

RESEARCH ARTICLE

A non-linear data mining parameter selection algorithm for continuous variables

Peyman Tavallali^{1*}, Marianne Razavi¹, Sean Brady²

1 Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, California, United States of America, **2** Principium Consulting, LLC, Pasadena, California, United States of America

* ptavalla@caltech.edu, tavallali@gmail.com



Abstract

In this article, we propose a new data mining algorithm, by which one can both capture the non-linearity in data and also find the best subset model. To produce an enhanced subset of the original variables, a preferred selection method should have the potential of adding a supplementary level of regression analysis that would capture complex relationships in the data via mathematical transformation of the predictors and exploration of synergistic effects of combined variables. The method that we present here has the potential to produce an optimal subset of variables, rendering the overall process of model selection more efficient. This algorithm introduces interpretable parameters by transforming the original inputs and also a faithful fit to the data. The core objective of this paper is to introduce a new estimation technique for the classical least square regression framework. This new automatic variable transformation and model selection method could offer an optimal and stable model that minimizes the mean square error and variability, while combining all possible subset selection methodology with the inclusion variable transformations and interactions. Moreover, this method controls multicollinearity, leading to an optimal set of explanatory variables.

OPEN ACCESS

Citation: Tavallali P, Razavi M, Brady S (2017) A non-linear data mining parameter selection algorithm for continuous variables. PLoS ONE 12 (11): e0187676. <https://doi.org/10.1371/journal.pone.0187676>

Editor: Tiratha Raj Singh, Jaypee University of Information Technology, INDIA

Received: May 23, 2017

Accepted: October 24, 2017

Published: November 13, 2017

Copyright: © 2017 Tavallali et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All synthetic data generated or analyzed during this study are included in the Supporting Information. The human data used in this study comes from the Framingham Heart Study. This data is publicly available to qualified investigators. De-identified data can be provided to investigators of approved research proposals. Data can be requested by a submitting research application to one of the following: Directly from Framingham Heart Study (<https://www.framinghamheartstudy.org/>), BioLINCC (<https://biolincc.nhlbi.nih.gov/home/>), or dbGaP (<https://www.ncbi.nlm.nih.gov/gap>) Data

Introduction

It happens often that the physical or mathematical model behind an experiment or dataset is not known. Hence, model selection (sometimes called subset selection) becomes an important feature during the data analysis endeavor. The methodology of selecting the best model from a set of inputs has constantly been examined by many authors [1]. Identifying the best subset among many variables is the most difficult part of this effort. The latter is exacerbated as the number of possible subsets grows exponentially, if the number of variables (parameters) grows linearly. Furthermore, there is also a chance that the original input parameters to a model do not convey enough information. Some transformations of the original parameters, and specifically interactions between them, are needed to make the data more available for information extraction.

In other words, in a supervised learning terminology, there is a long and unpaved journey between the *inputs* (also called *predictors*, *features* or *independent variables*) and the *outputs* (also called *responses* or *dependent variables*). Thus, the difficulty is not only embedded in

sets used in this study can be found using the following links: 1- (<https://biolincc.nhlbi.nih.gov/studies/gen3/?q=framingham>) for the Gen3 cohort 2- (<https://biolincc.nhlbi.nih.gov/studies/framcohort/?=framingham>) for the Original Cohort 3- (<https://biolincc.nhlbi.nih.gov/studies/framoffspring/?q=framingham>) for the Offspring Cohort.

