TOWSON UNIVERSITY

MASTER'S PROJECT

Aiding Linear Television Media Planning Through Bayesian Inference and Forecasting

Author:
Matthew TIGER

Supervisors:
Mr. Jason MUHLENKAMP
Dr. M.D. VOISEI

A project submitted in fulfillment of the requirements for the degree of Master of Science

at

Towson University

May 1, 2018

Declaration of Authorship

I, Matthew TIGER, declare that this project titled, "Aiding Linear Television Media Planning Through Bayesian Inference and Forecasting" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this project has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:
Matthew Tiger

Date:
2018-05-01

Acknowledgements

I would like to use this space to thank my advisers Mr. Muhlenkamp and Dr. Voisei for their patience in working with me on this project; without it this would have been much more difficult. I would also like to thank the fine folks at Videology Inc. for allowing me the opportunity to work on this challenging problem. Last, but not least, I would like to thank Stephanie Romano for dealing with the long nights and providing much needed encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

ACM	**A**verage **C**ommercial Minute
CR	**C**redible **R**egion
MAE	**M**ean **A**bsolute **E**rror
MCMC	**M**arkov **C**hain **M**onte **C**arlo
HMC	**H**amiltonian **M**onte **C**arlo
NUTS	**N**o **U** **T**urn **S**ampler
CRPS	**C**ontinuous **R**anked **P**robability **S**core

1. Introduction

The world of advertising today consists of a multitude of options for delivering a message to a targeted audience, but the relevance of TV advertising remains as important as ever.

Thus, the need to correctly account for and forecast audience viewership is not surprising. Throughout this paper we discuss why that need exists, the challenges of providing such forecasts, and a solution to the issue at hand.

1.1 Background

The TV advertising landscape consists of TV sellers, who have airtime available for sale to be used to air advertisements, and TV buyers, who purchase airtime from TV sellers in order to air their desired advertisements.

In the normal course of events, TV buyers specify the target metrics that they wish to achieve through advertising such as meeting a certain number of impressions within their demographic target audience, i. e. a certain number of advertisement views. The TV sellers then create a media plan, which is a collection of advertising spots/units, to present to the TV buyer. The basis of the media plan is the forecasted number of demographic (or buy-demo) impressions for each unit. These demographic impressions are called the buy-demo impressions because the TV buyer is charged based off of the total number of forecasted demographic impressions in a media plan . The TV buyer then accepts or proposes changes to the media plan and eventually a deal is struck and the units are set to air. These buy-demo impressions are guaranteed to the TV buyer and additional units will be added in order to achieve that guarantee [5]. Thus, the need for accurate forecasts becomes clear under this paradigm.

In addition to having a demographic target audience, the TV buyer may also have some sub-population target in mind, called a strategic target audience. It is then the job of the TV seller to provide forecasted in-target impressions for each unit in order to generate the media plan. The forecasted strategic target impressions are generated based off the forecasted buy-demo impressions since it is the latter that form the currency of the media plan.

Note that the forecasts that are used by TV sellers are based off of television viewership as measured through audience measurement companies. Some companies provide such measurements using a statistical sample of the TV viewing population and track the viewership of that sample, extrapolating to the population at large. However, such approaches have disadvantages as will be discussed in the coming sections.

1.2 Challenges

For the audience measurement source used for this paper, the sample sizes present on an individual program airing are orders of magnitude smaller than the statistical panel sample. The issue is exacerbated on networks that have low viewership and is made worse when the strategic target is small relative to the national population. This is the major challenge of providing forecasts for media plans; the measurements provided through the measurement company are based off of a sample and results in noisy data for the typical past program airing.

A related problem exists in evaluating the performance of baseball players: players are judged by their batting average (percentage of hits) but this metric is not informative when the player has few at-bats. With more information about the league and past historical performances we are able to come up with a better estimate that takes such factors into account as well as the observed at-bats through Bayesian inference.

Thus, we adopt the use of a similar approach in order to create better forecasts when the data from past airings with small sample sizes.

1.3 Problem Description

Given the challenges outlined above, the problem this paper aims to solve is whether or not we can devise a model using the noisy measurement data as well as the TV supplier data that will accurately forecast the strategic target impressions for future program airings. In addition to accuracy, forecasting the range of possible outcomes for each future airing is desired so as to understand the probability that an advertiser will reach their targeting goals given a media plan.

To this end, we use the theory of Bayesian statistics to develop such a model and provide the desired forecasts.

2. Data

In order to provide the forecasts necessary for media planning, we make use of several data sets. Namely, TV supplier data, which contains the inventory to plan against, and audience measurement data, which contains historical program airings and the amount of viewing associated to those airings for various audiences. We discuss these data sets in more detail in the sections below.

2.1 TV supplier data

TV suppliers provide us with a holistic view of their inventory that is for sale. The most granular unit in their inventory is a selling title during a future broadcast week, i. e. a week that starts on a Monday and ends the following Sunday. This unit is a grouping of content that will air during a given broadcast week and the grouping can be based on either similar time of airing or similar content. The two most important pieces of data provided by TV suppliers are the linear media schedule, which contains the exact air times of selling titles, and the forecasted demographic impressions, both of which are explained below.

2.1.1 Linear Media Schedule

The linear media schedule lists all scheduled program airings on a given network. Associated to each airing is an identifier for the selling title and an identifier for the content. Note that the selling title indicates the grouping of content that is sold while the content indicates the nature of the programming that will be shown. Example records from a linear media schedule are shown in Table 2.1.

network	selling title	selling title name	content	content name	start datetime	end datetime
BCST	100	Adult Cartoon 8PM	10	Adult Cartoon	2017-04-02 20:00:00	2017-04-02 20:30:00
BCST	101	Adult Cartoon 8:30PM	10	Adult Cartoon	2017-04-02 20:30:00	2017-04-02 21:00:00

TABLE 2.1: Example linear media schedule data for two selling title airings.

As can be seen from Table 2.1, the same content identifier can be associated to multiple selling titles. This indicates that the TV supplier wishes to differentiate the content airing at 8:00 PM from the content airing at 8:30 PM, usually to price the airings differently.

2.1.2 Forecasted Demographic Impressions

In the advertising industry, an impression is synonymous with an advertisement viewing. Thus, this data contains for each selling title, broadcast week, and demographic audience the expected impressions per unit, i. e. the expected number of

audience members that will watch a commercial airing during the given selling title and broadcast week.

Example records from the forecasted demographic impressions data are shown in Table 2.2

selling title	broadcast week	demographic	impressions per unit
100	2017-03-27 06:00:00	F45-49	150000
100	2017-03-27 06:00:00	P18-49	1500000
101	2017-03-27 06:00:00	F45-49	120000
101	2017-03-27 06:00:00	P18-49	1000000

TABLE 2.2: Example forecasted demographic impressions data for two selling title weeks.

