Final Project

DSC4043/STA4063 - Bayesian Statisctics

E.Vanathey
(S/19/813)

AN APPLICATION OF BAYESIAN STATISTICS TO ENERGY EFFICIENCY DATA

Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya

2025

Abstract

This project uses Bayesian linear regression to predict the heating load of buildings based on their features like size and window area. The Bayesian method helps us understand how certain the predictions are and which features are most important. We calculate the probability that each feature affects heating load and give estimates with ranges showing possible values. The results show that factors such as compactness, wall area, roof area, height, and glazing strongly influence heating load. This approach gives clearer and more useful information to help design energy-efficient buildings

Contents

1 Introduction

- **Background -** Energy use in buildings is a major contributor to overall energy consumption worldwide. Efficient heating systems are essential to reduce costs and environmental impact. Predicting heating load accurately using building features like size, shape, and window area helps design energy-efficient buildings

- **Research Questions.**
  Which building characteristics have the strongest effect on heating load?

  How can a Bayesian regression model improve prediction and understanding of heating load?

  Can variable selection identify the most important factors while accounting for uncertainty?.

- **Objectives.**
  To fit a Bayesian linear regression model to heating load data.

  To estimate the effects and credible intervals of building features.

  To assess variable importance using marginal inclusion probabilities.

  To provide actionable insights for improving building energy efficiency..

2 Methodology

2.1 Description of Dataset

The data used in this study contains information about different buildings and their Heating load. It includes features like Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distribution. The goal is to use these features to predict the heating load needed for each building.

Figure 1: Dataset

```
'data.frame':   768 obs. of  9 variables:
 $ Relative.Compactness     : num  0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...
 $ Surface.Area             : num  514 514 514 514 564 ...
 $ Wall.Area                : num  294 294 294 294 318 ...
 $ Roof.Area                : num  110 110 110 110 122 ...
 $ Overall.Height           : num  7 7 7 7 7 7 7 7 7 7 ...
 $ Orientation              : int  2 3 4 5 2 3 4 5 2 3 ...
 $ Glazing.Area             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Glazing.Area.Distribution: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Heating.Load             : num  15.6 15.6 15.6 15.6 20.8 ...
```

2.2 Exploratory Data Analysis (EDA)

- Checked basic summary statistics like mean, median, and range for all building features and heating load.
- Created scatter plots to visually observe relationships between each predictor and the heating load.
- Calculated correlation coefficients to measure strength and direction of relationships.
- Found strong negative correlation between relative compactness and heating load.
- Observed positive correlation of glazing area and wall area with heating load.
- Examined correlations among predictors to detect multicollinearity issues.

2.3 Bayesian Methods and Models

Bayesian linear regression is used to model the relationship between heating load and the building features. This method lets us estimate the impact of each feature on heating load while considering uncertainty.Also used Bayesian variable selection to find which features are the most important to the model.

2.4 Prior Specification

For model selection used a BIC (Bayesian Information Criterion) prior that helps balance the simplicity and fit of the model. We gave equal initial chances (uniform prior) for different feature combinations when choosing the best model.

2.5 Model Checking

The model checked by looking at credible intervals and the probabilities that each feature should be in the model. Probabilities visualized with plots to see which features are confidently important. The process helped confirm that the model fits well and that the main predictors were correctly identified.