Funding: The research leading to this manuscript was not funded. The author Sean Brady (S.B.), having the affiliation at Principium Consulting, LLC, has not financially contributed to this research. This author participated in the original idea of the study through discussions with the first author, Peyman Tavallali (P.T.). S.B. helped draft the manuscript, and revised the manuscript critically for important intellectual content. S.B.'s contribution to this study has solely been individual, non-profit, scientific, and unfunded. S.B. nor Principium Consulting, LLC did not provide any financial support in any form for this study. No funder provided support in the form of salaries for authors, and no funder had any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: We declare no competing interest. The Principium Consulting, LLC is not doing research or business in the field of statistics learning. There are no marketed products, employment, consultancy, patents, and products in development relating to the material of this manuscript. The collaboration with S.B. does not alter our adherence to PLOS ONE policies on sharing data and materials. There are no restrictions on sharing of data and/or materials regarding the manuscript.

picking the right machine learning algorithm for the problem at hand, but also in picking proper transformations and interactions of the inputs or their subsets. There are different methods capable of addressing transformations and subset selection. However, to the best of our knowledge, none of these methods solves both issues simultaneously.

In our discussions in this paper, we denote the vectorial form of an input variable x by an $N \times 1$ vector \mathbf{x} as a collection of N observations. The assembly of p such inputs and an intercept is denoted by an $N \times (p + 1)$ matrix $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$. The vectorial form of the output y is denoted by an $N \times 1$ vector \mathbf{Y} . For example, based on this description, a linear model is defined as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where ε is the $N \times 1$ noise vector, and $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is a $(p + 1) \times 1$ vector of coefficients with the first element β_0 as the *intercept* (or *bias*) of the model. In what follows next, we review a series of methods and algorithms that are used to find some subset(s) of the inputs that could possibly relate the inputs to outputs in an efficient way.

Subset selection

There are currently various methods for selecting predictors, such as the traditional best subset selection, forward selection, backward selection and stepwise selection methods [1, 2]. In general, the best subset procedure finds for each $k \in \{1, 2, \dots, p\}$, the subset of inputs of size k that minimizes the Residual Sum of Squares (RSS) [3–6]. There are fast algorithms optimizing the search [7]. However, searching through all possible subsets could become laborious as p increases.

A number of automatic subset selection methods seek a subset of all inputs, that is as close as possible to the best subset method [1]. These methods select a subset of predictors by an automated algorithm that meets a predefined criterion, such as the level of significance (set by the analyst). For example, the forward selection method [1] starts with no predictors in the model. It then adds predictors one at a time until no available predictors can contribute significantly to the response variable. Once a predictor is included in the model, it remains there. On the other hand, the backward elimination technique [1] works in the opposite direction and begins with all the existing predictors in the model, then discards them one after another until all remaining predictors contribute significantly to the response variable. Stepwise subset selection [8] is a mixture of the forward and backward selection methods. It modifies the forward selection approach in that variables already in the model do not always remain in the model. Indeed, after each step in which a variable is added, all variables in the model are reevaluated via their partial F or t statistics and any non-significant variable is removed from the model. The stepwise regression requires two cutoff values for significance: one for adding variables and one for discarding variables. In general, the probability threshold for adding variables should be smaller than the probability threshold for eliminating variables [1].

Subset selection methods are usually based on targeting models with the largest R_{adj}^2 , or in other words smallest Root Mean Square Error (RMSE). However, there are other methods in which the selection model is based on Mallows's C_p [9–12]. These criteria highlight different aspects of the regression model. As a results, they can lead to models that are completely different from each other and yet not optimal.

Unfortunately, none of these subset selection methods address the issue of multicollinearity.

Ridge regression

There are also other issues regarding the traditional subset selection regression methods. They could lead to models that are unreliable for prediction because of over-fitting issues. More specifically, they could generate models that have variables displaying a high degree of multicollinearity. Such methods can lead to R^2 values that are biased and yield to confidence limits that are far too narrow or wide. Moreover, the selection criterion primarily relies on the correlation between the predictor(s) and the dependent variable. Thus, these methods (e.g. Stepwise method [13]) do not take into consideration the correlation within the predictors themselves. The latter is a source of multicollinearity that is not addressed automatically by these mentioned methods [13].

