A Bayesian Analysis of Casino Generated Trips

Kathryn Haglich

12/8/2019

Introduction

When developments are planned and built, the Massachusetts Department of Transportation (DOT) requires developers to analyze changes in traffic flow caused by the addition of the new developments. This is to ensure that the surrounding infrastructure can support the trips generated by the development. Utilizing current trips as a baseline, predictions are made based on characteristics of the development (square footage, development type, etc.) which become the inputs of models from the Institute of Transportation Engineers (ITE) manual. These models are national standards with the models built on a national scope heavily influenced by auto-centric states. These models overestimate predictions for the more pedestrian and public transportation focused state of Massachusetts.

This problem is especially apparent when estimating trips generated by casinos particularly in Massachusetts. The ITE model extrapolates when analyzing Massachusetts casino data and overestimates the trips generated by these establishments. Other attempts to improve the ITE models are based on data collected from casinos in Las Vegas, and still fail to model accurately the data of interest. However, these models do indicate that linear regression models comparing the number of trips generated to the casino's total number of gaming positions are appropriate. Participants of the Masters of Science in Statistical Practice (MSSP) program from Boston University have analyzed this inquiry and produced linear regression models significantly improving trip generation predictions using traditional frequentists approaches.

The goal of this report is to perform a full Bayesian analysis to explore the relationship between the number of gaming positions and the total amount of trips generated by casinos. In doing so, predictions can be improved and interpreted from a a more holistic and probabilistic perspective, as demonstrated with a case study focused on Encore Boston Harbor. By improving the models and producing more accurate traffic estimates, the DOT can ensure that infrastructure to new casinos can be designed more efficiently and mitigate environmental impact and the developer's economic costs.

Background

The Massachusetts DOT is interested in understanding the relationship between various casino attributes and the trips generated by the casino. *Generated trips* is defined as the number of vehicles that are going to or coming from the development. Prior modeling has determined that the number of *gaming positions* is an excellent predictor of generated trips. This metric is defined as the sum of the casino's total number of slot machines plus seven times the total number of tables as an estimation for seat positions. Using the same data set as this report, the MSSP collaborators explored the following various linear regression models:

$$y_{sat} = 167.53 + 1.04x_{fri} \tag{1}$$

$$y_{sat} = 1.12x_{fri} \tag{2}$$

$$y_{trips} = 40.46 + 31.41x_{gp} - 117.56x_{sat} + 5.74x_{gp}x_{sat} \tag{3}$$

$$y_{FriTrips} = 32.16x_{gp} \tag{4}$$

Equation 3 could be parsed into two equations for Friday and Saturday, respectively: $y = 40.46 + 31.41x_{gp}$ and $y = -77.10 + 37.15x_{gp}$. From equations 1 and 2, it was concluded that there is no difference between the number of trips generated on Friday and the number of trips generated on Saturday. These findings were supported by equation 3 with the gaming position coefficient the only significant coefficient as reported in the literature. Finally, their analysis determined that equation 4 is the most appropriate model for this data and for predictions.

The data was provided by the Massachusetts DOT to the MSSP collaborators, one of whom was allowed to use the data set for this project. However, the data limited the analysis to exploring the relationships between number of gaming positions, day of the week (predictors), and total trips generated (outcome). The gaming position predictor variable was scaled by dividing by 100 to ease the interpretation of the variable's coefficient. Additionally, previous literature has determined that the observation for Pocono Manor in Pennsylvania (8000 gaming positions and 6420 total trips generated on Saturday) was an outlier that should be removed from further modeling.

Bayesian Methods

To begin establishing the Bayesian linear regression model, it is assumed that all observations are independent with constant variance σ^2. Thus, the likelihood of the outcome variable ($y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$) given the coefficient matrix, β, σ^2, and the predictor variables of interest $X_{n\times p}$, is the following distribution:

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I_n)$$

where n is the number of observations and p is the number of variables. Flat priors were implemented because there was no previous literature or experience that would gave solid evidence to establish the contrary.

$$\beta|\sigma^2 \sim N(\beta_0, \sigma^2\Sigma_0)$$
$$\sigma^2 \sim Inv\chi^2(\nu, \tau^2)$$

The resulting posterior for $\beta, \sigma^2|y, X$ is not in closed form and cannot be answered analytically. Instead, the Expectation Maximization (EM) algorithm that finds the Maximum A Posterior (MAP) values is implemented within a Gibbs sampler.

Traditionally, the EM is an iterative process that goes through a consecutive sequence of parameter estimates that improves the marginal likelihood with each pass until it reaches convergence. This optimization algorithm returns point values that maximize the likelihood. Mathematically, this is represented as

$$(\beta, \sigma^2)_{MAP} = argmax_{\beta,\sigma^2}\{log(P(\beta, \sigma^2|y))\}$$

The initial values of $RSS(\beta) = \sum(y - \bar{y})^2$ and $\hat{\sigma}_0^2 = \frac{RSS(\beta)+\nu\tau^2}{n+\tau}$. At iteration t,

$$\hat{\beta}_t = (\frac{1}{\hat{\sigma}_t^2}X^TX + \Sigma_0^{-1})^{-1}(\frac{1}{\hat{\sigma}_t^2}X^Ty + \Sigma_0^{-1}\beta_0)$$

$$\hat{\sigma}_t^2 = \frac{RSS(\beta_t) + \nu\tau^2}{n + \tau}$$

While calculating $\hat{\sigma}_t^2$ is straightforward, determining $\hat{\beta}_t$ is more challenging and requires the Cholesky decomposition to simplify the computations.

