Bayesian Regression :From Scratch

Abhishek Kumar
ai21btech11003@iith.ac.in

Ranveer Sahu
es21btech11025@iith.ac.in

Abstract

The Project delves into Simple Linear Regression from the both Frequentist and Bayesian Approches establishing the correctness/Usefulness/Scratch of Metropolis Hastings Algorithm by the agreement of the both approches .

1. Data Description

Keeping in mind the computational intensity of the MH algorithm and its convergence(needs large number of iterations), we take a very simple dataset . It has one column (Sales) as the label and one column feature (TV budget expenditure) . We want to regress Sales Vs TV budget expenditure through the simple linear regression.

Linearity:-Before modelling the Sales and expenditure dependence linearly , it makes sense to investigate a bit about their linear/monotonic dependence. For it ,we do scatter plot and find pearson correlation coefficient(captures linearity/monotonicity) and spearman correlation cefficient.The pearson correlation coefficient between variables feature and label is $r = 0.794562$ and spearman correlation coefficient is $r = 0.8006144$.

Regression Assumptions:-We now do the important checks (assumptions) before doing regression.We first check the normality of the label data through histogram . Then we do QQ plot . From figure 2 and figure 3(line close to the line (y=x)) , data label appears do be normal .To Conclude , we come to **Hypothesis Tests** (Shapiro Wilk and Anderson darling test) with level of significance = 0.05. Shapiro wilk test and AD test both yields a smaller p-value . We failed to accept the Null hypothesis of data being normal .

Box cox tranformation to gaussianize the data label .

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln(y), & \text{if } \lambda = 0. \end{cases}$$

where y is the original variable and λ is the transformation

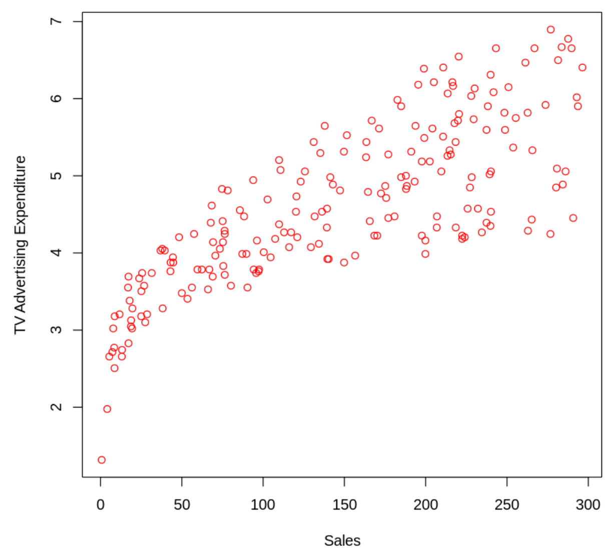

Figure 1. Linearity(Scatter Plot)

parameter. Basically this approach tries varies values of λ and check the gaussian likelihood the tranformed data. This tranformation give $\lambda = 0.5858$.We apply the suggested tranformation to data label. After it , p-values of both shapiro wilk and AD test are greater than 0.05 . we successfully gaussianed it .

Zero mean of the data label assumption : The data is not zero mean .we shift the data to make its mean zero.We assume data label to be Homoscedastic.we are now done with all the assumptions of the regression

2. Frequentist Mean Regression

It starts assuming the model parameters fixed and it aims to find their values.The linear regression model for the ith data point can be written as:

