STATS451 - Analysis Report

Team Zellner: Wenxing Wang, Yueying Hu

Jiazhang Yu, Haoran Cheng

April 2022

1 Introduction

For the last two decades, heart disease has been the leading cause of mortality for people in the worldwide. The number of deaths from heart disease has increased by almost 2 million since 2000, to the point that it now accounts for approximately 16% of all deaths. Heart disease is a broad term that refers to a variety of illnesses that affect the heart. It encompasses a variety of vascular symptoms, including coronary artery disease, arrhythmias, and congenital heart problems. The term ”heart disease” is often interchangeable with the term ”cardiovascular disease”, which generally refers to conditions involving narrowed or blocked blood vessels that can cause heart attacks, angina, or strokes(Heart Disease Fact). According to a study by the Centers for Disease Control and Prevention (CDC), more than half of Americans have at least one of three major risk factors for cardiovascular disease: hypertension, high cholesterol, or smoking (Pytlak, Kamil). Unhealthy lifestyles such as excessive alcohol consumption or insufficient physical exercise are among the main causes for heart disease. Thus, in order to advise people on how to minimize the risk of heart disease, the purpose of our research is not only to identify the key indicators for cardiovascular disease, but more importantly, to avoid them in future healthcare. Fortunately, We have achieved at the satisfactory outcomes by using a variety of models and methodologies, which will be addressed in further detail in later sections.

2 Dataset

2.1 Variable Selection & Preprocessing

We obtained our data from the online data science community Kaggle. The original dataset contains 319795 observations and 18 variables. Since most of the variables are categorical and are described in texts, we first used as.factor() to transform variables with less than 15 unique values to categorical variables, and then used as.numeric() to turn these categories into continuous integers in order to conduct Lasso regression. We ended up with 10 variables selected by Lasso to perform further modeling. Finally, we transformed the categorical variables (Smoking, Stroke, Diffwalking, Diabetic, KidneyDisease, SkinCancer, Sex, AgeCategory) to dummy variables using dummy_cols() from the library fastDummies, which gave us 21 predictors in total. The variables of interest are:

- **HeartDisease**: 1 if the subject has ever reported having coronary heart disease or myocardial infarction, 0 otherwise;
- **BMI**: Body Mass Index (range from 12.02 to 94.85);
- **Smoking**: 1 if the subject has smoked 100 cigarettes in their entire life, 0 otherwise;
- **Stroke**: 1 if the subject has ever had a stroke, 0 otherwise;
- **PhysicalHealth**: days in the past month that the subject had physical illness or injury (range from 0 to 30)
- **DiffWalking**: 1 if the subject has serious difficulty walking or climbing stairs, 0 otherwise;
- **Sex**: 1 if the subject is male, 0 otherwise;
- **Diabetic**: 1 if the subject has ever had diabetes, 0 otherwise;
- **KidneyDisease**: 1 if the subject has ever had kidney disease not including kidney stones, bladder infection or incontinence, 0 otherwise;
- **SkinCancer**: 1 if the subject has ever had skin cancer, 0 otherwise;
- **AgeCategory2**: 1 if the subject is 25-29 years old, 0 otherwise;
- **AgeCategory3**: 1 if the subject is 30-34 years old, 0 otherwise;

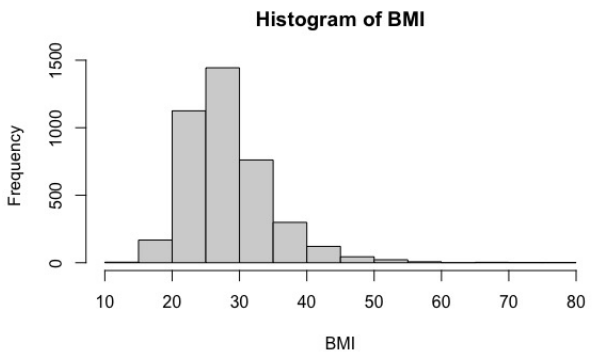
- **AgeCategory4**: 1 if the subject is 35-39 years old, 0 otherwise;
- **AgeCategory5**: 1 if the subject is 40-44 years old, 0 otherwise;
- **AgeCategory6**: 1 if the subject is 45-49 years old, 0 otherwise;
- **AgeCategory7**: 1 if the subject is 50-54 years old, 0 otherwise;
- **AgeCategory8**: 1 if the subject is 55-59 years old, 0 otherwise;
- **AgeCategory9**: 1 if the subject is 60-64 years old, 0 otherwise;
- **AgeCategory10**: 1 if the subject is 65-69 years old, 0 otherwise;
- **AgeCategory11**: 1 if the subject is 70-74 years old, 0 otherwise;
- **AgeCategory12**: 1 if the subject is 75-79 years old, 0 otherwise;
- **AgeCategory13**: 1 if the subject is 80 years old or older, 0 otherwise;

