TBD*

TBD

TBD

31 October 2020

Abstract

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

1 Introduction

Here's a dumb example of how to use some references: In paper we run our analysis in `R` (R Core Team 2020). We also use the `tidyverse` which was written by Wickham et al. (2019) If we were interested in baseball data then Friendly et al. (2020) could be useful.

2 Data

Our data is of penguins (Figure 1).

Talk more about it.

Also bills and their average (Figure 2). (Notice how you can change the height and width so they don't take the whole page?)

Talk way more about it.

3 Model

The model we use to predict the result of the election is MRP, multilevel regression with post-stratification. Logistic regression allows us the simplicity of analyzing the categorical response variable while incorporating both numerical and categorical explanatory variables; while combined with post-stratification it yields a prediction result closer to the population parameter.

We are interested in how variables such as age, education, race, family income, and geographic region affect citizens' voting intentions. The response variable that we are interested in is the chances of a citizen voting for Trump in the coming presidential election. It is a categorical variable with two levels: 1 represents intending to vote for Donald Trump, and 2 means planning to vote for Joe Biden.

First, we train a multilevel logistic regression model using raw data obtained from the voter study group towards the variable of interest using the 7 explanatory variables. We then apply this model to our post-stratification data set to get the estimates.

*Code and data are available at: LINK.

