Liver Cirrhosis Survival Prediction

Liver cirrhosis is a life-threatening condition that affects millions of individuals worldwide, presenting a significant public health challenge. This chronic liver disease is characterized by the progressive scarring of the liver tissue, which ultimately leads to liver failure if left untreated. Among the critical aspects of liver cirrhosis is the unpredictability of patient survival, which depends on a myriad of factors, including disease severity, comorbidities, treatment regimens, and individual patient characteristics.

Problem Description

This study aims to employ Bayesian analysis to develop predictive models for assessing the survival probability of patients who were subjects of research for primary biliary cirrhosis (PBC) of the liver by the Mayo Clinic. Traditional statistical methods and machine learning approaches have been applied to address this issue, but Bayesian analysis offers a unique advantage by incorporating prior knowledge, updating beliefs, and providing probabilistic predictions.

The goals are as follows:

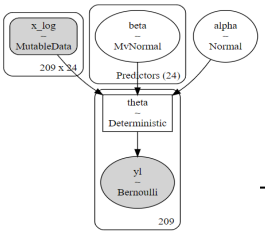
1. To predict the survivability estimates of patients with various medical data under purview.
2. To study the impact of stage and drug groups that affect the survivability estimates of patients.
3. To study how uncertainties inform survivability estimates.
4. To evaluate the best-performing model using accuracy and information criteria.

Data Description

The data was sourced from Kaggle. The 418 observations represented the 418 patients involved in the Mayo Clinic study. There were 20[1] columns in the dataset originally, two of which were dropped due to their lack of relevance to our analysis. These two columns were the ID and N_Days: the patient ID and the number of days between registration and determination of status, respectively. The response variable is the status, indicating the status of the patient. Out of the 17 predictor variables, eight are categorical and nine are continuous. The predictors, seen in Appendix _, consist of sensitive data (sex, age), medical test data (Ascites, Hepatomegaly, Cholesterol, etc.), and patient data (use of D-penicillamine, stage of disease). Data exploration found no significant correlation between the predictors, as seen in the heatmap in Appendix _. For this reason, all predictors were considered in model development. The response variable, status, was originally in three categories: C (Censored), CL (Censored due to transplant), and D (Death). The response was transformed into two categories: 0 (Death), and 1 (Survived). This transformation creates a binary outcome for the survivability of the patient as the impact of transplant on survivability is not investigated in this project.

The data was cleaned through several missing value imputations: the predictor mean was utilized for missing numerical values and the predictor mode was used for missing categorical values. Additionally, the age predictor was originally in the number of days. For easier interpretability, we converted age to be in years. Feature scaling was used for some models, this was performed after looking at the varying degrees of magnitude, ranges, and units. The scaled data was used for models that are sensitive to varying degrees of data. Categorical variables were dummy encoded, resulting in the data increasing from 17 columns to 24 columns. Finally, the data was split into a training and testing set to allow for out-of-sample predictions.

Probability Models



4 models were used to carry out the analysis:

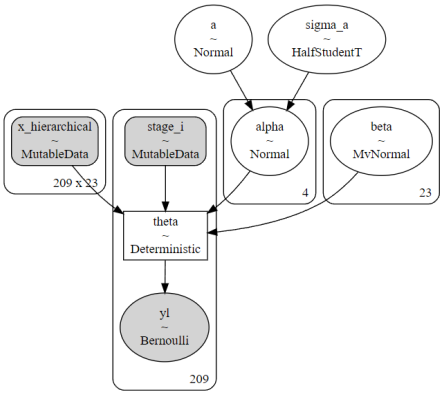
a) Bayesian Logistic Regression Model
Logistic regression has binary outcomes of the prediction - death or survival, which made it the most intuitive choice for modeling. Additionally, due to the

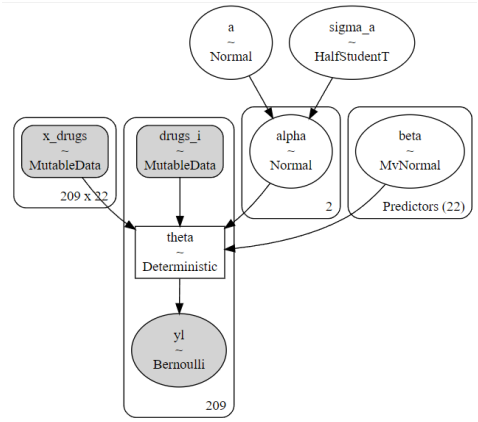
1 Appendix E

limited sample size, incorporating prior information provided an edge over finding more stable estimates. A multivariate Gaussian distribution for the coefficients of the parameters and a Bernoulli distribution for the likelihood in the model were used. A vector of zeros was used for the mean vector of the Multivariate Gaussian, indicating it is uninformed.

b) Hierarchical Model: Grouped By Stage

One of the 2 Hierarchical models we made was grouped by the stage of liver cirrhosis. The stage variable was categorical and had values 1,2,3, and 4, where 1 depicted the first stage of the disease and 4, the worst. We aimed to study the nested relationship between the stage variable and our predictions. A multivariate Gaussian distribution for the coefficients of the parameters and the Bernoulli distribution for the likelihood of this model were used. A vector of zeros was used for the mean vector of the Multivariate Gaussian, indicating it is uninformed.



