PLOS ONE

CARRoT: R package for predictive modelling by means of regression adjusted for multiple regularisation methods

--Manuscript Draft--

Manuscript Number:	PONE-D-23-16556
Article Type:	Research Article
Full Title:	CARRoT: R package for predictive modelling by means of regression adjusted for multiple regularisation methods
Short Title:	CARRoT: R package for predictive modelling
Corresponding Author:	Alina Bazarova Forschungszentrum Julich Julich Supercomputing Centre Jülich, GERMANY
Keywords:	predictions; events per variable rules; regression; lasso; statistical significance; R-squared; R-package
Abstract:	We present an R-package for predictive modelling CARRoT (Cross-validation, Accuracy, Regression, Rule of Ten). CARRoT is a tool for initial exploratory analysis of the data which performs exhaustive search for a regression model yielding the best predictive power with heuristic ‘rules of thumb’ and expert knowledge as regularization parameters. It uses multiple hold-outs in order to internally validate the model. The package allows to take into account multiple factors such as collinearity of the predictors, event per variable rules (EPVs) and R-squared statistics during the model selection. In addition, other constraints, such as forcing specific terms and restricting complexity of the predictive models can be used. The package allows taking pairwise and three-way interactions between variables into account as well. These candidate models are then ranked by predictive power which is assessed via multiple hold-out procedures and can be parallelised in order to reduce the computational time. Models which exhibited the highest average predictive power over all hold-outs are returned. This is quantified as absolute and relative error in case of continuous outcomes, accuracy and AUROC values in case of categorical outcomes. In this paper we briefly present statistical framework of the package and discuss the complexity of the underlying algorithm. Moreover, using CARRoT and a number of datasets available in R we provide comparison of different model selection techniques: based on EPVs alone, on EPVs and R-squared statistics, on lasso regression and on including only statistically significant predictors.
Order of Authors:	Alina Bazarova
	Marko Raseta
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate.	Alina Bazarova is supported by Helmholtz Association Initiative and Networking Fund

Unfunded studies Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement. View published research articles from *PLOS ONE* for specific examples.	The authors have declared that no competing interests exist.

NO authors have competing interests Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	N/A

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating ‘data available on request from the author’ is not sufficient. If your data are only available upon request, select ‘No’ for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.)*. • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	All the data are publicly available within the packages of statistical software R. The corresponding packages are indicated within the manuscript and therefore the datasets can be retrieved by calling them from R.

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

CARRoT: R package for predictive modelling by means of regression adjusted for multiple regularisation methods

Alina Bazarova[1,2]*, Marko Raseta[3],

1Jülich Supercomputing Center, Forschungszentrum Jülich, Jülich, Germany
2 Helmholtz AI, Germany
3 Department of Molecular Genetics, Erasmus MC, Rotterdam, Netherlands

* Corresponding author al.bazarova@fz-juelich.de

Abstract

We present an R-package for predictive modelling *CARRoT* (*C*ross-validation, *A*ccuracy, *R*egression, *R*ule *o*f *T*en). *CARRoT* is a tool for initial exploratory analysis of the data which performs exhaustive search for a regression model yielding the best predictive power with heuristic 'rules of thumb' and expert knowledge as regularization parameters. It uses multiple hold-outs in order to internally validate the model. The package allows to take into account multiple factors such as collinearity of the predictors, event per variable rules (EPVs) and R-squared statistics during the model selection. In addition, other constraints, such as forcing specific terms and restricting complexity of the predictive models can be used. The package allows taking pairwise and three-way interactions between variables into account as well. These candidate models are then ranked by predictive power which is assessed via multiple hold-out procedures and can be parallelised in order to reduce the computational time. Models which exhibited the highest average predictive power over all hold-outs are returned. This is quantified as absolute and relative error in case of continuous outcomes, accuracy and AUROC values in case of categorical outcomes. In this paper we briefly present statistical framework of the package and discuss the complexity of the underlying algorithm. Moreover, using *CARRoT* and a number of datasets available in R we provide comparison of different model selection techniques: based on EPVs alone, on EPVs and R-squared statistics, on lasso regression and on including only statistically significant predictors.

Introduction

Linear regression remains a first-choice tool for building predictive models in a vast variety of applied problems. Interpretability and simplicity of implementation often make it into a 'gold standard' to compare with when developing more complex models for prediction. In the assessment of medical datasets this approach becomes particularly important as the comprehensive risk scores for patient stratification can be easily derived from the regression coefficients and are therefore accepted and well-understood by practitioners. Moreover, possible non-linear relations between explanatory and dependent variables can also be discovered by including higher order interactions terms between independent variables as new predictors analogously to multidimensional Taylor series approximation.

Often in the studies of a very specific medical condition it is the case that feasible sample size is rather moderate whilst the number of explanatory variables is high [1, 2], therefore including too many of them into the predictive model may lead to overfitting. Events per Variable (EPV) rules are a common tool in Medical Statistics used to avoid the latter, initially introduced as 'one in ten rule' [3] and verified via extensive simulations. Since then albeit substantial criticism both on the number of events ([4], [5], [6], [7]) and on the general concept of the EPV ([8], [9]) the bespoke rule remains widely used and highly cited.

The latter two papers address such problems as handling separated datasets and biased regression coefficients. Furthermore, the authors assess robustness and performance of multiple different models in the context of EPV rules via a number of simulation studies. In conclusion they suggest targeting certain values of out-of-sample mean absolute prediction error and mean squared prediction error to determine the sample size. Of note is that the above two papers do not aim to study prediction problems in small sample sizes. Moreover, although there is a heterogeneity between performance of different models in cases EPV rule is set to low values it seem to significantly decrease as the latter grows. In turn, the impact of EPV rule on the predictive power decreases with the fraction of events across a number of different regression models presented in the manuscripts. This observation is especially important for the studies with low prevalence outcomes.

In [10], [11] authors provide detailed step-by-step instructions of estimating sample size for fitting binary and continuous outcomes. These are mainly based on R^2-statistics assessing goodness of fit and global shrinkage factors aiming to reduce overfitting. However, the authors also point readers to certain examples where low R^2 values do not imply poor predictive performance of the model.

In this paper rather than estimating optimal sample size in order to provide a robust model fit we concentrate on a somewhat inverse problem. Namely, given a dataset of fixed sample size and a certain number of predictive variables determine a model leading to the best predictive performance. The latter along with the model robustness is assessed by means of multiple internal validations essential for model development ([12], [13]).

CARRoT is implemented as a package within statistical software R [14]. It performs exhaustive search over all subsets of explanatory variables, also known as 'best subset regression' ([15]) subject to custom constraints. These are firstly the ones translating available sample size into the properties of feasible models. Namely, user can specify a certain EPV rule (with the default 'one in ten') to be satisfied or alternatively verify whether the criteria described in [10], [11] are met for each model. These two methods can be applied both separately and in combination. In section "Results and discussion" we provide examples run on publicly available datasets illustrating the usage of each one of them.

Secondly, *CARRoT* provides the option of expert knowledge integration thereby introducing further constraints into the process of model selection. In other words, *CARRoT* is adaptive as its prediction algorithm can be tailored to specific situations. This arises in practice when there is an expert consensus on which variables will necessarily be a part of predictive model and the scientific question to be answered is whether adding a number of new independent variables results in increased predictive power. Another common situation occurs during a General Practitioner appointment when a number of measurements have to be taken and decision made on the spot. Indeed, as the study [16] demonstrates, length of the GP appointment does not exceed 5 minutes for more than 50% of world population. Consequently, this practical temporal constraint naturally gives rise to an upper bound on the number of explanatory variables the model can comprise. Moreover, *CARRoT* also allows one to reduce the search space by focusing only on those models where correlation between the explanatory variables is bounded by a specified constant.