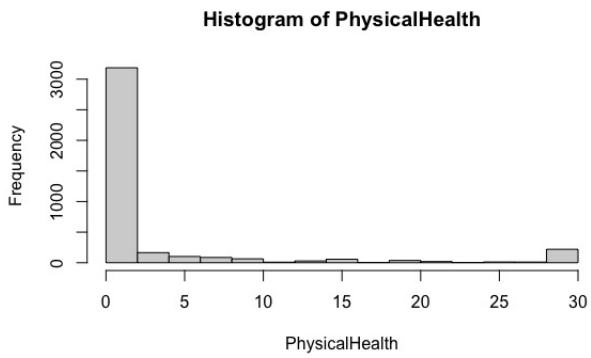
2.2 Exploratory Data Analysis

Since this is a relatively high dimensional space, we would like to use Markov Chain Monte Carlo simulation to assess the performance of our proposed models. To effectively facilitate our computational process, we randomly sample 4000 observations as our dataset of interest. This sample is a reasonable representation of the whole dataset, since its descriptive statistics are very similar to that of the whole dataset, as presented in Table 1 and Table 2.

Table 1. Comparison of Continuous Variables Statistics

Variable	Sample Mean	Whole Mean
BMI	27.34	27.34
PhysicalHealth	3.322	3.372





Next, we plotted the histograms of the two continuous variables to check if any transformation is needed. Although the histogram of BMI is slightly skewed to the right, it is mainly affected by a few outliers. Most of the data still center around the mean (27.34), and thus it is safe for us to assume a normal distribution for BMI. For PhysicalHealth, most data have the value of 0, followed by 30 (indicating a chronic disease). However, the rest of the values (from 1 to 29) also have varying frequencies. It would be tedious to do a factor transformation here, nor is it scientifically supported to divide these values into intervals. Therefore, we decide to simply leave this variable as a continuous one. Finally, we normalized both of the continuous variables with scale().

Table 2. Comparison of Categorical Variables Statistics

Variable	Sample Pct.	Whole Pct.	Variable	Sample Pct.	Whole Pct.
HeartDisease	8.925%	8.56%	AgeCategory4	5.975%	6.426%
Smoking	41.55%	41.25%	AgeCategory5	6.325%	6.569%
Stroke	3.5%	3.774%	AgeCategory6	6.55%	6.814%
DiffWalking	14%	13.89%	AgeCategory7	7.625%	7.937%
SexMale	46.98%	47.53%	AgeCategory8	9.225%	9.305%
Diabetic	13.68%	12.16%	AgeCategory9	10.42%	10.53%
KidneyDisease	4.175%	3.683%	AgeCategory10	10.5%	10.68%
SkinCaner	9.4%	9.324%	AgeCategory11	9.7%	9.714%
AgeCategory2	5.475%	5.302%	AgeCategory12	6.875%	6.717%
AgeCategory3	6.2%	5.864%	AgeCategory13	7.9%	7.553%

3 Modeling Approach

3.1 Full Model

Since we did not obtain any informative prior by reviewing the current literature, we decided to use default prior adjusted by stan_glm() to fit a Bayesian logistic regression model on HeartDis-

ease with all 10 selected predictors. To express our model, we define $Y = \begin{cases} 1 & Has\ heart\ disease \\ 0 & Does\ not\ have\ heart\ disease \end{cases}$, and X_d as the design matrix of the 10 predictors. We further define $p(x) = P[Y = 1|X = X_d]$, so that our logistic regression model is defined as

$$log(\frac{p(x_i)}{1-p(x_i)})|\mu_i, \sigma \overset{ind}{\sim} Normal(\mu_i, \sigma), i = 1, \ldots, n, \tag{1}$$

where $\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{i,21} X_{i,21}$.

