Predicting Stroke Likelihood Using Bayesian Logistic Regression

DS 6040: Bayesian Machine Learning

Hyun Ko, Griffin McCauley, Eric Tria

I. Problem Description

According to the World Health Organization (2021), stroke is the second leading cause of death in the world and third in terms of causing disability. It is a common occurrence as one out of every four people are in danger of having a stroke in their lifetime. There are both lifestyle and medical risk factors for stroke. The lifestyle risk factors include obesity, physical inactivity, tobacco use, and alcohol abuse. The medical risk factors include high blood pressure, high cholesterol, diabetes and a history of stroke or heart attack.

Boehme et al. (2018) discusses that these risk factors for stroke can also be identified as modifiable or non-modifiable. The modifiable risk factors are characteristics that can be directly addressed such as diet or comorbid conditions. Non-modifiable risks factors are things such as age or gender that cannot be directly addressed. With these identifiable factors available, it makes it possible to estimate the risk of having a stroke as a part of primary care. These risk estimations are useful so that preventive procedures and actions can be used to hopefully lessen the percentage of people suffering from strokes. There does, however, exist uncertainty in diagnosing strokes. Fallik (2017) notes that misdiagnosis of stroke, especially for those with milder symptoms, is an observed occurrence in emergency departments with almost 10 percent of stroke patients being initially misdiagnosed as of 2017. Stroke is also noted as one of the most common misdiagnoses among major diagnosis errors, which highlights the need for methods that can assist in diagnosis.

This analysis makes use of a Kaggle dataset with over 5000 observations and various clinical features that can be used for predicting the likelihood of a patient getting a stroke. The available parameters include continuous health metrics namely average glucose level and body mass index. The data also includes information on whether a patient has other prior health issues such as hypertension, heart disease, and smoking status. Rounding out the features are those that are more focused on the demographics of a patient including gender, age, marital status, and employment status. This project aims to identify which of these features are useful in building a model that effectively predicts stroke likelihood using various Bayesian methods and techniques.

II. Probability Model

From the outset, the goal of attempting to identify whether a given individual is likely to have a stroke or not naturally presents itself as a classification problem. Unlike some classification settings, however, which might require multilevel outputs for a variety of group memberships, since the target variable in the dataset we are working with is binary and simply indicates whether that patient has suffered a stroke or not, our situation would specifically be a binary classification problem with only two outcomes. To determine the most suitable probability model to address this problem of binary stroke classification, we considered several approaches including quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), Naïve Bayes, and support vector machines (SVM) but ultimately elected to pursue this analysis using Bayesian logistic regression.

By applying Bayesian logistic regression rather than one of the other techniques, this allowed us to infer the relationship between the predictor variables in the dataset and the resulting stroke likelihood more directly while also providing us the ability to explicitly measure and account for the uncertainty and variance both in our posterior parameter estimates and in the model's stroke predictions. This probability model can formally be expressed in the following linear form: $\log\left(\frac{p}{1-p}\right) = a + X \cdot b$ where p is the probability of the binary outcome occurring, a is the model intercept, b is the vector of slope parameters, and X is the vector of features for an observation. For a graphical representation of this model, see Figure 1.

Through implementing an HMC sampling approach, we can derive full posterior distribution estimates for each of the parameters of this model, and, thus, determine both their maximum a posteriori values and their associated uncertainties. After performing these operations for both a full and reduced set of observation features, we will also be able to produce a final ensemble model through pseudo-Bayesian model averaging which weights the predictions of each of the base learners according to their widely applicable information criterion (WAIC) scores. These weights are computed by applying a soft-max activation across the WAIC scores of all base learners in the ensemble. Through the use of these Bayesian logistic regression probability models and the aggregating, Bayesian model averaging ensemble, we will be afforded the opportunity to perform for a meaningful analysis of both the predictive performance and the uncertainty of stroke probabilities generated by our methods.

