On Variable selection in Bayesian regression

Eddie Conti[1], Gabriela Zemencikova[2], Garcia Vega Ignacio[3], and Brandon Alfaro Jersai[4]

[1]econtico8@alumnes.ub.edu
[2]gzemenze7@alumnes.ub.edu
[3]igarcive11@alumnes.ub.edu
[4]bralfarc7@alumnes.ub.edu

June 11, 2024

1 Introduction to Linear Regression

Linear Regression is a statistical model which estimates the linear relationship between independent variables $x_1, \ldots, x_n$ and dependent variable Y. This model has been widely adopted due to his intepretability and explainability and because it is relatively easy to fit. In a general setting, given $\{y_i, x_{i1}, \ldots, x_{ip}\}$ for $i = 1, \ldots, n$ data points, we want to determine $\beta_0, \ldots, \beta_p$ such that

$$y_i = \beta_0 + \ldots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n,$$

where ϵ_i is a noise and its distribution is very important to understand how the model is performing.
The term 'linearity' refers to the linear relationship between parameters and independent variables, but the variables themselves need not be linear. To better understand this, we can take as example the case of polynomial fitting, which is a particular example of linear regression because we want to estimate $\alpha_0, \ldots, \alpha_p$ such that

$$y_i = \sum_{j=0}^{p} \alpha_j x_i^j \qquad i = 1, \ldots, n.$$

The method to estimate the parameters $\beta = (\beta_0, \ldots, \beta_p)$ is the least-square. We minimize the squared l_2 norm of the prediction against the real value:

$$\mathcal{L}(\beta) = \min_{\beta} ||X\beta - Y||^2$$

where

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Now, by simply imposing

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = 0$$

we can obtain the optimal set of parameters as

$$\beta = (X^T X)^{-1} X^T Y.$$

1.1 Bayesian Linear Regression

The previous model assumes that the set of parameters β are optimal and does not take into account the possibility of uncertainty in collecting the data and estimating the model. In the Bayesian Linear Regression we apply the Bayesian framework to compute the parameters. In particular, the regression coefficients β are assumed to be random variables with a specified prior distribution. The prior distribution, as always, can bias the solutions for the regression coefficients. As a consequence, the Bayesian estimation process does not produce a single point but an entire posterior distribution which allow us to determine the most appropriate parameters along with the uncertainty sorrounding the quantity.

To be more precise, the frequentist setting, is modelling $\mathbb{E}[y|x]$ and assumes that $\epsilon_1, \dots, \epsilon_n \sim N(0, \sigma^2)$. In this scenario, instead, we assume a different relationship between the observed data $y_1, \dots, y_n$ conditional upon $(x_1, \dots, x_n)$ and values β and σ^2:

$$\begin{aligned} p(y_1, \dots, y_n | x_1, \dots, x_n, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i | x_i, \beta, \sigma^2) \\ (2\pi\sigma^2)^{-n/2} \cdot \exp - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2. \end{aligned} \tag{1}$$

In other terms, we can recognize the normal distribution and say that

$$y | X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$$

where I is the identity matrix. Therefore the output, y is generated from a normal distribution characterized by a mean and variance. To obtain the posterior distribution we observe that

$$\begin{aligned} p(\beta, \sigma^2 | y, X) \propto p(y | X, \beta, \sigma^2) p(\beta, \sigma^2) \\ p(\beta | \sigma^2, y, X) \propto p(\beta, \sigma^2 | y, X) \end{aligned} \tag{2}$$

which, in other words, establish that the posterior distribution on parameter β is determined by the likelihood multiplied by the join prior distribution of parameters β, σ^2. Let us show how to derive formulation (2). From Bayes' formula

$$p(\beta, \sigma^2|y, X) = \frac{p(y, X|\beta, \sigma^2)p(\beta, \sigma^2)}{p(y, X)}, \tag{3}$$

now,

$$p(y, X|b, \sigma^2 = p(y|X, \beta, \sigma^2)p(X|\beta, \sigma^2) = p(y|X, \beta, \sigma^2)p(X) \tag{4}$$

because data X are independent of parameters β, σ^2. Now, substituting (4) in (3), we obtain

$$\begin{aligned} p(\beta, \sigma^2|y, X) = \frac{p(y|X, \beta, \sigma^2)p(X)p(\beta, \sigma^2)}{p(y, X)} &= \frac{p(y|X, \beta, \sigma^2)p(\beta, \sigma^2)}{p(y|X)} \\ &\propto p(y|X, \beta, \sigma^2)p(\beta, \sigma^2) \end{aligned}$$

because $p(y|X)$ is independent from parameters β, σ^2 and its role is to normalize the quantity. Similarly,

$$p(\beta|\sigma^2, y, X) = \frac{p(\beta, \sigma^2|y, X)}{p(\sigma^2|y, X)} \propto p(\beta, \sigma^2|y, X)$$

since b is independent of $p(\sigma^2|y, X)$ which is a normalization factor.

In principle, we can use any prior on our parameters that we would like. However, the functional form of most priors, when multiplied by the functional form of the likelihood in results in an posterior with no closed-form solution. As a consequence, it is a good idea to use conjugate prior, meaning that the posterior has the same functional form. A common setting for Bayesian linear regression is to consider all the parameters coming from a normal distribution. Alternatively and more interestingly from a mathematical point of view we can consider the following setting:

$$\begin{aligned} p(\beta, \sigma^2) &= p(\beta|\sigma^2)p(\sigma^2) \\ \beta|\sigma^2 &\sim N(b, \sigma^2 B) \\ \sigma^2 &\sim IG(\alpha_0, \beta_0) \end{aligned} \tag{5}$$

where b is a vector of dimension p and $B = (1/\lambda)I$. Let us denote for simplicity $\lambda I = \Lambda$ We are now ready to derive posterior distribution: according to our assumptions

$$\begin{aligned} p(y|X, \beta, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \cdot \exp\Big(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\Big) \\ p(\beta|\sigma^2) &= (2\pi\sigma^2)^{-p/2}|\Lambda|^{1/2} \cdot \exp\Big(-\frac{1}{2\sigma^2}(\beta - b)^T\Lambda(\beta - b)\Big) \\ p(\sigma^2) &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0 - 1}\exp -\frac{\beta_0}{\sigma^2}. \end{aligned} \tag{6}$$