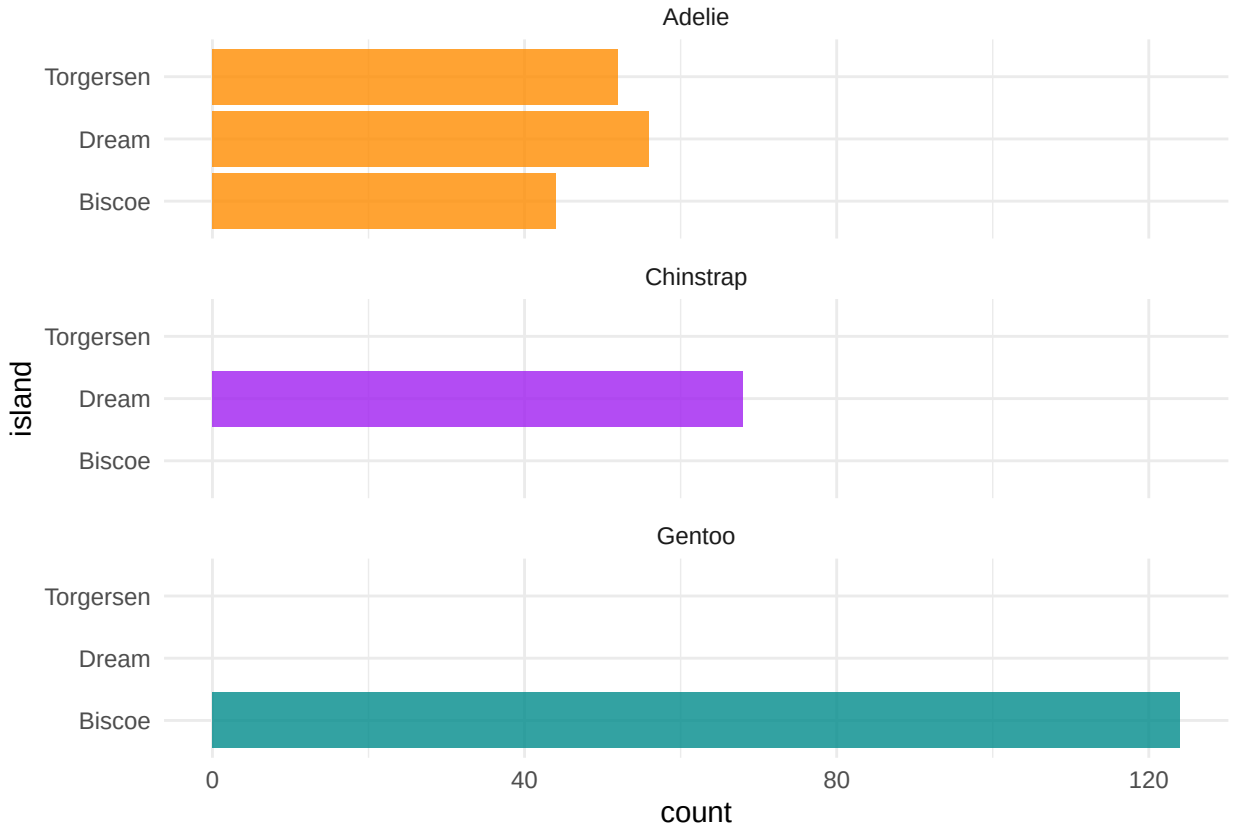

Figure 1: Bills of penguins

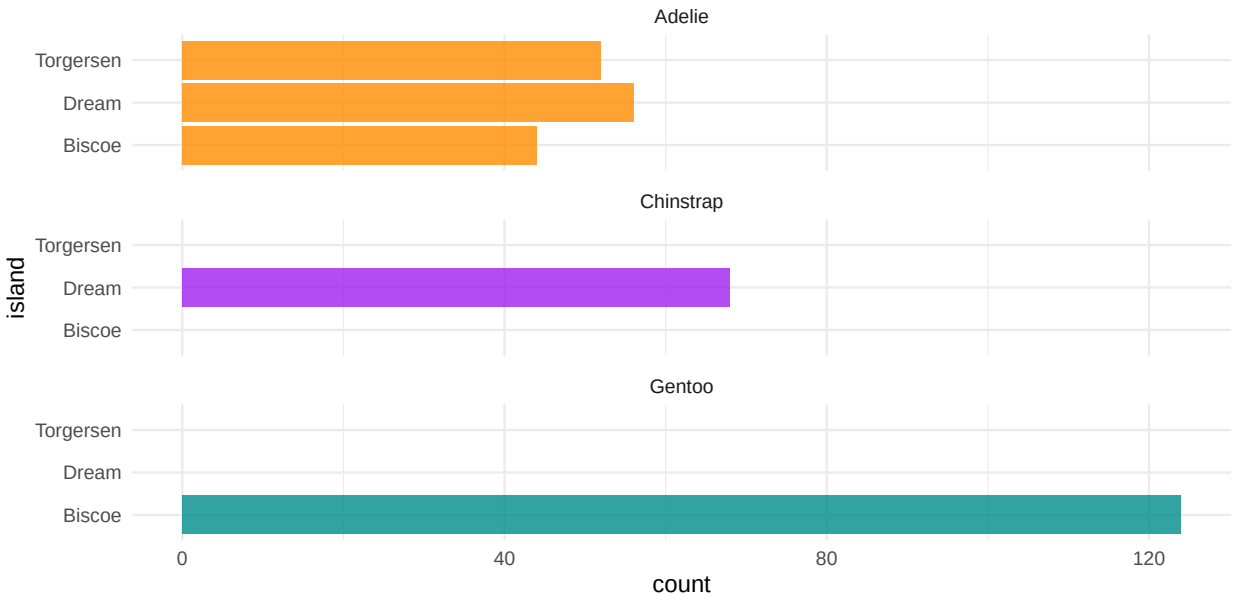

Figure 2: More bills of penguins

Logistic regression estimates $\beta_0 ... \beta_k$ in equation (1)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1}$$

where p is the probability of event A that we are interested in, $\beta_0$is the intercept, $x_1 ... x_K$ are our variables of interest and $\beta_1 ... \beta_k$ are parameters for each of these variables. Based on the result, we are able to estimate p for a particular case given all the variables.

The logic behind post-stratification is that we divide the sample data into strata based on a few attributes such as state, income, education level. Then we come up with a weight for each stratum to adjust its influence towards our prediction. Namely, if a stratum is over-represented, we want to reduce its impact on the prediction result; if a stratum is under-represented, we want to increase its influence. We need a post-stratification dataset that is large enough to do this. In this report, we use the ACS(reason).

The post-stratification estimate is defined by $\hat{y}_{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. An illustration using the education variable is $\hat{y}_{edu}^{PS} = \frac{\sum_{j \in J_{edu}} N_j \hat{y}_j}{\sum_{j \in J_{edu}} N_j}$, we can get an estimation given any level of education of respondents, J_{edu} denotes all subsets of education levels, $\cup_{i \in J} J_i^{edu}$ = all possible education levels.

The multilevel regression before post-stratification produces a proportion for the post-stratification based on regression, instead of merely averaging the sample in each stratum. Specifically, all possible combinations of cells are from combining:

-
-
-
-
-
-

We fit the logistic regressions for estimating candidate support in each cell. We used `function` from `package` to fit (2).

$$P(Trump) = logit^{-1}(\beta_0 + \beta_a ge + \beta_{j[i]}^{edu} + \beta_{j[i]}^{race} + \beta_{j[i]}^{income} + \beta_{j[i]}^{state} + \beta_{j[i]}^{mask}) \tag{2}$$

(2) is the model that we fit using the survey data, with $\beta_{j[i]}^{var} \sim N(0, \sigma_{var}^2)$. β_0 is the intercept,each of $\beta_{j[i]}^{var}$ represents the parameter for the i-th respondent in j-th cell. For example, $\beta_{j[i]}^{state}$ can take values from β for all states in the US. Then the model sums up estimates for each cell to the population.

We find the mask parameter interesting, firstly because US is the country with the most COVID cases in the world, the influence of COVID on election choices is inevitable. Secondly, since Trump and Biden have very opposites towards the pandemic and how to handle it, this parameter will likely represent part of respondent's voting intention in the same way as any more common political indicators.

4 Results

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

5.3 Third discussion point

5.4 Weaknesses and next steps

5.4.1 1: Nested Bayesian Model /,

Appendix

References

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean "Lahman" Baseball Database*. https://CRAN.R-project.org/package=Lahman.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.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.