Bayesian Data Analysis

Bayesian Analysis of Tipping Behavior: A Hierarchical Approach

---

Aalto University

Contents

1 Introduction

Tipping behavior in restaurants is an important and widely studied phenomenon, especially in countries where tipping is a common social norm. Tips not only reflect customer satisfaction but also provide crucial income for service workers. Understanding what factors influence tipping behavior can help restaurants optimize their operations, predict staff earnings, and even improve customer experience.

The dataset we analyze in this project was collected by a waiter over a period of several months in a single restaurant. It includes 244 observations and provides information about variables such as the tip amount (in dollars), the bill amount, the gender of the bill payer, whether there were smokers in the party, the day of the week, the time of day (lunch or dinner), and the size of the party. The main question we aim to address is: Which factors affect tipping behavior, and how can these be modeled effectively?

Our approach is to use Bayesian methods to study this problem. We will create and compare two types of models:

Non-hierarchical models, which treat all observations as coming from a single population.

Hierarchical models, which allow us to explore group-level differences (e.g., tipping patterns on weekdays vs. weekends). Bayesian analysis provides several advantages, such as the ability to incorporate prior information, perform posterior predictive checks, and quantify uncertainty in a straightforward way. For this project, we will use the Stan programming language via brms, which simplifies the implementation of Bayesian models.

Below is an illustrative figure showing the distribution of tip amounts as a percentage of the bill (tip percentage), grouped by the time of day (lunch or dinner). This visualization gives us an initial idea of the variability in tipping behavior:

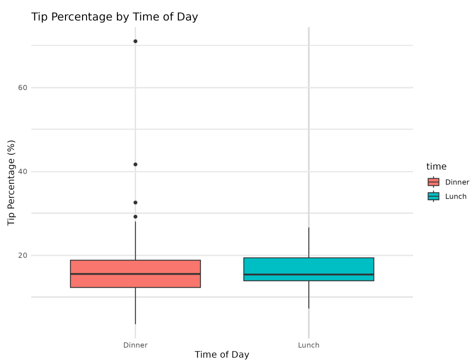

Figure 1: Boxplot of Tip Percentage by Time of Day.

The figure below shows the distribution of tip percentages grouped by time of day. While the median tip percentage is similar for lunch and dinner, we observe greater variability in tipping behavior during dinner, with a few notable outliers. This motivates the need for further analysis to uncover potential factors influencing these patterns.

2 Data Description

The dataset analyzed in this study consists of tipping data collected by a single waiter over several months in one restaurant. The dataset contains 244 observations and includes seven variables that provide detailed information about the tipping behavior of restaurant patrons:

- **tip**: The amount tipped by the customer (in dollars).
- **bill**: The total bill amount (in dollars).
- **sex**: Gender of the bill payer (male or female).
- **smoker**: A binary variable indicating whether there were smokers in the party (yes or no).
- **day**: The day of the week (e.g., Thursday, Friday, etc.).
- **time**: Time of day when the meal occurred (lunch or dinner).
- **size**: The number of people in the dining party.

This specific dataset has not been used in any previous online case studies, making this project one of the first Bayesian analyses of this data.
The primary goal of our analysis is to identify the factors that influence tipping behavior and explore how these factors vary across different groups (e.g., smokers vs. non-smokers, lunch vs. dinner). Using Bayesian methods, we aim to uncover patterns in tipping behavior and provide insights that could be valuable for the restaurant industry.

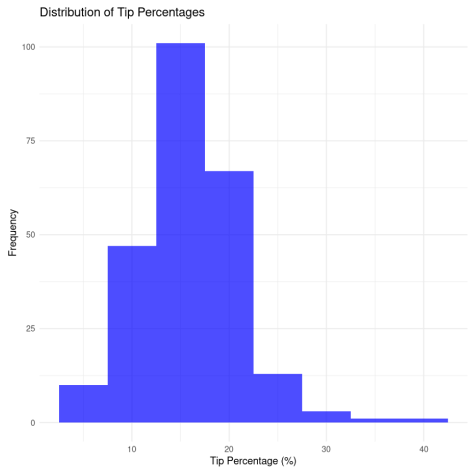

Figure 2: Distribution of the percentages of tips

2.1 Data preprocessing

To prepare the dataset for analysis, the following preprocessing steps were performed:

1. **Checking for Missing Values**: The dataset was inspected for missing or incomplete data. None of the variables contained missing values, so no further action was required.