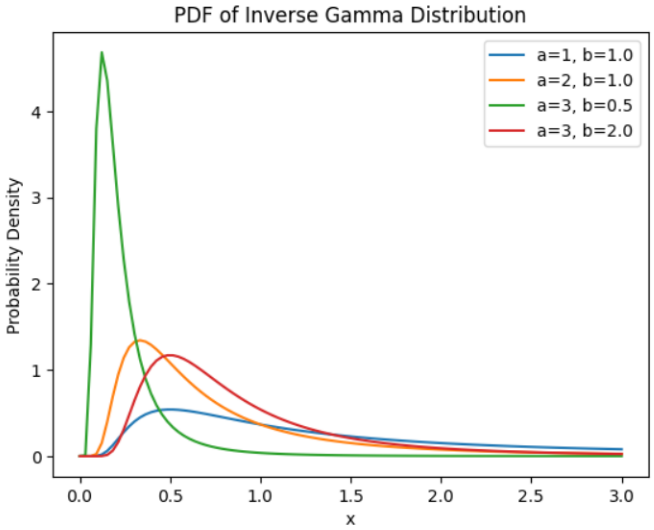


Figure 1: Example of pdf of inverse gamma distribution

Now in view of (2) and that our prior is $p(\beta, \sigma^2)$ as in (5), then the first two equation in are multiplied, as a consequence let us work out the exponential term. First of all we note that

$$\begin{aligned}(y - X\beta)^T(y - X\beta) &= (y - X\beta \pm X\tilde{\beta})^T(y - X\beta \pm X\tilde{\beta}) \\ &= (y - X\tilde{\beta})^T(y - X\tilde{\beta}) + (\tilde{\beta} - \beta)X^TX(\tilde{\beta} - \beta)\end{aligned}$$

As a consequence,

$$\begin{aligned}&(y - X\beta)^T(y - X\beta) + (\beta - b)^T\Lambda(\beta - b) \\ &= (y - X\tilde{\beta})^T(y - X\tilde{\beta}) + (\tilde{\beta} - \beta)X^TX(\tilde{\beta} - \beta) + (\beta - b)^T\Lambda(\beta - b) \\ &= y^Ty + b\Lambda b - b_n^T\Lambda_n b_n + (\beta - b_n)^T\Lambda_n(\beta - b_n)\end{aligned}$$

where Λ_n and b_n are defined as

$$\Lambda_n = X^TX + \Lambda, \qquad b_n = \Lambda_n^{-1}(b\Lambda + X^Ty).$$

The main idea of this manipulation is that we have a term that is quadratic on parameters β. Now, finally,

$$\begin{aligned}p(y|X, \beta, \sigma^2) \propto & (2\pi\sigma^2)^{-p/2}|\Lambda|^{1/2} \cdot \exp\Big(-\frac{1}{2\sigma^2}(\beta - b_n)^T\Lambda_n(\beta - b_n)\Big) \\ & (2\pi\sigma^2)^{-n/2} \cdot \exp\Big(-\frac{1}{2\sigma^2}[y^Ty + b\Lambda b - b_n^T\Lambda_n b_n]\Big) \\ & \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0-1}\exp -\frac{\beta_0}{\sigma^2}.\end{aligned} \tag{7}$$

We can combine the bottom two lines in (7) to obtain

$$(\sigma^2)^{-\alpha_0-n/2-1}\exp\Big(-\frac{1}{\sigma^2}\Big[\beta_0+\frac{1}{2}(y^Ty+b\Lambda b-b_n^T\Lambda_n b_n)\Big]\Big)$$

Now, we define

$$\alpha_n=\alpha_0+\frac{n}{2},\qquad \beta_n=\beta_0+\frac{1}{2}(y^Ty+b\Lambda b-b_n^T\Lambda_n b_n)$$

and we clearly recognize an un-normalized inverse gamma distribution. As a consequence, our posterior

$$\begin{aligned} p(\beta,\sigma^2|X,y)&\propto p(\beta|X,y,\sigma^2)p(\sigma^2|X,y)\\ \beta|X,y,\sigma^2&\sim N(b_n,\sigma^2\Lambda_n^{-1})\\ \sigma^2|y,X&\sim IG(\alpha_n,\beta_n). \end{aligned}\tag{8}$$

1.1.1 Posterior predictive checks

The posterior predictive distribution for new data $\tilde{X}$ and $\tilde{y}$ can be obtained as

$$p(\tilde{y}|y)=\int\int p(\tilde{y}|\tilde{X},\beta,\sigma^2)p(\beta|X,y,\sigma^2)p(\sigma^2|y,X)d\beta\,d\sigma^2.$$

The integration weights the model's prediction of new data by the posterior's parameter estimates from observed data. We can compute $p(\tilde{y}|y)$ by means of Monte Carlo simulation:

- sample $\theta^{(1)}=(\beta,\sigma^2)^{(1)}$ from $p(\beta,\sigma^2|y)$;
- sample a new data set $\tilde{y}$ from $p(y|\theta^{(1)})$;
- run a regression of $\tilde{y}$ on X.

Now, in the Bayesian framework, we can test the goodness of fit of our model by additionally testing how well a data set generated under the assumed model matches with the real data, by means of posterior predictive checks. The main idea is to calculate some statistic for which we have some idea what an 'extreme' value is in the true data set (i.e. invalidate the model). Then calculate the same statistic in the posterior predictive 'data sets' and check how extreme it is. For example, we know that the proportion of standardized residuals greater than 3 is likely to be very small if the regression model is a good fit. So we can calculate a statistic

$$T^{(i)}=\{\text{proportion of std. residuals }>k\text{ in data set }\tilde{y}^{(i)}\}$$

where $k=3$, and compare its distribution over all i with the observed data.

2 Model comparison and variable selection

Often in regression analysis, there are a large number of possible regressor variables measured, even though the majority of the regressors have no true relationship to the response variable Y. In the classical statistical framework, forward, backward and stepwise selection of variables is commonly used, however the model still may pick up spurious correlations. In the Bayesian framework we the approach is straightforward: if we believe that many of the regression coefficients are potentially equal to zero, then we simply come up with a prior distribution that reflects this possibility. As a consequence, we simply modify the regression coefficients as follows

$$\beta_j = z_j \cdot b_j \quad b_j \in \mathbb{R}, \quad z_j \in 0, 1.$$

Therefore,

$$y_i = z_1 b_1 x_{i1} + \ldots + z_p b_p x_{ip} \quad i = 1, \ldots, n.$$

It is clear now that each value of $z = (z1, \ldots, z_p)$ corresponds to a different model. Bayesian model selection proceeds by obtaining a posterior distribution for z. Given a prior distribution $p(z)$ over models, this allows us to compute a posterior probability for each regression model:

$$p(z|y, X) = \frac{p(z)p(y|X, z)}{\sum_z p(z)p(y|X, z)}. \tag{9}$$

The term $p(y|X, z)$ is called marginal likelihood, defined as the probability of an observation after integrating out the model's parameters

$$p(y|X, z) = \int\int p(y|X, \beta, \sigma^2, z)p(\beta|X, \sigma^2, z)p(\sigma^2)d\beta\, d\sigma^2. \tag{10}$$

Once this quantity is computed we can easily compare models according to (9).

