The Cost of Basic Goods

CS146 Location-Based Assignment

Introduction

In this assignment, we model the cost of groceries in different parts of the world. To what extent do grocery prices vary by country and store brand? Are grocery prices and the geographical distribution of different grocery stores correlated with other cost-of-living measures – for example, rent and real estate prices?

Carefully read the entire assignment description at least once. Make sure you have a complete list of all requirements and deliverables before you start gathering data or implementing your model to avoid missing important details when heading into the city or writing your report.

Data collection

You have each been assigned (or chosen) supermarkets to visit in your city.

COVID-19 considerations

- If it is safe to do so, visit your two assigned/chosen grocery stores. If you are selecting your own stores, you should choose stores with a reasonably wide product range so you are likely to find all the products listed below.
- If it is not safe to go out, use whichever online grocery store you use to do your grocery shopping. Look up the prices online rather than visiting the store. You need to record the same information as everyone else but should indicate (in the Address field) what the URL of the online store is rather than its physical address.

Grocery price data

Your first data gathering task is to collect price data from each supermarket for the products listed below. For each product, there might be different brands with different prices. Please record 3 different prices for each product. It is also important to record the quantity of the product so that prices can be normalized. For example, one brand might sell apples in 1 kg bags while another brand might sell apples in 2 kg bags.

Gather price information on all 10 of these items.

- Apples
- Bananas
- Tomatoes
- Potatoes
- Flour, white
- Rice, basmati
- Milk, full cream
- Butter
- Eggs
- Chicken breasts

Rental price data

We want to correlate grocery price data with rental price data. Find a resource online that has information about average rental prices in the neighborhood of your grocery store. Here are some resources to help you out. For other cities, find and use any resource you can. If you find useful resources for this part, feel free to share those with other students.

- A 2016 map of Buenos Aires rental prices, organized by Metro station
- A 2017 map of Berlin rental prices, organized by U-Bahn and S-Bahn station
- An aggregated map of rental price ranges in London

Submitting your data

- The deadline for submitting your data is Friday, October 29.
- Your assigned grocery stores for data collection are in this spreadsheet. If you are not in Berlin or London, you will need to choose your own supermarkets. Pick your nearest/favorite supermarkets that have a good range of products.
- Data that have already been submitted are automatically collected in this spreadsheet.
- Enter the data you collect using this form.
    - Please complete a new form for each store you visit.
    - If a store does not have 3 different brands for a product, leave the superfluous quantity and price fields blank.
- You have to use everybody's data to build your statistical model and answer the questions below.

Questions to answer

- What is the basic average price for each product? You need to think carefully about how to anchor the basic price for each product since this will depend on the currency used as well as the distribution of prices.
- How much does each of the following factors modify the basic price of the product (up or down)?
    - The geographical location (country) of the grocery store.
    - Brand of the grocery store. Since we are getting data from multiple countries, you will need to specify whether the store brand is considered budget (cheap), mid-range, or luxury (expensive). This should be based on what you think the general public perception of the store brand is.

  Explain in your report how strong each of these effects is. Which has the greatest influence on price variation between shops?
- Does price variation by geographical location correlate with variation in rental prices, or not?

Building a model

You are encouraged to use the model structure described below for this assignment. However, you may modify the model if you have ideas on how to improve it. If you do decide to modify the model, you have to do so in a way that still allows you to address all points in the *Questions to Answer* section above. You should also motivate any changes to the model.

Implement your model in PyStan, generate samples from the posterior, present your posterior results, and use your posteriors to answer the questions provided.

The basic idea of the model is that each type of product (apples, bananas, etc.) has a base price, with multipliers depending on store brand and geographical location.

- The base price of each product.
    - Price is a positive real number.
    - It is up to you to choose a good prior.
- The multiplier for each store type (budget/mid-range/luxury).
    - This is a scale parameter (positive real number).
    - The prior can be centered on 1, to achieve an average multiplier of 1. This would make the base price (approximately) match the average price in mid-range stores.
- The multiplier for the country or state.
    - This is a scale parameter (positive real number).
    - The prior should also be centered on 1. Expensive countries/states will have multipliers above 1 and inexpensive countries will have multipliers below 1.

For example,

- The base price of 1 liter of full cream milk might be 0.70 €.
- ALDI stores are considered inexpensive and might be only 0.9 times as expensive as the average store.
- Stores in Germany might be 1.4 times more expensive than stores in the average country or state.

So as a result, 1 liter of full cream milk in an ALDI store in Berlin should cost about 0.70 € × 0.9 × 1.4 ≈ 0.88 € with some random variation around that value.

Stretch goal (optional)

Instead of having an independently sampled multiplier for each geographical region in the model, use a correlated prior where the correlation coefficient is determined by the distance between geographical locations.

The motivation for this is that we would expect two stores or two neighborhoods that are close together to have similar price multipliers, whereas we expect there to be little correlation between price multipliers in locations that are far apart.

- Modify your model to incorporate a spatially correlated distribution over location multipliers.
- Carefully explain how you modified your model.
- Quantify how much of a difference incorporating correlation makes to your results.

An excellent attempt on the stretch goal will get you a score of 5 on #RightDistribution and possibly on #InterpretingProbabilities if your presentation of and insight into results is also excellent. A very good (but not excellent) attempt will score a 4. It is not possible to score less than a 4 on a stretch goal, but you will get no score if the attempt is not good enough.

Deliverables

Submit a PDF report with your model description and results. You must also include a zip file with all code (in a Jupyter notebook) and data files needed to reproduce your results. Your instructor should be able to run your code and reproduce the same results you got.

PDF report

- Write this as if you are writing a report to a client or a supervisor. The purpose of the report is to explain your statistical model and present your results. Make it professional.
- Describe your statistical model. Be clear and detailed about the variables in your model and what quantities they represent, as well as how your variables are related mathematically in your model. Outline any assumptions you made and how they are incorporated into your statistical model.
- Present and explain your results in detail. Address all questions posed in this assignment.
- Do not simply make a PDF out of your Python notebook. A client or supervisor will not read through pages of code to get to your explanation of the model and the results.
- A good report length is approximately 10 pages – much more or less than this means you are either not providing enough information, or you are not being concise enough.

Python notebook

- The purpose of the Python notebook is to show how you implemented your model and calculated your results.
- It has to be reproducible (as far as possible given that there is random variation in the posterior samples from Stan). Make sure your instructor can run your notebook from start to finish without any bugs or errors.
- Remember to include your data in the zip file you upload, so your code can be run to reproduce your results.
- Make your code readable – use good variable names, add code comments and docstrings, and organize your code using functions.