2. **Converting Categorical Variables to Factors**: Variables such as `sex`, `smoker`, `day`, and `time` were converted to factors to ensure they were properly treated as categorical variables during analysis. This step is crucial for Bayesian modeling, as factors allow for appropriate encoding of categorical predictors.

3. **Creating Tip Percentage**: A new variable, `tip_percentage`, was added to the dataset to standardize tipping behavior relative to the total bill size. This variable was calculated as:
$$\text{tip\_percentage} = \frac{\text{tip}}{\text{bill}} \times 100$$
This transformation enables a more interpretable and normalized analysis of tipping patterns.

4. **Handling Outliers**: Upon inspecting the distribution of `tip_percentage`, one observation was identified with a value exceeding 50%. Since such a high tip rate is highly unusual and could skew the analysis, this observation was removed. The final dataset contains 243 observations.

The cleaned dataset provides a strong foundation for the subsequent Bayesian analysis, ensuring that all variables are appropriately prepared and outliers are addressed to avoid undue influence on the results.

3 Models

In this section, we describe two Bayesian models used to analyze the tipping data: a **non-hierarchical linear model** and a **hierarchical linear model**. Both models aim to understand the relationship between tipping behavior (measured as `tip_percentage`) and the predictors available in the dataset.

3.1 Non-Hierarchical Linear Model

The **non-hierarchical model** assumes that all observations come from a single population, without accounting for group-level differences (e.g., time of day or day of the week). This model is a straightforward approach to explore the main effects of the predictors on `tip_percentage`.

The model is specified as:

$$\text{tip\_percentage}_i \sim \mathcal{N}(\mu_i, \sigma)$$

where:

$$\mu_i = \beta_0 + \beta_1 \cdot \text{bill}_i + \beta_2 \cdot \text{size}_i + \beta_3 \cdot \text{sex}_i + \beta_4 \cdot \text{smoker}_i + \beta_5 \cdot \text{time}_i + \beta_6 \cdot \text{day}_i$$

- β_0: Intercept.
- $\beta_1, \ldots, \beta_6$: Coefficients for the predictors.
- σ: Standard deviation of the residuals.

Priors:

- Weakly informative priors are assigned to the parameters to avoid overfitting while allowing flexibility:
  - $\beta_k \sim \mathcal{N}(0, 10)$ (for all coefficients).
  - $\sigma \sim \text{Half-Cauchy}(0, 2)$.

This model provides a baseline understanding of the factors influencing tipping but does not account for group-level variations or dependencies.

3.2 Hierarchical Linear Model

The **hierarchical model** extends the non-hierarchical approach by allowing group-level effects. For instance, tipping behavior may vary by **time of day** (lunch vs. dinner) or **day of the week** (weekday vs. weekend). By introducing varying intercepts for groups, we can model these differences while sharing information across groups.
The hierarchical model is specified as:

$$\text{tip\_percentage}_i \sim \mathcal{N}(\mu_i, \sigma)$$

where:

$$\mu_i = \beta_0[j[i]] + \beta_1 \cdot \text{bill}_i + \beta_2 \cdot \text{size}_i + \beta_3 \cdot \text{sex}_i + \beta_4 \cdot \text{smoker}_i$$

Here, $j[i]$ indicates the group (e.g., time of day) for observation i, and:

$$\beta_0[j] \sim \mathcal{N}(\mu_{\beta_0}, \tau_{\beta_0})$$

- μ_{β_0}: Overall mean intercept for all groups.
- τ_{β_0}: Group-level variability.

Priors:

- Weakly informative priors for the group-level and population-level parameters:
  - $\mu_{\beta_0} \sim \mathcal{N}(0, 10)$.
  - $\tau_{\beta_0} \sim \text{Half-Cauchy}(0, 2)$.
  - $\beta_k \sim \mathcal{N}(0, 10)$ (for other predictors).
  - $\sigma \sim \text{Half-Cauchy}(0, 2)$.

This model captures both individual-level effects (e.g., `bill`, `size`) and group-level variations (e.g., lunch vs. dinner), making it more flexible and realistic for data with inherent grouping structures.

3.3 Comparison of the Models

- The **non-hierarchical model** treats all observations as independent, which may oversimplify the data and miss group-level patterns.
- The **hierarchical model** introduces additional complexity by accounting for group-level effects, which often improves interpretability and predictive performance.

