The Application of Bayesian Models in the Study of Medical Insurance Cost

Zhiwei Gong
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
zgong9@jhu.edu

Xu He
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21218
xhe51@jhu.edu

Bingxu Han
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
bhan22@jhu.edu

Abstract

This project focuses on modeling personal medical cost paid by insurance and deducing significant contributors to the payment. Data analysis of variable distributions and correlations indicates that smoke and age have greater impacts on the cost. The ordinary least square regression models are performed to predict the medical cost, and the backward step-wise regression is used to select the best subset of our model by eliminating one extra parameter each time with log transformation to normalize the data. Through Bayesian Information Criterion and Akaike's Information analysis of the three models, the original model is selected based on better fit with lowest score. However, Residual plots suggest that the selected model is not homoscedastic or normally distributed, which violates OLS assumptions. Therefore, an improved model is proposed by adding nonlinear terms to the regression. The new model shows a more statistically significant result with a lower R-square and higher correlation accuracy. Finally, Markov chain Monte Carlo method is performed to estimate the data and demonstrates a fast convergence reaching the stationary state, indicating the validity of proposed model.

1 Introduction

The U.S. health system comprises a mix of public and private financed coverage. Public coverage aims to provide medical care for elderly and low-income groups, and private coverage is mainly given by employers [1]. Beginning in 2010 after the enforcement of Affordable Care Act (ACA), public and private insurers set down their premium rate and benefits packages which make insurance more affordable.

The premium rate is based on several determinants including location, age, plan category, and dependent numbers. Factors such as population growth and inflammation result in an expected premium rate rise over time, especially with COVID-19 pandemic this year [2]. What is more, in 2021, US health care has reached $12,914 per person, accounted about 18.3% of the total spending [3]. Such high medical costs can be due to the rise in drug prices and hospital care. For hospitalization, surgical procedures in the U.S. greatly exceed those of other countries since hospital and physican costs are billed as fixed numbers depending on the service rather than patient status [4]. According to ACA, health plans must cover an actuarial value of 60% and essential benefits including emergency

service, mental health or rehabilitation; therefore, having medical insurance helps to transfer finical risks and unexpected medical costs since U.S. medical system is complex [5].

Additionally, multisystem regulation adds to the complexity of choosing insurance. People can get private insurance from different providers and in each sector, enrollees need to select among different tiers of coverage based on their status. However, these plans may not include all the benefits they want. Therefore, it is necessary to evaluate medical cost based on different factors to decide which insurance to purchase. In this report, we aim to find out important contributors to medical cost in order to help people choose appropriate plans based on medical cost prediction. As such, we can greatly reduce the impact of premium rate increases and maximize the benefit received from healthcare insurance.

2 Related work

Several factors are specifically associated with insurance costs, such as children numbers, tobacco status and age. According to Wentworth (2015), multi-children family is likely to split children into different insurance programs with distinct service levels upon their coverage, which leads to medical care and payment variations[5]. Another important factor is tobacco consumption. Research suggested that smokers are more likely to get asthma with chronic respiratory symptoms and require hospitalization for treating inflammation[6]. Thus, smokers are charged more for health insurance due to the increased chance of getting medical care. Similar effect with age since the elderlies typically need more medical care. Therefore, those factors are likely to contribute more to coverage.

Several regression methods have been applied to analyze cost-related problems. For example, in right-censored medical cost data, the researchers utilized IPW (the inverse probability weighted) regression with covariates to deduce the mean cost of hospitalization. And through generalized survival-adjusted estimators, they showed that medical cost accumulation intensity indirectly affects survival [7]. In another study exploring spending with children neurologic impairment (NI) disease, researchers assessed cost trends with logistic and linear regression using 15 different parameters. By evaluating annual cost versus each service category using logistic regression, they suggested that inpatient service cost remained the largest annual spending but decreased over time with even distribution across inpatient and outpatient care later on [8].

