Stat 447C Project Proposal

Takumi Horiba(66627217)
Riu Sugimoto (8172226)

Git Repository

- GitHub Repo: https://github.com/takumihoriba/bayes-stats

Project Theme

- *A careful and scientific comparison of a Bayesian estimator with a non-Bayesian estimator. Focus on Bayesian and Frequentist regression on housing price data.*

Candidate Datasets

1. ***Ames Housing Dataset***

   URL: https://www.kaggle.com/datasets/marcopale/housing

   https://github.com/topepo/AmesHousing

   Description:

   - Contains ~80 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.
   - Structure: includes both numeric (e.g., lot area, living area) and categorical features (e.g., neighborhood, style).

House Style <chr>	Overall Qual <dbl>	Overall Cond <dbl>	Year Built <dbl>	Year Remod/Add <dbl>	Roof Style <chr>	Roof Matl <chr>	Exterior 1st <chr>	Exterior 2nd <chr>	Mas Vnr Type <chr>
1Story	6	5	1960	1960	Hip	CompShg	BrkFace	Plywood	Stone
1Story	5	6	1961	1961	Gable	CompShg	VinylSd	VinylSd	None
1Story	6	6	1958	1958	Hip	CompShg	Wd Sdng	Wd Sdng	BrkFace
1Story	7	5	1968	1968	Hip	CompShg	BrkFace	BrkFace	None
2Story	5	5	1997	1998	Gable	CompShg	VinylSd	VinylSd	None
2Story	6	6	1998	1998	Gable	CompShg	VinylSd	VinylSd	BrkFace

   - 6 rows | 18-27 of 82 columns

2. SF Rents
   - https://github.com/rfordatascience/tidytuesday/blob/main/data/2022/2022-07-05/readme.md

     Description:

     - This dataset contains information on San Francisco rental listings, including variables such as price, number of bedrooms/bathrooms, square footage, neighborhood, and date of listing.
     - Additional data files include SF permits and new construction records, potentially useful for analyzing housing trends, rent changes over time, and the effect of new construction on housing markets.

post_id <chr>	date <dbl>	year <dbl>	nhood <chr>	city <chr>	county <chr>	price <dbl>	beds <dbl>	baths <dbl>	sqft <dbl>
pre2013_134138	20050111	2005	alameda	alameda	alameda	1250	2	2	NA
pre2013_135669	20050126	2005	alameda	alameda	alameda	1295	2	NA	NA
pre2013_127127	20041017	2004	alameda	alameda	alameda	1100	2	NA	NA
pre2013_68671	20120601	2012	alameda	alameda	alameda	1425	1	NA	735
pre2013_127580	20041021	2004	alameda	alameda	alameda	890	1	NA	NA
pre2013_152345	20060411	2006	alameda	alameda	alameda	825	1	NA	NA

6 rows | 1-10 of 17 columns

permit_number <dbl>	permit_type <dbl>	permit_type_definition <chr>	permit_creation_date <S3: POSIXct>	block <chr>	lot <chr>	street_number <dbl>	street_number_suffix <chr>	street_name <chr>
2000010368	3	additions alterations or repairs	2000-01-03	0113	025	9	NA	Calhoun
2000010353	6	demolitions	2000-01-03	1785	001A	2921	NA	Irving
2000010498	3	additions alterations or repairs	2000-01-04	3705	042	865	NA	Market
2000010484	3	additions alterations or repairs	2000-01-04	6540	040	525	NA	Jersey
2000010480	3	additions alterations or repairs	2000-01-04	0013	013	145	NA	Jefferson
2000010475	3	additions alterations or repairs	2000-01-04	0241	003	600	NA	California

6 rows | 1-9 of 44 columns

A tibble: 6 × 10

cartodb_id <dbl>	the_geom <lgl>	the_geom_webmercator <lgl>	county <chr>	year <dbl>	totalproduction <dbl>	sfproduction <dbl>	mfproduction <dbl>	mhproduction <dbl>	source <chr>
1	NA	NA	Alameda County	1990	3601	2166	1378	57	DOF_E-8
2	NA	NA	Alameda County	1991	226	-236	395	67	DOF_E-8
3	NA	NA	Alameda County	1992	2652	2018	563	71	DOF_E-8
4	NA	NA	Alameda County	1993	3049	2693	282	74	DOF_E-8
5	NA	NA	Alameda County	1994	2617	2753	-233	97	DOF_E-8
6	NA	NA	Alameda County	1995	3515	3001	445	69	DOF_E-8

6 rows

Code

```
library(readr)
# 1. Read the local AmesHousing CSV file
ames_data <- read_csv("C:/2024winter/stat 447/project/archive/AmesHousing.csv")
head(ames_data)



# 2. Load the TidyTuesday data from GitHub

rent <-
read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2022/2022-07-05/rent.csv')

permits <-
read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2022/2022-07-05/sf_permits.csv')

new_construction <-
read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2022/2022-07-05/new_construction.csv')
head(rent)
head(permits)
head(new_construction)
```

Short Summary of Potential Approaches

- **Frequentist**: Ordinary Least Squares or Regularized Linear Regression (e.g., Ridge, LASSO).
- Evaluate predictive performance via standard metrics (RMSE, MAE, R^2), along with cross-validation for model selection.
- **Bayesian**: Bayesian Linear Regression with a suitable prior (e.g., normal prior on coefficients). Use posterior predictive checks to evaluate model fit and calibration. Compare performance via cross-validation.
- Will focus on how well each approach handles model uncertainty, coefficient interpretability, and predictive accuracy.
- Possibly if time allows , we might Incorporate random intercepts (and possibly slopes) across different groups and Investigate partial pooling to handle sparse data in certain groups, which can improve estimates by sharing information across levels.

Equal Contribution Plan

We plan to divide tasks and collaborate closely so that both team members contribute equally:

- Takumi Horiba
    - Lead on data acquisition, cleaning, and exploratory data analysis (EDA) for the Ames Housing dataset.
    - Implement the Bayesian linear regression model and run posterior predictive checks.
    - Contribute to the final report writing and interpretation of results.
- Riu Sugimoto
    - Lead on the frequentist regression models (OLS, Ridge, LASSO), including hyperparameter tuning and cross-validation.
    - Conduct model diagnostics (e.g., residual analysis, checking assumptions) and performance comparisons.
    - Contribute to the final report writing and result presentation.

We will both:

- Discuss the modeling approach, interpret findings, and ensure consistent documentation.
- Contribute to the GitHub repository commits, code reviews, and final presentation materials.