Mini Essay 10[*]

Julian Marrero

March 19, 2024

1 Building the simplified model

Considering (Maher 1982), (Smith 2002), or (Cohn 2016), here is a simplified approach to the model.

1.1 Model Construction

Apply a regression model using the election data, treating votes for Buchanan as a function of demographic variables, and identify Palm Beach as an outlier. Create a simplified model that accounts for polling variability due to methodological decisions, similar to the variability observed across different pollsters in the NYT experiment.

1.2 Incorporation of Variability

Introduce a term or parameter to account for potential ballot design effects (like the "butterfly ballot"), possibly through interaction terms or a separate model component for affected counties. Emphasize the role of decision-making in pollster methodology (weight adjustments, likely voter models) and its impact on polling outcomes, as seen in the variance among the pollsters' results with the same data.

[*] Code for R scripts are available at: https://github.com/julianmarrero/STA302-Mini-Essay-10

2 Simplification and Integration

Combine the insights from both analyses to understand how both election outcomes and polling predictions can be influenced by factors beyond raw numbers—such as ballot design and pollster methodology decisions. Highlight the importance of considering and modeling these external factors to improve the accuracy and reliability of both election predictions and polling analyses.

3 Estimation of Model

3.1 Obtaining Dataset

I do not have access to a dataset for Florida's election and polling, however a dataset containing the election votes of Toronto officials can be openly accessed from the Open Data Toronto website. A dataset containing the election votes for Toronto officials is included in the Data folder, and is used for the estimation of the model. The dataset can be found at the following path of the repository

`STA302-Mini-Essay-10/Data/2022_Toronto_Poll_By_Poll_All_Offices.xlsx`

3.2 Estimating the Model

The R script conducting the estimation of the model can be found at the following path of the repository:

`STA302-Mini-Essay-10/Scripts/01-model_estimation.R`

4 Choice of Regression Type

Choosing between logistic, Poisson, and negative binomial regression models depends on the nature of the dependent variable you're analyzing and the specific characteristics of your data. Here's a brief discussion on when to use each.

4.1 Logistic Regression

Logistic regression is best suited for binary outcome variables (e.g., win vs. lose, yes vs. no). It's used when you're interested in predicting the probability of a binary outcome based on one or more predictor variables. It directly models the probability of the outcome; coefficients can be interpreted as odds ratios after exponentiation (Hosmer Jr., Lemeshow, and Sturdivant 2013). A limitation is that the model assumes a linear relationship between the logit of the outcome and the predictors. It can only be used with dichotomous dependent variables (Hosmer Jr., Lemeshow, and Sturdivant 2013).

4.2 Poisson Regression

Poisson regression is ideal for count data that represent the number of times an event occurs in a fixed interval of space or time. It assumes that the data follow a Poisson distribution where the mean and variance are equal (Dobson and Barnett 2008). An advantage is that the model is specifically tailored for count data, allowing for the modeling of rates and counts where the occurrences of events are independent. However, a limitation is the assumption that the mean equals the variance (equidispersion) often doesn't hold in real data, leading to overdispersion which Poisson regression cannot adequately address (Dobson and Barnett 2008).

4.3 Negative Binomial Regression

Negative binomial regression is a more flexible alternative to Poisson regression that is used for count data susceptible to overdispersion, where the variance exceeds the mean (Hilbe 2011). It introduces an extra parameter to model the overdispersion. The model can handle overdispersed count data effectively, making it more suitable for real-world data where the Poisson assumption of equidispersion is violated. The interpretation of coefficients is less straightforward compared to Poisson regression. The model is also more complex due to the additional dispersion parameter (Hilbe 2011).

4.4 Regression Type Used in Estimation

In my R script `01-model_estimation.R`, I used Poisson regression due to its fit of the dataset. However, I checked the dataset for overdispersion using the declared variable `dispersion_param` to assess the adequacy of the Poisson model for the given data by checking for overdispersion. If the dispersion parameter is significantly greater than 1, then I would consider using negative binomial regression.

References

Cohn, Nate. 2016. "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results." https://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html.

Dobson, Annette J., and Adrian Barnett. 2008. *An Introduction to Generalized Linear Models*. 3rd ed. Chapman; Hall/CRC.

Hilbe, Joseph M. 2011. *Negative Binomial Regression*. 2nd ed. Cambridge University Press.

Hosmer Jr., David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. John Wiley & Sons.

Maher, M. J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36 (3): 109–18. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

Smith, Richard L. 2002. "A Statistical Assessment of Buchanan's Vote in Palm Beach County." *Statistical Science* 17 (4): 484–89. https://doi.org/10.1214/ss/1049993203.