My title*

My subtitle if needed

First author Another author

November 26, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

1.1 Estimand

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 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

In this paper, a Bayesian logistic model will be utilized to understand how factors such as application type, submission date, and planning district influence the decisions of the application committee. This model estimates the probability of an application being approved (P(Approval)) based on the given predictors.

The goal of this modeling strategy is twofold: (1) to estimate how each predictor affects the likelihood of approval, and (2) to quantify the uncertainty associated with these estimates. This approach is well-suited for incorporating prior knowledge while addressing the hierarchical nature of the predictors, such as geographic regions.

Here we briefly describe the Bayesian analysis model used to investigate, and include justification for model and the variables, as well as discuss underlying assumptions, potential limitations, software used to implement the model, and evidence of model validation and checking.

Background details and diagnostics are included in Appendix B.

3.1 Model set-up

In the Bayesian logistic regression model, the response variable y follows a Bernoulli distribution, reflecting its binary nature. Coefficients β_k (where k=0,1,2,3) of every predictor variables follow a Normal distribution. Specifically:

$$
\begin{aligned}
y_i | p_i &\sim \text{Bern}(p_i) && (1)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (2)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (5)
\end{aligned}
$$

Combining all the components, the complete model can be expressed as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \times \text{application type} + \beta_2 \times \text{submission date} + \beta_3 \times \text{planning district} \quad (6)$$

- p_i: the probability for the binary outcome variable to be 1 (1 = approval, 0 = refusal).
- β_0: Intercept, representing the baseline log-odds of approval for the reference levels of the predictors.
- β_1: Coefficients for application type, dummy-coded with CO as the reference category.
- β_2: Coefficient for date, measured as the year of application to capture temporal trends.
- β_3: Coefficients for planning district, dummy-coded with North York as the reference category.

In the above Bayesian logistic regression model, the coefficients $\beta = \beta_0, \beta_1, \beta_2, \beta_3$ are treated as random variables, each following a specified prior distribution. These priors represent our initial beliefs about the plausible values of the coefficients before observing the data. In this model, each β_k (where k=0,1,2,3) is assigned a normal distribution, with a zero mean and 2.5 standard deviation. The choice of these parameters reflects a weakly informative prior, centered around zero to express no strong prior belief about the direction or magnitude of the effects, while allowing for reasonable variability in the coefficient estimates.

The normal distribution is particularly suitable for these priors because it reflects a belief that most effects are likely to be small or moderate, centered around zero, while allowing for deviations in either direction. The standard deviation of 2.5 is a practical choice for weakly informative priors, offering a balance between constraining the parameters and allowing the data to influence the estimates. This ensures that the model avoids overfitting, particularly when data is sparse or multicollinearity exists among predictors.

After observing the data, the posterior distributions of β are obtained by combining these priors with the likelihood of the observed outcomes. The resulting posterior distributions provide updated beliefs about the coefficients, incorporating both prior knowledge and evidence from the data. This approach allows for robust inference, offering a clear quantification of parameter uncertainty and supporting more nuanced interpretations of the predictors' effects.

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 Bayesian logistic regression model is an appropriate choice for analyzing how application type, submission date, and planning district influence the likelihood of application approval. This choice is driven by the binary nature of the response variable, the model's ability to incorporate prior knowledge, and its probabilistic framework, which quantifies uncertainty more effectively than frequentist alternatives. Logistic regression naturally constrains predicted probabilities to the [0,1] interval, ensuring interpretability as likelihoods of approval, while the Bayesian approach enhances flexibility and provides richer insights through posterior distributions.

Compared to traditional generalized linear models (GLM), the Bayesian approach allows for the inclusion of prior beliefs through weakly informative priors ($N(0, 2.5)$) for all coefficients. These priors help stabilize the model, particularly in cases where data is sparse or multicollinear, without overwhelming the contribution of the observed data. Additionally, Bayesian methods yield full posterior distributions rather than single-point estimates, allowing for a clearer understanding of parameter uncertainty and a better capacity to incorporate uncertainty into decision-making processes. These advantages make the Bayesian logistic regression model preferable to a standard GLM, which provides point estimates and relies on asymptotic approximations for inference.

The model also stands out against simpler approaches, such as simple linear regression (SLR). While SLR might be used for binary outcomes in certain contexts, it is theoretically inappropriate because it does not constrain predictions to the [0,1] interval. This can result in nonsensical predicted probabilities outside this range. Moreover, SLR assumes a linear relationship between predictors and the response variable, which is unsuitable for binary outcomes. The logit transformation in logistic regression, by contrast, ensures a proper probabilistic framework while preserving interpretability in terms of log-odds.

The predictors included in the model further justify the choice of Bayesian logistic regression. Application type captures fundamental differences in how various categories of applications, such as minor variances and consents, might affect approval outcomes. Submission date is treated as a continuous variable, reflecting temporal trends without arbitrary grouping, while planning district accounts for geographic variability through categorical indicators. These predictors are naturally suited to a logistic framework, and the Bayesian approach accommodates any inherent variability in their effects.

In summary, Bayesian logistic regression provides a robust, clarified, and flexible framework for analyzing the factors influencing application decisions. Its advantages over GLM and SLR include the ability to incorporate prior information, quantify uncertainty, and handle the non-linear nature of the binary response variable. By balancing methodological rigor with practical transparency, this model ensures that the analysis remains both statistically sound and actionable for stakeholders.

