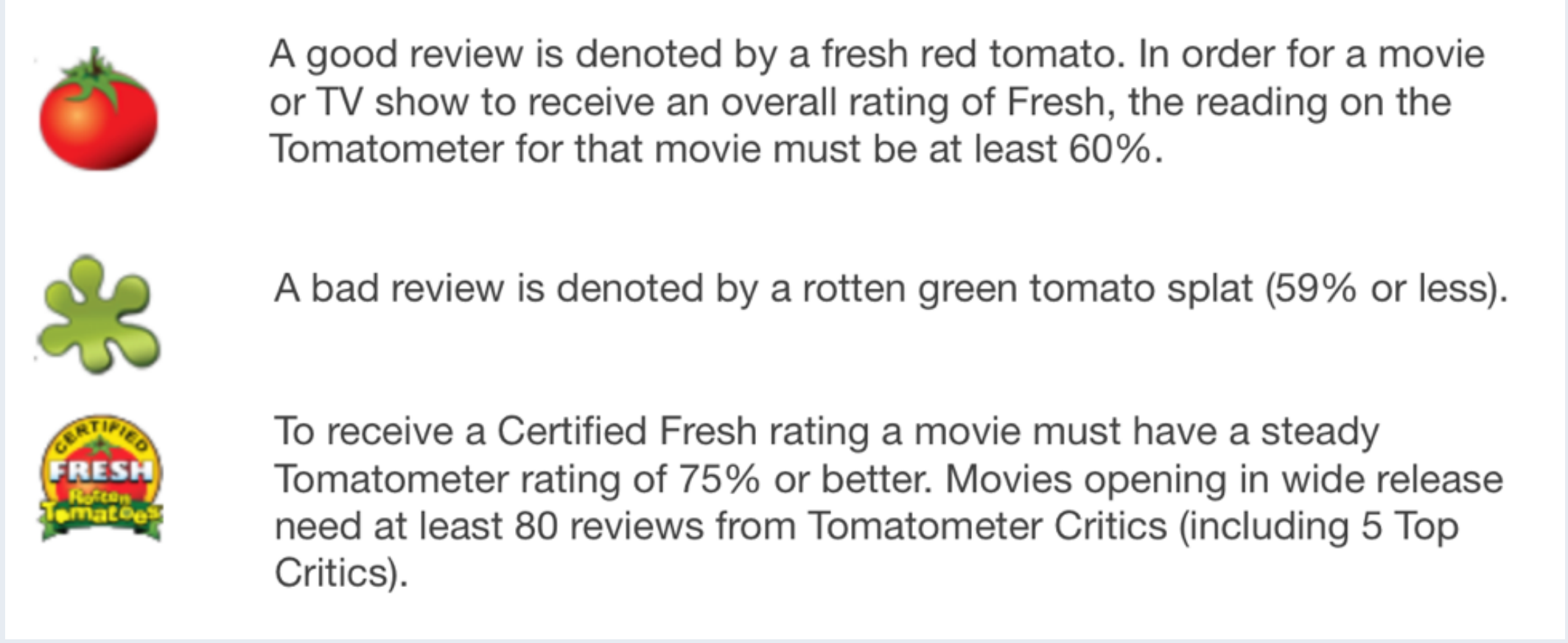
Movie Ratings Prediction by Bayesian Model

Xingmeng Zhao

Department of Mathematics, University of Colorado Denver

The Big Question

How can we tell the greatness of a movie when it released in cinema?



A good review is denoted by a fresh red tomato. In order for a movie or TV show to receive an overall rating of Fresh, the reading on the Tomatometer for that movie must be at least 60%.

A bad review is denoted by a rotten green tomato splat (59% or less).

To receive a Certified Fresh rating a movie must have a steady Tomatometer rating of 75% or better. Movies opening in wide release need at least 80 reviews from Tomatometer Critics (including 5 Top Critics).