$$y_i = a + b \cdot x_i + e_i,$$

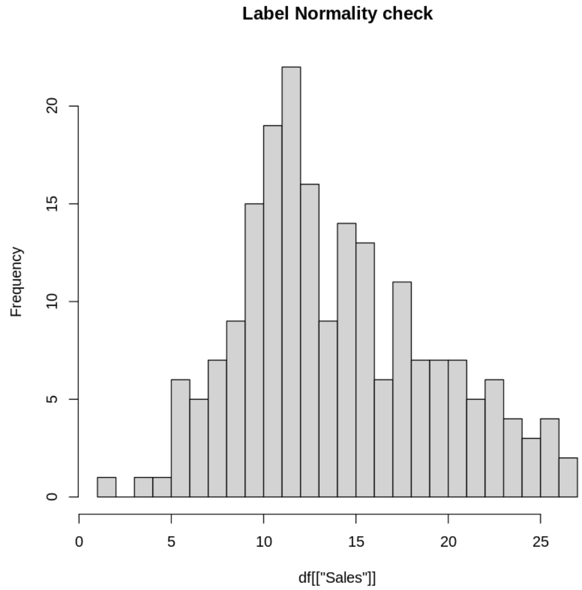

Figure 2. Histogram

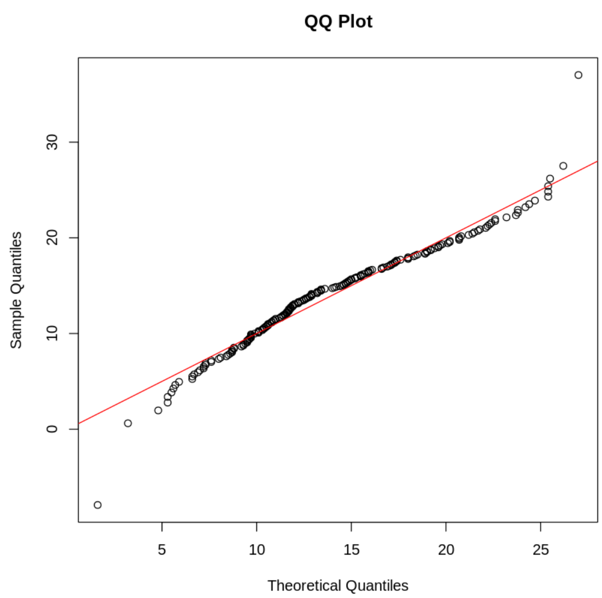

Figure 3. QQ plot

where y_i is the response variable for the ith data point, x_i is the predictor variable for the ith data point, a is the intercept, b is the slope, and e_i is the error term for the ith data point.

Additionally,

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$y_i \overset{\text{iid}}{\sim} \mathcal{N}(a + b * x_i, \sigma^2)$$

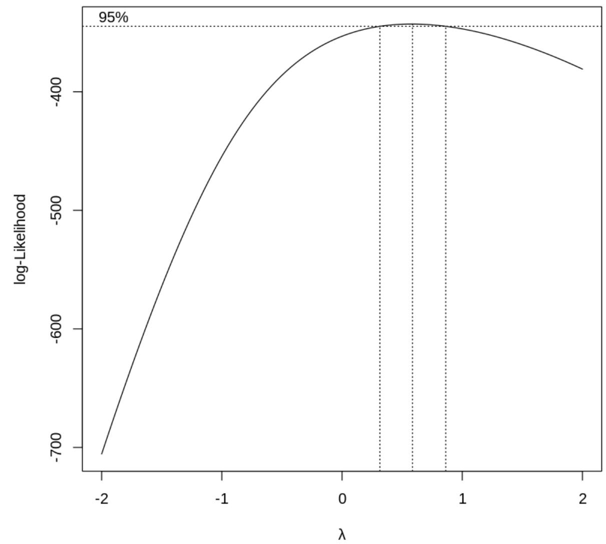

Figure 4. Box Cox Transformation

To find the **least squares estimates** $\hat{a}$ and $\hat{b}$, we optimize the sum of squared residuals:

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0,$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0.$$

The least squares estimates $\hat{a}$ and $\hat{b}$ are given by:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{Cov}[x, y]}{\mathrm{Var}[x]} = r_{xy}\frac{s_y}{s_x},$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

where r_{xy} is the sample correlation coefficient between x and y, s_x is the standard deviation of x, and s_y is the standard deviation of y.We obtain the values of $\hat{a}$ and $\hat{b}$ as intercept and TV in figure 4 respectively.Results are shown in figure 6 and figure 7. Till now we use R programming to get the above results(rwork.ipynb).We now shift to Python for the rest of the works(Pythonwork.ipynb).