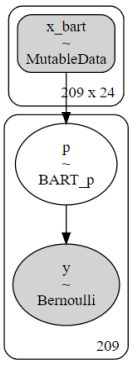


c) Hierarchical Model: Group By Drug

The other hierarchical model we built was grouped by drug usage to treat liver cirrhosis. This was a categorical variable as well that had 2 values - D-penicillamine and Placebo. We wanted to study if there was an impact of the consumption of the drug or placebo on the survivability estimates of the patient, therefore we chose a hierarchical model so that we could explore the effect of this partial pooling. For this model as well, a multivariate Gaussian distribution on the coefficients of the parameters and a Bernoulli likelihood were used. A vector of zeros was used for the mean vector of the Multivariate Gaussian, indicating it is uninformed.

d) BART

Another model that we used for Bayesian Analysis on this problem was BART. It seemed like a reasonable choice because it is effective in capturing the non-linear relationship between the predictors and the response variable. BART is also robust to outliers and noisy data and hence seemed promising. The package did not allow us to control a lot of parameters. The only thing we controlled in the BART model was the number of trees. As the number of trees increased, the model performance depleted, therefore 50 trees for this classification problem were parameterized.



Approach

Sampling

For the analysis, the posterior distributions of all the models using NUTS sampling were sampled. The trace plots and rank plots[2] indicated convergence for sampling chains. Great values for the shrinkage factor and the associated effective sample sizes were obtained, indicating good sampling. The influence of parameters was also checked with forest plots.[3]

Prior and Posterior Predictive Checks

All models excluding the BART model utilized uninformed priors, essentially because BART is a blackboxed model that does not allow for the user to add much information, resulting in high variance for the prior predictive checks.[4] The posterior predictive checks[5] showed significantly less variance and resulted in overlapping observations and posterior predictive means for all models. The prior predictive