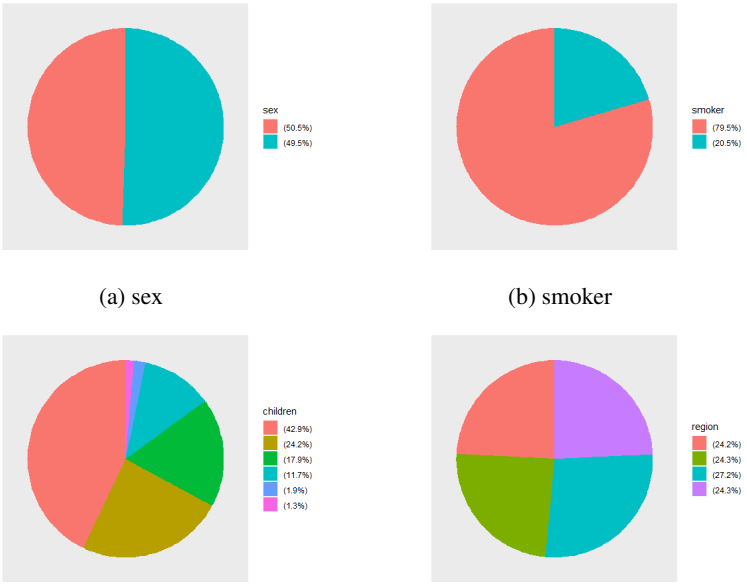
By evaluating the dataset, the team decided to implement ordinary least square regression (OLS) model with different predictor variables to estimate medical cost. The advantage of OLS includes fast computation and easy interpretation as long as the assumptions are met. By modeling different predicted variables and comparing distributions with original data, we can easily deduce weights of the variables affecting the medical cost.

3 Data Exploration and Analysis

This dataset is from Kaggle's Medical Cost Personal Datasets.[9] The dataset has 1338 rows and 7 columns. The data include age, sex, bmi (Body mass index), children (number of children), smoker (smoking or not smoking), region and charges. Specifically, "age" and "gender" are direct variables representing the customer's age and gender information. "Smoke" is a Boolean variable, indicating whether the individual smokes. "bmi" and "charges" are floating variables, representing the individual's body mass index and medical insurance costs, respectively. Children "represents the number of individual children‘ Region 'indicates the region of origin.

Based on the general situation of the data, Figure 1a, 1b, 1c and 1d respectively show the data distribution of gender, smoker, number of children and regions. The gender data is balanced. The number of smokers accounted for only 20.5 % of the total population, about one fifth of the total. As for the number of children, 42.9% of people have no children, 24.2% have only one child, 17.9% have two children and 11.7% have three children. Only 3.2% have four or five children.

Through the above data exploration, the string data is converted to integer type instead. At the same time, the correlation coefficient matrix of the data is obtained (Figure 2). Through the correlation coefficient matrix, it can be found that the correlation coefficients of the target variables charges, age, bmi and smoker are high, of which the correlation coefficient of smoker is the highest, 0.79. Therefore, it can be preliminarily assumed that smoker has the greatest impact on charges. In addition,



(a) sex (b) smoker

(c) children (d) region

Figure 1: Pie chart of sex, children, smoker and regional components

it can be found that the age distribution is similar to the uniform distribution, and bmi is similar to the normal distribution. These distributions will be explored and verified later.

