Bayesian Statistics Final Assignment

Alex Carriero (9028757): a.carriero@students.uu.nl

1. Introduction

The focus of the paper is the analysis of publicly available medical insurance data provided by an anonymous American insurance company. For insurance companies, it is vital to identify factors that are associated with increased medical costs and also, to understand how these factors interact. The data set provides information regarding patient age, smoking status, body mass index (BMI) and total medical bills accrued in a year's time. Of course, it is no secret that factors like age, smoking and BMI affect one's health, and thus, the medical charges one accrues in a years time. To provide an enriched perspective, my analysis will take a deeper look at the relations among these factors.

My interest in the data is two fold: I would like to know which of these factors has the largest impact on health care costs and further, I would like to know if the effect of BMI on health care costs depends on whether or not a person smokes. Naturally, the answers to these research questions are of importance to medical insurance companies, as they must make predictions regarding how much each person is likely to cost them per year. However, the same conclusions are applicable to government funded health care programs, which could largely benefit from making accurate predictions surrounding expected annual medical costs.

To best answer these research questions, Bayesian regression techniques are employed. First, I present two specific hypotheses, followed by the methods that are used to test the hypotheses. Research methods and results are discussed and justification for utilizing Bayesian as opposed to frequentist techniques is provided. Finally, conclusions to the research questions are drawn.

2. Hypotheses

H_1: Out of the three predictors considered, age has the largest impact on health care costs.

This hypothesis is investigated by comparing the standardized regression coefficients from a Bayesian regression model including all three predictors. Additionally, evidence in favor of this hypothesis vs. its complement is evaluated using the Bayes Factor.

H_2: The effect of BMI on heath care costs depends one's smoking status (i.e., an interaction effect is present between BMI and smoking status).

This hypothesis is investigated by comparing a Bayesian regression model which includes an interaction effect to one without, by means of DIC. The parameter estimate and credible interval for the interaction effect is also considered and further, the Bayes Factor is employed to test this hypotheses against its complement.

3. Methods

3.1 Data Set

The medical insurance data set is comprised of observations from 1338 individuals measured on four variables: total health care costs per year, age (years), BMI (kg/m^2), and smoking status. Smoking status is a binary variable while the others are continuous. The response variable, total health care costs per year, is heavily right skewed; therefore, it is transformed by the natural logarithm such that it becomes symmetric and approximately exhibits a normal distribution. For the analysis, all data are standardized, as standardized regression coefficients will be utilized to investigate hypothesis 1.

3.2 Model Selection

Two Bayesian regression models are fit in this analysis. Their regression equations are:

Simple model: $\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

Interaction model: $\log(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{3i} x_{4i} + \varepsilon_i$

where, i is an individual indicator, y represents total annual medicals costs, x_1 represents age (years), x_2 represents BMI (kg/m^2), x_3 represents smoking status (1 = smoker, 0 = non-smoker), and ε represents random error. For both equations, it is assumed that the random errors follow a normal distribution with mean 0 and variance σ^2 ($\varepsilon_i \sim N(0, \sigma^2)$). The density of the data for both models is specified to be a normal distribution with variance σ^2 and mean $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ for the simple model and $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3 x_4$ for the interaction model.

3.3 Prior Specification

In Bayesian analysis we assume that the parameters are themselves random variables with their own distributions that we, the researchers, are afforded the opportunity to choose. This report is of course not the first time someone has analyzed health insurance data including these factors, thus, I would like to incorporate prior information in my analysis. The Medical Expenditure Panel Survey (MEPS) collects data on a yearly basis regarding health insurance costs and other person criteria. Data from the year 2018 (the same year as my data set) were downloaded, total annual medical charges were log-transformed, all data were standardized, and a linear regression was conducted using smoking, age and BMI as predictors. Informative priors for β_0 through β_3 were selected based on the regression estimates; the prior means are identical to regression estimates and all variances are set to 0.01, rather than their estimates, to allow for uncertainty. For BMI (β_2) I have selected the t-distribution to allow for extra uncertainty, while the others follow a normal distribution. For β_4, I specify a weakly informative prior, the standard normal distribution; it is centered around zero such that the data must be compelling to show evidence of an interaction (i.e., parameter estimates are shrunk towards zero unless there is strong evidence of an effect), this is done to help protect against false positive findings (Lemoine, 2019). Finally, I allow the error variance of the model to follow an uninformative, inverse gamma prior. The priors for each variable are as follows:

$\beta_0 \sim N(-0.31, 0.01)$, $\beta_1 \sim N(0.32, 0.01)$, $\beta_2 \sim t(\nu = 1338 - q + 1, 0.01, 0.01)$, $\beta_3 \sim N(0.08, 0.01)$, $\beta_4 \sim N(0, 1)$, $\sigma^2 \sim IG(0.001, 0.001)$, where q represents the number of estimated parameters.

3.4 MCMC Sampler

3.4.1 Specifications

The MCMC sampler utilized for this analysis operates based on a hybrid of Gibbs and Metropolis-Hasting (MH) sampling. For the regression parameters β_0, β_1, β_3, β_4 and σ^2, the priors specified are conjugate priors for the density of the data, thus, they are estimated using Gibbs sampling. Conversely, the prior specified for β_2 is not a conjugate prior for the density of the data, and therefore, a MH step utilizing a random walk is implemented in the sampler for the estimation of this parameter. With respect to the binary predictor variable, smoking is coded as 1 = smoker and -1 = non-smoker within in the sampler. For both models, 2 chains with different starting values are utilized and 10000 burn-in iterations are allowed so that the MCMC sampler may converge to a steady state, after the burn-in, a further 10000 samples are taken.

3.4.2 Convergence

It is vital that MCMC sampler convergence be carefully assessed as poor convergence will result in poor estimation of model parameters. For both models in this analysis, MCMC sampler convergence is assessed utilizing 4 methods: trace plots, autocorrelation plots, Gelman Rubin Statistics and MCMC error. Trace plots and autocorrelation plots (displayed to 40 lags) are generated for each model parameter. The Gelman Rubin Statistic is also computed and reported for each model parameter, using the the total sample (discarding burn-in iterations); additionally, the Gelman Rubin Statistic is calculated throughout the sample at intervals of 100 iterations for each parameter and displayed in a graph. Finally, MCMC error is calculated for each model parameter and checked to ensure it is less than 5% of a parameter's respective standard deviation. To visualize the conditional posterior distributions, density graphs of the MCMC samples for each parameter are generated; the plots are not included in the report but can be viewed in Appendix II if they are of interest. Convergence results are displayed for the interaction model only.

3.5 Model Assumptions

This Bayesian linear regression analysis hinges on the assumption of normally distributed regression residuals. This assumption is investigated by means of a posterior predictive check (ppc). Using 1000 sets of sampled parameter values from the MCMC sampler, 1000 simulated data sets are generated. For each set of sampled parameter values, fitted responses are calculated and two sets of regression residuals are subsequently computed, based on the observed and simulated responses, respectively. A discrepancy measure is then calculated for each set of residuals and a posterior predictive p-value is computed by calculating the proportion of pairwise comparisons (observed vs. simulated) that results in the simulated discrepancy measure being greater than that of the observed.

For this posterior predictive check an original discrepancy measure is utilized. It is known that when data are normally distributed, 68.3% of the observations fall within one standard deviation of the mean. For a given set of residuals, I compute the proportion of residuals that fall within one standard deviation of their mean and retrieve the square distance of this proportion from 0.683. If the residuals are normally distributed I expect this distance to be quite near zero, conversely, if the data are non-normal this distance will be larger than zero. The efficacy of this discrepancy measure is demonstrated in a small simulation whereby normal and right skewed data are generated and the discrepancy measure preforms as expected. Furthermore, if the observed regression residuals are normally distributed, I expect to see a posterior predictive p-value of approximately 0.5.