2 Appendix A

3 Appendix F

4 Appendix B

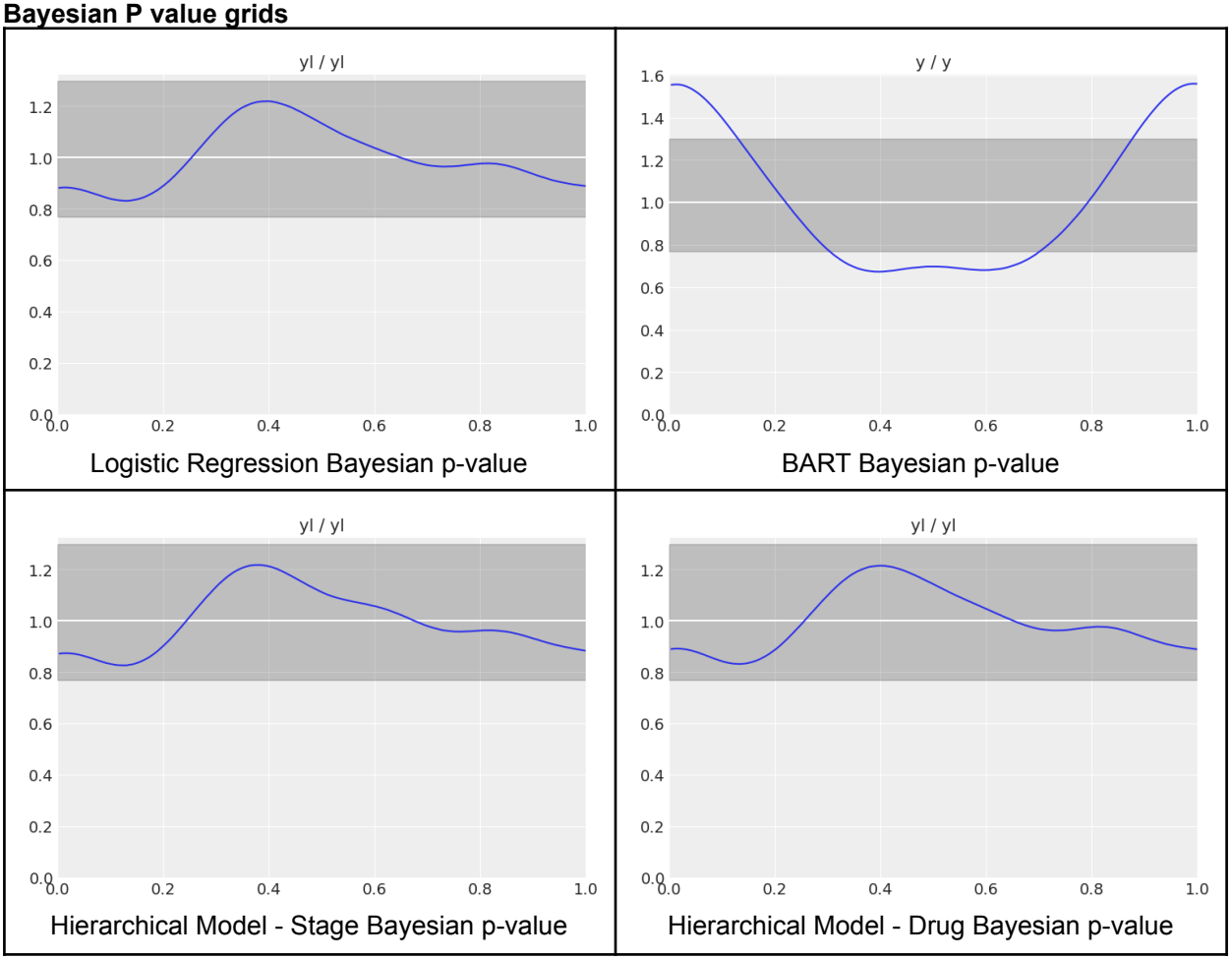
5 Appendix C

plots demonstrate our lack of information on the parameters. The posterior predictive plots indicate that the parameters are informed and the models closely fit the training data.

Model Assessment

We could assess from various metrics that the models perform well. Our Bayesian p-value plots indicate the good performance of all the models.

Bayesian P value grids



Model Evaluation

Evaluation of the performance of all our models against the WAIC and LOO was performed to identify the model that performs the best. Both information criteria indicate that the **BART Model** performs the best. The best model was selected, i.e. the BART Model, based on the criteria.

The model was also evaluated based on the accuracy scores and ROC[6] curves, which also indicated the same model for best performance. The BART model had the highest accuracy score out of all. However, if assessed through ROC curve only, both BART and Hierarchical Model-By Drugs have an equal value for AUC.

	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
model_bart	0	-92.599367	13.564613	0.000000	1.000000e+00	6.096528	0.000000	True	log
model_log	1	-115.050154	21.644978	22.450786	0.000000e+00	10.413621	6.629706	True	log
model_drugs	2	-115.461343	22.104171	22.861975	4.174439e-14	10.494310	6.730683	True	log
model_stage	3	-116.721670	23.148554	24.122303	0.000000e+00	10.511723	6.764628	True	log

6 Appendix D

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
model_bart	0	-92.747110	13.712355	0.000000	1.000000e+00	6.105953	0.000000	False	log
model_log	1	-115.660694	22.255518	22.913584	1.696421e-13	10.466597	6.694419	False	log
model_drugs	2	-116.044084	22.686912	23.296975	0.000000e+00	10.522932	6.768542	False	log
model_stage	3	-117.445190	23.872074	24.698080	2.310152e-12	10.583087	6.848904	True	log

Figure: WAIC and LOO Evaluation tables

Results

Uncertainties Informing Survivability Estimates

The posterior log odds for the best-performing model, BART, were assessed, and it was observed that the age variable has a statistically significant influence over the odds of survival. That could contextually be translated by a doctor to a patient, in a way that the doctor tells the patient to be more cautious of the disease if they are old. Similarly, the stage and use of the drug, D-penicillamine, impact the survival estimates of a patient. A doctor might make the patient cautious of the stage of cirrhosis, and while prescribing them the drug. However, it is also evident from the posterior log odds that the impact of a placebo is not statistically significant in informing the survivability estimates of a patient.