III. Approach

With the conceptual model for our analysis established, we were able to advance to the stage of executing our ideas in code. The first step of this process was to properly load and clean our data from Kaggle. The raw data was originally comprised of eleven predictor variables and one binary, target variable, but some of these features needed to be removed or refactored to support our model meaningfully. Features such as *id*, *work_type*, and *smoking_status* were dropped since *id* was simply a unique patient identifier that should not be statistically significant in any regression and because *work_type* and *smoking_status* were both multi-class categorical variables which induced excessive sparsity in our data frame when we attempted to one-hot encode them. (Note that the multi-class categorical predictors were not initially removed when we embarked on this analysis, but, once we realized that they were causing sampling convergence issues due to the sparsity created by one-hot encoding them, we decided to remove them even before fitting the full parameter model since their excessive variance and statistical insignificance were dramatically impairing the predictive performance of the model.) We also converted the three two-level categorical predictors of *gender*, *ever_married*, and *Residence_type* into binary variables and removed the 201 observations which contained NA values. These modifications resulted in a final data matrix containing three numerical predictors and five binary variables which we proceeded to partition into training and validation sets, using an 80-20 split.

Having now configured the data as desired, we began designing and constructing the framework for the Bayesian logistic regression model using the pymc package in Python. Since the parameters are treated as random variables in a Bayesian context, we had to assign appropriate priors to them, and, in order to make these priors uninformative, we chose to apply Gaussian priors with mean zero and standard deviation 100 to the intercept and slope parameters, respectively. These parameters along with the observed tuples of the predictor variables can then be related to the expectation of the response variable via the logistic regression equation provided in the previous section. Combined, these parameter priors and the probability model definition yield the final graphical representation depicted in Figure 1.

Once the model object was built, it was time to perform the posterior sampling using Hamiltonian Monte Carlo (HMC) methods. Even with only eight predictor variables and 3926 observations, this process of generating 10000 samples still took an exorbitant amount of time to run, but, thankfully, we were able to set a random seed and export the results of the sampling trace in order to ensure consistency and reproducibility going forward. Despite the extreme duration of the sampling, upon observing the associated trace plots, we were able to conclude that the process did appear to demonstrate strong convergence for all the parameters and that none of the predictors were producing excessive variance in the posterior distribution estimates (Figure 2). Although there did not appear to be any convergence related issues present in the process, we still needed to assess the statistical significance and value of the parameter estimates through investigating the forest plots and summary table for the trace. These two graphics allowed us to determine that most of the predictor variables' parameters had converged to values which were not statistically significantly difference from zero according to their 94% highest density intervals (HDIs) (Figure 3, 4). Since the only two parameters whose values were statistically significant corresponded to the features of *age* and *hypertension*, these were the ones we retained when assembling our reduced model.

The reduced model now only had three remaining parameters which we needed to be approximated via HMC sampling (Figure 5). Contracting the full set of eight predictors down to two greatly reduced the complexity of the model and significantly diminished the time required to collect the 10000 samples, and, as could be seen from the trace plots, forest plots, and summary table, the posterior distribution estimates for the intercept and slope terms demonstrated good convergence properties, had reasonably sized spreads and variances, and all produced statistically significant values (Figure 6, 7, 8). Having successfully generated sampling traces for both models of interest, we could then compute and compare the models' WAIC scores and determine how the two models' predictions should be properly weighted in the final, pseudo-Bayesian model averaging ensemble. Based on the output of the WAIC comparison function, it turned out that, despite the full model having six additional features that seemed to primarily add noise to the model without providing any statistically significant relationships on their own, the full model was still slightly preferable to the reduced model according to their WAIC scores (Figure 9). Although we found the preference suggested by these WAIC weights to be slightly unintuitive at first, after considering the situation more thoroughly, we realized that being minutely in favor of the full model was not particularly surprising. Since we had dropped so many parameters all at once without fully considering potential instances of multi-collinearity or interaction effects, it was very reasonable to expect that the full model may have been able to occasionally represent some nuanced relationships and associations better than the reduced model. At this point, having acquired sampling traces and their associated WAIC-based weights for both the full and reduced models, we were able to finally assess the predictive performance of our models and attempt to gauge the certainty with which they produced their estimates.

IV. Results

Due to the huge class imbalance on our response variable, with 95% of our responses being negative and only 5% being positive, a naïve majority classifier model that diagnoses every patient as negative can already achieve 95% accuracy; but these results are meaningless in terms of their ability to help identify high-risk individuals. Thus, instead of adopting accuracy as a single factor explaining our model performance, we also considered AUC score, precision, and recall. Each of these results have been derived and calculated from confusion matrices.