The default prior in **rstanarm** tends to be weakly informative, which assumes independence among prior distributions of coefficients and the standard deviation. That is,

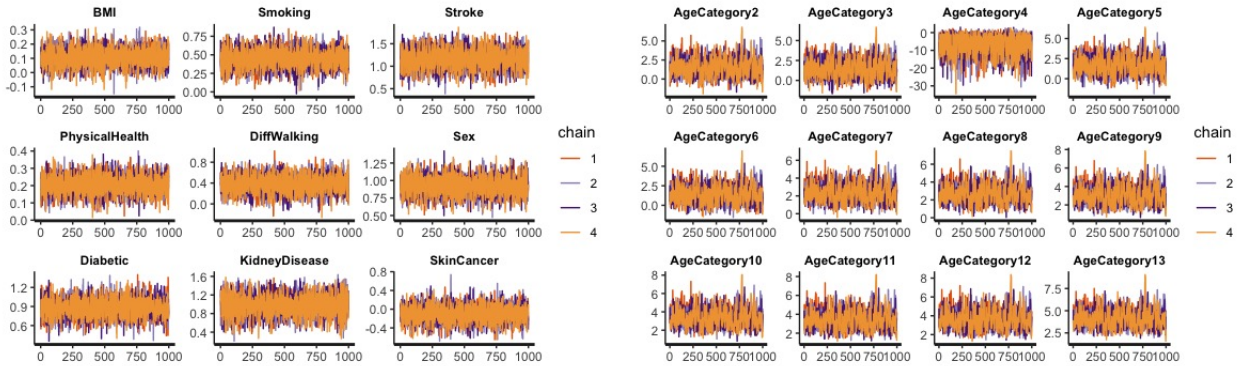
$$\pi(\beta_0, \beta_1, \ldots, \beta_{21}, \sigma) = \pi(\beta_0, \beta_1, \ldots, \beta_{21})\pi(\sigma); \tag{2}$$

$$\pi(\beta_0, \beta_1, \ldots, \beta_{21})\pi(\sigma) = \prod_{j=0}^{21} \pi(\beta_j),\ \beta_j \sim Normal(\mu_j, s_j). \tag{3}$$

Here, running prior_summary(), we found that rstanarm chose $\beta_j \sim Normal(0, 2.5)$ for us.

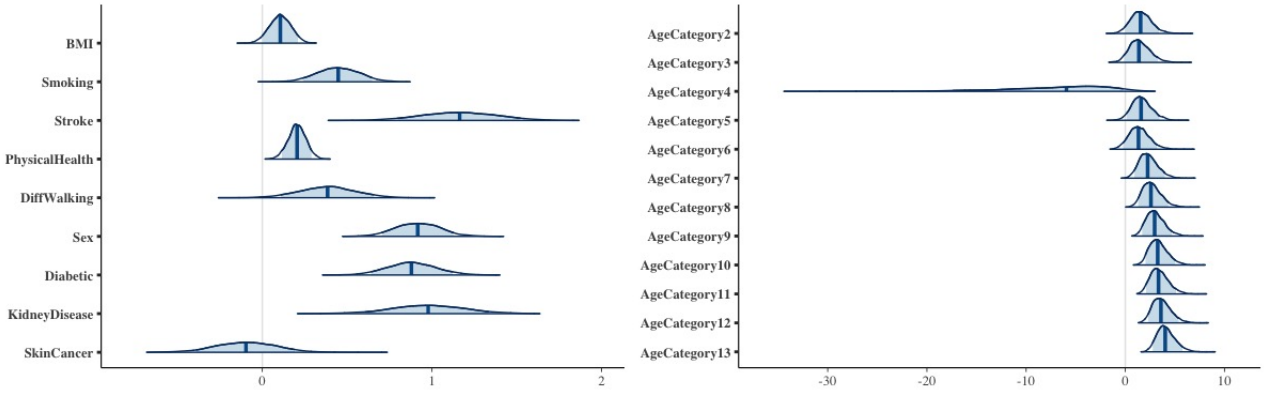
3.1.1 Diagnostics

We used 2000 samples for each of the 4 chains in our MCMC computation. From the trace-plots returned by Bayesian logistic regression, we found that all the chains converged to a common distribution for each coefficient and are in equilibrium. The MCMC diagnostics also showed that our model fit well, with $\hat{R}$ = 1.0 for each coefficient.