The marginal likelihood (10) in the case of $\sigma^2 \sim IG(\alpha_0, \beta_0)$ has a closed form. Using (1.1) and previous notation, the term under the double integral is

$$\begin{aligned}
&(2\pi\sigma^2)^{-p/2}|\Lambda|^{1/2} \cdot \exp\Big(-\frac{1}{2\sigma^2}(\beta - b_n)^T\Lambda_n(\beta - b_n)\Big) \\
&(\sigma^2)^{-\alpha_n - 1} \exp -\frac{\beta_n}{\sigma^2} \\
&(2\pi)^{-n/2} \exp \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}.
\end{aligned}$$

By splitting the integral and recognizing Gamma and Gaussian kernel, we can obtain the following, closed, formulation

$$p(y|X, z) = \frac{1}{(2\pi)^{n/2)}}\sqrt{\frac{|\Lambda|}{|\Lambda_n|}\frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}}\frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)}}.$$

2.1 Gibbs sampling for p large

If we consider a linear regression with p regressors, there are 2^p (i.e. the number of possible subsets of $\{1, \ldots, p\}$) models. In the case we are using $p = 64$ regressors then the number of models is around 1.8×10^{19}. As a consequence it becomes practically unfeasible to compute the marginal probability of each model. A good option is to determine a list of high-probable models and this can be done using Gibbs sampling.
Let us introduce the algorithm in the 2 dimensional case to be able to visualize it. Let us assume a pair of random variables (X, Y) whose joint distribution is not easy to sample from. On the other hand, instead, sampling from $p(Y|X)$ and $p(X|Y)$ is relatively easy. In this scenario we start from $(x^{(0)}, y^{(0)})$, then we sample $x^{(1)} \sim p(X|y^{(0)})$ and finally $y^{(1)} \sim p(Y|x^{(1)})$ resulting in a new point $(x^{(1)}, y^{(1)})$. The intuition behind this method is the following: if a point $(\hat{x}, \hat{y})$ is highly probable w.r.t. $p(X, Y)$, then it means that both the entries $\hat{x}$ and $\hat{y}$ are highly probable w.r.t. $p(X|\hat{y})$ and $p(Y|\hat{x})$. The following picture visually explain the intuition behind Gibbs sampling.

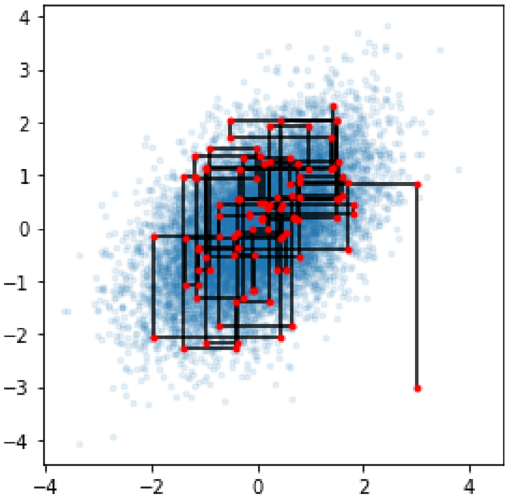


Figure 2: We 'cut' horizontally the joint distribution and move according the marginal distributions.

In our scenario, we can proceed in the following way to perform Gibbs sampling:

- Initialize $z_1 = \ldots = z_p = 0$;
- for each iteration update z_j by sampling from $p(z_j|X, y, z_{-j})$;

- z_{-j} denotes all the entries of z except z_j.

The full conditional for z_j is, as we showed before,

$$p(z_j|X, y, z_{-j}) \propto_{z_j} p(y|X, z)p(z).$$

This can be written as

$$p(z_j = 1|X, y, z_{-j}) = \frac{r_j}{1 + r_j}, \qquad p(z_j = 0|X, y, z_{-j}) = \frac{1}{1 + r_j}$$

where

$$r_j = \frac{p(y|X, z_{-j}, z_j = 1)}{p(y|X, z_{-j}, z_j = 0)} \frac{p(z_{-j}, z_j = 1)}{p(z_{-j}, z_j = 0)}.$$

An alternative possibility to pick the most appropriate model is to compare the BIC, i.e., the Bayesian information criteria, defined as

$$BIC = -2\ln(\text{Likelihood}) + (p + 1)\ln(n).$$

Here n is the number of observations in the model, and p is the number of predictors. This method is actually very simple but tends to be one of the most popular criteria ([1]). In the experiments, for simplicity, we will use this method.

3 Experiments on Housing dataset

For the purpose of the analysis, we are going to use a dataset that is originally from the second chapter of Géron's book [5]. The dataset includes information from the 1990 California census describing the housing market conditions. The dataset provides details about houses within various California districts, along with summary statistics based on the 1990 census data. Note that the dataset is not pre-cleaned, necessitating some preprocessing steps.

The columns included in the dataset are:

`longitude`: The longitude coordinate of the district. In other words, we can understand it as a measure of how far west a house is; a higher value is farther west.

`latitude`: The latitude coordinate of the district. Similarly, it can be understood as a measure of how far north a house is; a higher value is farther north.

`housing_median_age`: The median age of the houses in the district. The lower the number the newer the building.

`total_rooms`: The total number of rooms in the district.

total_bedrooms: The total number of bedrooms in the district.

population: The population of the district.

households: The number of households in the district.

median_income: The median income of the households in the district (measured in tens of thousands of US Dollars).

median_house_value: The median house value in the district (measured in US Dollars).

ocean_proximity: The proximity of the district to the ocean.

3.1 Data-Wrangling & Exploratory Analysis

We start by looking at the missing values. The column with missing values is total_bedrooms. The amount of missing values is relatively small compared to our dataset. However, we make sure that for all categories in ocean proximity, the number of missing values is kept small by taking the ratio. The maximum value of missing data for a category is slightly over 1% which is not too big and the dataset itself is big enough to conduct a linear regression, therefore we proceed by dropping the missing values.

Figure 3 displays histogram of each variable in the dataset. We can obtain several insides. Firstly, housing_median_age presents an odd behaviour around the value 52 together with median_house_value around 500k, and thus it is an alert that there may be potential outliers in the data. Secondly, there are graphs that look like a shifted normal distribution. Thirdly, 4 variables, specifically households, population, total_bedrooms and total_rooms have similar distributions that is skewed to the right. Lastly, when looking at the range of the values it highly differs, therefore, we would proceed with normalizing the dataset by scaling.