Indeed, when collinearity among the predictors exists, the variance of the coefficients is inflated, rendering the overall regression equation unstable. To address this issue, a number of *penalized regression* or *shrinkage* approaches are available. For example, the Ridge method tries to eliminate the multicollinearity by imposing a penalty on the size of the regression coefficients [2]. Indeed, a model is fitted with all the predictors, however, the estimated coefficients are shrunken towards zero relative to the least squared estimates. Therefore, biased estimators of regression coefficients are obtained, reducing the variance and thus leading to a more stable equation.

Solving for β in Eq (1) using the Least Squares (LS) method would be equivalent to solving

$$\hat{\beta}^{LS} = \arg \min_{\beta} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 . \tag{2}$$

Here, $\| \mathbf{x} \|_2 = (\sum |x_j|^2)^{\frac{1}{2}}$ is the L_2 norm of \mathbf{x} . Ridge regression, on the other hand, places a constraint on the estimator β in order to minimize a penalized sum of squares [14, 15]

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 + \lambda \| \beta \|_2^2 . \tag{3}$$

The complexity parameter $\lambda \geq 0$ controls the amount of shrinkage. Large values of this parameter would result in a large shrinkage. The value of the constant λ is predefined by the analyst and is usually selected in order to stabilize the ridge estimators, producing an improved equation with a smaller RMSE compared to the least-squares estimates. One weakness of the Ridge method is that it does not select variables. Indeed, unlike the subset selection method, it includes all of the predictors in the final model with shrunken coefficients. The other weakness is that multicollinearity is not fully addressed. In fact, the Ridge estimate of variables in (3) only shrinks the coefficients even for the inputs with multicollinearity. However, the Ridge Method does not fix multicollinearity, it only alleviates it. This issue has been shown and addressed in [16].

Lasso

To obtain variable selection procedures, there are shrinkage methods available such as Least Absolute Shrinkage and Selection Operator (Lasso), where the penalty involves the sum of the absolute values of the coefficients β excluding the intercept [17]. Lasso is closely related to sparse optimization found in works by Candes and Tao [18]. Taking $\beta^- = (\beta_1, \dots, \beta_p)^T$, the Lasso method can be presented as the following optimization problem

$$\hat{\beta}^{Lasso} = \arg \min_{\beta} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 + \lambda \| \beta^- \|_1, \tag{4}$$

where $\|\beta^-\|_1 = \sum_1^p |\beta_j|$ is the L_1 norm of β^- and $\lambda > 0$. The advantage of Lasso is that much like the best subset selection method, it performs variable selection.

The parameter λ is usually selected by cross validation. For a small λ , the result is equal to the least squares estimates. As the value of λ augments, shrinkage happens in such a way that only a sparse number of variables having an active role in the final model would show up. Thus, Lasso is a combination of both shrinkage and variables selection.

LAR

Least Angle Regression (LAR) is a new model of automatic subset selection based on a modified version of forward procedure [19]. The LAR method follows an algorithmic procedure: First, the independent variables are standardized in order to obtain a mean zero. At this stage, the β coefficients are all equal to zero. Then the predictor that most correlates to the response variable is selected; its coefficient is then shifted from zero towards its least squares value. Now, once a second predictor becomes as correlated with the existing residual as the first predictor, the procedure is paused. The second predictor is then added to the model. This procedure then continues until all desired predictors are included in the model, leading to a full least-squares fit.

The method of Least Angle Regression with Lasso modification is very similar to the above procedure, however it includes an extra step: if a coefficient approaches zero, LAR excludes its predictor from the model and recalculates the joint least squares path [2]. LAR methods and its variations are better subset selector algorithms compared to most of the subset selection methods.

Dantzig

Another selection approach is the Dantzig selector [20], which can be formulated as

$$\min_{\beta} \|\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)\|_{\infty} \tag{5}$$

subject to $\|\beta\|_1 \leq t$. Here, $\|\cdot\|_{\infty}$ is the L_{∞} norm, that is the maximum of its argument. The objective of this method is to minimize the maximum inner product of the existing residual with all the independent variables. This approach has the capacity of recovering an underlying sparse coefficient vector.