3. Bayesian Regression - Core of the Project

It starts with the assumption that a and b and σ are not fixed parameters .We should start bayesian analysis with likelihood because everything follows its own path itself afterwards .

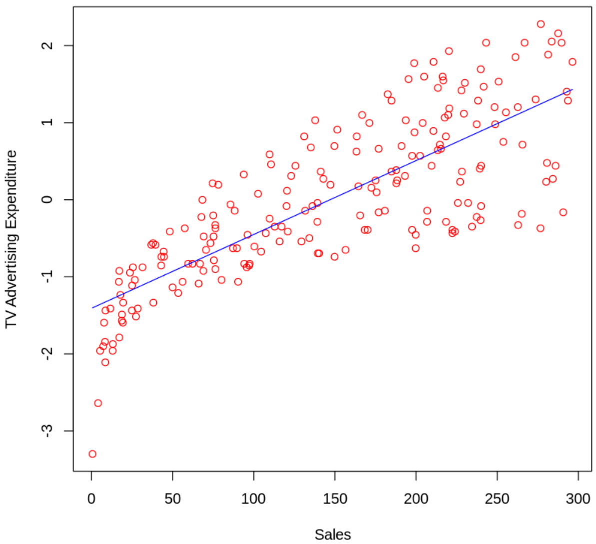

Figure 5. Fitted line

```
lm(formula = Y ~ TV, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89627 -0.37579  0.00599  0.45446  1.27556

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.4094421  0.0885788  -15.91   <2e-16 ***
TV           0.0095853  0.0005206   18.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6305 on 198 degrees of freedom
Multiple R-squared:  0.6313,    Adjusted R-squared:  0.6295
F-statistic: 339.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Figure 6. Results

3.1. Likelihood

$$f(y_1 \dots, y_n \mid a, b, \sigma^2) = \prod_{i=1}^{n} f(y_i \mid a, b, \sigma^2) \tag{1}$$

$$\implies f(y_1 \dots, y_n \mid a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma^2}} \tag{2}$$

$$\implies L(a, b, \sigma^2 \mid y_1 \dots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma^2}} \tag{3}$$

3.2. Priors

This is one of the important steps for the convergence of the MCMC algorithm .We assume (a,b) and σ^2 to be independent of each other.We choose priors with help of likelihood function to get the best priors in the context.
For prior of the parameter (a,b) ,we represent the $(a, b) \mid \sigma^2$ as 2d gaussian:

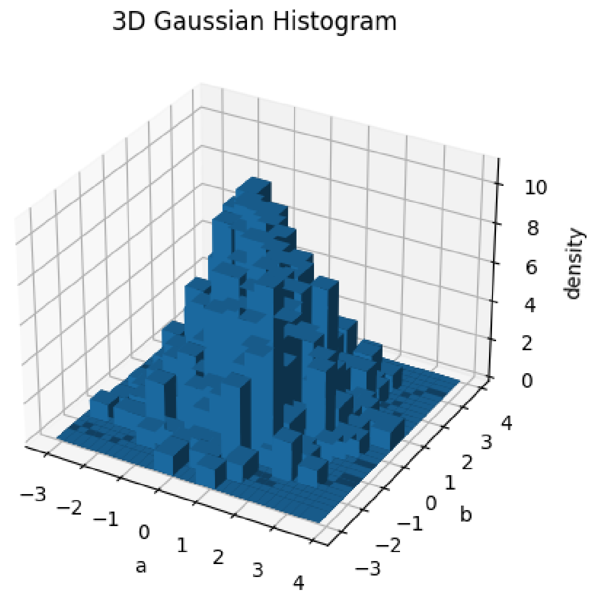

Figure 7. Prior for $\boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$

$$f(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right)$$

where $\boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, d=2 Prior for (a,b) is shown in figure 7.

For prior of the σ^2 , we assume (a, b) to be known .From likelihood function

$$f(\sigma^2 \mid a, b) = k * \sigma^{-n} e^{-\frac{1}{2\sigma^2} c} \tag{4}$$

where k and c are some constant .This pdf look very similar to Inverse Gamma distribution is given by:

$$f(\sigma^2 \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha - 1} e^{-\frac{\beta}{\sigma^2}}, \quad \sigma > 0$$

where σ^2 is the random variable, α and β are the shape and scale parameters, respectively, and $\Gamma(\alpha)$ is the Gamma function.we take $\alpha = 2$, $\beta = 2$(unbiased).This prior distrobution is shown in the figure 8.

3.3. Proposal Distribution:-

$$\boldsymbol{\theta'} \sim g(\boldsymbol{\theta'} \mid \boldsymbol{\theta}) \tag{5}$$

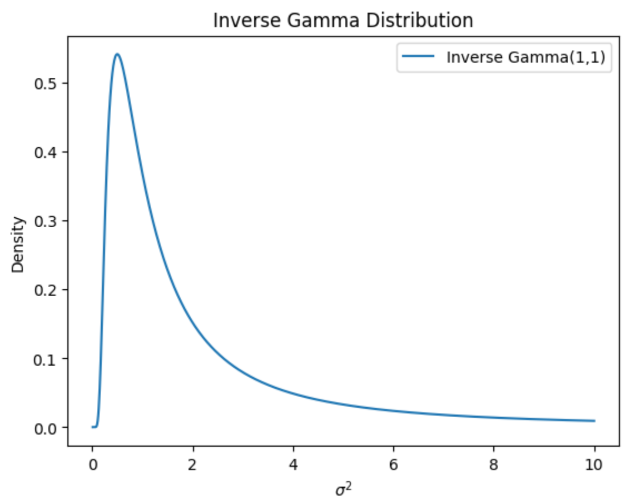

Figure 8. Prior for σ^2

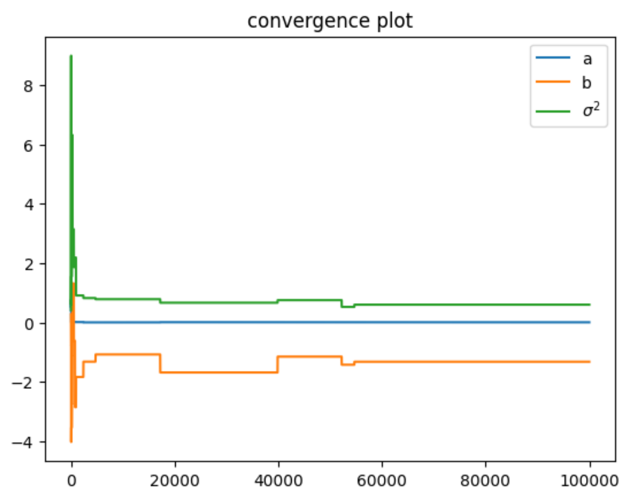

Figure 9. Convergence

we define g (pdf) as follows:-

$$\boldsymbol{\theta}' = \begin{bmatrix} a' \\ b' \\ \sigma^{2\prime} \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} a' \\ b' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right) \tag{7}$$

$$\sigma^{2\prime} \sim \text{Inv-Gamma}(2, 2) \tag{8}$$

3.4. Metro -Polis Hastings Algorithm(Markov Chain Monte Carlo):-

1. Initialize the starting state $\boldsymbol{\theta}_0$.
2. Set the number of iterations T.
3. Set the proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$.
4. For $t = 1$ to T:
   (a) Sample a candidate state $\boldsymbol{\theta}'$ from the proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_{t-1})$.
   (b) Calculate the proposal ratio:

   $$\text{Proposal Ratio} = \frac{q(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_{t-1})}$$

   (c) Calculate the likelihood ratio:

   $$\text{Likelihood Ratio} = \frac{f(y_1 \ldots, y_n \mid \boldsymbol{\theta}')}{f(y_1 \ldots, y_n \mid \boldsymbol{\theta}_{t-1})}$$

   (d) Calculate the prior ratio (if applicable):

   $$\text{Prior Ratio} = \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_{t-1})}$$

   (e) Calculate the acceptance probability:

   $$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}_{t-1}) = \min\left(1, \frac{\text{Likelihood} \times \text{Prior}}{\text{Proposal}}\right)$$

