# My title*

**My subtitle if needed**

Tim Chen Steven Li Tommy Fu

October 20, 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….

## 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

### 2.2 Measurement

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

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

## 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 (Figure 1), from Horst, Hill, and Gorman (2020).

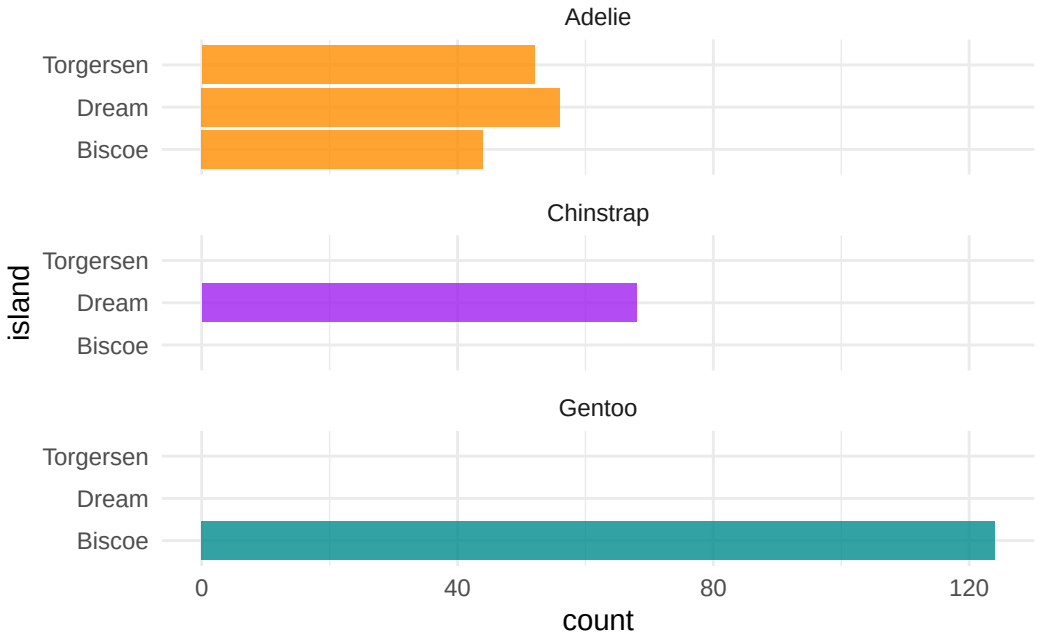


Figure 1: Bills of penguins

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

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in **?@sec-model-details**.

## 3.1 Model set-up

Define $y_i$ as the number of seconds that the plane remained aloft. Then $\beta_i$ is the wing width and $\gamma_i$ is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

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

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance $\theta$.

# 4 Results

Our results are summarized in Table 1.

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

| | First model |
|---|---|
| (Intercept) | 1.12 |
| | (1.70) |
| length | 0.01 |
| | (0.01) |
| width | −0.01 |
| | (0.02) |
| Num.Obs. | 19 |
| R2 | 0.320 |
| R2 Adj. | 0.019 |
| Log.Lik. | −18.128 |
| ELPD | −21.6 |
| ELPD s.e. | 2.1 |
| LOOIC | 43.2 |
| LOOIC s.e. | 4.3 |
| WAIC | 42.7 |
| RMSE | 0.60 |

# 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

## .1 Idealized Methodology and Survey Design for Forecasting the US Presidential Election

Objective: The goal of this methodology is to forecast the outcome of the U.S. presidential election using a representative sample, ensuring accuracy, transparency, and cost-efficiency with a budget of $100,000.

1. Sampling Approach 1.1. Target Population The target population is the U.S. voting-eligible population (VEP), which consists of all U.S. citizens over the age of 18 who are eligible to vote in the upcoming election. This includes both registered voters and likely voters, ensuring inclusivity in the survey while focusing more on likely voters, as they have the highest chance of participating in the election.

1.2. Sample Size Calculation Given a total population of over 250 million U.S. adults and assuming a margin of error of $\pm 2\%$ at a 95% confidence level, we aim for a sample size of n = 2,500 respondents. This will provide adequate precision in estimating key outcomes, such as the overall percentage of votes for the candidates.

Budget considerations: We estimate that survey responses can be collected at a cost of approximately $25 per respondent (including recruitment, incentives, and data validation costs), which gives a budget allocation of approximately $62,500 for data collection. 1.3. Sampling Frame We will use multi-stage stratified sampling to ensure a representative and balanced sample that reflects the demographics of the U.S. electorate.