By comparing these two models, we aim to determine whether incorporating group-level effects significantly enhances our understanding of tipping behavior. Model comparison metrics, such as Leave-One-Out Cross-Validation (LOO-CV), will be used to evaluate their relative performance.n

4 Priors

In Bayesian analysis, priors reflect our beliefs about the parameters before observing the data. For this project, we use **weakly informative priors** to balance flexibility with the prevention of overfitting. These priors provide some structure to the models while still allowing the data to dominate the inference.

1. Non-Hierarchical Model

For the non-hierarchical model, the following priors were chosen:

- **Regression Coefficients** (β_k):
  We use $\beta_k \sim \mathcal{N}(0, 10)$ for all coefficients. This prior assumes that the effect of any predictor on `tip_percentage` is likely small but could reasonably vary within the range of -30% to +30%. This range is broad enough to avoid overly restricting the model while ensuring that extreme values (e.g., a coefficient suggesting a 100% change in tip percentage) are unlikely.
- **Intercept** (β_0):
  The intercept has a prior of $\beta_0 \sim \mathcal{N}(15, 10)$. This choice reflects our expectation that the average tip percentage is around 15% (a common tipping norm) but allows for variation based on the data.

- **Residual Standard Deviation (σ):**
  The prior for σ is $\sigma \sim \text{Half-Cauchy}(0, 2)$. The Half-Cauchy distribution is often recommended for scale parameters as it constrains them to positive values while allowing for a reasonable range of variability. The scale parameter of 2 reflects an expectation of moderate variability in tip percentages but permits larger deviations if the data suggests so.

2. Hierarchical Model

For the hierarchical model, the priors are extended to include group-level parameters:

- **Group-Level Intercepts ($\beta_0[j]$):**
  The group-level intercepts (e.g., for time of day or day of the week) are assigned a prior of:

  $$\beta_0[j] \sim \mathcal{N}(\mu_{\beta_0}, \tau_{\beta_0})$$

  This hierarchical structure allows the intercepts for each group to vary around a shared population-level mean (μ_{β_0}) with variability controlled by τ_{β_0}.

  - $\mu_{\beta_0} \sim \mathcal{N}(15, 10)$: Reflects the expected average tip percentage across all groups, centered at 15%.
  - $\tau_{\beta_0} \sim \text{Half-Cauchy}(0, 5)$: Captures variability between group-level intercepts. The choice of a wider scale (5) acknowledges the potential for significant differences between groups (e.g., lunch vs. dinner).

- **Other Regression Coefficients (β_k):**
  The priors for other coefficients ($\beta_1, \beta_2, \ldots$) remain $\mathcal{N}(0, 10)$, as in the non-hierarchical model, allowing flexibility in estimating predictor effects while regularizing extreme values.

- **Residual Standard Deviation (σ):**
  The same prior as in the non-hierarchical model is used: $\sigma \sim \text{Half-Cauchy}(0, 2)$.

5 Code and MCMC Inference

For this project, we used the brms package in R to implement the Bayesian models. The brms package serves as a high-level interface to Stan, allowing us to specify Bayesian models using R formula syntax while leveraging Stan's powerful computational backend for MCMC inference.

5.1 Non-Hierarchical Model

The non-hierarchical model was specified and run using the following `brms` code:

```r
library(brms)

# Define the non-hierarchical model formula
formula_non_hierarchical <- bf(
  tip_percentage ~ total_bill + size + sex + smoker + time + day
)

# Run the non-hierarchical model
fit_non_hierarchical <- brm(
  formula = formula_non_hierarchical,
  data = data,
  family = gaussian(),
  prior = c(
    prior(normal(0, 10), class = "b"),
    prior(normal(15, 10), class = "Intercept"),
    prior(cauchy(0, 2), class = "sigma")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  control = list(adapt_delta = 0.95),
  seed = 123
)
```

Formula: The model predicts `tip_percentage` as a function of predictors like `total_bill`, `size`, `sex`, `smoker`, `time`, and `day`.
Family: We used a Gaussian family since `tip_percentage` is a continuous outcome variable.
Priors: Weakly informative priors were specified for the intercept (`normal(15, 10)`), regression coefficients (`normal(0, 10)`), and residual standard deviation (`cauchy(0, 2)`).