Data Description

These dataset were obtained from IMDB and Rotten Tomatoes. The data represent 456 randomly sampled movies released between 1972 to 2014 in the Unites States. This data frame contains 456 observations (rows), each representing a movie, and 27 variables (columns):

- **audience_score** : Audience score on Rotten Tomatoes (response variable)
- **type**: Type of movie (Documentary, Feature Film, TV Movie)
- **genre**: Genre of movie (Action Adventure, Comedy, Documentary, Drama, Horror, Mystery Suspense, Other)
- **runtime** : Runtime of movie (in minutes)
- **year** : Year the movie is released
- **critics_score** : Critics score on Rotten Tomatoes
- **audience_rating** : Categorical variable for audience rating on Rotten Tomatoes (Spilled, Upright)
- **imdb_rate**: the movie rate score on IMDB
- **director** : Director of the movie

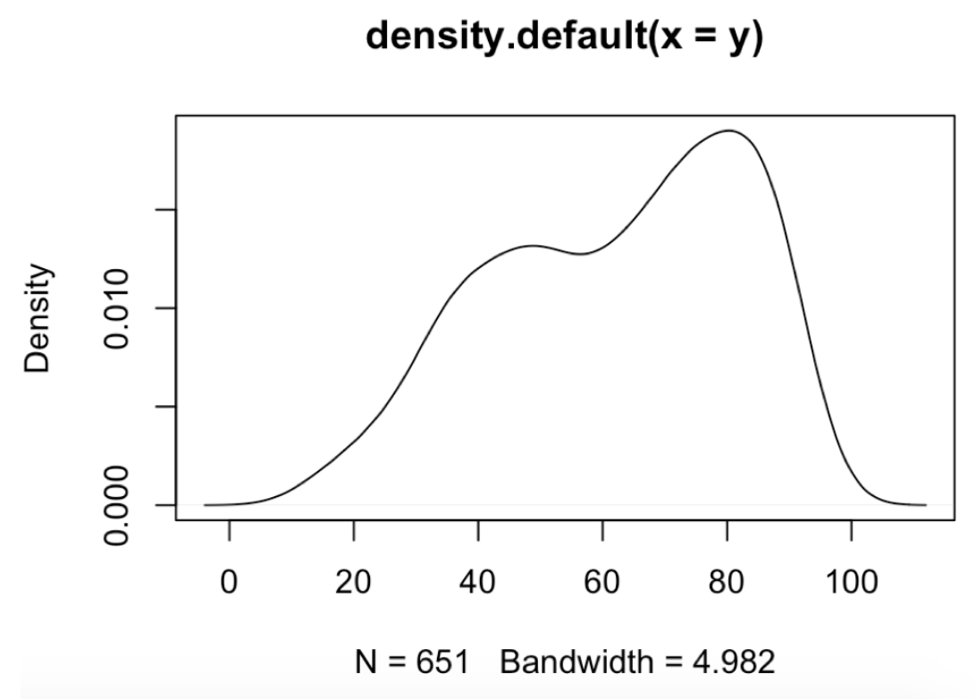
Model Choosing

Based on the response type is Count Data and responses defined on 0, 1, 2, ..., 100, we try to use Poisson regression models known as Poisson log-linear models.

Model Selection

To build the predictive model, we began with all the requested variables and used BIC, WAIC and LOOIC criteria to find the optimal variables and optimal model to include.

- First, at begining we have 27 variables and some of them are character variables, like the name of movie. Therefore, we kicked of all these character variables. And then, we have 16 numerical predictors.
- Based on BIC criteria, we use Generalized Linear Models(glm) backward selection to get the best subset which finally includes 7 predictors, for example, imdbrate, audience, runtime, genre, year, imdbvote, critics, audience.
- Using WAIC and LOOIC criteria, we calculate the most combination of these predictors. Finally, we found y ∼ imdbrate * audience + runtime + genre + year. This combination has the lowest LOOIC and WAIC value which is equal to 4353.5.



Poisson Regression Models

We know that The Poisson log-linear model is summarized by the expression:

$Y_i \mid \lambda_i \sim \text{Poisson}(\lambda_i)$

with $\log \lambda_i = \beta_0 + \sum_{i=1}^{p} \beta_j X_{ij} = x_i^T \beta$

- **Data distribution** :

  $audience\_score_i \mid \lambda_i \sim \text{Poisson}(\lambda_i)$

  with $\log \lambda_i = \beta_0 + \beta_1 imdbrate_i + \beta_2 runtime_i + \beta_3 genre_i + \beta_4 year_i + alpha2 * audience_i + delta2 * imdbrate_i * audience_i$
- **Prior distributions** :

  $\beta_j \sim N(0, 25)$, j= 0,1,2,3,4.

  $alpha2 \sim N(0, 25)$; $delta2 \sim N(0, 25)$.