Another notable point is that separate misclassification costs for false positives and false negatives have been adopted. In light of the gravity of diagnosis, the impact of misdiagnosing a stroke patient as negative is much more significant than misdiagnosing a healthy person as positive. Accordingly, we set the cost of a false negative 50 times higher than that of a false positive. In other words, the threshold of diagnosing someone as positive should be quite low, 0.04 for all models, so that a person misdiagnosed as having stroke might go through additional testing to verify illness state. This value of 0.04 was empirically determined to be the optimal threshold to minimize misclassification cost in our problem, but, in general, the theoretically optimal value would be approximately 0.02.

The full HMC sampling model has a total of eight predictors: five of them are categorical (*gender, hypertension, heart_disease, ever_married, residence_type*) and three of them are continuous (*age, avg_glucose_level, and bmi*). The threshold of diagnosing a person as positive is 0.04, and this results in 71.3% of people being correctly diagnosed as having stroke and produces a recall of 94.3%. Table 1 shows the result of our full HMC sampling model with a total of eight predictors. Since the initial purpose of our project was to minimize the misclassification cost, recall is maximized at the expense of precision and accuracy. Our models prioritize detecting a sick person correctly rather than being conservative regarding misdiagnosing a healthy person. Figure 10 shows the odds ratio of a person having a stroke based on each single predictor. From the plot, we can see that *age, hypertension,* and *heart_disease* seem to have a positive association with the response variable.

The reduced HMC sampling model with only the two predictors *age* and *hypertension* also used the same threshold of 0.04 and was able to correctly diagnose 68.6% of the observations in test data and again yield a recall of 94.3%. Table 1 shows the results of the reduced sampling model, and every metric shows slightly worse performance results compared to the full sampling model. Figure 11 shows the odds ratio of a person having stroke based on each individual predictor, and, from the plot, *age and hypertension* seem to have a similar range compared to the full sampling ones.

As mentioned, the weights for Bayesian model averaging are calculated based on the WAIC scores, and in this case, 0.56 is applied to the full sampling model and 0.44 to the reduced sampling model. It is known that model averaging generally improves prediction results and reduces variance. According to Table 3, the performance of the BMA model is in between the full sampling one and the reduced sampling one as expected.

In order to check the variability embedded in each model, graphs of the posterior predictions from the entire sampling trace are displayed as well. Since the reduced model includes only two predictors, *age* and *hypertension*, and *hypertension* is a categorical variable, we can only visually inspect the predictive variability as a function of *age*. To take *hypertension* into account, separate plots have been produced controlling for having hypertension or not. For both graphs, *age* is on the x-axis and the probability of having a stroke is on the y-axis. Figure 12 displays the probability of having a stroke across *age* without *hypertension* for all three models. For all of these models, both the probability of having a stroke and variability of the prediction value tend to increase as *age* increases. It is unsurprising that as we *age*, we are more likely susceptible to having a stroke.

Among all models, the full sampling model has the largest variability across all ages. Specifically, for those around the age of 80, the probability ranges from about 0.07 to 0.35, which is quite a large spread. Even though the full sampling model has eight predictors, many of the predictors are deemed to be statistically insignificant based on the 94% HDI as shown in the summary table. Hence, additional noise is present in the full sampling model. However, the reduced sampling model has the least variability among all models because it has only two statistically significant predictors with reduced noise. The range of predicted stroke probabilities for a person around the age of 80 is between 0.17 and 0.3 which is tighter than that of the full sampling model. The Bayesian averaging model successfully deals with the huge variability in full sampling model and has comparatively smaller ranges of variability across all ages.

For observations with *hypertension*, the probability of having a stroke for all models increases. Previously, the y-axis ranged up to 0.4, but now it ranges up to 0.5. This is unsurprising since *hypertension* and *stroke* have a strong association: a patient who has *hypertension* tends to have a higher probability of having a stroke than a patient who does not have *hypertension*. Like the previous result, the full sampling model had the highest variability, followed by the Bayesian averaging model, and the reduced model last. Finally, Figure 14 displays the overlapping graphs of all models by hypertension, and Figure 15 shows the mean predictions for each model.

To summarize, according to all the models, *age* and *hypertension* are the key factors for predicting a stroke. Classification-wise, the full HMC sampling model with eight predictors provides the best results and does an excellent job of correctly diagnosing stroke patients. From the Bayesian perspective, however, the Bayesian averaging model is superior to the rest in terms of dealing with the uncertainties embedded in our data while also producing high recall and comparable accuracy, which this model the optimal choice.

