Bayesian Regression: California Housing Price

Zhiquan Shen
Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
zshen37@jh.edu

Jiyue Zhang
Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
jzhan380@jh.edu

Abstract

This paper delves into Bayesian regression using the California Housing Price data set, sourced from the 1990 census. We also compare this approach with ordinary least square regression for predicting housing prices. The study underscores the practical application of Bayesian techniques and provides insights from the comparative analysis with traditional methods.

1 Introduction

The purpose of this project is to predict housing prices using the California Housing Price dataset from the 1990 census, and gain insights into the factors that determine housing prices in California. The project aims to contribute to a deeper understanding of the dynamics influencing housing markets. Through the application of Bayesian regression, we seek to offer valuable insights in the field of real estate economics.

1.1 Dataset

The data is from 1997 paper titled *Sparse Spatial Autoregressions* by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal, which is based on the 1990 California census data. The data contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). Our data has 10 variables:

Longitude: A measure of how far west a house is; a higher value is farther west

Latitude: A measure of how far north a house is; a higher value is farther north

Housing Median Age: Median age of a house within a block; a lower number is a newer building

Total Rooms: Total number of rooms within a block

Total Bedrooms: Total number of bedrooms within a block

Population: Total number of people residing within a block

Households: Total number of households, a group of people residing within a home unit, for a block

Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

Median House Value: Median house value for households within a block (measured in US Dollars)

Ocean_Proximity: Location of the house with respect to ocean/sea, whether each block group is near the ocean, near the Bay area, inland or on an island.

There are 207 missing data in variable “total bedrooms”, so we removed all these data.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.2 Exploring Correlations

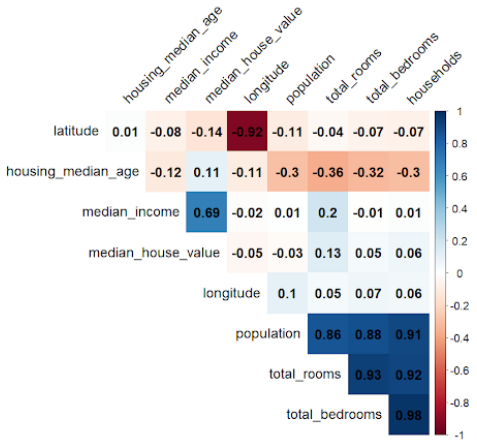

Figure 1: Correlation Matrix

By looking at the correlation matrix (Figure 1), we observed a significant correlation among population, households, total rooms, and total bedrooms, which makes sense since an increased population typically requires more bedrooms. To avoid multicollinearity, we choose to remove certain predictors. Therefore, we plan to compare the 4 models with one of population, households, total rooms, and total bedrooms in the model, and the remaining predictors being housing median age, median income and ocean proximity (Table 1). Following consideration of R^2-Adjested, Akaike Information Criterion(AIC), and Bayesian Information Criteria (BIC), the decision was made to retain total bedrooms among the four strongly correlated variables (Table 2). The choice was made since Model 1 has the lowest AIC, BIC and highest R^2-Adjested, which indicates a better model.

There appears to be a notable negative correlation between the longitude and latitude in our dataset. This relationship can be attributed to California's geographical location on the west coast, with latitude ranging approximately from 32.5 to 42 and longitude from -114.8 to -124.4. Consequently, the geographical positioning gives rise to a strong negative correlation between longitude and latitude. Given their limited range and negative correlation, we decided not to include longitude and latitude as predictors in our model.