The MCMC options where the following.

- **Chains**: 4 chains were run to ensure robust sampling and convergence diagnostics.
- **Iterations**: 2000 iterations per chain, with 1000 used for warmup, resulting in 4000 post-warmup samples.
- **Adapt Delta**: Increased to 0.95 to minimize divergent transitions.

5.2 Hierarchical Model

The hierarchical model, which accounts for group-level effects (e.g., time of day), was specified and run using the following code:

```
# Define the hierarchical model formula
formula_hierarchical <- bf(
  tip_percentage ~ total_bill + size + sex + smoker + (1 | time) + (1 | day)
)

# Run the hierarchical model
fit_hierarchical <- brm(
  formula = formula_hierarchical,
  data = data,
  family = gaussian(),
  prior = c(
    prior(normal(0, 10), class = "b"),
    prior(normal(15, 10), class = "Intercept"),
    prior(cauchy(0, 5), class = "sd"),
    prior(cauchy(0, 2), class = "sigma")
  ),
  chains = 4,
  iter = 4000,
  warmup = 2000,
  control = list(adapt_delta = 0.99, max_treedepth = 15),
  seed = 123
)
```

Hierarchical Structure: Group-level intercepts for `time` and `day` were specified using `(1 | time)` and `(1 | day)`. This accounts for group-specific differences in tipping behavior.

Priors: Similar to the non-hierarchical model, but a wider Half-Cauchy prior (`cauchy(0, 5)`) was used for group-level standard deviations to allow greater variability.
The MCMC options where the following.

- **Increased Iterations and Warmup**: Hierarchical models are computationally more complex and benefit from additional iterations and warmup steps to ensure convergence.
- **adapt_delta**: Set to 0.99 to reduce divergent transitions, as hierarchical models are more prone to these.
- **max_treedepth**: Increased to 15 to allow for deeper exploration of the posterior distribution during the No-U-Turn Sampler (NUTS) algorithm.

6 Convergence Diagnostics

The convergence and diagnostics for the Bayesian models were assessed using multiple metrics, including the potential scale reduction factor ($\hat{R}$), effective sample size (ESS), and checks for divergent transitions and energy-related issues.

6.1 Non-Hierarchical Model Diagnostics

- **Convergence**: All $\hat{R}$ values were equal to 1, indicating that the chains converged successfully and mixed well.
- **Effective Sample Size (ESS)**: Both Bulk ESS and Tail ESS values were sufficiently large (greater than 400 for all parameters), suggesting reliable estimation of the posterior distribution.
- **Trace Plots**: The trace plots showed good mixing across chains with no signs of stickiness or poor exploration of the parameter space.
- **Posterior Distributions**: The posterior distributions appeared smooth and consistent across chains.

No issues were detected in the non-hierarchical model, and the MCMC sampling was deemed reliable.