   (f) Generate a uniform random number u from $[0, 1]$.
   (g) If $u \leq \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}_{t-1})$, set $\boldsymbol{\theta}_t = \boldsymbol{\theta}'$; otherwise, set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.
5. we can drop the samples collected in the intial phase of the algorithm(because they are not from the posterior distribtuion because of convergence yet to come) called burn in samples.
6. since we are using mcmc to sample , Hence the samples are from 1st order markov pdf .To distort this dependence of next state on the previous state (not completely) we use remove each (9/10) th of the 10 samples .The period of removal is known as thinning period.

4. MCMC Convergence

We apply log to each of the pdf ratio for better numerical stability and hence better convergence.

The convergence of the MCMC algorithm can be seen in figure 9 (key point to observe the convergence is that the there is no change in the values of the parameters (a,b,σ^2) after 60,000 th iteration .

Plots for the distributions for a , b, c can be seen in the figures 10 ,11,12 respectively.The posterior histograms are not continuous , reason is that the samples are not perfectly independent.

5. The Agreement

Now we come to end . We here want to show the agreement of the two approches .This agreement relies on the key fact the likelihood fucntion .

$$f(a, b, \sigma^2 \mid y_1, \ldots, y_n) \propto f(a, b, \sigma^2) * L(a, b, \sigma^2 \mid y_1, \ldots, y_n) \tag{9}$$

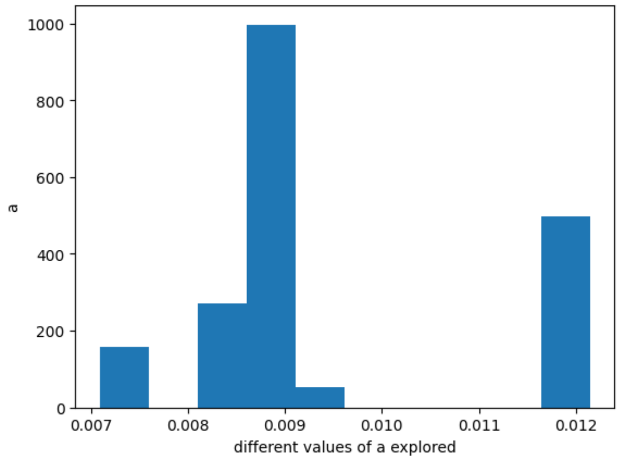

Figure 10. posterior for a

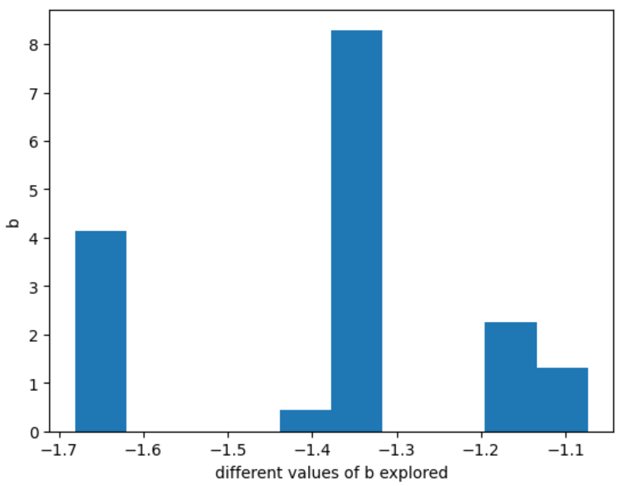

Figure 11. posterior for b

Above equation no. 9 clearly represents that posterior is formed with help of weights given by prior and likelihood(pdfs) to the parameters values (a, b, σ^2).No matter what prior you started yours bayesian analysis , if you have good amount of data the effect of prior is nullified and the only effect of its remain is that the algorithm takes more or less no of iterations to converge .Because of the large amount of the data , the process becomes data driven and in total MH Algorithm give more and more weights to the likelihood function .