Table 1: The Four Models to be Compared

Model	Predictor
1	Total bedrooms, Housing median age, Median income, Ocean proximity
2	Total rooms, Housing median age, Median income, Ocean proximity
3	Population, Housing median age, Median income, Ocean proximity
4	Households Housing median age, Median income, ocean proximity

Table 2: Model Selection with AIC, BIC and R^2 Adjusted

Model	Predictor to choose	AIC	BIC	R^2 Adj
1	Total bedrooms	515352.9	515424.2	0.6463
2	Total rooms	515601.6	515672.9	0.6005
3	Population	515782.4	515853.7	0.5969
4	Households	515444.8	515516.1	0.6036

2 Literature Review

2.1 Zestimate

There were many previous studies related to housing price prediction. One of the most famous is Zestimate home valuation model made by Zillow, which is one of the biggest real-estate company in the United States. Zillow uses information sourced from county and tax assessor records, along with direct feeds obtained from numerous multiple listing services and brokerages. The Zestimate model takes into accout home facts, locations and market trends, including the following:

Home characteristics: square footage, location, number of bathrooms

On-market data: listing price, description, comparable homes in the area and days on the market

Off-market data: tax assessments, prior sales, Market trends, including seasonal changes in demand

Unfortunately, our dataset does not include information related to market trends or taxes. However, we take into consideration the fact that key home characteristics, such as location and the number of rooms, are key factors in house price prediction.

2.2 Bayesian Method

When it comes to regression, besides the widely used ordinary least square regression, bayesian regression is also a very good choice. While some doubting that none of the regressions model selected are actually true, Key et al (1999) in their study of mortality in vegetarians and non-vegetarians indicated that examined mortality in vegetarians and non-vegetarians, revealing that Bayesian model selection are meaningful if we chose the model with best predictive performance. At the same time, the predictive performance of bayesian regression can be improved by averaging predictive distributions under the different models according to their posterior probability(Raftery et al.1997).

As for the choice of piror for the bayesian regression, when in the absence of precise prior information, Kass and Wasserman (1995) suggests using a kind of weakly informative prior called unit information prior, setting $\Sigma^{-1} = (X^T X)/n\sigma^2$ and $\beta_0 = \beta^{ols}$ centering the prior distribution of β around the OLS estimate. The prior distribution, at the same time, can be set as $\nu_0 = 1, \sigma_0^2 = \sigma_{ols}^2$ This setting of prior can be thought of as having weak but unbiased information.

3 Method

We opted for Bayesian linear regression as our method for analyzing and predicting housing prices.The primary purpose of employing a Bayesian linear regression model is to analyze and predict housing prices based on four factors that we have chosen from the OLS method, total bedrooms, median house age, income, and proximity to the ocean. We also want to compare these two methods to see if there are any different insights from the Bayesian perspective of view.

Based on the analysis of the Normal Quantile-Quantile plot of the data (Figure 2), it is evident that the variables under investigation exhibit heavier tails than the normal distribution. In anticipation of the Markov Chain Monte Carlo procedure, we preprocess the data by scaling the numeric columns to standardize the values, facilitating more stable and efficient statistical analysis.

Additionally, the categorical variable Ocean_proximity is transformed into a factor to enable its proper inclusion in the regression model. The dataset is then split into training (80%) and testing (20%) sets for later model evaluation.

In our Bayesian framework, we adopt unit information prior as suggested by Kass and Wasserman (1995) for the regression coefficients, reflecting minimal initial assumptions due to the lack of strong prior knowledge. This approach balances the need for flexibility in the model with the necessity of guiding the inference process.

Unit Information Prior

$$\beta_0 = \beta^{ols}, \Sigma^{-1} = (X^T X)/n\sigma^2, \nu_0 = 1, \sigma_0^2 = \sigma_{ols}^2$$

To estimate the posterior distributions of the model parameters, we employ Markov Chain Monte Carlo (MCMC) methods. The MCMC approach, involving thousands of iterations with an initial

