Project BDA

2702305576_Grace

2025-01-02

California Housing Prices

- Number of Rows: 20,640
- Number of Columns: 10
- Source: California Housing Prices Dataset

```
library(rjags)
library(coda)
library(mice)
```

Dataset Overview

This dataset contains housing data for districts in California, including attributes such as location, income levels, and housing characteristics. The main objective is to analyze how these factors influence house pricing.

```
data <- read.csv('housing.csv')
head(data, 5)

##   longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1   -122.23    37.88                 41         880            129
322
## 2   -122.22    37.86                 21        7099           1106
2401
## 3   -122.24    37.85                 52        1467            190
496
## 4   -122.25    37.85                 52        1274            235
558
## 5   -122.25    37.85                 52        1627            280
565
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252             452600        NEAR BAY
## 2       1138        8.3014             358500        NEAR BAY
## 3        177        7.2574             352100        NEAR BAY
## 4        219        5.6431             341300        NEAR BAY
## 5        259        3.8462             342200        NEAR BAY

# Checking for missing value
sum(is.na(data))

## [1] 207
```

```r
# Convert 'ocean_proximity' to a factor so we can impute mice
data$ocean_proximity <- as.factor(data$ocean_proximity)
impute_methods <- ifelse(sapply(data, is.numeric), "pmm", "")

imputed_data <- mice(data, m = 1, maxit = 50, method = impute_methods, seed =
500)
```

```
##
##  iter imp variable
##   1   1  total_bedrooms
##   2   1  total_bedrooms
##   3   1  total_bedrooms
##   4   1  total_bedrooms
##   5   1  total_bedrooms
##   6   1  total_bedrooms
##   7   1  total_bedrooms
##   8   1  total_bedrooms
##   9   1  total_bedrooms
##   10   1  total_bedrooms
##   11   1  total_bedrooms
##   12   1  total_bedrooms
##   13   1  total_bedrooms
##   14   1  total_bedrooms
##   15   1  total_bedrooms
##   16   1  total_bedrooms
##   17   1  total_bedrooms
##   18   1  total_bedrooms
##   19   1  total_bedrooms
##   20   1  total_bedrooms
##   21   1  total_bedrooms
##   22   1  total_bedrooms
##   23   1  total_bedrooms
##   24   1  total_bedrooms
##   25   1  total_bedrooms
##   26   1  total_bedrooms
##   27   1  total_bedrooms
##   28   1  total_bedrooms
##   29   1  total_bedrooms
##   30   1  total_bedrooms
##   31   1  total_bedrooms
##   32   1  total_bedrooms
##   33   1  total_bedrooms
##   34   1  total_bedrooms
##   35   1  total_bedrooms
##   36   1  total_bedrooms
##   37   1  total_bedrooms
##   38   1  total_bedrooms
##   39   1  total_bedrooms
##   40   1  total_bedrooms
##   41   1  total_bedrooms
```

```
##   42   1  total_bedrooms
##   43   1  total_bedrooms
##   44   1  total_bedrooms
##   45   1  total_bedrooms
##   46   1  total_bedrooms
##   47   1  total_bedrooms
##   48   1  total_bedrooms
##   49   1  total_bedrooms
##   50   1  total_bedrooms
```

```r
df <- complete(imputed_data)
sum(is.na(df))
```

```
## [1] 0
```

```r
str(df)
```

```
## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : Factor w/ 5 levels "<1H OCEAN","INLAND",..: 4 4 4 4
4 4 4 4 4 4 ...
```

Column Descriptions:

1. **longitude**
    - Longitude coordinate of the district, representing its geographical location (Float).
2. **latitude**
    - Latitude coordinate of the district, representing its geographical location (Float).
3. **housing_median_age**
    - Median age of houses in the district. Used to approximate the age of housing stock (Float).
4. **total_rooms**
    - Total count of rooms across all houses in the district (Integer).
5. **total_bedrooms**
    - Total count of bedrooms across all houses in the district. May contain missing values (Float).
6. **population**
    - Total number of residents in the district (Integer).