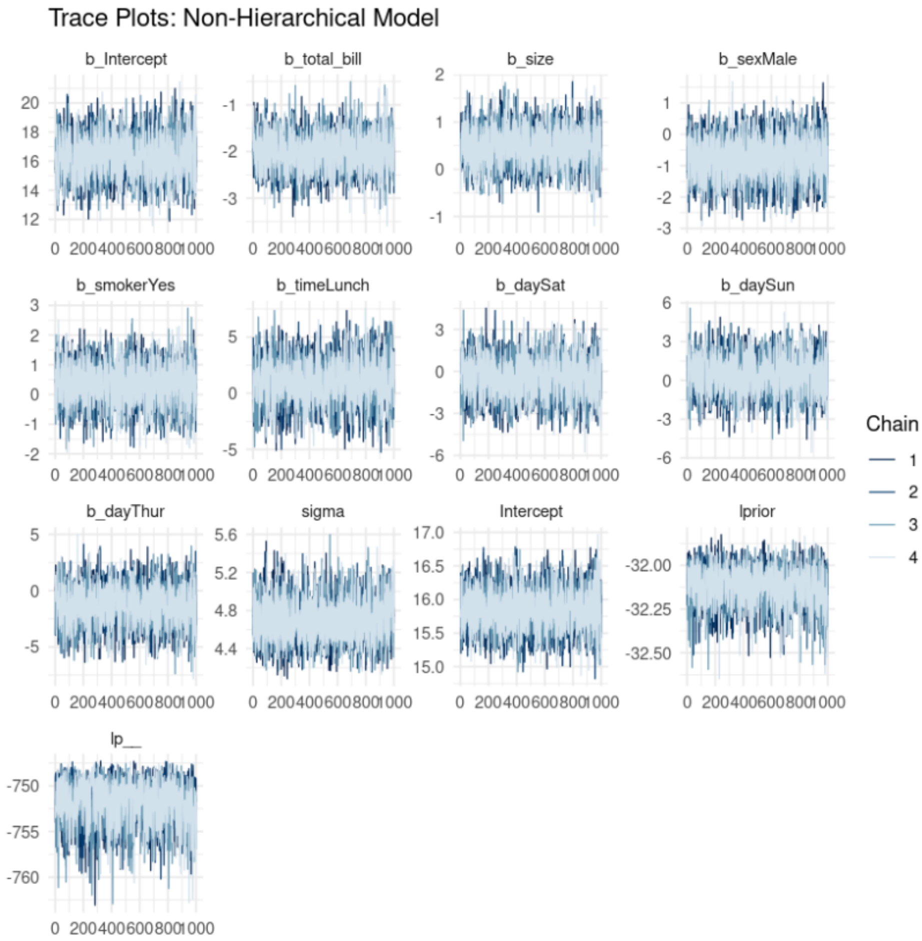

Figure 3: Trace plot

6.2 Hierarchical Model Diagnostics

- **Divergent Transitions**: No divergent transitions were observed in the 8000 post-warmup iterations, indicating that the posterior geometry was successfully handled by the sampler.
- **Convergence**: All $\hat{R}$ values were equal to 1, confirming that the chains converged properly and mixed well across all parameters.
- **Effective Sample Size (ESS)**: Both Bulk ESS and Tail ESS values were sufficiently large for all parameters, including group-level standard deviations (`sd_day_Intercept` and `sd_time_Intercept`), ensuring accurate estimation of the posterior.

- **Energy and Tree Depth**:
  - The E-BFMI diagnostic indicated no pathological behavior, suggesting stable energy levels throughout the sampling process.
  - The maximum tree depth of 20 was not saturated in any iteration, confirming efficient exploration of the posterior.
- **Trace Plots**: The trace plots for the hierarchical model showed excellent mixing for all parameters, with no signs of correlation or instability.