3 Result and Discussion

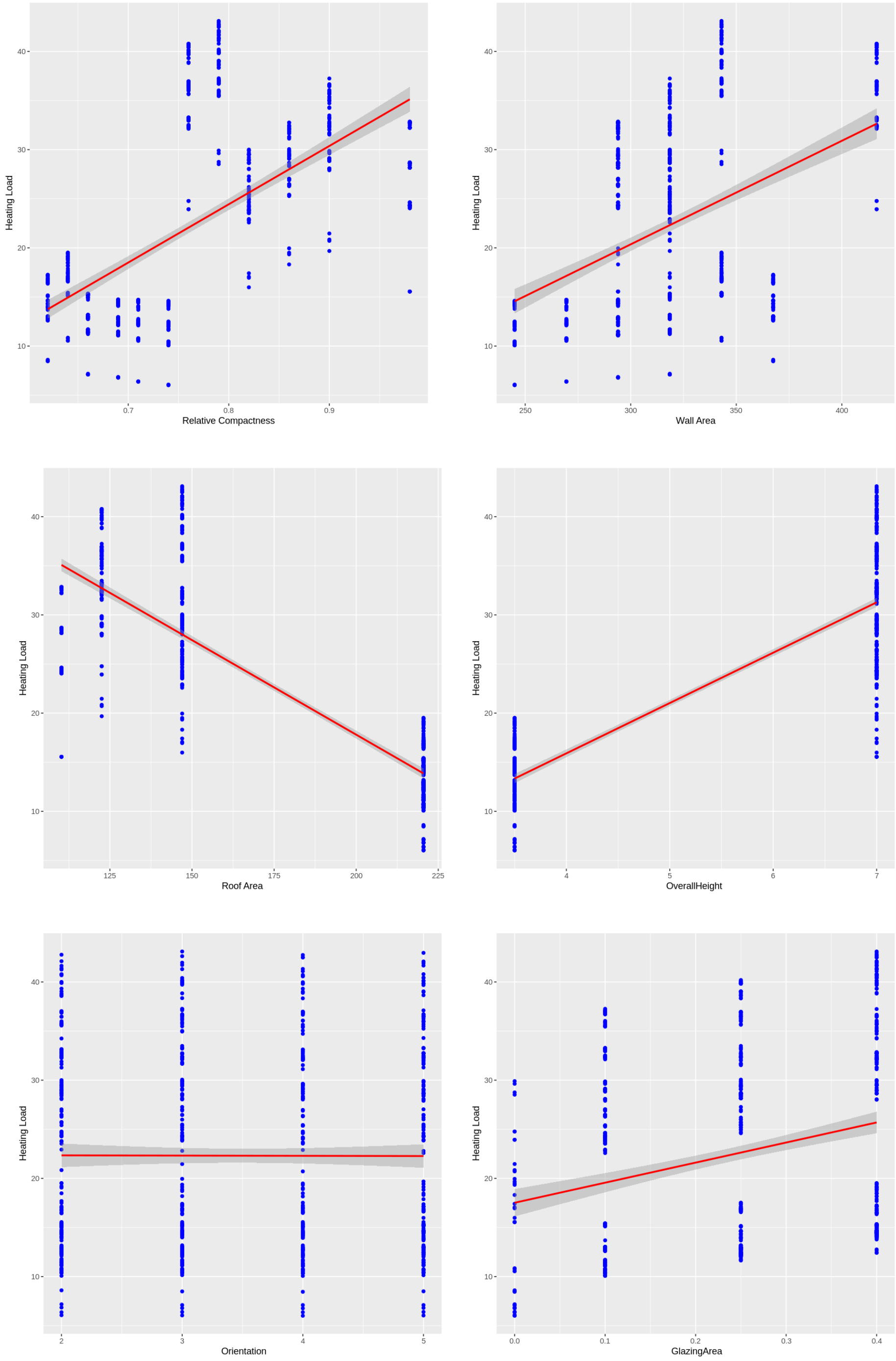
3.1 EDA Findings

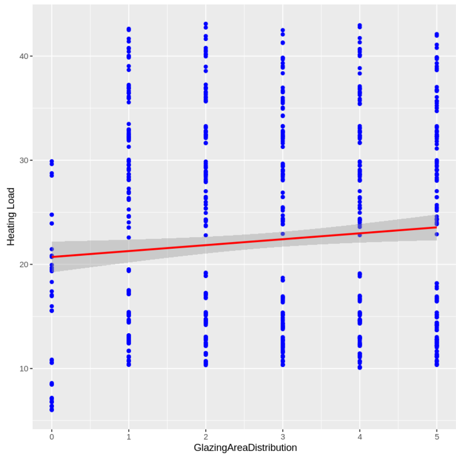
The exploratory data analysis revealed important patterns in the building features affecting heating load. Relative compactness showed a strong negative correlation, indicating that more compact buildings require less heating. Wall area, roof area, overall height, and glazing area had positive correlations with heating load, suggesting larger surfaces and taller buildings increase heating needs. Scatter plots visually confirmed these trends, while orientation and glazing area distribution showed little influence.

