Package ‘dbglm’

October 13, 2022

Title Generalised Linear Models by Subsampling and One-Step Polishing

Version 1.0.0

Description Fast fitting of generalised linear models on moderately large datasets, by taking an initial sample, fitting in memory, then evaluating the score function for the full data in the database. Thomas Lumley <doi:10.1080/10618600.2019.1610312>.

Imports DBI, tidypredict, rlang, methods, tidyverse, dbplyr, vctrs, knitr, dplyr, purrr, tibble, tidyr, stringr

Suggests RSQLite, duckdb, bigquery, testthat (>= 3.0.0)

License MIT + file LICENSE

Maintainer Shangqing Cao <caoalbert@g.ucla.edu>

RoxygenNote 7.1.1

Encoding UTF-8

Depends R (>= 2.10)

Config/testthat/edition 3

NeedsCompilation no

Author Thomas Lumley [aut, cph], Shangqing Cao [ctb, cre]

Repository CRAN

Date/Publication 2021-06-23 08:00:02 UTC

R topics documented:

dbglm .................................................. 2
fleet1 ................................................... 3

Index 4
dbglm

Fast generalized linear model in a database

Description

Fast generalized linear model in a database

Usage

dbglm(formula, family = binomial(), tbl, sd = FALSE, weights = .NotYetImplemented(), subset = .NotYetImplemented(), ...)

Arguments

... This argument is required for S3 method extension.
formula A model formula. It can have interactions but cannot have any transformations except factor
family Model family
tbl An object inheriting from tbl. Will typically be a database-backed lazy tbl from the dbplyr package.
sd Experimental: compute the standard deviation of the score as well as the mean in the update and use it to improve the information matrix estimate
weights We don’t support weights
subset If you want to analyze a subset, use filter() on the data

Details

For a dataset of size N the subsample is of size N^(5/9). Unless N is large the approximation won’t be very good. Also, with small N it’s quite likely that, eg, some factor levels will be missing in the subsample.

Value

A list with elements
tildebeta coefficients from subsample
hatbeta final estimate
tildeV variance matrix from subsample
hatV final estimate

References

Data of vehicles registered in New Zealand as of November 2017

Description
Data of vehicles registered in New Zealand as of November 2017

Usage
data(fleet1)

Format
A tibble with 10000 rows and 34 variables:

- **basic_colour** character colour of the car
- **power_rating** numeric horsepower of the car
- **gross_vehicle_mass** numeric mass of the vehicle in kg
- **number_of_seats** numeric number of seats in the car

Source
Index

* datasets
  fleet1, 3

dbg1m, 2
dbsample (dgb1m), 2

fleet1, 3