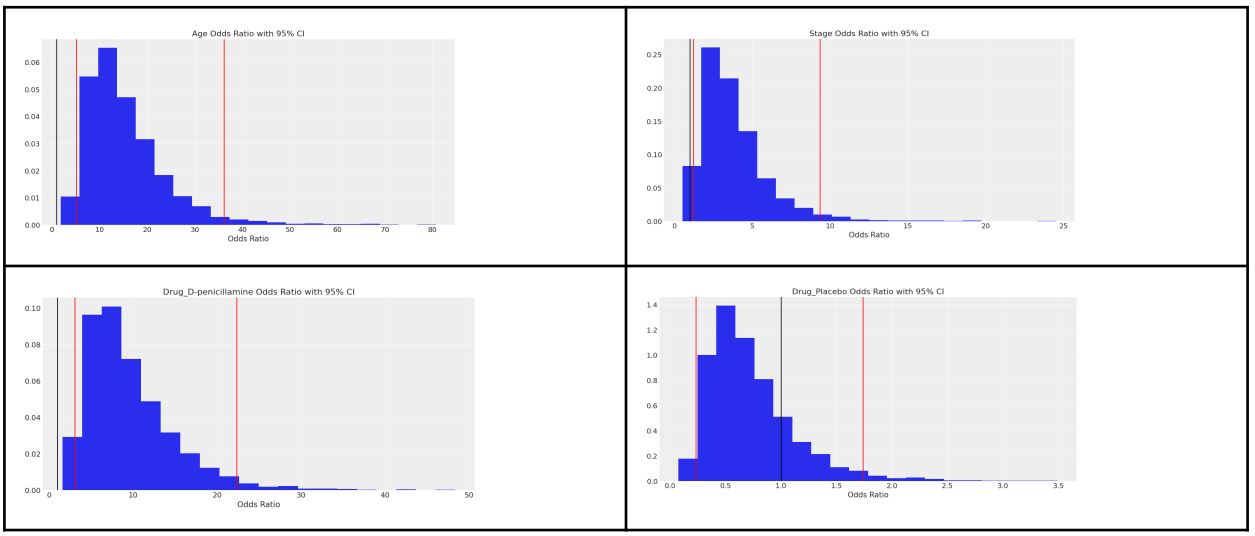


Table: Odds ratio plots for predictors of interest

Conclusions

The objectives of the study were met successfully. The hierarchical models provided valuable insights into the impact of liver cirrhosis stages and drug usage on survival predictions. The BART model, despite its simplicity in parameter control, demonstrated effectiveness in capturing non-linear relationships in the data, when assessed by the evaluation metrics which were Information Criteria - WAIC and LOO, accuracy score and ROC curves. The model that *performed the best* was the *BART model*. Additionally, through the use of odds ratio plots, we were able to demonstrate the uncertainty surrounding the survival estimates. It was also evident from the odds ratio that the stage and the use of drug against the placebo significantly impacted the survivability estimates.

Limitations

A common limitation of medical datasets is the few number of observations. The dataset for this study had only 418 patients as observations which made it difficult to build more robust and generalizable models. An extension of the same limitation was the limited assessment of uncertainty that could be carried out. The credible intervals for probabilities of survival were extensive and were hard to measure against any continuous variable due to the lack of them in the dataset. Furthermore, the most impactful predictors were categorical.

GitHub Repository:
https://github.com/sunidhigoyal05/ds6040-project.git

References

[1]Gelman, Andrew. "STS149A.pdf." Andrew Gelman's Research Papers, http://www.stat.columbia.edu/~gelman/research/published/STS149A.pdf.

[2] Martin Osvaldo A, Kumar Ravin; Lao Junpeng Bayesian Modeling and Computation in Python Boca Ratón, 2021. ISBN 978-0-367-89436-8

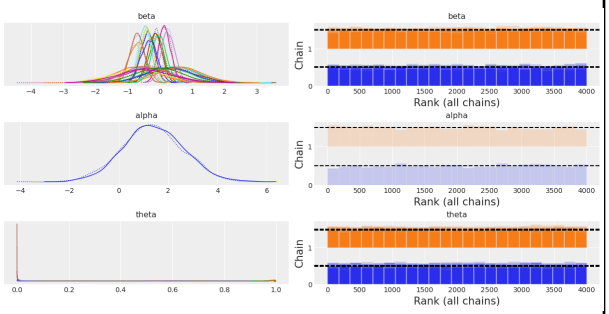
[3]Joe Beach Capital. "Cirrhosis Patient Survival Prediction." Kaggle, https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-prediction.

[4]PyMC Development Team. "API Reference." Bayesian Additive Regression Trees (BART) - PyMC3 Documentation, https://pymc.io/projects/bart/en/latest/api_reference.html.

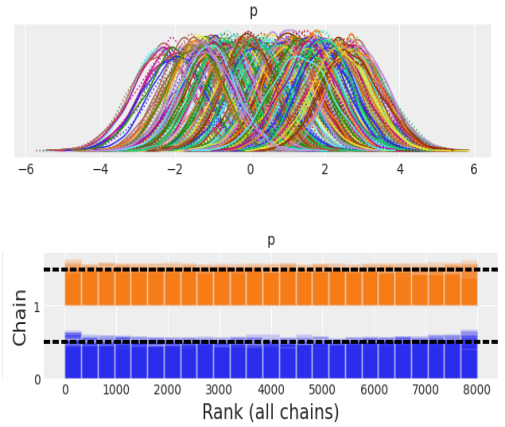
[5]Abril-Pla, Oriol. "PyMC3 & ArviZ: Bayesian Estimation and Diagnostics." Oriol Abril-Pla's Blog, https://oriolabrilpla.cat/en/blog/posts/2020/pymc3-arviz.html.

Appendices

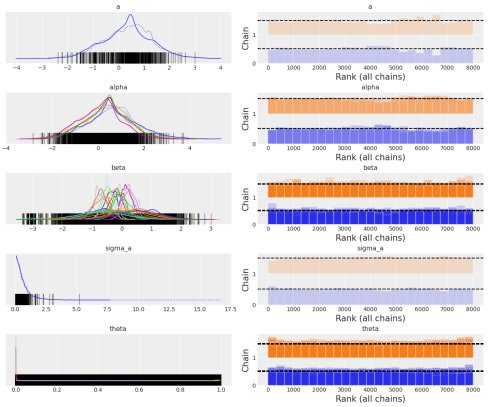
Appendix A



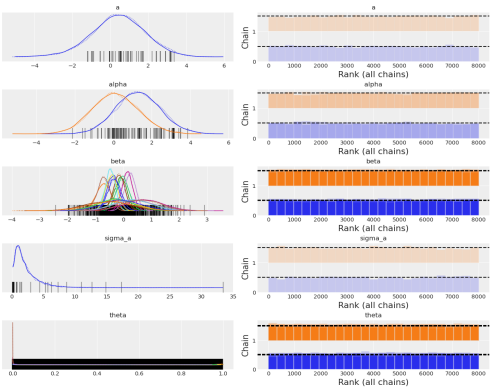
Sampling Evaluation for Logistic Regression Model



