Comparing Multiple Linear Regression using Frequentist and Bayesian approaches on Energy Efficiency data

Viviane Callier, Ben Graf, Cong Zhang

Introduction

Heating load, as one important measure of energy efficiency, refers to the amount of heat energy needed to maintain ambient temperature in an acceptable range. Energy efficiency reduces pollution, lowers building management expenses, and decreases demand for energy imports. In this study, we conduct multiple linear regression on energy efficiency data by adopting two different statistical methods, namely Frequentist approach and Bayesian approach.

Data

We used the Energy Efficiency Data Set from Tsanas and Xifara, 2012, which has 768 observations. The Heating Load variable was selected as the response, and four potential explanatory variables – Relative Compactness, Wall Area, Overall Height, and Glazing Area – were included. The descriptive statistics for all the variables included in the current study are summarized in Table 1.

Table 1: Descriptive statistics for all variables

Variables	N	Mean	SD	Min	Max
Heating Load	768	22.31	10.09	6.01	43.10
Relative Compactness	768	0.76	0.11	0.62	0.98
Wall Area	768	318.50	43.63	245.00	416.50
Overall Height	768	5.25	1.75	3.50	7.00
Glazing Area	768	0.23	0.13	0.00	0.40

Research Question

Do the Frequentist and Bayesian approaches to multiple linear regression provide similar coefficient estimates? Is one approach superior to the other?

Frequentist Analysis

We propose the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \text{ where } \varepsilon \sim \mathrm{N}(0, \sigma^2), p = 4$$

The matrix form for this model is:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

where
$\mathrm{Y} =$ dependent variable $(\mathrm{n} \times 1)$
$\mathrm{X} =$ the matrix of independent variables $(\mathrm{n} \times (\mathrm{p} + 1))$
$\beta =$ vector of regression model parameters $((\mathrm{p} + 1) \times 1)$
$\varepsilon =$ vector of errors $(\mathrm{n} \times 1)$
$n = 768$
$p = 4$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The $\boldsymbol{\beta}$ parameter is estimated by:

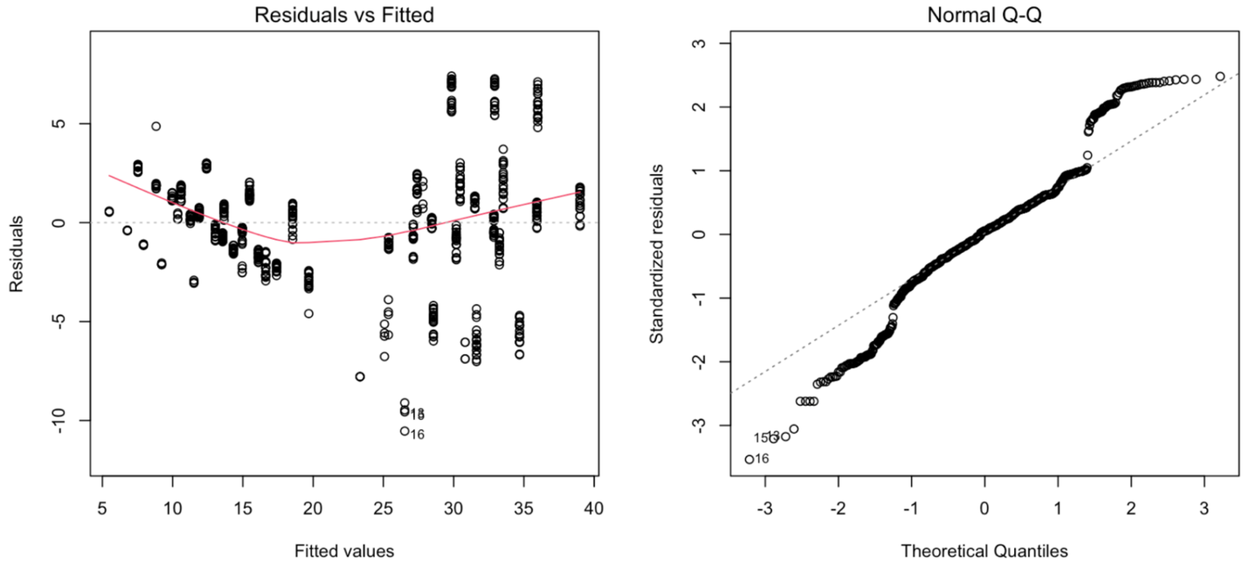
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

The results of multiple linear regression model are summarized in Table 2.

