
SM-4290 Research Project

Journal Article on Bayesian Variable Selection Linear Models

16B2070

MUHAMMAD NUR WAIZUDDIN BIN MOHD NOH

SUPERVISOR: DR MD HAZIQ MD JAMIL

18/4/2020

# Bayesian Variable Selection Linear Model

Muhammad Nur Waizuddin Bin Mohd Noh, Faculty of Science, University of Brunei Darussalam

**Abstract**

It is complicated to build a possible good fit model when a data contains a large number of variables that may cause some methods are not feasible. This arise a problem on how to reduce and select the most important variable that can enhance the quality of final model. This paper concentrate on a statistical method Bayesian Variable Selection (BVS) where Bayesian has different types of method that has been improvised from time to time. However, in this paper the one particular Bayesian method to be describe is the Kuo and Mallick (1998) method. The aim for this paper is to simply understand on Bayesian Variable Selection and understanding the difference between a method that is relying on distributions such as prior distribution and a method that is based on likelihood such as Akaike information criterion (AIC) by attempting to real world data to select variable using both method. Moreover, using Bayesian as a model selection has a major difference in interpreting the results comparing to a method that depends on AIC such as Stepwise Regression.

Keywords: Bayesian Variable Selection, Stepwise Regression, linear models.

## *1. Introduction*

Selecting variables for a good fit model has variety of methods and techniques. In this paper, Bayesian Variable Selection (BVS) method will be used throughout this paper. However, to focus on what BVS is reliable in selecting variable cases, one of the most popular classical method Stepwise Regression will also be used for comparison of the outcome. Bayesian Variable Selection itself has various types of procedure and techniques but in this paper, the method by Kuo and Mallick (1998) that inspired from George and McCulloch (1993) is the one we interested. Moreover, to analyses the difference between BVS and other method, we pick Stepwise Regression for comparison purposes.

There are three types of Stepwise Regression process, in this case we are only using the 'both directions' rather than forward selection and backward elimination. What both direction method does is sequentially adding and removing all of the variables in the data until it reaches the lowest possible AIC for a good fit linear model.

On the other hand, Bayesian selecting variable base on the highest probability of posterior model probability (PMP) in the form of zero or one as well as the Posterior

Inclusion Probability (PIP). This method has been calculated by a random sampling MCMC (Markov Chain Monte Carlo) algorithm. After the simulation is complete, the PMP will listed the top 5 model with the highest probability of a good fit linear model. By analyzing the Posterior Inclusion Probabilities (PIP) now we will only fit variable that the number is more than 0.5 or close to one in a model.

There are various types of Bayesian that has been introduced and improved. For instance, George and McCulloch (1993) developed stochastic search variable selection (SSVS) method and quoted that "SSVS is based on embedding the entire regression setup in a hierarchical Bayes normal mixture model where latent variables are used to identify subset choices" (George & McCulloch, 1993, p. 881). The SSVS method also mentioned in O'Hara and Sillanpää (2009) that to get convergence of random variables, it requires tuning the models such as adjusting the prior distribution. Then Kuo and Mallick (1993) come up with a method inspired by complexity of SSVS method and introduced alternative process. One of the aims that the method was created is to avoid the complicated tuning factor for hierarchical setup. Furthermore, Ntzoufras (2011) agreed that The Kuo and Mallick (1998) method is the simplest that only need to specify the usual prior ($\beta$, $\gamma$ and $\sigma$).

In this paper, two datasets will be demonstrated in order to apply the classical method Stepwise Regression and Bayesian Variable Selection. Throughout this paper, both methods will be calculated using R programme statistical computing. Next, the result will be compiled, calculate the AIC of the complete model using both BVS and Stepwise Regression to determine which method has the advantage on predicting the best for linear models. Finally, analyzing and interpreting the selected variables then each of the variables selected for the final product will be clarified to determine which method is best at choosing subset variable reasonably.

## 2. Materials and Methods

There are $2^p$ number of best possible model to build, where p is number of variables recorded in a dataset. For example, there are 10 independent variables (X1, X2, ..., X10) to be selected from a data and $2^{10}$ is 1024. Therefore, there are 1024 possible models available to pick the most preferred one. The aim of using Bayesian Variable Selection approach is to pick the highest likeliest variable and models. In this case the Stepwise Regression is commonly decided by the likelihood such as the lowest AIC (Akaike information criterion) will be picked and consider as preferred. Meanwhile, BVS will pick top 5 models determined by the PMP.

