Use of Bayesian Models to Reduce Sample Size in Functional Observational Battery Experiments

Nathaniel Bowerman, Steve Denham

charles river

1 Background

Charles River Laboratories repeatedly performs the same types of experiments, and thus has a large set of historical data for many types of experiments. If this historical data could be incorporated into the statistical model used, it could increase the power or reduce the required sample size for a new study. Bayesian models provide a well established framework for the incorporation of historical data into statistical models.

The type of experiment we focused on for this study was Functional Observational Battery (FOB). FOB experiments are used to assess behavioral changes in a study population by measuring the response of experimental units to various stimuli. FOBs have the advantage of containing both continuous and categorical response variables, but we chose to focus on categorical response variables for the current study.

The goal of this project was to determine if we can use data from previous studies in a Bayesian model to reduce the sample size required, compared to a Frequentist approach, while maintaining similar power and accuracy.

2 Methods

To find Bayesian models that we could use to analyze FOB studies, we analyzed a single study was using both Frequentist and Bayesian methods. Some characteristics of this dataset:

-Rat study
-10,243 observations total
-One vehicle control group (Group 1), and 3 drug groups with different dosages (Groups 2-4)
-260 analysis subsets split by:
  -Male/female
  -Five data collection timepoints in dataset
  -26 activity and measurement combinations
-127 subsets showed no variation between scores and group and were removed from analysis

Table 1. Example table used for analysis byvar=14

		Score				
		0	1	2	3	Total
Group	1	0	8	1	1	10
	2	0	10	0	0	10
	3	1	7	2	0	10
	4	2	7	1	0	10
	Total	3	32	4	1	40

Typically, the data is split into subsets by sex, timepoint, and activity/measurement and analyzed separately. An example of the frequency table for a subset is given at the right. For each analysis subset, Exact Mantel-Haenszel, ordinal logistic regression, and nominal logistic regression models were fit using a Frequentist framework. The Bayesian models used are listed below. All models were fit using 100,000 posterior iterations and a burn-in of 15,000 iterations.

Model 1

- Bayesian equivalent to logistic regression (ordinal groups)
- $Y_{i,j} \sim \text{multinomial}(n_{i,j}, p_{i,j})$
  - Score=i, group=j
- $pi_{,j} = \gamma_{i,j} - \gamma_{(i-1),j}$
- $\gamma_{i,j} = \text{logistic}(\alpha_i + \beta^* (i-1))$
- $\alpha_i \sim \text{normal}(0,100,\text{lower}= \alpha_{i-1})$
- $\beta \sim \text{normal}(0,100)$

Model 2

- Bayesian equivalent to probit regression
- $Y_{i,j} \sim \text{multinomial}(n_{i,j}, p_{i,j})$
  - Score=i, group=j
- $p_{i,j} = \gamma_{i,j} - \gamma_{(i-1),j}$
- $\gamma_{i,j} = \Phi(\alpha_i + \beta_j)$
- $\alpha_i \sim \text{normal}(0,100,\text{lower}= \alpha_{i-1})$
- $\beta_j \sim \text{normal}(0,100)$

Model 3

Model 4

2 Methods (cont.)

The results were first examined to determine if the models were converging properly. The significant and non-significant subsets for the Frequentist and equivalent Bayesian models were compared to determine if the results for the Bayesian models matched the Frequentist and if there was a change in power. Significance for Bayesian models was determined by examining the 95% HPD for the parameter of interest.

Next, we used the data from FOB studies conducted in 2017 to generate priors for Bayesian models. This dataset consisted of:

-All negative control and pre-treatment observations from 2017
-369,699 observations
-89 studies
-140 data subsets split by:
  -Sex: Male, Female
  -Species: Mouse, Rat, Monkey
  -Measure*Activity: 40 combos total

We fit a separate model for each sex, species, and activity/measurement combination to generate priors for the control group parameters in the model. There are a total of 240 possible sex, species, and activity/measure combinations. In the case where there is no historical data, an uninformative prior would be used.