V. Conclusions

After conducting Bayesian analysis on the given dataset, the group was able to develop a logistic regression model that can provide decent predictions for stroke likelihood. Given the available features in the dataset, we were able to see which features would lead to less variability in the proposed models through graphs of the posterior. The comparison of the full and reduced models presented interesting results for the analysis. Although the full model was ranked slightly higher in terms of WAIC ranking, the full model also displayed more variability in the response due to statistically insignificant predictors adding noise to the model's predictions. Bayesian model averaging was used to build an ensemble model from the full and reduced ones with the goal of improving accuracy, reducing variance, and accounting for uncertainty in the choice of model. From the results, we can see that the ensemble model had better overall performance than the reduced one and less variability than the full one.

This analysis was able to answer the initial problem presented. Given that stroke is one of the most common misdiagnoses, we can say that the predictive ability of the generated model is good. There are, however, limitations to this analysis. The dataset used for training consisted of roughly 5000 observations, which is a decent amount for creating an initial model but would be lacking for a model that would be ready for deployment. The dataset also only consisted of 11 possible predictors to be used, whereas different studies such as that from Boehme et. Al. (2017) show that there are more factors that can be attributed to causing stroke. This analysis also does not consider the difference in terms of regions which The GBD 2016 Lifetime Risk of Stroke Collaborators (2018) notes as an observed occurrence. These are some of the improvements that can be built on top of the analysis our group conducted in order to provide a predictive model that can be deployed to assist in preventive treatment for stroke, one of the leading causes of death in the world.

VI. References

Boehme, A., Esenwa, C., Elkind, M., (2017, February 3). Stroke Risk Factors, Genetics, and Prevention. National Library of Medicine. Retrieved December 6, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5321635/

Fallik, D., (2017, April 6). Nearly 10 Percent of Strokes Are Misdiagnosed (At Least Initially) in the Emergency Department. Neurology Today. Retrieved December 6, 2022, from https://journals.lww.com/neurotodayonline/fulltext/2017/04060/Nearly_10_Percent_of_Strokes_Are_Misdiagnosed__At.8.aspx

Singh, P. (2021, October 28). World Stroke Day. World Health Organization. Retrieved December 6, 2022, from https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

Soriano, F., (2021, January 26). Stroke Prediction Dataset. Kaggle. Retrieved December 6, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

The GBD 2016 Lifetime Risk of Stroke Collaborators, (2018, December 20). Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016. The New England Journal of Medicine. Retrieved December 7, 2022, from https://www.nejm.org/doi/full/10.1056/NEJMoa1804492