2.2 Audience measurement data

As mentioned previously, the audience measurement data contains the programs that have aired historically and who watched them according to the audience measurement source. This data is stored in two separate tables, the program table and the viewing table.

The program table details the historical program airings and meta data about those airings. A program in the table corresponds to an identifier of a logical grouping of aired content per the measurement source while the telecast corresponds to an identifier of a single airing of a program. Some information included is the network the program aired on, the start and end datetime of the program airing, as well as the genre associated to the program and whether or not this is a repeat airing of the telecast. Note that the genre information serves to categorize the programs aired by grouping together programs with similar content. Some example program data is provided in Table 2.3.

network	program	telecast	program name	start datetime	end datetime	genre	is first run	is live
BCST	1000	301	Adult Cartoon	2017-04-02 20:02:00	2017-04-02 20:30:00	Animation	1	0
BCST	1000	302	Adult Cartoon	2017-04-02 20:30:00	2017-04-02 21:00:00	Animation	1	0

TABLE 2.3: Example program table data for two historical program airings.

Associated with the program table is the viewing table which details the respondents' viewing of historical programs. This data is stored at the minute level and contains which respondents were watching each minute of a given telecast and how many commercial seconds aired during the minute as well as the total number of commercial seconds that aired. Additionally, limited identifying information about the respondent is stored in this table such as the respondent's age and gender. Some example viewing data is provided in Table 2.4.

Of primary interest from this data set is the impressions per unit, or average commercial minute (ACM), for a given audience. The definition of a program airing's ACM measurement is the weighted average of the in-target commercial viewing seconds over the number of commercial seconds that aired during the program's airing.

Formally, let A be the set of panelists that are considered in-target out of a total of n possible panelists. Suppose that program airing i has t_i minutes of programming and let s_{ij} be the number of commercial seconds during minute j of the airing. Further, let $p_{ijk} = 1$ if panelist k was watching during the j-th minute of program

program	telecast	respondent	minute	comm secs	weight	age	gender	total comm secs
1000	301	2	3	60	2050	48	F	120
1000	301	2	4	45	2050	48	F	120
1000	301	2	15	60	2050	48	F	120
1000	301	2	16	30	2050	48	F	120
1000	302	2	22	15	2050	48	F	100
1000	302	2	23	60	2050	48	F	100

TABLE 2.4: Example viewing table data for one respondent's viewing on historical program airings from Table 2.3.

airing i and 0 otherwise and let w_k be the weight assigned to the panelist by the measurement source. Then, m_i^A, the ACM of program airing i is

$$m_i^A = \left\lceil \frac{\sum_{k=1}^{n} w_k \sum_{j=1}^{t_i} s_{ij} p_{ijk} \mathbf{1}_A(p_k)}{\sum_{j=1}^{t_i} s_{ij}} \right\rceil. \tag{2.1}$$

Throughout, we will refer to impressions and ACM interchangeably as the ACM is in fact a measurement of the number of impressions a commercial airing receives.

A related quantity for program airing i, the impression concentration of target A relative to target B where $A \subseteq B$, is

$$c_i = \frac{m_i^A}{m_i^B}.$$

This quantity is used to measure the proportion of impressions attributed to target A relative to target B.

As an example, if we consider target A to be the set containing only respondent 2, then from Table 2.4 we see that for program 1000 and telecast 302, $m^A_{(1000,302)} = 1538$.

2.3 Data Used

For the purposes of this project, we limit the data used from the above data sets to three networks labeled BCST, ETMT, and SPTS. These networks are a broadcast network that has large reach, a cable entertainment network, and a cable sports network, respectively. Further, we limit the audience measurement and linear media schedule data to only airings that occur between 2015-12-28 06:00:00 and 2018-01-01 06:00:00 which coincides with broadcast years 2016 through 2017. We shall consider the in-target audience, target A, to be Females aged 45 through 49 and the demographic audience, target B, to be Persons aged 18 through 49. The demographic audience was chosen as it is representative of the demographic audience typically chosen when constructing a media plan while the in-target audience was chosen as it is small relative to the total number of TV viewers in the United States, roughly 3.3%. As a consequence, the size of the in-target panel in the audience measurement data is small and the target is exemplary of the issues this project aims to solve.

The data used for training and testing the model is combined from the above tables to create a data set that associates to each selling title airing in the linear media schedule the program airing in the measurement data whose start and end datetimes most overlap with the start and end datetimes in the linear media schedule. We then assign the selling title airing the ACM of the matched program airing for target A and target B. Example data is shown in Table 2.5. Throughout this paper, we will

consider the training set to be the airings occurring during broadcast year 2016 and the test, or hold-out, set to be the airings occurring during broadcast year 2017.

network	selling title	content	start datetime	end datetime	program	telecast	ACM A	ACM B
BCST	100	10	2017-04-02 20:00:00	2017-04-02 20:30:00	1000	301	110560	1203560
BCST	101	10	2017-04-02 20:30:00	2017-04-02 21:00:00	1000	302	210560	1501000

TABLE 2.5: Example data set used for model for two selling title airings.

3. Model

Using the data set from Section 2.3, we present a model that will forecast the ACM of target A given the ACM of target B.

3.1 Preliminaries

In the following section we will provide preliminary information necessary to understand the model.

3.1.1 Units of Observation and Analysis

As mentioned in Section 2.3, the units of observation for this model are individual content airings in the linear media schedule. The outcome variable that is measured for each unit i is m_i^A, the ACM of target A, which we denote as y_i for notational convenience. Similarly, we denote m_i^B, the ACM of target B for unit i by n_i.

The desired units of analysis for this model are the selling title airings during a given broadcast week. These are the units that comprise a media plan allocation and are of importance in evaluating the performance of the forecasting model. However, for the purposes of this paper, we will instead use the units of observation as the units of analysis in order to gain a better understanding of basic model performance.

3.1.2 Covariates

There are two types of covariates that will be used in constructing the model: time covariates and program covariates. Time covariates relate to the time in which the unit of observation airs while program covariates relate to the a unit of observation's content.

The covariates that are most important in measuring audience on TV are time covariates such as day of week and time of day as well as program covariates such as the genre of the program airing [5]. The forecasting models used by Danaher et al. that had the highest forecast accuracy when measured by Mean Absolute Error (MAE) allowed for random effects for each program. Further, the forecast accuracy was greatest when a separate model was fit for each network. Thus, we proceed to include the time and program covariates mentioned, plus a content covariate that indicates per the linear media schedule the program that will air. See below for a detailed outline of the covariates that will be used. Binary covariates are "shifted to have a mean of 0 and to differ by 1 in their upper and lower conditions" per Gelman et al [8].

1. Broadcast Month - The broadcast month associated to the start of a program airing. The broadcast and Gregorian calendars are related in that the first week

of every broadcast month always contains the Gregorian calendar first of the month [2].

