Hello, I’m Jason and I’m 3rd year in the undergraduate Applied Mathematics and Statistics program.

For the final project, I am working by myself on an interesting problem: can we predict heart failure based on publicly-accessible data? This is important, because heart failure is a major global killer and early detection could save many lives.

**PIC of summary**

To do this I am using a dataset from Kaggle, which has 13 variables, including a dependent variable. The dependent variable is binary, which means logistic regression is a sensible model to address this problem. The dataset has only 300 observations.

The independent variables include binary variables, discrete variables but also one continuous variable.

**PIC of model diagram**

This is the model diagram. Normal priors are used because regression parameters are continuous and exist in the interval of negative to positive infinity. The bournoulli likelihood reflects the dependent variable being strictly binary. This will help to model the dependent variable as a distribution.

I am going with the assumption that there is no prior information.

My first question was, do any variables correlate strongly? This would allow me to eliminate variables easily.

To compare the numerical variables, the standard Pearson Correlation Coefficient was used.

For binary-to-binary variable comparisons, I used the Mathews Correlation Coefficient and for numerical-to-binary comparisons of correlation, I used the biserial correlation method.

**PIC of heatmap**

I couldn’t find a nice way to visualize these correlations together, so I made a function that does the appropriate correlation method for each pair of variables, based on whether they’re both numerical, both binary or a combination. Then I made this heatmap. Each label has its variable type as either N for numerical or B for binary.

All results showed a correlation of 0.4 or lower, so there’s no good reason yet to drop any variables, yet.

To see if we can drop some variables, I used R and JAGS to construct a model that uses all variables. The resulting posterior plots should give me ideas about what to keep.

I eventually found that some settings allowed for good diagnostics:

Here, we can see in every case that the chains overlap, the autocorrelation is low, the shrink factor is below 1.1 and the chains overlap with a low MCSE. Not all diagnostics are shown here, because there are 13 variables and they’re all much the same.

**PICS of diagnostics**

Once I was satisfied this had gone well, then I plotted the posteriors together.

**PIC of posteriors**

Several of these posteriors show that 0 falls almost in the middle of the posterior distributions. For the 5 variables of anaemia, diabetes, high blood pressure, platelets and smoking, there is insufficient evidence to reject the null hypothesis that the parameter values are 0. So, these variables won’t be included in the final model selection. Another 4 variables are just inside the 95% HDI, so final model comparisons will test models that include at least 4 and at most 8 of these parameters.

**PIC of TODO:**

**Observed:**

* low correlations between variables

**TODO:**

* Include predicted values in the model text
* Run several models for model selection
* Use Bayes Factor and sensitivity analysis to determine final model

This is as far as I have gotten so far. The next steps will be to run several models at the same time, with different combinations of the remaining parameters and variables. I will add some predictions to the model text to evaluate the success of each model. Then I can use a Bayes Factor analysis and sensitivity analysis to determine the final model.