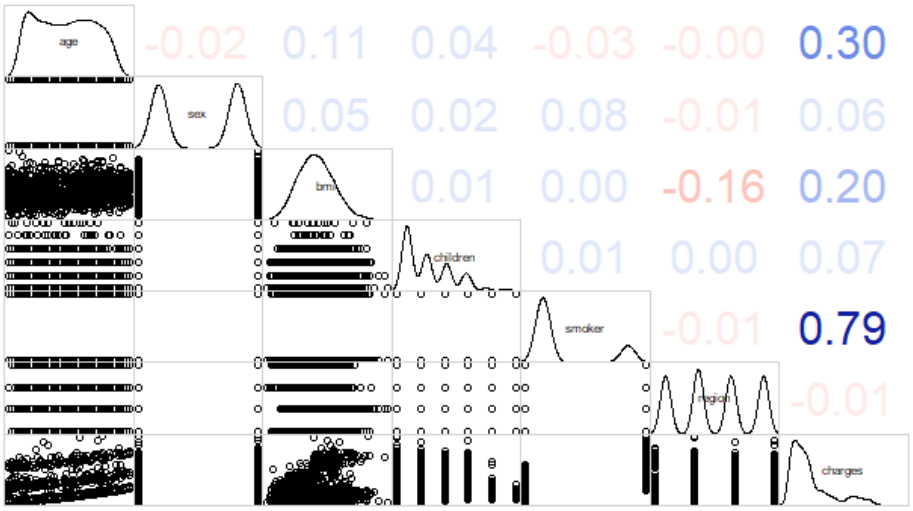
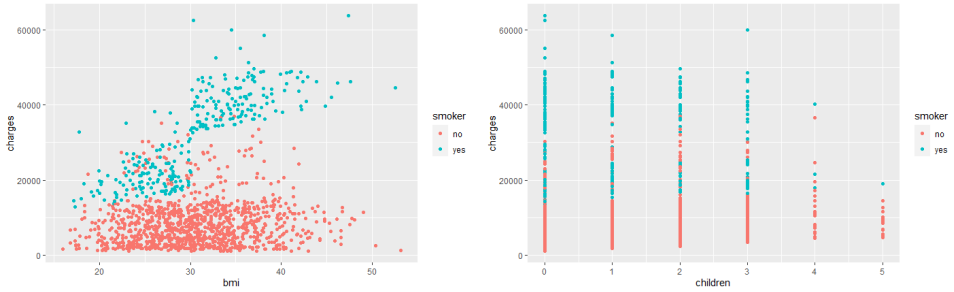


Figure 2: Correlation matrix of age, sex, bmi, childrren, smoker, region and charges

According to the correlation coefficient matrix, draw a scatter map of the number of children and bmi to charges under different smoker conditions. Figure 3a shows the scatter plot of bmi for charges under different smoker conditions. Obviously, the overall bmi of smokers is higher than that of non-smokers. In addition, among smokers, the higher the bmi, the higher the corresponding charges. Interestingly, there is a large gap around the boundary of bmi equal to 30. Figure 3b shows the scatter

plot of the number of children for charges under different smoking conditions. It can be found that the more children there are, the less people smoke. At the same time, with the same number of children, the charges corresponding to smokers are greater than those corresponding to non-smokers.



(a) Scatter plots of charges with bmi (b) Scatter plots of charges with children

Figure 3: Scatter plots under different smoker conditions

4 Methodology

In this project, in order to determine if there is a relationship between attributes and medical costs and to predict the costs, we used the ordinary least square (OLS) linear regression models to model our data and fit the model using the Bayesian approach. For our Bayesian model, Monte Carlo Markov Chain (MCMC) is performed to model our data and estimate the regression parameters.

4.1 OLS Regression Models

We first use the OLS linear regression models for our Bayesian model. According to the results of data analysis, especially for the correlation matrix we get, we can see that the variable "region" has the lowest correlation value with our target variable "charges". So we did not include the "region" for our linear regression models. Since from Figure 2, we can see that the distribution of "charges" is right-skewed, which will result in a poor fitting model, given that all the values of "charges" are positive, we perform the *log-transformation* on it. Then the distribution of the new variable "$log(charges)$" is less skewed and more symmetric (Figure 4).

