TBD*

TBD

Kar Hian Ong and Fadlan Arif

01 November 2020

Abstract

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

1 Introduction

With the US federal election 2020 swinging by, there is intense debate on who will be the upcoming US president of the United States. The election uses the electoral college system where each state is given a certain number of votes. Candidates with the 270 votes or more will win the election. Us is govern by 2 major parties, Democrats and Republicans. The candidate representing Democrats is Joe Biden while the candidate representing Republicans is Donald Trump. Both candidates have held positions prior to the 2020 election making them a strong contender for this election. Donald Trump is the current sitting president of the United States while Joe Biden was the vice president of the previous administration.

This research is about predicting who will win the popular vote for the US election. The election result has massive impact for the global population. The sitting president will set the tone for international and domestic policies.

Our approach towards handling our data is through the scope of logistic regression. Logistic regression allows us to look at a binary variable, to see the probability of that event happening ot not. We chose logisitic regression as the majority of the votes land on the democratic or republican candidates thus making that our variable of interest. We mainly focus on those planning to vote for these candidates, as these willl most likely be the deciding vote in the election. After narrowing down the election to two options, have turned the variable of interest into a binary one, thus allowing us to use logistic regression. The formula we use is:

$$log(\frac{p}{1-p}) = B_0 + B_1 x_1 + ... + B_k x_k \tag{1}$$

Where p is the probability that the event of interest occuring (voting for Joe Biden), B_0 is the y-intercept and $B_i, 1 \leq i \leq k$, coefficient represents change in log odds for one unit increase in x_i.

Within this paper we look at the probability of and individual voting for Joe Biden, stratified by their age group. We were able to stratify by cross examining the survey data about election opinions with the US census for 2018. With this information, we were able to more accurately represent the survey data to help it be proportionate to the population of America.

There are several sections to this paper. Section 2 will talk about the survey data and ACS data, section 3 will detail the model that us use to predict the winner of the US election, section 4 we will get to see the results, section 5 will be discussion on the findings and weaknesses.

*Code and data are available at: https://github.com/karhian/2020-election

2 Data

The survey data is provided by Nationscape (Tausanovitch and Vavreck (2020)).Nationscapre conducts interview to 50,000 americans corvering the campaign and election. All responses take the survey online and must complete an attention check before taking the survey. Nationscape handles non-response by using a set of weights. The weights are devived from the the adult population of the 2017 American Community Survey of the U.S. Census Bureau.

The stratification dataset is by IPUMS USA (Steven Ruggles and Sobek (2020)).

The cleaning of survey data and stratification data uses `haven` (Wickham and Miller (2020)) package to read the data

Our data is of penguins (Figure 1).