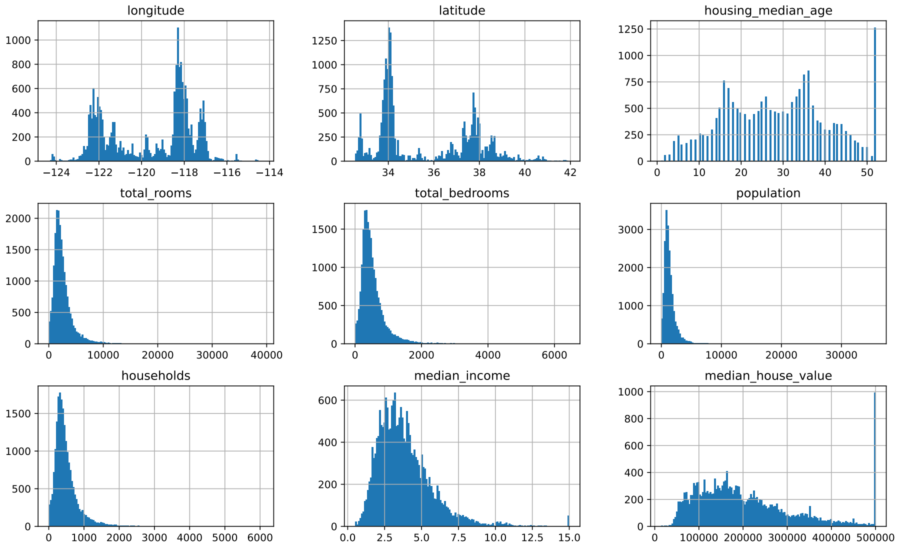


Figure 3: Data Distributions of variables before data-cleaning

To closely examine potential outliers we proceed by using box plots (see Figure 4). It better helps to observe the variability of the data and its distribution. From the plot, we can identify that the variable `median_house_value` is indeed an outlier and those values are dropped from the dataset.

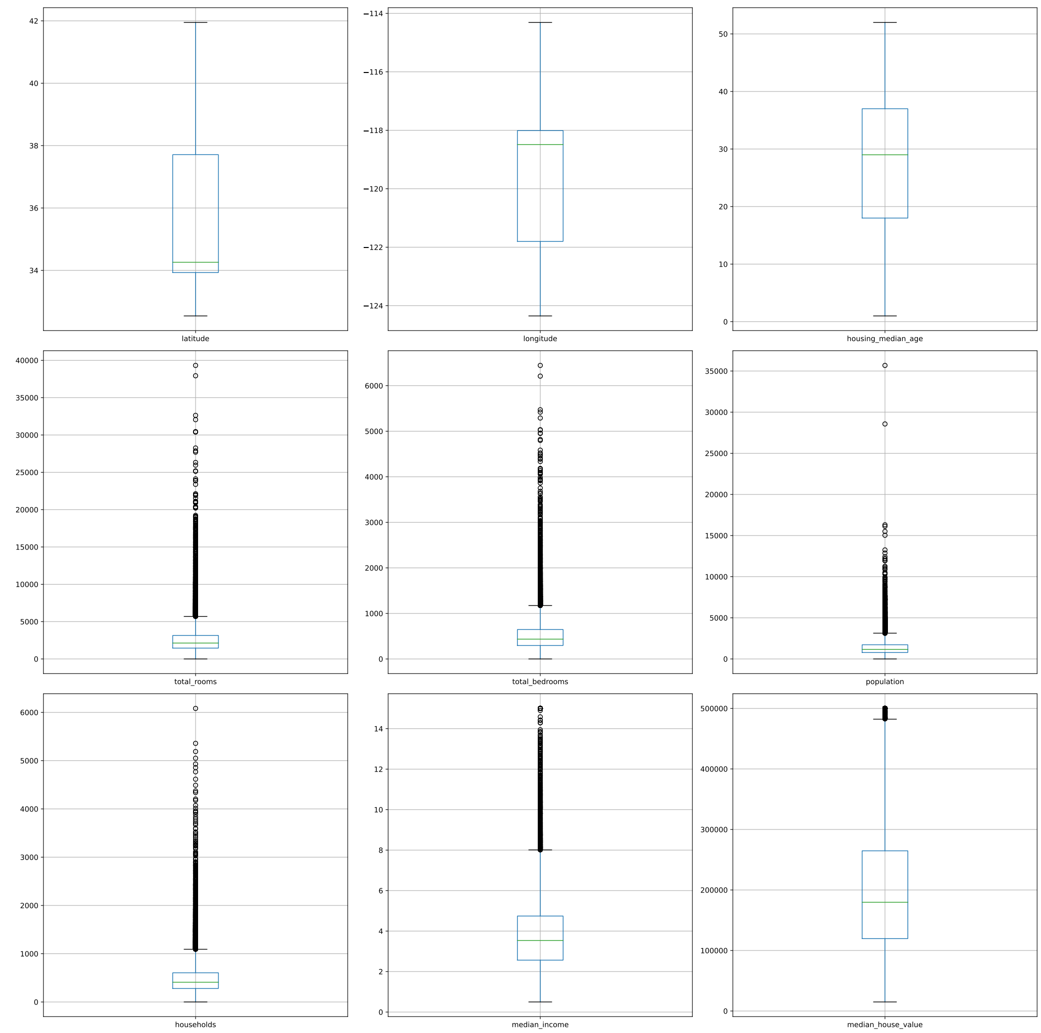


Figure 4: Box plots grouped by variables

Next in order to build a good model, we need to better understand the interaction between variables, Figure 5 displays the correlation plot to check for multicollinearity. Furthermore, it is crucial for the feature selection. It can be observed that `longitude` and `latitude` are highly negatively correlated (-0.92). `Median_house_value`, the target variable is very mildly correlated with all other values (in absolute terms between 0.03-0.14) with the exception of `median_income` which correlation is 0.69 suggesting it to be and important feature. `Total_rooms`, `households`, `population` and `total_bedrooms` are highly correlated to each other, which only confirms our past observations.