Knockoff filter

This method is recently introduced as a new variable selection method to control the false discovery rate (FDR) [21], for linear models. For a selected subset of variable indices \hat{S} , the FDR is formally defined as

$$FDR = \mathbb{E} \left(\frac{\#\{j|\beta_j = 0, j \in \hat{S}\}}{\max(\#\{j|j \in \hat{S}\}, 1)} \right). \tag{6}$$

It is also well-suited for high-dimensional linear models in which the number of features are more than the number of data points. This method is capable of being combined with different methods, such as Lasso explained above, to perform a more reliable variable selection in the context of controlling the FDR.

PCR

Lastly, Principal Component Regression (PCR) is a method that involves an orthogonal transformation to address multicollinearity [2, 22, 23]. This approach is closely related to the

Singular Value Decomposition (SVD) method [24]. PCR applies dimensionality reduction and decreases multicollinearity by using a subset of the principal components in the model [2]. PCR is one of very few methods that tries to eliminate multicollinearity with linear transformations and, at the same time, perform a regression.

The various approaches described so far aim to select the best set of relevant variables from an original set. With the exception of the PCR method, in which there are linear transformations, variables transformations are not incorporated among predictors in any of the methods mentioned above. These traditional methods do not offer the option of automatic variable transformation to address polynomial curvilinear relationships. No non-linear interpretable interaction of the predictors is available in them. An analyst usually needs to manually apply polynomial, logarithmic, square-root and interaction-between-variables transformations in order to address non-linearity of the data.

Non-Linear transformation. There are a number of non-linear transformation procedures currently available such as Box-Cox or Box-Tidwell [25, 26]. These methods are relatively efficient in finding the dependent and independent variables transformations. In Box-Tidwell method [26], independent variables are transformed using a recursive Newton algorithm. As a result, it becomes susceptible to round-off errors which would in turn result in unstable and improper transformations [1]. Despite the relative success of these methods, there is no automatic variable selection embodiment with them.

Artificial Neural Networks (ANN) are the current state of the art method in transformations and capturing non-linearity [2, 27]. ANN is a machine learning method that finds some non-linear transformations of the inputs using layers of nodes. One recent exemplary example is the Deep Neural Networks (DNN) used in speech recognition [28]. Despite the efficient performance in capturing the non-linearity of the data, the model itself is not comprehensible particularly if there is a physical component to the data that one needs to interpret or understand. In other words, ANN is a perfect black box model, but not a good interpretable medium for understanding physical and mathematical mechanism(s) behind the observed data.

Subset selection and transformation. As mentioned earlier, only the PCR method performs linear transformations automatically, and also picks variables. However, PCR is not enough when non-linearity is present. On the other hand, ANN has the best capability in capturing non-linearities, but acts like a black box and does not lend insight into the physical and mathematical mechanism(s) behind the observed data.

To produce an enhanced subset of the original variables, an effective selection method should have the potential of adding a supplementary level of regression analysis that would capture complex relationships in the data via mathematical transformation of the predictors and exploration of synergistic effects of combined variables in an interpretable fashion. The method that we present here has the potential to produce an optimal subset of variables, which is even interpretable in the presence of non-linear interaction between the inputs, resulting in a more efficient overall process of model selection.

The core objective of this paper is to introduce a new estimation technique for the classical least square regression framework. This new automatic variable transformation and model selection method could offer an efficient and stable model that minimizes the mean square error and variability, while combining all possible subset selection methodologies and including variable transformations and interaction. Moreover, this novel method controls multicollinearity, leading to an optimal set of explanatory variables. The final model is also easy to interpret. In other words, we will depict a method that tries to address variable selection, interpretation, non-linear interaction and transformation at the same time.