7. **households**
    - Total number of households in the district, where each household represents a group of people living in the same housing unit (Integer).
8. **median_income**
    - Median household income in the district, scaled between ~0.5 and ~15 (Float).
9. **median_house_value**
    - Median house price in the district, expressed in US dollars (Float). This serves as the target variable.
10. **ocean_proximity**
    - Categorical feature indicating the district's distance to the ocean, with values like:
        - **<1H OCEAN**: Less than one hour from the ocean.
        - **INLAND**: Located inland.
        - **NEAR OCEAN**: Close to the ocean.
        - **NEAR BAY**: Near the bay area.
        - **ISLAND**: Island region.

```
summary(df)

##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
##  Median :-118.5   Median :34.26   Median :29.00      Median : 2127
##  Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
##  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##  total_bedrooms     population      households     median_income
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
##  1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  median_house_value   ocean_proximity
##  Min.   : 14999      <1H OCEAN :9136
##  1st Qu.:119600      INLAND    :6551
##  Median :179700      ISLAND    :   5
##  Mean   :206856      NEAR BAY  :2290
##  3rd Qu.:264725      NEAR OCEAN:2658
##  Max.   :500001

long <- df[,1]
lat <- df[,2]
age <- df[,3]
room <- df[,4]
bedroom <- df[,5]
```

```r
pop <- df[,6]
household <- df[,7]
income <- df[,8]
price <- df[,9]
ocean <- df[,10]

df <- as.matrix(df)
Y <- as.numeric(price)
X <- cbind(long, lat, age, room, bedroom, pop, household, income, price)
names <- c("Intercept", "Longitude", "Latitude", "Age", "Room", "Bedroom",
"Population", "Household", "Income", "Price")
cor(X)

##                  long         lat         age        room      bedroom
## long       1.00000000 -0.92466443 -0.10819681  0.04456798  0.068358801
## lat       -0.92466443  1.00000000  0.01117267 -0.03609960 -0.066169120
## age       -0.10819681  0.01117267  1.00000000 -0.36126220 -0.320547306
## room       0.04456798 -0.03609960 -0.36126220  1.00000000  0.930084573
## bedroom    0.06835880 -0.06616912 -0.32054731  0.93008457  1.000000000
## pop        0.09977322 -0.10878475 -0.29624424  0.85712597  0.877837647
## household  0.05531009 -0.07103543 -0.30291601  0.91848449  0.979773178
## income    -0.01517587 -0.07980913 -0.11903399  0.19804965 -0.008010765
## price     -0.04596662 -0.14416028  0.10562341  0.13415311  0.050602620
##                    pop  household       income       price
## long       0.099773223  0.05531009 -0.015175865 -0.04596662
## lat       -0.108784747 -0.07103543 -0.079809127 -0.14416028
## age       -0.296244240 -0.30291601 -0.119033990  0.10562341
## room       0.857125973  0.91848449  0.198049645  0.13415311
## bedroom    0.877837647  0.97977318 -0.008010765  0.05060262
## pop        1.000000000  0.90722227  0.004834346 -0.02464968
## household  0.907222266  1.00000000  0.013033052  0.06584265
## income     0.004834346  0.01303305  1.000000000  0.68807521
## price     -0.024649679  0.06584265  0.688075208  1.00000000
```

I am using median_income, latitude, and total_rooms as the predictors for house pricing, with median_house_value being the target variable.

```r
# JAGS Model Specification
model_code <- "
model {
  for (i in 1:N) {
    Y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * X1[i] + beta2 * X2[i] + beta3 * X3[i]
  }
  beta0 ~ dnorm(0, 0.01)
  beta1 ~ dnorm(0, 0.01)
  beta2 ~ dnorm(0, 0.01)
  beta3 ~ dnorm(0, 0.01)
  tau <- 1 / sigma2
  sigma2 ~ dgamma(2, 0.1)
```

```r
}
"

jags_data <- list(
  N = nrow(X),
  Y = Y,
  X1 = X[, 8],  # Income
  X2 = X[, 2],  # Lat
  X3 = X[, 4]   # Room
)

model <- jags.model(textConnection(model_code), data = jags_data, n.chains =
3)

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 20640
##    Unobserved stochastic nodes: 5
##    Total graph size: 122928
##
## Initializing model

update(model, 1000)

samples <- coda.samples(model, c("beta0", "beta1", "beta2", "beta3",
"sigma2"), n.iter = 10000)
```

