A Bayesian approach on multicollinearity problem with an Informative Prior

I G N M Jaya¹, B Tantular², Y Andriyana³
¹,²,³Department Statistics, Universitas Padjadjaran, Indonesia

Corresponding author: mindra@unpad.ac.id

Abstract. Multicollinearity is a severe problem in multiple regression. High collinearity in some explanatory variables leads to the high standard error estimates. It becomes a problem for the hypothesis test on the slope of regression. The ridge regression is the most popular method used to minimize the standard error estimates. However, the hypothesis testing in regression model has been not solved yet. It does not provide the statistical hypothesis test. Therefore, the alternative method is needed. The method must be able to obtain the parameter estimates with a high level of precision and also facilitates the hypothesis test of regression parameters simultaneously. We proposed the Bayesian method with an informative prior as an alternative solution. The Monte Carlo simulation concludes the Bayesian method outperformed to ridge regression in term of Bias, Mean Square Error and power of the test. Based on the simulation result, the Bayesian method can be used to solve hypothesis testing in regression analysis with multicollinearity problem effectively.

1. Introduction
Multicollinearity is a term in the economic literature to present the perfect and near linear relationship between the independent variables in linear regression analysis [1]. Although this condition is not a part of the classical assumption of linear regression by, however, the occurrence of collinearity problem may lead to serious problem in regression inference such as imprecise of the regression estimate ([1]; [2]; [3]; [4]). The estimates have sizeable standard error and very high determination coefficient. In order of goodness of fit evaluation, the model with the multicollinearity problem seemed to meet the goodness of fit criteria; however, the majority of the independent variables are not statistically significant.

Several methods have been introduced to overcome the impact of occurrence of the collinearity, i.e. principal component regression, partial least square, and ridge regression. The last is the most popular in many fields [5]. Ridge regression procedure was introduced by Hoerl and Kennard (1975) [6]. The idea is added the positive value ($k$) to the diagonal element of $(XX)$. This idea decreased the standard error of parameter estimate for the high value of ($k$) significantly. However, there is a cost that has to be paid. Increasing ($k$) lead to a large bias estimate. The iteration procedure was developed to select the optimum $k$ and provide the ridge trace depends on constant values between zero and one versus estimated curves [6]. Shih and Shih (1979) [7] applied ridge regression for water resource studies. Hadgu (1984) [8] used it for modelling disease data. In economic studies, the ridge regression procedure also much more used ([9]; [10]; [11]).
However, the ridge regression procedure has not been able to resolve the inference problem fully. The ridge regression produces a reliable estimate without significant test statistics while the main problem in collinearity is producing invalid statistical tests.

Here we developed Bayesian approximation using informative prior. A Bayesian method is well known and has been applied in many fields. Using the Bayesian method, we can improve the quality of parameter estimate by mean including the outside information via prior distribution [12] and used central limit theorem, the Bayesian method facilitates the hypothesis test of regression parameters.

We proposed Monte Carlo simulation study to proof that the Bayesian method outperformed to Ridge regression approach to solve the multicollinearity problem.

The remainder of the paper is organized as follows. Section 2 method, explain the Bayesian regression and simulation study. Section 3 presents result of the simulation study and the real example, while section 4 summarizes, concludes and presents suggestions for future work.

2. Methods
2.1. Bayesian Regression

Bayesian regression offers the alternative method in estimating and inferencing regression model. In a few decades, Bayesian regression becomes the most popular method in statistical analysis. The most advantage of the Bayesian method presents a reliable estimate because of the additional information given prior. Bayesian inference combines subjective and objective information. Before the data are collected, subjective information is presented [12]. However, the reliable results are obtained with complexity computation process. Bayesian approaches involve three components, likelihood function, prior and posterior distributions. Prior distributions is a probability distribution of the parameters to be estimated. While the likelihood is related to the probability distribution of observational and posterior distribution, it is a combined distribution of the required parameters of the data. The prior is determined before the observation data is held so that likelihood is often expressed as a validating function of the prior. The posterior distribution is defined using Bayesian Law. Suppose that states the observation data by stating the parameter to be estimated. Posterior distribution can be stated in Bayesian rule.

