Alexis J. Carr
Malik Nabeel Aamer
Ruiqi Guo
Statistical Computing in R & Python
20 March 2019

Project Proposal: Implementing a Zero-Inflated Poisson Regression Package in Python

Background

Count data, at times, may prove difficult to run standard statistical analyses on, because a prevalence 0's that skews the dataset. Zero-inflated Poisson regression solves this problem by allowing for the presence of greater variability in a dataset than would be expected based on a given statistical model. This model assumes that a sample is a mixture of two individual sorts — one of whose counts are generated through standard Poisson regression. The other group can be termed as absolute zero, where there exists zero probability of a count greater than 0. Observed values of 0 could come from either group. This model is most suitable for application when a conventional negative binomial model might not be a good fit. Its application ranges from software fault prediction to evaluation of prognostic factors of certain diseases like hepatitis C.

This model is easily implemented in R via the `pscl` package. However, there is currently not a Python implementation of a similar package. Our project seeks to translate the zero-inflated model functionality of the pscl package into Python.

Objectives

For this project, we aim to accomplish the following goals. Goals I and II constitute our minimum viable product, while we aim to complete Goal III if time permits.

- **Goal I. Translate the R `pscl` package function `zeroinf`, namely the sub-functions listed below.** Said sub-functions will use the results of `GLM.fit` for their starting values. We will limit ourselves to implementing the `logit` link for simplicity.
    - likelihood and log-likelihood (handled by `ziPoisson` in `pscl`)
    - gradient likelihood (handled by `gradPoisson` in `pscl`)
    - maximum likelihood
    - print function that has a similar output to that in R

- **Goal II. Compare the functionality of our package for both *veracity* and *speed* against the R *`pscl`* package.**
    - The veracity of the results will be tested on the datasets supplied with the `pscl` package, including but not limited to the absentee ballot dataset.
    - The speed will be clocked using the linux command-line tool `time` to time both the R and Python packages real runtimes via the terminal.

- **Goal III. Implement at least one of the following:**
    - zero-inflated negative binomial regression
    - zero-inflated geometric regression

Sources

Jackman, S., Tahk, A., Zeileis, A., Maimone, C., Fearon, J., Meers, Z., ... & Imports, M. A. S. S. (2017). Package 'pscl'. See http://github. com/atahk/pscl.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of statistical software, 27(8), 1-25.