27    1.68     827     1.0
country_multiplier[1]  0.94   3.3e-3    0.09    0.75    0.93    1.17     750     1.0
country_multiplier[2]  0.88   2.9e-3    0.11    0.66    0.88    1.15    1374     1.0
country_multiplier[3]  0.67   2.3e-3    0.08    0.51    0.67    0.87    1118     1.0
country_multiplier[4]  1.67   5.6e-3    0.19    1.27    1.66    2.16    1118     1.0
country_multiplier[5]  0.97   3.3e-3    0.09    0.77    0.96    1.21     804     1.0
country_multiplier[6]  1.08   3.7e-3     0.1    0.87    1.07    1.35     755     1.0
country_multiplier[7]  0.86   2.9e-3     0.1    0.66    0.85    1.11    1050     1.0
```

Figure 2: The results given by the model.

A more visual representation of the results is seen in figures 3, 4, and 6.

Starting with the brand multiplier, we can see that the algorithm agrees with what we would expect: the cheapest stores have a multiplier below 1 (0.76), the mid-range multiplier is close to 1 (0.95), and the multiplier for the luxury brands above 1 (1.27). In addition, we see that there is a larger range for the effect of luxury brands.

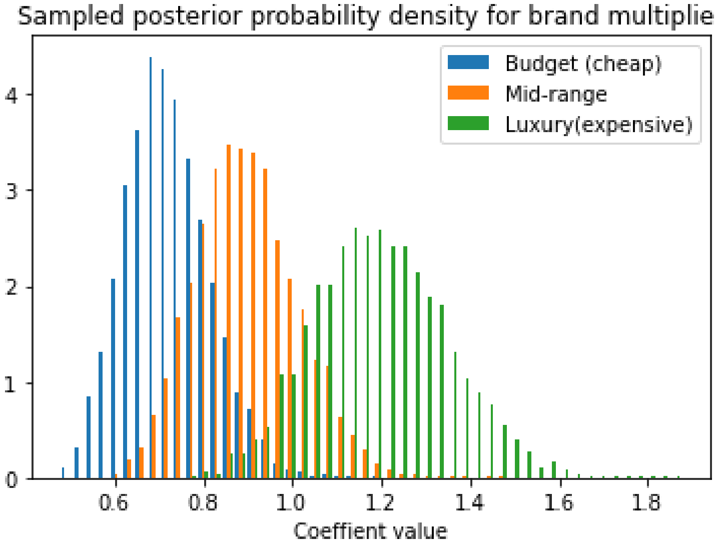
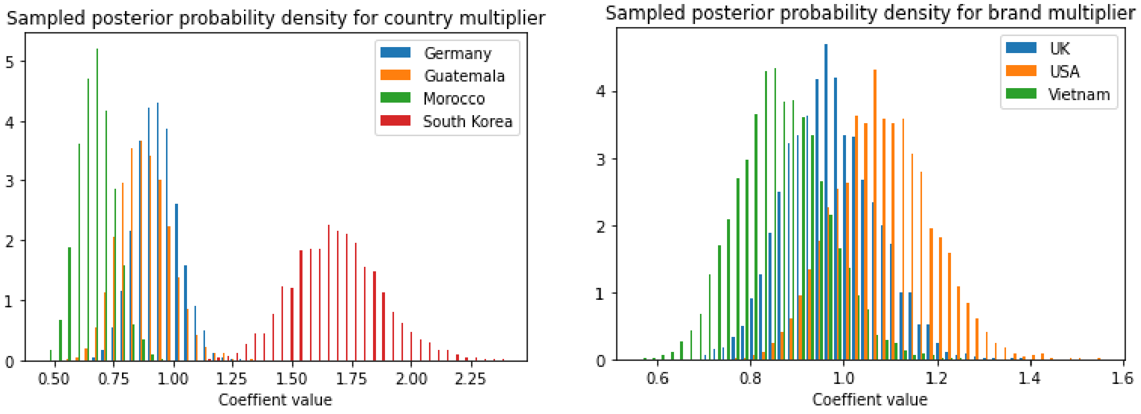


Figure 3: The distribution of brand coefficients shows the three categories of stores are recognizable but there is significant overlap.

