Package 'pmsampsize'

July 23, 2025

Version 1.1.3

Date 2023-12-05

Title Sample Size for Development of a Prediction Model

Maintainer Joie Ensor <j.ensor@bham.ac.uk>

Depends R (>= 2.1)

Suggests stats

Description Computes the minimum sample size required for the development of a new multivariable prediction model using the criteria proposed by Riley et al. (2018) <doi:10.1002/sim.7992>. pmsampsize can be used to calculate the minimum sample size for the development of models with continuous, binary or survival (time-to-event) outcomes. Riley et al. (2018) <doi:10.1002/sim.7992> lay out a series of criteria the sample size should meet. These aim to minimise the overfitting and to ensure precise estimation of key parameters in the prediction model.

License GPL (>= 3)

RoxygenNote 7.2.3

Encoding UTF-8

NeedsCompilation no

Author Joie Ensor [aut, cre]

Repository CRAN

Date/Publication 2023-12-06 09:10:02 UTC

Contents

pmsampsize *pmsampsize - Sample Size for Development of a Prediction Model*

Description

pmsampsize computes the minimum sample size required for the development of a new multivariable prediction model using the criteria proposed by Riley *et al*. 2018.

Usage

```
pmsampsize(
  type,
  nagrsquared = NA,
  csrsquared = NA,
  rsquared = NA,
  parameters,
  shrinkage = 0.9,
  prevalence = NA,
  cstatistic = NA,
  seed = 123456,
  rate = NA,
  timepoint = NA,
  meanfup = NA,
  intercept = NA,
  sd = NA,
  mmoe = 1.1
)
```

Arguments

type	specifies the type of analysis for which sample size is being calculated • "c" specifies sample size calculation for a prediction model with a continuous outcome • "b" specifies sample size calculation for a prediction model with a binary outcome • "s" specifies sample size calculation for a prediction model with a survival (time-to-event) outcome
nagrsquared	for type="b" or type="s" this specifies the expected value of the Nagelkerke's R-squared of the new model, which is the Cox-Snell R-squared scaled to lie in the [0,1] range. It is interpretable in the same way as the standard R-squared, i.e. the percentage of variation in outcome values explained by the model. Please read the description of rsquared for additional details about specifying the expected R-squared performance
csrsquared	for type="b" or type="s" this specifies the expected value of the Cox-Snell R-squared of the new model. The Cox-Snell R-squared is the generalised version of the well-known R-squared for continuous outcomes, based on the likelihood.

Please read the description of rsquared for additional details about specifying the expected R-squared performance. The papers by Riley et al. (see references) outline how to obtain the Cox-Snell R-squared value from published studies if they are not reported, using other information (such as the C-statistic [see cstatistic() option below]).

rsquared — for type="c" this specifies the expected value of the R-squared of the new model, where R-squared is the percentage of variation in outcome values explained by the model. For example, the user may input the value of the R-squared reported for a previous prediction model study in the same field. If taking a value from a previous prediction model development study, users should input the model's adjusted R-squared value, not the apparent R-squared value, as the latter is optimistic (biased). However, if taking the R-squared value from an external validation of a previous model, the apparent R-squared can be used (as the validation data was not used for development, and so R-squared apparent is then unbiased). Users should be conservative with their chosen R-squared value; for example, by taking the R-squared value from a previous model, even if they hope their new model will improve performance.

parameters — specifies the number of candidate predictor parameters for potential inclusion in the new prediction model. Note that this may be larger than the number of candidate predictors, as categorical and continuous predictors often require two or more parameters to be estimated.

shrinkage — specifies the level of shrinkage desired at internal validation after developing the new model. Shrinkage is a measure of overfitting, and can range from 0 to 1, with higher values denoting less overfitting. We recommend a shrinkage = 0.9 (the default in pmsampsize), which indicates that the predictor effect (beta coefficients) in the model would need to be shrunk by 10% to adjust for overfitting. See references below for further information.

prevalence — (type="b" option) specifies the overall outcome proportion (for a prognostic model) or overall prevalence (for a diagnostic model) expected within the model development dataset. This should be derived based on previous studies in the same population.

cstatistic — (type="b" option) specifies the C-statistic reported in an existing prediction model study to be used in conjunction with the expected prevalence to approximate the Cox-Snell R-squared using the approach of Riley et al. 2020. Ideally, this should be an optimism-adjusted C-statistic. The approximate Cox-Snell R-squared value is used as described above for the csrsquared() option, and so is treated as a baseline for the expected performance of the new model.

seed — (type="b" option) specifies the initial value of the random-number seed used by the random-number functions when simulating data to approximate the Cox-Snell R-squared based on reported C-statistic and expect prevalence as described by Riley et al. 2020

rate — (type="s" option) specifies the overall event rate in the population of interest, for example as obtained from a previous study, for the survival outcome of interest. NB: rate must be given in time units used for meanfup and timepoint options.

timepoint — (type="s" option) specifies the timepoint of interest for prediction. NB: time units must be the same as given for meanfup option (e.g. years, months).