Results

Figure 1. Trace Plots for the Regression Parameters of the Interaction Model

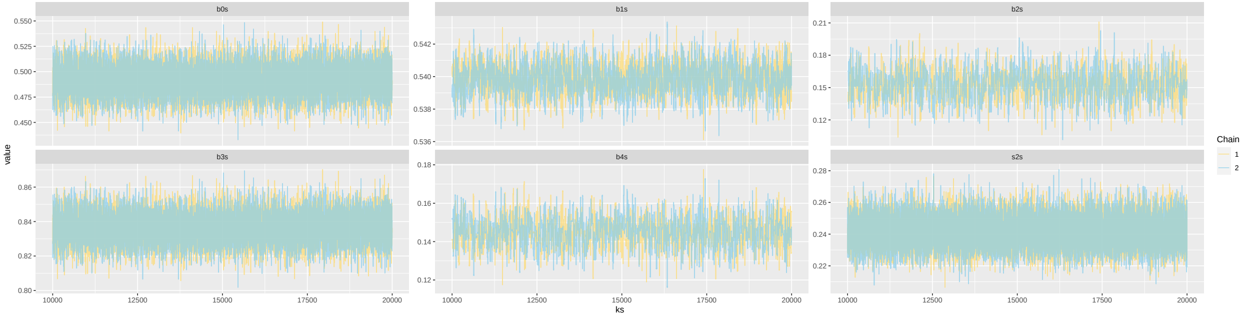


Figure 2. AutoCorrelation for the Regression Parameters of the Interaction Model

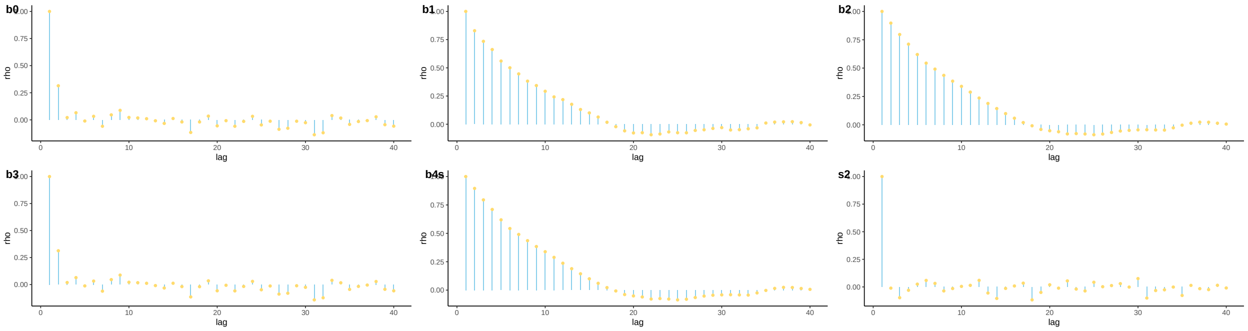
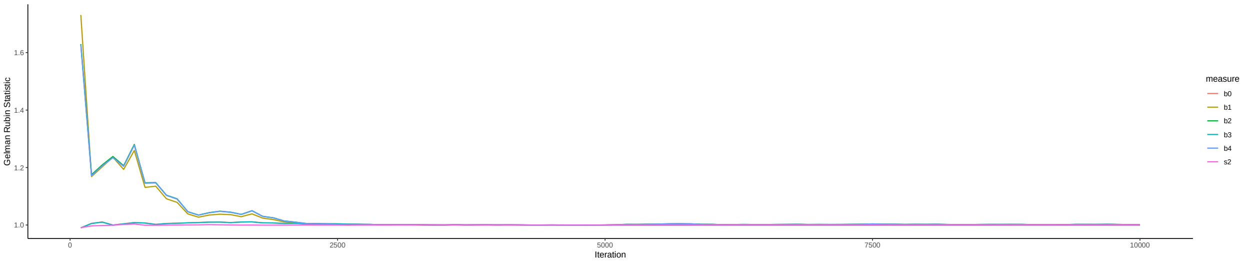