Table 2: Multiple linear regression model results

Variables	estimate	std_error	t-value	p_value	lower_ci	upper_ci
intercept	-11.953	2.696	-4.434	0.000***	-17.245	-6.661
Relative Compactness	-14.532	3.109	-4.674	0.000***	-20.636	-8.429
Wall Area	0.035	0.004	7.936	0.000***	0.026	0.044
Overall Height	5.607	0.192	29.266	0.000***	5.231	5.983
Glazing Area	20.438	0.812	25.180	0.000*	18.845	22.031

We also checked the constant variance and normality assumptions. The plot of residuals vs fitted values shows a non-constant variance, and it is obvious from the Normal Q-Q plot that the normality assumption is violated as well.



In the Frequentist approach, the estimate of σ is Root MSE, which is 2.995.

Bayesian Analysis

For semi-conjugate (independent) priors, we select:

$$\begin{aligned} \boldsymbol{\beta} \quad &\sim N(\boldsymbol{b_0}, \boldsymbol{\Sigma_0}) \\ 1/\sigma^2 \quad &\sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

The likelihood function is:

$$L(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i)^2\right]$$

Because Posterior $\propto$ Likelihood $\times$ Prior, the conditional posterior distributions for **β** and $1/\sigma^2$ are:

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{Y} \sim N(\boldsymbol{b_n}, \boldsymbol{\Sigma}_n)$$

$$1/\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{Y} \sim \text{Gamma}\left((\nu_0 + n)/2, \left(\nu_0\sigma_0^2 + \sum^{n} (\boldsymbol{Y_i} - \boldsymbol{\beta X_i})^2\right)/2\right)$$

where

$$\begin{aligned} \boldsymbol{b_n} &= \left(\boldsymbol{\Sigma_0^{-1}} + \boldsymbol{X}^T\boldsymbol{X}/\sigma^2\right)^{-1}\left(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{b_0} + \boldsymbol{X^T}\boldsymbol{Y}/\sigma^2\right) \\ \boldsymbol{\Sigma_n} &= \left(\boldsymbol{\Sigma_0^{-1}} + \boldsymbol{X}^T\boldsymbol{X}/\sigma^2\right)^{-1} \end{aligned}$$

The priors described at the beginning of this section can be simplified for coding purposes to the following:

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y} \sim MVN(\mathbf{a}, \sigma^2 \mathbf{R})$$

$$\frac{1}{\sigma^2} \mid \boldsymbol{\beta}, \mathbf{Y} \sim \text{Gamma}(a_0, b_0)$$

1. Noninformative prior (Jeffreys prior)

Because we do not have any prior expertise regarding the parameters, we began with the noninformative prior in our study:

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

This corresponds to the same priors listed above with hyperparameters:

$$\mathbf{a} = (0,0,0,0,0),\ \mathbf{R} = \begin{pmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{pmatrix}, a_0 = 0, b_0 = 0.$$

We obtained the initial value for $\boldsymbol{\beta}$ and $\frac{1}{\sigma^2}$ by MLE estimates :

$$\boldsymbol{\beta}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\frac{1}{\sigma^2}^{(0)} = \frac{n}{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}$$

We generated 5000 iterations through the Gibbs Sampling method. The estimated $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are shown in Table 3. The 95% Highest Posterior Density (HPD) credible intervals are also included. All of these estimates allow 50% of the Gibbs samples to be treated as burn-in and do not include them in the calculation.

