Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

aggregress: A Package for Regression Analysis with Aggregated Data

Lealand Morin
University of Central Florida

Abstract

Package **aggregress** makes adjustments to the lm object output when the inputted data are in aggregated form. That is, when the data frame has only the unique values of all variables and the frequencies of these observations recorded in a column of weights. For example, in the case of linear regression, the resulting object is indistinguishable from that from the **lm** object original unaggregated data. It includes adjustments to the statistics and diagnostics for a regression model.

Keywords: regression, lm, glm, aggregate, group by.

1. Dealing with Aggregate Data in R

Package **aggregress** makes adjustments to the lm object output when the inputted data are in aggregated form. That is, when the data frame has only the unique values of all variables and the frequencies of these observations recorded in a column of weights. For example, in the case of linear regression, the resulting object is indistinguishable from that from the **lm** object original unaggregated data. It includes adjustments to the statistics and diagnostics for a regression model.

To quote the help files for the `lm` function: "Therefore, the sigma estimate and residual degrees of freedom may be suboptimal; in the case of replication weights, even wrong. Hence, standard errors and analysis of variance tables should be treated with care." This packages makes adjustments so that the sigma estimate and residual degrees of freedom are not suboptimal or wrong; they are correct.

There are several R packages (R Core Team 2017) available for working with aggregated data. The packages **markovchain** (Spedicato, Kang, Yalamanchi, Yadav, and Cordon 2020) is not one of them.

Figure 1: Caption goes here

2. Model

Is there really a model?

It's more about what others are missing.

Computational benefit: $O(\tilde{n})$ vs. $O(n)$, where n is the sample size and $\tilde{n}$ is the number of rows in the aggregated dataset. This number is defined by `n_tilde <- length(unique(data[, column_names))` where `column_names` is a character vector of the names of the depenedent and independent variables in the regression model.

This can considerably reduce the memory requirements for estimating a model. This is especially true when using categorical variables, since the length of the dataset is limited by the number of permutations of the discrete dataset.

There is a provision to categorize a continuous variable onto a discrete grid of values. However, it is well known that this sort of measurement error in the covariates can lead to a downward bias in the magnitude of the regression coefficients. To account for this, the bias correction of () is applied.

What changes with the linear model, the linear probability model and logistic regression?

What adjustments are made for heteroskedasticity?

What adjustments are made for measurement error induced by aggregation?

How can I calculate the AUROC for a logistic regression with aggregated data?

3. Example

A demonstration of analysis is shown in `aggregress_demo.R` and it serves as an example of what a typical session of model specification, estimation and testing can include. This procedure includes the following steps:

1. Organizing data
2. Choosing estimation options
3. Lag selection
4. Model estimation
5. Hypothesis testing

3.1. Organizing data

3.2. Choosing options

3.3. Lag-order selection

3.4. Model estimation

3.5. Hypothesis testing

4. Summary and discussion

This is a good package because...

Computational details

The results in this paper were obtained using R 3.5.1. with the **aggregress** package Version 0.0.0.9000. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

The development version of this package is available by using the **devtools** package, with which the latest version can be installed by

`devtools::install_github(LeeMorinUCF/aggregress)`.

Acknowledgments

To be amended: I am particularly grateful to .... For useful comments and helpful suggestions on earlier versions of this paper, I thank ....

References

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Spedicato GA, Kang TS, Yalamanchi SB, Yadav D, Cordon I (2020). *The **markovchain** Package: A Package for Easily Handling Discrete Markov Chains in R*. R package version 0.8.5, URL https://CRAN.R-project.org/package=markovchain.

Affiliation:

Lealand Morin
Department of Economics
University of Central Florida
E-mail: Lealand.Morin@ucf.edu