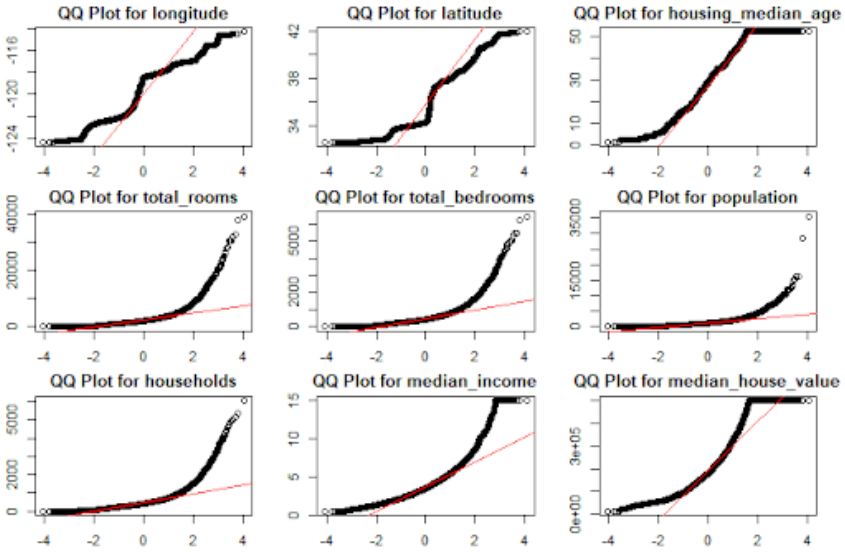

Figure 2: Normal QQ-Plot for variables

burn-in period, will allow us to draw samples from the posterior distributions of the coefficients. This iterative process is crucial for ensuring the convergence and representativeness of the posterior samples, ultimately leading to more reliable and interpretable results. Therefore, we check the convergence by trace plot to ensure the number of iterations is enough.

4 Data Analysis and Results

4.1 Ordinary Least Squares Regression

As mentioned in the previous section about strong correlation between population, households, total rooms, and total bedrooms, to avoid multicollinearity, we first compared the four models(see table 1) to see which in four of the predictors to choose. After selecting the best of the four models, we used backward selection to further improve the model. However, none of the predictors is eliminated after back ward selection.

As a result, our regression model is set to be the following:

$$Y_i = \beta_0 + \beta_1 \times X1 + \beta_2 \times X2 + \beta_3 \times X3 + \beta_4 \times X4 + \beta_5 \times X5 + \beta_6 \times X6 + \beta_7 \times X7$$

X1 = Housing median age
X2 = Total Bedrooms
X3 = Median Income
X4 = 1 if subject's ocean_proximity is in land; 0 otherwise
X5 = 1 if subject's ocean_proximity is island; 0 otherwise
X6 = 1 if subject's ocean_proximity is near bay; 0 otherwise
X7 = 1 if subject's ocean_proximity is near ocean; 0 otherwise

Our ordinary least squares regression result shows the expected value of the coefficients being $b_0 = 25257.988, b_1 = 1258.566, b_2 = 26.702, b_3 = 38658.632, b_4 = -69261.233, b_5 = 184497.594, b_6 = 11347.681, b_7 = 17626.609$

Under t-test with α as low as 0.0001, every coefficient is still significant. At the same time, the result of F-test shows a p-value as low as $2.2e^{-16}$, therefore indicating overall significance of the model.

The OLS regression model indicates that in California in 1990, houses in wealthier blocks with older housing age and more bedrooms generally have higher housing prices. Additionally, houses located on an island, near a bay, or near the ocean are generally more expensive. Conversely, houses located inland are relatively cheaper.

```
Call:
lm(formula = median_house_value ~ housing_median_age + total_bedrooms +
    median_income + ocean_proximity, data = housing_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-540141  -45775  -12519   29822  487890

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              25257.988   2394.300  10.549  < 2e-16 ***
housing_median_age        1258.566     45.994  27.364  < 2e-16 ***
total_bedrooms              26.702      1.282  20.831  < 2e-16 ***
median_income            38658.632    281.146 137.504  < 2e-16 ***
ocean_proximityINLAND   -69261.233   1248.148 -55.491  < 2e-16 ***
ocean_proximityISLAND   184497.594  32447.082   5.686 1.32e-08 ***
ocean_proximityNEAR BAY  11347.681   1743.426   6.509 7.75e-11 ***
ocean_proximityNEAR OCEAN 17626.609  1608.565  10.958  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72520 on 20425 degrees of freedom
Multiple R-squared:  0.6055,    Adjusted R-squared:  0.6053
F-statistic:  4478 on 7 and 20425 DF,  p-value: < 2.2e-16
```

Figure 3: Ordinary Least Square Regression Summary

4.2 Bayesian Regression

4.2.1 Trace Plot and Regression Result

To ensure the convergence of the chains with 10,000 iterations, we examine the trace plots for the model parameters (Figure 4). All the trace plots demonstrate stability without evident trends or drifts, and display good mixing, as indicated by their dense and fuzzy appearance. The absence of visible patterns or cyclicality suggests that the MCMC simulation has achieved successful convergence and mixing.