Sampling Evaluation for BART



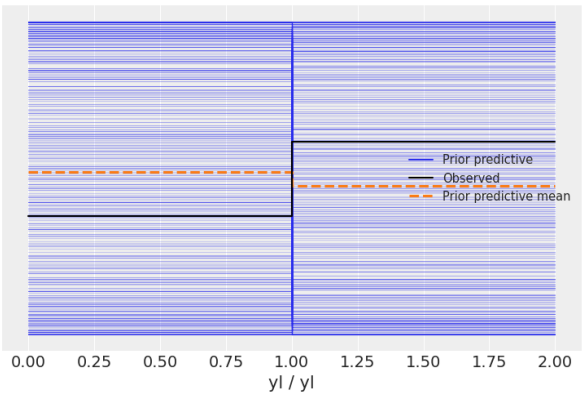
Sampling Evaluation for Hierarchical Model - Stage



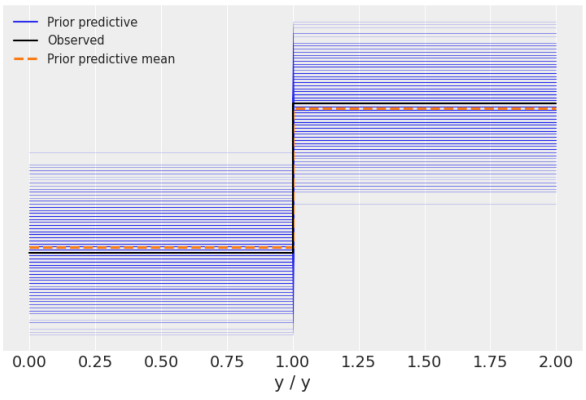
Sampling Evaluation for Hierarchical Model - Drug

Table: Sampling evaluation for each model

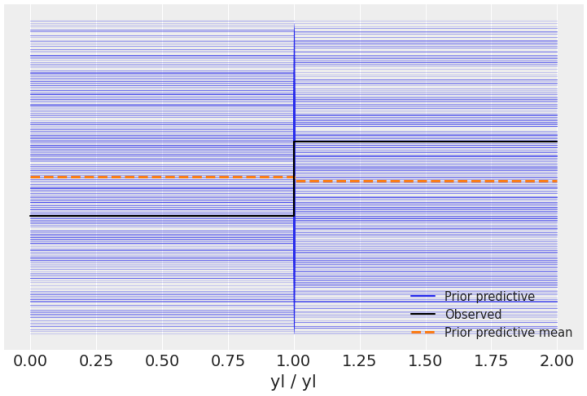
Appendix B



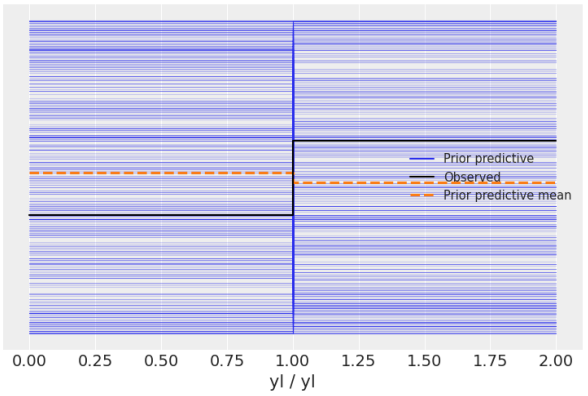
Logistic Prior Predictive Check



BART Prior Predictive Check



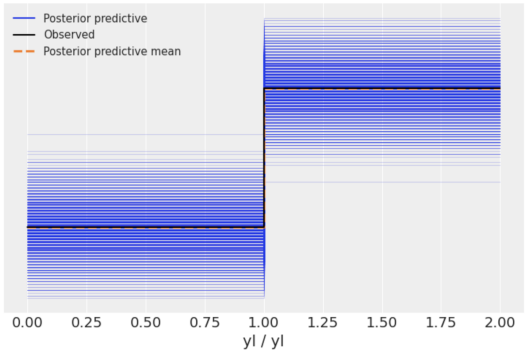
Hierarchical Model - Stage Prior Predictive Check



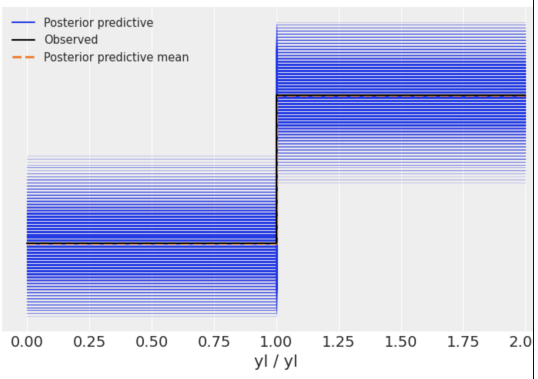
Hierarchical Model - Drug Prior Predictive Check