As Bayesian Variable Selection approaches in Linear Regression, we consider that it is the Multiple Linear Regression and the equation can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ ... + \beta_p x_p + \varepsilon \qquad (2.1)$$

$$y = \sum_{j=1}^{k} \beta_j x_j + \ \varepsilon \qquad (2.2)$$

y – Response Variable

$x_p$ – Explanatory variable

$\varepsilon$ - Random error term

This is the standard formula for Linear Regression. However, Bayesian Variable Selection from Kuo and Mallick (1998) improved the equation as

$$yi = \sum_{j=1}^{p} \beta_j \gamma_j \, x_{ij} + \varepsilon_i \qquad (2.3)$$

This formula is similar to (2.2) with an addition of $\gamma$ variable to be multiplied by each $\beta$ and $x$ introduced by (Kuo and Mallick,1998). Moreover, this equation assumes the error ( $\varepsilon_i$) set to a normal distribution with mean 0 $(0, \sigma^2)$. The $\gamma$ variable act as an indicator for the potential variable $x_{ij}$ either selected or not. When $\gamma_j = 1$ , the covariates have to be included when setting up the model and if $\gamma_j = 0$, we neglect the variable.

***Prior Distributions***

Refer to the Bayesian model (2.3), the y$i$ and $x_{ij}$ here can be identify using information for a chosen dataset. With the parameters $\{\beta_0, \dots, \beta_p, \gamma_{1,\dots}\gamma_p, \sigma^2\}$ are unknown exist in the model, the setting for Bayesian have to assume the prior distributions on the parameters consider the parameters are random.

The prior distributions are,

$$\beta_{j \sim N(b,B)} \qquad (2.4)$$

$$r_j = \begin{bmatrix} 1 & w \cdot p & \pi_j \\ 0 & w \cdot P & 1 - \pi_j \end{bmatrix} \qquad (2.5)$$

We have Normal Distribution for $\beta_j$, as for the indicator $r_j$ we have Bernoulli Distribution with probability of $\pi_j$ for the variable existence and $1 - \pi_j$ for variable that are failed to be included.
O'Hara and Sillanpää (2009) and Ntzoufras (2011) consider the model parameters with Normal Distribution and Bernoulli Distribution are independent prior. This statement can be expressed by

$$f(\beta, \gamma) = f(\beta) \cdot f(\gamma) \qquad (2.6)$$

The left-hand side of the expression is the joint probability density function (p.d.f) whereas the right-hand side is the product of individual whatever inside the joint p.d.f. The multiplication of both individuals is independent each other.

O'Hara and Sillanpää (2009) also mentioned the Kuo and Mallick (1998) equation (2.3) with an adjustment of introducing new parameter $\theta_j$, where $\theta_j = \beta_j \gamma_j$ . This prior is called "slab and spike" prior.

$$y_i = \alpha + \sum_{j=1}^{p} \theta_j x_{i,j} + e_i \qquad (2.7)$$

$\alpha$ is the intercept of $y_i$ and $e_i$ are the errors $(N, \sigma^2)$. There are two important auxiliary variables in this equation. As $\theta_j = \beta_j \gamma_j$ , the $\gamma_j$ still act as binary zero and one to indicate the presence and absence of input variable similar to (2.5). The second auxiliary variable is when $\gamma_j$ = 0, the new parameter $\theta_j$ is 0 and $\theta_j = \beta_j$ if $\gamma_j$ = 1.

***Posterior distributions***

Another significant aspect that we need to concern when using BVS method for a model selection is the Posterior Inclusion Probabilities (PIP) and Posterior Model Probabilities (PMP) that is calculated by MCMC algorithm. By assuming that the parameters in Bayesian have prior distribution, Bayesian also focus on posterior distribution of parameters. The purpose of posterior distribution is to calculate the probability after the evidence of variable $x_{ij}$ is taken into account when indicator variables is 1 ($\gamma_j$ = 1) in the prior distribution.