Assumes, $\mathbf{D} = \{\mathbf{y}, \mathbf{X}\}$ denotes the observation and $\theta = \{\beta, \sigma^2\}$ are parameters interest. The posterior distribution can be defined as $f(\theta \mid \mathbf{D}) = \frac{f(\mathbf{D}, \theta)}{f(\mathbf{D})}$ where, $f(\theta, \mathbf{D})$ is join density function between $\mathbf{D}$ and $\theta$, and $f(\mathbf{D})$ denotes marginal density function of observed variable $\mathbf{D}$. Using Bayes rule, $f(\mathbf{D}, \theta) = f(\mathbf{D} \mid \theta) f(\theta)$ where $f(\mathbf{D} \mid \theta)$ is the likelihood function $f(\theta)$ denote the prior distribution and the posterior distribution can be defined as:

$$f(\theta \mid \mathbf{D}) = \frac{f(\mathbf{D} \mid \theta) f(\theta)}{f(\mathbf{D})}$$

$$= \frac{f(\mathbf{D} \mid \theta) f(\theta)}{\int \Omega f(\mathbf{D} \mid \theta) f(\theta) d\theta}$$

$$= c f(\mathbf{D} \mid \theta) f(\theta)$$

$$\propto f(\mathbf{D} \mid \theta) f(\theta)$$

The first complexity in Bayesian approach defines the prior distribution. Here we use Normal-Gamma conjugate prior with an informative prior parameter.

2.1.1. Normal-Gamma Prior
Let \( y \sim N(X\beta, \sigma^2 I) \), where \( X \) design matrix with dimension \( n \times (k+1) \). There are two parameters we have to estimate, \( \beta \) and \( \sigma^2 \). In order to simplify the computation process, we define precision parameter \( \nu \) i.e. \( \nu = 1/\sigma^2 \). The likelihood function can be written as
\[
L(\beta, \sigma^2) = f(y \mid \beta, \sigma^2) \propto \nu^{n/2} \left\{ \nu (y - X\beta)'(y - X\beta)/2 \right\}
\]
with the prior distributions \( \nu \sim \Gamma(a,b) \) and \( \beta \mid \nu \sim N(\beta_0, V/\nu) \) and the quantities \((a, b, \beta_0, V)\) are the so-called supper-parameters. The Bayesian specification of Bayesian regression can be presented as below:
\[
\begin{align*}
    y \mid (\beta, \nu) &\sim N(X\beta, I/\nu) \\
    \beta \mid \nu &\sim N(\beta_0, V/\nu) \\
    \nu &\sim \Gamma(a,b)
\end{align*}
\]
Based on the specification in (3) the joint posterior distribution can be defined as [12]:
\[
f(\beta, \nu \mid y) \propto \nu^{(n+a)/2-1} \exp(-\nu \tilde{b}) \times \nu^{(k+1)/2-1} \exp\left[ -\frac{\nu}{2} \left( \beta - \tilde{\beta} \right)' \left( X'X + V^{-1} \right)^{-1} \left( \beta - \tilde{\beta} \right) \right]
\]
where
\[
\begin{align*}
    \tilde{a} &= \frac{n}{2} + a \\
    \tilde{b} &= b + \frac{y'y - \left( X'y + V^{-1} \beta_0 \right)' \left( X'X + V^{-1} \right)^{-1} \left( X'y + V^{-1} \beta_0 \right)}{2} \\
    \tilde{\beta} &= \left( X'X + V^{-1} \right)^{-1} \left( X'y + V^{-1} \beta_0 \right)
\end{align*}
\]
The first component of \( \tilde{\beta}, (X'X + V^{-1})' \) identically with ridge regression component. Using this fact, we can use it to solve the collinearity problem using Bayesian approach and directly obtain the hypothesis testing by mean asymptotic t distribution.