Since this is a linear regression model, the criteria for convergence is determined by the residual sum of squares (RSS):

$$\epsilon > \left| \frac{RSS(\beta_t) - RSS(\beta_{t-1})}{RSS(\beta)} \right|$$

where, in this case, $\epsilon = 1 \times 10^{-8}$.

When incorporated with the Gibbs sample, the convergence criterion is ignored; all values of β and σ are stored, and the algorithm continues for the specified number of iterations and chains (2000 and 4, respectively, for this project).

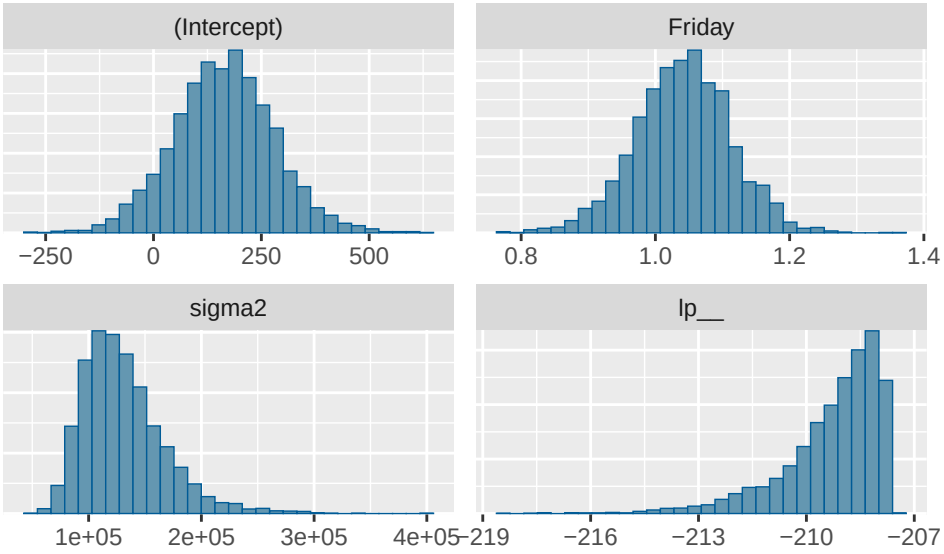
Models

The following models are Bayesian equivalences to the frequentist regressions represented by equations 1-4. All simulations ran for 2000 iterations and 4 chains with a burn-in period of 1000 iterations, and the convergence plots can be found in the appendix.

Similarities between Friday and Saturday Trips

First, a linear model was created that regressed the total number of trips on Saturday against the total number of trips on Friday to explore any initial differences between tips generated on the different days. No difference would indicate that the number of trips does not depend on the day of the week.

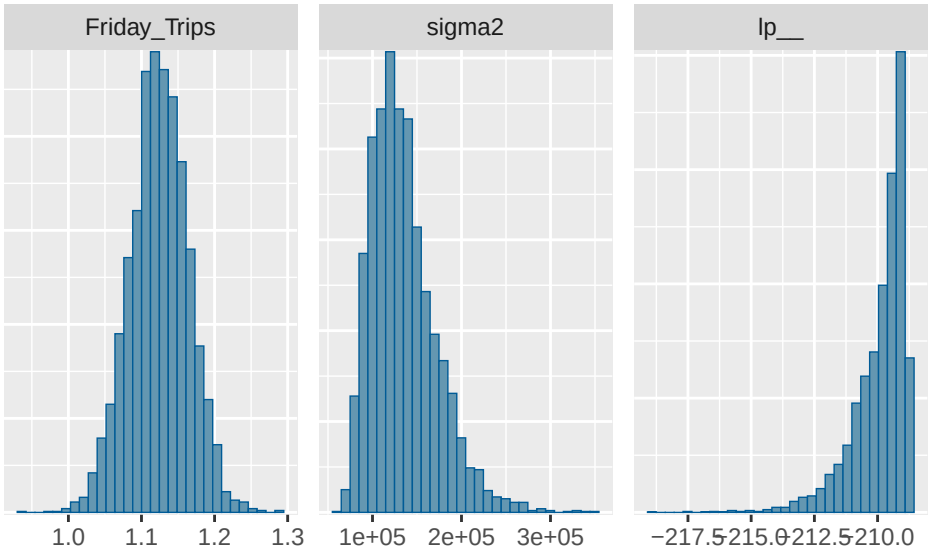
	Coefficient	Standard Error
(Intercept)	167.53	109.44
Friday	1.04	0.07



The distribution of the slope coefficient is extremely close to one with a mean centered at 1.04 and 0.07 standard error. This one to one relationship indicate that there is no difference between trips generated on Friday and trips generated on Saturday implying independence between the days.