Table: Prior Predictive Checks

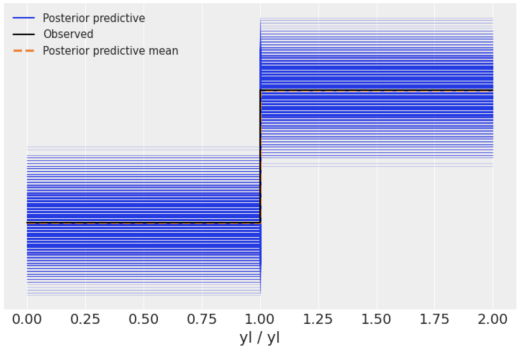
Appendix C



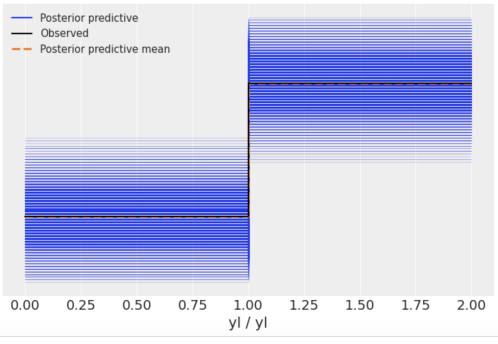
PPC for Logistic Regression Model



PPC For BART



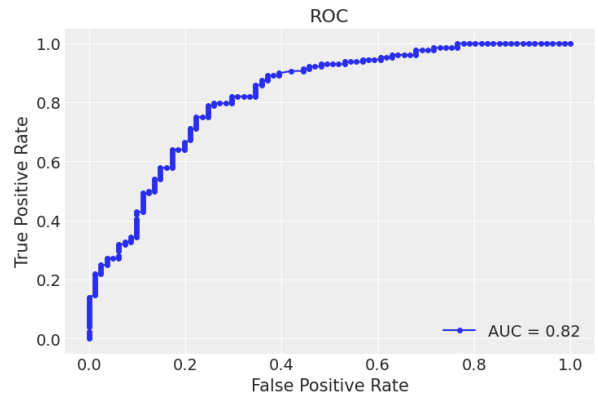
PPC for Hierarchical Model - Stage



PPC for Hierarchical Model - Drug

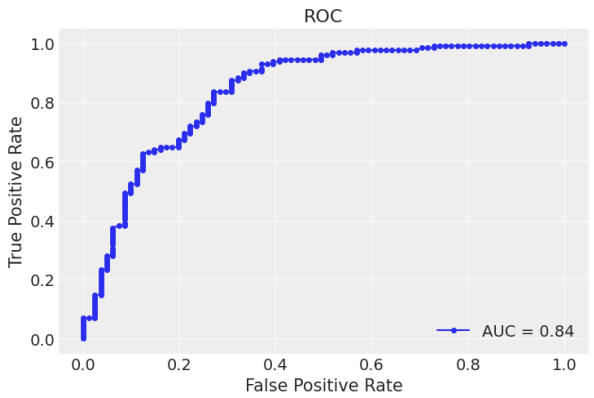
Table: Posterior Predictive Checks