2. Day of Week - The day of the week associated to the start of a program airing. These are encoded from 0 - 6 with 0 being Monday and 6 being Saturday.

3. Stratified Hour - We define this covariate as the following groupings of hours which are adapted from the measurement source:

   - morning - The hour associated to a program airing start time is between 6 and 9, inclusive.
   - daytime - The hour associated to a program airing start time is between 10 and 14, inclusive.
   - early fringe - The hour associated to a program airing start time is between 15 and 18, inclusive.
   - prime 1900 - The hour associated to a program airing start time is 19.
   - prime 2000 - The hour associated to a program airing start time is 20.
   - prime 2100 - The hour associated to a program airing start time is 21.
   - prime 2200 - The hour associated to a program airing start time is 22.
   - late fringe - The hour associated to a program airing start time is 23, 0, or 1.
   - graveyard - The hour associated to a program airing start time is between 2 and 5, inclusive.

4. Content - The identifier denoting the content associated with the program airing per the linear media schedule. This covariate will be set to -1 if the content is considered a "new program", i. e. the content only airs in the hold-out set and not in the train set.

5. Lead-in Content - This covariate intends to capture the "lead-in effect" where "viewing a program on the same [network] prior to the current program enhances the viewing of the current program" by allowing for random effects based off the of the content identifier that aired immediately prior to the unit of observation [5]. This covariate will be set to 0 if there was no prior airing within 15 minutes of the start of a program airing.

6. First-run - This covariate denotes if this is the first time that the program content has aired where the covariate is 1 if the content has not aired previously and 0 if the specific program has aired before, i. e. the program is a "repeat". This information is obtained through the measurement source.

7. Live-program - This covariate denotes if the program is airing as the event is occurring per the measurement source where 1 indicates that the program is airing live and 0 otherwise. A typical example of a program airing that has a live-program value is a sports event such as a soccer game. Note that this value per the measurement source is 0 for all airings on the BCST network so it is excluded from that network's model definition.

8. Genre - The genre associated to the airing which are categorizations of programs based off of the associated content per the measurement source. Some examples of the genre are Animation, General Drama, and Sports Entertainment.

3.1.3 Bayesian Inference

Bayesian inference is a statistical practice that allows one to provide probability statements that express one's uncertainty on the occurrence of events. The normal course of practicing Bayesian inference involves specifying a model for the data y known as the *likelihood*. The likelihood describes the generative process for the observed data as some probability distribution conditional on some unknown parameter vector θ and is denoted by $p(y|\theta)$. Once the likelihood has been established, prior belief about the parameter θ is defined in the form of a probability distribution. This distribution is known as the *prior distribution* and is denoted by $p(\theta)$. Once the model is defined, we are interested in determining the probability distribution of the parameter θ that determines the generative process of the data conditional on the data that has been observed. This distribution, $p(\theta|y)$, is the *posterior density* and is the main goal of Bayesian inference.

Bayes' Theorem states that the posterior density is proportional to the product of the likelihood and the prior distribution, i. e.

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{3.1}$$

which is a straightforward consequence of the definition of conditional probability.

It is formula 3.1 that allows for the computation of the posterior density.

3.2 Assumptions

We assume that the response variables y_i are exchangeable given the parameters of the model and the covariates of the unit of observation. A sequence of random variables is exchangeable if the "joint probability density $p(y_1, \ldots, y_k)$ is invariant to permutations of the indexes [8]." This allows us to model the data as independently and identically distributed given the covariates and unknown parameters. Note that exchangeability implies that only the model parameters and covariates determine the likelihood distribution.

3.3 Description

We will now provide a formal description of the model. We aim to predict the response variable y_i of each unit of observation i is the ACM of target A given n_i, the ACM of target B where $A \subseteq B$, and the covariates of the unit i. Since target A is contained in B and we are given n_i, it is natural to model the likelihood of y_i as a binomial distribution with unknown probability parameter π_i. This parameter represents the proportion of successes, y_i, given the number of trials, i. e. n_i. The prior distribution of the parameter π_i is modeled as a Beta distribution which is common throughout the literature [8, 13]. The hyper parameters of the Beta distribution α_i and β_i are reparameterized in terms of the Beta distribution's mode ω_i and concentration κ_i. The mode of the beta distribution ω_i is determined by the logistic regression equation of the model covariates while the concentration of the beta distribution κ_i has an exponential distribution dependent on the model covariates. We place Student's T distributons on the coefficients of the model covariates. Further, since each model covariate is categorical, we place a hierarchical structure on the model such that each category within a covariate receives its own coefficient but the coefficients themselves are drawn from the same distribution. This type of model is known as a hierarchical logistic regression model.

Let X_i be the vector of covariates for the unit of observation i. Then the above description of the model $\mathcal{M}$ can be written as follows:

$$\begin{aligned} y_i|X_i, n_i, \pi_i, \omega_i, \kappa_i &\sim \text{Bin}(n_i, \pi_i) \\ \pi_i|\omega_i, \kappa_i &\sim \text{Beta}\left(\omega_i\kappa_l + 1, (1-\omega_i)\kappa_l + 1\right) \\ \omega_i &= \text{logit}^{-1}\left(\beta_0 + \sum_{j=1}^{m} \beta_j X_{ij}\right), \quad \beta_j \sim t_4(0, \sigma_j^2) \\ \kappa_i|X_i &\sim \text{Exp}(\lambda_l X_{il}) \end{aligned}$$

where $\text{logit}^{-1}(\alpha) = \frac{\exp\alpha}{1+\exp\alpha}$.

3.4 Prior Distribution Choice

We must now justify the choice of prior distributions for the major components of model $\mathcal{M}$: the linear equation coefficients β_j and the concentration parameter κ_i.

When choosing the prior distribution of the coefficients β_j we wish to provide weakly informative priors, i. e. priors that are "intentionally weaker than whatever actual prior knowledge is available [8]." These priors act as constraints on the possible parameters that are to be sampled. From the literature, weakly informative priors for coefficients of logistic regression are chosen to be Student's T distributions with mean 0, degrees of freedom set to 4, and scale 2.5 [8]. The distribution is centered at 0 since apriori we do not know if the coefficient will be positive or negative. The degrees of freedom and scale are chosen such that the mass of the corresponding distributions falls between -5 and 5, which are roughly .01 and 0.99 on the logit scale, respectively. Thus, the prior is constrained such that each coefficient is allowed to account for a change in the logit probability from 0.01 to 0.99, which is a larger effect than is to be expected by any single coefficient. We allow the intercept of the model, β_0, to account for larger changes by adopting the same Student's T distribution with mean 0 and degrees of freedom 4, but with scale 5.