Thus , the algorithm tend towards then likelihood function

Hence the algorithm generates such more and more samples which maximises the likelihood function(high pdf).Thus , theorectically maximum likelihood estimates

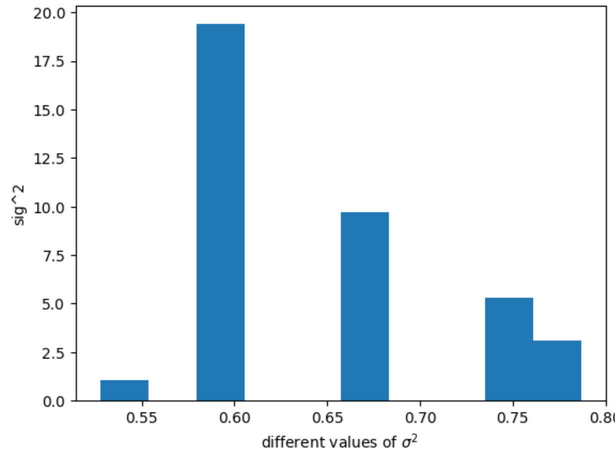

Figure 12. posterior for c

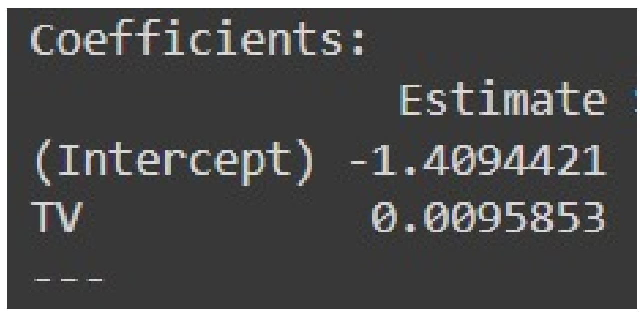
Coefficients:
Estimate
(Intercept) -1.4094421
TV 0.0095853

Figure 13. parameters values from the frequent regression

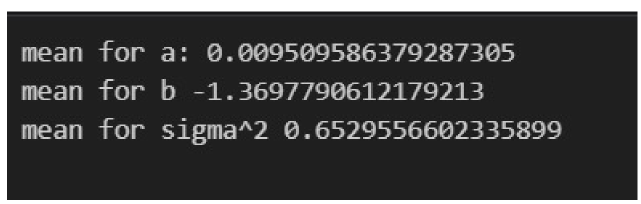
mean for a: 0.009509586379287305
mean for b -1.3697790612179213
mean for sigma^2 0.6529556602335899

Figure 14. mean value of parameters(MCMC)

for the parameters (a,b,σ^2) should be the expected values from the posterior distribution most times. Expected value of the posterior distribution is nothing but the mean of the distribution .

In total , we must be able to represent this essential result in ours work where we use bayesian and frequentist approches to regress because the frequentist approch from the tries to find such parameters which maximises the likelihood function (least squares method).

5.1. Demonstration of the agreement / Concrete:-

We can see clearly that the values of a and b from the both approches are very close to each other in figure 14,13.

References

[1] Kaggle. Simple Dataset. https://www.kaggle.com/datasets/devzohaib/tvmarketingcsv?resource=download.

[2] Wikipedia. Probability Distributions. https://en.wikipedia.org/wiki/List_of_probability_distributions, .

[3] Wikipedia. Mean Regression. https://en.wikipedia.org/wiki/Regression_toward_the_mean, .

[4] Wikipedia. Metropolis Hastings Algorihtm. https://en.wikipedia.org/wiki/MetropolisHastings_algorithm, .