To ensure internal validation *CARRoT* performs hold-out procedure a specified number of times by randomly splitting the input dataset into the training and test sets. For each training set the package fits regression models based on all possible subsets of variables subject to constraints, and then assesses their predictive power on the corresponding test set. The average predictive power for each model is then computed over all the test sets and variables constituting the models with the highest average predictive power are then returned.

There are a number of R packages which perform an exhaustive search over all possible subsets of variables and then output the 'best' one, based on a certain type of statistics. However such packages unlike *CARRoT* lack a built-in cross-validation procedure and therefore do not work directly with predictive power. Furthermore, although the option of restricting the number of variables is usually available in the existing packages, it is not equivalent to the EPV constraint when categorical variables are present in the dataset. In such packages R^2 statistics is frequently used to rank the models, however this is not enough to ensure that the model is feasible for the given sample size either. Therefore, to the best of our knowledge, this is the first package that allows one to restrict the model selection based on sample size for regression criteria within a hold-out framework.

Methods

EPV rule

EPV rules were originally introduced as overfitting reduction tool, however, this procedure naturally gives rise to the upper bound on the number of variables to be included in the model. Let $X_1, \dots, X_n$ be the dependent variables and Y be an independent variable. Then, let

$$Y = \sum_{k=1}^{p+q} \beta_{i_k} X_{i_k}, \quad \{i_1 \dots i_{p+q}\} \subset \{1 \dots n\} \tag{1}$$

where $X_{i_1}, \dots, X_{i_p}$ are continuous and $X_{i_{p+1}}, \dots, X_{i_{p+q}}$ are discrete categorical variables each with l_{i_j}, $j \in \{p+1 \dots p+q\}$ numbers of categories. Assume that the desired EPV rule is the rule of r. Then, the model (1) is feasible if the following is satisfied

$$r \left(p + \sum_{j=p+1}^{p+q} (l_{i_j} - 1) \right) \leq N \tag{2}$$

where N is the given sample size. Note that categorical variables with 5 or more numerical categories can be treated as continuous [17]. The default value for r is 10 which correponds to 'one in ten' rule by [3].

Predictive power and hold-out procedures

During each hold-out the dataset of size N is split into the test and training sets of size $\lfloor N\alpha \rfloor$ and $N - \lfloor N\alpha \rfloor$ respectively, where α is a user defined parameter (with 0.9 as the default value). For each such partition we fit all feasible regression models to the training set and assess their predictive power on the corresponding test set.

Let $\mathcal{S}$ stand for the set of all feasible subsets of predictors for a given hold-out, $\{Y_{j_i}\}_{i=1}^{\lfloor N\alpha \rfloor}$ for the values of outcome on the test set and $\{\hat{Y}^S_{j_i}\}_{i=1}^{\lfloor N\alpha \rfloor}$ for their estimates based on the linear regression model corresponding to the subset of predictors $S \in \mathcal{S}$, the corresponding vectors of absolute and relative errors will respectively read

$$\mathbf{E}^S_a = (Y_{j_1} - \hat{Y}^S_{j_1}, \dots, Y_{j_{\lfloor N\alpha \rfloor}} - \hat{Y}^S_{j_{\lfloor N\alpha \rfloor}})^T, \mathbf{E}^S_r = ((Y_{j_1} - \hat{Y}^S_{j_1})/Y_{j_1}, \dots, (Y_{j_{\lfloor N\alpha \rfloor}} - \hat{Y}^S_{j_{\lfloor N\alpha \rfloor}})/Y_{j_{\lfloor N\alpha \rfloor}})^T$$

We define absolute and relative errors for the model corresponding to subset S as $||\mathbf{E}_z^S||_1/\lfloor N\alpha \rfloor$, where $||\cdot||_1$ is an L_1-norm and $z \in \{a, r\}$ respectively.

For continuous variable prediction each linear regression model $S \in \mathcal{S}$ is ranked by relative and absolute errors averaged over the performed hold-outs. Models S_a and S_r minimizing absolute and relative errors respectively are then reported.

For categorical variable predictions the output of the multinomial regression are the odds ratios transformed into probabilities and consequently into categories as follows

$$P(\hat{Y}_i^S = k) = \frac{\exp\left(\hat{P}_{ik}\right)}{1 + \sum_l \exp\left(\hat{P}_{lk}\right)},\ k = \{0, \dots, L-2\}, \tag{3}$$

$$\hat{Y}_i^S = \arg\max_k \{P(\hat{Y}_i^S = k)\} \tag{4}$$

where $\hat{P}_{ij}$ stands for the odds ratio of the jth category against the reference one and L is the number of outcome categories. The reference category is always the one with the highest value or, if the outcome is non-numeric, the least frequent one.

Predictive power of the models for categorical dependent variables is quantified using accuracy. To be more specific, the accuracy vector for a model $S \in \mathcal{S}$ is defined by

$$\mathbf{A}^S = (I\{|\hat{Y}_{j_1}^S - Y_{j_1}| = 0\}, \dots, I\{|\hat{Y}_{j_{\lfloor N\alpha \rfloor}}^S - Y_{j_{\lfloor N\alpha \rfloor}}| = 0\})^T, \tag{5}$$

where I is the indicator function of the event. Average accuracy is defined analogously to the linear case.

Predictive power of the models for binary outcomes can also be quantified by computing AUROC values defined in a standard fashion with its average value derived exactly as before.

Each hold-out procedure yields $|\mathcal{S}|$ values of predictive power corresponding to every feasible subset of predictors. The number $|\mathcal{S}|$ may vary in case of categorical outcomes since two training sets belonging to different partitions do not necessarily contain the same number of least likely outcomes. This means that predictive models evaluated during one procedure may not be evaluated during the other. If the model did not appear in all hold-outs its average predictive power will still be calculated and scaled appropriately. In practice there are situations when, due to a very rare type of dataset partition into training and test set, certain model has high predictive power simply because it was not cross-validated sufficient number of times. This can be resolved by setting an upper bound on the number or weight of the predictors in a model, as it will eliminate those extremely unlikely situations when such models are feasible.

Additional sample size criteria

The authors of [10, 11] propose to estimate sample size required to fit a particular regression model based on multiple quantities. These are global shrinkage factor, difference between apparent and adjusted R^2 statistics, residual standard deviation estimate and mean predicted outcome estimate for the continuous case. For binary case, instead of standard deviation and mean an estimation of overall risk in the population are used. Shrinkage factor and difference between apparent and adjusted R^2 statistics are the two measures to evaluate the problem of overfitting on the relative and absolute scale respectively. Required sample size ensures that the first one is close to 1 whilst the second one is close to 0 (> 0.9 and < 0.05 respectively as the authors recommend). The remaining criteria guarantee an adequate estimation of the basic parameters from the population.

When calculating appropriate sample size for a regression model according to [10, 11] one also needs to take into account the anticipated values of R^2-statistics, mean outcome and the population variance or alternatively an outcome proportion. This typically requires either using the data from previous studies or an educated guess and can therefore be rather demanding. In *CARRoT* the hold-out concept of splitting the data into training and test set naturally allows to decide whether a model is feasible for the target sample size based on the values computed on the training set. Namely, the above quantities are computed on the training set and, in case they satisfy the given constraints, predictive power of the model is assessed on the corresponding test set. Otherwise it is declared infeasible for the current hold-out and the validation step is skipped.