VII. Appendix

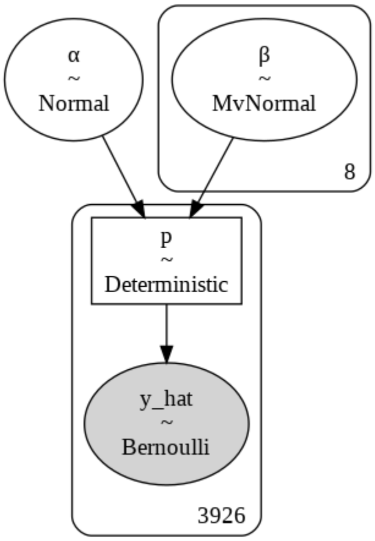


Figure 1

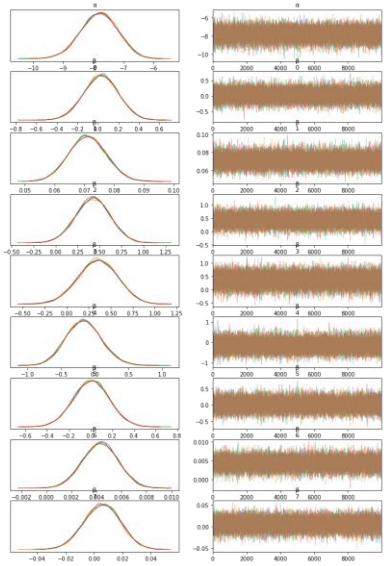


Figure 2

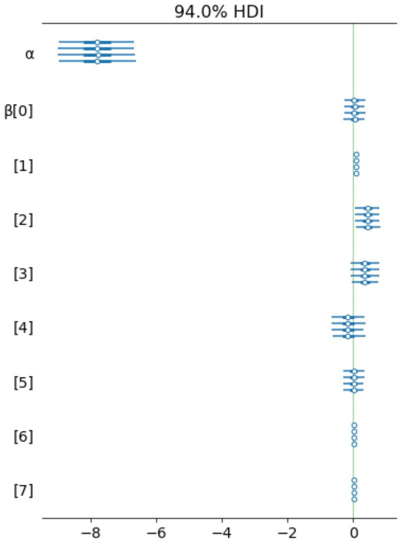
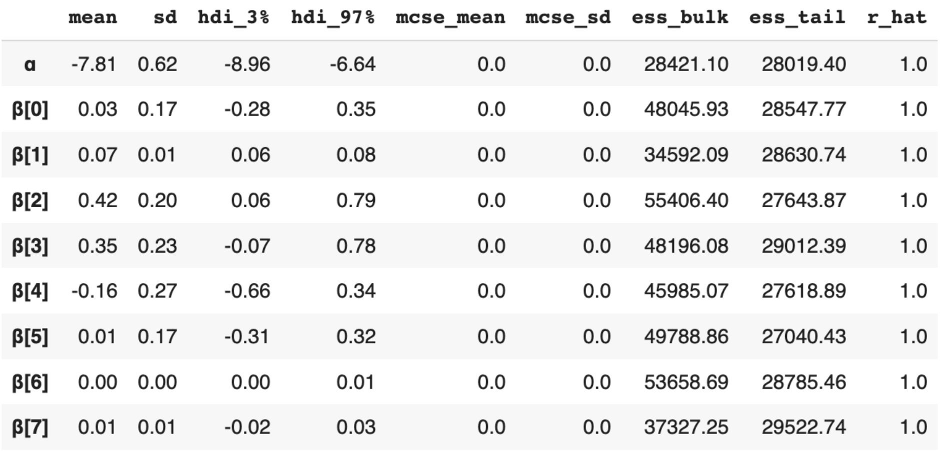


Figure 3



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	-7.81	0.62	-8.96	-6.64	0.0	0.0	28421.10	28019.40	1.0
β[0]	0.03	0.17	-0.28	0.35	0.0	0.0	48045.93	28547.77	1.0
β[1]	0.07	0.01	0.06	0.08	0.0	0.0	34592.09	28630.74	1.0
β[2]	0.42	0.20	0.06	0.79	0.0	0.0	55406.40	27643.87	1.0
β[3]	0.35	0.23	-0.07	0.78	0.0	0.0	48196.08	29012.39	1.0
β[4]	-0.16	0.27	-0.66	0.34	0.0	0.0	45985.07	27618.89	1.0
β[5]	0.01	0.17	-0.31	0.32	0.0	0.0	49788.86	27040.43	1.0
β[6]	0.00	0.00	0.00	0.01	0.0	0.0	53658.69	28785.46	1.0
β[7]	0.01	0.01	-0.02	0.03	0.0	0.0	37327.25	29522.74	1.0