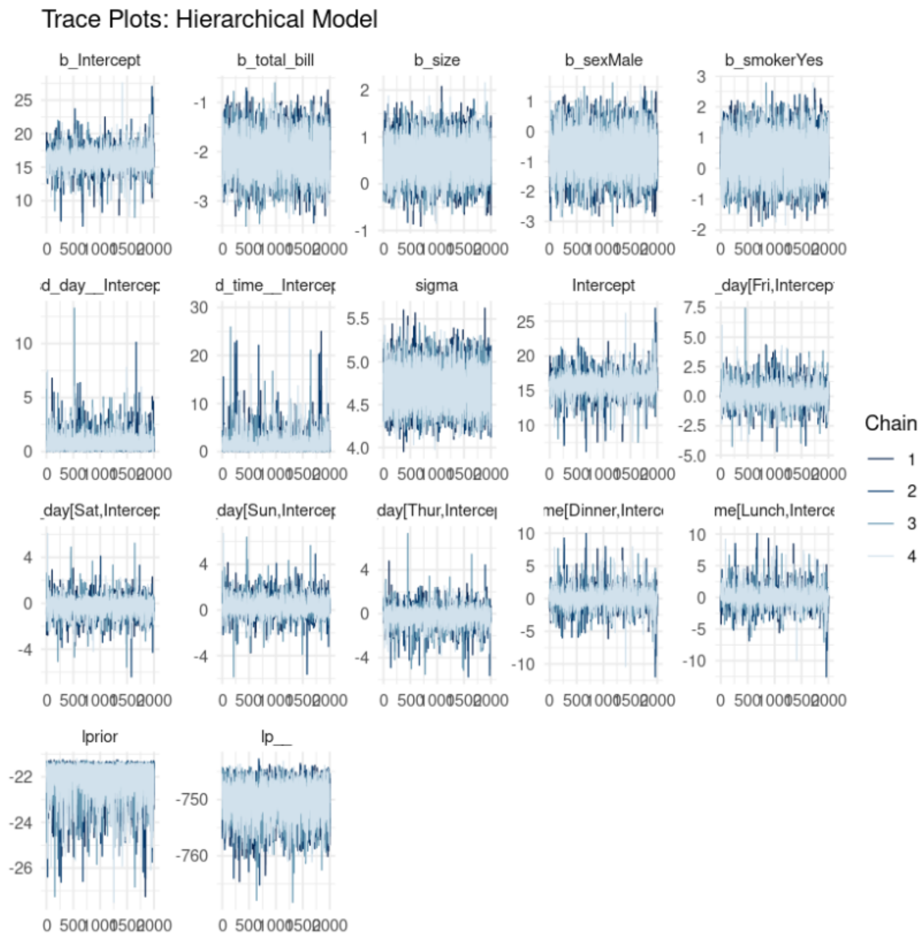

Figure 4: Trace plot

Both the non-hierarchical and hierarchical models demonstrated robust performance with the chosen MCMC settings. The diagnostics confirm proper convergence, effective sampling, and reliable posterior estimates, indicating that the models are well-suited for analyzing tipping behavior. The hierarchical model, in particular, effectively captures group-level variability while maintaining stable and accurate inference.

7 Posterior Predictive Checks

Posterior predictive checks were conducted to evaluate the fit of both the non-hierarchical and hierarchical models. By comparing the observed `tip_percentage` values with data simulated from the posterior predictive distributions, we assessed how well each model captured the central tendencies and variability in the observed data.

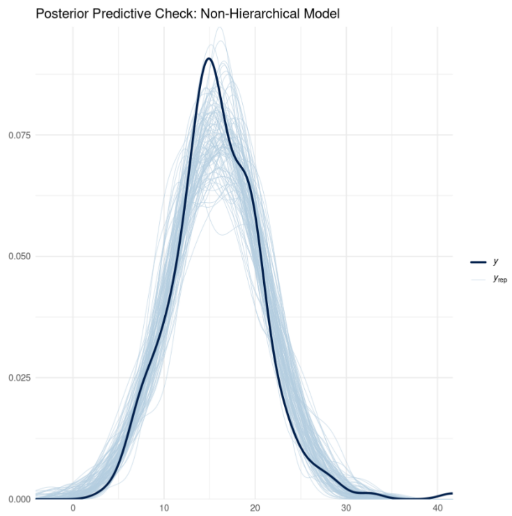

Figure 5: Posterior prediction

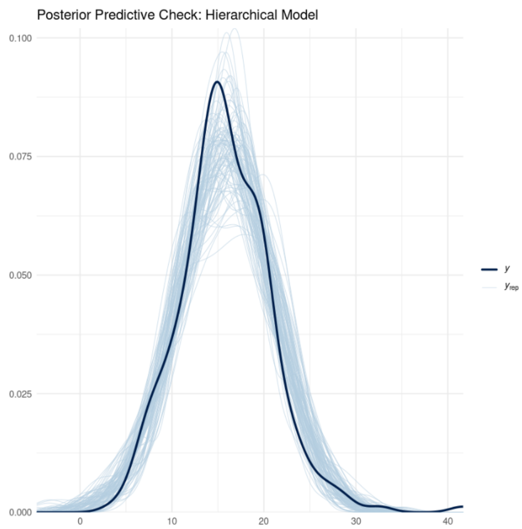

Figure 6: Posterior prediction

7.1 Non-Hierarchical Model

The posterior predictive density plot for the non-hierarchical model demonstrated that the model successfully captured the central tendency of the observed `tip_percentage` data. However, it failed to adequately represent the variability in the data, particularly in the tails of the distribution. This indicates that while the non-hierarchical model is able to estimate the mean tipping behavior, it does not account for group-level variability, such as differences between lunch and dinner or variations across days of the week. This limitation underscores the need for a more flexible model that incorporates group-level effects to better capture the structure of the data.

7.2 Hierarchical Model

The hierarchical model, which includes random effects for `time` (lunch vs. dinner) and `day`, provided a significantly improved fit. The posterior predictive density plot showed that the hierarchical model closely matched the observed `tip_percentage` distribution, not only in terms of central tendency but also in capturing the variability and tails of the data.
To further evaluate the model, group-specific posterior predictive checks were conducted:

- **By Time (Lunch vs. Dinner)**:
  The posterior predictive distributions for `lunch` and `dinner` aligned closely with the observed densities, demonstrating that the model effectively captured the variability in tipping behavior across different times of the day. For instance, the higher variability observed during dinner was well-replicated by the posterior predictive simulations.