Figure 4 shows the effect of the country multiplier. Many of the factors are close to each other and to 1, with Morocco having the smallest around 0.67. The exception is Korea which has a very high multiplier at 1.67. This value appears as even more of an outlier when comparing the relationship between GDP per capita and coefficient determined by our model, showing South Korea has an exceptionally high cost of food. This may partially be driven by the high cost of especially produce in the country, but could also be an artifact of small sample size in the country, with only one supermarket visited.



(a) The posterior distribution over the coefficient of the first set of countries

(b) The posterior distribution over the coefficient of the second set of countries

Figure 4: The distribution over the country multipliers for all seven countries, split into two plots for increased visibility.

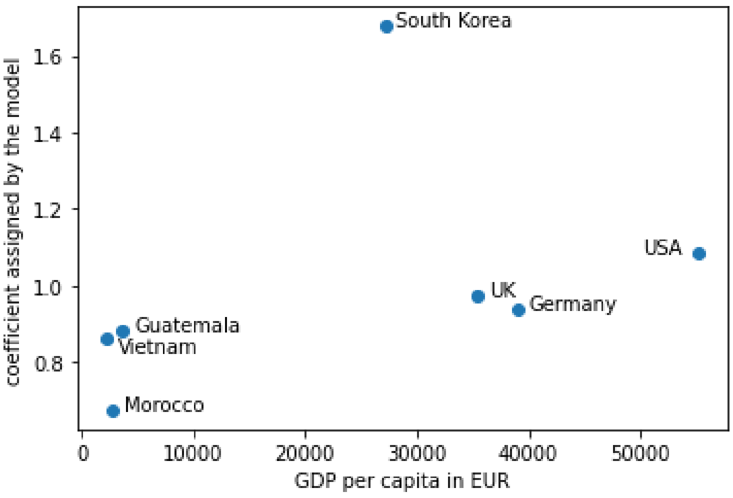
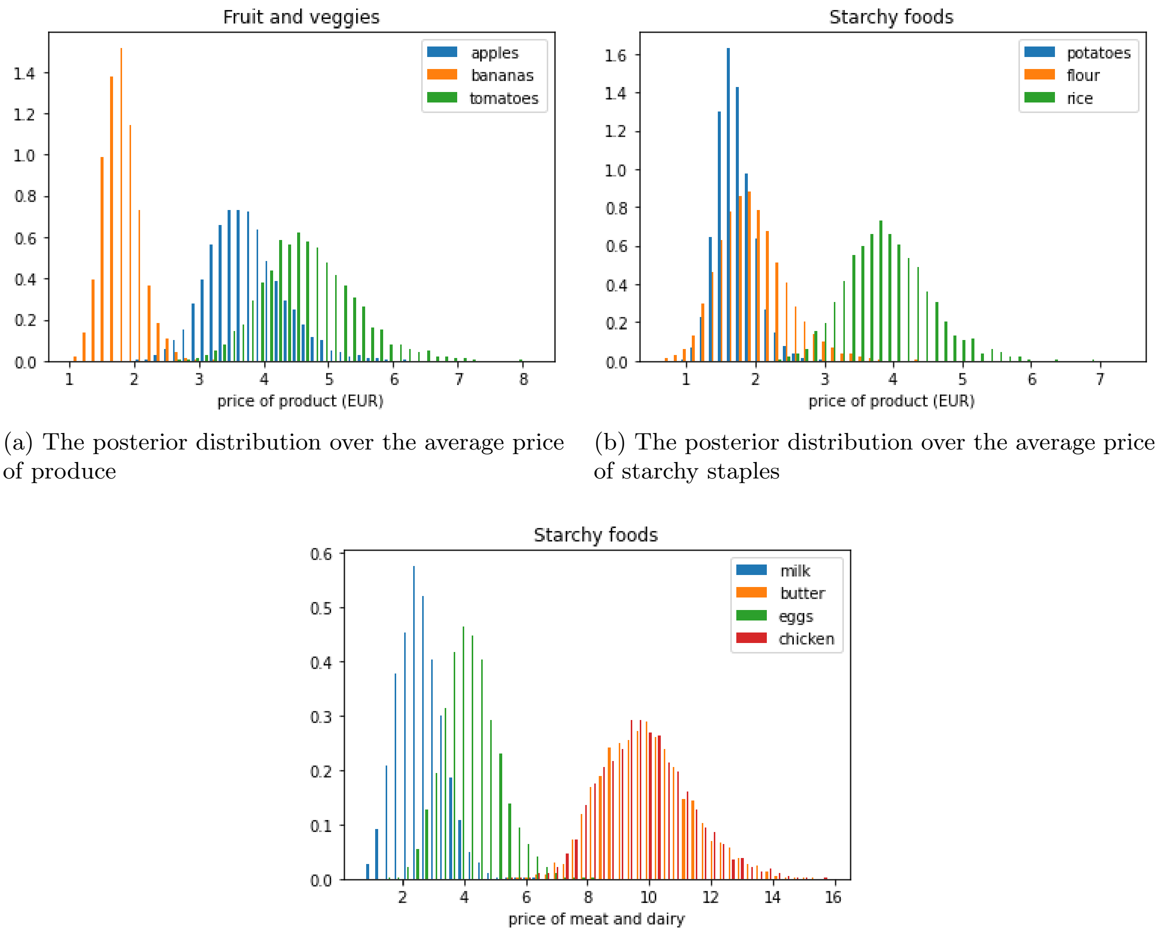


Figure 5: Assuming cost of living is approximately proportional to GDP, South Korea's high coefficient seems especially

Figure 6 shows the distribution over the prices for each of the goods. Some interesting observations here are the variation in the width of these distributions, suggesting some of them are much more consistent in cost than others. For example, the price of bananas is much more consistent than the price of tomatoes.



(a) The posterior distribution over the average price of produce

(b) The posterior distribution over the average price of starchy staples