Table 3: Bayesian Regression model, prior 1

	Estimates	95% HPD Lower	95% HPD Upper
$\widehat{\boldsymbol{\beta}}_0$	-12.222	-17.240	-7.364
$\widehat{\boldsymbol{\beta}}_1$	-14.032	-19.583	-8.188
$\widehat{\boldsymbol{\beta}}_2$	0.03520	0.02677	0.04329
$\widehat{\boldsymbol{\beta}}_3$	5.580	5.234	5.951
$\widehat{\boldsymbol{\beta}}_4$	20.254	18.693	21.919
$\widehat{\boldsymbol{\sigma}}$	3.003	2.859	3.152

The estimated parameters for the 5000 iterations are plotted in Figure 1. All converge almost instantly, and none of the values straddle zero, which is consistent with the 95% HPD results in Table 3.

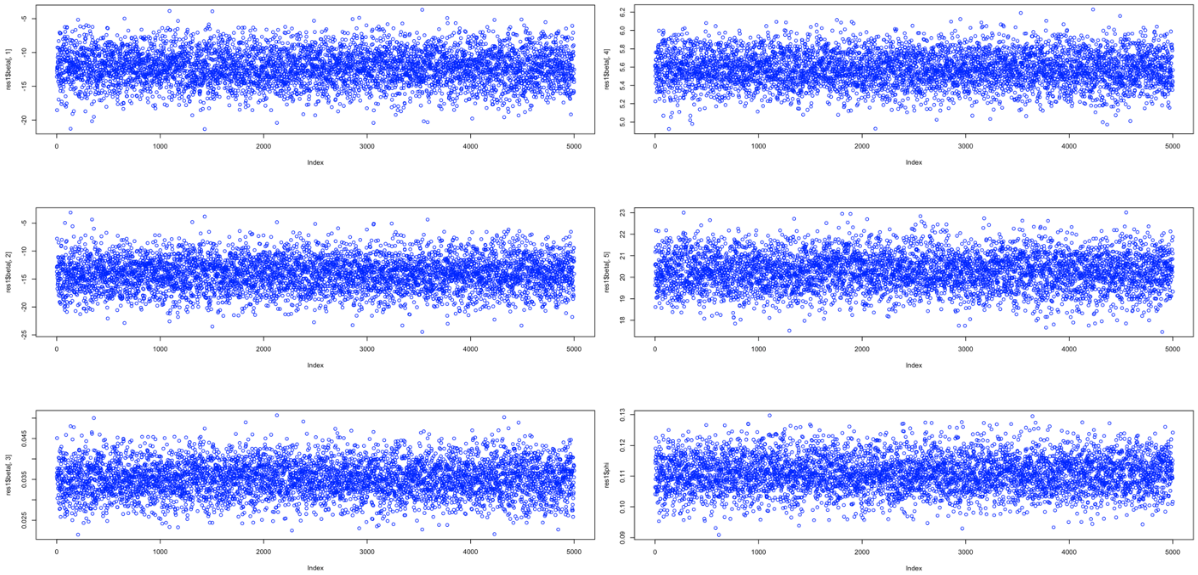


Figure 1: Estimated βs and $1/\sigma^2$

Root MSE for the model is 2.995. Autocorrelation is not an issue, as the lowest effective size is 2245 (out of a possible 2501). In running the Geweke diagnostic, all values are less than ± 2, indicating that lack of convergence has not been detected.

2. Prior from Frequentist values

We next considered a prior reminiscent of the values obtained from the Frequentist approach. The specific hyperparameters used were:

$$\mathbf{a} = (-12, -15, 0, 6, 20),\ \ \mathbf{R} = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}, a_0 = 2, b_0 = 4.$$

The procedure was otherwise the same. In this case, the resultant estimates were as follows:

Table 4: Bayesian Regression model, prior 2

	Estimates	95% HPD Lower	95% HPD Upper
$\widehat{\boldsymbol{\beta}}_0$	-11.955	-16.532	-7.709
$\widehat{\boldsymbol{\beta}}_1$	-14.557	-19.526	-9.438
$\widehat{\boldsymbol{\beta}}_2$	0.03501	0.02775	0.04279
$\widehat{\boldsymbol{\beta}}_3$	5.609	5.303	5.951
$\widehat{\boldsymbol{\beta}}_4$	20.421	18.819	21.950
$\widehat{\boldsymbol{\sigma}}$	2.979	2.837	3.132

We will not repeat the plots of the parameter estimates here, as they are very similar to Figure 1 and can be found in the Appendix. Root MSE is again 2.995. Autocorrelation is even less of an issue, with the lowest effective size 2474. The Geweke diagnostic again does not detect lack of convergence.