After this, we split the 2017 data into pre-treatment and negative control groups for each subset. We used a Cochran Mantel-Haenszel test to determine if there is a relationship between pre-treatment to negative control data for each subset.

3 Results

The Bayesian models for the single study were examined to determine if the models converged properly. A summary of the convergence diagnostics are given in the table below and an example of the diagnostic plots that are given for each parameter are given to the right for a single parameter.

Table 2. Summary of efficiency for all single study subsets

Model	Minimum	Mean	Median	Maximum
EMHnoAlpha	0.001	0.055	0.038	0.305
EMHwAlpha	0.001	0.052	0.044	0.208
LogReg	0.001	0.079	0.067	0.804
ProbitReg	0.001	0.073	0.065	0.905

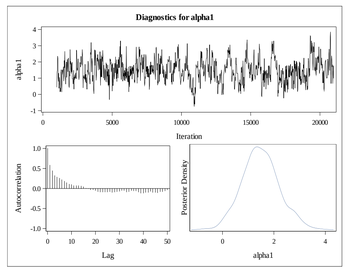

Figure 1: Example diagnostic plots for Bayesian parameter

3 Results (cont.)

The comparison between Frequentist and Bayesian models is summarized in the tables to the right. The single study results for Frequentist and Bayesian methods show that for Probit and Logistic Regression the Bayesian models have more power and are able to pick up all significant subsets that the Frequentist approach did.

To determine if the Bayesian models were accurately fitting the observed data, the correlation between predicted probabilities for each score and group comparison were examined. The results plotted to the right show that most models show a high correlation between observed and predicted probabilities with Probit regression having the least accurate model. Bayesian logistic regression was slightly more accurate than Model 4 which was slightly more accurate than Model 3.

There were 48.28% of subsets that showed a difference between pre-treatment and vehicle control groups. This demonstrates that the vehicle control has an effect on the score and thus these two groups should not be combined in the historical control data.

Tables 3-6. Comparison of significant subsets to equivalent Frequentist model for each Bayesian model

FreqLogistic	BayesLogistic Non-sig	BayesLogistic Sig	BayesLogistic Total
Non-sig	99	11	11
Sig	0	23	2
Total	99	34	13

FreqProbit	BayesProbit Non-sig	BayesProbit Sig	BayesProbit Total
Non-sig	94	14	108
Sig	0	25	25
Total	94	39	133

FreqEMH	BayesEMHnoAlpha Non-sig	BayesEMHnoAlpha Sig	BayesEMHnoAlpha Total
Non-sig	93	12	1
Sig	11	17	
Total	104	29	1

FreqEMH	BayesEMHAlpha Non-sig	BayesEMHAlpha Sig	BayesEMHAlpha Total
Non-sig	96	9	105
Sig	13	15	28
Total	109	24	133

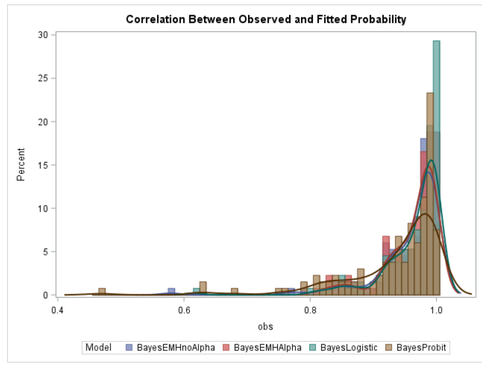

Figure 2: Histogram of correlation between observed and fitted cell probabilities by subset for each Bayesian model

5 Conclusions/Future Work

We were able to demonstrate that the Bayesian models fitted converged well and were accurate. Equivalent Bayesian methods showed more power than the Frequentist approach for the same dataset for non-informative priors. It would be expected that power would increase with an informative power, indicating that there is precedent for using an informative power to reduce sample size, compared to a Frequentist approach.

In order to maximize our use of historic data to generate priors for our typical study design, the vehicle control group data should come from the same distribution independent of the vehicle being used. We showed that this was not the case, so we will be moving to a mixed model approach to capture these effects.