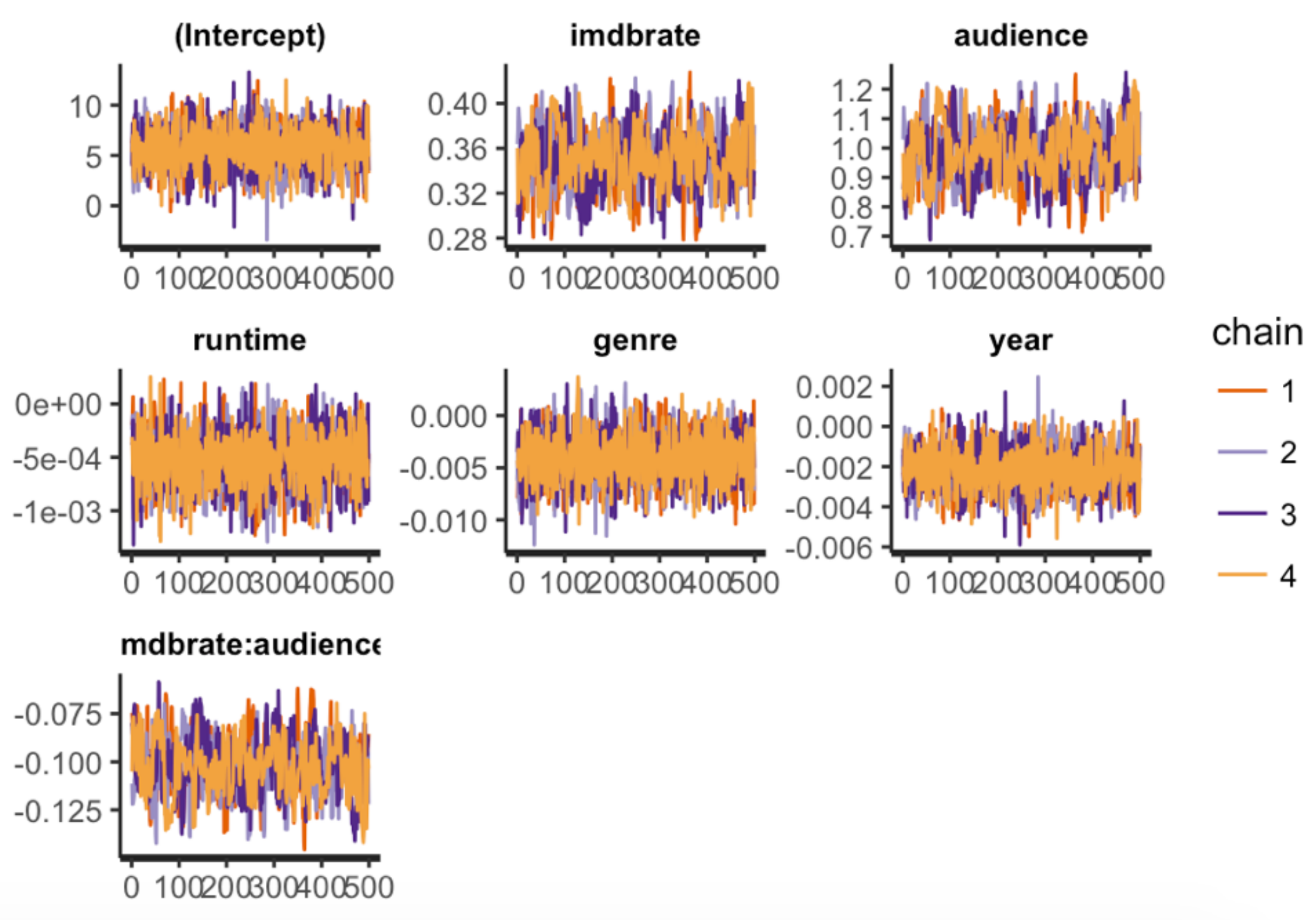


Figure 1: The trace plots of MCMC chain

Prediction Model

A point estimate of the model based on the posterior means is

$log\lambda_i = 0.85 + 0.34\ imdbrate_i - 0.0005\ runtime_i - 0.005\ genre_i + 0.0003\ year_i + 0.96\ audience_i - 0.10\ imdbrate_i * audience_i$

Model Checking

We need check the fit of our posterior model to the data and to our substantive knowledge. Because a poor model will lead to erroneous conclusions.

- Using pp_check function to generates graphical (top left plot) comparisons of the data y and replicated datasets yrep.
- Using top right plot to compare density curves for replicated samples from $p(y^{rep} \mid y)$ to the observed data.