Materials and methods

Problem definition

We assume \mathcal{T} to be the set of all transformations on a given set of inputs $\{x_i\}$, for $i \in \{1, \dots, p\}$ and $x_i \in \mathbb{R}$. One possible formulation, to find the best subset and transformation estimating a dependent variable $y \in \mathbb{R}$, can be expressed as

$$\begin{aligned} & \underset{f \in \mathcal{T}}{\text{minimize}} && \|y - f(\{x_j\}_{j \in \Omega})\| \\ & \Omega \subset \mathcal{P}(\{1, \dots, p\}) \end{aligned} \tag{7}$$

Here, one desirable candidate for the norm $\|\cdot\|$ could be the L_2 norm, since the purpose is regression. Also, $\mathcal{P}(\cdot)$ is the power set. This is an NP hard problem. As a result, we need to find approximations of this problem to make it traceable.

In the first step, we confine ourselves to a set of certain functions in \mathcal{T} that are easy to interpret from a casual physical perspective. We call this set \mathcal{F} . For example we could pick only the polynomial transformations. Consequently, the set of all transformed variables would be

$$Z = \{\mathcal{F}(\{x_i\}_{i \in \{1, \dots, p\}})\}. \tag{8}$$

This step would reduce the search space for (7). However, there are sources of redundancy which we could minimize or eliminate. Knowing this, the next step could be to pick transformed variables that have a significant absolute value correlation ρ_{zy} with the output y . This set can be expressed as

$$Z^\delta = \{z \in Z \mid \rho_{zy} \geq \delta, \delta > 0\}. \tag{9}$$

Also, there is a chance that many of the elements in Z^δ are strongly correlated with each other. Later, this could be a serious source of multi-collinearity. So, we could further trim Z^δ by only picking the most correlated variables to the output among two correlated variables. This would reduce the set Z^δ to

$$\begin{aligned} Z^r = & \left\{ z \in Z^\delta \mid z = \underset{\alpha, \beta \in Z^\delta}{\text{arg max}} (\rho_{\alpha y}, \rho_{\beta y}) \text{ s.t. } \rho_{\alpha\beta} > \varsigma > 0 \right\} \cup \\ & \{z \in Z^\delta \mid \forall \alpha \in Z^\delta, \alpha \neq z, \varsigma > 0 \Rightarrow \rho_{\alpha z} < \varsigma\}. \end{aligned} \tag{10}$$

Here, $\rho_{\alpha\beta}$ is the absolute value correlation between α and β .

At this stage, using (8)-(10), and considering that we are looking for a linear estimator among these reduced transformations, the optimization problem (7) would become

$$\underset{\{\beta_i\}_{i \in \mathbb{T}, \mathbb{T} \subset \{1, \dots, |Z^r|\}}}{\text{minimize}} \left\| y - \sum_{i \in \mathbb{T}} \beta_i z_i^r \right\|_2. \tag{11}$$

Here, $Z^r = \{z_i^r\}_{i \in \{1, \dots, |Z^r|\}}$ and $|Z^r|$ is the cardinality of Z^r . The optimization problem (11) is nothing but a subset selection model and could be approximated by any methods of subset selection [1, 2]. Hence, we now have a model (11) that not only takes care of some desirable interpretable transformations, but also extracts the most meaningful set of parameters.

Note. As we intend to provide a data mining method rather than a pure statistical one, the easy interpretation would act as a constraint on the types of transformations in (7). For example, in a medical investigation, the investigator is mainly looking for basic algebraic interactions between the inputs which can provide physiological view of the system under scrutiny. Hence, the non-linear transformations and interaction between terms must be as basic as

possible, such as exponents, logarithms, multiplications and etc. On the other hand, a linear model, like (11), should be used to keep the interpretability of the model intact providing a robust and accurate model. By this formulation, we are trying to deploy an interpretable and accurate data mining model, instead of a black-box pure statistical learning method. Our effort is not to compete with statistical learning methods, but to provide an easy and a faithful-fit data mining method. In the next section, we are going to discuss our methodology in more practical detail.