Similarly we choose the prior distribution of κ_i to be weakly informative as well. The concentration parameter κ_i can be interpreted as the average number of trials present in the units of observations. A standard choice for such a parameter would be the the exponential distribution with expected value set to be weakly informative for the data. Since the number of trials is constrained by the un-weighted ACM m_i^{β}, we know that this value can not be larger than the number of panelists which for this particular measurement source is on the order of 10^5. Thus, setting the value of $\lambda_l = 10^{-4}$ allows for the possibility of the concentration parameter to be equal to the total number of panelists, a number that is in practice much smaller.

Further, as can be seen from Figure 3.1, the number of trials appears to be dependent on the first-run covariate. Thus, we allow for heterogeneous concentration parameters by modeling the κ_i parameter as conditional on the covariates.

3.5 Inference

In order to compute the posterior density of our model, we perform the inference using the probabilistic programming language `pymc3` available through the programming language Python. This language allows us to specify the model $\mathcal{M}$ along with the observed data and also allows for the computation of the numerical approximation to the posterior density.

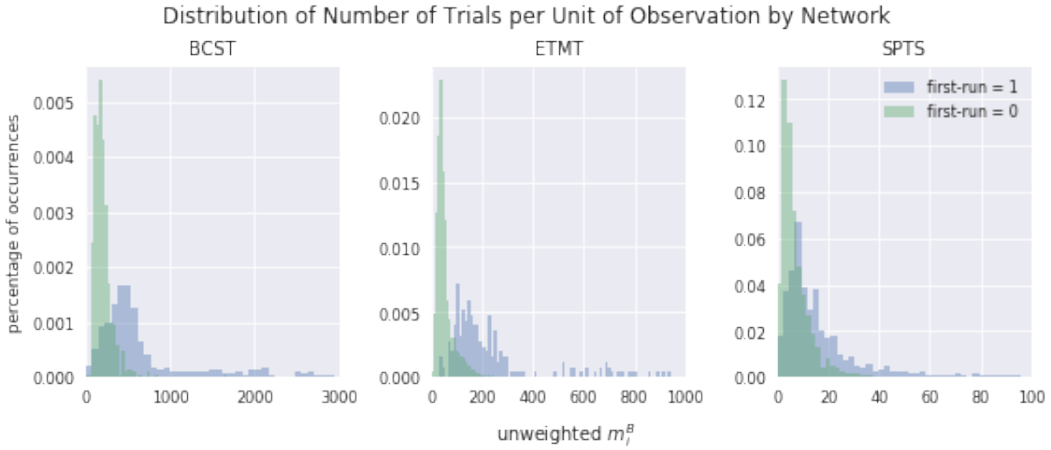

FIGURE 3.1: Distributions of the number of trials per each unit of observation factored by the covariate first-run. These distributions have much longer tails but have been truncated to illustrate the differences between each covariate factor's distribution. Note that the first-run units have a much wider distribution with a larger mean.

3.5.1 Computation

The computations of the approximation are performed through a Markov Chain Monte Carlo (MCMC) sampling algorithm. The algorithm itself is a variant of the Hamiltonian Monte Carlo (HMC) algorithm called the No U-Turn Sampler (NUTS). The description of the NUTS and HMC algorithms is beyond the scope of this paper, but the main idea is that the sampler borrows physical principles from Hamiltonian mechanics in combination with Monte Carlo simulation in order to compute the posterior density. The algorithm uses the Markov property to chooses the next point in the parameter space conditional only on the previous point [12].

An advantage of NUTS over regular HMC is that that the NUTS sampler has "several self-tuning strategies for adaptively setting the tuneable parameters of Hamiltonian Monte Carlo." This translates into the modeler only needing to specify the model and inferences are performed without much effort on the part of the modeler.

The primary parameter available to the modeler to control the adaptation routines of the NUTS sampler is the `target_accept` parameter. The value of this parameter determines the acceptance rate of the NUTS algorithm which is how often the sampler will accept target distributions during sampling. Higher values of this parameter correspond to smaller step sizes and allow the sampler to better explore the parameter space of the posterior density.

The other parameters that are set include the number of "tuned" samples, the number of drawn samples, and the number of sampled Markov chains. The tuned samples are the simulations that are discarded and serve only to move the sampler away from the starting values of the Markov chain. Sampling from more than one Markov chain allows for convergence checks of the simulation as will be discussed in the next section. Once the sampler has been tuned, the algorithm draws the specified number of samples for each chain declared. After drawing the number of specified samples, the drawn parameters from the separate Markov chains are joined together to form the sampled posterior density.

Throughout this paper, we set the above parameters as follows:

- `target_accept`: 0.95
- tuned samples: 3000
- drawn samples: 500
- number of chains: 4

3.5.2 Convergence

Once we have performed the inference, we are interested in whether or not the simulation has approximately converged to the posterior density. Using samples from a simulation that has not converged carries the danger of biased inference which can lead to incorrect conclusions about the model.

The main diagnostic check to determine if the sampler has approximately converged is the *Gelman-Rubin* statistic, denoted by $\hat{R}$. This is a measure of the between-sequence and within-sequence variance across the Markov chains and is used to determine if stationarity and mixing has been achieved.

Formally, for a model parameter ψ, suppose we have N simulations and let $\{\phi_{ij}\}_{i=1}^{N}$ for $j = 1 \dots M$ simulated chains. The between-sequence and within-sequence variance is computed as

$$B = \frac{N}{M-1} \sum_{j=1}^{M} \left(\frac{1}{N} \sum_{i=1}^{N} \psi_{ij} - \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} \psi_{ij} \right)^2$$

$$W = \frac{1}{M(N-1)} \sum_{j=1}^{M} \sum_{i=1}^{N} \left(\psi_{ij} - \frac{1}{N} \sum_{i=1}^{N} \psi_{ij} \right)^2$$

respectively [8]. Note that, $\mathrm{Var}(\psi|y)$, the marginal posterior variance of the estimand ψ, can be estimated by a weighted average of W and B, i. e.

$$\hat{\mathrm{Var}}^{+}(\psi|y) = \frac{N-1}{N} W + \frac{1}{N} B.$$

which overestimate the marginal posterior variance, but is unbiased under stationarity or in the limit $N \to \infty$. However, the within variance W should underestimate the marginal posterior variance since "the individual sequences have not had time to range over all of the target distribution". Thus, in the limit as $N \to \infty$,

$$\hat{R} = \sqrt{\frac{\hat{\mathrm{Var}}^{+}(\psi|y)}{W}} \to 1, \tag{3.2}$$

and we can check for approximate convergence by determining if $\hat{R}$ is close to 1 [8].

Another diagnostic check is the number of effective samples produced by the simulation, denoted by n_{eff}. Since the simulation draws within each sequence will have autocorrelation, the number of posterior samples is less than the simulation draws. Asymptotically, the effective sample size is

$$n_{\mathrm{eff}} = \frac{MN}{1 + 2\sum_{t=1}^{\infty} \rho_t} \tag{3.3}$$

where ρ_t is the autocorrelation of the simulation sequence at lag t [8]. The interested reader can consult Gelman et al. for a technical discussion of how to calculate, $\hat{n_{\mathrm{eff}}}$,

an estimate of n_{eff} for finite sequences. The recommendation from BDA3 is that $\hat{n_{\text{eff}}} > 10M$, i. e. there are at least 10 independent draws for each sequence.