Before the data observation, (Kuo and Mallick,1998) stated that the $\pi_j$ in prior distribution (2.5) is equal to 0.5 for the probability of variable $x_{ij}$ to be included yet the number is not fixed and can be raise depending on how confident the researcher believes towards the variable to be included. The conditional probability posterior distribution of $f(\beta|y)$ is directly proportional to the prior $f(\beta, y)$ when $\gamma_j$ = 1 / $f(\ y)$ when $\gamma_j$ = 0. The posterior distribution concern on $f(\beta|y)$, where the probability of variable included after observation has been done. We denote this by $\hat{\pi}_j$.

***Posterior Inclusion Probabilities (PIP) and Posterior model probabilities (PMP)***

When using BVS method on R software, the result from the random sampling by MCMC will display two significant parts to indicate researcher which variables are the best to construct a linear model. The Posterior Inclusion Probabilities (PIP) and the Posterior Model Probabilities (PMP). The PIP is the result of $\hat{\pi}_j$ of $x_{ij}$ variable where number of the $x_{ij}$ selected divided by the total number of MCMC simulation. Meaning that the probability for $\hat{\pi}_j$ is measured within the range from 0 to 1 that evaluates how likely variable $x_{ij}$ is important to build the linear model. If PIP only focus on the probability on one certain variable, the Posterior Model Probability interests in the probability of each unique model assembled out of $2^p$ models from only 1 variable $x_1$ is included for the model to full set variables $\{x_1, x_{2,\dots}, x_P\}$ are included when building the model and will also be divided to the number of MCMC simulation.

To demonstrate on how the random sampling of MCMC scheme works in a random dataset by BVS using R, two random datasets have been chosen. First real-world data set is Mortality and air pollution data and second is the Ozone. The similarity between both of this dataset is the response variable is continuous. Stepwise Regression method will also be applied to these data examples to distinguish variable selection process of Bayesian that by observing the PIP and PMP and Stepwise Regression using AIC.

**2.1 Mortality and air pollution data**

This dataset can be found in R in 'iprior' package and first appeared in McDonald and Schwing (1973). There are 16 variables has been recorded. The mortality variable which is the total age adjusted mortality rate has been set to be the dependent variable. The other 15 variables can be grouped into 3 categories. weather, socioeconomics and pollution. The main subject to focus on this dataset is the correlation between mortality rate and pollution. The three potential pollution variables are Hydrocarbons (HC), Nitrogen Oxide (NOx) and Sulfur Dioxide (SO2). Despite pollution is the main cause of mortality in Standard Metropolitan Statistical Areas (SMSA) other factors such as weather

and socioeconomic also affect the rate of mortality. Table 1 shows that the variables recorded in this dataset as well as the description. This dataset has been approached before by McDonald and Schwing (1973) where in the paper the use Total Squared error and Ridge Trace to eliminate ineffective variable and the result from each method is almost the same. However, the aim for us to select this random dataset is to apply Stepwise Regression and BVS method.

Table 1: Description of variables recorded in Mortality and air pollution data

| Variable | Description |
|---|---|
| Mortality | Total age adjusted mortality rate |
| Precipitation | Mean annual precipitation (in) |
| Relative humidity | Percent relative humidity, annual average at 1 p.m. |
| January temperature | Mean January temperature (F) |
| July temperature | Mean July temperature (F) |
| Population density | Population per square mile in urbanised area |
| Household size | Population per household |
| Education | Median school years completed for those over 25 |
| Sound housing units | Sound housing units (no defects) (%) |
| Age >65 years | Population that is 65 years of age or over (%) |
| Non-white | Urbanized area population that is non-white (%) |
| White collar | Employment in white-collar urbanized occupations (%) |
| Income <$3,000 | Families with income under $3,000 (%) |
| HC | Relative population potential of hydrocarbons |
| $NO_x$ | Relative population potential of oxides of nitrogen |
| $SO_2$ | Relative population potential of Sulphur dioxide |