`meanfup` (`type="s"` option) specifies the average (mean) follow-up time anticipated for individuals in the model development dataset, for example as taken from a previous study in the population of interest. NB: time units must be the same as given for timepoint option.

`intercept` (`type="c"` options) specifies the average outcome value in the population of interest e.g. the average blood pressure, or average pain score. This could be based on a previous study, or on clinical knowledge.

`sd` (`type="c"` options) specifies the standard deviation (SD) of outcome values in the population e.g. the SD for blood pressure in patients with all other predictors set to the average. This could again be based on a previous study, or on clinical knowledge.

`mmoe` (`type="c"` options) multiplicative margin of error (MMOE) acceptable for calculation of the intercept. The default is a MMOE of 10%. Confidence interval for the intercept will be displayed in the output for reference. See references below for further information.

Details

`pmsampsize` can be used to calculate the minimum sample size for the development of models with continuous, binary or survival (time-to-event) outcomes. Riley *et al*. lay out a series of criteria the sample size should meet. These aim to minimise the overfitting and to ensure precise estimation of key parameters in the prediction model.

For continuous outcomes, there are four criteria:
i) small overfitting defined by an expected shrinkage of predictor effects by 10% or less,
ii) small absolute difference of 0.05 in the model's apparent and adjusted R-squared value,
iii) precise estimation of the residual standard deviation, and
iv) precise estimation of the average outcome value.

The sample size calculation requires the user to pre-specify (e.g. based on previous evidence) the anticipated R-squared of the model, and the average outcome value and standard deviation of outcome values in the population of interest.

For binary or survival (time-to-event) outcomes, there are three criteria:
i) small overfitting defined by an expected shrinkage of predictor effects by 10% or less,
ii) small absolute difference of 0.05 in the model's apparent and adjusted Nagelkerke's R-squared value, and
iii) precise estimation (within +/- 0.05) of the average outcome risk in the population for a key timepoint of interest for prediction.

With thanks to Richard D. Riley, Emma C Martin, Gary Collins, Glen Martin & Kym Snell for helpful input & feedback

Value

A list including a matrix of calculated sample size requirements for each criteria defined under *'Details'*, and a series of vectors of parameters used in the calculations as well as the final recommended minimum sample size and number of events required for model development.

Author(s)

Joie Ensor (University of Birmingham, j.ensor@bham.ac.uk),

References

Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical research ed)*. 2020

Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, Collins GS. Minimum sample size required for developing a multivariable prediction model: Part I continuous outcomes. *Statistics in Medicine*. 2018 (in-press). doi: 10.1002/sim.7993

Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, Collins GS. Minimum sample size required for developing a multivariable prediction model: Part II binary and time-to-event outcomes. *Statistics in Medicine*. 2018 (in-press). doi: 10.1002/sim.7992

Riley, RD, Van Calster, B, Collins, GS. A note on estimating the Cox-Snell R2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Statistics in Medicine*. 2020

Examples

```
## Examples based on those included in two papers by Riley et al.
## published in Statistics in Medicine (2018).
## NB: Survival example based on Riley et al. BMJ paper (2020).

## Binary outcomes (Logistic prediction models)
# Use pmsampsize to calculate the minimum sample size required to develop a
# multivariable prediction model for a binary outcome using 24 candidate
# predictor parameters. Based on previous evidence, the outcome prevalence is
# anticipated to be 0.174 (17.4%) and a lower bound (taken from the adjusted
# Cox-Snell R-squared of an existing prediction model) for the new model's
# R-squared value is 0.288

pmsampsize(type = "b", csrsquared = 0.288, parameters = 24, prevalence = 0.174)

# Now lets assume we could not obtain a Cox-Snell R-squared estimate from an existing
# prediction model, but instead had a C-statistic (0.89) reported for the existing prediction
# model. We can use this C-statistic along with the prevalence to approximate the Cox-Snell
# R-squared using the approach of Riley et al. (2020). Use pmsampsize with the cstatistic()
# option instead of rsquared() option.

pmsampsize(type = "b", cstatistic = 0.89, parameters = 24, prevalence = 0.174)

## Survial outcomes (Cox prediction models)
# Use pmsampsize to calculate the minimum sample size required for developing
# a multivariable prediction model with a survival outcome using 30 candidate
# predictors. We know an existing prediction model in the same field has an
# R-squared adjusted of 0.051. Further, in the previous study the mean
# follow-up was 2.07 years, and overall event rate was 0.065. We select a
# timepoint of interest for prediction using the newly developed model of 2
# years
```

```
pmsampsize(type = "s", csrsquared = 0.051, parameters = 30, rate = 0.065,
           timepoint = 2, meanfup = 2.07)

## Continuous outcomes (Linear prediction models)
# Use pmsampsize to calculate the minimum sample size required for developing
# a multivariable prediction model for a continuous outcome (here, FEV1 say),
# using 25 candidate predictors.  We know an existing prediction model in the
# same field has an R-squared adjusted of 0.2, and that FEV1 values in the
# population have a mean of 1.9 and SD of 0.6

pmsampsize(type = "c", rsquared = 0.2, parameters = 25, intercept = 1.9, sd = 0.6)
```

Index