3. Prior from thin air

For our third prior, we picked numbers haphazardly to see whether we could achieve similar results. The specific hyperparameters used were:

$$\mathbf{a} = (5,3,-2,1,-7),\ \mathbf{R} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}, a_0 = 3, b_0 = 1.$$

The procedure was again the same. This led to the following coefficient estimates:

Table 5: Bayesian Regression model, prior 3

	Estimates	95% HPD Lower	95% HPD Upper
$\widehat{\boldsymbol{\beta}}_0$	-12.169	-16.177	-7.958
$\widehat{\boldsymbol{\beta}}_1$	-13.229	-17.945	-8.767
$\widehat{\boldsymbol{\beta}}_2$	0.03473	0.02737	0.04168
$\widehat{\boldsymbol{\beta}}_3$	5.524	5.222	5.823
$\widehat{\boldsymbol{\beta}}_4$	19.341	17.789	21.044
$\widehat{\boldsymbol{\sigma}}$	3.114	2.958	3.269

Root MSE was slightly higher at 3.000. Autocorrelation is still no issue, with the lowest effective size 2462. The Geweke diagnostic, with a value of 2.009, suggests the estimate of σ may not quite have converged. Extending the burn-in to 52% (vice 50%) remedied the issue, and the estimate and HPD credible interval in Table 5 above did not change.

Comparisons

Table 6 summarizes the coefficient estimates for the four methods tested in this study and also indicates, for the Bayesian approaches, how far off they are from the Frequentist approach:

Table 6: Comparison of Regression Methods

	Frequentist	Bayesian prior 1	Bayesian prior 2	Bayesian prior 3
$\hat{\beta}_0$	-11.953	-12.222 (2.3%)	-11.955 (0.0%)	-12.169 (1.8%)
$\hat{\beta}_1$	-14.532	-14.032 (3.4%)	-14.557 (0.2%)	-13.229 (9.0%)
$\hat{\beta}_2$	0.03498	0.03520 (0.6%)	0.03501 (0.1%)	0.03473 (0.7%)
$\hat{\beta}_3$	5.607	5.580 (0.5%)	5.609 (0.0%)	5.524 (1.5%)
$\hat{\beta}_4$	20.438	20.254 (0.9%)	20.421 (0.1%)	19.341 (5.4%)
$\hat{\sigma}$	2.995	3.003 (0.3%)	2.979 (0.5%)	3.114 (4.0%)

Bayesian prior 2 has the closest estimates to the Frequentist values, and this makes sense, given that the $\boldsymbol{\beta}$ means for this prior were selected to be very close to the Frequentist coefficient estimates. Prior 1 (Jeffreys) also does quite well, with the largest difference from the Frequentist values being a mere 3.4% off. Prior 3 struggles the most. This is not a surprise, as an "informative" prior was selected, but we chose to select values haphazardly. Still, it is in the right ballpark, with its worst parameter being 9.0% away from its Frequentist counterpart.

Conclusion

In this study, we conducted Frequentist approach and Bayesian approach with three different priors to estimate the coefficients for the multiple linear regression model. We found that, for this dataset, the two approaches produced very similar coefficient estimates. The Bayesian approach was almost identical to the Frequentist approach when highly informative priors were used, but the noninformative (Jeffreys) prior did quite well too.

We also discovered that the normality and constant variance assumptions necessary for linear regression under the Frequentist approach are not satisfied here, so statistical inference is potentially hazardous. For these reasons, the Bayesian method may be the safer approach.

References

Tsanas, Athanasios, and Angeliki Xifara. "Energy Efficiency Data Set." Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK (2012). Downloaded from UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Energy+efficiency.

Tsanas, Athanasios, and Angeliki Xifara. "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools." *Energy and Buildings* 49 (2012): 560-567.

Appendix

Our complete code and analysis can be found in an Appendix included as a separate file.