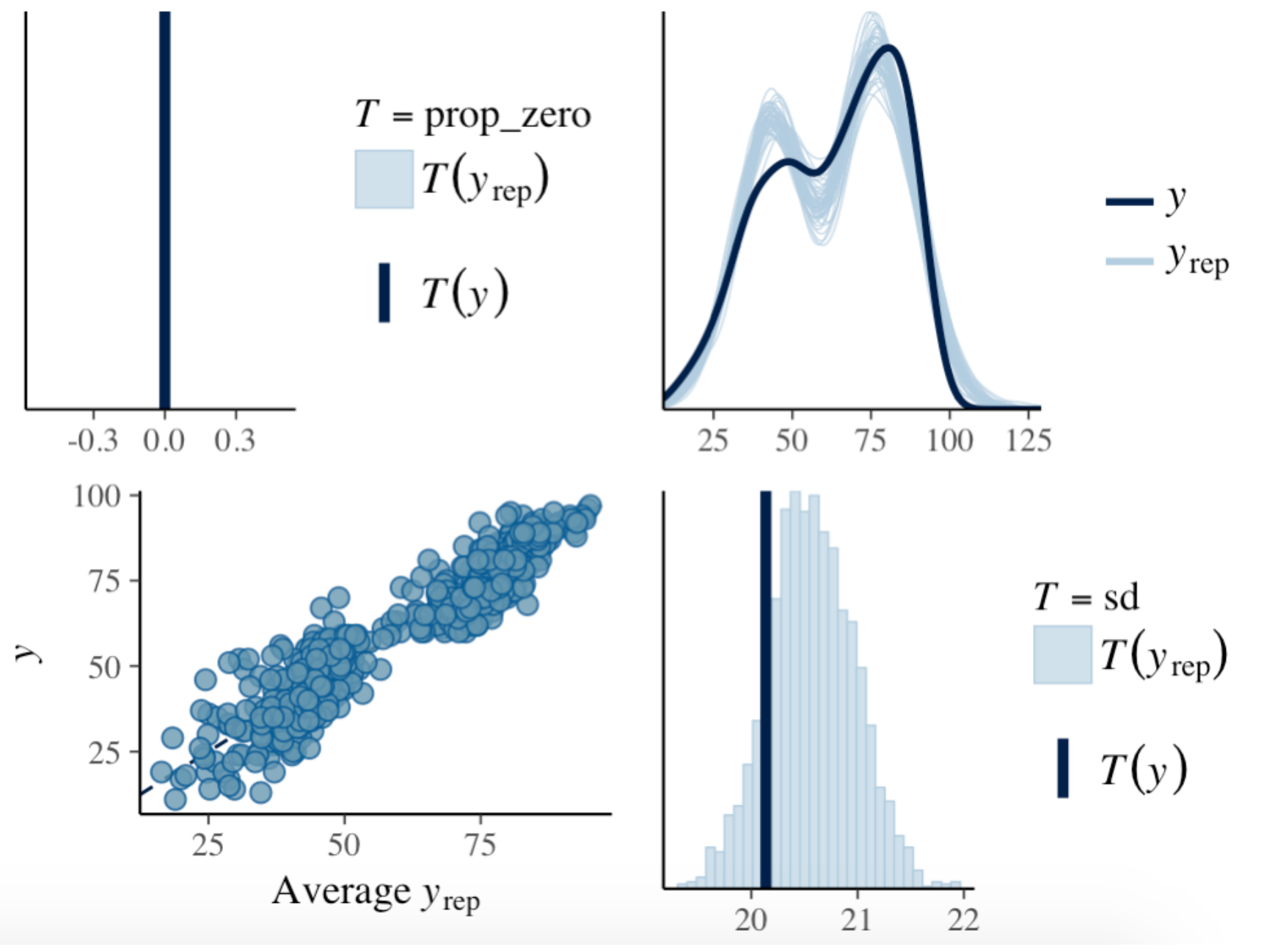
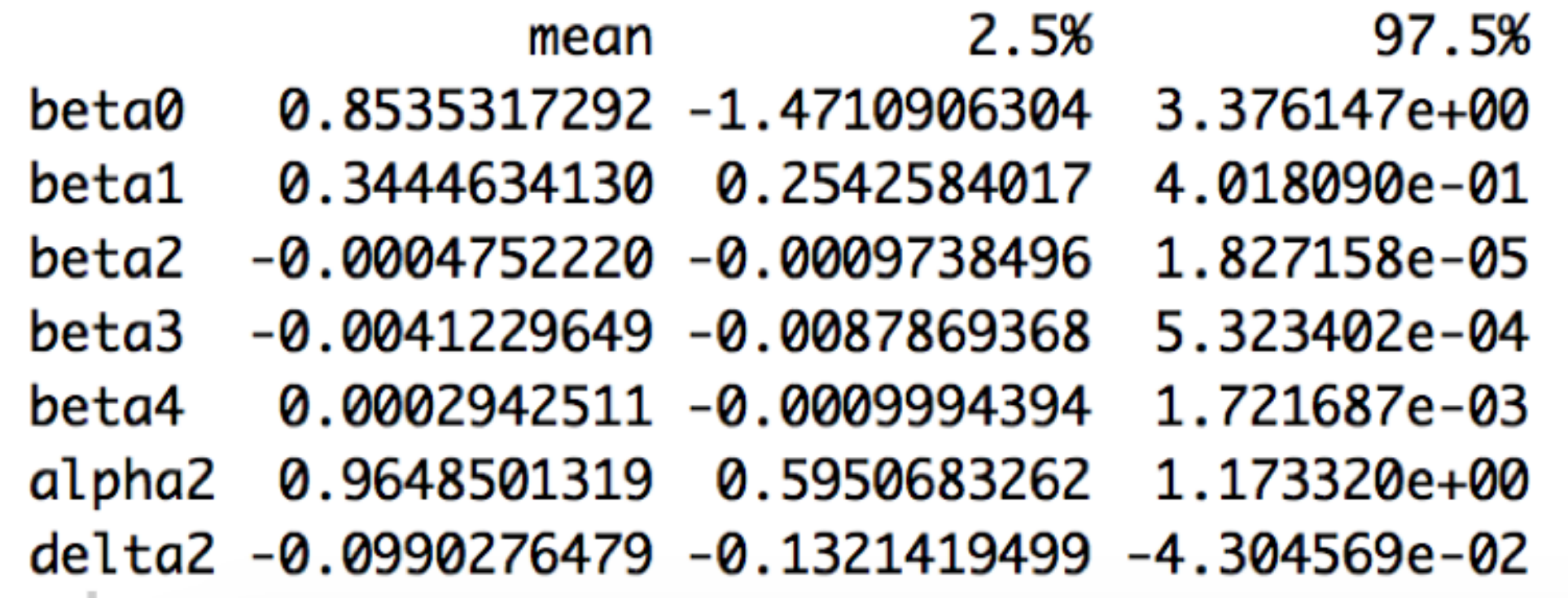