Output Posterior

Empirical and Quantiles

```r
summary(samples)

##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##             Mean        SD  Naive SE Time-series SE
## beta0  2.066e+02 1.003e+01 5.792e-02      5.933e-02
## beta1  7.877e+03 9.334e+00 5.389e-02      8.557e-02
## beta2  4.272e+03 1.971e+00 1.138e-02      2.448e-02
## beta3  7.785e+00 1.863e-02 1.076e-04      2.064e-04
## sigma2 3.424e+07 1.362e+04 7.863e+01      9.910e+01
##
## 2. Quantiles for each variable:
```

```
##
##              2.5%       25%       50%       75%     97.5%
## beta0  1.869e+02 1.998e+02 2.066e+02 2.134e+02 2.262e+02
## beta1  7.859e+03 7.871e+03 7.877e+03 7.883e+03 7.895e+03
## beta2  4.268e+03 4.271e+03 4.272e+03 4.273e+03 4.276e+03
## beta3  7.749e+00 7.772e+00 7.785e+00 7.798e+00 7.822e+00
## sigma2 3.421e+07 3.423e+07 3.424e+07 3.425e+07 3.426e+07
```

Interpretation

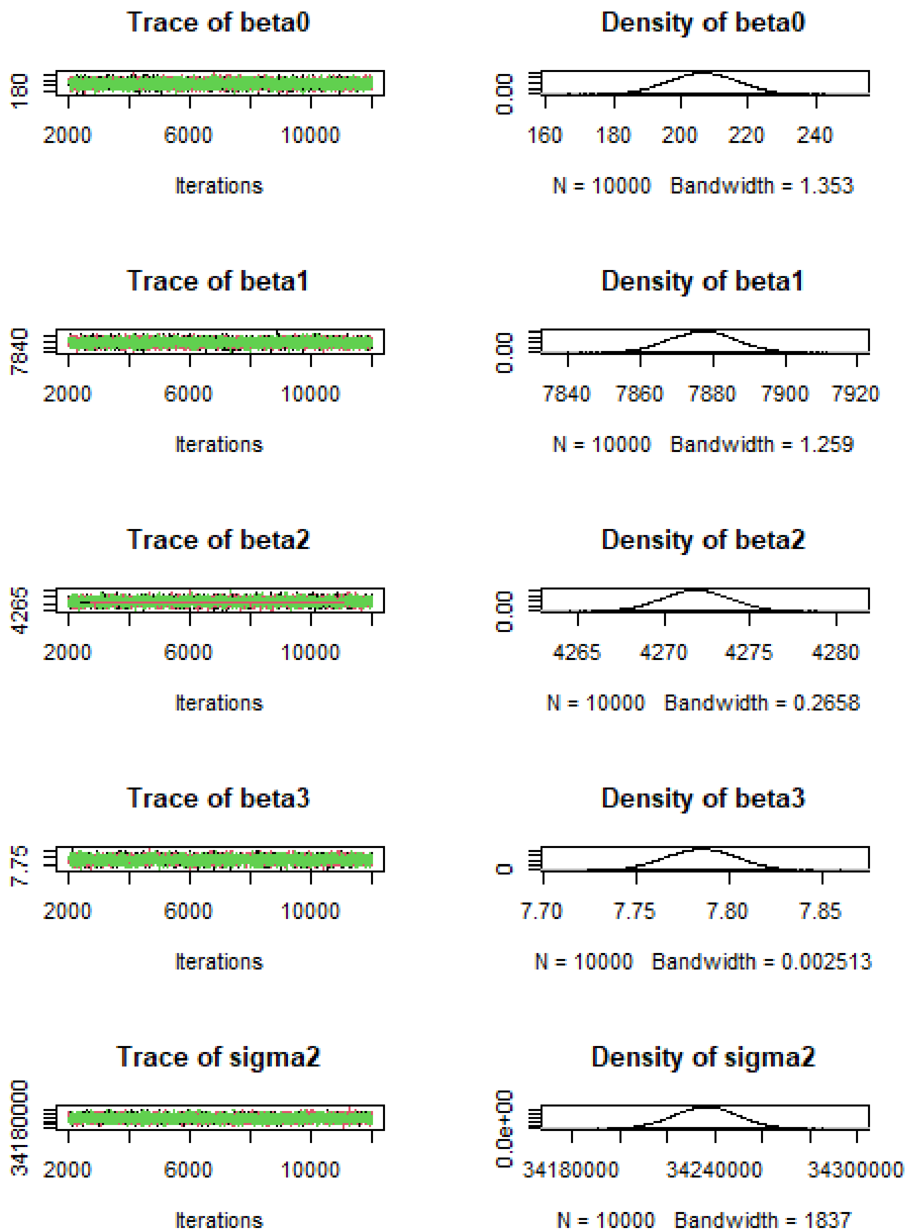
The residual variance, sigma2, is estimated at 34,240,000, indicating substantial variability in median house prices not explained by the predictors. This is expected since I am only using three predictors: median_income, latitude, and total_rooms.