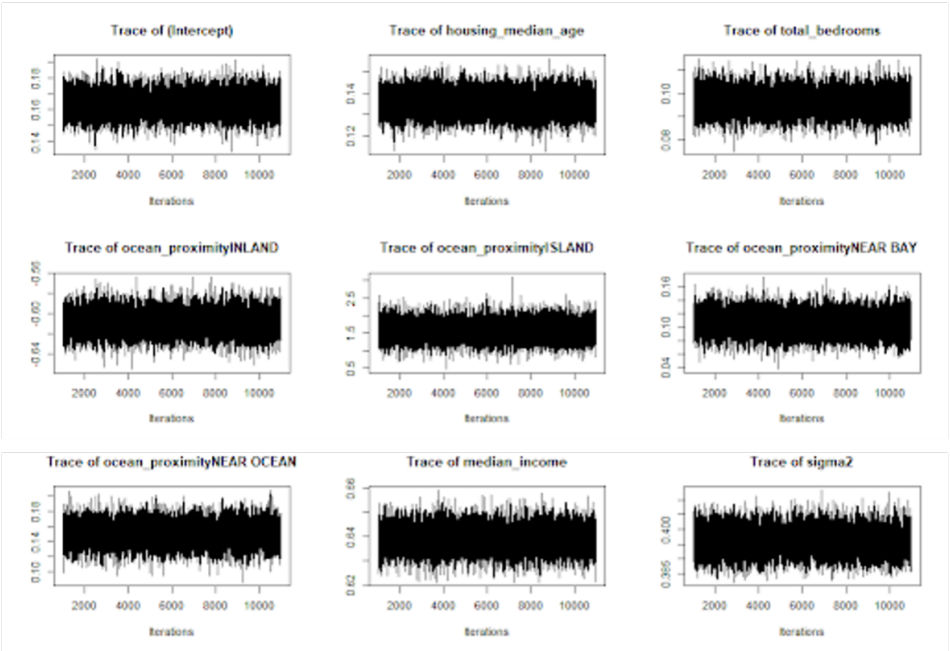

Figure 4: Trace plots for Intercept, housing median age, total bedrooms, ocean proximity (levels), median income, and σ^2.

From the model, the empirical means and standard deviations provide estimates of the central tendency and variability of each parameter (Figure 5). For instance, the mean coefficient for median income is 0.63909 with a standard deviation of 0.005293, indicating a robust and significant positive relationship between median income and housing prices. Conversely, the negative coefficient for 'ocean_proximityINLAND' (mean = -0.60838, SD = 0.011946) suggests that houses further inland are typically associated with lower prices compared to the baseline category of ocean proximity. Additionally, the model's variance parameter, σ^2, has a mean of 0.39610 and a relatively narrow standard deviation of 0.004381, reflecting a reasonable level of precision in the model's predictions.

The quantile summaries provide further insight into the uncertainty and credible ranges for each parameter's estimates (Figure 6). For example, the 2.5% and 97.5% quantiles for 'ocean_proximityISLAND' range from 1.04694 to 2.1430, indicating a high degree of uncertainty, yet the consistently positive values confirm that properties on islands are significantly more expensive.

In summary, the Bayesian regression analysis, supported by empirical means, standard deviations, and quantile ranges, along with the verification of MCMC convergence through trace plots, paints a comprehensive picture of the factors affecting housing prices in the dataset.