Stage 1: Stratification by state (to ensure geographic representativeness), including oversampling battleground states like Pennsylvania, Wisconsin, Michigan, and Florida. Stage 2: Stratification within each state by demographic variables such as age, gender, race/ethnicity, and education. These are key factors that influence voting behavior and must be accurately represented. Stage 3: Random sampling of individuals within each stratum using lists from voter registration databases and online panel recruitment for hard-to-reach groups. This method ensures that all relevant subpopulations are sufficiently represented, and weighting adjustments can be made later if necessary to correct for any sampling imbalances.

2. Recruitment of Respondents 2.1. Recruitment Channels We will employ a mixed-method recruitment strategy:

Online Panels: Leveraging existing voter panels such as YouGov and Ipsos, we will recruit participants with a focus on capturing a diverse range of demographics. Random-Digit Dialing (RDD): To reach populations not heavily engaged online, especially older voters, we will conduct telephone interviews using RDD (both landline and cell phones). Incentives: Offering a small monetary incentive (e.g., $10 per completed survey) will encourage participation and

increase response rates. 2.2. Survey Mode A mixed-mode survey approach will be used to reach a broad range of participants:

Online surveys for digitally engaged respondents. Telephone surveys (both live interviews and Interactive Voice Response (IVR)) for those less likely to respond online, particularly among older voters and rural populations. 3. Survey Instrument Design 3.1. Survey Questions The survey will consist of a mix of closed-ended and open-ended questions, focusing on:

Voter preferences: Asking respondents whom they plan to vote for, using both traditional candidate name options and ranked-choice format (if applicable to state-specific elections). Voter likelihood: A question to assess how likely the respondent is to vote, with responses on a Likert scale (e.g., "How likely are you to vote in the upcoming election?"). Demographics: Standard questions on age, gender, race/ethnicity, education, income, and region to ensure representativeness. Issue-based questions: Gathering respondents' positions on key policy issues (e.g., economy, healthcare, climate change), as this may correlate with candidate preference. 3.2. Survey Length The survey will be concise to avoid fatigue, aiming for a completion time of no more than 10-15 minutes.

4. Data Validation and Quality Control 4.1. Response Validation Deduplication: Using respondent IP addresses, phone numbers, and other unique identifiers, we will ensure no duplicate responses are included. Screening questions: Filtering out respondents who are ineligible to vote or not part of the target population (e.g., non-U.S. citizens, individuals under 18). Attention checks: Including questions that help detect respondents who are not paying attention (e.g., "Select option 3 to continue"). 4.2. Weighting and Post-Stratification We will employ post-stratification weighting to adjust the data to match known population benchmarks, using factors such as:

Age Gender Race/Ethnicity Education level Geographic location (state) Voting history (using voter file data) This will help correct any sampling biases and ensure that our sample is demographically aligned with the general U.S. electorate.

Weighting will be performed using R's survey package, which allows us to adjust for complex survey designs.

4.3. Polling Aggregation To improve the robustness of our forecast, we will aggregate our poll results with other reliable polls. Poll aggregation will use a Bayesian model, weighting polls based on:

Sample size Methodology (online, IVR, live phone interviews) Pollster accuracy ratings (using metrics like FiveThirtyEight's pollster ratings) The model will also account for the recency of the polls and house effects (consistent biases shown by particular pollsters).

5. Data Analysis and Forecasting 5.1. Predictive Model We will employ a multilevel regression and post-stratification (MRP) model to forecast election outcomes. This model will leverage both individual-level survey data and state-level covariates (e.g., past voting behavior, demographic composition) to make predictions for each state.

```
# # Loading necessary libraries
# library(survey)
# library(dplyr)

# # Assuming 'poll_data' is a dataframe with survey responses

# # Define survey design
# design <- svydesign(ids = ~1, strata = ~state, weights = ~weight, data = poll_data)

# # Post-stratification weighting
# # Assume we have a dataframe 'pop_data' with known population distributions
# post_strat_weights <- postStratify(design, ~age + gender + race, pop_data)

# # Run the weighted analysis (e.g., estimate support for candidate)
# result <- svymean(~candidate_support, post_strat_weights)
# print(result)

# # Aggregate polls using Bayesian updating
# # Assume 'aggregated_polls' is a dataframe with multiple polls' data
# library(rstanarm)

# # Fit Bayesian model
# fit <- stan_glm(candidate_support ~ poll_method + recency, data = aggregated_polls, family
# summary(fit)

# # Predict election outcome
# predictions <- posterior_predict(fit, newdata = new_poll_data)
# print(predictions)
```

## .2 Budget Allocation

Category Cost Estimate Survey recruitment & incentives $62,500 Data validation & quality control $10,000 Poll aggregation & modeling software $7,500 Statistical consulting & data analysis $10,000 Miscellaneous (overheads, report writing) $10,000 Total $100,000

## 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/.