2.1.2. Informative Prior
The informative prior was selected here for parameter \( \beta \) and \( \sigma^2 \). For \( \beta \sim N(\beta_{OLS}, V/\nu) \) with small value for diagonal \( V \) and for \( \sigma^2 \sim IG(1,1) \). Gibbs sampler algorithm for Bayesian regression
- Define the initial value for vector parameter \( \beta_{(0)} = \tilde{\beta}_{(OLS)} \)
- Do Markov Chain Monte Carlo iteration sampling
  - Iterate \( m = 1 \) until \( M \) e.g., \( M=10000 \)
    - Substitute \( \beta = \beta_{(m-1)} \) and generate random variable \( \nu_{(m)} \sim \Gamma(a,b) \)
    - Substitute \( \nu \sim \nu_{(m)} \) and generate random variable \( \beta_{(m)} \sim N(\tilde{\beta}, V^{-1} \left( X'X + V^{-1} \right)^{-1}) \)
  - Stop the iteration processes
- Burn-in: discard the \( L \) first iteration \( \{ \beta_{(m)}, \nu_{(m)} \} \) (e.g. \( L=1000 \))
- Obtain \( \hat{\beta}_L = \sum_{m=L}^{M} \beta_{m,k} / (M - L) \)
- Obtain \( \text{se}(\hat{\beta}_L) = \sqrt{\sum_{m=L}^{M} (\beta_{m,k} - \hat{\beta}_L)^2} / (M - L) \)
- Obtain asymptotic \( t_L = \hat{\beta}_L / \text{se}(\hat{\beta}_L) \)
2.2. Simulation Study

We compare Ridge regression and Bayesian with informative prior by means Monte Carlo simulation study. Here, we simulate the multiple regression with two independent variables. The parameters intercept and slopes are fixed 1, and the correlation between two independent variables are set varies start from 0 until 1. The independent variables follow multivariate normal distributions with mean equal to zero, and the variance and the covariance matrix is defined as the correlation matrix. The error of the regression model follows the univariate normal distribution with mean and variance are zero and one respectively.

Simulation design

1. Set \( \beta_0 = \beta_1 = \beta_2 = 1 \)
2. Set \( N = \{10, 30, 50, 100, 500\} \)
3. Set \( \rho = \{0, 0.3, 0.5, 0.8, 0.9, 0.99, 0.999\} \)
4. Generating matrix of independent variables
   \[
   x_1 \sim N(0,1) \quad x_2 = \rho x_1 + N(0,1) \sqrt{(1-\rho^2)}
   \]
5. Generating vector \( \epsilon \sim N(0, 0.1I) \)
6. Define vector \( y = 1 + X_1 + X_2 + \epsilon \)
7. Run the ordinary least square (OLS), ridge regression and Bayesian regression.
8. Calculate the Bias, mean square error estimate (MSE) and power testing.
   a. Bias
   \[
   Bias = \frac{\sum_{i=1}^{m} |\hat{\theta} - 1|}{m}
   \]
   with \( \hat{\theta} = \{\hat{\beta}_1, \hat{\beta}_2\} \) and \( m \) denotes the number of iteration.
   b. MSE
   \[
   MSE = \frac{\sum_{i=1}^{m} (\hat{\theta} - 1)^2}{m}
   \]
   c. Power test
   \[
   Power = \frac{\# \text{hypothesis null are rejected}}{m}
   \]

3. Results and Discussion

3.1. Simulation Results

Monte Carlo simulation was used to evaluate the Bias, MSE and the power testing of OLS, Ridge Regression and Bayesian regression approach for multicollinearity problem. Figure 1-4 present the simulation of three estimation methods, seven different value of multicollinearity parameters, and five different numbers of sample size. For large sample size (i.e., 500), the effect of multicollinearity for parameter estimate is relatively small. However, for a small sample size with a large value of collinearity parameter, the effect of multicollinearity is very serious [12]. The bias and mean square error tend to large [3]. However, we can see that the effect of multicollinearity is relatively small for the Bayesian approach for all conditions of simulation. It indicates that the Bayesian approach is the best solution for a multicollinearity problem.
Figure 1. (a) Bias and (b) MSE of Regression Parameters

Figure 1 shows the bias and means square error estimate of regression parameter by mean Monte Carlo simulation study. It presents clearly that the Bayesian approach is the best estimation method to solve the multicollinearity problem. Figure 1 indicates that the Bayesian method has the smallest Bias and mean square error compare than OLS and ridge regression for all simulation conditions. Ridge regression has the largest Bias. It is a consequence of ridge regression to obtain a smaller standard error estimate. It is comparing with OLS, ridge regression has the smallest variance estimate. For OLS, there are some points in simulation approaches which have a large mean square error and it indicates the OLS method provides largest standard error estimate.