The distribution of the intercept has a mean of 167.53 and standard error of 109.44. In the context of the problem, it is illogical in the context of the problem to have a nonzero intercept. In relation to the scale of possible trip values (which range from 225 to 4000), an argument can be made that an intercept with the given distribution is not necessary to include in the model, as further analysis shows. Thus, when a second model is fitted without the intercept, the coefficient has the following distribution:

	Coefficient	Standard Error
Friday_Trips	1.12	0.04

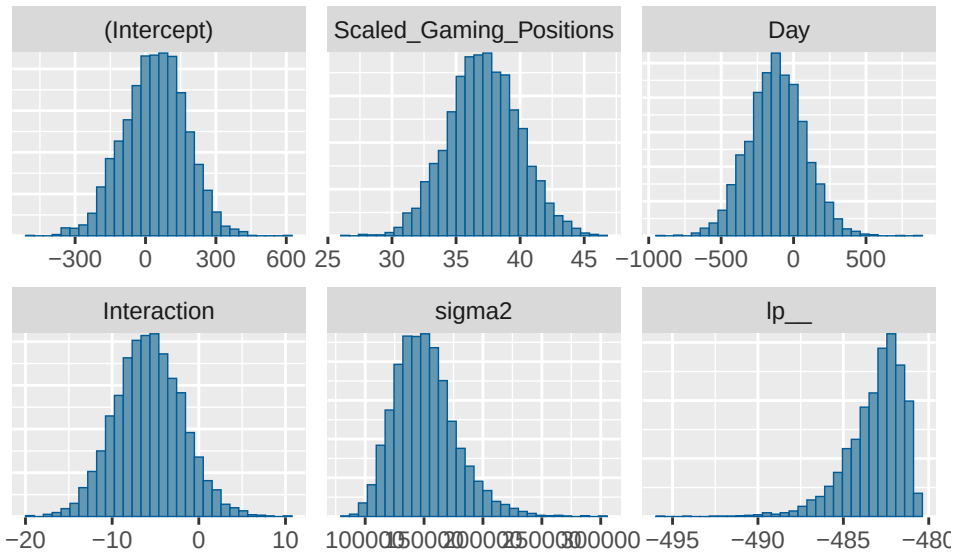


The coefficient is 1.12 with the density concentrated between 1 and 1.12. This model confirms what was shown previously: that the trips generated on different days are independent of one another. Thus, there is no need for any special considerations as to whether the remaining analysis is performed on trips measured on Friday or Saturday. The number of trips generated by a casino on Friday is expected to be similar to the number of trips generated by the same casino on Saturday. This is in accordance with and draws the same conclusions as the frequentist models represented by equations 1 and 2.

Regression of Total Trips on Gaming Positions and Day

For the following regression models, the data was adapted so that two variables were created: a binary variable to indicate the day of interest Friday (0) or Saturday (1), and a numerical variable to indicate the number of trips for that day. Thus, there are two observations for each casino, and the previous scaling for trips was preserved. An interaction term between the day and the number of gaming positions was also calculated and included in the model.

	Coefficient	Standard Error
(Intercept)	40.46	126.59
Scaled_Gaming_Positions	37.15	2.68
Day	-117.56	188.17
Interaction	-5.74	3.75

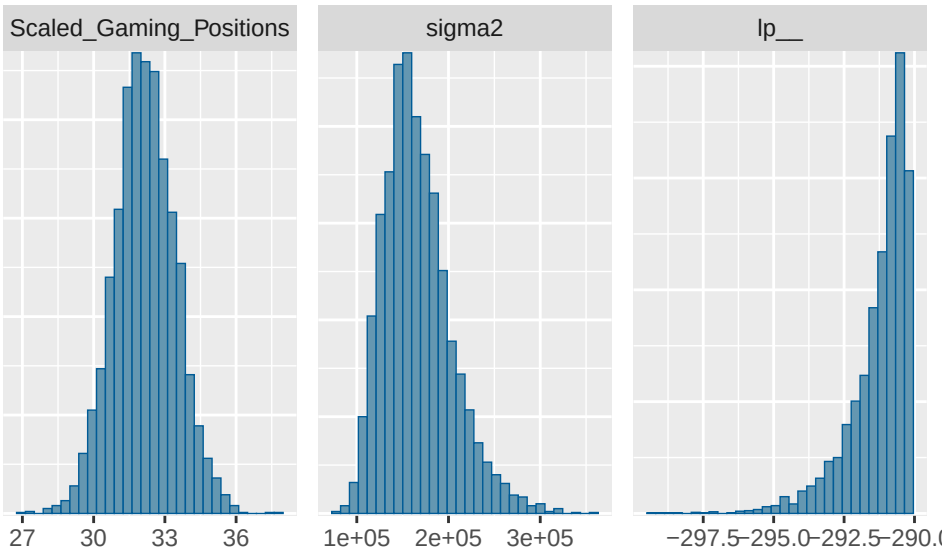


The intercept, while nonzero, is within one standard error away from 0, and the density shows a high likelihood of the intercept being extremely close to 0. This confirms our previous assertions that it would not be of best interest to include the intercept in the model. A similar story can be told for the day and interaction coefficients. The densities show non-zero probabilities that the these coefficients could be negative or positive which suggests that a concrete relationship between these variables and the total trips generated. This gives evidence that the day of the week and the interaction between day and gaming positions are not needed in the model.