### 2.2 Ozone dataset

This dataset can be found from the package mlbench in R software. According to Casella and Moreno (2006), this data was firstly observed by (Breiman and Freidman, 1985). There is a missing data from the original dataset and this can be solved by using complete cases in R programme where it detects and eliminate the incomplete information from the data. Table 2 shows the variables that contains the dataset This data consists of 11 variable which is 1 response variable and 10 explanatory variables. For this Ozone case, the response variable will be the daily maximum one-hour-average ozone reading at Upland, CA. The explanatory variables have been recorded on 5 location in California, USA with different information. For instance, temperature in Sandberg, CA as variable for $X_7$, Wind speed at LAX Los Angeles international airport for $X_5$ and more information are available in table 2

Table 2: Description of variables recorded in Ozone dataset.

| Variable | Description |
|---|---|
| $y$ | Daily maximum one-hour-average ozone reading (ppm) at Upland, CA |
| $X_1$ | Month: 1 = January, ..., 12 = December |
| $X_2$ | Day of month: 1*; 2;: : :* |
| $X_3$ | Day of week: 1 = Monday*; : : ;* 7 = Sunday |
| $X_4$ | 500-millibar pressure height (m) measured at Vandenberg Air Force Base |
| $X_5$ | Wind speed (mph) at Los Angeles International Airport (LAX) |
| $X_6$ | Humidity (%) at LAX |
| $X_7$ | Temperature (F) measured at Sandberg, CA |
| $X_8$ | Inversion base height (feet) at LAX |
| $X_9$ | Pressure gradient (mmHg) from LAX to Daggett, CA |
| $X_{10}$ | Visibility (mi) measured at LAX |
| $X_{11}$ | Temperature (F) measured at El Monte, CA |
| $X_{12}$ | Inversion base temperature (degrees Fahrenheit) at LAX |

For this dataset, Breiman and Freidman (1985) uses ACE algorithm. However, Casella and Moreno (2006) constructing Bayesian criterion for model selection using intrinsic prior and the most selected variables from both papers are X7-X10.

## Results and Discussion

After selecting variable using Stepwise regression and Bayesian Variable Selection for linear models in R software, we will compare the results between the two methods. Table 3 shows all of the variable selected from both methods in Air Pollution and Mortality dataset and same procedure in Ozone dataset illustrate in Table 4.

### *Results on Air Pollution and Mortality*

Table 3: Results showing which subset variable is selected in Air Pollution and Mortality dataset

| | Stepwise Regression | BVS |
|---|---|---|
| Rain | ● | ● |
| JanTemp | ● | ● |
| JulTemp | ● | |
| Over65 | ● | |
| Popn | ● | |
| Educ | ● | |
| NonW | ● | ● |
| HC | ● | |
| SO2 | | ● |
| NOx | ● | |
| **AIC** | 601.9077 | 603.2024 |

Table 4: Summary results of Stepwise Regression on Air Pollution and Mortality dataset

| | Estimate | Standard Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1934.0683 | 333.4988 | 5.799 | 4.48e-07 *** |
| Rain | 1.8564 | 0.8373 | 2.217 | 0.03119 * |
| JanTemp | -2.2620 | 0.6957 | -3.252 | 0.00206 ** |
| JulTemp | -3.3199 | 1.3971 | -2.376 | 0.02136 * |
| Over65 | -10.9202 | 7.1398 | -1.529 | 0.13245 |
| Popn | -137.3881 | 59.7751 | -2.298 | 0.02576 * |
| Educ | -23.4217 | 7.0620 | -3.317 | 0.00170 ** |
| NonW | 4.6623 | 0.9689 | 4.812 | 1.42e-05 *** |
| HC | -0.9222 | 0.3192 | -2.889 | 0.00571 ** |
| NOx | 1.8711 | 0.6095 | 3.070 | 0.00346 ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