Figure 3. Gelman Rubin Plot for the Regression Parameters of the Interaction Model



An assessment of convergence is demonstrated for the interaction model; please note that convergence criteria are the same or better for the simple model. Figure 1 displays trace plots of each of the estimated parameters, showing the sampled values for each iteration (excluding burn-in) in both chains. Here we see that the trace plots look like fat hairy caterpillars indicating that the MCMC sampler has reached a steady state. Figure 2 highlights sampler autocorrelation, by lag 20 all autocorrelation is near zero, I find this level of autocorrelation acceptable. Figure 3 summarizes the Gelman Rubin statistic throughout the iterations for each parameter, here we see that by 2500 iterations, all Gelman Rubin statistics are stable and very close to 1. After 10000 iterations the Gelman Rubin statistics are 0.999, 1.001, 1.001, 0.999, 1.001, 1.000 for β_0 through σ^2, respectively. Finally, MCMC error is included in Table 1 and 2; here we see that it never surpasses 5% of a parameter's respective standard deviation.

Table 1: Summary Statistics for MCMC Sampled Parameter Values in the Simple Model.

	Mean	SD	Naive SE	2.5%	Median	97.5%
b0	0.49245	0.01467	0.00015	0.46352	0.49256	0.52082
b1	0.53651	0.00150	0.00001	0.53364	0.53652	0.53943
b2	0.07143	0.01355	0.00014	0.04512	0.07124	0.09758
b3	0.83685	0.00865	0.00009	0.81978	0.83691	0.85362
s2	0.25598	0.00993	0.00010	0.23716	0.25558	0.27633

Table 2: Summary Statistics for MCMC Sampled Parameter Values in the Interaction Model.

	Mean	SD	Naive SE	2.5%	Median	97.5%
b0	0.49172	0.01427	0.00014	0.46382	0.49182	0.51919
b1	0.53988	0.00088	0.00001	0.53815	0.53990	0.54157
b2	0.15247	0.01356	0.00014	0.12598	0.15234	0.17927
b3	0.83632	0.00841	0.00008	0.81987	0.83639	0.85252
b4	0.14472	0.00756	0.00008	0.12986	0.14463	0.15964
s2	0.24156	0.00929	0.00009	0.22429	0.24132	0.25999

Using a posterior predictive check, I assess the normality assumption associated with linear regression. The posterior predictive check results in a p-value of 0, for both the simple and interaction model, indicating that unfortunately, neither model is a good fit to the data. In other words, the observed residuals do not behave as if they were generated with the assumed data generating mechanism. Rather, the residuals are heavily right skewed, and thankfully, my discrepancy measure had the power to detect this.

Parameter estimates and credible intervals for the sampled parameter values are displayed in Tables 1 and 2. With regards to the first hypothesis, from Table 1, we see that the largest parameter estimate in the simple model belongs to smoking status ($\beta_3 = 0.84$), with associated credible interval: $(0.82, 0.85)$. The credible intervals in the tables represent the range in which I believe there is a 95% probability the regression coefficients lie, based both on the data and prior information. Since the regression coefficients are standardized, there is evidence that the factor with the largest impact on health care costs is smoking, not age. Further, we see that the credible interval for smoking does not overlap with the credible intervals for age or BMI indicating that it is clearly the most important factor. To test my specific hypothesis that $\beta_2 > \beta_1$, $\beta_2 > \beta_3$ (i.e., age is the important factor) against its complement, I calculate a Bayes Factor. The Bayes Factor (BF) is zero, indicated that these data offer no evidence in favor or my hypothesis. This is in disagreeance with my prior information, as the MEPS analysis showed age as the most important factor.