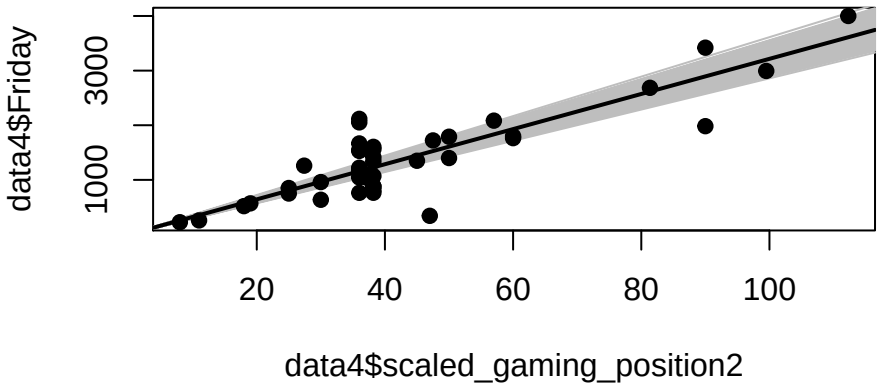
Regression of Friday Trips on Gaming Positions

Logically following the analysis from the previous models, there is enough evidence to suggest that the most appropriate models the total number of trips generated on the number of gaming positions while fixing the intercept at 0. This results in the following model parameters:

	Coefficient	Standard Error
Scaled_Gaming_Positions	32.16	1.23



With a mean coefficient value of 32.16 and standard error of 1.23, this estimated coefficient is identical to the 32.16 estimated frequentist value in equation 4. The histogram shows that the density is concentrated between 30 and 35 indicating that it is highly likely that the true value of the slope coefficient lies in this range.



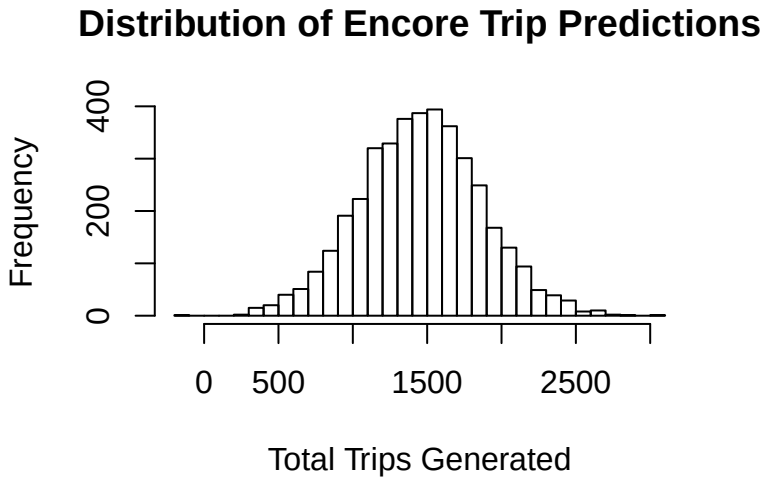
The graph above visualizes the regression line with the mean estimated coefficient overlaying the regression lines generated by all values of β from the Gibbs sampling. As a result of fixing the intercept at 0, the differing regression models give fairly similar predictions for casinos with fewer gaming positions, and the uncertainty increases as the number of gaming positions increases.

Encore Predictions

The Department of Transportation's primary interest in this project is to predict the trips generated by Encore Boston Harbor, which has 4515 gaming positions. The frequentist equivalent of model 4 that the number of trips generated is 1453 trips with a 95 % confidence interval of $(1,342, 1,563)$. To provide a Bayesian perspective, predicted gaming positions was done by first sampling β and σ values from their respective distributions as determined by model 4. (Computationally, β and σ were randomly drawn from the values from the Gibbs sampler 4000 times with replacement.) Then, for each β, σ pair, a y value was drawn from the following distribution:

$$y \sim N(45.15\beta, \sigma^2)$$

The average value of the y samples is 1456.984 with a standard deviation of 409.3236. The complete distribution is displayed below.



This shows that density is concentrated primarily between 1300 and 1600 trips, which is consistent with the frequentist confidence interval. However, the 95% predictive interval is much wider, $(654.71, 2259.258)$ as a result of the standard deviations. Additionally, the mean of the predicted y values is within 4 trips of the frequentist prediction. This indicates that predicting, and then observing, value for trips generated by Encore around 1460 is intuitively acceptable prediction.

Conclusions

Overall, the strong concordance between this Bayesian analysis and prior frequentist models indicates that these are the appropriate approaches to modeling this data. The first three models dictate that the most logical regression model is one where the only predictor is the number of gaming positions and the exclusion of the day of the week as a parameter. The last model supports this with an estimated mean slope value equal to the frequentist estimate, which intuitively suggests that this value is an extremely good estimate for the coefficient. In the context of the project, this indicates that within probabilistic uncertainty, for every 100 unit increase in the number of gaming positions, the number of trips generated by the casino is likely to increase by 32-33 trips. This is seen in the application of the last model to the Encore data point: the distribution is concentrated with the mean predicted value extremely close to the frequentist point estimate.

Not only does the Bayesian approach confirm the frequentist analysis, it provides the Mass DOT with an understanding of the models' parameter and prediction distributions. While the true value of the number of trips generated by Encore is unknown to the author of this report, the Bayesian distributions can provide an explanation for variations between the predictions and the observed trips. This is a flexibility that a point estimate cannot provide and can make it possible for the DOT to account for the natural model uncertainties in their decisions. And those decions, in accordance to the model, will be able to insure that the raods around the development are deigned to be safe, environmentally beneficial, cost-effective, and sound to handle the generated trips.

Future Research

Future research should include a more thorough analysis of variable selection and model comparison. Due to time constraints and technical difficulties, a formal analysis implementing Bayes Factor to justify model selection and the exclusion of the intercept, day variable and interaction variables was not completed. This would be the next logical steps to a more thorough analysis.