	Mean	SD	Naive SE	Time-series SE
(Intercept)	0.16346	0.007494	7.494e-05	7.494e-05
housing_median_age	0.13551	0.005708	5.708e-05	5.708e-05
total_bedrooms	0.09616	0.005206	5.206e-05	5.206e-05
ocean_proximityINLAND	-0.60838	0.011946	1.195e-04	1.195e-04
ocean_proximityISLAND	1.59509	0.281384	2.814e-03	2.859e-03
ocean_proximityNEAR BAY	0.10569	0.017102	1.710e-04	1.710e-04
ocean_proximityNEAR OCEAN	0.14997	0.015648	1.565e-04	1.565e-04
median_income	0.63909	0.005293	5.293e-05	5.293e-05
sigma2	0.39610	0.004381	4.381e-05	4.381e-05

Figure 5: Empirical mean and standard deviation for each variable

	2.5%	25%	50%	75%	97.5%
(Intercept)	0.14908	0.15835	0.16343	0.16853	0.1783
housing_median_age	0.12421	0.13164	0.13557	0.13934	0.1465
total_bedrooms	0.08601	0.09260	0.09619	0.09962	0.1066
ocean_proximityINLAND	-0.63152	-0.61640	-0.60845	-0.60021	-0.5850
ocean_proximityISLAND	1.04694	1.40498	1.59654	1.78109	2.1430
ocean_proximityNEAR BAY	0.07189	0.09402	0.10559	0.11719	0.1391
ocean_proximityNEAR OCEAN	0.11902	0.13944	0.14996	0.16052	0.1805
median_income	0.62887	0.63548	0.63911	0.64273	0.6495
sigma2	0.38760	0.39315	0.39603	0.39901	0.4049

Figure 6: Quantile for each Variable

4.2.2 Predictive Analysis Using Test Data

In the realm of statistical modeling, the true test of a model's predictive capability lies in its performance on new, unseen data. The test dataset serves this exact purpose by acting as a stand-in for future data, allowing us to assess the model's real-world applicability.

Our predictive model, built upon the foundation of our original Bayesian analysis, incorporates an added layer of realism through a noise component. This noise is drawn from a normal distribution centered around zero, with a variance mirroring that of the posterior residuals, reflecting the inherent unpredictability in genuine data scenarios.

Upon examination of the two residual plots presented as Figure 7, we gain valuable insights into the model's behavior. The left plot maps the residuals against their corresponding index, where we observe a lack of discernible patterns or systematic arrangement along the index axis. This randomness in distribution is a favorable sign, indicative of a model that is well-calibrated and free of overt biases.

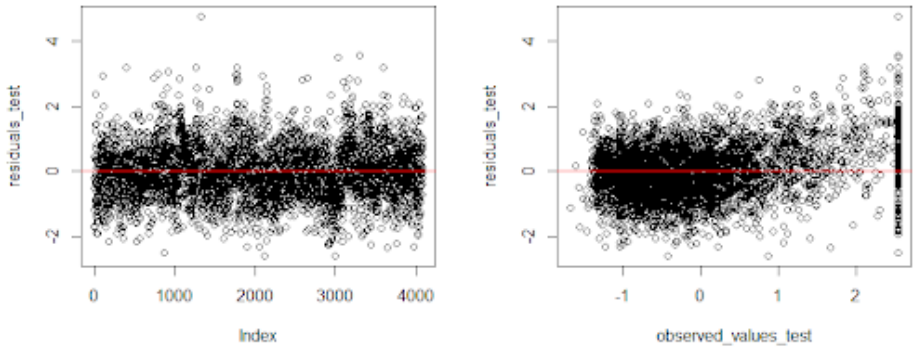

Figure 7: Bayesian Regression Residual Plot

Turning our attention to the right plot, which juxtaposes the residuals against the observed values, we encounter a "funnel" shape in the cloud of points. This pattern, characterized by a widening spread in residuals as we ascend the scale of observed values, hints at potential heteroscedasticity—a phenomenon where the error variance is not uniform across all levels of the independent variable.

Although the horizontal red line in the plot suggests an absence of systematic bias across the spectrum of house values, the diverging spread of residuals at the upper end signals that our model may offer a snugger fit for properties with lower median values. Consequently, the precision of our predictions might diminish as we venture into the domain of higher-valued properties.