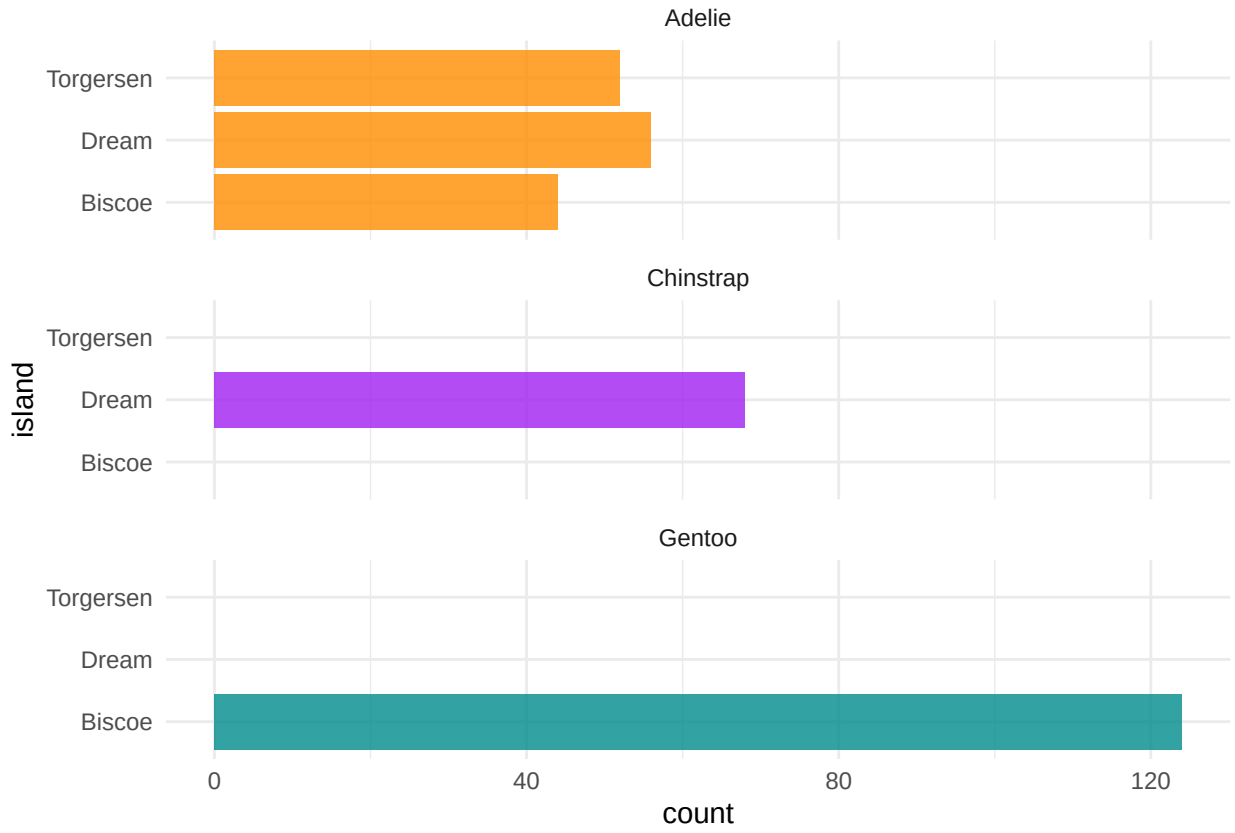

Figure 1: Bills of penguins

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.

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \tag{2}$$

Equation (2) seems useful, eh?

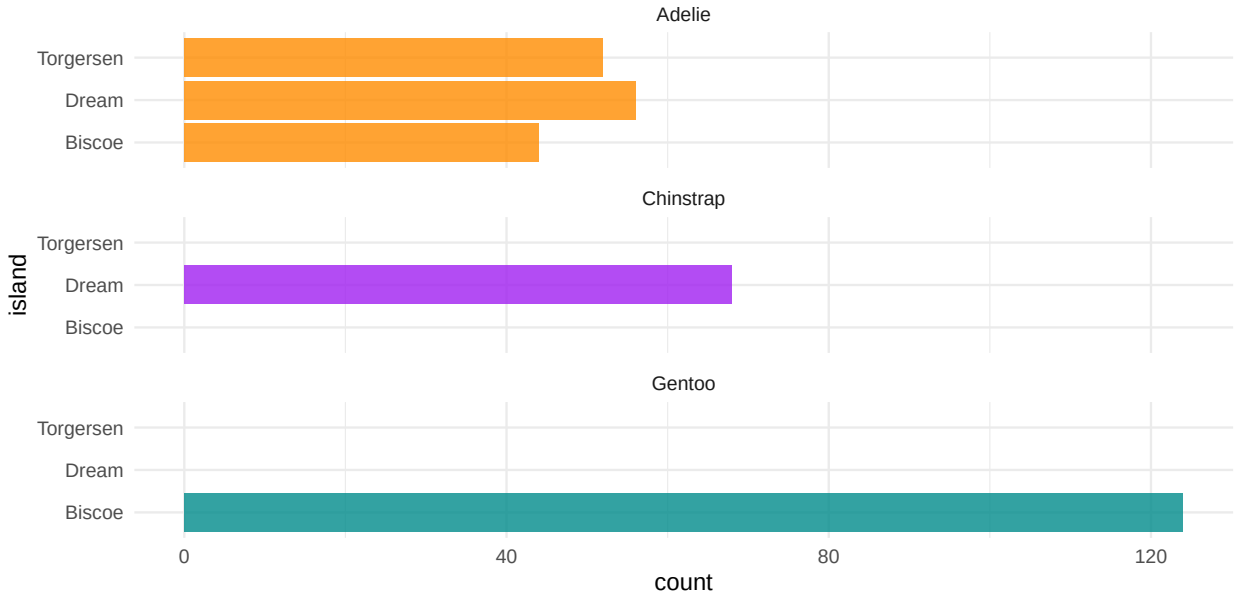

Figure 2: More bills of penguins

3 Model

After molding the data to our needs, we compute the generalised linear model with vote2020_bin as the dependent variable and age and education level being out independent variable.

We then called onto the summary(first_logit) function to retrieve all the needed coefficients and assigned simpler variable names to each value to form our regression formula:

- b0 <- first_logit$coef[1] #intercept
- age <- first_logit$coef[2]
- doctorate <- first_logit$coef[3]
- gradeschool <- first_logit$coef[4]
- masters

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

Weaknesses and next steps should also be included.

Appendix

References

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. *IPUMS Usa: Version 10.0 [Dataset]. Minneapolis, Mn: IPUMS*. https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. *Democracy Fund + Ucla Nationscape,October 10-17, 2019 (Version 20200814)*. https://www.voterstudygroup.org/publication/nationscape-data-set.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files*. https://CRAN.R-project.org/package=haven.