Introduction

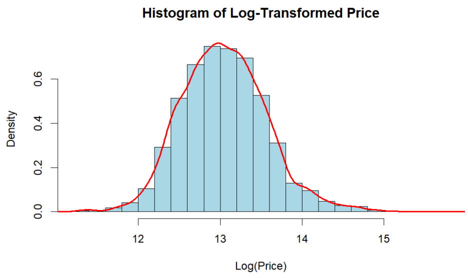
We obtain the dataset used for this project from kc_house_data. The dataset originally contains 21,613 rows and 21 columns, with the aim to predict the variable 'price' by using other predictors. But then, we sampled 5000 data and subset 11 predictors variables and its output based on correlation matrix. The final data contains:

1. price - Price of each home sold
2. bedrooms - Number of bedrooms
3. bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
4. sqft living - Square footage of the apartments interior living space
5. floors - Number of floors
6. waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not
7. view - An index from 0 to 4 of how good the view of the property was
8. grade - An index from 1 to 13, where 1 falls short of building construction and design and 13 have a high quality level of construction and design
9. sqft above - The square footage of the interior housing space that is above ground level sqft
10. basement - The square footage of the interior housing space that is below ground level
11. lat - Lattitude
12. sqft living15 - The square footage of interior housing living space for the nearest 15 neighbors

Model

Bayesian Linear Regression

We chose to use Linear Regression because the model provides a straightforward and interpretable way to model the relationship between a continuous dependent variable (house price) and multiple predictors (features such as bedrooms, bathrooms, etc.). So before we continue the model, we first ensure the data is normally distributed using histogram.



Bayesian Beta Regression

This model is suitable for the dataset because the response variable, is likely a positive continuous value (house price) and it is possible to transform it into a range between 0 and 1. This regression model helps us to estimate the relationship between predictors/features and the scaled response variable (house price proportion) while providing uncertainty quantification through posterior distributions. The goal is to predict the normalized house price based on the predictors.

Algorithm

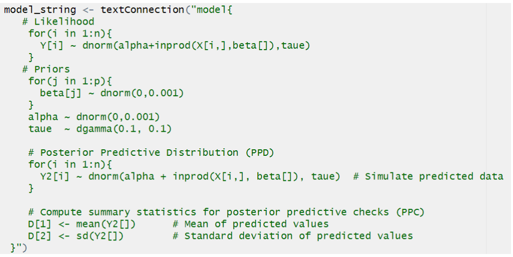
Bayesian Linear Regression → Uninformative Gaussian

Posterior:

$$Yi \sim Normal(\beta 0\ +\ \sum_{j=1}^{p} Xij\ \beta j,\ \sigma^{2})$$

Prior:

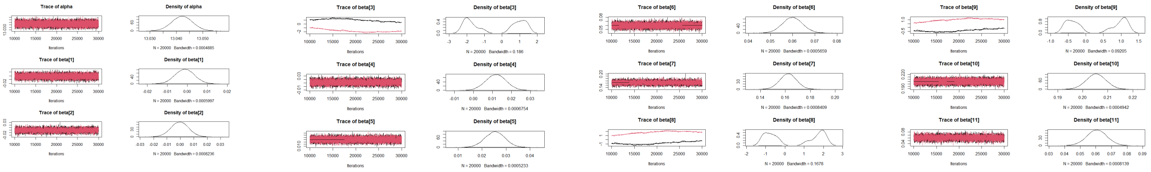
$$\beta j \sim Normal(0,\ 1000)$$



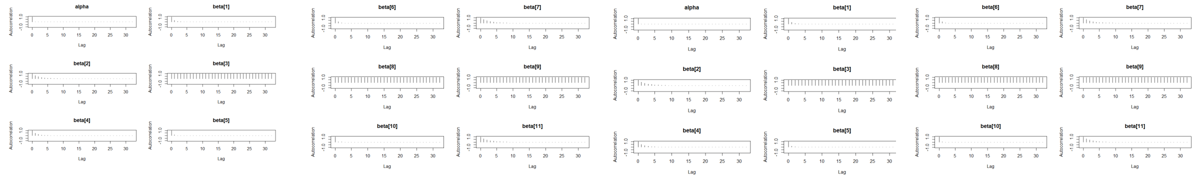
```
model_string <- textConnection("model{
   # Likelihood
    for(i in 1:n){
      Y[i] ~ dnorm(alpha+inprod(X[i,],beta[]),taue)
    }
   # Priors
    for(j in 1:p){
      beta[j] ~ dnorm(0,0.001)
    }
    alpha ~ dnorm(0,0.001)
    taue  ~ dgamma(0.1, 0.1)

    # Posterior Predictive Distribution (PPD)
    for(i in 1:n){
      Y2[i] ~ dnorm(alpha + inprod(X[i,], beta[]), taue)  # Simulate predicted data
    }

    # Compute summary statistics for posterior predictive checks (PPC)
    D[1] <- mean(Y2[])      # Mean of predicted values
    D[2] <- sd(Y2[])        # Standard deviation of predicted values
 }")
```