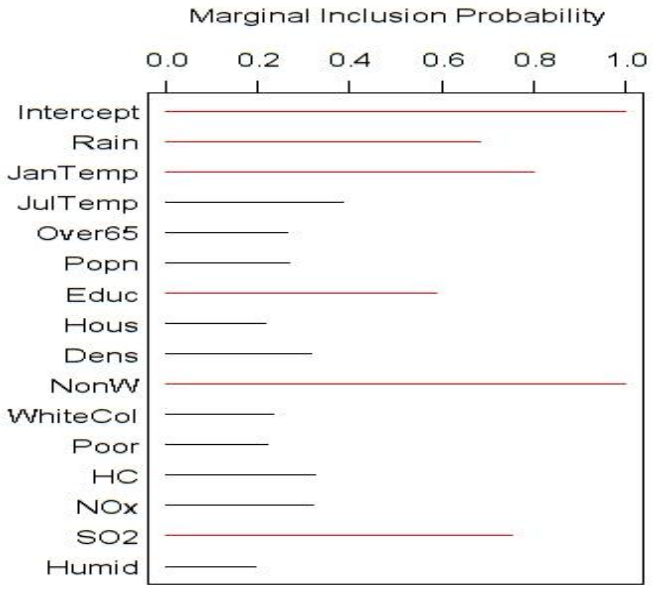


Fig 1: plots on marginal Posterior Inclusion Probabilities (PIP) in Air pollution and Mortality dataset.

In Fig 1, the graph is plotted after observation to all potential variables in Air Pollution and Mortality data using MCMC algorithm. This plot indicates the covariate shown in red has greater than 0.5 of marginal PIPs. The variables with high inclusion probability is more likely to be included in a model consider as important for explaining and predicting. In this case, BVS method chose 5 variables to be included in the model that shown in red line in figure 1.

***Results on Ozone***

Table 5: Results showing which subset variable is selected in Ozone dataset

| | Stepwise Regression | BVS |
|---|---|---|
| Month | ● | ● |
| Presvand | ● | |
| HumLAX | ● | ● |
| TempSand | ● | |
| TempElMon | ● | ● |
| ibhLAX | ● | |
| **AIC** | 1181.49 | 1185.346 |

Table 6: Summary results of Stepwise Regression on Ozone dataset

| | Estimate | Standard Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 52.6428370 | 35.0855204 | 1.500 | 0.135116 | |
| Month | -0.3334315 | 0.0956622 | -3.486 | 0.000606 | *** |
| PresVand | -0.0136194 | 0.0065952 | -2.065 | 0.040234 | * |
| HumLAX | 0.0979485 | 0.0173681 | 5.640 | 5.89e-08 | *** |
| TempSand | 0.1237666 | 0.0576373 | 2.147 | 0.032995 | * |
| TempElMon | 0.4745550 | 0.0908036 | 5.226 | 4.41e-07 | *** |
| ibhLAX | -0.0003360 | 0.0002157 | -1.558 | 0.120847 | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

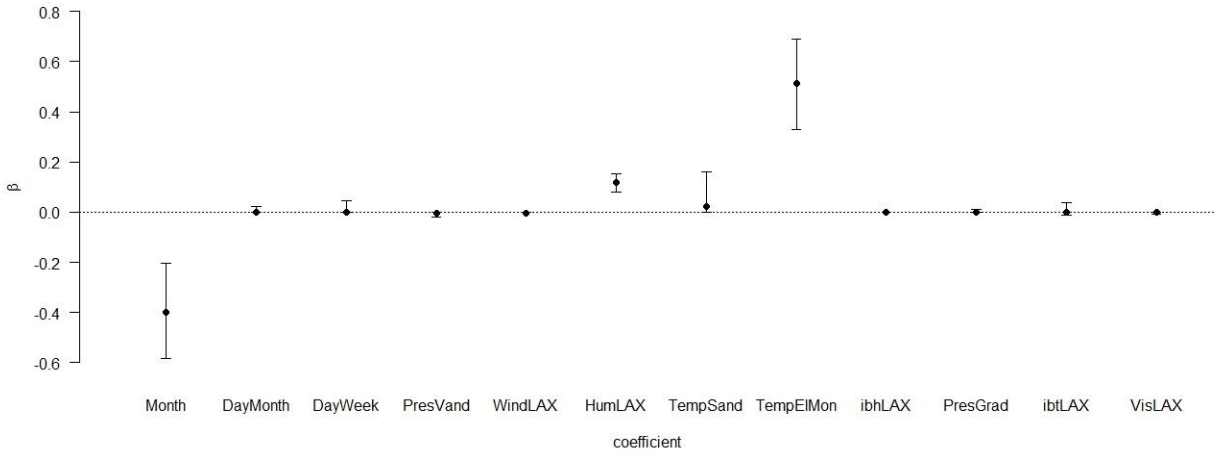


Fig 2: Plots of the posterior distributions of the coefficients under model averaging in Ozone datasets.