In conclusion, while the model exhibits commendable predictive power, the residual plots underscore a need for cautious interpretation, especially when extrapolating to properties at the higher end of the market value range.

4.2.3 Result

The Bayesian linear regression model reveals several key factors that influence housing prices:

Median Income: There is a significant and robust positive correlation between median income and median house value. As median income increases, house values tend to rise, confirming the economic principle that higher earnings provide greater purchasing power in the housing market.

Housing Median Age: Older properties are generally valued higher. This could be due to a variety of factors, such as desirable locations, the charm and quality of historical construction, or limited supply.

Total Bedrooms: An increase in the number of bedrooms is associated with higher house values, which may reflect the market's demand for larger family accommodations or space that can provide multifunctional use.

Ocean Proximity: Living close to the ocean has a distinct positive impact on house values, with island properties commanding a premium. Conversely, properties located inland tend to be valued lower, highlighting the desirability of coastal living.

Our model offers a comprehensive view of the factors affecting house values within the studied district. Median income stands out as a predominant influence on housing prices, with ocean proximity also being a crucial determinant. Despite the model's generally sound predictive capabilities, the heteroscedasticity observed warrants further investigation. Adjustments to the model or its assumptions may be necessary to enhance its predictive reliability across all market segments, ensuring that the model remains robust and unbiased regardless of the value of the property in question.

4.3 Comparison Between OLS Regression and Bayesian Regressions

The Ordinary Least Squares (OLS) regression analysis reinforces the conclusions drawn from the Bayesian linear regression model. Both analytical approaches identify median income, housing median age, total bedrooms, and ocean proximity as significant determinants of housing prices. The coefficients and p-values from the OLS model are consistent with the Bayesian posterior means and credible intervals, providing robust evidence for the impact of these variables on housing values.

The OLS model boasts an R-squared value of approximately 60.55%, denoting that more than half of the variance in housing prices is captured by the included predictors. This metric, while not directly translatable, would benefit from a Bayesian analogue such as the posterior predictive R-squared to allow for a direct comparison.

Despite this strong model performance, the OLS analysis is not without its caveats. The wide range of residuals, as reported by the OLS model, echoes the potential heteroscedasticity observed in the Bayesian residual analysis. Such findings call for a deeper investigation into the model's assumptions and perhaps an exploration of alternative specifications or transformations that could provide a more homoscedastic variance structure.

5 Enhanced Understanding through Bayesian Insights

While the OLS and Bayesian models converge on similar findings, the Bayesian approach affords a more nuanced appreciation of uncertainty and model robustness. The Bayesian framework's ability to quantify uncertainty and incorporate prior knowledge lends itself to a more informed decision-making process, especially valuable in the complex and often unpredictable real estate market.

In summary, both models underscore the prominence of median income and ocean proximity as predictors of housing value, with the Bayesian analysis adding depth to our understanding of predictive uncertainty. Nonetheless, the observed heteroscedasticity signals the need for continued refinement of the models to ensure their reliability and validity across different market segments and value ranges.

References

[1]Hoff, Peter D. *A First Course in Bayesian Statistical Methods*, Springer (2009).

[2]Raftery, Adrian E. David Madigan, and Jennifer A. Hoeting. "Bayesian Model Averaging for Linear Regression Models." Journal of the American Statistical Association, vol. 92, no. 437, 1997, pp. 179-191.

[3] Kass, Robert E. and Larry A. Wasserman. "The Selection of Prior Distributions by Formal Rules." Journal of the American Statistical Association, vol. 91, 1996, pp. 1343-1370.

[4] Key, T. J. et al. "Mortality in vegetarians and nonvegetarians: detailed findings from a collaborative analysis of 5 prospective studies." American Journal of Clinical Nutrition, vol. 70, no. 3 Suppl, 1999, pp. 516S-524S.

[5] Pace, R. Kelley. and Ronald Barry.S̈parse Spatial Autoregressions." Statistics and Probability Letters, vol. 33, no. 3, 1997, pp. 291-297.

[6] Zillow. "Zestimate Home Value." Zillow, https://www.zillow.com/z/zestimate/.