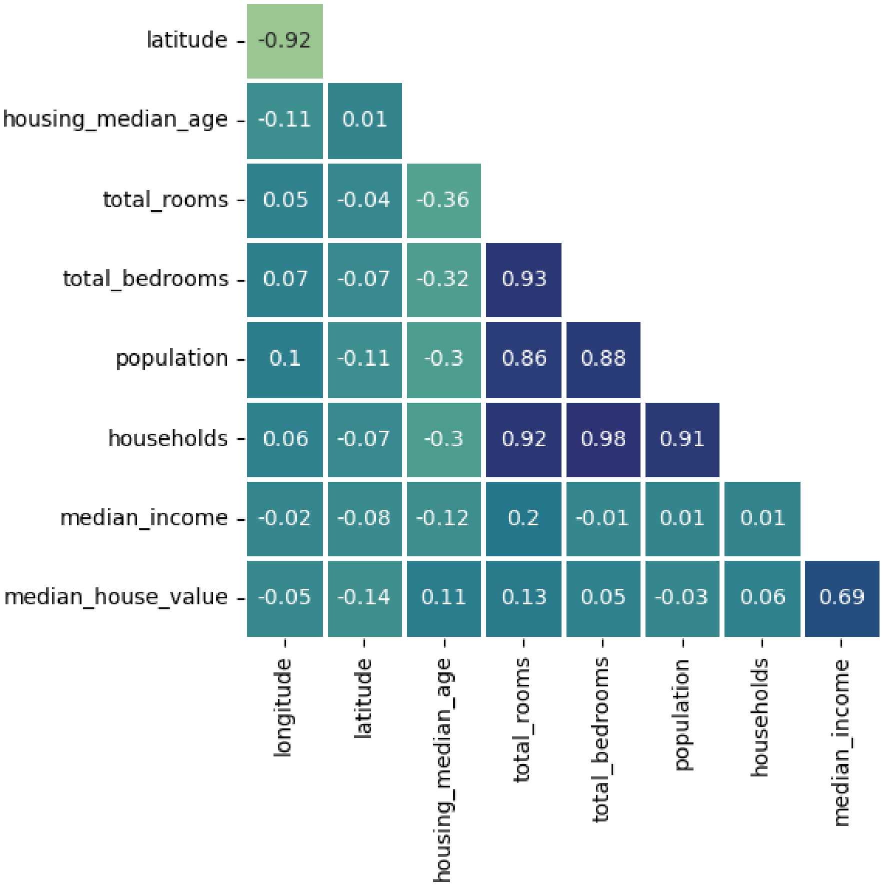
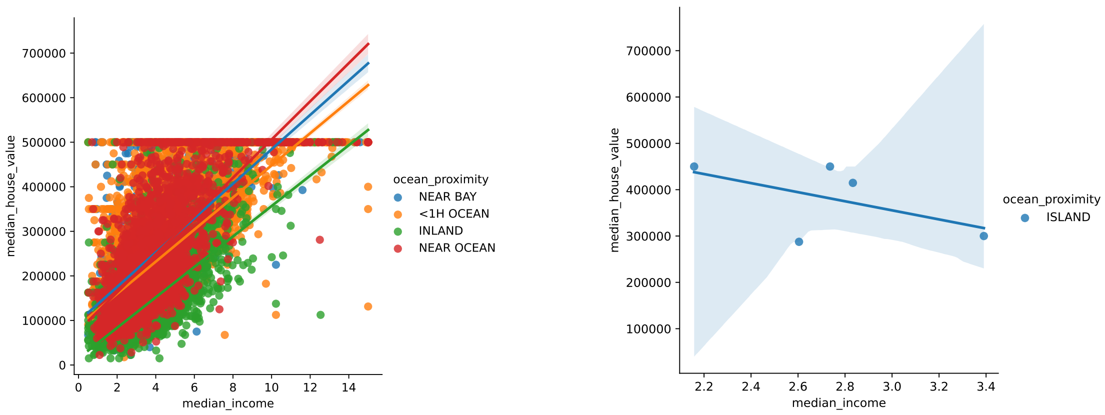


Figure 5: Correlation plot

Since `median_income` is potentially a crucial predictor, we segment the data by `ocean_proximity` and plot a linear regression of the target variable against the income (see Figure 6). Figure 6a display a positive linear relationship of the data. Note that the plot further confirms the 500k income value being an outlier. However, in the `Island` category, the relationship appears less evident (Figure 6b), which may be most likely due to the limited amount of data available for that particular category.



(a) Excluding `ISLAND`

(b) Including `ISLAND`

Figure 6: Regression Plot of Median House Value vs Median Income

After the data cleaning process the Table 1 presents summary statistics by category of ocean proximity for the median house value.

We can observe that districts categorized as `<1H OCEAN` have the highest count of observations, i.e. 8505 districts, which results in the largest sample size out of all. There is a considerable variability present within the median house value variable, having the standard deviation of $86730.26. The category `ISLAND` comprises only 5 districts, resulting in a small sample size. Despite this, these districts exhibit a notably higher mean median house value of $380440, suggesting relatively high house prices. The standard deviation of $80559.56 indicates some variability in house values within this category. `INLAND` has the second largest count of districts, i.e. 6469. These districts on average have low house price than any other category. `NEAR BAY` includes 2077 districts with moderately higher prices than the inland and lower than near the ocean. The last category `NEAR OCEAN` has a similar price than those near the ocean or bay.

In summary, these statistics provide insights into the housing market dynamics within each category of ocean proximity, helping to understand the distribution and range of median house values across different geographical areas. To support the table, we plot histograms of each variable which is displayed in Figure 7.

As mentioned before, after completing data preprocessing steps, i.e. dropping missing values, removing outliers, and identifying the essential variables, the final preparatory step before moving to the modeling phase involved scaling all the numerical variables. Additionally, to accommodate categorical variables in the analysis, we employed one-hot

	count	mean	std	min	25%	50%	75%	max
ocean_proximity								
<1H OCEAN	8505	224112.93	86730.26	17500	162500	208100	271500	500000
INLAND	6469	123331.27	65869.59	14999	77500	108300	148200	500000
ISLAND	5	380440.00	80559.56	287500	300000	414700	450000	450000
NEAR BAY	2077	236910.83	102997.30	22500	157500	220000	313600	500000
NEAR OCEAN	2419	227359.69	101990.51	22500	144600	216700	291350	500000

Table 1: Summary statistics by `ocean proximity` for median house value

encoding to transform them into a suitable format for regression analysis.

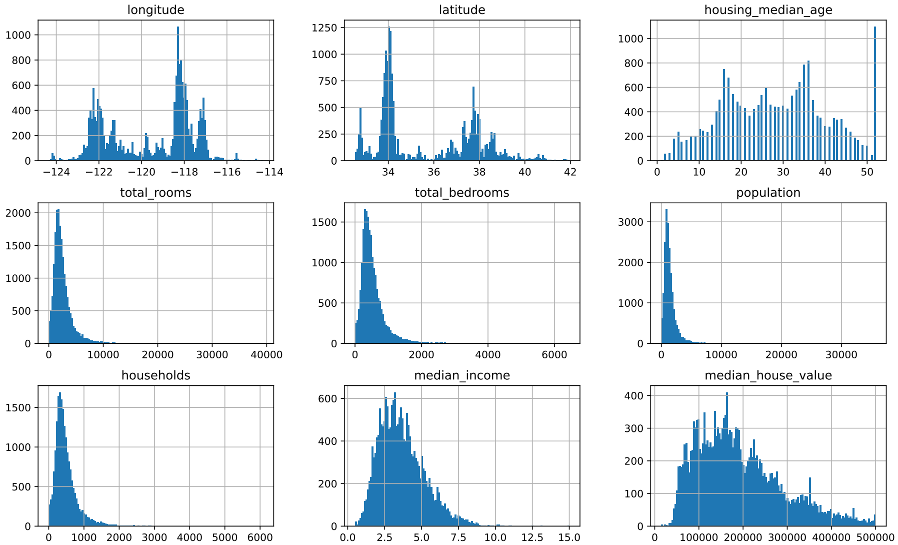


Figure 7: Data Distributions of variables after data-cleaning

4 Modelling

In the first step of our analysis, we performed a standard linear regression to understand the relationship between the median house value and various predictors in the dataset. This initial model serves as a baseline for comparison with the Bayesian framework that we will set up later.