Where the intercept $\beta 0 \sim Normal(0,\ 1000)$ and the variance $\sigma \sim Gamma(0.1,\ 0.1)$

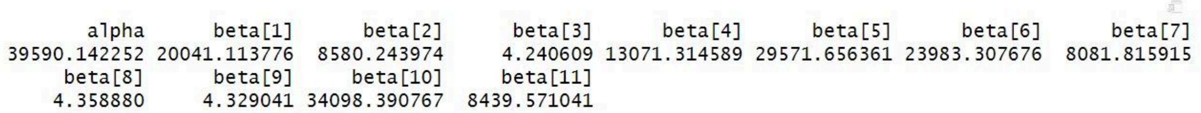
Convergency test:



Here are the graphics diagnostic where we use the chains = 2 and both of the model iteration are convergent, except for beta[3] (sqft_living), beta[8] (sqft_above), beta[9] (sqft_basement).

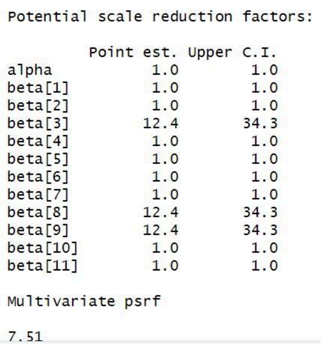


For most parameters, the **autocorrelation** drops rapidly to near-zero within a lag of 1–2, indicating excellent mixing of the MCMC chains and efficient parameter space exploration. However, beta[3], beta[8], and beta[9] stands out with consistently high autocorrelation across all lags, which suggests poor mixing for this parameter. These results confirm that the MCMC algorithm is performing well, producing independent and reliable samples for posterior analysis.



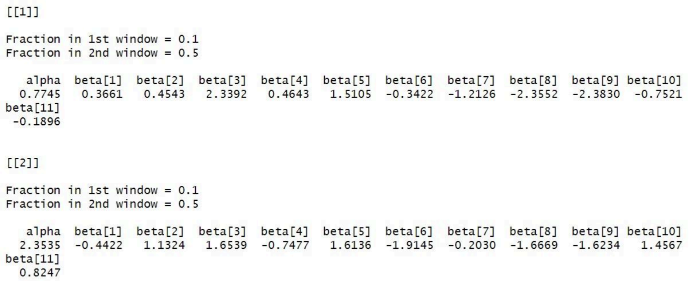
```
       alpha      beta[1]      beta[2]      beta[3]      beta[4]      beta[5]      beta[6]      beta[7] 
39590.142252 20041.113776  8580.243974     4.240609 13071.314589 29571.656361 23983.307676  8081.815915 
     beta[8]      beta[9]     beta[10]     beta[11] 
    4.358880     4.329041 34098.390767  8439.571041 
```

The Effective Sample Size also show that most of the chain provides sufficient independent information about the parameter, so that the posterior estimates (mean, standard deviation, credible intervals) can be trusted. But beta[3], beta[8], and beta[9] indicating possible autocorrelation or poor mixing for this parameter.



```
Potential scale reduction factors:

         Point est. Upper C.I.
alpha           1.0        1.0
beta[1]         1.0        1.0
beta[2]         1.0        1.0
beta[3]        12.4       34.3
beta[4]         1.0        1.0
beta[5]         1.0        1.0
beta[6]         1.0        1.0
beta[7]         1.0        1.0
beta[8]        12.4       34.3
beta[9]        12.4       34.3
beta[10]        1.0        1.0
beta[11]        1.0        1.0

Multivariate psrf

7.51
```