Figure 4

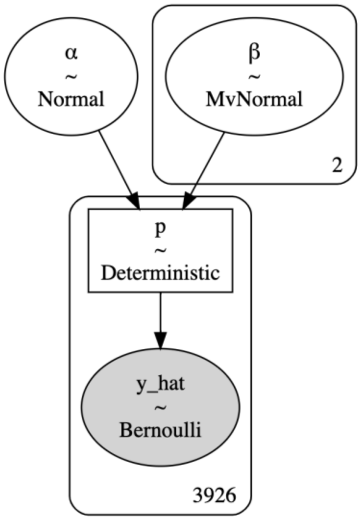


Figure 5

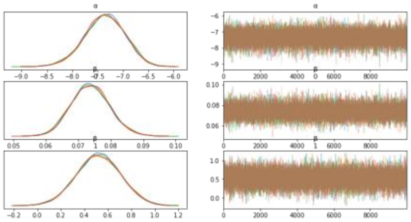


Figure 6

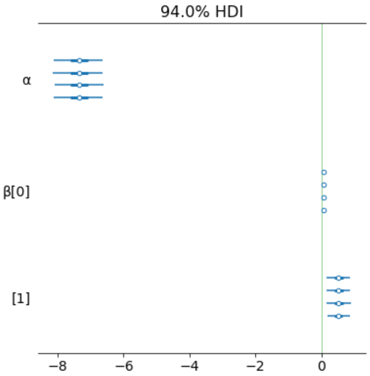
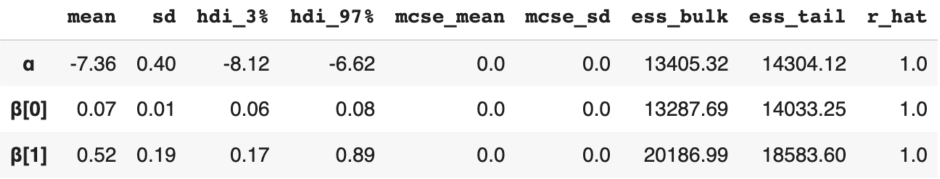
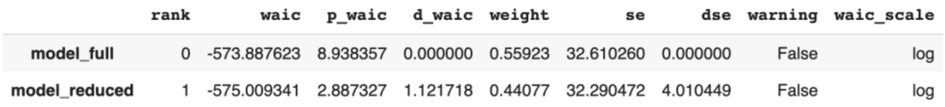


Figure 7



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	-7.36	0.40	-8.12	-6.62	0.0	0.0	13405.32	14304.12	1.0
β[0]	0.07	0.01	0.06	0.08	0.0	0.0	13287.69	14033.25	1.0
β[1]	0.52	0.19	0.17	0.89	0.0	0.0	20186.99	18583.60	1.0

Figure 8



	rank	waic	p_waic	d_waic	weight	se	dse	warning	waic_scale
model_full	0	-573.887623	8.938357	0.000000	0.55923	32.610260	0.000000	False	log
model_reduced	1	-575.009341	2.887327	1.121718	0.44077	32.290472	4.010449	False	log

Figure 9

	AUC	Confusion Matrix	Precision	Accuracy	Recall
Full Sampling	0.854	[[33, 2], [280, 667]]	0.105	0.713	0.943
Reduced Sampling	0.836	[[33, 2], [306, 641]]	0.097	0.686	0.943
BMA	0.847	[[33, 2], [292, 655]]	0.102	0.701	0.943