We utilized the `lm` function in R to fit a linear model with `median_house_value` as the dependent variable and all other variables in the dataset as predictors. This method estimates the coefficients for each predictor by minimizing the sum of squared residuals, providing a straightforward interpretation of the influence of each variable on the target. As result we obtained:

- Residual standard error: 60700 on 19461 degrees of freedom
- Multiple R-squared: 0.6142
- Adjusted R-squared: 0.614
- F-statistic: 2384 on 13 and 19461 DF
- p-value: $< 2.2 \times 10^{-16}$

With parameters being:

Table 2: Summary of Linear Regression Results with t-values

Parameter	Estimate	Std. Error	t value	Signif. Code
(Intercept)	2.208e+05	7.656e+03	28.844	$< 2 \times 10^{-16}$
X	6.460e-01	8.065e-02	8.010	1.21×10^{-15}
longitude	-2.313e+05	9.329e+03	-24.791	$< 2 \times 10^{-16}$
latitude	-2.028e+05	8.517e+03	-23.813	$< 2 \times 10^{-16}$
housing_median_age	4.970e+04	2.094e+03	23.736	$< 2 \times 10^{-16}$
total_rooms	-2.956e+05	2.934e+04	-10.076	$< 2 \times 10^{-16}$
total_bedrooms	5.892e+05	4.037e+04	14.596	$< 2 \times 10^{-16}$
population	-1.082e+06	3.467e+04	-31.218	$< 2 \times 10^{-16}$
households	2.751e+05	4.054e+04	6.786	1.19×10^{-11}
median_income	5.572e+05	5.504e+03	101.221	$< 2 \times 10^{-16}$
ocean_proximity_.1H.OCEAN	-1.594e+03	1.477e+03	-1.079	0.2805
ocean_proximity_INLAND	-4.132e+04	2.033e+03	-20.329	$< 2 \times 10^{-16}$
ocean_proximity_ISLAND	1.647e+05	2.720e+04	6.057	1.41×10^{-9}
ocean_proximity_NEAR.BAY	-4.668e+03	2.131e+03	-2.190	0.0285
ocean_proximity_NEAR.OCEAN	NA	NA	NA	NA

The linear regression results indicate that all variables are significant predictors, with the exception of the one-hot encoded variable `ocean_proximity_1H.Ocean`. It is also necessary to clarify that the last variable produces NA because it is redundant. This occurs because the value of this variable can be precisely predicted by the other dummy variables. Essentially, in dummy encoding, $[0, 0, 0, 0]$ is equivalent to $[0, 0, 0, 0, 1]$.

The standard error of the residuals is 60700, which gives us an idea of the average distance between the observed values and the model's predictions.

Next we proceeded to plot the model, resulting:

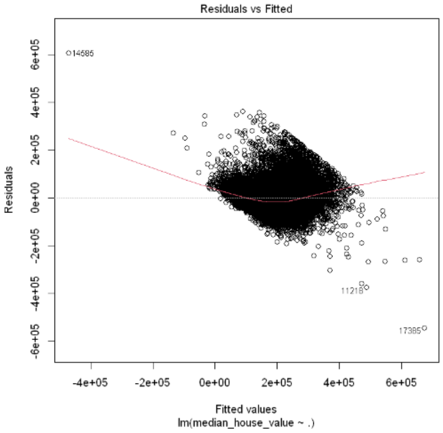


Figure 8: Residuals vs Fitted Values

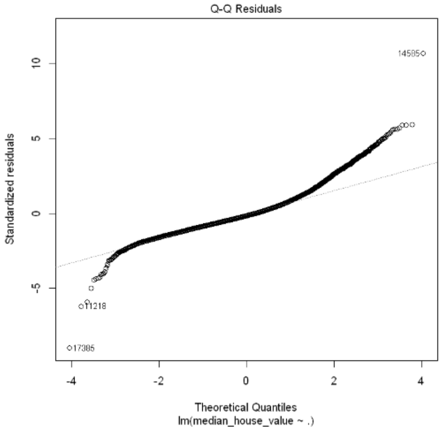


Figure 9: Q-Q Plot of Residuals

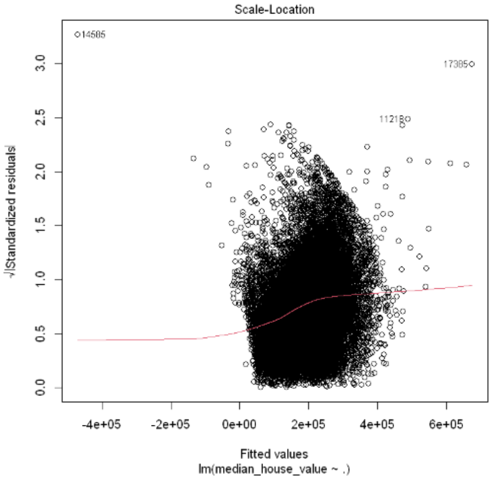


Figure 10: Scale-Location Plot

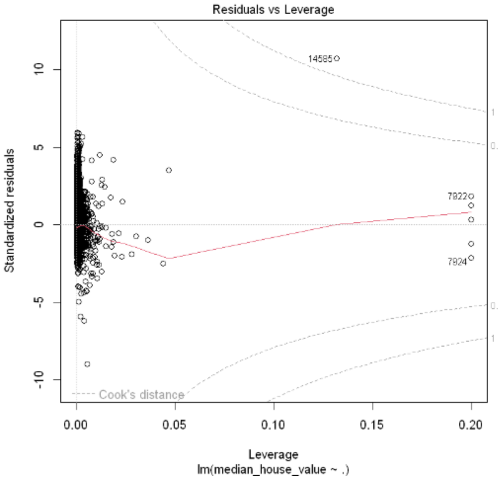


Figure 11: Residuals vs. Leverage Plot

The preceding graphs illustrate that the residuals from our linear regression model are generally normal, suggesting a good fit. This is corroborated by the extremely small p-value in the summary. Overall, our model accounts for 65% of the variability in `median_house_value`.

From this initial regression, we can conclude that while most predictors are significant, there are issues with multicollinearity and non-significant predictors that need to be addressed. Specifically, `median_income` stands out as a crucial variable, significantly affecting the median house value. This insight will guide us in refining the model and setting up the Bayesian framework for further analysis.
For additional analysis we decided to check the variance-covariance matrix since is useful to understand the relation between variables and their contribution. As result we obtained a high variance for each variable, meaning that the model is unstable. This is due to the large size of the dataset and the high values assumed by the variable of interest.
To obtain better results we implemented a new linear regression model (`m2_reg`) with a log transformation applied to the target variable `median_house_value`. The results of the linear regression analysis are shown next:

- Residual standard error: 0.3126 on 19461 degrees of freedom
- Multiple R-squared: 0.6566
- Adjusted R-squared: 0.6564
- F-statistic: 2863 on 13 and 19461 DF
- p-value: $< 2.2 \times 10^{-16}$

As we can see in the previous values, the results of the new model showed significant improvements over the initial regression (0.656 vs 0.6142). And the parameters:

Table 3: Summary of Linear Regression Results with t-values

Parameter	Estimate	Std. Error	t value	Signif. Code
(Intercept)	1.238e+01	3.942e-02	313.913	$< 2 \times 10^{-16}$
X	4.142e-06	4.153e-07	9.974	$< 2 \times 10^{-16}$
longitude	-1.396e+00	4.804e-02	-29.068	$< 2 \times 10^{-16}$
latitude	-1.270e+00	4.386e-02	-28.952	$< 2 \times 10^{-16}$
housing_median_age	1.703e-01	1.078e-02	15.794	$< 2 \times 10^{-16}$
total_rooms	-1.620e+00	1.511e-01	-10.724	$< 2 \times 10^{-16}$
total_bedrooms	3.214e+00	2.079e-01	15.463	$< 2 \times 10^{-16}$
population	-5.141e+00	1.785e-01	-28.795	$< 2 \times 10^{-16}$
households	1.203e+00	2.087e-01	5.763	8.36×10^{-9}
median_income	2.872e+00	2.835e-02	101.321	$< 2 \times 10^{-16}$
ocean_proximity_.1H.OCEAN	3.622e-02	7.608e-03	4.761	1.94×10^{-6}
ocean_proximity_INLAND	-2.640e-01	1.047e-02	-25.224	$< 2 \times 10^{-16}$
ocean_proximity_ISLAND	6.795e-01	1.400e-01	4.852	1.23×10^{-6}
ocean_proximity_NEAR.BAY	9.220e-03	1.098e-02	0.840	0.401
ocean_proximity_NEAR.OCEAN	NA	NA	NA	NA

As we can see the predictors became statistically significant in predicting the log-transformed house values, indicating a stronger relationship between predictors and the target. It also clear that the most important feature became again the median income, with a t value of 101.3.

Moreover, when we run again the variance-covariance matrix and the results exhibited reduced instability compared to the initial model. This reduction in variance suggests that the log transformation helped mitigate the instability caused by the large dataset and high predictor values. It also shown, in this case, a bit of relation between variables total_rooms, households, population, and total_bedrooms.

Since median income is the most important variable we decide to run the model only using this feature, this gave us:

Table 4: Coefficients of the Linear Regression Model

Predictor	**Estimate**	**Std. Error**	**t value**
Intercept	11.34282	0.00663	1710.9
median_income	3.15399	0.02713	116.2

With values:

- Residual standard error: 0.4097 on 19473 degrees of freedom

- Multiple R-squared: 0.4097
- Adjusted R-squared: 0.4096
- F-statistic: 1.351×10^4 on 1 and 19473 degrees of freedom
- p-value: $< 2.2 \times 10^{-16}$

Additionally we plot both median income and log of the house value we obtain:

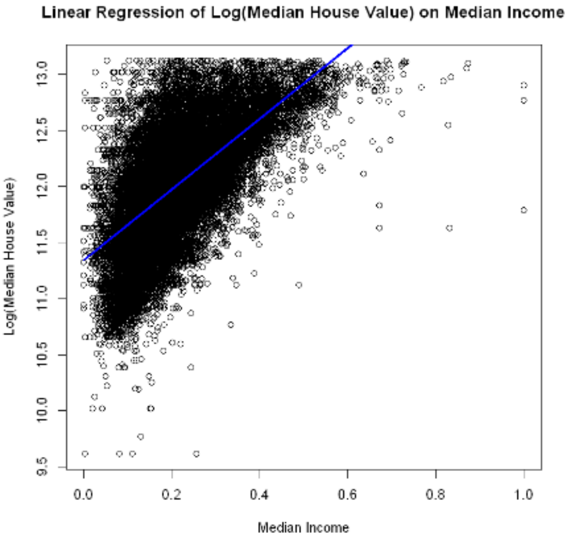


Figure 12: Enter Caption

As we can see using only median income produces a worse model, passing from 0.65 to a 0.4 (far worst result). It is clear that we need to do a complete model selection procedure to identify the best subset of predictors. We started with a base linear model (`ls0`) and systematically added predictors using the step-wise regression method. The results were:

In the table 5, the stepwise function starting from the intercept identifies the best model. As expected, `median_income` emerges as the most important feature. Interestingly, the final model includes variables that are correlated. It would be interesting, though beyond the scope of this project, to explore compressing the variables `latitude` and `longitude` into a single geographical variable. Similarly, combining `total_rooms`, `households`, `population`, and `total_bedrooms` into a single variable before performing the linear regression could be worth investigating. The overall goodness of fit of this model is 0.3134 on 19462 degrees of freedom with an R-squared of 0.6549.

Variable	Coefficient
(Intercept)	12.50022
median_income	2.87794
ocean_proximity_INLAND	-0.27158
total_bedrooms	3.21479
population	-5.16213
housing_median_age	0.15522
total_rooms	-1.55944
households	1.16073
longitude	-1.48658
latitude	-1.32786
ocean_proximity_NEAR.BAY	-0.02821
ocean_proximity_ISLAND	0.65889
ocean_proximity_.1H.OCEAN	0.02060

Table 5: Summary of Best Model from Stepwise Regression

Next, for the Bayesian setting, we utilized the `stan_glm` function from the `rstanarm` package to fit a Bayesian generalized linear model. This approach allows us to estimate the regression coefficients while considering their uncertainty. Using a normal prior distribution we have the following results:

Variable	**Median**	**MAD_SD**
(Intercept)	12.427	1.602
X	0.000	0.000
longitude	-1.398	0.048
latitude	-1.271	0.044
housing_median_age	0.170	0.011
total_rooms	-1.622	0.152
total_bedrooms	3.218	0.212
population	-5.142	0.179
households	1.199	0.210
median_income	2.872	0.029
ocean_proximity_.1H.OCEAN	-0.014	1.593
ocean_proximity_INLAND	-0.312	1.592
ocean_proximity_ISLAND	0.634	1.604
ocean_proximity_NEAR.BAY	-0.041	1.600
ocean_proximity_NEAR.OCEAN	-0.052	1.594

Table 6: Median and MAD_SD for each variable

From the table, we observe that `longitude`, `latitude`, `housing_median_age`, `total_rooms`, `total_bedrooms`, `population`, `households`, and `median_income` have relatively small MAD_SD values compared to their median estimates. This indicates a high degree of certainty in these coefficients.

In particular, `median_income` shows a strong positive effect with a median of 2.872 and a very low MAD_SD of 0.029, highlighting its importance in predicting `median_house_value`. Conversely, the dummy variables for ocean proximity exhibit high MAD_SD values, reflecting greater uncertainty and variability in their estimates. This further supports the idea of compressing and refining the feature set for a more robust model.

We then visualized the distribution of the model parameters using kernel density plots, which offer insights into the uncertainty and variability of the coefficient estimates. These plots were generated using the mcmc_dens function on the Bayesian model.