In Fig 2, The solid line at 0.0 is the posterior probability that the coefficient is zero. The height of each coefficient represents the probability that it is nonzero coefficient. In this plot, there are 3 obvious coefficient (Month, HumLAX, and TempElMon) that is distant from the 0.0 line.

From the final results of both Air Pollution and Mortality and Ozone datasets, we realize that the AIC for Stepwise Regression is always lower than BVS method. In what situation does BVS is good at? After interpretation from the results, we will notice that the BVS model has chosen the variable that is reliable as well as understandable to the situation and explain efficiently. On the other hand, Stepwise Regression has good result from $R^2$ and AIC but is only good for predicting rather than explaining.

***Analysis of Mortality and Ozone dataset***

The number of variables selected between the two methods is distinguishable. In Table 3 Stepwise Regression clearly has more explanatory variable selected compare to Bayesian Variable Selection approaches. Next, the classical method has lower AIC than BVS. Clearly it shows that the method to predict variable selection in favor of Stepwise Regression.

However, if we look at the summary of Stepwise Regression calculated by R programme as shown in table 5 and 6, if we look at the Over 65 dependent variable from Air Pollution and mortality dataset as well as ibhLAX variable from the Ozone dataset, both has the similarity which is both variables are selected but not significant for the p value. This will create uncertainty whether or not these variables to be included.

As we know that the Air Pollution and Mortality dataset focuses more on the factor of potential pollution that affect the number of the Response variable. There are three potential pollution that stated in table 1, Hydrocarbons (HC), Nitrogen Oxide (NOx) and Sulphur Dioxide (SO2). Stepwise Regression method has selected HC and NOx as the pollution that has a potential to affect the independent variable. By observing the

estimate section from table 5, The HC variable shows an -0.9222 basically means that the lower number of Hydrocarbons in a Standard Metropolitan Statistical Areas (SMSA). This statement is inexplicable. Moreover, besides the potential pollution section this situation also happens in weather section. If we look at the January Temperature (JanTemp) variable and July Temperature variable (JulTemp) in Table 3, BVS method selected JanTemp only and for Stepwise Regression does select both. By observing the summary of Stepwise Regression method in table 5 where both variables are showing negative numbers in the estimate section. For JulTemp variable, the information cannot be accepted because July in SMSA, USA is summer season where the temperature should be not low.

Now we analyses the Ozone dataset. This circumstance will be slightly different from Air Pollution and Mortality dataset yet undergo the same procedure. The comparison between Stepwise Regression and Bayesian Variable Selection regarding the AIC, indeed that the Stepwise Regression conquered with 1181.49. Since there are only total of 6 variables selected out of 12 variables in Stepwise Regression, BVS only pick 3 variables that are more acceptable. In Table 2, there are two locations that temperature recorded for the data. The first one is the Temperature in Sandberg, California ($X_7$) and the second one is in El Monte, California ($X_{11}$). By referring to table 4, after a process of sequentially adding and removing all the 12 variables, Stepwise Regression has chosen both locations to be included in a model.

On the other hand, using MCMC algorithm and observing the PMP and PIP with Bayesian Variable Selection only include one congressional district El Monte, California over Sandberg, California. To figure out the explanation, we calculate the correlation between the two explanatory variables $X_7$ and $X_{11}$ using R software. The $X_7$ and $X_{11}$ indicate high correlation with more than 0.5. Thus, BVS only include one point of interest El Monte that is the closest range to Upland, California where the daily ozone reading measured or the response variable.

From the results above, we can conclude that every method has their own ability from its own perspective. Some method approaches are good for predicting like the classical method Stepwise Regression. In spite of good at predicting, Stepwise Regression or other method that the result is based on likelihood is lack of explaining which the Bayesian Variable Selection are capable. This will raise an issue choosing a good fit model based on good prediction or good explanation.