Therefore, in case of linear regression, apparent R^2 of the model on the training set has to satisfy the following condition which follows directly from the formulae for global shrinkage factor and difference between apparent and adjusted R^2 statistic provided in [10, 11]

$$R^2 > \max\left(1 - \exp\left(\frac{2-p}{0.1n_{tr}}\right), \frac{p - 0.05(n_{tr} - 1 - p)}{p}\right) \tag{6}$$

where p is the weight of the model and n_tr is the size of the training set. Similarly, for error margins of the residual variance and mean outcome we use the following formulae

$$M_v = \sqrt{\max\left(\frac{\chi^2_{1-\frac{0.05}{2},n_{tr}-p-1}}{n_{tr}-p-1}, \frac{n_{tr}-p-1}{\chi^2_{1-\frac{0.05}{2},n_{tr}-p-1}}\right)} \tag{7}$$

$$M_o = t_{1-\frac{0.05}{2},n_{tr}-p-1}\frac{\hat{\sigma}}{\sqrt{n}\hat{\alpha}}, \tag{8}$$

where the notation is as in (6), M_v, M_o, $\hat{\sigma}^2$, $\hat{\alpha}$ are margins of error and training set estimates of residual variance and mean outcome respectively. The package allows to specify the desired margin $M_o = M_v$ in (7) to be chosen.

For the binary case we ensure that the following conditions are satisfied

$$S_{VH} = 1 - \frac{p}{-2(\ln L_0 - \ln L_m)} > 0.9 \tag{9}$$

$$\frac{1 - \exp\left(-2(\ln L_0 - \ln L_m)/n_{tr}\right)}{1 - \exp\left(2L_0/n_{tr}\right)}(1 - S_{VH}) < 0.05, \tag{10}$$

where S_{VH} is the Van Houwelingen's shrinkage factor, L_0 and L_m are the log-likelihoods of the model with no predictors and the full model currently being assessed, respectively. The error margin for ouctome proportion in the population can be customized and is computed as follows

$$M_b = 1.96\sqrt{\frac{\hat{\phi}(1-\hat{\phi})}{n}}, \tag{11}$$

where M_b is the error margin and $\hat{\phi}$ is the estimated outcome proportion on the training set. Note that all formaulae (6)-(11) are either taken or derived from the ones in [10, 11].

Other features and computational complexity

In case when all predictors in a given dataset are either continuous or binary the number of all subsets constrained by an EPV rule will be $\sum_{i=1}^{w}\binom{n}{i}$, and therefore if $w > n/2$ the problem of finding all subsets becomes computationally infeasible with the increase in sample size. However, if $w < n/2$ and is fixed, the sum can be bounded by $O(\binom{n}{w})$, and therefore complexity grows polynomially when it is the number of variables rather than the sample size that is increasing. In practice the set of explanatory variables is often a mixture of those with different weights allowing for more computational feasibility. Note that when working with medical datasets, e.g., creating a score, the maximal number of predictors is often bounded by 20 (e.g. due to the time constraint and necessity to take measurements on the spot, [16]) therefore making the problem computationally feasible.

Additional features of the package ease the computational burden of the algorithm. In case there is a 'fixed' part of the predictive model (consensus between medical experts reached, meaning specified variables have to be a part of the model), the number of all feasible subsets will be reduced by $\sum_{i=w-l+1}^{w} \binom{n}{i} - 1$ (where l is the weight of the "fixed" subset). This reduction is not insignificant as it removes the largest terms of the partial sum which takes predictors without this constraint into account. Similarly, on top of specifying minimum and maximum number of predictors, as well as the maximum weight of a model (where weight of a model equals the sum of weights of variables it comprises), *CARRoT* also allows users to eliminate models containing correlated predictors from the selection process by specifying the highest allowed correlation between the independent variables (default is set to 1).

Higher order models

Subject to feasibility restricted by computational complexity (e.g. $w \ll n$), in order to increase the predictive power, one can include second order terms, such as squares and all pairwise interactions between the variables. As previously indicated, the complexity growth is significantly slower with the number of variables when the sample size is fixed, although here the increase is still highly significant growing from $O(n^w)$ to $O(n^{2w})$. Function `quadr` transforms the set of given numeric variables into the one containing quadratic terms in a following way:

$$(\mathbf{a}_1, \dots, \mathbf{a}_k) \rightarrow (\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{a}_{11}, \dots, \mathbf{a}_{1k}, \mathbf{a}_{22}, \dots, \mathbf{a}_{2k}, \dots, \mathbf{a}_{k-1,k-1}, \mathbf{a}_{k-1,k}, \mathbf{a}_{kk}), \quad (12)$$

where $\mathbf{a}_i$ is the ith predictor (m-dimensional vector), $\mathbf{a}_{ij}$ is the variable obtained by coordinatewise multiplication of $\mathbf{a}_i$ and $\mathbf{a}_j$.

Similarly, there is also a function `cub` which includes cubic terms, transforming the variables analogously to (12) and adding the following k terms in the acsending order to the right of the matrix on the right hand side of (12)

$$(\mathbf{a}_{iii}, \dots, \mathbf{a}_{iik}, \dots, \mathbf{a}_{i,k-1,k-1}, \mathbf{a}_{i,k-1,k}, \mathbf{a}_{ikk})),\ i \in \ \{1, \dots, k\}$$

Functions `quadr` and `cub` create additional variables by introducing higher order terms and are specifically designed for numerical continuous and binary variables and is not well-defined for non-numeric variables. Importantly, by allowing the predictive model to take a form of a polynomial, we enable it to unravel non-linear relationship between variables analogous to Taylor series approximation.

Software implementation

The software package implementing the methods above is available under `https://cran.r-project.org/web/packages/CARRoT/index.html` with the main function : `regr_ind` delivering the analysis. The remaining ones mostly play auxiliary roles. For the linear mode specified as `mode='linear'` the function `lsfit()` is used to call linear regression. Multinomial and binary modes are specified as `mode='binary'` and `mode='multin'`, respectively. We note that in both cases the function `multinom()` from the package *nnet* is used to run the corresponding regression.

We use R dataset `swiss` to illustrate the linear mode of `regr_ind`. We attempt to predict `fertility` (first variable in the dataset) which is a continuous variable using the remaining five variables. Sample size equals 47 which means that all models with up to 4 predictors are feasible for EPV= 10. The following line of code calls `regr_ind` with 1000 cross-validations.

```
>set.seed(657)}
>result<-regr_ind(vari=swiss[,2:6],outi=swiss[,1],crv=1000,mode='linear')
```

The partition into training and test set is determined by the parameter `part` with a default value 10 as in the above case, which means that the size of the training set is 1/10 of the size of the whole dataset. The number of hold-outs `crv` is set to a 1000. The first line of the output will be an array of six numbers, first two corresponding to the average absolute and relative errors attained on the test sets by the best performing models respectively. Furthermore, the second pair of numbers refers to the same types of errors reached by this model on the training sets. Comparing the difference between the average errors on the training and the test sets allows to evaluate degree of overfitting while training. Finally, the last two numbers are the average absolute and relative errors produced by the empirical prediction on the test set for each cross-validation. The idea behind this is to determine whether the prediction produced by the best feasible subset of the dependent variables is able to reach a higher predictive power than the most straightforward prediction, i.e., the sample mean.

```
[1] 6.05029812 0.08904263 5.42811143 0.07987293 9.72291238 0.15296233
```

The list object `result` consists of three lists: `[[1]], [[2]] and [[3]]`, which correspond to the array of predictive powers printed out earlier and two lists of models exhibiting the lowest absolute and relative errors respectively, each model defined as a set of indices of the dependent variables constituting it. Note that the variable indices in the output are given with respect to the input dataset, `swiss[,2:6]` in this case, and not the entire dataset.

```
>result
[[1]]
[[1] 6.05029812 0.08904263 5.42811143 0.07987293 9.72291238 0.15296233}
[[2]]
[[2]][[1]]
[1] 1 3 4 5
[[3]]
[[3]][[1]]
[1] 1 3 4 5