Appendix D



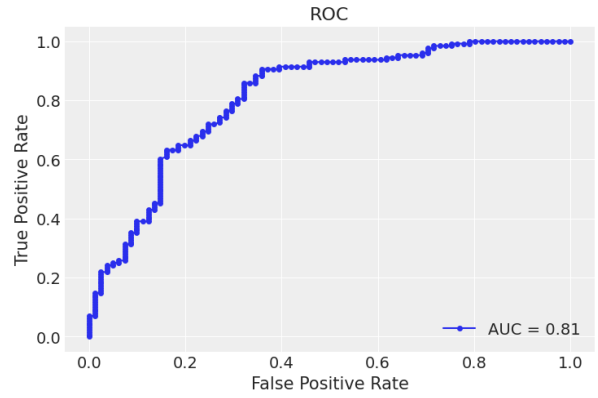
accuracy = 0.770

Logistic ROC Curve



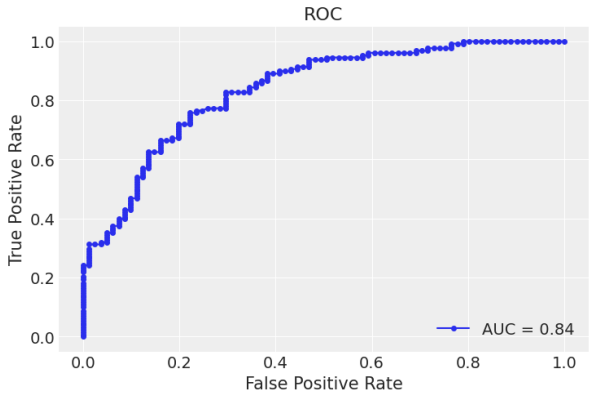
accuracy = 0.804

BART ROC Curve



accuracy = 0.785

Hierarchical Model - Stage ROC Curve

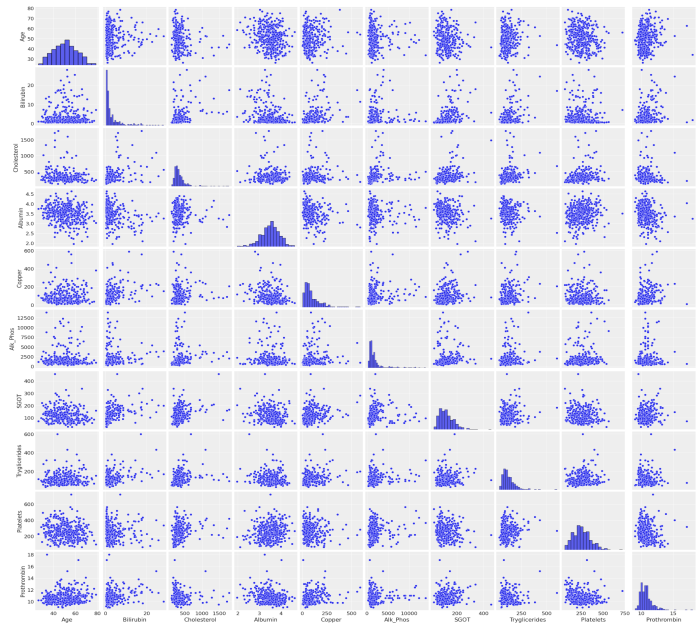


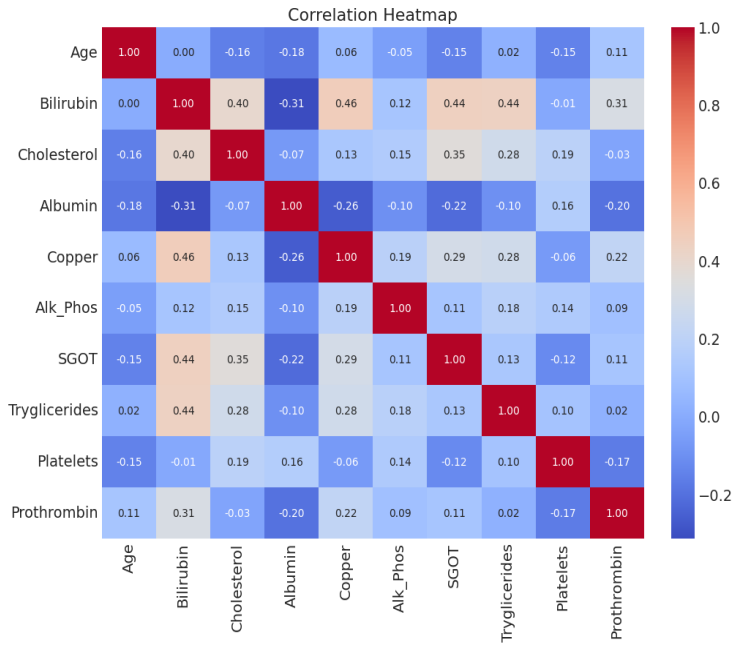
accuracy = 0.785

Hierarchical Model - Drug ROC Curve