(c) The posterior distribution over the average price of meat and dairy products

Figure 6: The distribution over the average price of each product, divided into three plots for better visibility

In summary, the product type has the largest impact on the price, followed by the country and brand coefficients which are similar in order of magnitude and range.

3 Model 2: brand and rent

3.1 The Model

3.1.1 Variables

The variables used in the rent variation of our model are the same product and brand variables as described in section 2.1.1. However, instead of a categorical variable for the country, each data point is assigned a value corresponding to the rent in EUR around the supermarket. This variable replaces the country variable in our model.

3.1.2 Mathematical expression of the model

As in our first model, we are predicting the price of a product based on the supermarket brand perception, and an average price. However, instead of a country coefficient, we integrate the rent price into the model.

The new mathematical relationship is given by equation 3.1.2. Instead of the country coefficient for the country of the product we fit a linear regression with slope m and intercept b to the rent parameter.

$$price \sim Normal(p_{product} * b_{brand}(*m * rent_{product} + b), \sigma_{product} \tag{6}$$

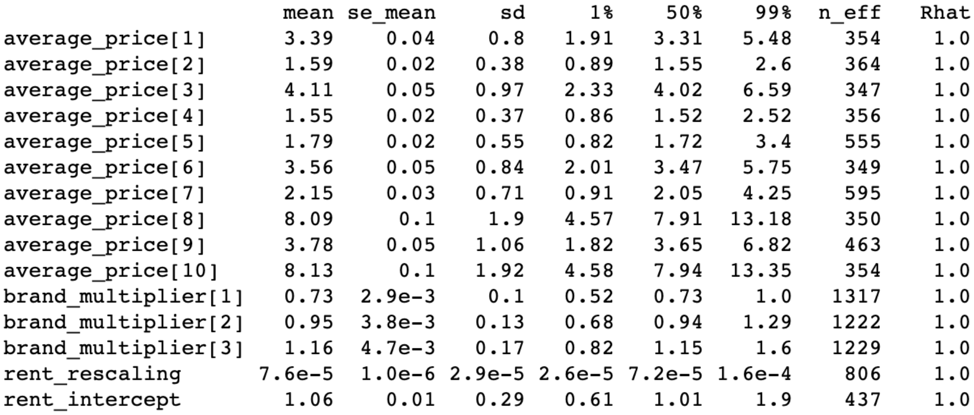
The model will still return an average price for each product and a coefficient for the brand rating, but instead of 7 coefficients for the 7 countries, we will get one slope and one intercept for the rent variable.