```

In this particular case the same model exhibits the lowest average absolute and relative error. In order to display the names of the variables constituting the best model

```
>names(swiss[,2:6])[result[[2]][[1]]]
[1] "Agriculture"    "Education"    "Catholic"    "Infant.Mortality"
```

By varying parameters such as the partition into training and test set (`part`), EPV rule (`rule`) and choosing the option of model assessment based on R^2-statistic and global shrinkage factor (`Rsq=TRUE`) and others one can assess robustness of the model. For example, if on top of the default EPV= 10 we use `Rsq=TRUE`, the difference between the error on the training and test set, as well as the overall performance of the best model is only slightly better than in the default case `Rsq=FALSE`. The selected model is the same as in previous case.

```
>set.seed(657)
result<-regr_ind(swiss[,2:6],swiss[,1],1000,mode=’linear’,part=10,Rsq=T)
[1] 6.04956000 0.08900914 5.43180952 0.07990912 9.72291238 0.15296233

>names(swiss[,2:6])[result[[2]][[1]]]
[1] "Agriculture"    "Education"     "Catholic"      "Infant.Mortality"
```

In the above example we did not set the error margin for mean and variance estimators discussed in Section ”Additional sample size criteria” as the sample size of 47 subjects is rather small and low error margin will simply lead to all modes being infeasible. The output of the model will read

```
>set.seed(657)
result<-regr_ind(swiss[,2:6],swiss[,1],1000,\\
                                 mode=’linear’,part=10,Rsq=T,marg=0.1)
[1] NA NA 9.72291238 0.15296233
```

`NA`s in the output refer to non-existence of any feasible model, performance on the training set is therefore omitted and the last two values as before correspond to predictive power of the empirical prediction. On the other hand if one chooses EPV= 20, the increase in the absolute error is smaller than in case EPV= 10 (8% vs 11%) whilst the absolute error itself increases by 20%. The selected model is now constituted by only two variables due to an EPV restriction.

```
>set.seed(657)
result<-regr_ind(swiss[,2:6],swiss[,1],1000,mode=’linear’,rule=20)
[1] 7.3065447 0.1061592 6.7645138 0.0981193 9.7229124 0.1529623

>names(swiss[,2:6])[result[[2]][[1]]]
[1] "Education" "Catholic"
```

As a binary case example we use dataset `Pima.tr` from the package *MASS* ([18]), also known as ”Diabetes in Pima Indian Women”. We use first seven variables of the dataset in order to predict whether a person has diabetes. Note that there are 68 subjects with diabetes overall.

```
>set.seed(100556)
regr_ind(Pima.tr[,1:7],Pima.tr[,8],crv=100,cutoff=0.5,\\
                                mode=’binary’,objfun=’acc’,part=5,maxw=5)
[1] 0.7635000 0.7841875 0.6602500
[[1]]
[1] 0.7635000 0.7841875 0.6602500
[[2]]
[[2]][[1]]
[1] 2 6 7

>names(Pima.tr[,1:7])[c(2,6,7)]
[1] "glu" "ped" "age"
```

In this case the best model is chosen via accuracy maximisation. This is indicated by `objfun=’acc’`, although could have been skipped since this is the default value for parameter `objfun`. The output consists of two lists. One is a list of accuracies, where the first and the second values correspond to the accuracy of the best model on the test and training set respectively and the third one corresponds to the accuracy of the empirical prediction. The second list is a list of the corresponding best models. Cut-off value for the deterministic prediction is specified via `cutoff=0.5`. The size of the sample is 200 with 68 events which means that if one always predicts the absence of diabetes accuracy of the prediction will be 0.66 (although for different partitions of the set into training and test sets the value may vary slightly, e.g. it reads 0.6603 in the example above). Input variable `part` indicates that the training set is 1/5 (20%) of the whole dataset. Note, that there are 68 adverse outcomes the probability of at least 60 of them occurring in the randomly chosen 160 subjects is around 0.026. This implies that on average out of 100 hold-outs there will be 2-3 for which a model of weight 6 will be feasible and for the remaining hold-outs it will not be assessed. In case such model performs well on those hold-outs it will significantly bias the output since as compared to smaller models it was not cross-validated as many times. We tackle this problem by specifying parameter `maxx=5` thereby restricting the maximum weight of the model to 5. Note that empirical prediction computes occurence of the value 0, and hence if if in a given dataset value 1 arises more frequently than 0 the output will be below 0.5. Similarly, one can change the objective function parameter to `'roc'` in order to identify a model with the higest average AUROC value. Note that in this case empirical prediction is not reported.

```
>set.seed(100556)
regr_ind(Pima.tr[,1:7],Pima.tr[,8],100,cutoff=0.5,mode='binary',\\
                                        objfun='roc',part=5,maxw=5)
[1] 0.8223213 0.8504157
[[1]]
[1] 0.8223213 0.8504157
[[2]]
[[2]][[1]]
[1] 2 5 6 7

>names(Pima.tr[,1:7])[c(2,5,6,7)]
[1] "glu" "bmi" "ped" "age"}
```

To illustrate multinomial mode of the package we use the dataset `bladder1` from the package *survival* ([19]). We predict the `status` variable given the previous six variables, merging into one category events of type 2 and 3. The notation is as in the binary case

```
>bladder1[which(bladder1[,8]==3),8]=2
>set.seed(901)
>regr_ind(bladder1[,2:7],bladder1[,8],100,mode='multin')
[1] 0.8226667 0.8290152 0.6406667
[[1]]
[1] 0.8226667 0.8290152 0.6406667
[[2]]
[[2]][[1]]
[1] 4 5}
```

Hold-out procedures within `regr_ind` can also be parallelised by means of parameter `parallel=TRUE`, which is set to `FALSE` by default and specifying the number of `cores`. This is accomplished via the function `foreach` from the package *doParallel* and significantly enhances the computations as in the following example where "`...`" stands for the same values of parameters as in the case of no parallelisation.

```
>set.seed(100556)
>system.time(regr_ind(Pima.tr[,1:7],Pima.tr[,8],crv=100,\\
                             mode='binary',objfun='roc',part=5,maxw=5)))
```

```
[1] 0.8223213 0.8504157
   user  system elapsed
48.730   0.224  49.002

>system.time(regr_ind(vari=Pima.tr[,1:7],outi=Pima.tr[,8],\\
                                    ... ,parallel=TRUE,cores=7))
[1] 0.8374064 0.8465505
   user  system elapsed
0.145   0.068  13.812
```

To take into account higher order interactions between variables functions `quadr` and `cub` can be used. An example is the dataset `hodg` from the package *KMsurv* ([20]), where the time to relapse of the disease or death is the dependent variable given the graft type, disease type, Karnofsky score and waiting time to transplantation.

```
>set.seed(12379)
>regr_ind(hodg[,c(1:2,5:6)],hodg[,3],crv=1000,mode=’linear’)
[1] 379.211400   6.183534 341.180737   5.655696 445.683189  11.160090
[[1]]
[1] 379.211400   6.183534 341.180737   5.655696 445.683189  11.160090
[[2]]
[[2]][[1]]
[1] 1 2 3
[[3]]
[3][[1]]
[1] 3}

>set.seed(12379)
>regr_ind(quadr(hodg[,c(1:2,5:6)]),hodg[,3],crv=1000,mode=’linear’))
[1] 354.332000   3.411531 318.508921   3.158049 445.683189  11.160090
[[1]]
[1] 354.332000   3.411531 318.508921   3.158049 445.683189  11.160090
[[2]]
[[2]][[1]]
[1]  1  7 12
[[2]][[2]]
[1]  5  7 12
[[3]]
[3][[1]]
[1] 10 12 14}
```

Note that in this case introduction of quadratic terms in the above case reduces the absolute error by around 7% and the relative one by more than 55%. One of the quadratic models for the best absolute error also contains linear term graft type (index 1) which was previously a part of the best linear model. The function `find_int` can be used to ’decipher’ the index of the coefficient into the interaction between the original variables by using the size of the initial dataset:

```
>find_int(ind=7,N=4)
[1] 1 3
```

Therefore, index 7 given 4 variables is the interaction between the first one and the third one, which correspond to graft type and score in this particular case.