Gelman-Rubin diagnostic shows that almost all of the variable above has PSRF ~ 1and that shows the parameters are likely converged, but for beta[3] (Sqft Living), beta[8] (Sqft Above), and beta[9] (Sqft Basement) they have a high Point Estimates around 12.0 and Upper C.I. values exceeding 10. These are clear indicators of poor convergence for these parameters. This means they are highly correlated with one another (multicollinearity). But, the high multivariate PSRF shows that the chain in some parameter are not converged. Even the univariate PSRF individually are near 1, the multivariate PSRF shows that the joint posterior may reveal some convergency problem that univariate diagnostic miss.



```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10] 
  0.7745   0.3661   0.4543   2.3392   0.4643   1.5105  -0.3422  -1.2126  -2.3552  -2.3830  -0.7521 
beta[11] 
 -0.1896 


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10] 
  2.3535  -0.4422   1.1324   1.6539  -0.7477   1.6136  -1.9145  -0.2030  -1.6669  -1.6234   1.4567 
beta[11] 
  0.8247 
```

The **Geweke diagnostic** compares the means of the MCMC chain's early and late samples, with z-scores near 0 indicating good convergence. In the results, most parameters show reasonable convergence, but beta[3], beta[8], and beta[9] exhibit higher z-scores, suggesting poor convergence.

Bayesian Beta Regression

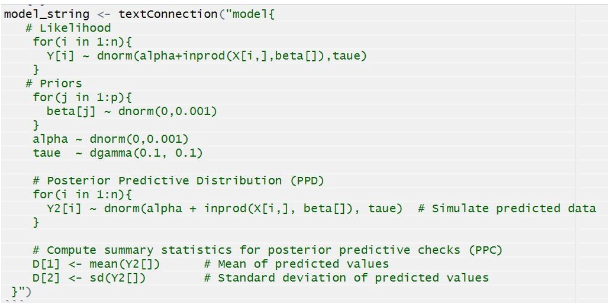
Posterior:

$$Yi|\beta,\ r \sim Beta(r.qi,\ r(1\ -\ qi))$$

Prior:

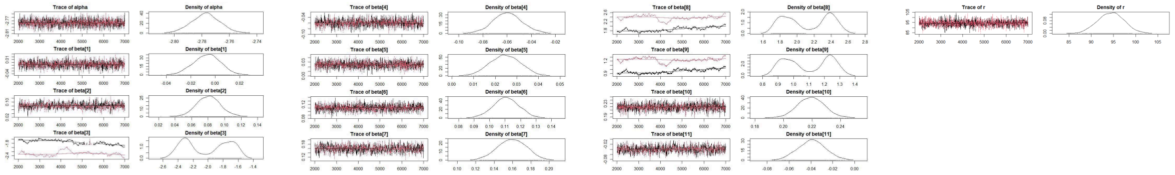
$$\beta|\sigma^2 \sim Normal(\mu, \sigma^2\Omega)$$

Where the intercept $\beta0 \sim Normal(0,\ 1000)$ and the variance $\sigma \sim Gamma(0.1,\ 0.1)$

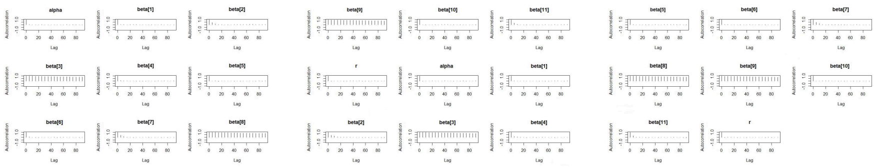


```
model_string <- textConnection("model{
   # Likelihood
    for(i in 1:n){
      Y[i] ~ dnorm(alpha+inprod(X[i,],beta[]),taue)
    }
   # Priors
    for(j in 1:p){
      beta[j] ~ dnorm(0,0.001)
    }
    alpha ~ dnorm(0,0.001)
    taue  ~ dgamma(0.1, 0.1)

    # Posterior Predictive Distribution (PPD)
    for(i in 1:n){
      Y2[i] ~ dnorm(alpha + inprod(X[i,], beta[]), taue)  # Simulate predicted data
    }

    # Compute summary statistics for posterior predictive checks (PPC)
    D[1] <- mean(Y2[])      # Mean of predicted values
    D[2] <- sd(Y2[])        # Standard deviation of predicted values
 }")
```

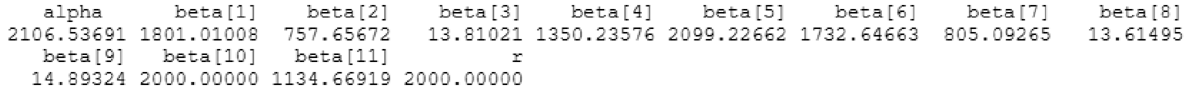
Convergency test:



Here are the graphics diagnostic where we use the chains = 2 and both of the model iteration are convergent, except for beta[3] (sqft_living), beta[8] (sqft_above), beta[9] (sqft_basement). In which they have double peak and the data is dispersion.