For each network model, we have that $0.99 \leq \hat{R} \leq 1.01$ and $\hat{n_{\text{eff}}} > 400$ for all model parameters. Thus, we conclude that the chains have mixed, are stationary, and that we have enough parameter samples to use to perform inference.

3.6 Validation

"If the model fits, then replicated data under the model should look similar to observed data." [8] Generating data using the posterior density of a model and then checking some aspect of the generated data set is similar to the analogous aspect of the observed data is called a posterior predictive check.

Let y be the observed data, θ be the vector of parameters, and X be the model covariates. For model $\mathcal{M}$, we have that $\theta = (\pi, \omega, \kappa)$. Define y^{rep} to be the replicated data that could have generated given the vector of parameters θ obtained through inference, i. e.

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta. \tag{3.4}$$

Note that since we use the posterior density obtained through inference to generate the replicated data sets, the replicated values are necessarily dependent on the observed data just as the posterior density is.

The main purpose of posterior predictive checks is to measure the discrepancies between the data and the model and to determine if the discrepancies could have occurred by chance under the assumptions of the model [8, 13].

3.6.1 Replicated versus Actual Data Distributions

The first posterior predictive check we will perform is to assess whether the distribution of replicated data sets is similar to the distribution of observed data. We choose to graphically display these distributions for the assessment of the parameters m_A^{rep} vs m^A, the observed data, and $c^{\text{rep}} = \pi$ vs c, the probability of a success for a given trial. Note that for each model we draw 500 simulated values from posterior predictive distribution.

As can be seen from Figure 3.2, the replications m_A^{rep} that have been drawn from the posterior density reasonably mimic the observed data. Thus, we see that in this aspect, the model is a good fit for the observed data.

Similarly, we see from Figure 3.3 that the replications c^{rep} reasonably mimic the observed data with one notable exception; on the SPTS network, the replicated parameters c^{rep} do not capture the mode of 0 in the observed data set. This is one aspect of the data that we do not wish to capture as we expect the probability of success to be positive, if only very small. The reasoning behind this is that the mode of 0 is present only due to the right censoring of the data, an artifact of the small number of trials present per unit of observation on the network. Thus, we accept this particular misfit of the model to the observed data and from the graphical replications declare that the model fits reasonably well.

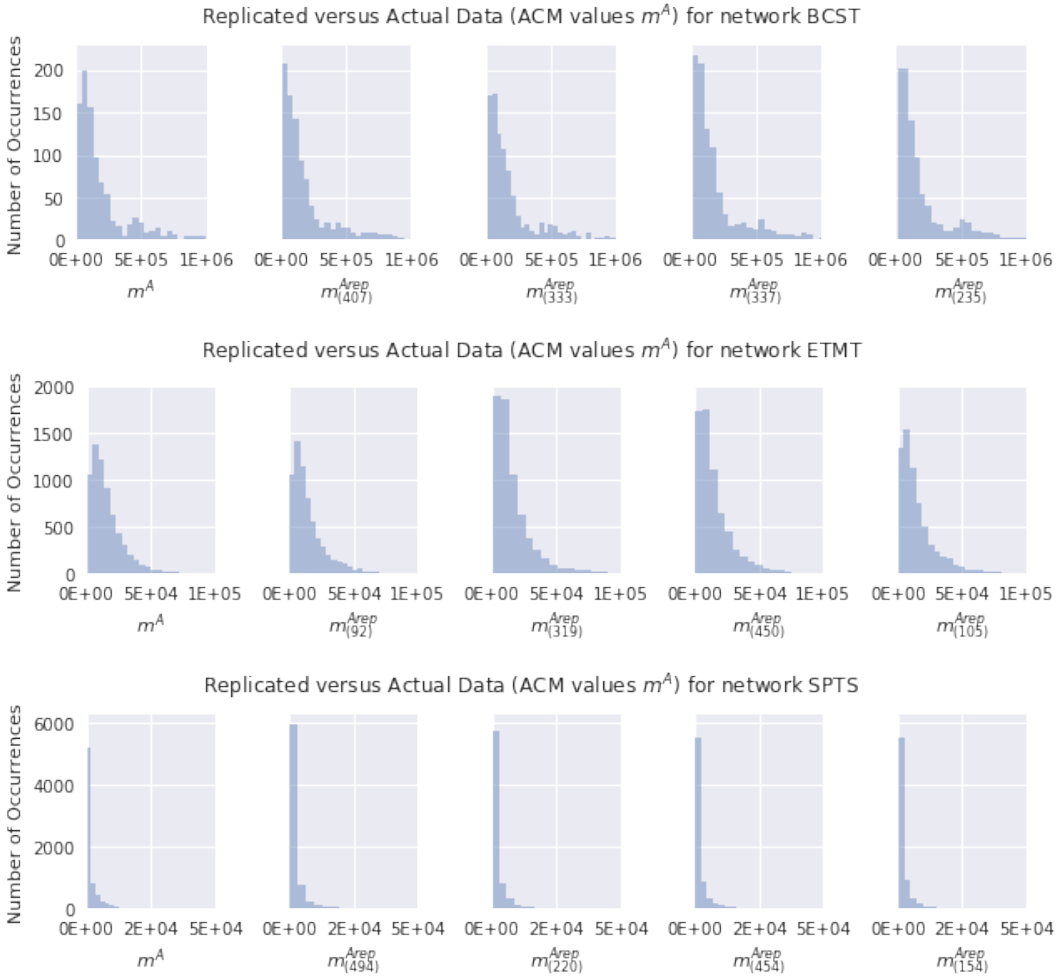

FIGURE 3.2: Actual m^A data (left) compared to replicated data sets $m^{A\text{rep}}$ (right four). Note that some parts of each distribution may be slightly truncated in order to display the critical features of the distribution. The replicated data sets largely mimic the actual data set.

3.6.2 Test Statistics

We can also more concretely quantify model discrepancies by defining a test quantity $T(y, \theta)$ and then measuring the discrepancy between the observed data and the replicated data.

Formally, we can compute a posterior predictive p-value defined as

$$p_B = \Pr\left(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y\right). \tag{3.5}$$

Values that exist in the extremes of 0 and 1 indicate poor model fit in regard to the particular aspect of the model that the test quantity aims to capture. As we have simulated values of the posterior density, the estimated p-value is just the proportion of the simulations such that the test quantity is greater than or equal to the test quantity as measured in the observed data, i. e.

$$\hat{p_B} = \frac{1}{S}\sum_{i=1}^{S}[T(y^{\text{rep}}_{(i)}, \theta_{(i)}) \geq T(y, \theta_{(i)})]$$

FIGURE 3.3: Actual c data (left) compared to the replicated data c^{rep} (right four). The replicated data sets largely mimic the actual data set with the exception of the replicated data sets of network SPTS.

for S simulations.

