Analyzing Toronto Traffic Speeds: Insights into Patterns and Predictive Modeling*

Bayesian Linear Regression Reveals Key Determinants of Mean Speed and Post-Pandemic Trends

Andy Jiang

November 29, 2024

This paper investigates traffic speed data from Toronto collected between 2017 and 2024, focusing on trends in speed percentiles and their relationship to traffic volume. Using a Bayesian linear regression model, we predict mean vehicle speeds based on key predictors such as 5th, 50th, and 95th percentile speeds and total traffic volume. Results highlight significant changes in traffic patterns post-2020, with increased extreme speeds linked to higher traffic volumes. This study provides insights into urban mobility dynamics, offering a basis for traffic management strategies and policy-making.

Table of contents

*Code and data are available at: [https://github.com/AndyYanxunJiang/toronto-vehicles-speed-analysis).

1 Introduction

Traffic speed analysis is a critical tool for understanding urban mobility, road safety, and congestion. In metropolitan areas like Toronto, shifting traffic patterns reflect changes in urban infrastructure, population growth, and behavioral adaptations to external events, such as the COVID-19 pandemic. The advent of technologies like radar-equipped speed signs has enabled the collection of detailed traffic data, providing new opportunities to analyze and predict traffic behavior at a granular level.

This paper focuses on traffic speed data collected in Toronto's designated Safety Zones between 2017 and 2024. The dataset comprises monthly summaries of speed percentiles, vehicle counts across specific speed ranges, and total traffic volumes. Notably, the study explores trends in extreme speeds (vehicles exceeding 100 km/h) and evaluates how they have evolved before and after 2020, a period marked by rapid population growth and post-pandemic recovery. These analyses aim to address a key gap: understanding how percentile-based speed measures and traffic volume influence overall mean speeds.

To investigate these relationships, we employ a Bayesian linear regression model to predict mean vehicle speed (`spd_mean`) based on predictors including 5th, 50th, and 95th percentile

speeds and traffic volume. Bayesian modeling allows for the incorporation of uncertainty and prior information, making it well-suited for analyzing urban traffic data.

Our findings reveal significant changes in traffic behavior post-2020, including higher extreme speed occurrences at increased traffic volumes, reflecting broader shifts in urban mobility. This study highlights the importance of percentile-based metrics in predicting mean speeds and contributes to ongoing efforts to enhance traffic safety and efficiency. The paper is structured as follows: the data section provides an overview of the dataset and key variables, the model section outlines the Bayesian regression framework, results are discussed in detail, and the discussion section reflects on the implications and limitations of the findings.

2 Data

All data analysis was conducted using R (R Core Team 2023) with the support of the following packages: tidyverse (Wickham et al. 2019), here (Müller 2020), ggplot2 (Wickham 2016), rstanarm (Goodrich et al. 2022), kableExtra (Zhu 2024), modelsummary (Arel-Bundock 2022), dplyr (Wickham et al. 2023), and lubridate (Grolemund and Wickham 2011). These packages provided a robust framework for data manipulation, visualization, Bayesian modeling, and reporting, facilitating reproducible and efficient analysis.

2.1 Overview

This paper analyzes traffic speed data collected in Toronto over the period 2017–2024. The dataset contains detailed information on percentile speeds, total traffic volume, and counts of vehicles traveling within specific speed ranges. These data provide an opportunity to study the evolution of traffic patterns in Toronto, especially in the context of changing demographics, urbanization, and post-pandemic recovery.

The dataset aggregates metrics monthly, offering insights into how traffic behavior has changed over time. This analysis is particularly focused on extreme speeds (vehicles exceeding 100 km/h) and their relationship with total traffic volume, as well as trends in key speed percentiles.

2.2 Measurement

The dataset analyzed in this paper originates from the Safety Zone Watch Your Speed Program (WYSP), a City of Toronto initiative designed to monitor and influence driver behavior in designated Safety Zones. Data are collected using radar-equipped speed display signs installed on hydro poles or streetlights, which measure the speeds of oncoming vehicles with an accuracy of ± 1 km/h and display the speeds on LED screens to encourage compliance with speed limits. These measurements are aggregated into monthly summaries, including key metrics such as

percentile speeds (e.g., 5th, 50th, 95th percentiles), vehicle counts within specific speed ranges, and total observed traffic volume. The dataset is curated and published by Open Data Toronto, which ensures its metadata quality, completeness, and accessibility.