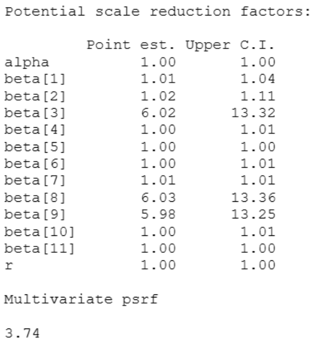


For most parameters, the **autocorrelation** indicates excellent mixing of the MCMC chains and efficient parameter space exploration. Most of the beta (predictor variable) shows that the autocorrelation is low as shown as the data having a small elevation/peak. However, beta[3], beta[8], and beta[9] stand out with consistently high autocorrelation across all lags, which suggests poor mixing for this parameter.



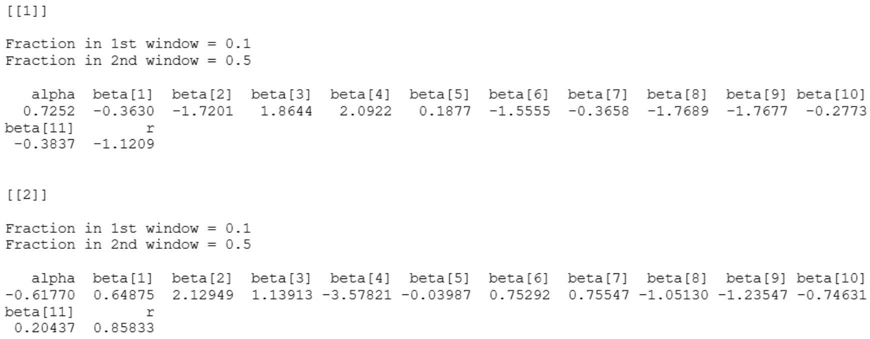
```
      alpha     beta[1]     beta[2]     beta[3]     beta[4]     beta[5]     beta[6]     beta[7]     beta[8] 
 2106.53691  1801.01008   757.65672    13.81021  1350.23576  2099.22662  1732.64663   805.09265    13.61495 
    beta[9]    beta[10]    beta[11]           r 
   14.89324  2000.00000  1134.66919  2000.00000 
```

The Effective Sample Size also show that for most parameters the ESS is relatively high, indicating good convergence and reliable sampling for these parameters. Some values (beta[3] and beta[8]) have much lower ESS values (~13), which has significant difference with other variable. The ESS for r is 2000, which is likely the total number of posterior samples. This suggests rrr has very good mixing and no issues.



```
Potential scale reduction factors:

         Point est. Upper C.I.
alpha          1.00       1.00
beta[1]        1.01       1.04
beta[2]        1.02       1.11
beta[3]        6.02      13.32
beta[4]        1.00       1.01
beta[5]        1.00       1.00
beta[6]        1.00       1.01
beta[7]        1.01       1.01
beta[8]        6.03      13.36
beta[9]        5.98      13.25
beta[10]       1.00       1.01
beta[11]       1.00       1.00
r              1.00       1.00

Multivariate psrf

3.74
```

Gelman-Rubin diagnostic shows that almost all of the variable above has PSRF ~ 1 and that shows the parameters are likely converged, but for beta[3] (Sqft Living), beta[8] (Sqft Above), and beta[9] (Sqft Basement) they have a high Point Estimates around 13.0 and Upper C.I. values exceeding 10. These are clear indicators of poor convergence for these parameters. This means they are highly correlated with one another (multicollinearity). But other varibale stay within the range of [1,1.1] showing the convergency of the variable. The multivariate PSRF also decreasing rather than the linear regression model.



```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5 

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10] 
  0.7252  -0.3630  -1.7201   1.8644   2.0922   0.1877  -1.5555  -0.3658  -1.7689  -1.7677  -0.2773 
beta[11]        r 
 -0.3837  -1.1209 


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5 

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10] 
-0.61770  0.64875  2.12949  1.13913 -3.57821 -0.03987  0.75292  0.75547 -1.05130 -1.23547 -0.74631 
beta[11]        r 
 0.20437  0.85833 
```

The **Geweke diagnostic** compares the means of the early and late portions of the MCMC chains. A z-score close to 0 showing good convergence, while higher values indicates poor convergence. In [[1]], showing that beta[3], beta[8], and beta[9], has high z-scores, pointing to poor convergence. And similarly, in [[2]], beta[3], beta[8], and beta[9], deviate more significantly. **r** Values**,** `[[1]]: r = -1.1209` and `[[2]]: r = 0.85833,` represent correlation or other summary statistics but do not directly indicate convergence.