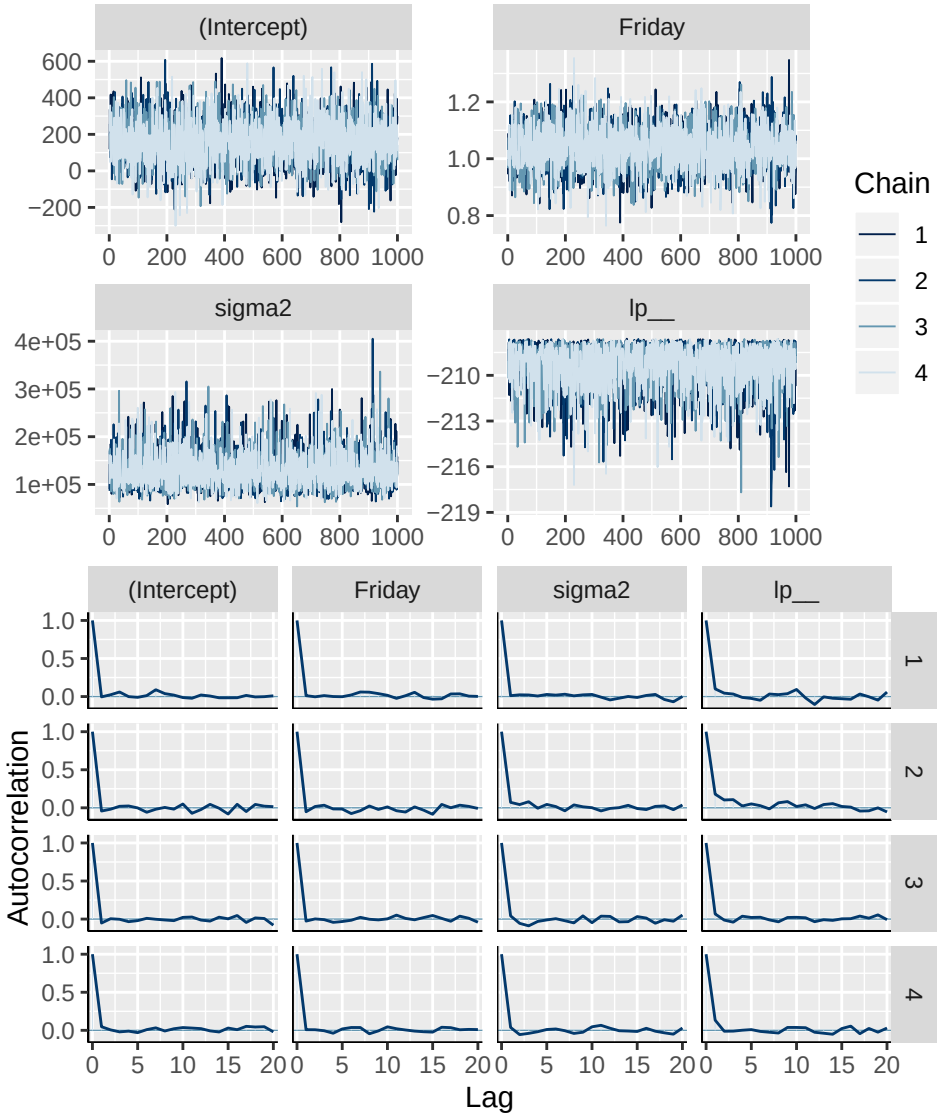
The model could be expanded to consider other predictor variables. Prior literature suggest that some characteristics that may influence trip generation my include urban/non-urban square footage, number of hotel rooms on premises, etc. These were not included in these models because the data was not available. If/when it does, it may provide additional information that could lead to better models.