However, the dataset has certain limitations. The number of vehicles recorded by the speed signs is not equivalent to true traffic volume, as it may exclude vehicles that pass too quickly or fall outside the radar's range. Additionally, the presence of the signs might influence driver behavior, leading to changes in speed as drivers approach them. Furthermore, the data is limited to Safety Zones where the speed signs have been installed, restricting its spatial coverage. Despite these limitations, the dataset is well-documented, complete, and up-to-date, providing valuable insights into traffic speed patterns across Toronto.

2.3 Speed Dataset Preview

The traffic speed dataset contains information collected from multiple locations and months, detailing percentile speeds, total traffic volume, and counts of vehicles traveling within specific speed ranges. The dataset aggregates these metrics monthly, offering insights into traffic behavior and variations over time.

This dataset includes: - Percentile speeds (5th, 10th, … 95th) representing speed thresholds exceeded by corresponding proportions of vehicles. - Traffic volume as the total number of vehicles observed monthly. - Speed bins (e.g., 0-4 km/h, 5-9 km/h) for detailed breakdowns of vehicle counts across speed ranges. - Table 1 showcases a preview of the cleaned dataset with selected columns to highlight key features.

Table 1: Preview of cleaned traffic speed dataset.

Month	5th Percentile	Median (50th Percentile)	95th Percentile	Vehicles (30-34 km/h)	Vehicles (50-54 km/h)	Vehicles (>100 km/h)	Total Volume
2024-09-01	7	21	32	2636	5	0	57758
2024-09-01	5	18	29	1463	1	0	48365
2024-09-01	0	0	28	162	0	0	2116
2024-09-01	21	40	55	16492	11560	4	163591
2024-09-01	0	28	46	3524	444	0	30606
2024-09-01	17	38	49	8954	1957	1	64690

Month	5th Percentile	Median (50th Percentile)	95th Percentile	Vehicles (30-34 km/h)	Vehicles (50-54 km/h)	Vehicles (>100 km/h)	Total Volume
2024-09-01	0	29	40	8312	72	0	27790
2024-09-01	0	11	26	535	0	0	59467
2024-09-01	0	49	65	1472	15730	7	62706
2024-09-01	6	33	44	8899	257	0	34309

2.4 Random Sample from the Dataset

Table 2 provides a random sample of 10 observations, offering an unbiased snapshot of key metrics such as speed percentiles, total traffic volume, and counts of vehicles exceeding 100 km/h. This selection illustrates the dataset's diversity without emphasizing specific patterns.

Table 2: Random sample of traffic speed dataset.

Month	5th Percentile	Median (50th Percentile)	95th Percentile	Total Volume	Vehicles (>100 km/h)
2024-03-01	15	39	51	33333	0
2020-02-01	11	28	42	40020	0
2024-09-01	9	24	38	117573	0
2022-12-01	0	20	34	31121	4
2021-07-01	14	37	51	160808	1
2023-09-01	30	50	65	182810	23
2022-06-01	8	22	37	58684	0
2022-02-01	15	25	47	195854	0
2020-11-01	14	36	47	87896	0

Month	5th Percentile	Median (50th Percentile)	95th Percentile	Total Volume	Vehicles (>100 km/h)
2024-02-01	0	0	28	3472	0

2.5 Extreme Speeds Dataset

Vehicles traveling at extreme speeds (over 100 km/h) represent a critical factor in understanding traffic safety and violations. Table 3 highlights the top 5 months with the highest counts of vehicles exceeding 100 km/h. This provides insights into when extreme speeds are most prevalent.

Table 3: Top 5 months with the highest counts of vehicles exceeding 100 km/h.

Month	Total Volume	Vehicles (>100 km/h)	95th Percentile Speed
2021-11-01	45735	3423	122
2023-01-01	62476	1640	85
2023-01-01	87523	1475	77
2020-06-01	273454	1391	87
2020-07-01	224624	1304	52

2.6 Proportion of Vehicles at Moderate Speeds

Another key insight from this dataset is the proportion of vehicles traveling within a moderate speed range (50–70 km/h). This measure helps understand how typical driving speeds align with safe and efficient traffic flow.