With regards to the second hypothesis, from Table 2 we see that the estimate of the regression coefficient for the interaction effect is $\beta_4 = 0.15$ with associated credible interval: $(0.13, 0.16)$. Based on the data and prior information, I believe there is a 95% probability that the interaction effect lies in this range and thus, there is evidence that the interaction effect is non-zero. In other words, the effect of BMI on total health care charges depends on whether or not a person smokes; specifically, the effect of BMI on annual health costs is greater for smokers than for non-smokers. To provide further evidence, I compared the simple model with the interaction model by means of DIC. The DIC values were 1975.3, 1897.5, for the simple and interaction model, respectively; lower DIC indicates superior model fit and thus, there is sufficient evidence of an interaction effect. Finally, I compute a BF to test the hypothesis that $\beta_4 > 0$ against its complement; the Bayes Factor is very large (BF > 10000) and thus, again, there is evidence in favor of an interaction effect.

5. Frequentist vs. Bayesian

To compare my results with frequentist regression, I fit both the simple and interaction model using the lm() function in base R. I retrieve the same parameter estimates accurate to two decimal places, however, the standard errors in the Bayesian analysis are smaller. This is advantageous, especially to insurance companies (or governments) who's prime interest lies in producing precise predictions. Upon reflection three additional advantages come to mind. First, I was able to incorporate prior information into my analysis, this is a huge advantage as in this context, there is already much medical insurance information available that can be used to help answer my research questions, assuming it is good quality, it would be a shame to waste this information. Further, the use of a weakly informative prior for the interaction effect could be used to help protect against false positive findings. Secondly, the credible intervals for my parameter estimates have an intuitive meaning. Finally, in this analysis, the model assumptions clearly do not match the data generating process, as the posterior predictive check for normality in residuals demonstrates strong evidence against this assumption. Thus, even with the log-transformation it is inappropriate to assume a normal data generating process for this data. Thanks to the flexibility offered by Bayesian methods, we can simply change the assumed data generating process. For instance, the next steps for this analysis could be to investigate a skew-normal or skew-t distribution as the data generating mechanism, for the raw outcome variable, which is heavily right skewed.

6. Discussion

In summary, my research questions were investigating utilizing model parameter estimates, credible intervals and model comparisons via DIC and BF. Based on prior knowledge, (i.e.,the data from the MEPS survey of the same year) I hypothesized that age was the most impactful factor, as this was the case in the MEPS data, however, my data set told a different story. The analysis showed that smoking was the most impactful factor for prediction of annual health costs and there was no evidence in favor of my hypothesis. Additionally, I investigated the presence of a potential interaction effect between BMI and smoking status. In this case, there was overwhelming evidence of interaction effect, indicating, that the effect of BMI on annual health costs depends on one's smoking status. Unfortunately, the models fit in this analysis failed the posterior predictive check for normality of residuals, thus, these methods were not the best way to analyze this data. Furthermore, since informative priors were implemented, a sensitivity analysis should also be conducted to assess the robustness of the estimates under different prior specifications. In conclusion, the research questions should be further investigated using Bayesian regression methods in which a more reasonable data generating mechanism is assumed.