***Predicting vs Explaining***

Predicting and explaining are two different nature. Shmueli (2010) believes that these two terms exist in statistics perception when inaccurate construct of measurable data and create a huge difference between generating predicting model at measurable level and producing a model that has the ability to explain in any fields. Moreover, this controversy has been argued between modern linguistics perspective and research perspective shared by (Wong, 2019),

Noam Chomsky made a point based on linguistic perspective disagrees the complexity of statistical analysis by analyzing each variable depend on theories and resolving the approximation of data with incomplete information because it does not determine real knowledge of a language. Chomsky also admit that predictive accuracy

defined the success in natural language processing (NLP) and not science. This statement implies on the culture of algorithmic in statistical modelling that is complicated to clarify numerous parameters. Nevertheless, the Director of Research, Peter Norvig disputed the science does not have a big effect on NLP. Norvig believes that most excellent applications such as search engine or speech recognition used in human language processing are influence by the probabilistic models that is more advance than classic logical based. Despite the fact that Chomsky is more on intellect side rather than Norvig with science machine with improved algorithm method, both agrees on having the entire variables without understanding the variable gives a better prediction.

Overall, it is on the Researcher determination which method is eligible depend on the aim for statistical analysis is to predict or explain. If the researcher is looking towards of having a model that has a better explanation, it would be suggested using a statistical method that concern on estimating a distribution and unknown parameters such as Bayesian. Even though Prediction is the common thing in today's world, it is also better to have a model that can explain on a certain phenomenon to get better accuracy.

## Conclusion

To summaries the whole point in this paper, Bayesian Variable Selection is a method that uses algorithmic modelling that is measured by using MCMC simulation. Bayesian concerns to look at entire model space by estimating the prior distribution and also posterior distribution. By analyzing and interpret the two example datasets and comparing the method between the classical method Stepwise Regression and Bayesian Variable Selection we discover that the product from Stepwise Regression is better at predicting variables without understanding all of them. However, the objective of the final model by BVS is more on explaining the variable selected.


### Acknowledgements

I would like to express my sincere and deep thanks to my advisor Dr Haziq Jamil Assistant Professor in Statistics. Mathematical Sciences, Faculty of Science (FOS), University of Brunei Darussalam (UBD) for his introduction of statistics and helping me to explore on a specific topic in statistics the Bayesian Variable Selection. I would also like to thank for his patience through restriction of face to face meeting and able to continue meeting using online platform as well as carefully supervised me throughout my final year project. From this project I would be able to understand more the theory behind a software that simply does a calculation also guiding me as a beginner student to use R programme.

**References**

Breiman, Leo and Jerome H. Friedman (1985). “Estimating Optimal Transformations for Multiple Regression and Correlation”. In: *Journal of the American Statistical Association* 80.391, pp. 590–598. doi: 10.1080/01621459.1985.10478157.

Casella, George and Elías Moreno (2006). “Objective Bayesian Variable Selection”. In: *Journal of the American Statistical Association* 101.473, pp. 157–167. doi: 10.1198/016214505000000646.

George, Edward I. and Robert E. McCulloch (1993). “Variable Selection Via Gibbs Sampling”. In: *Journal of the American Statistical Association* 88.423, pp. 881–889. doi: 10.2307/2290777.

Kuo, Lynn and Bani Mallick (1998). “Variable selection for regression models”. In: *Sankhyā: The Indian Journal of Statistics, Series B* 60.1, pp. 65–81.

McDonald, Gary C. and Richard C. Schwing (1973). “Instabilities of Regression Estimates Relating Air Pollution to Mortality”. In: *Technometrics* 15.3, pp. 463–481. doi: 10.2307/1266852

Ntzoufras, Ioannis (2011). *Bayesian Modeling Using WinBUGS*. Wiley. isbn: 978-0-470-14114-4. doi: 10.1002/9780470434567.

O'Hara, Robert B. and Mikko J. Sillanpää (2009). "A Review of Bayesian Variable Selection Methods: What, How and Which". In: *Bayesian Analysis* 4.1, pp. 85–117. doi: 10.1214/09-BA403.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310

Wong, P. (2019, July 25). *Predicting vs. Explaining.* Retrieved from https://towardsdatascience.com/predicting-vs-explaining-69b516f90796