Statistics and Probability Final Project

Yichen Zhu

Release: Dec 4, 2024 DDL: Jan 11, 2025

1 Dataset

The File "housing.csv" is a dataset regarding housing prices in Melbourne, collected at 2017. It contains the housing prices, as well as many variables that may help estimate the housing prices. The list of variables, as well as there descriptions, are detailed as follows.

- Rooms: Number of rooms
- Price: Price in dollars
- Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
- Type: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
- SellerG: Real Estate Agent
- Date: Date sold
- Distance: Distance from CBD
- Regionname: General Region (West, North West, North, North east ... etc)
- Propertycount: Number of properties that exist in the suburb.
- Bathroom: Number of Bathrooms
- Bedroom2 : Scraped Number of Bedrooms (from different source)
- Car: Number of carspots
- Landsize: Land Size
- BuildingArea: Building Size
- CouncilArea: Governing council for the area

The dataset is already well cleaned, but certain variables still have missing values. This is inevitable for real world datsets.

2 General Suggestions

This final project is an open problem. While I provide some general comments and suggestions here, they are **my** "prior" knowledge, which may or may not be helpful when **you** are performing data analysis. In the end, it is also practically impossible to study all the methods & objectives within one project, so don't worry if you do not follow my suggestions.

Objective The general objective is to study the housing prices and how they relate to the covariates in the dataset, which is of great interests for both businesses and individuals. There are numerous ways to detail this general objective, which include, but not limited to,

- How do covariates like type, area, number of rooms, etc, affect housing prices?
- Do different real estate agents result in differences in housing prices?
- What regions of the city of Melbourne have the most expensive housing prices? And can you provide some explanations?
- Is there any temporal trend within the time frame of this dataset?
- Based on the properties that are already sold (or sold prior, etc.), can you give a reasonable suggested buying/selling prices for the properties that have yet been sold at the time of data collection?
- Some covariates have missing values. How do you intend to deal with missing values? Does the "missing" phenomenon itself contains useful information?

The list can be potentially extended indefinitely, but in the end, it is ultimately up to you to decide what you would like to study.

Methods The most basic method would be a linear regression with the housing price as response and variables of your choice as regression covariates. Of course, this is a huge, rich and complex datasets. It has the sample size to support more advanced models. Some potential ways to enhance/improve linear regression include, but not restricted to,

- Interaction: different variables may or may not be independent. It can be either beneficial or redundant to include interaction terms between different variables
- Categorical variables: A naive way is to put an indicator for each category. Another possible approach is to employ a Bayesian hierarchical models, which essentially assumes there is indeed some information that can be shared across different categories.
- Temporal trend: the trend with respect to time, if exists, is usually not linear. Time random effects or auto-regression models might see their uses in this dataset.
- Spatial information: it is clearly housing in different parts of cites generally have different prices. This is often referred to as the spatial patterns. One possible approach is to set a random effects for each council area and put a hierarchical Bayesian prior for such random effects. More complex approaches can lead to to, eg, Gaussian graphical models.
- Missing values: the naive, yet still feasible way, is to remove all rows involving missing values, or removing certain columns with too many missing values. More elegant approaches including setting a Bayesian models and integrate out missing values, or impute (with Bayesian or frequentist approaches) missing values with proper models.

3 Submission and Grading

Submission The project is assigned and submitted in groups. The submission consists of two items:

- A .pdf file, the report of the project. The report shall clearly state your method and the result when applying to the dataset, with human-readable graphical demonstration of your results. It can NOT contain any codes. The length of the report shall be 3-4 pages. While not strictly mandatory, it is highly recommended to write the report in LaTeX. If you do not have some preferred LaTeX templates, you can always use the NeurIPS template available at `https://neurips.cc/Conferences/2023/PaperInformation/StyleFiles`.
- A Rmarkdown or Jupyter notebook, containing all the codes that reproduce the results in your report. For methods involving random procedure, please make sure your result is not an outlier from a carefully picked random number, since the same random number will produce different results on different machines. The code should be run "as is", without any debugs from my side.

Grading The grading of the project consists of two parts, the statistical modeling part, and the statistical computing&coding part.

- The statistical modeling part evaluates your knowledge, experience and creativity when trying to solve this problem. You need to clearly states what model you use, why you choose such model, and demonstrate your model actually produces convincing result regarding certain questions of housing prices. While in general real-world problems require more advanced models, it is not true that more complex models are always better. In fact, under the same effectiveness of statistical results, it is preferred to use the simpler models.
- The statistical computing part evaluates your mastery of computing methods and coding. You are free to use whatever packages available. Depending on the complexity of your methods and the availability of your time, you can also choose to code the basic optimization or MCMC by yourself. Demonstration of effective coding for statistical computing will lead to higher grades. However, I do realize it is not realistic for you to code too basic or too complex methods, and will adjust the grades accordingly.

As a final remark, you do not need to do perfectly on both the statistical modeling and statistical computing part. Mastering on one part and being ok on the another part is sufficient to get full grades.

Plagiarism This is a public available dataset, potentially having related papers & projects circulating on the Internet. While discussions, in person or online, are always welcomed, plagiarism (in the forms like copying texts, copying codes, etc) is strictly forbidden, and will immediately result the **fail** for this course. In fact, this is a complex enough problem, such that it is highly unlikely that two people will think exactly the same as a priori.