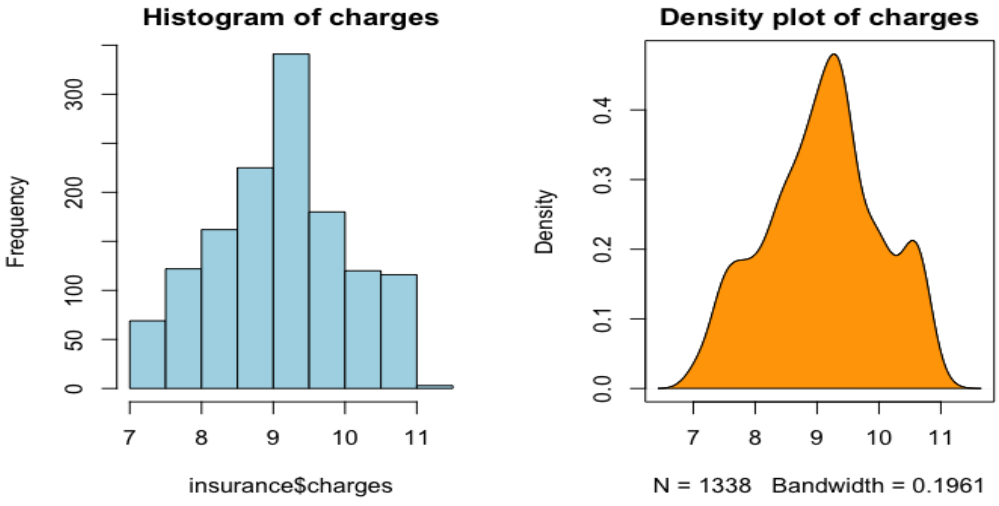


Figure 4: Distribution of the variable "charges" after log-transformation

Then we propose three multiple linear regression models as follows:

$$log(charges)_i = \beta_0 + \beta_1 * age_i + \beta_2 * smoker_i + \beta_3 * bmi_i + \beta_4 * sex_i + \beta_5 * children_i + \epsilon \quad (1)$$

$$log(charges)_i = \beta_0 + \beta_1 * age_i + \beta_2 * smoker_i + \beta_3 * bmi_i + \beta_4 * sex_i + \epsilon \tag{2}$$

$$log(charges)_i = \beta_0 + \beta_1 * age_i + \beta_2 * smoker_i + \beta_3 * bmi_i + \epsilon \tag{3}$$

where each row of our dataset is indexed by i.

Based on the three models we proposed, stepwise regression is used to provide a number of stopping rules for selecting the best subset of variables for our model. In step-wise selection, variables are added to or deleted from a model one at a time, until some stopping criterion is reached. For our three models, we use the backward stepwise. The model selection criteria we used are Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC), which will be shown in the **Results** section.

4.2 Bayesian Analysis - MCMC Estimation

In our project, we performed MCMC estimation on model Eq.(3). The method of MCMC estimation for an OLS linear regression model involves constructing a Markov Chain whose states correspond to possible values of the model parameters. These parameters include the intercept term and the coefficients for each predictor variable. The transitions between states in the Markov Chain are governed by the posterior distribution of the model parameters given the data.

The priori distribution we assume is as follows:

$$\beta_{prior} = \begin{pmatrix} \beta_{Intercept} \\ \beta_{age} \\ \beta_{smoker} \\ \beta_{bmi} \end{pmatrix}, \mu_{prior} = 1, \sigma_{prior} = (\mathbf{X}^T * \mathbf{X})^{-1} * \sum \frac{res^2}{n-p} * n \tag{4}$$

where $\beta_{Intercept}$ is the intercept of model, β_{age} is the parameter of age, β_{smoker} is the parameter of smoker, β_{bmi} is the parameter of bmi. **x** is variable of intercept, age,smoker and bmi, res is residual error of model.n is the length of variable, p is the dimension of x.

Then we update β through the following formula:

$$\beta_v = ({\sigma_{prior}}^{-1} + \frac{\mathbf{X}^T * \mathbf{X}}{\sigma_2})^{-1}, \beta_e = \beta_v * ({\sigma_{prior}}^{-1} * \beta_{prior} + \mathbf{X}^T * \mathbf{Y} * \sigma_2) \tag{5}$$

where σ_2 is the residuals of model: y = 0+x.

$$\beta_{update} = mvnorm(\beta_e, \beta_v) \tag{6}$$

Then we update σ through the following formula:

$$\mu_n = \mu_{prior} + n, ss_n = \mu_{prior} * 15^2 + \sum (\mathbf{Y} - \mathbf{X} * \beta_{update})^2 \tag{7}$$

$$\sigma_{update} = \frac{1}{gamma(\frac{\mu_n}{2}, \frac{ss_n}{2})} \tag{8}$$

The details of implementation are shown in the attached code file, and the results will be discussed in the **Results** section.