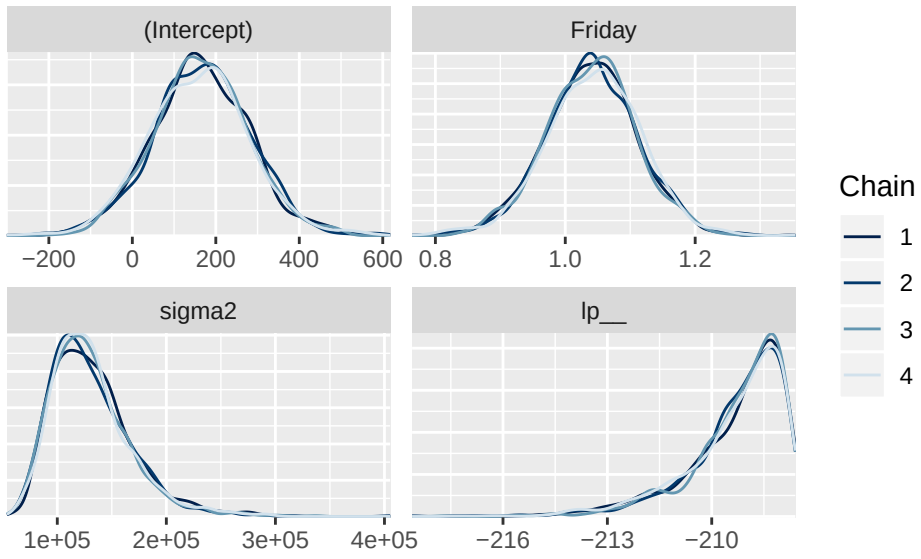
Appendix

Conversion and Density Plots

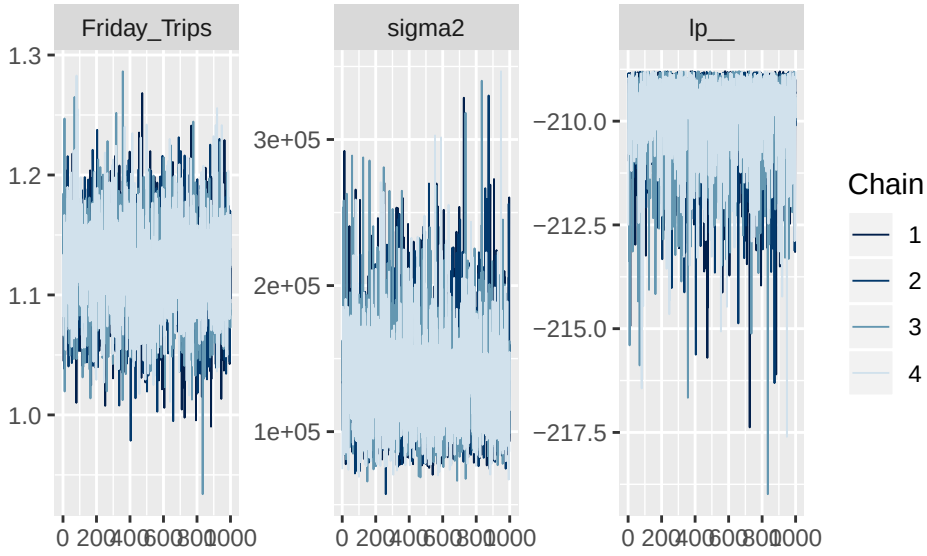
The following plots include trace and lag plots for the Markov Chains. All of them indicate that the chains mixed well and converged quickly after the burn in period. Additionally, the density overall plots are included for each model and their respective parameters as well.

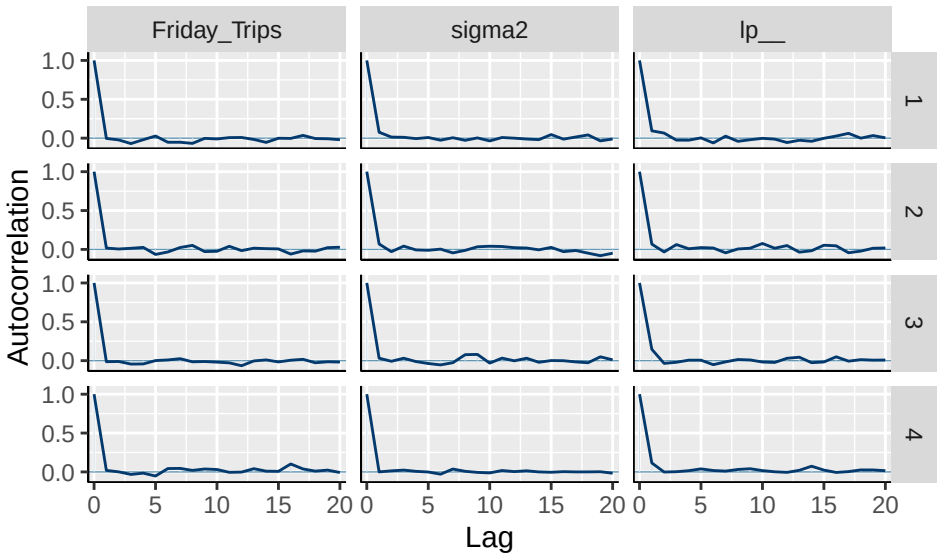
Regression of Saturday Trips on Friday Trips with Intercept