We define the following test quantities to use in evaluating the fit of model $\mathcal{M}$:

- $T_1(y, \theta) := \min(y)$,
- $T_2(y, \theta) := \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$,
- $T_3(y, \theta) := \max(y)$,
- $T_4(y, \theta) := \text{std}(y) = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N-1}}$,
- $T_5(y, \theta) := \overline{\pi} = \frac{1}{N} \sum_{i=1}^{N} \pi_i$.

In Table 3.1 we summarize the results of these test quantities when compared to the observed data. As can be seen from the table, for some test quantities, the estimated p-value lie in the extremes. This suggests that the model does not fit those aspects of the data. For instance, on the ETMT network, the mean value of the replicated data are all greater than the mean of the observed data.

Note that there are pathological examples where the estimated p-values are 1, but do not adequately measure the discrepancy between the replicated data, e. g. for

Test quantity	BCST network $T(y,\theta)$	BCST network 95% int. for $T(y^{\text{rep}},\theta)$	BCST network p_B	ETMT network $T(y,\theta)$	ETMT network 95% int. for $T(y^{\text{rep}},\theta)$	ETMT network p_B	SPTS network $T(y)$	SPTS network 95% int. for $T(y^{\text{rep}},\theta)$	SPTS network p_B
$T_1(y,\theta)$ (min)	3701	[6245, 14270]	0.99	0	[9, 182]	1.0	0	[0, 0]	1.0
$T_2(y,\theta)$ (mean)	227457.84	[2266852.49, 236367.09]	0.95	16357.80	[16705.39, 1748.11]	1.0	3972.45	[3714.91, 4559.66]	0.73
$T_3(y,\theta)$ (max)	4311038	[3443885, 4989241]	0.34	452762	[307901, 760822]	0.78	526816	[607186, 2239365]	0.99
$T_4(y,\theta)$ (std)	334052.86	[325128.37, 364859.10]	0.90	17686.89	[20021.24, 23205.09]	1.0	22300.18	[20012.59, 39808.44]	0.88
$T_5(y,\theta)$ (mean π)	0.105	[0.104, 0.107]	0.60	0.0903	[0.089, 0.092]	0.43	0.056	[0.065, 0.0689]	1.0

TABLE 3.1: Evaluation of Test Quantities across networks.

the test statistic T_1 on the SPTS network, the estimated p-value is 1, but the replicated data matches the observed data in regards to this test statistic. Thus, we ignore such cases in evaluating the model fit.

From these measured test quantities, we see that the model for BCST best fits the data and that the model does not adequately describe some aspects of the data for the other two networks. For the p-values of the test quantities that do lie in the extremes, the magnitude of the test quantity discrepancy is small. Thus, these models could be expanded to better accommodate the observed data, but we elect to use these models as is, noting that the model does not adequately describe all aspects of the observed data.

3.6.3 Regression Fit

The last check that we will perform is the analysis of the regression fit to the model. To evaluate this fit, we will analyze the standardized residuals of the model.

For a model with unknown parameters θ and predictors x_i, the *predicted* value is $\mathrm{E}(y_i|x_i,\theta)$ and the *residual* is $r_i = y_i - \mathrm{E}(y_i|x_i,\theta)$. The *standardized residual* is given by $r_i/\mathrm{std}(y)$ which allows us to ignore the magnitude of the data itself. Using the simulated posterior density, we can compute $\mathrm{E}(y_i|x_i,\theta)$ to be the mean of the replicated test, or hold-out, data itself.

We can evaluate the standardized residuals by graphically comparing the replicated data from the hold-out set and their realized standardized residuals versus the observed data and their standardized residuals in the hold-out set. As we can see from Figure 3.4, the standardized residuals have some misfits in the extremes of the data for the ETMT and SPTS network; as the replicated data becomes larger, so does the model's standardized residuals.

The misfit seen graphically can be quantified through the test quantity

$$T(y,\theta,x) = \frac{\bar{r}}{\mathrm{std}(y)},$$

which is the mean standardized residual of the data set. By proceeding as we did in Section 3.6.2, we can use 3.5 to calculate the estimated p-value of observing standardized residuals of the replicated test data set more extreme than in the observed data.

The graphical analysis of the above test quantities is shown in Figure 3.5. As we can see, for the BCST and SPTS networks, the standardized residuals of the replicated data sets do not differ in the extremes of the observed data set, while the standardized residuals do for the ETMT network. However, the magnitude of the discrepancy is small and we accept this deficiency of the model for the ETMT network. Note the results of the test quantities can be found in Table 3.2.

Network	$T(y,\theta,x)$	95% int. for $T(y^{\mathrm{rep}},\theta,x)$	p_B
BCST	-0.014	[-0.043, 0.041]	0.73
ETMT	-0.070	[-0.030, 0.036]	1.0
SPTS	-0.013	[-0.022, 0.018]	0.89

TABLE 3.2: Evaluation of test quantity $T(y,\theta,x)$ using the hold-out data set across networks.

FIGURE 3.4: Actual standardized residuals data (left) compared to the standardized residuals of the replicated data (right four). For the ETMT and SPTS network, the model's residuals become larger as the replicated becomes larger which is indicative of model misfit.

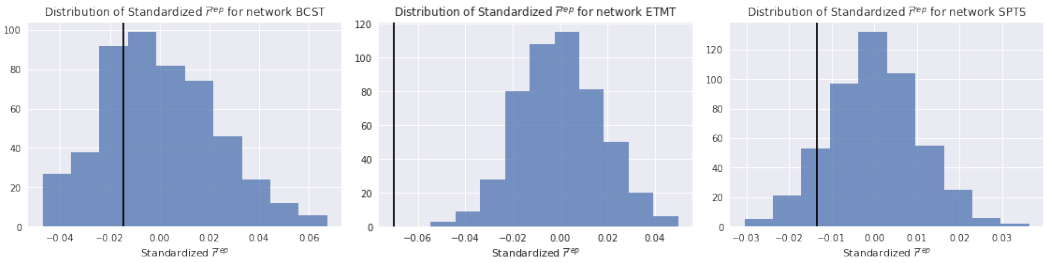

FIGURE 3.5: Graphical comparison of distribution of test quantities $T(y^{\text{rep}}, \theta, x)$ versus observed test quantity $T(y, \theta, x)$ (black) in hold-out data set across networks.

4. Results

Now that we have validated the model and identified its areas of discrepancies, we wish to evaluate the performance of the model in terms of its ability to generate forecasts. In order to make this evaluation, we will compare the performance of the model $\mathcal{M}$ to the baseline model $\mathcal{M}_0$ used throughout the TV industry modified from the literature.