Table 4 highlights monthly proportions of vehicles in this speed range.

Table 4: Monthly proportions of vehicles traveling at moderate speeds (50–70 km/h).

Month	Moderate Speed Vehicles	Total Volume	Proportion (%)
Jan	32248526	208619163	15.46
Feb	31376308	213960232	14.66
Mar	35727340	235271611	15.19
Apr	31383270	194534832	16.13
May	32859985	207592245	15.83
Jun	32981330	201948797	16.33
Jul	33826108	196693112	17.20
Aug	33837333	195652825	17.29

Month	Moderate Speed Vehicles	Total Volume	Proportion (%)
Sep	36352915	240764265	15.10
Oct	32154660	203667881	15.79
Nov	33186997	217011827	15.29
Dec	32651478	211458699	15.44

2.7 Extreme Speeds and Traffic Volume

Figure 1 illustrates the relationship between extreme speeds (vehicles traveling over 100 km/h) and total traffic volume, categorized by pre-2020 and post-2020 periods. The post-2020 data points show a significant increase in both traffic volume and extreme speed counts, with some months recording over 2,000 vehicles exceeding 100 km/h. This increase may reflect changes in traffic behavior following the COVID-19 pandemic and a rapid population influx into Toronto through immigration. In contrast, pre-2020 data points display lower traffic volumes and fewer extreme speed counts, rarely exceeding 2,000. The fitted lines for each period highlight this divergence, with a steeper slope for post-2020 data, suggesting that higher traffic volumes post-pandemic are associated with a broader range of extreme speed occurrences.

Figure 1: Relationship between extreme speeds (over 100 km/h) and total traffic volume, categorized by pre-2020 and post-2020 time periods.

2.8 Trends in Key Percentiles

Figure 2 examines trends in key speed percentiles (5th, 50th, and 95th percentiles) from 2017 to 2024, along with the average speed spread (the difference between the 95th and 5th percentiles). The 95th percentile speed demonstrates a clear decline, falling below 50 km/h over the years. Similarly, the 50th percentile speed dropped significantly between 2017 and 2018, stabilizing at around 30 km/h with minor fluctuations. The 5th percentile speed remained steady until mid-2022, when it experienced a noticeable decline to 5 km/h, down from its previous 10 km/h range. Despite these shifts, the speed spread remains relatively consistent, ranging between 35 and 40 km/h, indicating that the overall variability in speeds has not significantly changed. These trends may reflect broader changes in traffic regulations, driver behavior, or urban infrastructure developments in Toronto.