Results and discussion

Predictive power

In this section we provide examples of usage of *CARRoT* on a number publicly available datasets and compare its performance to other widely used variable selection techniques within the presented framework. In addition we discuss the application of *CARRoT* in several published studies.

Firstly, we assess over a 100 datasets appropriate for predictive modelling available as part of different R-packages. These include 88, 32 and 8 instances used for continuous, binary and multinomial predictions, respectively. We choose the default *CARRoT* mode with EPV= 10 and $90\% - 10\%$ hold-out split for these datasets and report the predictive power quantified as absolute/relative error, accuracy/AUROC and accuracy only for each one of the respective package modes in table S1. The number of hold-outs is set to 1000 and the cut-off value in the binary mode equals 0.5 while neither R^2-statistic nor global shrinkage factor are not taken into account.

Furthermore, we amend the above predictive models based solely on EPV rule with constraints on the R^2-statistic and global shrinkage factor by setting `Rsq` parameter to `TRUE` and investigate the subsequent change in the predictive power. Note that this method applies only to datasets where continuous and binary modes have been assessed.

Finally, using the same framework of multiple hold-outs we evaluate two types of models widely used in practice, namely variable selection based on statistically significant univariate regressions and lasso regression.

For the first method we build our comparison as follows: for each hold-out we identify variables which are statistically significant (at 0.05 level) on the training set based on the results of univariate regressions and compute the predictive power of the model which contains all such variables on the test set. Note that, for different partitions of data into training and test set, different variables may exhibit significance especially in the presence of data heterogeneity. No particular predictors are fixed in this case and therefore instead of computing the average predictive power of a particular model we obtain the average predictive power achievable with this approach. In reality this predictive power can be reached if and only if there is a certain set of predictors which are consistently statistically significant independently of the nature of the hold-out split.

Similar holds for the case of lasso regression. We use hold-out splits in order to choose the penalty parameter λ which maximizes the predictive power. The set of penalty parameter values is therefore fixed across all hold-outs and the highest predictive power corresponds to the most optimal penalty parameter rather than to a set of predictors. Hence, as in the previous case of significant variables selection, we evaluate the approach as a whole since a given parameter λ does not guarantee the same set of non-zero regression coefficients for each split of the data into training and test set.

For the first method the p-value of F-statistic was computed by extracting the corresponding value from `lm` for linear and `Anova` from package *car* ([21]) for multinomial regression. For the latter one we used `glmnet` from the corresponding package ([22]). We make sure to use same exact hold-out splits for all types of models.

Note that in 76% of the cases predictions provided by *CARRoT* are strictly better than the ones obtained by the models built on statistically significant explanatory variables as presented in the Table 1. On the other hand lasso-based model exhibits strictly lower predictive power 50% of the time and in case the *CARRoT* model is constrained by R^2-statistics the corresponding value goes down as low as 22% (see Tables 2, 3). The remaining cases of the alternative models yielding higher or equal results largely fall into the following categories:

(i) Predictive power of *CARRoT* output is equal to the predictive power of the other model.

Usually this means that the output of *CARRoT* includes the same set of predictors selected by another model (and possibly other sets of predictors with the same exact predictive power). In particular, when comparing the model with additional R^2 constraints to the simple EPV=10 the latter result implies that the model satisfying EPV=10 constraint and exhibiting the highest predictive power also satisfies the R^2 constraint. Note that this occurs 67% of the time.

(ii) Predictive power of *CARRoT* output is lower than the predictive power of another model since the sample size is too small for *CARRoT* to fit the latter model.

This applies to lasso-based and significant predictors based models which are therefore likely to be overfitted.

Examples of such datasets are `immer`, `oats`, `mpg` see Table S1. In each one of these datasets the sample size is low compared to the number of categories of the predictors and the lasso-based and significance-based models violate the 'rule of ten'.

(iii) Predictive power of *CARRoT* output is lower than the predictive power of another model due to the fact that the set of significant predictors changes depending on the partition into the training and test set.

The latter is again applicable to lasso-based and significance-based models and indicates that the given dataset is rather heterogeneous and that the model built on the basis of lasso regularisation or significant predictors is highly unreliable and hence unsuitable for medical applications in particular.

Examples of such datasets are `urine`, `coop` (see Table S1) where the set of significant predictors or shrinked to 0 coefficients corresponding to the same penalty parameter λ changes depending on the partition into the training and test set).

Similar situation arises when the *CARRoT* based model with additional R^2 constraint exhibits higher predictive power than the model based on EPV alone: this indicates that, due to the R^2 constraint, the best model was not feasible for a number of hold-out splits. This further indicates the presence of heterogeneity in the given dataset, which one can observe e.g. in `uis`

(iv) Coefficient estimators of the lasso model, known to be biased towards zero lead to a higher predictive predictive power, even though the selected variables are the same as the ones picked by *CARRoT*, e.g. in `environmental`. In such cases other factors, such as overfitting of the model need to be assessed in order to conclude which coefficients are more reliable.

Statistics on the results from Table S1 is summarised in Tables 1 - 4. Observe that, on average, absolute error exhibited by the models selected by *CARRoT* is 1.38 times smaller than the one obtained when using the model constituted of all significant predictors whereas the corresponding relative error is on average 1.66 times smaller. Accuracy and AUROC values produced by *CARRoT* are higher than the ones obtained by the significant predictors model. Indeed, the ratios read 1.03 and 1.02 for binary accuracy and binary AUROC modes, respectively. Lasso-based model performs mostly no better than *CARRoT* and analogous values from the Table 4 are mainly above 1. Note that all the outputs were rounded to the second decimal place prior to performance comparison.

Table 1. Comparison of predictive power. Sigificance-based models. The table reports percentage of time one model yields better/worse predictive power than the other one.

		B better than S	B equals S	B worse than S	E better than S	E better than B
Linear	Absolute	85%	12%	3%	20%	3%
	Relative	71%	25%	4%	21%	0%
Binary	Accuracy	83%	13%	3%	16%	0%
	AUC	71%	23%	6%	-	-
Multin	Accuracy	44%	33%	22%	0%	0%
Overall		76%	19%	5%	19%	1%

'S' stands for the model constituted by all significant predictors, 'B' stands for the one provided by *CARRoT*, 'E' stands for the empirical model (where applicable). For example, in 85 % of the cases the *CARRoT* exhibits lower absolute error than the model based on statistically significant explanatory variables. All other values should be interpreted in a similar fashion.

Table 2. Comparison of predictive power. Lasso-based models.

		B better than L	B equals L	B worse than L	E better than L	E better than B
Linear	Absolute	57%	10%	33%	8%	3%
	Relative	51%	28%	21%	1%	0%
Binary	Accuracy	48%	35%	16%	3%	0%
	AUC	35%	42%	23%	-	-
Multin	Accuracy	44%	11%	44%	0%	0%
Overall		50%	24%	26%	4%	1%

Notation as in Table 1. 'L' stands for the lasso-based regression model.

Table 3. Comparisons of predictive power. R^2 based models.

		B better than R	B equals R	B worse than R	E better than R	E better than R
Linear	Absolute	30%	62%	8%	3%	1%
	Relative	20%	65%	15%	1%	0%
Binary	Accuracy	16%	78%	6%	0%	0%
	AUC	16%	78%	6%	-	-
Overall		22%	67%	10%	2%	1%

Notation as in Table 1. 'R' stands for the model with EPV=10 and additional constraints based on R^2-statistics.