Results

DIC test

Bayesian Linear Regression	Bayesian Beta Regression
[1] 1112.348	[1] 2467.758

The Bayesian Linear Regression is preferable for prediction since it has a significantly lower DIC, indicating a better trade-off between model complexity and goodness of fit. Bayesian Beta Regression may not fit the data as well in this context.

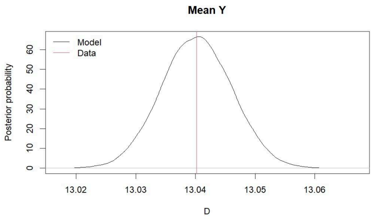
WAIC test

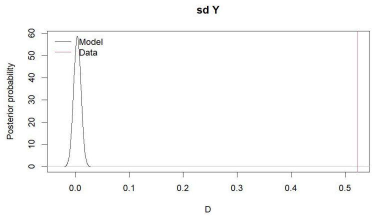
Bayesian Linear Regression	Bayesian Beta Regression

[1] 397.511596580712	[1] 548.129753922

The Bayesian Linear Regression model is better for prediction since it has a lower WAIC value, indicating superior fit and predictive performance compared to the Bayesian Beta Regression model.

The Posterior Predictive Checks:



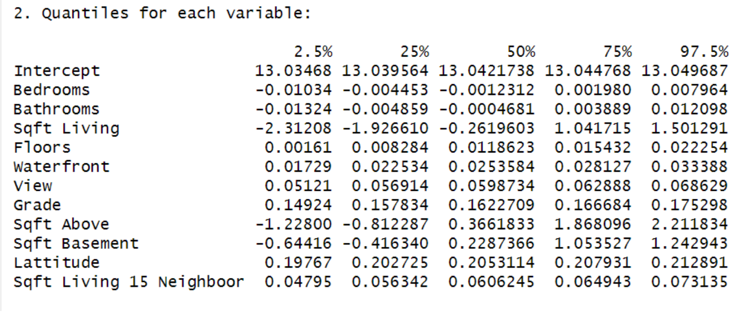


The mean posterior predictive check shows that the model accurately captures the observed data, as the observed mean aligns well with the model's posterior predictive distribution. This shows that the model's assumptions and parameter estimates are reasonable for the data.

On the other hand, the standard deviation posterior predictive check shows a narrow distribution of the model, in which the posterior distribution is very concentrated around small values of D, suggesting low uncertainty and strong confidence in small values of D. The red line is far from the model's peak, indicating a significant difference between the model's predicted D and the actual value observed in the data.

Conclusion

Based on the comparison test, we find that the Bayesian Linear Regression model have a higher accuracy in predicting the housing price rather than Bayesian Beta Regression.



```
2. Quantiles for each variable:

                            2.5%       25%        50%       75%     97.5%
Intercept               13.03468 13.039564 13.0421738 13.044768 13.049687
Bedrooms                -0.01034 -0.004453 -0.0012312  0.001980  0.007964
Bathrooms               -0.01324 -0.004859 -0.0004681  0.003889  0.012098
Sqft Living             -2.31208 -1.926610 -0.2619603  1.041715  1.501291
Floors                   0.00161  0.008284  0.0118623  0.015432  0.022254
Waterfront               0.01729  0.022534  0.0253584  0.028127  0.033388
View                     0.05121  0.056914  0.0598734  0.062888  0.068629
Grade                    0.14924  0.157834  0.1622709  0.166684  0.175298
Sqft Above              -1.22800 -0.812287  0.3661833  1.868096  2.211834
Sqft Basement           -0.64416 -0.416340  0.2287366  1.053527  1.242943
Lattitude                0.19767  0.202725  0.2053114  0.207931  0.212891
Sqft Living 15 Neighboor 0.04795  0.056342  0.0606245  0.064943  0.073135
```

From the **summary** of the dataset above, we can conclude that with 95% credible interval, we can say:

1. Bedrooms, bathrooms, sqft living, sqft above, and sqft basement has no significant influence toward the price of the housing.
2. Floors, waterfront, view, grade, and latitude: These have positive median values, indicating a positive relationship with the response variable. Their 95% credible intervals are mostly above zero, suggesting stronger evidence for a positive effect.