3.1.2 Inference

As can be seen from the distribution plots, all of the predictors that we used except SkinCancer are significant at 90% level, since their 90% credible intervals do not contain 0. Among the rest of predictors, Stroke, KidneyDisease, Sex, and Diabetic seem to be the strongest, while BMI, PhysicalHealth, and AgeCategories own a weaker power of explanation. Interestingly, AgeCategory4 appears to exert an opposite effect on the probability of heart disease compared to all other age categories, which is worth further investigation. Also, we can see a gradual shift of mean to the right from AgeCategory7 to AgeCategory13, indicating a possible increasing risk when people get older after 50. However, whether this trend is statistically significant should be tested in future work by estimating the difference of magnitude between coefficients.



We will give a few examples to interpret the estimates of coefficients in Table 3. $\hat{\beta}_0 = -6.510$ means that the expected probability of getting heart disease for a female aged 18-24 (the baseline Sex and AgeCategory) with average BMI, average reported PhysicalHealth value, no smoking habits, no history of stroke, diabetes, kidney disease or cancer, and no difficulty walking is $\frac{e^{p(x)}}{1+e^{p(x)}} = \frac{e^{-6.510}}{1+e^{-6.510}} = 0.0015$. $\hat{\beta}_1 = 0.106$ means that for a 1-standard-deviation increase in BMI, the odds of getting heart disease are expected to increase by a factor of $e^{0.106} = 1.11$. $\hat{\beta}_2 = 0.447$ means that compared to non-smokers, the odds of getting heart disease are expected to be 56.36% higher for smokers ($e^{0.447} - 1 = 0.5636$). The interpretation of other coefficient can thus be generalized by following β_1 for continuous ones or β_2 for categorical ones.

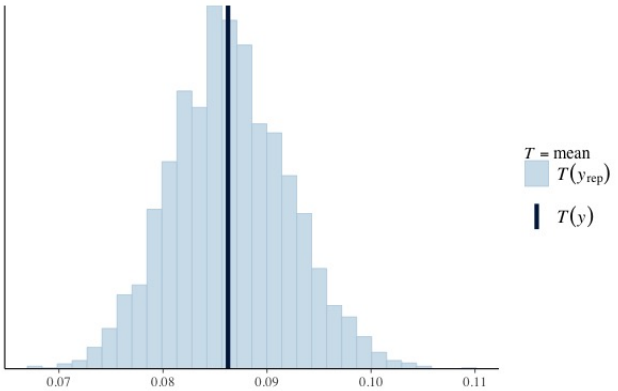
Table 3. Estimates of Coefficients

Variable	mean	sd	10%	50%	90%
(Intercept)	-6.510	0.922	-7.733	-6.414	-5.402
BMI	0.106	0.063	0.024	0.106	0.188
Smoking	0.447	0.124	0.288	0.447	0.605
Stroke	1.161	0.219	0.879	1.163	1.443
PhysicalHealth	0.205	0.052	0.137	0.205	0.271
DiffWalking	0.383	0.165	0.169	0.385	0.597
Sex	0.917	0.132	0.750	0.916	1.083
Diabetic	0.880	0.145	0.697	0.879	1.066
KidneyDisease	0.976	0.217	0.704	0.978	1.250
SkinCancer	-0.097	0.179	-0.322	-0.096	0.127
AgeCategory2	1.592	1.060	0.291	1.541	2.942
AgeCategory3	1.454	1.051	0.195	1.370	2.812
AgeCategory4	-7.008	5.362	-14.637	-5.933	-1.100
AgeCategory5	1.668	1.011	0.464	1.591	3.003
AgeCategory6	1.402	1.039	0.128	1.331	2.740
AgeCategory7	2.328	0.948	1.202	2.250	3.592
AgeCategory8	2.671	0.941	1.552	2.585	3.906
AgeCategory9	3.039	0.929	1.910	2.956	4.248
AgeCategory10	3.354	0.928	2.253	3.266	4.582
AgeCategory11	3.440	0.927	2.323	3.349	4.649
AgeCategory12	3.676	0.940	2.568	3.591	4.921
AgeCategory13	4.115	0.928	3.010	4.023	5.349