Figure 2: Summary

```
Relative.Compactness  Surface.Area     Wall.Area       Roof.Area    
Min.   :0.6200       Min.   :514.5   Min.   :245.0   Min.   :110.2  
1st Qu.:0.6825       1st Qu.:606.4   1st Qu.:294.0   1st Qu.:140.9  
Median :0.7500       Median :673.8   Median :318.5   Median :183.8  
Mean   :0.7642       Mean   :671.7   Mean   :318.5   Mean   :176.6  
3rd Qu.:0.8300       3rd Qu.:741.1   3rd Qu.:343.0   3rd Qu.:220.5  
Max.   :0.9800       Max.   :808.5   Max.   :416.5   Max.   :220.5  
Overall.Height  Orientation     Glazing.Area    Glazing.Area.Distribution
Min.   :3.50   Min.   :2.00   Min.   :0.0000   Min.   :0.000            
1st Qu.:3.50   1st Qu.:2.75   1st Qu.:0.1000   1st Qu.:1.750            
Median :5.25   Median :3.50   Median :0.2500   Median :3.000            
Mean   :5.25   Mean   :3.50   Mean   :0.2344   Mean   :2.812            
3rd Qu.:7.00   3rd Qu.:4.25   3rd Qu.:0.4000   3rd Qu.:4.000            
Max.   :7.00   Max.   :5.00   Max.   :0.4000   Max.   :5.000            
 Heating.Load  
Min.   : 6.01  
1st Qu.:12.99  
Median :18.95  
Mean   :22.31  
3rd Qu.:31.67  
Max.   :43.10  
```

Figure 3: Scatterplots and correlation plot bthe fitted linear regression lines





3.1.1 Correlation Matrix

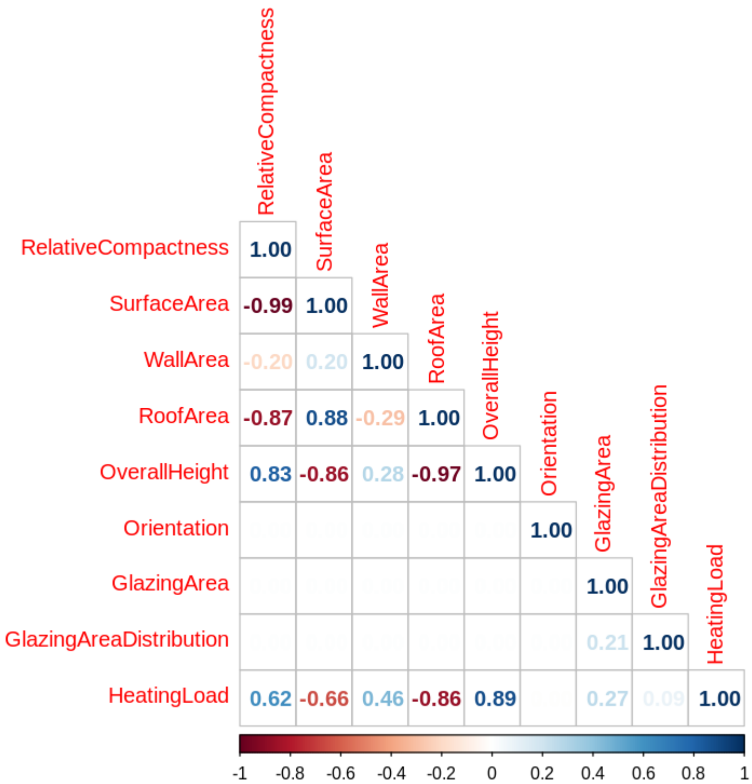


Figure 4: Correlation Matrix

- Relative Compactness has a strong negative correlation with Heating Load (r= -0.62), meaning compact buildings lose less heat and require less energy to heat.
- Glazing Area, Overall Height, and Roof Area are strongly positively correlated with Heating Load (r = 0.46, 0.89, and 0.86 respectively), so larger window areas, higher roofs, and taller buildings increase heating needs.
- Wall Area and Heating Load show a moderate positive correlation (r=0.66).
- Orientation and Glazing Area Distribution show weak or minimal correlation with Heating Load, implying little to no impact.

3.1.2 Multicollinearity

Table 1: VIF value for the variables

Var	RelativeCompactness	SurfaceArea	WallArea	OverallHeight	Orientation	GlazingArea	GlazingAreaDistribution
VIF	105.524054	201.53113	7.49298	31.20547	0.99999	1.047508	1.047508

We remove the variables which has the VIF value greater than 10. First we remove the variable SurfaceArea(201.53113) then again check the multicollinearity between predictors.