Figure 2: The series plots of model checking

Conclusions and Interpretation

Results for regression coefficients and 95% confident interval.



```
             mean          2.5%        97.5%
beta0   0.8535317292 -1.4710906304  3.376147e+00
beta1   0.3444634130  0.2542584017  4.018090e-01
beta2  -0.0004752220 -0.0009738496  1.827158e-05
beta3  -0.0041229649 -0.0087869368  5.323402e-04
beta4   0.0002942511 -0.0009994394  1.721687e-03
alpha2  0.9648501319  0.5950683262  1.173320e+00
delta2 -0.0990276479 -0.1321419499 -4.304569e-02
```

Figure 3: posterior means and 95 percent central posterior intervals

Calculate exponentiated regression coefficients is:

exp_beta0= 2.35, exp_beta1 =1.41, exp_beta2= 0.99, exp_beta3=0.99, exp_beta4=1.00, exp_alpha2=2.62, exp_delta2=0.91

Interpreting Results:

- The expected Audience score on Rotten Tomatoes for Spilled is about two and half as many as the Upright if other coefficients are same.
- Each minute of movie runtime increasing will reduce the expected Audience score by about 1%, assuming the movie type and other features do not change.
- Every unit of imdb rate increasing and the audience rate is Upright will increase the expected Audience score by about 41%, assuming the movie type and other features do not change.
- Every unit of imdb rate increasing and the audience rate is Spilled will reduces the expected of Audience score by about 10%, assuming the movie type and other features do not change.

References

[1] The 'Roger-Ebertron': Modeling Roger Ebert's Movie Ratings by Ozzie Liu.

[2] https://www.rottentomatoes.com/about/.

[3] https://cran.r-project.org/web/packages/rstanarm/vignettes/count.html.

[4] Bayesian Modeling and Prediction for Movies by Alex Knorr.

[5] http://www2.stat.duke.edu/ mc301/data/movies.html.