Figure 2. Power of hypothesis test of regression parameters

Figure 2 presents the power of the hypothesis test of regression parameters ($\beta_1, \beta_2$). Bayesian method again becomes the best method for the power of hypothesis testing. To generalized the result for all different values of number sample size and multicollinearity parameter, we present the non-linear modelling which is shown in 3 dimensions.
Figure 3. Bias and MSE of Regression Parameters
Figure 3 presents the Bias and MSE of Bayesian, OLS and Ridge regression methods. Bayesian and OLS methods present the smallest Bias and MSE which indicate by flat curve and the ridge regression perform cubic curve. The Bias and MSE are increasing for all methods when the degrees of collinearity is increasing. Figure 4 shows the power test of Bayesian, OLS, and Ridge regression methods. The Bayesian method outperformed to OLS and Ridge regression. It indicates the curve is flat and its value closed to one. For the high-level collinearity ($\rho > 0.700$), the power test of OLS and Ridge regression goes down significantly while for the Bayesian method, the power test slightly change with small value.

Figure 4. Power of hypothesis test of the regression parameters
The Monte Carlo simulation shows that the Bayesian method outperformed to OLS and Ridge regression methods in order to Bias, MSE and Power test. The prior knowledge utilized in the Bayesian regression increase the precision and the accuracy of the parameter estimates. This result is consistent with previous research which informs that the Bayesian method is better than OLS for informative prior selection ([14]; [15]).

3.2. Example: Pollution for 41 Cities in the United States
The application is taken from Sokal and Rohlf (1981) [13]. The data were collected to investigate the determinants of pollution for 41 cities in the United States. There are six independents variable considered such as a negative value of average annual temperature (negtemp), number manufacturing enterprises employing 20 (manuf), population size (pop), average yearly wind speed (wind), average annual precipitation (precip) and average number of days with precipitation per year (days). The dependent variable is the levels of SO2.

3.2.1. Identification of collinearity
The correlation matrix of independent variables can be seen below:
Figure 5 shows the correlation matrix between independent variables. The population (pop) and manufacture (manuf) have a very high correlation. It indicates there is strong collinearity between population and infrastructure.

Table 1. Variance inflation factor (VIF)

|        | Annual Temperature | Manufacture | Population | Wind Speed | Annual Precipitation | Days |
|--------|--------------------|-------------|------------|------------|----------------------|------|
| VIF    | 3.764              | 14.704      | 14.341     | 1.256      | 3.405                | 3.444|

3.2.2 Define prior parameter for super-parameter of variance \( \mathbf{V} \)

We need to define the prior parameter of variance and covariance matrix \( \mathbf{V} \). We set \( \mathbf{V} = \text{diag}(1); \mathbf{V} = \text{diag}(10); \mathbf{V} = \text{diag}(1000) \) and \( \mathbf{V} = \text{diag}(10000) \). Given sensitivity analysis below, we see the \( \mathbf{V} = \text{diag}(100); \mathbf{V} = \text{diag}(1000) \) and \( \mathbf{V} = \text{diag}(10000) \) present similar result, so that we use \( \mathbf{V} = \text{diag}(100) \).

(a) Regression Parameter Estimate \( (\hat{\beta}_k) \)    (b) Asymptotic t \( (t_k) \)

**Figure 6.** Sensitivity Analysis of Prior Parameter of \( \sigma^2 \) for (a) \( \hat{\beta}_k \) and (b) \( t_k \)
3.2.3 Estimated Bayesian Regression

The parameter estimates and testing are presented in Table 2.