Throughout this section, we perform the inference as described in the previous section and use the computed posterior density to generate the forecasts of the hold-out set. In order to remove any errors associated to inaccurate m^B forecasts, we use the observed m_i^B for each unit of observation from the hold-out set to generate the forecasts using model $\mathcal{M}$. We summarize the inferences produced by the probabilistic forecasts, as required in order to construct media plans, by taking their expected value when appropriate.

4.1 Industry Standard Model

Models used historically by the TV Industry rely on matching historical data to "previously broadcast programs on the basis of program genre and various attributes [5]." The model typically then is some average of the observed historical data.

To mirror this practice, we define the baseline industry model $\mathcal{M}_0$ as follows. Let x_i be the covariate vector of the units of observation in the hold-out set containing the content identifier and the stratified hour and let x_i' be the covariate vector containing only the stratified hour. Define the forecasted impression concentration of airing i to be the average c over the train set for each unit of observation that has the same covariate vector x_i. If no such average exists, i. e. the content has never aired before, define the forecasted impression concentration of airing i to be the average c over the train set for each unit of observation that has the same covariate vector x_i'. The forecasted m_i^A for airing i is then cm_i^B. This method ensures that we will generate a forecast for each unit of observation in the hold-out set and that we are generating the forecasts in a comparable way to model $\mathcal{M}$.

4.2 Predictive Accuracy

We wish to understand the predictive accuracy of model $\mathcal{M}$ in comparison to the baseline model $\mathcal{M}_0$ to see if the forecasts generated by model $\mathcal{M}$ are more accurate than those generated by model $\mathcal{M}_0$.

To evaluate this we will use for point estimates the measure of Mean Absolute Error (MAE) defined as $MAE = (1/N)\sum_{i=1}^{N} |y_i^{\text{act}} - y_i^{\text{pred}}|$ As we are generating probabilistic forecasts using model $\mathcal{M}$, it makes sense to evaluate their probabilities against the actual observed outcome. A generalization of the MAE exists for probabilistic forecasts known as the Continuous Ranked Probability Score (CRPS).

Let F be the cumulative distribution function of X, the forecasted quantity. If x is the observed value, define the CRPS between F and x as

$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - H(y - x))^2 \, dy \tag{4.1}$$

where H is the Heaviside step function [3]. As we are using a simulated probabilistic forecast, we can compute 4.1 by discretizing appropriately using quadrature rules.

To begin, we calculate the above error metrics of model $\mathcal{M}_0$ and $\mathcal{M}$ at the level of the units of observation in comparison to the observed data of the hold-out set. The results of these calculations are stored in Table 4.1.

network	level	$\overline{m_i^A}$	$\mathcal{M}_0$ MAE	$\mathcal{M}$ CRPS	$\mathcal{M}$ MAE
BCST	unit of observation	171450.06	23307.81	15247.91	21408.24
ETMT	unit of observation	13034.77	5123.70	3683.04	5226.42
SPTS	unit of observation	3164.84	1844.51	1426.60	1942.76

TABLE 4.1: Error metrics of the predicted data and the observed data in the hold-out set. Note that $\mathcal{M}$ MAE is computed using the point forecast of the probabilistic forecasts. The average value of m_i^A is presented to illustrate the magnitude of the errors.

From this table, we see that the errors of both models range in magnitude from 10-50% of the observed mean. When using the point forecasts of model $\mathcal{M}$, the model performs similarly to the base model $\mathcal{M}_0$ where the base model slightly outperforms model $\mathcal{M}$ on the cable networks (ETMT and SPTS) and does slightly worse than model $\mathcal{M}$ on the broadcast network BCST. However, we see that model $\mathcal{M}$ greatly outperforms the base model. Thus, in terms of this error metric, we conclude that model $\mathcal{M}$ has more predictive accuracy than the base model at the level of the units of observation.

In addition to the error metrics listed above, we are interested in whether or not the forecasted probability distributions correctly assign probabilities to the possible outcomes. That is, we are interested in knowing if our probabilistic forecasts are calibrated. To assess this, for each unit of observation we can calculate a $(1-\alpha)\%$ Credible Region (CR) for $\alpha = 0.5, 0.05, 0.01$ and then measure whether the observed outcome was within that CR. If the forecasts are perfectly calibrated, then the proportion of events that are within the $(1-\alpha)\%$ CR will be equal to $(1-\alpha)\%$. From a Bayesian point of view, a $(1-\alpha)\%$ Credible Region is the region of the posterior density that contains all events that have a $(1-\alpha)\%$ chance of occurring.

From table 4.2, we can see that the probabilistic forecasts for the units of observation are near perfectly calibrated for all networks except for SPTS. Upon closer inspection, we see that 89.66 % of the observed outcomes that are less than the lower limit of the $(1-\alpha)$ CR for $\alpha = 0.05$ have an associated lower limit of less than 10. This distribution of lower limits is shown in Figure 4.1. Thus, technically the observed outcomes are outside of the interval, but these lower limits are close enough to 0 for the practical purposes of TV planning. Removing the assumed right-censored data on this network, i. e. data in which $m_i^A = 0$, shows that 91.8% of units are within the 95% CR, a result consistent with the other networks. This information is useful as we will be able to forecast a range of possible outcomes for each unit of observation and be reasonably certain that the outcome will be within that range.

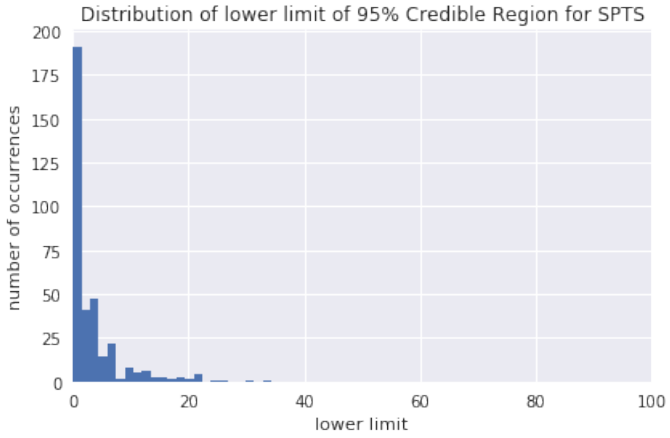

FIGURE 4.1: Distribution of the left end point of the 95% Credible Region for the SPTS network where the outcome was not within the Credible Region.

		Credible Region $(1-\alpha)\%$		
network	level	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
BCST	unit of observation	0.465	0.922	0.966
ETMT	unit of observation	0.48	0.91	0.95
SPTS	unit of observation	0.416	0.765	0.799

TABLE 4.2: Proportion of units of observation that are within the listed Credible Region.

4.3 Aggregated Predictive Accuracy

Now that we have evaluated the accuracy of the forecasts on the units of observation, we want to evaluate the accuracy of the forecasts on aggregations of those units in the form of media plans. Under the model assumptions, the units of observation are exchangeable and given the model parameters, they are independently and identically distributed. Thus, we can assume that the forecasts under the model are independent and we can create the distribution of aggregated impressions through simple sums of the distribution of outcomes at the units of observation.