We observe that in around 18% and 8% of the cases the absolute error of the significance-based and lasso-based models is higher than the one provided by the sample mean, respectively. In contrast, in all cases considered, the *CARRoT* output both with and without additional R^2-constraints, produces lower relative error than the empirical one whilst it yielded lower absolute error in at least 97 % of the cases. We note that that for a fixed model of significant predictors the numbers above might turn out to be even less favorable.

Overfitting analysis

Frequently observed low difference in the predictive power between models described above suggests a need for a closer analysis of overfitting. We choose a representative sample of 44 datasets and assess each model. For this purpose in addition to average predictive power on the test sets we computed average predictive power on the respective training sets. We report the ratio between predictive power on the test set and the training set with higher values of the latter indicating presence of overfitting.

Table 4. Comparisons of predictive power.

Linear					
S/B abs	S/B rel	L/B abs	L/B rel	R/B abs	R/B rel
1.38	1.66	1.04	1.09	1.66	1.06
Binary					
B/S acc	B/S AUC	B/L acc	B/L AUC	B/R acc	B/R AUC
1.03	1.02	1.01	1.0	1.02	1.02
Multinomial					
B/S acc		B/L acc		B/R acc	
1.0		0.96		1.74	

'S' stands for the model constituted by all significant predictors,'L' stands for lasso-based method, 'R' stands for EPV=10 with additional R^2 constraints, 'B' stands for the one provided by *CARRoT* with EPV=10, e.g., the value 'S/B absolute' of 1.28 means that, on average, the model constituted by all significant predictors leads to a 1.28 times higher absolute error than the one provided by *CARRoT*. In case of categorical variables 'B/S accuracy' value of 1.05 means that output of *CARRoT* yields 1.05 higher accuracy than the other model. All other values should be interpreted in a similar fashion.

Results of the overfitting analysis (Table S2)are summarised in the Tables 5-7. We observe that the highest average overfitting is exhibited by the lasso-based model, reaching 10% whilst the lowest value is attained by the *CARRoT* with additional R^2 constraint, although the latter outperforms the plain EPV-based only *CARRoT* by only a percent. Although in Table 7 the predictive power of EPV based only on *CARRoT* is not always strictly higher than the one of the lasso-based model the latter ones exhibit more overfitting, especially in case of binary outcomes where lasso-based models maximising AUC overfit more than *CARRoT* based ones in all cases (see Table 7). On the other hand, additional R^2 constraint significantly reduces the overfitting where positive values in the corresponding column of the upper part of Table 7 mainly account for the cases where the overfitting is equal between two models. We also assess the cases when EPV based only *CARRoT* yields both lower predictive power and higher degree of overfitting in the lower part of the Table 7. We report that EPV-based model performs worse than its significance-based and lasso-based counterparts on average only 1% and 7% of the time, respectively. The highest value of 16% corresponds to EPV-based model with R^2 constraint which reflects mainly its superiority with respect to overfitting.

Table 5. Overfitting comparison. The table reports the average absolute difference between overfitting defined in "Results and discussion" for different methods.

		Best	Significant	Lasso	R-squared
Linear	Absolute	6%	10%	12%	5%
	Relative	6%	11%	13%	5%
Binary	Accuracy	2%	3%	4%	2%
	AUC	2%	4%	4%	2
Multin	Accuracy	1%	7%	3%	-
Overall		5%	9%	10%	4%

We note that in more than 60% of the cases the EPV-based model output by *CARRoT* contains variables consituting a subset of those chosen by the lasso-based model. This together with the increased presence of overfitting in the latter suggests that in the presence of large numbers of variables when best subset selection constrained

Table 6. Combined overfitting and predictive power comparison. The table summarises predictive power and overfitting.

Overfitting		B less than S	B less than L	B less than R
Linear	Absolute	59%	63%	15%
	Relative	43%	65%	13%
Binary	Accuracy	57%	71%	0%
	AUC	47%	100%	29%
Multin	Accuracy	50%	63%	-
Overall		52%	68%	14%
Predictive power		B better than S	B better than L	B better than R
Linear	Absolute	70%	33%	30%
	Relative	65%	37%	37%
Binary	Accuracy	86%	57%	14%
	AUC	71%	57%	57%
Multin	Accuracy	50%	38%	-
Overall		67%	38%	35%

Predictive power and overfitting		B worse than S	B worse than L	B worse than R
Linear	Absolute	0%	7%	11%
	Relative	0%	11%	19%
Binary	Accuracy	0%	0%	43%
	AUC	14%	0%	0%
Multin	Accuracy	0%	13%	-
Overall		1%	7%	16%

Models notation as in Table 4. Upper table refers to percentage of time first model exhibits lower overfitting than the second one. Middle table refers to percentage of time the first model exhibits higher predictive power than the second one. Lower table refers to the first model exhibiting both lower predictive power and higher overfitting than the second one.

by EPV is computationally infeasible a hierarchical procedure of model selection can be used. Namely, lasso as the first step of variable selection and subsequent best subset regression which returns unbiased regression coefficients estimators.

The model corresponding to `Rsq=TRUE` exhibits less overfitting, however majority of the time outputs exactly the same result as the EPV= 10 model suggesting that the most of the models satisfying the latter constraint and exhibiting high predictive power also satisfy the necessary conditions on the R^2-statistics and global shrinkage factor. Moreover, the degree of overfitting on average is rather close for the two models as evidenced by the Table 5.

In order to assess how predictive power changes with the increase of EPV rules we ran *CARRoT* with EPV values of $7, 10, 20, 30$ and 40 predictors. In addition, we investigate the decrease in overfitting as EPV grows. The results are summarised in Tables 8 - 9. One can observe that for high values of EPV such as 30 and 40 the overfitting is as low as 3% and 4%, respectively. At the same time the average ratio of absolute error between EPV=30 and EPV=10, EPV=40 and EPV=10 is 1.09 and 1.4, respectively. On the other hand, the average values of overfitting and predictive power ratios for EPV=20 are comparable to those ones provided by EPV=10 with additional R^2 constraints suggesting that the latter to an extent may be replaced by EPV=20 being more user-friendly given its simplicity.

We provide several examples when including second-order interactions between independent variables significantly boosts the predictive power of the model and present our findings in the Table 10. Note that we used same exact partitions of the datasets into training and test ones for linear and quadratic models in order to see whether

Table 7. Combined overfitting and predictive power comparison. The table summarises predictive power and overfitting. Models notation as in Table 4. Upper table refers to percentage of time first model exhibits lower overfitting than the second one. Middle table refers to percentage of time the first model exhibits higher predictive power than the second one. Lower table refers to the first model exhibiting both lower predictive power and higher overfitting than the second one.

Overfitting		B less than S	B less than L	B less than R
Linear	Absolute	59%	63%	15%
	Relative	43%	65%	13%
Binary	Accuracy	57%	71%	0%
	AUC	47%	100%	29%
Multin	Accuracy	50%	63%	-
Overall		52%	68%	14%
Predictive power		B better than S	B better than L	B better than R
Linear	Absolute	70%	33%	30%
	Relative	65%	37%	37%
Binary	Accuracy	86%	57%	14%
	AUC	71%	57%	57%
Multin	Accuracy	50%	38%	-
Overall		67%	38%	35%

Predictive power and overfitting		B worse than S	B worse than L	B worse than R
Linear	Absolute	0%	7%	11%
	Relative	0%	11%	19%
Binary	Accuracy	0%	0%	43%
	AUC	14%	0%	0%
Multin	Accuracy	0%	13%	-
Overall		1%	7%	16%

Table 8. Overfitting comparison for different EPVs. The table reports the average absolute difference between overfitting defined in Results and discussion for different EPV rules.