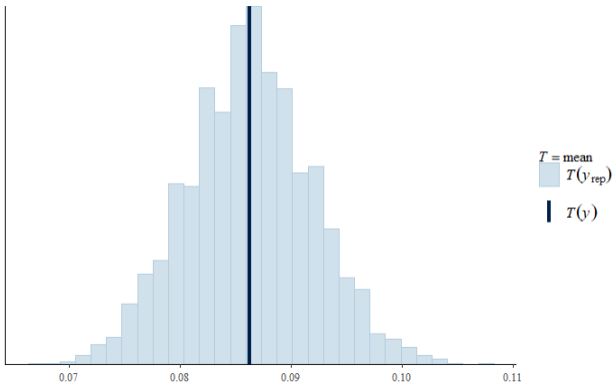
3.1.3 Posterior Predictive Checking

Finally, we ran a posterior predictive checking on this full model. The observed outcome variable is consistent with simulations of replicated data from this predictive distribution (the black

line lies around the mode of the distribution). Therefore, our model has a relatively good performance.



PPC for Full Model



PPC for Interaction Model

3.2 Interaction Model

From the relevant literature by Memorial Hermann Foundation, there is wide agreement that the sex and the age of patient is strongly related to their first heart attack. For males, the mean age of the first heart attack is 64.5. For females, the mean age of the first heart attack is 70.3. Thus, we come up with a new assumption that the different age categories might interact with the sex category. By applying the same default prior in the full model, we fit a new interaction model on the previous data, adding interaction terms between Sex and each AgeCategory. That is,

$$\begin{aligned} log(\frac{p(x_i)}{1-p(x_i)}) =& \beta_0 + \beta_1 X_{i,BMI} + \beta_{i,2} X_{i,Smoking} + \beta_{i,3} X_{i,Stroke} + \beta_{i,4} X_{i,PhysicalHealth} + \\ & \beta_{i,5} X_{i,DiffWalking} + \beta_{i,6} X_{i,Diabetic} + \beta_{i,7} X_{i,KidneyDisease} + \beta_{i,8} X_{i,SkinCaner} + \\ & \beta_{i,9} X_{i,Sex} + \sum_{j=2}^{13} (\beta_{j+8} X_{i,AgeCategory_j} + \beta_{j+20} X_{i,Sex} : X_{i,AgeCategory_j}) \end{aligned}$$

3.2.1 Diagnostics

We used 2000 samples for each of the 4 chains in our MCMC computation. From the traceplots returned by Bayesian logistic regression, we found that all the chains converged to a common distribution for each coefficient and are in equilibrium, even for the new coefficents of the interaction terms. The MCMC diagnostics showed that our model still fits the data, with $\hat{R}$ = 1.0 for each

coefficient. For the sake of space, we will display the traceplots which are very similar to previous ones in the Appendix.

3.2.2 Inference

In the plots (from the Appendix) of the 90% intervals for each coefficient in the interaction model, we can see that because of the interaction terms, the effects of Sex on the model become insignificant. The coefficients of other predictors exhibit similiar behavior as in the full model. For the new interaction terms of Sex and AgeCategory, they behave like the AgeCategory terms but all of the 90% confidence interval include zero, which suggests that they are all statistically insignificant.

The estimates of coefficients are presented in Table 4 in the Appendix, and we will simply give an example on how to interpret the coefficient for the interaction term here. $\beta_{22} = -1.000$ (Sex:AgeCategory2) indicates that for a male with average BMI and average physical health and an age around 25-29 who does not have a history of diabetes, kidney disease, stroke, skin cancer, smoking, and no difficulty walking, the odds of getting heart disease are expected to be $\frac{e^{-6.637+0.9+2.243-1.000}}{e^{-6.637+0.9}} = 3.466$ times of the odds for a male with the same physical condition but with an age around 18-24. To compare males and females with the same age, the odds ratio between a male with no history of diabetes, kidney disease, stroke, skin cancer, smoking, and no difficulty walking and a female with the same physical condition is $\frac{e^{-6.637+0.9+2.243-1.000}}{e^{-6.637+2.243}} = 0.9048$. Thus, the chance of getting the first attack heart attack on males with age 25-29 is 90.48% of the chances of the same aged females, with the same physical condition.