5 Results

5.1 Evaluation Metrics

In our project, the model selection criteria for stepwise regression to choose the important variables are BIC and AIC, both show better performance with smaller values. Figure 5 shows the results of our models Eq.(1), Eq.(2), and Eq.(3), respectively. According to the result, we can conclude that the first model Eq.(1) is the best.

	Rank <dbl>	Df.res <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>	R.squared <dbl>	Adj.R.sq <dbl>	p.value <dbl>	Shapiro.W <dbl>	Shapiro.p <dbl>
Model1	6	1323	1641	1641	1677	0.7627	0.7618	0	0.8405	1.569e-34
Model2	5	1324	1735	1735	1766	0.7450	0.7442	0	0.8812	1.049e-30
Model3	4	1325	1740	1740	1766	0.7436	0.7430	0	0.8833	1.717e-30

Figure 5: BIC and AIC values for our three proposed models

5.2 Diagnostic - Check Model Assumptions

In this section, we perform model diagnostics for our selected model. According to the residual plots in Figure 6, we can conclude that: (i) At Residuals x Fitted plot, the non horizontal line may indicate a non-linear relationship. (ii) At Normal QQ plot, we see that the residuals are not exactly on a straight line, indicating that they are not normally distributed. (iii) At Scale-Location plot, the non straight line indicates heteroscedasticity. With these violations of assumptions, insights and predictions from this model may be inefficient or wrong. Since the log-transformations have been performed, it's possible to add polynomial or quadratic terms to this initial model.

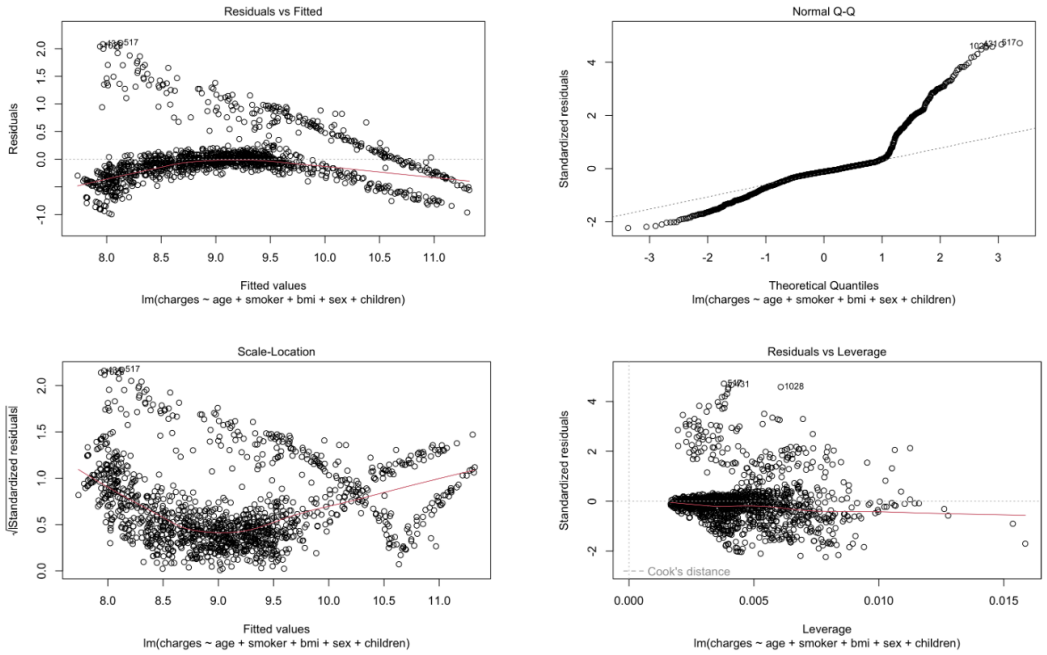


Figure 6: Residual plots of the first model

5.3 Improving Model Performance

In order to improve the performance of our model, we made the following adjustments: (i) Add the non-linear term for variable "age" and "bmi". (ii) Create an indicator for the new variable "obesity", where "obesity" is assigned to 1 if "bmi" is larger than 30, otherwise assigned to 0. (iii) Specify an interaction between "obesity" and "smoking". The new model is:

$$\text{log(charges)}_i = \beta_0 + \beta_1 * age_i + \beta_2 * smoker_i + \beta_3 * bmi_i + \beta_4 * sex_i + \beta_5 * children_i + \beta_6 * age_i^2 + \beta_7 * bmi_i^2 + \beta_8 * (smoke_i * obesity_i) + \epsilon$$

Table 1 shows the results by comparing the old model and the new model. As we can see, they both are statistically significant, and prediction power grows after the improvements are implemented. According to the values of R-square, Residual standard error, and correlation accuracy, our new model is better than the previous one. The R-square value goes from 0.7627 in previous model to 0.7903 in the new model, meaning the data deviance from the regression line has notably improved in the new model.

Table 1: Comparison between the performance of the original model and the improved model

	Original Model	Improved Model
R-square	0.7627	0.7903
Residual standard error	0.4473	0.4196
p-value	2.2e-16	2.2e-16
Correlation accuracy (%)	86.02	89.23

5.4 MCMC Estimation Results

In our project, the iteration number of MCMC sampling is set to be 5000. The posterior mean for β we get is (7.058, 0.035, 1.539, 0.012) for $\beta_0, \beta_1, \beta_2, and \beta_3$, repectively. And the posterior mean for σ^2 is 0.216. The correlation accuracy is 0.878. From Figure 7, we can see that it has been assessed the convergence of the Markov chain, and from Figure 8, the ACF plot, we can conclude that all autocorrelation values for β and σ^2 are small and they all have a fast convergence rate to the stationary distribution.

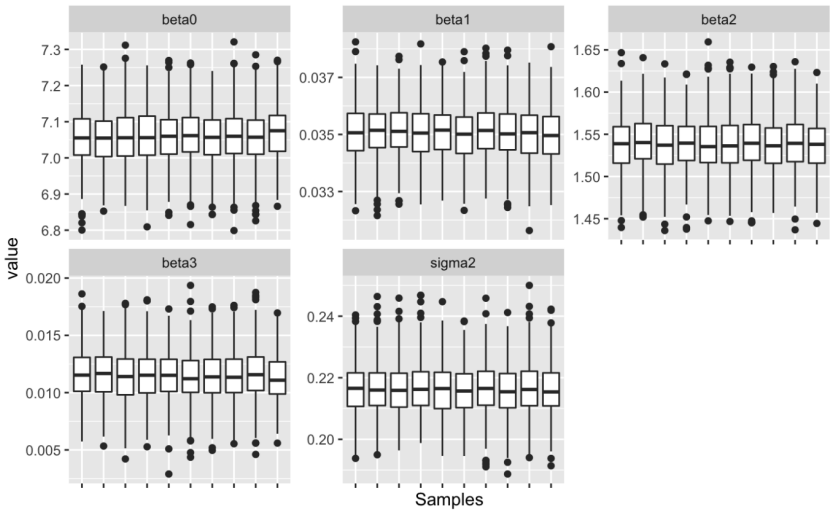


Figure 7: Assess the convergence of Markov chain after 5000 iterations

6 Conclusion and Discussion

We frame the problem of personal medical cost prediction as a regression problem. Inspired by the Bayesian approaches of OLS regression model and MCMC estimation, we formulated a new model by adopting these two methods to analyze important factors contributing to cost. We achieved better performance after selecting the best subset of variables and making some adjustments to the non-linear terms.

Despite the preliminary success in predicting the medical cost, the proposed Bayesian model violates some assumptions of the regression models. Therefore, applying the hierarchical linear model and generalized models could be a promising future direction.