Table 2: multicollinearity

RelativeCompactness	WallArea	OverallHeight	Orientation	GlazingArea	GlazingAreaDistribution
9.250283	3.161933	9.626102	1.00	1.047508	1.0475084

Here all variables VIF values are less than 10. So we can use them to predict.

3.2 Model Selection

Model 1 (Full Model, Start)
HeatingLoad$\sim$RelativeCompactness+WallArea+RoofArea+OverallHeight+Orientation+GlazingArea+Glazing
AIC = 1698.57

Model 2 (After Removing Orientation)
HeatingLoad$\sim$RelativeCompactness+WallArea+RoofArea+OverallHeight+GlazingArea+GlazingAreaDistribu
AIC = 1691.99

Model 3 (After Removing Orientation and WallArea, Final Selected Model)
HeatingLoad$\sim$RelativeCompactness+RoofArea+OverallHeight+GlazingArea+GlazingAreaDistribution
AIC = 1689.68
Since Model 3 have the smallest AIC value , We consider it as the Best Model.

3.3 Posterior and Credible Interval summary

Figure 5: Marginal Posterior Summaries of Coefficients

```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  1 models
                        post mean  post SD   post p(B != 0)
Intercept                22.30720   0.10588    1.00000
RelativeCompactness     -64.77399  10.28944    1.00000
WallArea                 -0.02648   0.01277    1.00000
RoofArea                 -0.17458   0.03415    1.00000
OverallHeight             4.16994   0.33799    1.00000
Orientation              -0.02333   0.09470    1.00000
GlazingArea              19.93268   0.81399    1.00000
GlazingAreaDistribution   0.20377   0.06992    1.00000
```

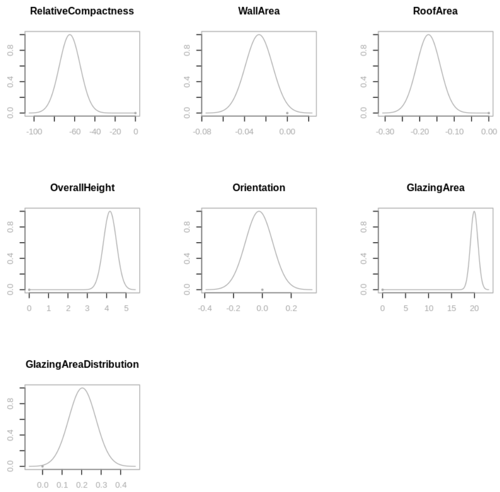
Figure 6: Credible interval summary

```
                               2.5%         97.5%         beta
RelativeCompactness     -84.97310070 -44.574882281 -64.77399149
WallArea                 -0.05155229  -0.001401579  -0.02647693
RoofArea                 -0.24162198  -0.107539115  -0.17458055
OverallHeight             3.50643397   4.833443661   4.16993881
Orientation              -0.20924198   0.162585727  -0.02332813
GlazingArea              18.33475226  21.530608099  19.93268018
GlazingAreaDistribution   0.06651684   0.341026700   0.20377177
attr(,"Probability")
[1] 0.95
attr(,"class")
[1] "confint.bas"
```

The results show Relative Compactness and RoofArea reduce heating load, while OverallHeight and GlazingArea increase it. WallArea and GlazingAreaDistribution have weaker effects, and Orientation's impact is uncertain. All variables show high inclusion probabilities, meaning they are likely important for predicting heating load.

Credible intervals confirm these patterns, with intervals for key predictors clearly different from zero, supporting their statistical importance. This means the model reliably identifies the main building features that influence heating load, giving clear guidance for energy-efficient design.

Figure 7: Coefficient summary



4 Conclusion and Recommendations

4.1 Key Findings

This Bayesian regression project showed that Relative Compactness, Roof Area, Overall Height, Glazing Area, and Glazing Area Distribution significantly affect building heating load. More compact buildings and those with larger roofs tend to use less heating energy, while taller buildings and those with more window area require more heating. Orientation was found to have little impact.