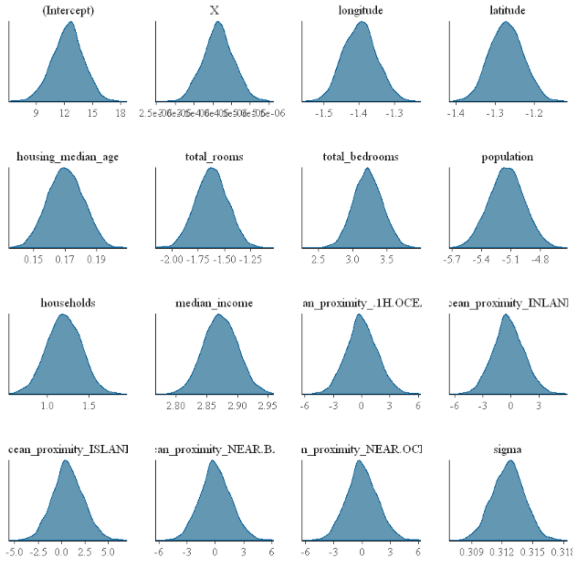


Figure 13: Enter Caption

From the density plots, we can make several observations:

- Narrow Peaks for Most Variables: The density plots for most variables, such as `longitude`, `latitude`, `housing_median_age`, `total_rooms`, `total_bedrooms`, `population`, `households`, and `median_income`, exhibit narrow peaks. This indicates a high level of certainty in the parameter estimates for these variables.

- `median_income`: The plot for `median_income` is particularly notable. It shows a very sharp peak concentrated aroun a small interval, reinforcing that this variable is a strong and significant predictor in the model with minimal uncertainty in its estimate.

- Dummy Variables for Ocean Proximity: The dummy variables representing different ocean proximities (`ocean_proximity_.1H.OCEAN`, `ocean_proximity_INLAND`, `ocean_proximity_ISLAND`, `ocean_proximity_NEAR.BAY`, and `ocean_proximity_NEAR.OCEAN`) display more spread-out distributions, suggesting higher variability and less certainty in their coefficient estimates. This indicates that these variables may not be as strong predictors as others.
- Intercept and Sigma: The intercept term has a very narrow peak, showing high certainty in its estimate. The sigma term, representing the residual standard deviation, also has a relatively narrow distribution, indicating a stable estimate of the model's overall error.

Overall, these density plots provided a visual confirmation of the stability and significance of the model's predictors, highlighting the robustness of variables like `median_income` while also pointing out the higher uncertainty associated with the dummy variables for ocean proximity.

Furthermore, we summarized the posterior distributions of the model parameters in a tabular format. This summary includes the median estimates, credible intervals, and diagnostic metrics such as the Rhat statistic and effective sample size (ESS). These metrics help assess the significance of each predictor and identify potential simplifications for the model.

Parameter	Median	CI	pd	ROPE_Percentage	ESS
(Intercept)	12.43	0.95	1.000	0.000	884
X	4.14e-06	0.95	1.000	1.000	5781
longitude	-1.40	0.95	1.000	0.000	2010
latitude	-1.27	0.95	1.000	0.000	2045
housing_median_age	0.17	0.95	1.000	0.000	5508
total_rooms	-1.62	0.95	1.000	0.000	2434
total_bedrooms	3.22	0.95	1.000	0.000	1641
population	-5.14	0.95	1.000	0.000	2744
households	1.20	0.95	1.000	0.000	1707
median_income	2.87	0.95	1.000	0.000	3021
ocean_proximity_.1H.OCEAN	-0.01	0.95	0.503	0.005	885
ocean_proximity_INLAND	-0.31	0.95	0.577	0.005	886
ocean_proximity_ISLAND	0.63	0.95	0.663	0.006	892
ocean_proximity_NEAR.BAY	-0.04	0.95	0.512	0.008	886
ocean_proximity_NEAR.OCEAN	-0.05	0.95	0.513	0.004	885

Table 7: Summary of Bayesian Regression Model Parameters (Rounded)

It is important to underline that several credible intervals of the coefficients contain zero, suggesting that we could potentially simplify the model. In particular, all the one-hot encoded variables seem to provide no useful information, suggesting that we can remove them to retrain the model.

Finally, we compared the classical linear regression model with the Bayesian one. For the classical model, we computed both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC value for the classical model was 9990.20, while the BIC value was 10108.36. These criteria help in model selection by balancing the goodness of fit with the complexity of the model.

Next, we utilized the Bayesian Information Criterion (BIC) within a Bayesian framework. The `bas.lm` function from the `BAS` package was employed to compute the Marginal Posterior Inclusion Probabilities, indicating the likelihood of each predictor variable being included in the model.

The results of this analysis are summarized in Table 8.

Predictor Variable	**Marginal Inclusion Probability**
Intercept	1.00000
X	1.00000
Longitude	1.00000
Latitude	1.00000
Housing Median Age	1.00000
Total Rooms	1.00000
Total Bedrooms	1.00000
Population	1.00000
Households	0.99999
Median Income	1.00000
Ocean Proximity 1H OCEAN	0.97141
Ocean Proximity INLAND	0.99058
Ocean Proximity ISLAND	0.98936
Ocean Proximity NEAR BAY	0.04714
Ocean Proximity NEAR OCEAN	0.06604

Table 8: Marginal Posterior Inclusion Probabilities for Predictor Variables

From the comparison (BIC values: 10108.36 for the classical linear regression model), we observed that the Bayesian model provides a more succinct representation of the data. We recall that when n is sufficiently large, then we have that

$$BIC \approx -2\ln(p(data|M))$$

To calculate the BIC for the Bayesian model, we utilized a method to approximate the BIC value based on the maximum marginal likelihood obtained from the model's output.

$$\text{BIC approximation} = \begin{cases} -2 \cdot \ln(\text{min\_likelihood}), & \text{if marg\_likelihood} < \text{min\_likelihood} \\ -2 \cdot \ln(\text{marg\_likelihood}), & \text{otherwise} \end{cases}$$

Where `min_likelihood` is a predefined threshold value below which the marginal likelihood is considered to be zero, and `marg_likelihood` is the maximum marginal likelihood obtained from the model's output. This resulted in a BIC value of 460.51. This suggests that the Bayesian approach effectively accounts for the complexity of the model while maintaining a good fit to the data.

Additionally by examining the Marginal Posterior Inclusion Probabilities, we determined the relevance of each predictor variable. We observed that all variables except for `ocean_proximity_NEAR.BAY` and `ocean_proximity_NEAR.OCEAN` were considered relevant in the Bayesian model.

This comparison highlights the advantage of Bayesian statistics, as it allows for a more nuanced approach to variable selection by considering the uncertainty in model parameters and automatically penalizing model complexity through the BIC criterion.

References

[1] M. Clyde et al., *An Introduction to Bayesian Thinking*, GitHub, 2022

[2] Mayetri G., *Advanced Bayesian Methods: Lecture Notes and Practice Exercises*, School of Mathematics and Statistics, University of Glasgow

[3] J. Fortiana, *Regularization & Sparsity*, Lectures notes Universitat de Barcelona, 2023

[4] J. W. Miller, *Model Selection and Variable Selection*, Department of Biostatistics Harvard T.H. Chan School of Public Health

[5] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly Media, 2017