Table 1

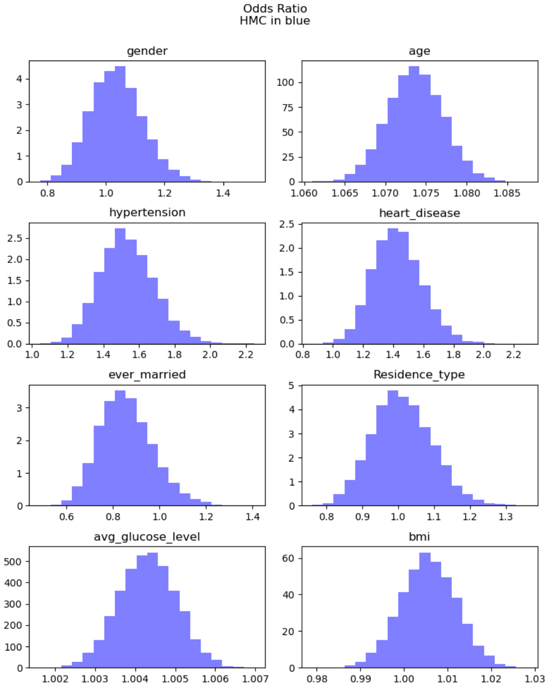


Figure 10

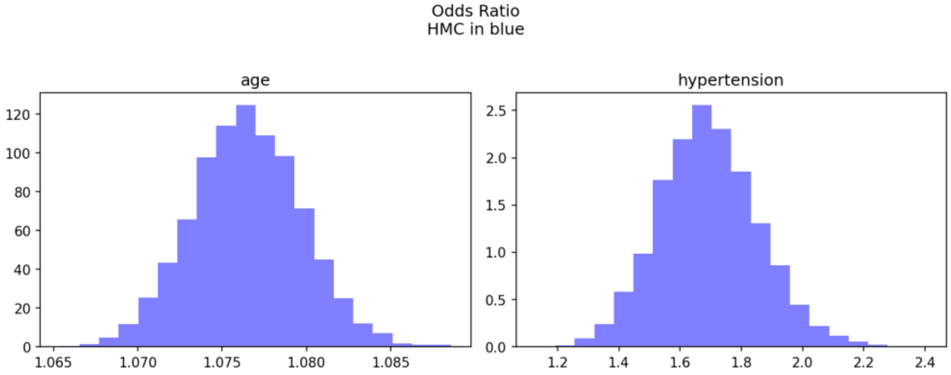


Figure 11

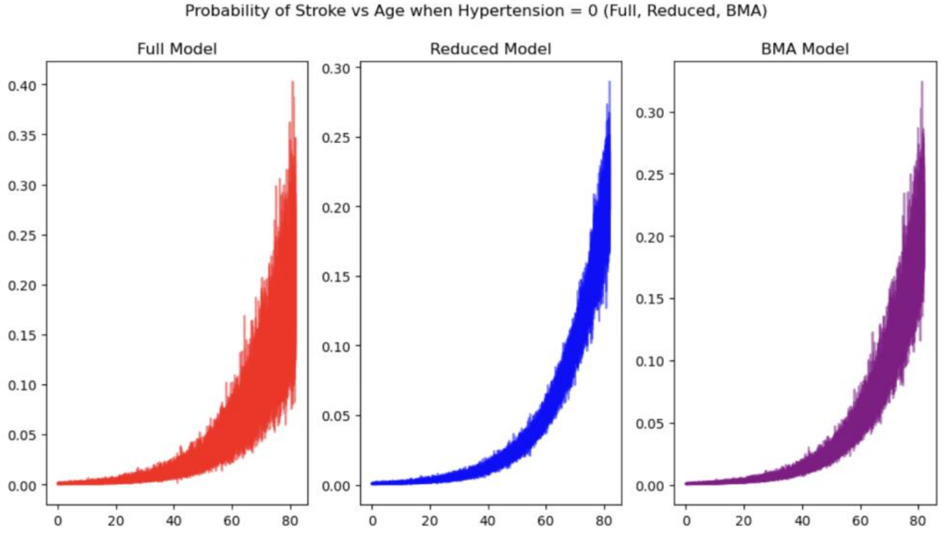


Figure 12

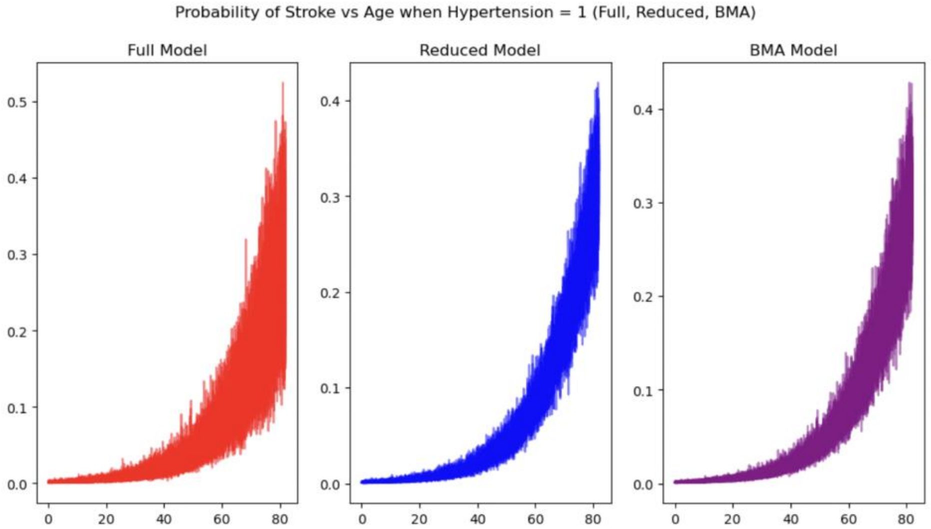


Figure 13

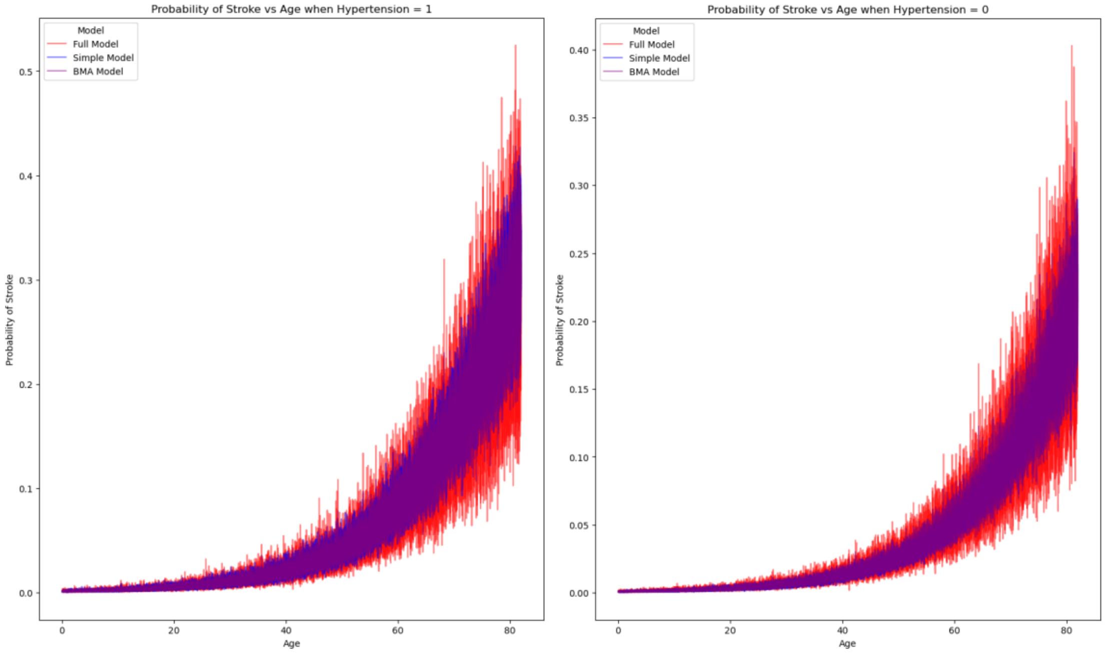


Figure 14

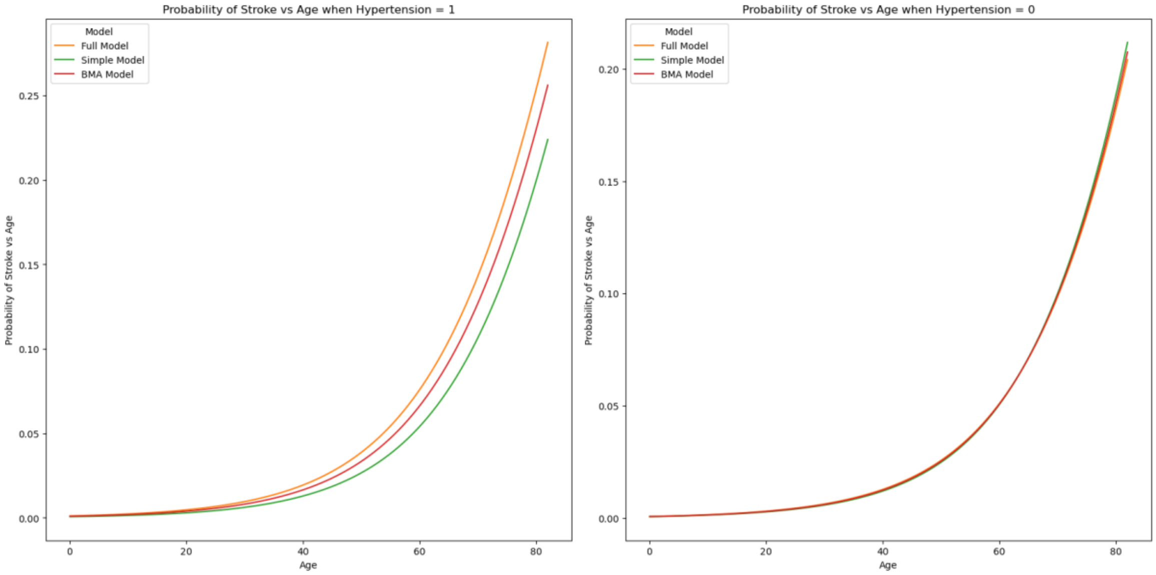

Probability of Stroke vs Age when Hypertension = 1
Model
Full Model
Simple Model
BMA Model
0.25
0.20
0.15
0.10
0.05
0.00
Probability of Stroke vs Age
0
20
40
60
80
Age
Probability of Stroke vs Age when Hypertension = 0
Model
Full Model
Simple Model
BMA Model
0.20
0.15
0.10
0.05
0.00
Probability of Stroke vs Age
0
20
40
60
80
Age

Figure 15