**Table 2. Regression Parameter Estimates and Testing**

| Parameters           | OLS       | Ridge     | Bayesian  |
|----------------------|-----------|-----------|-----------|
|                      | Estimate  | t-value   | Estimate  | t-value   | Estimate  | t-value   |
| (Intercept)          | 111.728   | 2.361*    | 97.273    | -         | 34.070    | 6.861*    |
| Annual Temperature   | 1.268     | 2.041*    | 1.250     | -         | 1.265     | 3.309*    |
| Manufacture          | 0.065     | 4.122*    | 0.028     | -         | 0.065     | 261.900*  |
| Population           | -0.039    | -2.595*   | -0.005    | -         | -0.039    | -171.700* |
| Wind Speed           | -3.181    | -1.753    | -2.743    | -         | -3.174    | -0.996    |
| Annual Precipitation | 0.512     | 1.412     | 0.355     | -         | 0.511     | 3.904*    |
| Days Precipitation   | -0.052    | -0.321    | 0.048     | -         | -0.052    | -1.971*   |

Note: *) significant for $\alpha = 5\%$

Table 2 presents the OLS, Ridge and Bayesian regression. Parameter estimates from OLS and Bayesian almost similar, while Ridge regression presents the different result. It indicates the Bayesian approach outperformed of Ridge regression to solve the multicollinearity problem. Using, Bayesian method the asymptotic t-student were calculated, and we found only one independent variable was not statistically significant while based on OLS we found three independent variables were not statistically significant.

4. Conclusion

The hypothesis testing is vital in regression analysis in term of the generalized the results. The hypothesis testing of regression parameters is invalid when multicollinearity occurs. Bayesian approach with informative prior can be used to solve this problem. The informative prior is defined for the super-parameter of the variance intercept and slopes. The Monte Carlo simulation concludes that the Bayesian approach presents the highest power test and the smallest Bias and Means Square Error. The sensitivity analysis shows the optimal value for diagonal super-parameters is 100 and the prior parameters for the intercept and slopes can be taken from ordinary least squares estimates. Based on the simulation result, this method can be used to solve hypothesis testing in regression analysis with multicollinearity problem effectively.

Acknowledgments

This work is supported by the Basic Research Grant of the Directorate General of the Higher Education, Indonesia (PDUPT, SP DIPA-042.06.1.401516/2018).

References

[1] Silvey S 1969 *J. R. Stat. Soc. Series. B. Stat. Methodol.* **31** 539
[2] Daud J 2017 *J. Phys.: Conf. Ser.* **949** 012009
[3] Clark PC 2007 *The Effects of Multicollinearity in Multilevel Models* (Pennsylvania: Wright State University)
[4] Yoo W, Mayberry R, Bae S, Singh K, He Q and Lillard J 2004 *Int. J. Appl. Sci. Technol.* **4** 9
[5] Bager A, Roman M, Algedih M and Mohammed B 2017 *Munich Personal RePEc Archive* **81390** 1
[6] Hoerl A, Kennard R and Baldwin K 1975 *Commun. Stat. Theory Methods* **4** 105
[7] Shih S and Shih W 1979 *J. Hydrol.* **4** 165
[8] Hadgu A 1984 *Stat Med* **3** 293
[9] Brown W G 1976 *Effect Of Omitting Relevant Variables Versus Use Of Ridge Regression In Economic Research* (Oregon : Oregon State University)

[10] Conniffe D, Stone J and O’Neill F 1976 *Irish Journal of Agricultural Economics and Rural Sociology* 6 89

[11] Ye Z S, Li J G and Zhang M 2014 *J. Appl. Stat* 41 1436

[12] Yang X and Su X G 2009 *Linear Regression Analysis Theory and Computing* (USA: World Scientific Publishing Co. Pte. Ltd).

[13] Sokal R and Rohlf F 1981 *Biometry* (San Francisco-California: Freeman)

[14] Elster C and Wübbeler G 2015 *Metrologia*. 54 1394

[15] Mettle F O, Asiedu L, Quaye E N B, and Asare-Kumi A A 2016 *British Journal of Mathematics & Computer Science* 15 1