EPV		7	10	20	30	40
Linear	Absolute	7%	6%	5%	3%	2%
	Relative	6%	6%	4%	3%	2%
Binary	Accuracy	3%	2%	1%	2%	1%
	AUC	3%	2%	1%	0.5%	1%
Multin	Accuracy	2%	1%	1%	2%	2%
Overall		5%	5%	4%	3%	2%

inclusion of quadratic terms yields higher performance characteristics. Indeed, in all cases considered, the predictive power increases by at least 5%, whilst there are cases where absolute error shrinks almost twice. Moreover, we report the degree of overfitting for both linear and quadratic cases. We find that in more than 50% of the datasets the latter did not grow despite the increased complexity of the predictive model and significant improvement in predictive power. The latter suggests that the presence of non-linearity in the relationship between the outcome and explanatory variables.

CARRoT was also used for building predictive models in a number of recently published studies. The Birmingham score for assessing patients with lower gastrointestinal bleeding was created based on the predictive model delivered by *CARRoT*, [23]. On top of outperforming the existing scores it is also simpler and easier

Table 9. Predictive power for different EPVs and R^2. The average ratios between predictive power of models with different EPVs as in Table 4.

EPV ratio		7/10	20/10	30/10	40/10	$R^2/10$
Linear	Absolute	0.99	1.04	1.09	1.4	1.06
	Relative	0.99	1.03	1.05	1.08	1.05
EPV ratio		10/7	10/20	10/20	10/20	$10/R^2$
Binary	Accuracy	0.99	1.01	1.03	1.04	1.0
	AUC	0.98	1.01	1.04	1.07	1.0
Multin	Accuracy	1.01	1.02	1.01	1.02	-
Overall		0.99	1.03	1.06	1.21	1.04

E.g. 7/10 means average ratio between predictive power provided by EPV=7 and EPV=10. The last column correponds to ratio of predictive power of EPV=10 model with additional R^2-statistics constraint and EPV=10 model only.

Table 10. Best performance based on linear and quadratic model.

Dataset	M	Out	Linear	OF	Quadratic	OF	Empirical	N
`kidrecurr` (`KMsurv`)	`l`	`time2`	92.71/3.27	0.04/0.02	85.98/1.64	0.05/0.04	100.57/3.74	5
`hodg` (`KMsurv`)	`l`	`time`	381.54/6.42	0.10/0.08	354.33/3.41	0.10/0.07	445.68/11.16	4
`myeloid` (`survival`)	`l`	`rltime`	137.75/0.59	0.03/0.01	128.70/0.53	0.01/0.02	193.89/0.91	3
`snails` (`MASS`)	`l`	`Temp`	1.54/NA	0.06/NA	1.23/NA	0.09/NA	2.99/NA	3
`environmental` (`lattice`)	`l`	`ozone`	16.74/0.70	0.03/0.03	14.02/0.47	0.06/0.05	27.06/1.59	3
`Phenobarb` (`nlme`)	`l`	`dose`	4.77/0.73	0.01/< 0.01	3.92/0.62	0.01/0.01	5.16/1.0	3(4)
`Wheat` (`nlme`)	`l`	`DryMatter`	2.06/0.29	0.07/0.04	1.10/0.16	0.10/0.07	3.22/0.50	3
`Wheat2` (`nlme`)	`l`	`yield`	4.54/0.41	0.01/0.05	4.07/0.33	0.03/0.01	5.42/0.57	4
`rock` (`datasets`)	`l`	`perm`	183.97/2.15	0.12/0.13	168.51/1.32	0.14/0.13	407.50/9.24	3
`channing` (`boot`)	`b`	`cens`	0.64/0.67	0.01/0.02	0.71/0.79	0.01/0.01	0.62/−	3(4)
`nodal` (`boot`)	`b`	`r`	0.74/0.69	0.01/0.01	0.78/0.73	0.01/0.01	0.62/−	5
`bladder` (`survival`)	`b`	`event`	0.72/0.77	0.02/0.01	0.77/0.81	< 0.01/0.02	0.67/−	4

Column 'Dataset' corresponds to the R dataset analyzed while the package it originates from is provided in the brackets. Column 'M' corresponds to the value of `mode` parameter of `regr_ind` or `regr_whole`. 'Out' is the name of the dependent variable from the corresponding dataset. Columns 'Linear', 'Quadratic' and 'Empirical' correspond to the values of predictive power attained by the best linear, best quadratic and empirical (were applicable) models, respectively. For the mode `linear`, the value in these columns stands for the absolute error/relative error, for the mode `binary` it corresponds to accuracy/AUROC while it simply represents accuracy for `multin` mode. Column 'Num' stands for the overall number of numeric predictive variables in the linear model while the number of all variables, including non-numerical ones (if any), is provided in the brackets. Columns 'OF' contain the overfitting measure for the respective model.

to implement in practice. *CARRoT*-based models exhibited high predictive power when assessing the binary and multinomial patient outcomes following mechanical thrombectomy in stroke patients [24]. In patients with a functional stroke condition *CARRoT* predicted reduction in the MRS and NIHSS scores following hypnotherapy sessions [25]. In [26] *CARRoT* was used to predict the cancer cell type based on the available explanatory variables and their second order interactions and yielded the accuracy of 92% with a quadratic model improving the linear one by 12%. In the metabolomics project the package exhibited up to 0.91 AUROC when discriminating the types of renal tumours [27].

We intentionally made *CARRoT* compute and output a limited number of values and at times deliberately avoided usage of already existing R tools with the idea of saving computational time. For instance, combination of generic R functions `predict()` and `glm()` performed significantly slower than *CARRoT* function `get_predictions()`. On the other hand, due to a relatively simple structure of the main functions, it is straightforward to build in new features and create additional objective functions if needed.

Moreover, if the set of independent variables is too large for performing exhaustive search, the concept of maximising predictive power over the test sets provided by *CARRoT* can be adapted to other types of variable selection, such as stepwise regression or backward elimination. All one needs to do is to restrict the number of variables in the model from below or above, fix certain variable in the model and then run *CARRoT* multiple times until the increment (decrease) in the predictive power is small (large) enough. This type of procedure was implemented in [28] where a modified version of the package was used in order to predict locations of the DNA replication origins and to take specific features of the dataset with no explicit negative instances into account.

There exists a number of R packages enabling exhaustive search for the best regression model such as `glmulti` [29], `bestglm` [30], `BeSS` [31], `meifly` [32], `lmSubsets` [33], `subselect` [34], `kofnGA` and [35] to name just a few. Some of these packages have features intersecting with those of *CARRoT*, e.g. the feature restricting the number of variables in a model is in some cases equivalent to EPV rules. However in *CARRoT* the feature of computing maximum number of variables based on the sample size given a certain EPV rule is automatically built-in, and moreover, in cases when the set of explanatory variables is a mixture of categorical and continuous ones the concept of restricting the number of variables from above is not necessarily equivalent to applying an EPV rule. Another important difference is that *CARRoT* has a built-in cross-validation procedure. Existing R packages for best subset regression mainly utilise Information Criteria or R-squared statistics for model ranking, whilst *CARRoT* works directly with the predictive power of each model as computed over the test sets. Large number of internal cross-validations ensures the approach provides robust results. Package *CARRoT* utilises linear regression models with Gaussian noise and multinomial logistic regressions. Note that the latter ones are not available as part of `glm` function, used in `glmulti` and `bestglm`, for dependent variables with more than two categories. In other words, *CARRoT* is a simple stand alone tool simultaneously combining the model selection, validation and controlling for overfitting.

Conclusions

We developed an R-package for predictive modelling *CARRoT* by integrating sample size calculation rules, expert knowledge, cross-validation, best subset regression in one simple stand-alone tool. The package provides predictions for both continuous and categorical dependent variables.