References

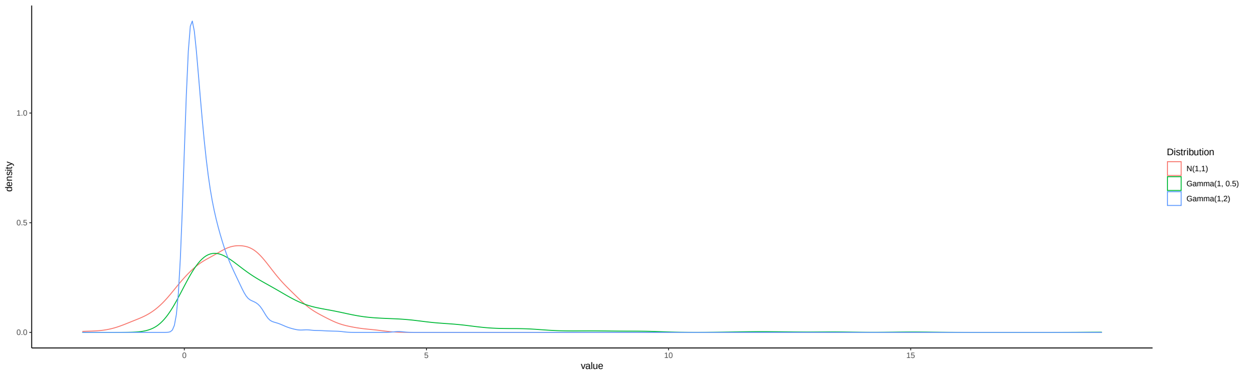
Data Retrieved from: https://www.kaggle.com/datasets/mirichoi0218/insurance

Data for Prior Retrieved from: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-209

Lemoine, N.P. (2019), Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos, 128: 912-928. https://doi.org/10.1111/oik.05985

Appendix

I: Demonstration that Discrepancy Measure has Power

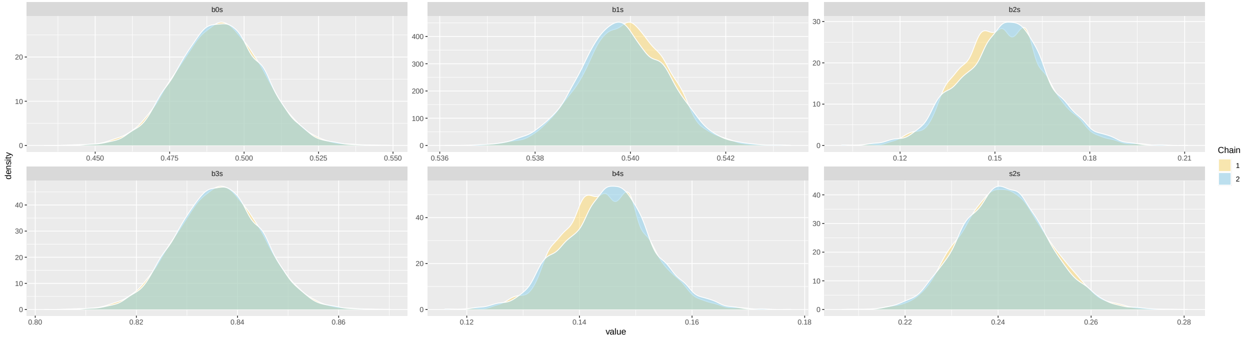


Distribution	Mean	Standard Deviation
Normal (1,1)	0.00014	0.00018
Gamma (1, 0.5)	0.02995	0.00646
Gamma (1, 2)	0.03000	0.00672

A small simulation is conducted to assess the power of my discrepancy measure. After using lm() to analyze the data, it is clear that the regression residuals are highly right skewed. Therefore, in this investigation I simulate many sets of normal ~N(1,1), right skewed ~Gamma(1, 0.5) and extremely right skewed ~Gamma(1,2) data and study the results of my discrepancy measure. More specifically, 1000 sets of "residuals" are generated for each of the three distributions, my discrepancy measure is then calculated for each of them. For each distribution, the mean discrepancy measure and the standard deviation of the discrepancy measure is reported in the table above. Recall that if residuals are normal, the discrepancy measure should be quite near to zero, otherwise, the measure should be larger than zero. From the results, it is clear that my discrepancy measure has the power to detect deviations from normality when applied to right skewed data.

II:

Figure 4. Density Plots of the Regression Parameters of the Interaction Model by Chain


b0s
b1s
b2s
density
b3s
b4s
s2s
value
Chain