- Median Income (beta1), the most influential predictor. Districts with higher median household income are strongly associated with higher median house prices. The posterior mean is 7877, reflecting a significant positive relationship.
- Latitude (beta2), latitude is positively associated with median house prices (posterior mean: 4272). This could reflect regional differences in housing demand or geographic desirability.
- Total Rooms (beta3), total rooms have a small but consistent positive effect (posterior mean: 7.785). While significant, its contribution is modest, possibly reflecting a limited role compared to other unmodeled housing characteristics.
- Residual Variance (sigma2), the large variance (34,240,000) since I am only using three predictors, this suggests other factors, which are not captured by the current model. But as expected.

Model Evaluation

Convergence Diagnostics

```r
plot(samples)
```



1. Trace Plots

- The chains for all parameters (beta0, beta1, beta2, beta3, sigma2) show good mixing and consistent fluctuations around a stable mean.

- Different chains (depicted in green, black, and red) overlap well, suggesting convergence from different initial values.
- No noticeable trends or drifts are present, confirming that the Markov chains have reached their stationary distribution.
- This indicates the posterior samples are reliable, and the MCMC sampler has converged for all parameters.

2. Density Plots

- beta0 (Intercept), the density is unimodal and symmetric, peaking around 206.5, matching the posterior mean. This shows the intercept is well-estimated with low uncertainty.
- beta1 (Median Income), the density is narrow and symmetric, centered near 7877, indicating the strong effect of income on house prices and high confidence in its estimate.
- beta2 (Latitude), the distribution is unimodal and slightly broader, centered around 4272. This reflects more moderate variability in the effect of latitude on house prices.
- beta3 (Total Rooms), the density is very narrow and symmetric, centered near 7.785. This indicates high confidence in the small but consistent effect of total rooms.
- sigma2 (Residual Variance), the distribution is smooth and slightly right-skewed, centered near 34,240,000, highlighting substantial variability in house prices unexplained by the predictors.

Overall Assessment

- The trace plots confirm the chains have converged and that the posterior space is well-explored.
- The density plots indicate that all parameters have well-defined posterior distributions consistent with their credible intervals.

Gelman-Rubin Diagnostic

```
gelman_diag <- gelman.diag(samples)
gelman_diag

## Potential scale reduction factors:
##
##         Point est. Upper C.I.
## beta0            1       1.00
## beta1            1       1.00
## beta2            1       1.01
```

```
## beta3           1       1.00
## sigma2          1       1.00
##
## Multivariate psrf
##
## 1
```

The Gelman Rubin Diagnostic (PSRF) confirm that the MCMC sampler has fully converged for all parameters. The posterior samples are reliable, and the model results can be confidently interpreted.

Autocorrelation diagnostics

```
acf_plot <- autocorr.diag(samples)
acf_plot

##                beta0         beta1        beta2        beta3       sigma2
## Lag 0   1.0000000000  1.0000000000  1.000000000  1.000000000  1.000000000
## Lag 1   0.0524653492  0.4598037860  0.670676602  0.588480889  0.203195506
## Lag 5  -0.0108585492 -0.0047045147  0.095575578  0.048281331 -0.003446531
## Lag 10 -0.0003424581 -0.0004885458  0.004048280 -0.003357387 -0.004151737
## Lag 50  0.0015319327  0.0033671763 -0.004855037 -0.007344229  0.001845582
```

- The diagnostics show that the chains for all parameters decorrelate quickly, demonstrating good mixing and independence of posterior samples.
- Parameters like beta2 and beta3 may exhibit slightly higher autocorrelation initially but still reach negligible values at higher lags.

ESS

```
ESS <- effectiveSize(samples)
ESS

##     beta0     beta1     beta2     beta3    sigma2
## 28662.467 11918.413  6506.808  8154.633 18974.164
```

- All parameters have sufficient ESS for reliable inference, with beta0 and sigma2 having the highest values.
- Slightly lower ESS for beta2 and beta3, it shows their relatively higher autocorrelation but is still acceptable for most analyses.