We proceed as we did in the previous section by computing error metrics and Credible Regions of unit aggregations. Specifically, for each network, we quantile the units of observation in the hold-out set by the point forecasts of impression concentration for both models into 20 bins. We then compute the forecasted impressions for each quantile and compare to the actual impressions observed. For model $\mathcal{M}_0$ this is merely the sum of the point estimates, but for model $\mathcal{M}$, this is the mean of the distribution of sums of impressions for the quantile.

The graphical results of this analysis are shown in Figure 4.2. From this graph we see that there is less dispersion in the forecasted versus actual values for model $\mathcal{M}$ compared to model $\mathcal{M}_0$, especially for the SPTS network. However, graphs can sometime be misleading and so we analyze the resulting error metrics.

Tables 4.3 and 4.4 show that the same general story arises when looking at the quantiled units in aggregate as when looking at the units of observation; model $\mathcal{M}_0$ performs better than model $\mathcal{M}$ when only looking at the point forecasts, but model $\mathcal{M}$ outperforms when using the probabilistic forecasts, drastically so on the SPTS network. One notable exception is that the aggregation of units on ETMT do not

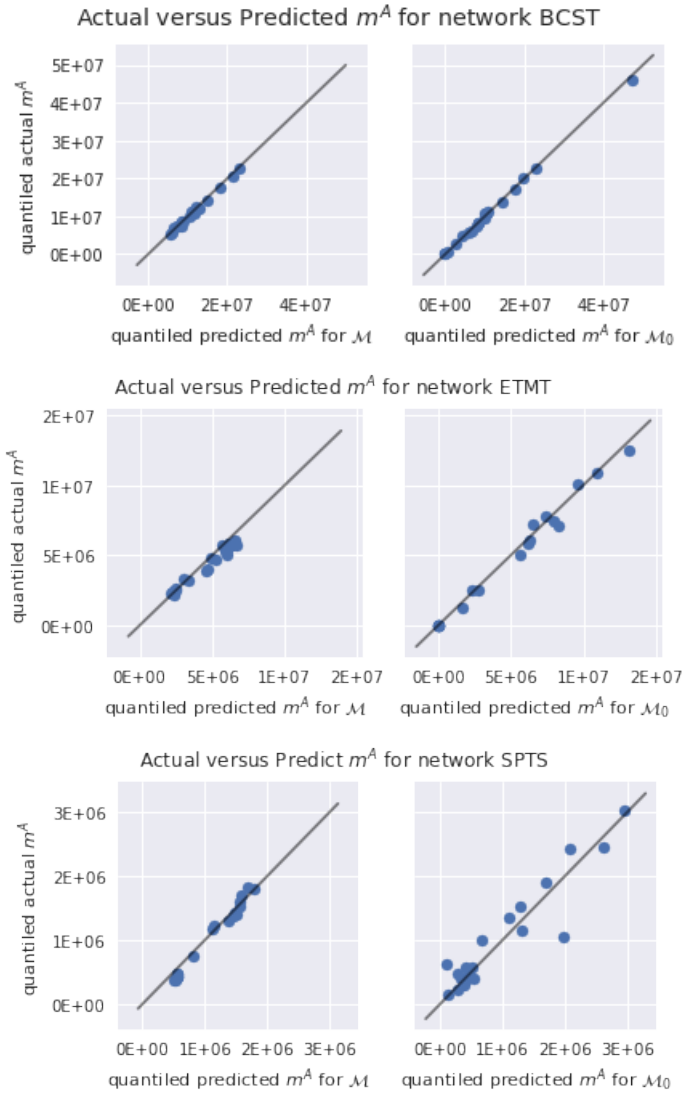

FIGURE 4.2: Comparison of forecasts versus actuals for quantiled units of observation across the three networks for both models.

network	level	$\mathcal{M}_0$ MAE	$\mathcal{M}$ CRPS	$\mathcal{M}$ MAE
BCST	quantiled	324034.14	277437.29	417797.60
ETMT	quantiled	297512.52	311316.38	399648.10
SPTS	quantiled	205096.15	75962.55	90668.35

TABLE 4.3: Error metrics of the predicted data and the observed data in the hold-out set for the quantiled units of observation across the three networks.

seem to have the same performance in regard to error metrics and as the individual units themselves. This suggests that some model assumptions must not hold with this network and that the model needs to be re-evaluated. This is in-line with the previous model checks that showed misfit with this particular model.

		Credible Region $(1-\alpha)\%$		
network	level	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
BCST	quantiled	0.45	0.95	1.0
ETMT	quantiled	0.1	0.55	0.75
SPTS	quantiled	0.4	0.65	0.65

TABLE 4.4: Proportion of quantiles that are within the listed Credible Region.

5. Conclusion

[8]

References

[1] George A. Barnet et al. “Seasonality in Television Viewing: A Mathematical Model of Cultural Processes”. In: *Communication Research* 18 (1991), pp. 755–772.

[2] *Broadcast Calendar / Nielsen Survey Dates*. Accessed: 2018-04-03.

[3] Jochen Bröcker. “Evaluating raw ensembles with the continuous ranked probability score”. In: *Quarterly Journal of the Royal Meteorological Society* 138 (2012), pp. 1611–1617.

[4] Peter J. Danaher and Tracey S. Dagger. “Using a nested logit model to forecast television ratings”. In: *International Journal of Forecasting* 28 (2012), pp. 607–622.

[5] Peter J. Danaher, Tracey S. Dagger, and Michael S. Smith. “Forecasting television ratings”. In: *International Journal of Forecasting* 27 (2011), pp. 1215–1240.

[6] Andrew Gelman. “Prior distributions for variance parameters in hierarchical models”. In: *Bayesian Analysis* 1 (2006), pp. 515–533.

[7] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. First. Cambridge, 2007.

[8] Andrew Gelman et al. *Bayesian Data Analysis*. Third. CRC Press, 2013.

[9] Dennis Gensch and Paul Shaman. “Models of Competitive Television Ratings”. In: *Journal of Marketing Research* 17 (1980), pp. 307–315.

[10] J. Ghosh, Y. Li, and R. Mitra. “On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression”. In: *ArXiv e-prints* (July 2015). arXiv: 1507.07170 [stat.ME].

[11] Joyee Ghosh, Yingbo Li, and Robin Mitra. “On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression”. In: *Journal of Marketing Research* 17 (1980), pp. 307–315.

[12] Salvatier J., Wiecki T.V., and Fonnesbeck C. “Probabilistic programming in Python using PyMC3.” In: *PeerJ Computer Science* 2:e55 (2016). DOI: "10.7717/peerj-cs.55".

[13] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Second. Elsevier, 2015.