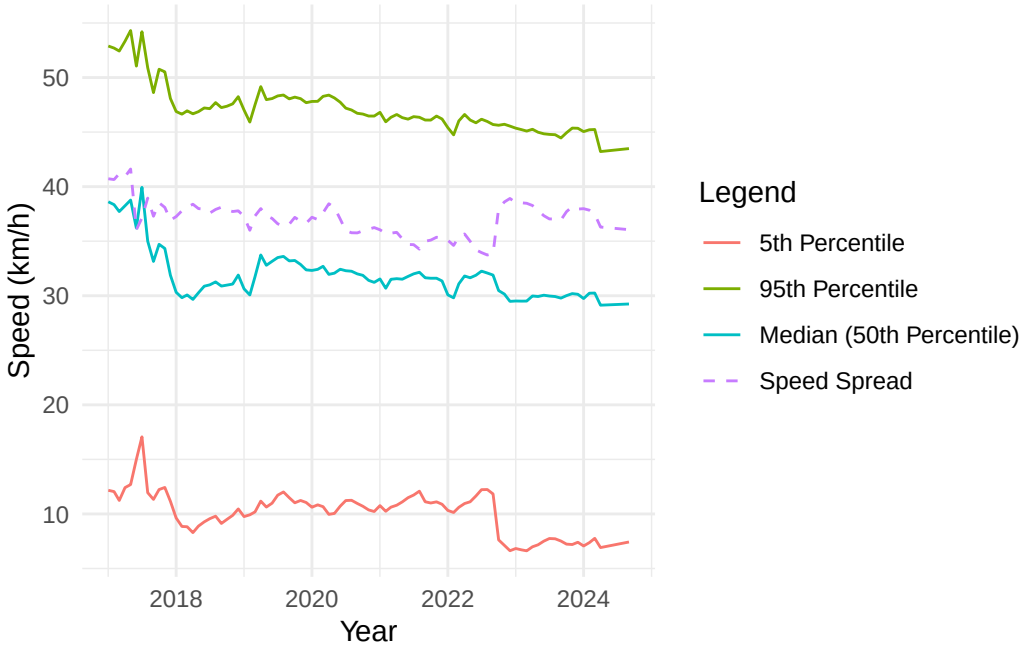


Figure 2: Trends in key speed percentiles (5th, 50th, and 95th) and the spread between the 95th and 5th percentiles over time.

2.9 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The purpose of this paper is to analyze the relationship between traffic volume, speed percentiles, and the mean speed of vehicles observed in Toronto. By utilizing a Bayesian linear regression model, we aim to investigate how various speed percentiles and total traffic volume predict the average vehicle speed (`spd_mean`). This model enables us to quantify the contributions of different speed measures and traffic density to overall traffic dynamics.

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

The Bayesian multiple linear regression model used in this paper predicts the mean speed (spd_mean) as a function of total traffic volume (volume) and speed percentiles (pct_05, pct_50, and pct_95). The model is defined as follows:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \cdot \text{volume}_i + \beta_2 \cdot \text{pct\_05}_i + \beta_3 \cdot \text{pct\_50}_i + \beta_4 \cdot \text{pct\_95}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\beta_4 &\sim \text{Normal}(0, 2.5) && (7)\\
\sigma &\sim \text{Exponential}(1) && (8)
\end{aligned}
$$

In the above model:

- μ_i is the predicted mean speed (spd_mean) for the i-th observation given the total traffic volume and speed percentiles.
- β_0 is the coefficient for the intercept.
- β_1 is the coefficient for the predicted change in the mean speed given a one-unit increase in traffic volume (volume).
- β_2 is the coefficient for the predicted change in the mean speed given a one-unit increase in the 5th percentile speed (pct_05).
- β_3 is the coefficient for the predicted change in the mean speed given a one-unit increase in the 50th percentile speed (pct_50).
- β_4 is the coefficient for the predicted change in the mean speed given a one-unit increase in the 95th percentile speed (pct_95).

- σ is the standard deviation of the residuals, representing unexplained variability in the mean speed.

We run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

The above model was chosen to investigate the relationship between the average speed (`spd_mean`) of vehicles and key predictors, including total traffic volume (`volume`) and speed percentiles (`pct_05`, `pct_50`, and `pct_95`). These variables were selected because they provide a comprehensive view of traffic patterns and vehicle behavior across different speed thresholds.

The percentiles represent critical thresholds of speed distribution, with the 5th percentile (`pct_05`) capturing the slowest vehicles, the 50th percentile (`pct_50`) representing the median speed, and the 95th percentile (`pct_95`) reflecting the fastest vehicles. These thresholds are expected to influence the average speed significantly. Similarly, total traffic volume (`volume`) accounts for overall road congestion and is anticipated to impact the mean speed inversely, as higher volumes may reduce average speeds due to congestion.

We employ a Bayesian linear regression model implemented using the `rstanarm` package (**citerstanarm?**). This approach provides a probabilistic framework for parameter estimation and allows the incorporation of prior information. Default priors were used for the regression coefficients (`Normal(0, 2.5)`) and residual standard deviation (`Exponential(1)`), as they are appropriate for general regression problems without overly constraining the estimates.

The linearity assumption is reasonable given the nature of traffic data, where incremental changes in predictors like percentiles and volume are expected to have proportional effects on the mean speed. However, this model does not capture potential nonlinearities or interactions, which could be explored in future work if evidence suggests their relevance.

We hypothesize that:

- Higher 5th, 50th, and 95th percentile speeds will correspond to higher average speeds.
- Higher traffic volumes will be associated with lower average speeds due to congestion effects. This model's simplicity balances interpretability with analytical rigor, making it a suitable choice for understanding traffic patterns and their determinants.