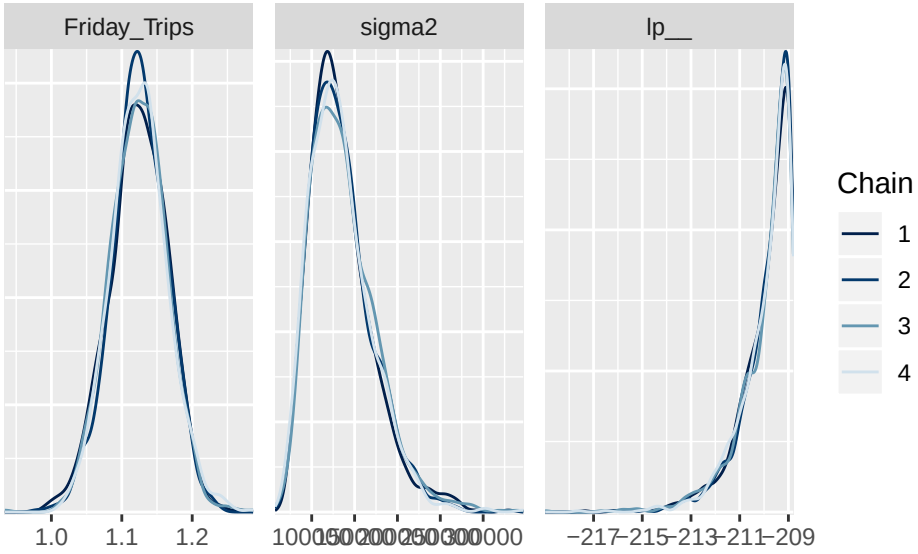




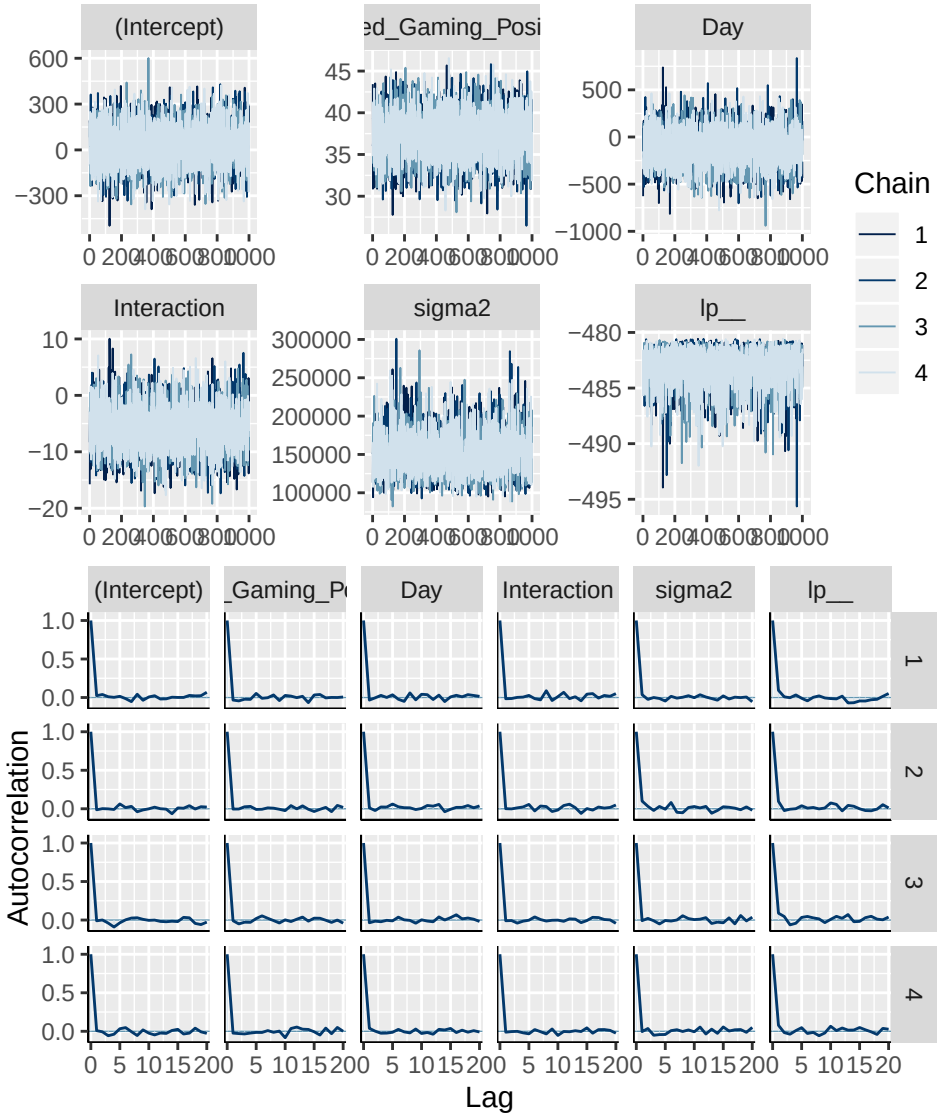
Regression of Saturday Trips on Friday Trips without Intercept

