Thesis Outline
Applied and Industrial Mathematics Program
Towson University

Matthew Tiger
Thesis advisors: Mr. Jason Muhlenkamp and Dr. M. D. Voisei
Spring 2018

April 8, 2018

I. **Introduction**

A. **Thesis Statement** We will use the theory of Bayesian statistics to develop a model that will forecast future telecast airings' audience impression concentration distributions corrected for small sample sizes present in the measurement training data. We will then aggregate those distributions in order to provide forecasted audience impressions distributions for media plans.

II. **Motivation and Background**

A. **Motivating Example** Provide example of Amazon product ratings. In this example, show that the rating of 4 stars, but only has 3 reviews is more worthy of suspicion than the rating of 4 stars with 3000 reviews.

B. **Problem Description** Suppliers give us forecasted TV average audience data for their selling titles. We use audience measurement data to come up with forecasted average audience percentages for the selling titles. These forecasted average audience percentages our used by our optimization engine to recommend media plans given certain objectives and constraints. The suppliers give these media plan as recommendations to advertisers and are mostly interested in the outcome of the media plan as a whole.

We need to explain the following terms in the problem description; their explanation should be used cohesively and will help guide the problem description:

i. Selling title
ii. Broadcast week
iii. TV rating

iv. Average Audience
v. Media plan
vi. Telecast
vii. Schedule
viii. Post-Campaign Reporting

C. **Data**

i. **Measurement Data** We are given audience measurement data that provides TV viewing habits of the panel sample. For each historical telecast, we are given the respondents that were watching and their total seconds of commercial viewing and total viewing.

ii. **Supplier Data** A given supplier provides us with a schedule that list all content airings historically and their tentative future airings. Additionally, for each selling title / week airing in the future

III. **Model**

A. **Description**

i. **Assumptions**

- The draws from the likelihood for a given telecast are iid Bernoulli variables. This assumption must be checked because we know that the draws from the likelihood might not be independent since co-viewing occurs within the same household.
- The hierarchical nature of the model which pools data.
- telecasts within a week are independent of each other. This means that the distribution of a selling title week is just the union of the forecasted distributions of each telecast within that week.
- list units of analysis and outcome variable.
- We want to use raw sample data to approximate average audience impression concentrations.
- We determined from literature and substantive experience that network, content, and air time are critical to determining the impression concentration values of audiences.
- Further, the data that we observe for each telecast airing is the result of Binomial experiment so it is natural to model the data-generating process under a hierarchical logistic regression model.
- Display Kruschke diagram for the model.

B. **Implementation**

i. **pymc3** We make use of computer software to aid in the numerical computation of the posterior distribution. Namely, we use pymc3 to sample from the posterior distribution which makes use of Hamiltonian Monte Carlo sampling techniques. However, Sampling with this method was not feasible for on-demand forecasts.

   a. We approximate sampling using ADVI mini-batch in order to improve sampling speed. ADVI works best when posterior is Gaussian without too many uncorrelations. We include python module used to sample from posterior in appendix.

C. **Validation** We want to make sure that the posterior distribution resulting from sampling matches not only our beliefs about the system, but also the observed data that is theoretically generated

   i. **Sensitivity Analysis** How much do our inferences change if the model changes or the prior distributions change? We will change the prior distribution used in the model, i. e. change the regressors' prior distributions to different parameters, and κ to different parameters and assess changes in inference. Feature selection is also encompassed here.

   ii. **Posterior Predictive Checks** "If the model fits, then replicated data generated under the model should look similar to observed data." To this end, we will use the posterior distribution to generate replicated data under the model and compare to observed values. We will check both raw impression concentration values, i. e. taking the raw proportion values of in-target sample / total sample, and the in-target sample values. We will look at the mean, min, max, and standard deviation of replicated vs. observed values and develop test-statistics to determine if they are statistically different.

D. **Predictive Accuracy**

   i. **Raw proportion vs. impression concentration values** Is using the raw proportion of successes to total trials a good approximator for the actual average audience concentration values? We observe that in the limit, this raw proportion tends to the average audience concentration that would be computed. We should rigorously prove that this view is valid.

   ii. Measures used for predictive accuracy in BDA3 that still need to be researched. We will use WAIC for model comparison and CRPS for probabilistic forecast evaluation and MAPE for point estimate forecasts with in-target sample predictions.

E. **Selling Title / broadcast week distributions** Using the posterior distribution created, the distribution of outcomes for a selling title / week is the union of distributions for each telecast associated to that

selling title within the week due to the assumption of independence of these telecasts.

F. **Summarizing inferences** We need to summarize the inferences generated by the model to a single point for the optimization engine. We need to describe that strategy here.

IV. **Results**

A. **Media plans** Now that we have a valid model, we want to evaluate its performance at the media plan level. We should look at random allocations of selling title weeks and look at the forecasted distribution and compare with actual.

V. **Conclusion** This model helps understand the uncertainty around our forecasts and gives insight into distribution of possible outcomes for each item at the allocation level. This can then be used to understand the uncertainty around the proposed media plan and can help suppliers anticipate any misses on their contractual guarantees.