Table 5: Model summary of the predicted impact of traffic volume and speed percentiles on mean speed.

	(1)
(Intercept)	−0.881
volume	0.000
pct_05	0.206
pct_50	0.652
pct_95	0.184
Num.Obs.	31 990
R2	0.992
R2 Adj.	0.992
Log.Lik.	−41 414.100
ELPD	−41 436.5
ELPD s.e.	385.2
LOOIC	82 872.9
LOOIC s.e.	770.4
WAIC	82 872.9
RMSE	0.88

4 Results

4.1 Model Coefficients

The results of the Bayesian linear regression model estimating the relationship between mean traffic speed (spd_mean) and its key predictors are summarized in Table 5.

The coefficient estimates reveal that all predictors significantly contribute to the estimation of mean speed (`spd_mean`). The results indicate:

Traffic Volume (`volume`): Higher traffic volumes are associated with slightly lower mean speeds, consistent with the expectation that congestion reduces average traffic speeds. 5th Percentile Speed (`pct_05`): The lowest observed speeds strongly correlate with mean speed, suggesting a substantial influence of slower-moving vehicles on the overall mean. 50th Percentile Speed (`pct_50`): The median speed demonstrates the largest positive relationship with mean speed, aligning with its central role in determining the distribution of traffic speeds. 95th Percentile Speed (`pct_95`): Faster-moving vehicles also positively impact the mean speed, but the effect

is less pronounced than that of the median speed. These results highlight the combined importance of different speed thresholds and traffic volume in shaping overall traffic behavior.

4.2 Actual vs. Predicted Mean Speeds

To evaluate the predictive performance of our Bayesian linear regression model, we compare the predicted mean speeds, derived from the model, with the actual observed mean speeds in the dataset. The comparison is visualized in Figure 3.

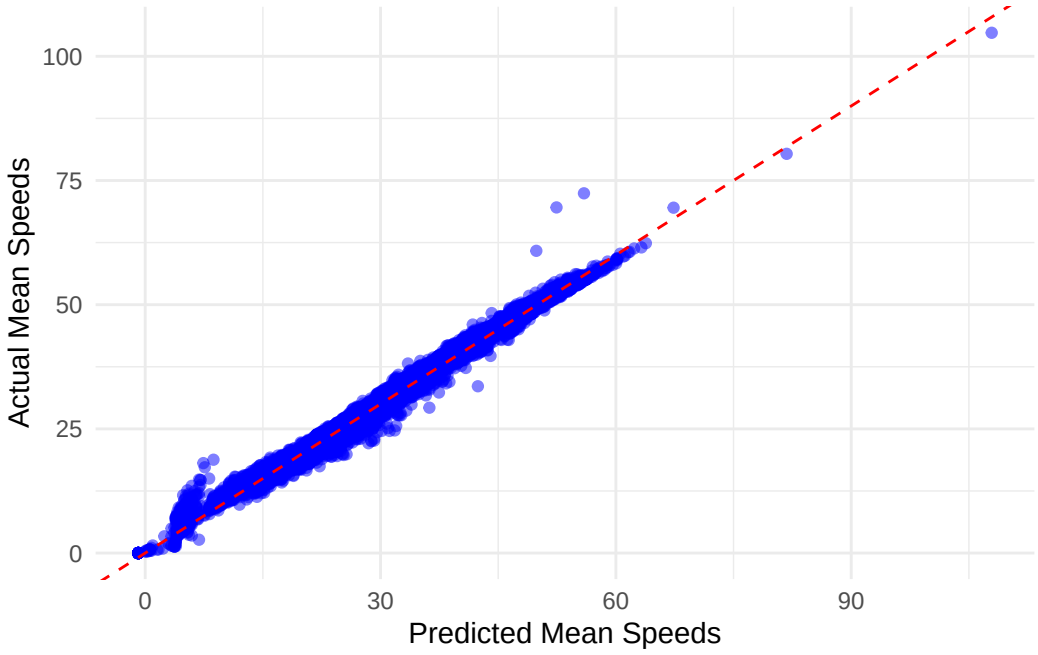


Figure 3: Comparison of actual and predicted mean speeds.