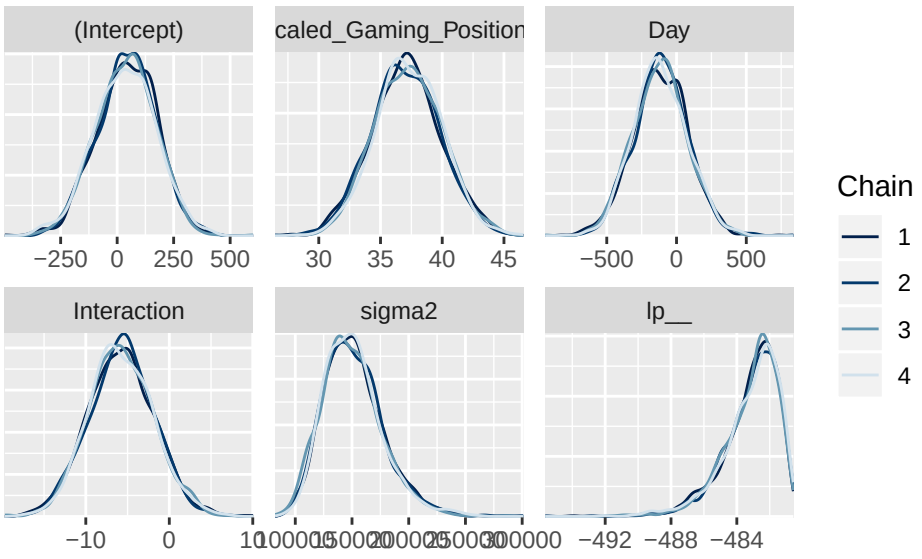



Friday_Trips
sigma2
lp__
Autocorrelation
1.0
0.5
0.0
0
5
10
15
20
Lag

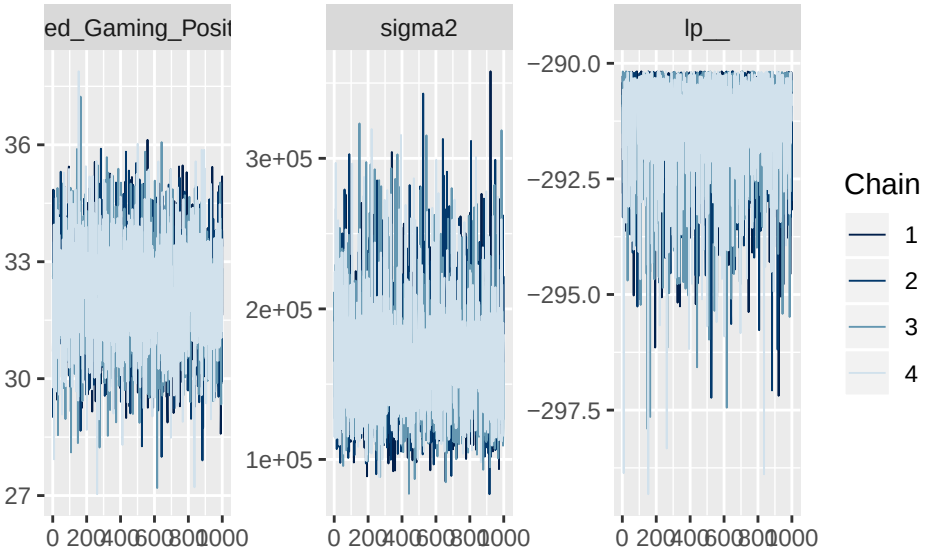

Friday_Trips
sigma2
lp__
1.0
1.1
1.2
−217
−215
−213
−211
−209
Chain
1
2
3
4

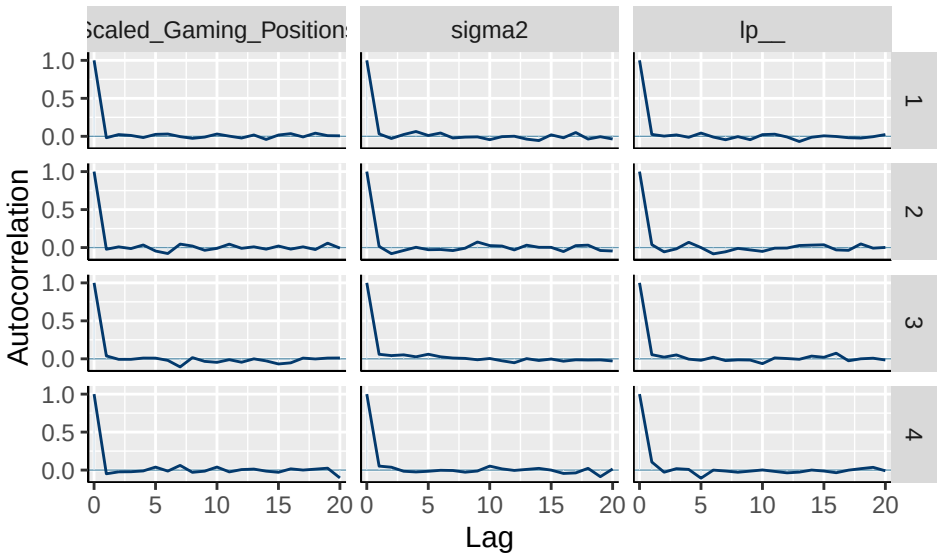
Regression of Total Trips on Gaming Positions and Day

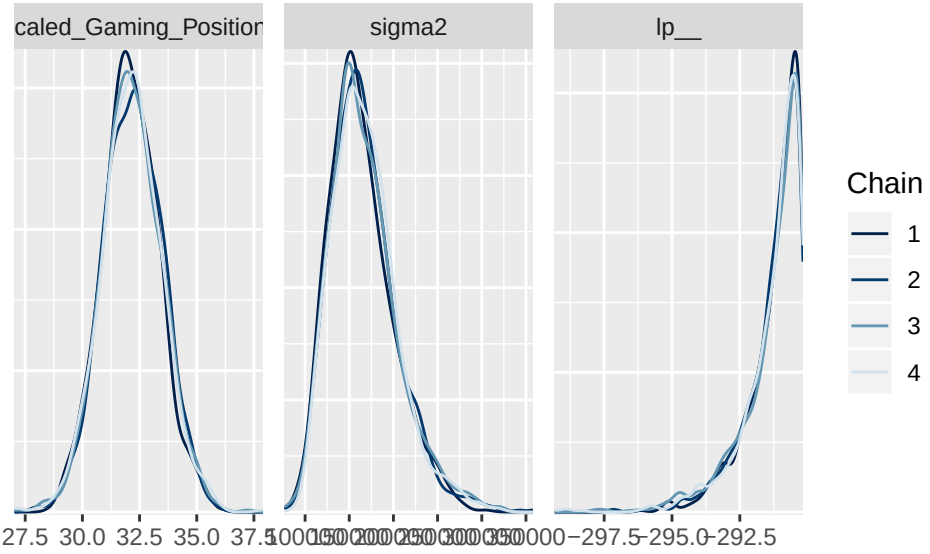




Regression of Friday Trips on Gaming Positions







GitHub Repository

Link to GitHub Repository for Code: https://github.com/hagk17/BayesianModelingProject