Which building characteristics strongly influence heating load?
Results identify Relative Compactness, Wall Area, Roof Area, Overall Height, and Glazing Area as having strong, significant effects supported by posterior means, credible intervals, and high marginal inclusion probabilities.

How can Bayesian regression improve prediction and understanding?
Modeling framework demonstrates the Bayesian advantage by providing uncertainty quantification through credible intervals and variable inclusion probabilities, allowing for probabilistic variable selection rather than binary judgments.

Can variable importance be assessed while accounting for uncertainty?
Yes, the marginal posterior inclusion probabilities plotted in your diagnostics quantify the confidence that each predictor belongs in the model, effectively managing variable selection under uncertainty.

4.2 Future Work

Future research should explore incorporating nonlinear and interaction terms to better capture complex relationships in building design. Utilizing expert knowledge for informative priors and extending analysis to larger, more varied datasets would improve robustness.

References

Al-Essa, L. A., Ebrahim, E. A. & Mergiaw, Y. A. (2024). Bayesian regression modeling and inference of energy efficiency data: the effect of collinearity and sensitivity analysis. *Frontiers in Energy Research*, *12*, 1416126.

Na, W. & Wang, M. (2022). A bayesian approach with urban-scale energy model to calibrate building energy consumption for space heating: A case study of application in beijing. *Energy*, *247*, 123341.

(Al-Essa, Ebrahim & Mergiaw, 2024)
(Na & Wang, 2022)

Appendix

Dataset (Click it energy.csv)
Here Heating Load and Cooling Load are the dependent variables. So I removed the Cooling Load column and performed the analysis.