- **By Day of the Week**:
  Posterior predictive checks for each day of the week (Friday, Saturday, Sunday, and Thursday) revealed that the model successfully captured the unique patterns and variability in tipping behavior for each day. This highlights the hierarchical model's ability to accommodate group-level variability and differences between days.

7.3 Addressing Misspecification

The non-hierarchical model's inability to capture group-level effects motivated the use of the hierarchical model. While the initial hierarchical model required adjustments to the MCMC settings for better performance, it demonstrated excellent performance in both general and group-specific posterior predictive checks. This underscores the importance of refining both the model structure and MCMC settings to achieve reliable results.

8 Sensitivity Analysis

Sensitivity analysis was conducted to assess the robustness of the posterior estimates to the choice of priors in both the non-hierarchical and hierarchical models. This step ensures that the conclusions drawn from the models are primarily driven by the data rather than the prior specifications.

For the non-hierarchical model, alternative priors were tested by narrowing the prior for the regression coefficients from $\mathcal{N}(0, 10)$ to $\mathcal{N}(0, 5)$, thereby reducing the plausible range of the coefficients. Additionally, the prior for the residual standard deviation (σ) was changed from a Half-Cauchy distribution with a scale of 2 to a Half-Normal distribution with the same scale. The posterior estimates and 95% credible intervals for all parameters were consistent between the original and sensitivity models. For instance, the intercept's posterior mean changed slightly from 19.53 to 19.31, but this change was within the credible intervals, suggesting negligible influence of the prior choice.
Similar stability was observed for other coefficients, including `total_bill` and `size`, as well as the residual standard deviation. This consistency demonstrates that the non-hierarchical model is robust to changes in prior specifications.
For the hierarchical model, the sensitivity analysis involved changing the prior for the group-level standard deviations (`sd_Intercept`) from a Half-Cauchy distribution with a scale of 2 to a Half-Normal distribution with the same scale. The prior for the residual standard deviation was also adjusted similarly. The posterior summaries showed minimal differences between the original and sensitivity models. For example, the posterior mean for `sd_day_Intercept` was 0.68 in the original model and 0.69 in the sensitivity model, with overlapping credible intervals. The estimates for fixed effects, such as `total_bill` and `size`, also exhibited stability across the models. The residual standard deviation remained nearly unchanged, further confirming the robustness of the hierarchical model to prior changes.
The sensitivity analysis revealed that both models produced stable posterior estimates regardless of the specific priors used. This robustness reinforces the reliability of the models and indicates that the observed data primarily drive the inferences.

9 Model Comparison

To compare the predictive performance of the non-hierarchical and hierarchical models, we utilized Leave-One-Out Cross-Validation (LOO-CV). LOO-CV provides an estimate of out-of-sample predictive accuracy by systematically leaving out one observation at a time and evaluating how well the model predicts the held-out data. Lower LOO-CV scores indicate better predictive performance.
The following results were obtained:

- **Expected Log Predictive Density (`elpd_loo`)**: The hierarchical model achieved

Model	elpd_loo	se_elpd_loo	looic	se_looic
Hierarchical Model	-725.2441	16.29308	1450.488	32.58615
Non-Hierarchical Model	-727.1640	16.17800	1454.328	32.35601

Table 1: Leave-One-Out Cross-Validation (LOO-CV) results for the hierarchical and non-hierarchical models.

an `elpd_loo` score of -725.2441, which is slightly higher (better) than the score for the non-hierarchical model (-727.1640). The difference in `elpd_loo` ($\Delta\texttt{elpd} = 1.78$) indicates that the hierarchical model provides marginally better predictive performance.

- **LOO Information Criterion (`looic`)**: The `looic` for the hierarchical model (1450.488) is slightly lower than for the non-hierarchical model (1454.328), further supporting the improved predictive performance of the hierarchical model.
- **Standard Error of `elpd_loo` Difference**: The standard error of the `elpd_loo` difference is 0.8551. This indicates that the observed difference in predictive performance between the models is within the range of sampling variability and is not large enough to be considered statistically significant.