3.1.3 Assumptions

The assumptions made in the original model still hold.

3.2 Results

The effect of the rent parameter is only minimally evident in the model. The slope is only around $7*10^{-}5$, meaning for every 1000EUR rent increase, the model predicts only 7 cents of price increase. The table in Figure 7 shows the exact results of our second model, which also converges as seen in Rhat values. However, the effect of the rent variable is not as good a fit as the country coefficient, increasing the overall uncertainty in the model. Figure 8 shows the distribution over the posterior of the linear regression parameters m and b. The slope is predicted to be very small and positive.[3]

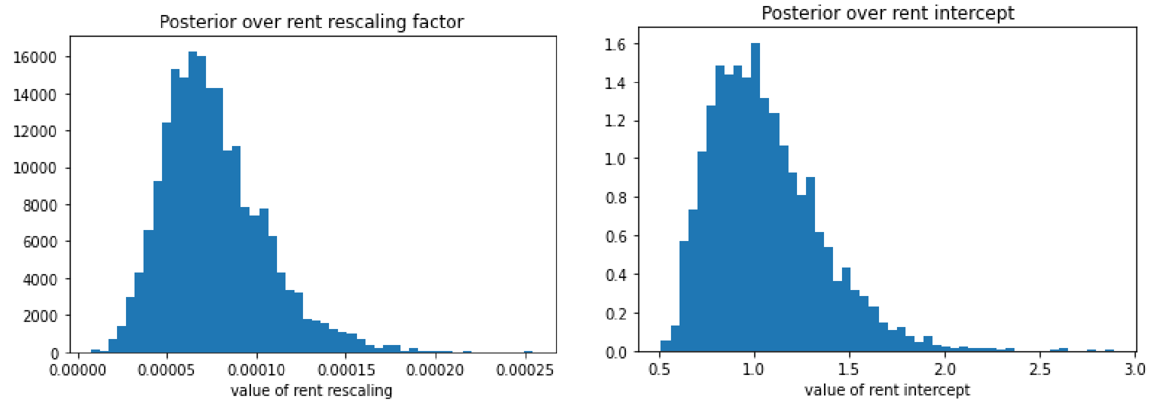


```
                       mean se_mean     sd     1%    50%    99%  n_eff   Rhat
average_price[1]       3.39    0.04    0.8   1.91   3.31   5.48    354    1.0
average_price[2]       1.59    0.02   0.38   0.89   1.55    2.6    364    1.0
average_price[3]       4.11    0.05   0.97   2.33   4.02   6.59    347    1.0
average_price[4]       1.55    0.02   0.37   0.86   1.52   2.52    356    1.0
average_price[5]       1.79    0.02   0.55   0.82   1.72    3.4    555    1.0
average_price[6]       3.56    0.05   0.84   2.01   3.47   5.75    349    1.0
average_price[7]       2.15    0.03   0.71   0.91   2.05   4.25    595    1.0
average_price[8]       8.09     0.1    1.9   4.57   7.91  13.18    350    1.0
average_price[9]       3.78    0.05   1.06   1.82   3.65   6.82    463    1.0
average_price[10]      8.13     0.1   1.92   4.58   7.94  13.35    354    1.0
brand_multiplier[1]    0.73  2.9e-3    0.1   0.52   0.73    1.0   1317    1.0
brand_multiplier[2]    0.95  3.8e-3   0.13   0.68   0.94   1.29   1222    1.0
brand_multiplier[3]    1.16  4.7e-3   0.17   0.82   1.15    1.6   1229    1.0
rent_rescaling       7.6e-5  1.0e-6 2.9e-5 2.6e-5 7.2e-5 1.6e-4    806    1.0
rent_intercept         1.06    0.01   0.29   0.61   1.01    1.9    437    1.0
```

Figure 7: The results of the rent model show a relatively small effect of rent and increased uncertainty in other terms as a result

Figure 9 shows the data from which this model is generated, which suffers from several outliers (likely due to human error) but otherwise very narrowly distributed prices regardless of rent price. The values in this plot were not adjusted for brand perception or product type. Part (b) of the figure shows two example products, which confirm the overall trend.