In Figure 3, each point represents an observation where the x-axis denotes the predicted mean speed, and the y-axis shows the actual mean speed. The red dashed line represents the ideal case where predictions perfectly match the actual values (i.e., a 45-degree reference line). Observations closely aligning with the dashed line indicate a strong agreement between predicted and actual values.

The model demonstrates strong predictive accuracy, as the majority of points cluster tightly along the dashed line with minimal deviation. This close alignment suggests that the model captures the relationship between the predictors (`pct_05`, `pct_50`, `pct_95`, and `volume`) and the response variable (`spd_mean`) effectively, with minimal systematic bias.

Notably, there are only a few points where slight deviations from the line are visible, mostly at higher mean speed values. This could be attributed to minor outliers or unique conditions not fully explained by the model's predictors. Overall, the results confirm the reliability of the Bayesian model in predicting mean traffic speeds, especially for the typical speed ranges observed in the dataset.

We explore the broader implications of these results, including potential improvements or limitations in the model, in Section 5.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

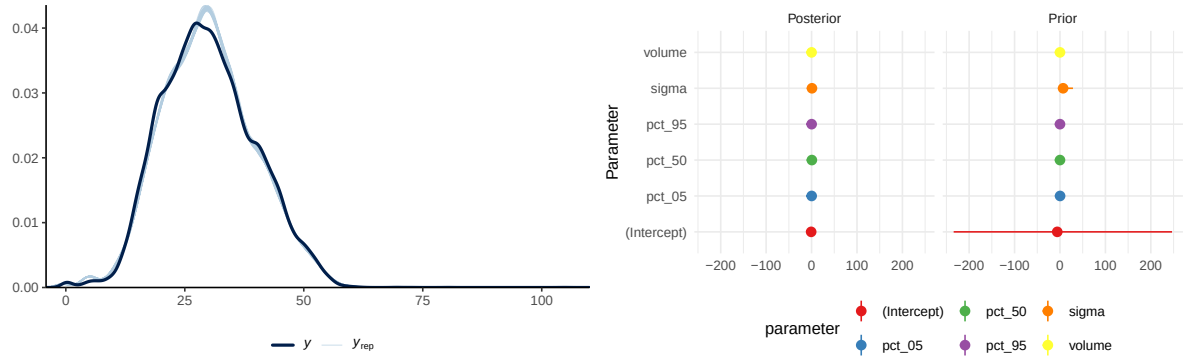
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check. This shows...

In Figure 4b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

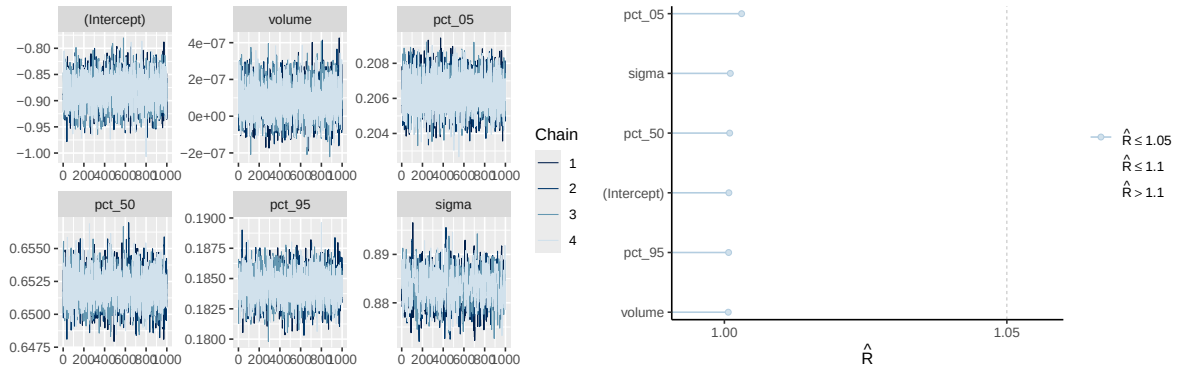
(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot. It shows... This suggests...

Figure 5b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.