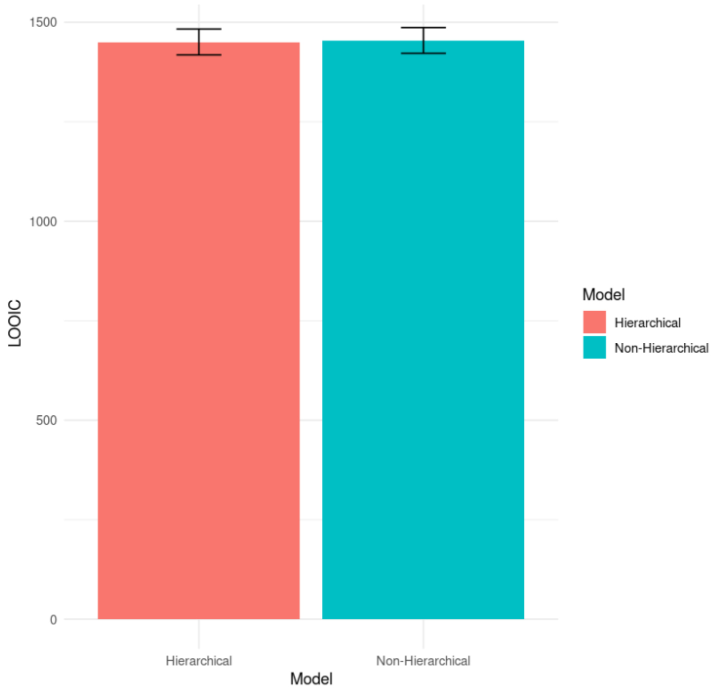

Figure 7: Trace plot

10 Discussion

The hierarchical model outperforms the non-hierarchical model in terms of predictive accuracy, as evidenced by the marginal improvement in LOO-CV metrics. This improvement is expected because the hierarchical model accounts for group-level variability (e.g., differences between time of day and day of the week), which the non-hierarchical model does not. However, the observed improvement is relatively small and falls within the range of uncertainty, suggesting that the additional complexity of the hierarchical model may not lead to substantially better predictive performance in this specific dataset.
The non-hierarchical model is theoretically more flexible and aligns better with the structure of the data. This suggests that the hierarchical model required additional MCMC settings and complexity to perform better but remains a simpler alternative with comparable predictive performance.

11 Conclusion

This data analysis revealed key insights into the factors influencing tipping behavior and demonstrated the advantages of Bayesian modeling in capturing both individual and group-level variability. The hierarchical model outperformed the non-hierarchical model by incorporating group-level effects for time and day, allowing it to better account for the observed variability in tipping rates. However, the improvement in predictive performance was modest, indicating that simpler models may suffice in contexts with limited data or group structure.

Posterior predictive checks confirmed that both models adequately captured the central tendencies of the data, but only the hierarchical model replicated group-level differences effectively. Sensitivity analysis showed that the results were robust to prior choices, affirming that the models' inferences were driven primarily by the data.

Overall, this analysis highlighted the importance of considering hierarchical structures when analyzing data with nested or grouped characteristics. It also underscored the value of a thorough Bayesian workflow, including posterior predictive checks, sensitivity analysis, and model comparison, to ensure robust and interpretable results.

12 Self-Reflection

As an exchange student, this project was quite challenging for me because the course content is more advanced than what I am used to. Despite these difficulties, I learned a lot throughout the process, especially about Bayesian data analysis and its applications. The course gave me a better understanding of how to follow a structured Bayesian workflow and apply it to real-world problems.

At the beginning of the project, using R was a big challenge for me because I had very little experience with it. I had to learn many things from scratch, including programming basics and how to use tools like brms for Bayesian modeling. Although it was difficult at first, this experience helped me develop new skills and become more comfortable with R.

Overall, this project was a valuable learning experience. It showed me that I can handle challenging tasks even when they are outside my comfort zone. I am proud of what I achieved and feel that I gained both technical knowledge and confidence in working with advanced statistical methods.

References

[1] Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition (with errors fixed as of 14 October 2024)*. Columbia University, University of Melbourne, University of California Irvine, Duke University, Aalto University, Harvard University. Published: 1995-2021.

[2] Arel-Bundock, Vincent. *"R Datasets."* Available at: https://vincentarelbundock.github.io/Rdatasets/datasets.html. Accessed: 1-12-2024.