Table: ROC Curves and Accuracy Scores

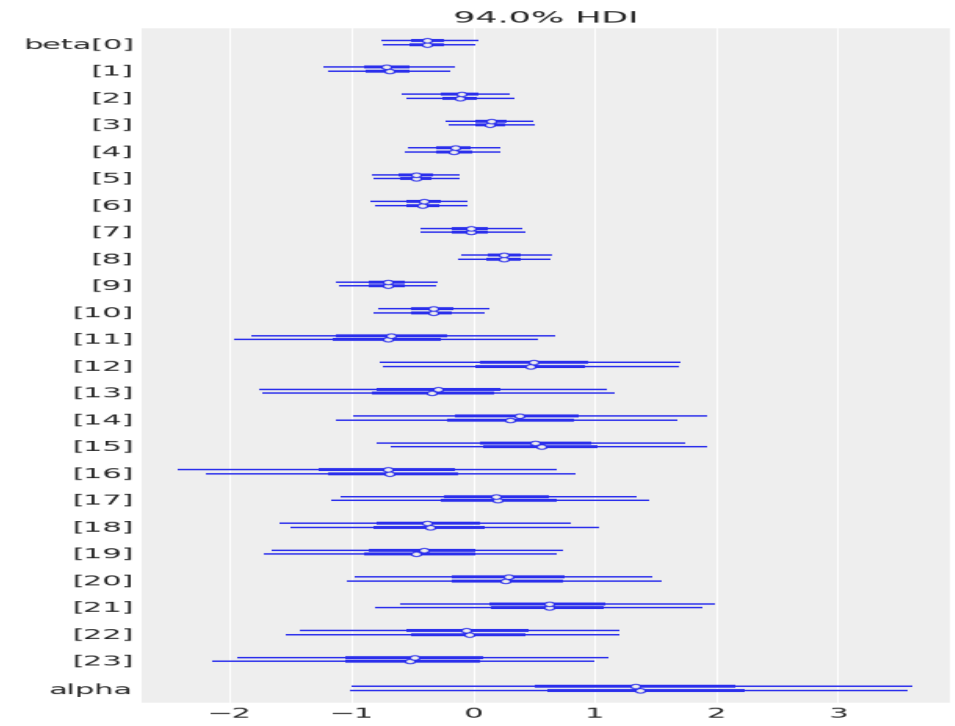
Appendix E



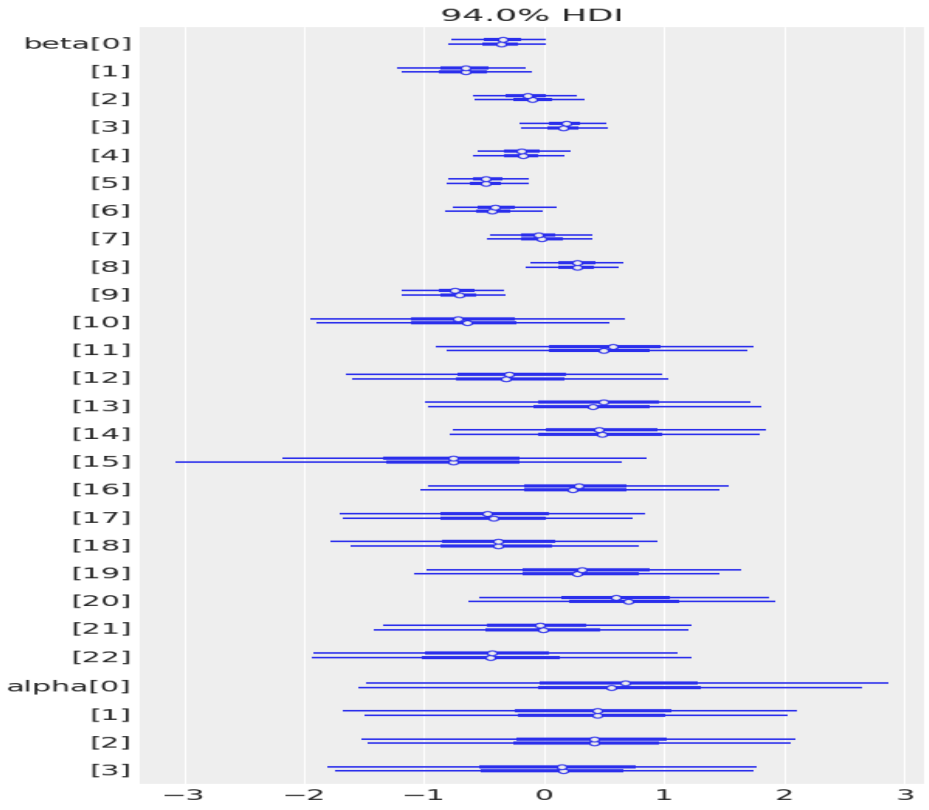


Exploratory Data Analysis

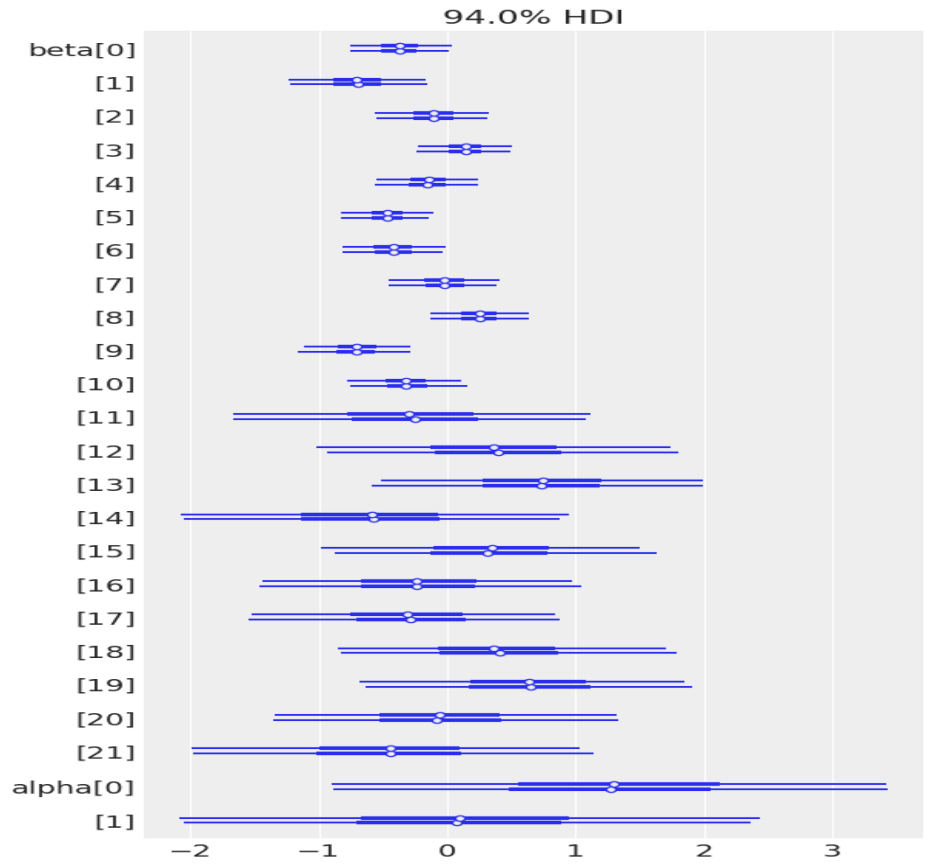
Appendix F



Forest Plot for the Logistic Regression Model



Forest Plot for the Hierarchical Model - Stage



Forest Plot for the Hierarchical Model - Drug Use