3.2.3 Posterior Predictive Checking

The posterior predictive checking graph is presented above together with the full model. The observed outcome variable is still consistent with simulations of replicated data from this predictive distribution. Therefore, the interaction model still possesses a relatively high predictive power. To compare two models, since two models are nested, we applied leave-one-out cross-validation (LOO, LOOIC) in **rstanarm** as the metric.

Since the out-of-sample predictive fit is estimated by Bayesian leave-one-out cross-validation,

we use the elpd (expected log point-wise predictive density) to measure the prediction accuracy. Given no difference between the full model and itself, the first row is zero. And the Interaction model has smaller elpd, which indicates larger LOOIC. Therefore, the interaction model does not perform better than the full model.

Table 5. Comparison of ELPD of both models

Model	elpd diff	se diff
Full model	0.0	0.0
Interaction model	-7.1	4.5

4 Discussion

Overall, both our proposed models generated a good performance in terms of convergence and posterior distributive checking. A high value of BMI, having the habit of smoking, and a history of other diseases (including stroke, diabetes, kidney disease) are positively correlated with the risk of heart disease. Also, people who are 18-24 years old or 35-39 old seem to have the lowest odds of getting heart disease, while people in other age groups being more vulnerable. However, we are not clear of the relative risk among other groups themselves, since we did not set each of them as baseline groups.

Unfortunately, based on this single dataset we obtained, we did not find the expected interaction between Sex and AgeCategory, given that all the interaction terms are statistically insignificant and that the Interaction Model does not have a lower LOOIC. This is probably due to selection bias during the collection of data, since respondents voluntarily participated in the questionnaire. It can also be attributed to the fact that this study is cross-sectional instead of longitudinal in nature, which does not focus on a single person at different periods. Therefore, the interaction effect involving age might be masked in this dataset. Future work, ideally a longitudinal experiment, should be dedicated before arriving a conclusion about the proposed interaction effect.

References

"Heart Disease Facts." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 7 Feb. 2022, https://www.cdc.gov/heartdisease/facts.htm.

Pytlak, Kamil. "Personal Key Indicators of Heart Disease." Kaggle, 16 Feb. 2022, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

"Five Facts About Heart Disease to Live By" Memorial Hermann Foundation, Memorial Hermann Foundation, https://memorialhermann.org/services/specialties/heart-and-vascular/healthy-living/education/heart-disease-and-age.

