Essay

Jena Shah

Poisson regression is a type of regression analysis that is used to model count data. In it, the outcome variable is a count of events or occurrences in a fixed period of time or space. It is based on the Poisson distribution, which describes the probability of observing a certain number of events in a fixed interval, given a known average rate of occurrence.

This analysis aims to look at the frequency of the letter 'e' in the first ten lines of each chapter of *Wuthering Heights* by Emily Bronte. Poisson regression is suitable for this analysis because the outcome variable is the count of 'e's in each line depending on the words per line, which is a count variable. This fits the criteria for Poisson regression.

Also, Poisson regression estimates the rate at which an event occurs, which in this case is the rate of 'e's per line depending on words per line. This helps with understanding the distribution of 'e's in the text. Poisson regression allows for modeling the relationship between the count of 'e's and the number of words in a line. This relationship is central to the analysis in the code.

Overall, Poisson regression is a suitable choice for this analysis due to its ability to model count data and estimate rates of occurrence, making it well-suited for understanding the distribution of 'e's in *Wuthering Heights*.

Code

Simulate data. The code first simulates data (`count_of_e_simulation`) to represent the number of words in a line and the number of 'e's in the first ten lines of each chapter.

Download data. It then downloads the text of *Wuthering Heights* using the `gutenberg_download` function.

Clean data. The data of the text is then cleaned and processed (`wutheights_reduced`) to extract the first ten lines of each chapter and count the number of 'e's in each line.

Model data. Finally, the code uses Poisson regression (`stan_glm`) to model the relationship between the number of words in a line (`word_count`) and the number of 'e's in the line (`count_e`), showing that the frequency of 'e's varies with the length of the line.

The model, along with all other parts of this mini-essay can be found here.

References

(Alexander 2023) (R Core Team 2023) (Wickham 2016) (Müller and Wickham 2023) (Arel-Bundock 2024)

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.