We compared *CARRoT*'s performance to other widely used interpretable methods of model selection such as Lasso regression, regression based on significant predictors as well as different methods within *CARRoT* itself both in terms of exhibited predictive power and degree of overfitting. Best subset regression restricted by EPV=10 shown very good results both in terms of predictive power and overfitting. With an additional constraint on the R^2-statistic overfitting decreases even further whilst predictive power does not encounter a significant drop. *CARRoT* drastically outperforms the models based on significant predictors alone (over 75% of the time) and yields strictly higher predictive power than lasso regression 50% of the time. Moreover, *CARRoT* yields twice lower overfitting than the latter. We therefore conclude that predictive models produced by *CARRoT* are both robust and reliable. We believe that inclusion of expert knowledge (if available for particular dataset of interest) will further emphasise the difference in predictive power between the *CARRoT* output and the traditional models as will including pairwise and/or three-way interactions between variables. The latter is a proxy for multidimensional Taylor series expansion and therefore *CARRoT* has strong potential to identify the non-linear relations (if present) between independent and dependent variables.

CARRoT allows users to select the best regression model based on a number of objective functions computed over the test sets depending on the outcome type. Other types of objective functions can be easily implemented in the future if necessary.

This package is a user-friendly universal tool designed to be able to deal with different data types, and so far it proved to be useful in Medical Statistics [26], [23], [27], [25]. It can also be used as a straightforward first step of data exploration before moving to more complex analysis techniques if necessary.

Supporting information

S1 Table. Detailed breakdown of performance comparison of CARRoT output and the other models. Performance in terms of absolute/relative error, accuracy/AUROC and accuracy only (for continuous, binary and multinomial outcomes respectively) of different prediction methods on 100 datasets available in R using the default 90%/10% training/validation split. The methods used are *CARRoT* with EPV=10, model, based on significant predictors only, lasso-based model, *CARRoT* with EPV=10 and additional R^2 constraint.

S2 Table. Detailed breakdown of overfitting comparison of *CARRoT* output and the other models. Overfitting in terms of absolute/relative error, accuracy/AUROC and accuracy only (for continuous, binary and multinomial outcomes respectively) computed both on training and test sets of different prediction methods on 43 datasets available in R using the default 90%/10% training/validation split. The methods used are *CARRoT* with EPV=10, model, based on significant predictors only, lasso-based model, *CARRoT* with EPV=10 and additional R^2 constraint.

Acknowledgments

This work is supported by the Helmholtz Association Initiative and Networking Fund

References

1. Collins J, Brown J, Schammel C, Hutson K, Edenfield WJ. Meaningful Analysis of Small Data Sets: A Clinician's Guide. Proceedings of Greenville Health System. 2017;2(1):16–19.

2. Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning–Whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging. 2017;30(4):392–399. doi:10.1007/s10278-017-9976-3.

3. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology. 1996;49(12):1373–1379. doi:10.1016/S0895-4356(96)00236-3.

4. Vittinghoff E, McCulloch CE. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. American Journal of Epidemiology. 2007;165(6):710–718. doi:10.1093/aje/kwk052.

5. Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000;19(8):1059–1079. doi:https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8¡1059::AID-SIM412¿3.0.CO;2-0.

6. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical Methods in Medical Research. 2017;26(2):796–808. doi:10.1177/0962280214558972.

7. Heinze G, Dunkler D. Five myths about variable selection. Transplant International. 2017;30(1):6–10. doi:10.1111/tri.12895.

8. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology. 2016;16(1):163. doi:10.1186/s12874-016-0267-3.

9. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research. 2019;28(8):2455–2474. doi:10.1177/0962280218784726.

10. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. Statistics in Medicine. 2019;38(7):1262–1275. doi:https://doi.org/10.1002/sim.7993.

11. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019;38(7):1276–1296. doi:https://doi.org/10.1002/sim.7992.

12. Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, et al. The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research. Int J Obes (Lond). 2016;40(6):887–894. doi:10.1038/ijo.2015.214.

13. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925–1931. doi:10.1093/eurheartj/ehu207.

14. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.

15. Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med. 2016;4(7):136.

16. Iacobucci G. GP appointments last less than five minutes for half the world's population. BMJ. 2017;359. doi:10.1136/bmj.j5172.

17. Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous?: A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods. 2012;17(3):354–373. doi:10.1037/a0029315.

18. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

19. Terry M Therneau, Patricia M Grambsch. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

20. by Klein O, Moeschberger, modifications by Jun Yan. KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis; 2012. Available from: `https://CRAN.R-project.org/package=KMsurv`.

21. Fox J, Weisberg S. An R Companion to Applied Regression. 2nd ed. Thousand Oaks CA: Sage; 2011. Available from: `http://socserv.socsci.mcmaster.ca/jfox/BOOKs/Companion`.

22. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software. 2011;39(5):1–13.

23. Smith SCL, Bazarova A, Ejenavi E, Qurashi M, Shivaji UN, Harvey PR, et al. A multicentre development and validation study of a novel lower gastrointestinal bleeding score—The Birmingham Score. International Journal of Colorectal Disease. 2020;35(2):285–293. doi:10.1007/s00384-019-03459-z.

24. Raseta M, Bazarova A, Wright H, Parrott A, Nayak S. A novel tool for the prediction of clinical outcomes following mechanical thrombectomy. Manuscript under review;.

25. Sanyal R, Raseta M, Natarajan I, Roffe C. The use of hypnotherapy as treatment for functional stroke: A case series from a single center in the UK. International Journal of Stroke. 2022;17(1):59–66. doi:10.1177/1747493021995590.

26. Rutter AV, Crees J, Wright H, Raseta M, van Pittius DG, Roach P, et al. Identification of a Glass Substrate to Study Cells Using Fourier Transform Infrared Spectroscopy: Are We Closer to Spectral Pathology? Appl Spectrosc. 2020;74(2):178–186. doi:10.1364/AS.74.000178.

27. Papathomas T, Tzortzakakis A, Sun N, Erlmeier F, Feuchtinger A, Trpkov K, et al. In Situ Metabolomics Expands the Spectrum of Renal Tumours Positive on 99mTc-sestamibi Single Photon Emission Computed Tomography/Computed Tomography Examination. European Urology Open Science. 2020;22:88–96. doi:https://doi.org/10.1016/j.euros.2020.11.001.

28. Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, et al. A predictable conserved DNA base composition signature defines human core DNA replication origins. Nature Communications. 2020;11(1):4826. doi:10.1038/s41467-020-18527-0.

29. Calcagno V. glmulti: Model selection and multimodel inference made easy; 2013. Available from: `https://CRAN.R-project.org/package=glmulti`.

30. McLeod AI, Xu C. bestglm: Best Subset GLM and Regression Utilities; 2018. Available from: `https://CRAN.R-project.org/package=bestglm`.

31. Wen C, Zhang A, Quan S, Wang X. BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. 2020;94(4):1–24. doi:10.18637/jss.v094.i04.

32. Wickham H. meifly: Interactive model exploration using GGobi; 2014. Available from: `https://CRAN.R-project.org/package=meifly`.

33. Hofmann M, Gatu C, Kontoghiorghes EJ, Colubi A, Zeileis A. lmSubsets: Exact Variable-Subset Selection in Linear Regression for R. Journal of Statistical Software. 2020;93(3):1–21. doi:10.18637/jss.v093.i03.

34. Orestes Cerdeira J, Duarte Silva P, Cadima J, Minhoto M. subselect: Selecting Variable Subsets; 2022. Available from: `https://CRAN.R-project.org/package=subselect`.

35. Wolters MA. A Genetic Algorithm for Selection of Fixed-Size Subsets with Application to Design Problems. Journal of Statistical Software, Code Snippets. 2015;68(1):1–18. doi:10.18637/jss.v068.c01.

Supporting Information