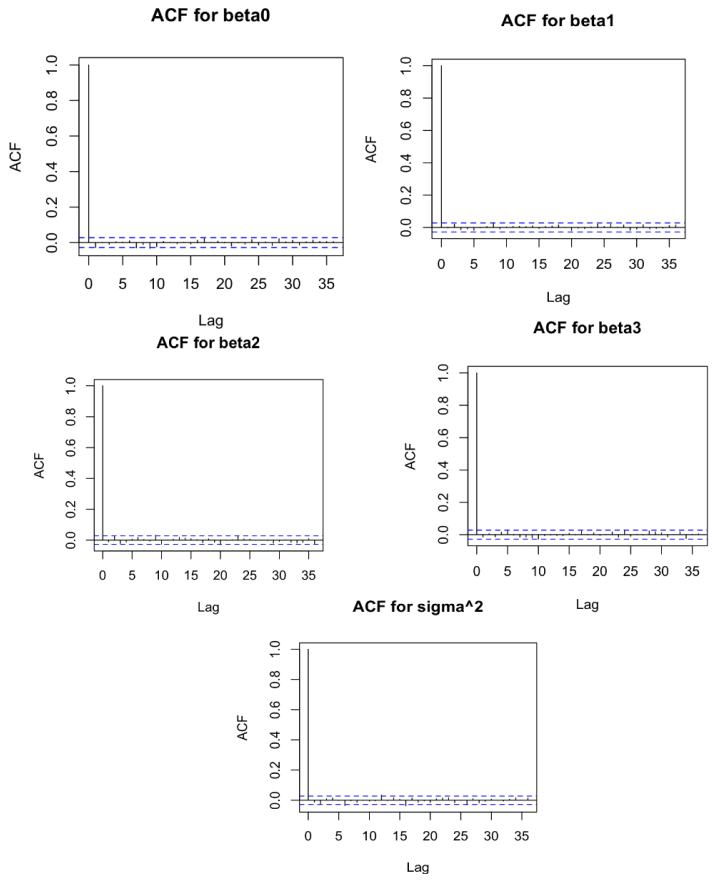


Figure 8: ACF plots for beta and sigma^2

References

[1]"United States," Jun. 05, 2020. https://www.commonwealthfund.org/international-health-policy-center/countries/united-states (accessed Dec. 19, 2022).

[2]"Legislative News, Studies and Analysis | National Conference of State Legislatures." https://www.ncsl.org/ (accessed Dec. 19, 2022).

[3]"National Health Expenditures 2021 Highlights".

[4]D. Gordon, "The Average Cost of Health Insurance in 2022," MoneyGeek.com, Dec. 28, 2021. https://www.moneygeek.com/insurance/health/average-cost-of-health-insurance/ (accessed Dec. 19, 2022).

[5]"Find out what Marketplace health insurance plans cover," HealthCare.gov. https://www.healthcare.gov/coverage/what-marketplace-plans-cover/ (accessed Dec. 19, 2022).

[6]N. C. Thomson, R. Polosa, and D. D. Sin, "Cigarette Smoking and Asthma," J. Allergy Clin. Immunol. Pract., vol. 10, no. 11, pp. 2783–2797, Nov. 2022, doi: 10.1016/j.jaip.2022.04.034.

[7]L. Deng, W. Lou, and N. Mitsakakis, "Modeling right-censored medical cost data in regression and the effects of covariates," Stat. Methods Appl., vol. 28, no. 1, pp. 143–155, 2019.

[8]N. D. Bayer, M. Hall, Y. Li, J. A. Feinstein, J. Thomson, and J. G. Berry, "Trends in Health Care Use and Spending for Young Children With Neurologic Impairment," Pediatrics, vol. 149, no. 1, p. e2021050905, Jan. 2022, doi: 10.1542/peds.2021-050905.

[9]Medical Cost personal Dataset:https://www.kaggle.com/datasets/mirichoi0218/insurance B. Lantz, Machine Learning with R. Packt Publishing Ltd, 2013.