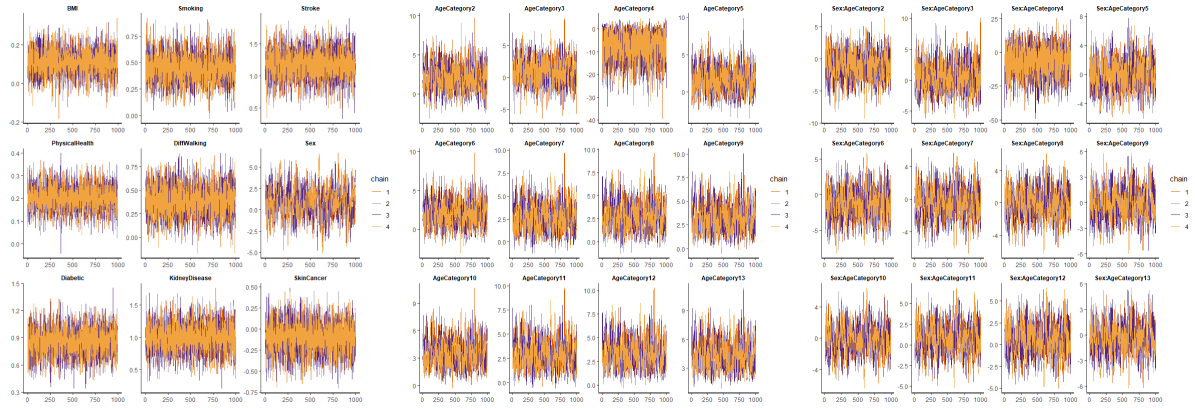
Appendix

Table 4. Estimates of Coefficients

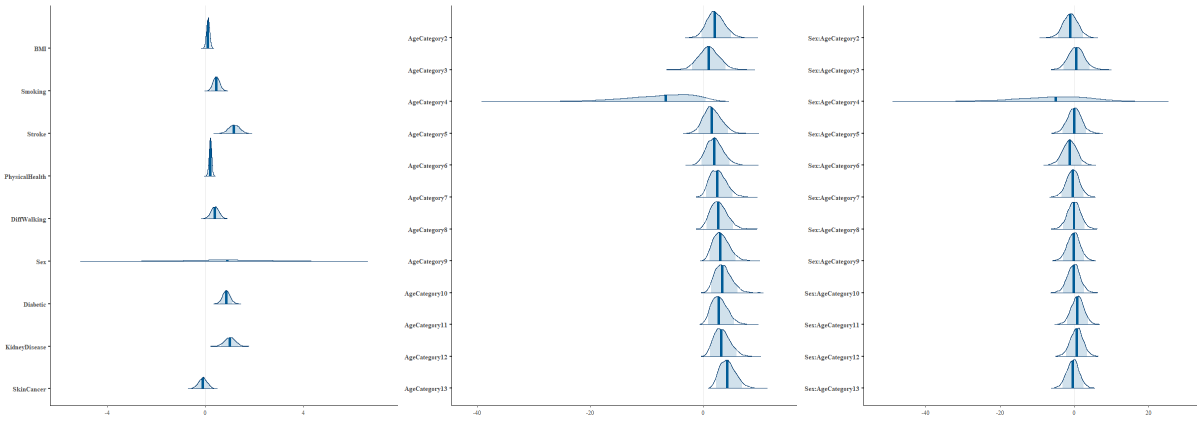
Variable	mean	sd	10%	50%	90%
(Intercept)	-6.637	1.409	-8.523	-6.518	-4.927
BMI	0.110	0.064	0.026	0.111	0.192
Smoking	0.448	0.134	0.280	0.446	0.619
Stroke	1.175	0.221	0.896	1.175	1.457
PhysicalHealth	0.204	0.052	0.137	0.205	0.268
DiffWalking	0.390	0.159	0.189	0.390	0.598
Sex	0.900	1.691	-1.197	0.891	3.134
Diabetic	0.853	0.148	0.666	0.851	1.042
KidneyDisease	1.002	0.213	0.724	1.003	1.275
SkinCancer	-0.104	0.178	-0.333	-0.101	0.127
AgeCategory2	2.243	1.595	0.304	2.141	4.332
AgeCategory3	1.015	1.852	-1.279	1.009	3.367
AgeCategory4	-7.741	6.277	-16.455	-6.617	-0.565
AgeCategory5	1.675	1.584	-0.234	1.568	3.775

(Table 4. Continued)

Variable	mean	sd	10%	50%	90%
AgeCategory6	2.092	1.530	0.198	2.011	4.094
AgeCategory7	2.682	1.465	0.912	2.578	4.619
AgeCategory8	2.824	1.463	1.043	2.720	4.786
AgeCategory9	3.247	1.430	1.519	3.108	5.141
AgeCategory10	3.580	1.435	1.850	3.461	5.510
AgeCategory11	2.960	1.432	1.228	2.833	4.834
AgeCategory12	3.409	1.441	1.658	3.293	5.379
AgeCategory13	4.413	1.428	2.676	4.288	6.319
Sex:AgeCategory2	-1.000	2.027	-3.507	-0.988	1.538
Sex:AgeCategory3	0.589	2.161	-2.107	0.551	3.317
Sex:AgeCategory4	-5.966	10.513	-20.227	-4.934	6.557
Sex:AgeCategory5	0.045	1.917	-2.424	0.054	2.428
Sex:AgeCategory6	-1.217	1.965	-3.707	-1.206	1.221
Sex:AgeCategory7	-0.387	1.777	-2.604	-0.376	1.826
Sex:AgeCategory8	-0.032	1.757	-2.221	-0.027	2.190
Sex:AgeCategory9	-0.140	1.727	-2.300	-0.114	2.045
Sex:AgeCategory10	-0.141	1.727	-2.311	-0.126	2.033
Sex:AgeCategory11	0.879	1.724	-1.287	0.886	3.063
Sex:AgeCategory12	0.654	1.735	-1.522	0.661	2.807
Sex:AgeCategory13	-0.351	1.715	-2.539	-0.314	1.808



Traceplots for the Coefficients in the Interaction Model



Distributions of the Coefficients in the Interaction Model