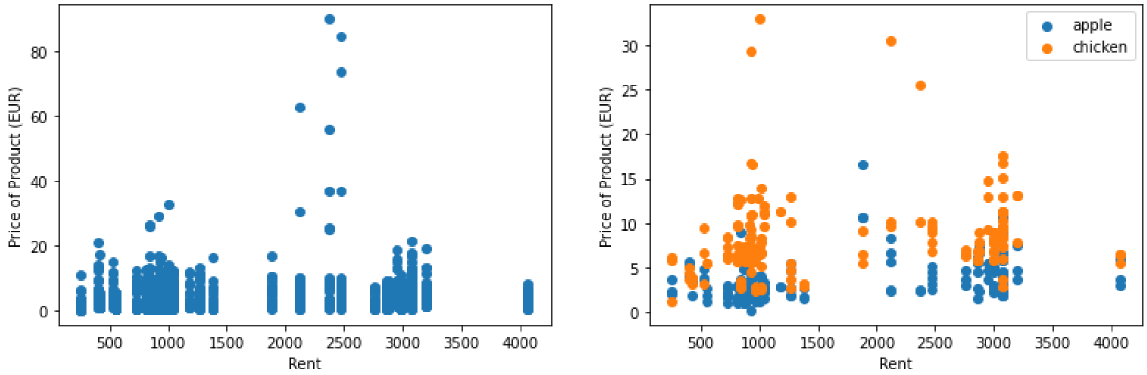
[3] #regression: I draw on the concepts from Formal Analysis and the concepts learned in this course to relate the product price with the rent parameter.



(a) The posterior distribution over the slope of the rent linear regression

(b) The posterior distribution over the intercept of the rent linear regression

Figure 8: The rent model predicts only a very small effect of rent prices.



(a) rent vs product price for all products

(b) rent vs product price for two example products

Figure 9: The bad fit for the rent model is consistent with the data set.

Because rent seems to be a worse predictor of price than the country, the brand perception is given slightly more significance in the second model. Since the model has less to work with, we also expect the uncertainty in all outputs to go up. Figure 10 shows this effect on the price of apples.

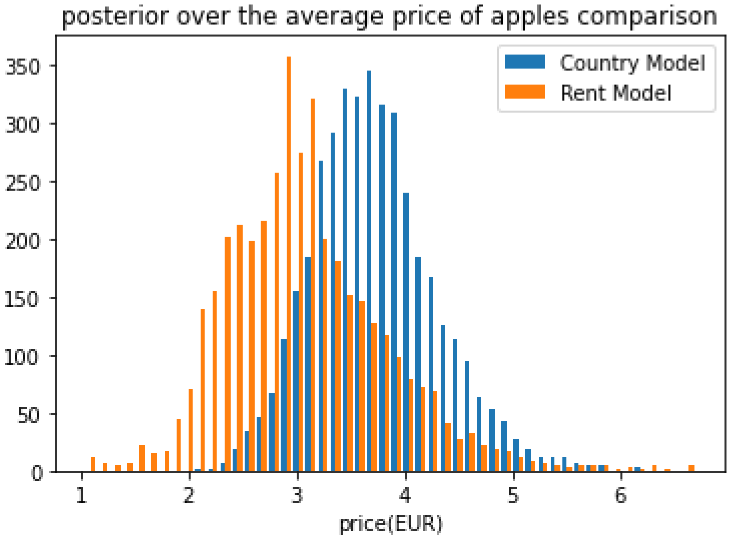


Figure 10: The uncertainty over the price of the products increases since rent is a worse predictor than country