R Code

```
install.packages("tidyverse")
install.packages("corrplot")
install.packages("BAS")
library(tidyverse)
library(corrplot)
library(BAS)
```

```
#Load the dataset
data <- read.csv("/content/energy.csv")
```

```
head(data)
```

```
#Structure of the variables
str(data)
```

```
#Remove the categorical variables and take all the numerical variables
energy<- data[, sapply(data, is.numeric)]
head(energy)
```

```
# Check for missing values
colSums(is.na(energy))
```

```
summary(energy)
```

```
names(energy) <- c("RelativeCompactness", "SurfaceArea", "WallArea",
    "RoofArea", "OverallHeight", "Orientation", "GlazingArea",
    "GlazingAreaDistribution", "HeatingLoad")
```

```
corrplot(cor(energy), method="number", type="lower")
```

```
install.packages("car")
```

```
library(car)
```

```
#check for multicollinearity
```

```
model_lm <- lm(HeatingLoad ~ RelativeCompactness + SurfaceArea +
    WallArea + OverallHeight + Orientation + GlazingArea +
    GlazingAreaDistribution, data = energy)
vif(model_lm)
energy <- subset(energy, select = -SurfaceArea)
colnames(energy)
model_lm <- lm(HeatingLoad ~ RelativeCompactness + WallArea +
    OverallHeight + Orientation + GlazingArea +
    GlazingAreaDistribution, data = energy)
vif(model_lm)
```

```
cor(energy)
```

```
library (ggplot2)
```

```
scPlot1 <- ggplot(data = energy , mapping = aes(x =
    RelativeCompactness , y = HeatingLoad)) + geom_point(color="blue")
    + xlab("Relative Compactness") + ylab("Heating Load") +
    geom_smooth(method=lm, color="red")
scPlot1
```

```
scPlot2 <- ggplot(data = energy , mapping = aes(x =WallArea  , y
    =HeatingLoad)) + geom_point(color="blue") + xlab("Wall Area") +
    ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot2
```

```
scPlot3 <- ggplot(data = energy , mapping = aes(x =RoofArea   , y
    =HeatingLoad)) + geom_point(color="blue") + xlab(" Roof Area ") +
    ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot3
```

```
scPlot4 <- ggplot(data = energy , mapping = aes(x =OverallHeight   , y
    =HeatingLoad)) + geom_point(color="blue") + xlab(" OverallHeight
    ") + ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot4
```

```
scPlot5 <- ggplot(data = energy , mapping = aes(x =Orientation   , y
    =HeatingLoad)) + geom_point(color="blue") + xlab(" Orientation") +
    ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot5
```

```
scPlot6 <- ggplot(data = energy , mapping = aes(x =GlazingArea   , y
    =HeatingLoad)) + geom_point(color="blue") + xlab(" GlazingArea") +
    ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot6
```

```
scPlot7 <- ggplot(data = energy , mapping = aes(x
    =GlazingAreaDistribution   , y =HeatingLoad)) +
    geom_point(color="blue") + xlab(" GlazingAreaDistribution") +
    ylab("Heating Load") + geom_smooth(method=lm, color="red")
scPlot7
```

```
energy %>%
  ggplot(aes(x= RelativeCompactness)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of Relative Compactness")
```

```
energy %>%
  ggplot(aes(x= WallArea)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of Wall Area")
```

```
energy %>%
  ggplot(aes(x= RoofArea)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of Roof Area")
```

```
energy %>%
  ggplot(aes(x= OverallHeight)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of Overall Height")
```

```
energy %>%
  ggplot(aes(x= Orientation)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of Orientation")
```

```
energy %>%
  ggplot(aes(x=GlazingArea )) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of GlazingArea")
```

```
energy %>%
  ggplot(aes(x= GlazingAreaDistribution)) +
  geom_density(fill="green", color = "red", alpha=0.8)+
  ggtitle("Data distribution of GlazingAreaDistribution")
```

```
#Perform BIC elimination from the ful model
df.lm = lm(HeatingLoad~., data=energy)
n = nrow(energy)
df.step = step(df.lm, k = log(n))
```

```
#Perform BIC elimination from the ful model
df.step = step(df.lm, k = log(n))
```

```
# Full model using all predictors
df.lm_full = lm(HeatingLoad ~ . , data = energy)
summary(df.lm_full)
```

```
#Perform BIC elimination from the ful model
df.step2 = step(df.lm_full, k = log(n))
```

```
library(BAS)
```

```
cog.bas = bas.lm(HeatingLoad~ . , data = energy,
                 prior = "BIC",
                 modelprior = uniform(),
                 include.always = ~ .)
cog.bas
```

```
summary(cog.bas)
```

```
# Find the best model based on maximum log marginal likelihood
best = which.max(cog.bas$logmarg)

# Get the variables included in the best model
bestmodel = cog.bas$which[[best]]
bestmodel

# Display the variables in the best model
bestgamma = rep(0, cog.bas$n.vars)
bestgamma[bestmodel + 1] = 1
bestgamma
```

```
cog.coef = coef(cog.bas)
cog.coef
```

```
#Coefficient summary
par(mfrow = c(3, 3), col.lab = "darkgrey", col.axis = "darkgrey", col
    = "darkgrey")
plot(cog.coef, subset = 2:8, ask = F)
```

```
#Credible interval summary
confint(cog.coef, parm = 2:8)
```

```
out = confint(cog.coef)[, 1:2]
# Extract the upper and lower bounds of the credible intervals
names = c("posterior mean", "posterior std", colnames(out))
out = cbind(cog.coef$postmean, cog.coef$postsd, out)
colnames(out) = names
round(out, 2)
```

```
# Fit the best BIC model by imposing which variables to be used using
    the indicators
cog.bestBIC = bas.lm(HeatingLoad ~ ., data = energy, prior = "BIC",
    n.models = 1, # We only fit 1 model
bestmodel = bestgamma, # We use bestgamma to indicate variables
modelprior = uniform())
cog.bestBIC
```

```
# Retrieve coefficients information
df.coef = coef(cog.bestBIC)
```

```r

# Retrieve bounds of credible intervals
out = confint(df.coef)[, 1:2]

# Combine results and construct summary table
coef.BIC = cbind(df.coef$postmean, df.coef$postsd, out)
names = c("post mean", "post sd", colnames(out))
colnames(coef.BIC) = names
coef.BIC
```

```r
# Get the names of the variables in the best model (excluding
    intercept)
best_model_vars <- names(energy)[bestgamma == 1][-1] # Exclude the
    intercept which is always included

# Construct the formula for the best model
best_model_formula <- as.formula(paste("HeatingLoad ~",
    paste(best_model_vars, collapse = " + ")))

# Fit the best model using bas.lm
model1 = bas.lm(best_model_formula, data = energy, prior = "BIC",
    modelprior = uniform())
model1
```

```r
#best fitted model
plot(model1, which = 4, ask = F, caption = "", sub.caption = "",
    col.in = "blue", col.ex = "darkgrey", lwd = 3)
```