3.1.2 Model Assumption

This Bayesian model relies on several assumptions that ensure the validity of its predictions and the transparency of its coefficients. While these assumptions are generally less restrictive than those for simpler models like linear regression, they remain critical to the robustness of the results. The key assumptions are listed below:

1. Binary Response Variable The model assumes that the response variable (y) is binary, taking values of either 1 (approval) or 0 (refused). This assumption aligns with the nature of the data, as application decisions are dichotomous outcomes. Any deviations from binary coding would invalidate the logistic framework and require alternative modeling approaches.

2. Independence of Observations: The model assumes that all observations are independent of one another. This is reasonable for application data, as each decision is typically made independently by the committee. However, if clustering or dependence exists (e.g., decisions within the same planning district are correlated), the model may need to incorporate hierarchical or random effects to address these dependencies.

3. No Perfect Multicollinearity Logistic regression assumes that the predictors are not perfectly correlated, as this would prevent the model from estimating unique coefficients. For example, if a predictor is a linear combination of other predictors, the model would fail to converge. In this analysis, categorical variables of application type and planning district are appropriately encoded to avoid such issues.

4. Proper Model Specification The model assumes that all relevant predictors are included and correctly specified. Omitting important variables or including irrelevant ones could lead to biased estimates or reduced transparency. For example, excluding interaction terms when they are theoretically justified might result in incomplete understanding of the predictors' effects.

3.1.3 Model Limitation

3.1.4 Model Implementation

4 Results

Our results are summarized in Table 1.

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.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	−4.53
	(0.88)
application_typeMV	−0.23
	(0.12)
date	0.00
	(0.00)
planning_districtNorth York	0.37
	(0.14)
planning_districtScarborough	0.01
	(0.18)
planning_districtToronto East York	0.17
	(0.15)
Num.Obs.	1940
R2	0.023
Log.Lik.	−1227.099
ELPD	−1233.1
ELPD s.e.	14.7
LOOIC	2466.3
LOOIC s.e.	29.5
WAIC	2466.2
RMSE	0.47

Appendix

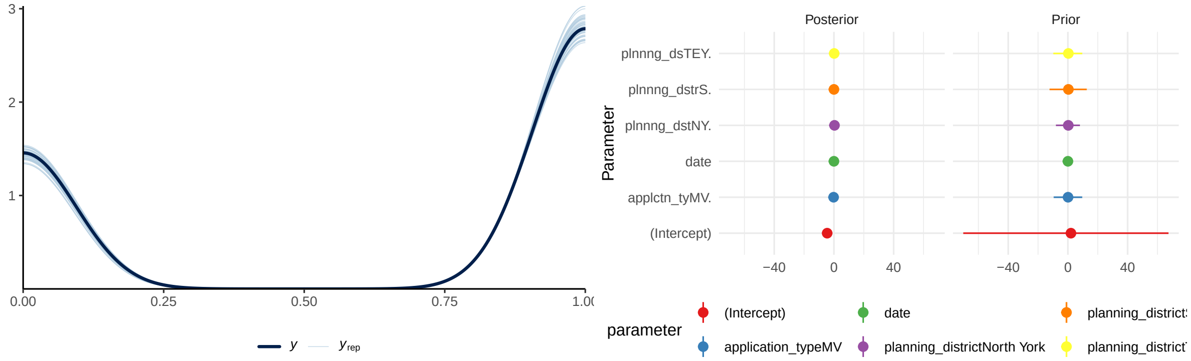
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 1a we implement a posterior predictive check. This shows...

In Figure 1b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

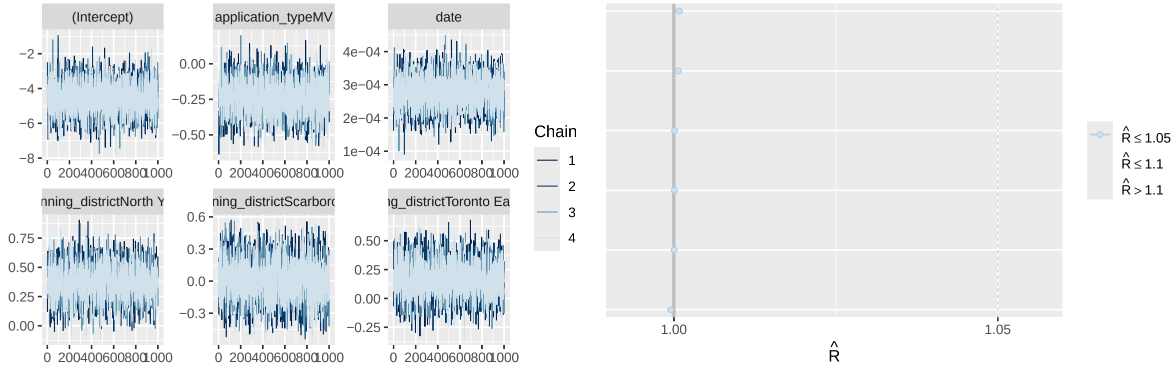
(b) Comparing the posterior with the prior

Figure 1: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 2a is a trace plot. It shows... This suggests...

Figure 2b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 2: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.