Methodology

As mentioned before, we are looking for transformations that are easy to interpret. There are four main transformation categories of this type capturing the non-linearity in a data set [2]. These transformations are as follow but not limited to

1. Logarithmic transformation of a positive variable; i.e. $\log x_j$,
2. Square-root transformation of a positive variable; i.e. $\sqrt{x_j}$,
3. Integer powers up to a certain amount $\alpha \in \mathbb{N}$; i.e. $\{\frac{1}{x_j^\alpha}, \frac{1}{x_j^{\alpha-1}}, \dots, x_j^{\alpha-1}, x_j^\alpha\}$,
4. Interactions between terms created in 1-3 up to a certain amount M ; e.g., for $M = 2$, possible candidates would be $\frac{1}{x_i} \frac{x_j^2}{x_j}$, $x_i x_j^2$, $x_i^2 x_j^2$, $x_i^2 (\log x_j)^2$ and $\frac{\sqrt{x_i}}{x_j}$.

We are going to use this set of transformations, namely $\mathcal{F}(\{x_j\}_{j \in \Omega}; \alpha, M)$, for the rest of this paper. After the construction of these interpretable interactions transformations, one can start to look for the best model, for \mathbf{Y} , among the set of all transformations 1-4. Here, $\mathbf{Y} \in \mathbb{R}^N$ is the vector form of the output y .

Denoting the set of variables created by transformations 1-4 as \mathbf{Z} , which is the matrix form of Z in (8), we are looking for the best model

$$\mathbf{Y} = \mathbf{Z}\beta_z + \varepsilon, \tag{12}$$

where some elements of β_z are zero. We note that we could further equip our algorithm with Standardized Regression (similar to the first step of the LAR method) to diminish the possibility of a numerically ill-conditioned variable matrix Z . In fact, some elements of β_z are zero since there is a chance that some columns of \mathbf{Z} are linearly dependent or that they do not contribute to any correlation with \mathbf{Y} . We can address these two issues, by a modified dictionary search [17] algorithm as follows. This part stands out for (9) and (10).

- Any column of \mathbf{Z} that has a non-significant correlation (less than δ) with \mathbf{Y} can be discarded; see (9).
- Any two columns of \mathbf{Z} that have a high correlation to each other (greater than ς) are redundant columns. Between these two columns, the one that has a higher correlation with \mathbf{Y} is picked and the other is discarded; see (10).

As a result of this methodology, we can now solve model (12) for only a reduced matrix. We denote this reduced matrix as \mathbf{Z}^r and its corresponding vector of coefficients as β_z^r .

The final task is to find the best subset of the columns in \mathbf{Z}^r to model the data in \mathbf{Y} . The latter can be done by any method of subset selection including the best subset selection method [1]. The subset selection method that we have used in our implementation is based on

targeting models with the largest R_{adj}^2 , or in other words smallest RMSE. As a reference point, we call our methodology the Parameter Selection Algorithm.

Parameter selection algorithm

The goal of the Parameter Selection Algorithm is to find the best interpretable model on the original observed variables \mathbf{X} , from a set of basic transformations, estimating \mathbf{Y} . Our method is summarized in Algorithm 1. Step 1 of this algorithm is input specification. Step 2 is where the dictionary of transformations and interactions is made. Steps 3 and 4 correspond to the elimination of columns of the dictionary which involve either a non-significant correlation to the output or multicollinearity between its elements. Step 5 is where the best model is finally found, subject to the constraint that the final set of variables has a Variation Inflation Factor (*VIF*) less than 10. *VIF* elements are the main diagonal values of the inverse of the multiplication of the input matrix transposed with the input matrix. For example if \mathbf{X} is the input, then $\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1}$ and $VIF_j = C_{jj}$ [1]. Although we eliminate similar-looking variables in step 4, checking for the *VIF* [29] is a necessary condition to make sure that no multicollinearity is introduced into the final model. In practice, step 5 can be solved by maximizing the R_{adj}^2 among all possible subsets of the variables in \mathbf{Z}^r [1].

Steps 2, 3 and 5 in this algorithm can be made parallel to decrease the computational time of the method. To our best knowledge, Algorithm 1 is the first linear data mining method that performs both variable transformation and model selection while adding interaction terms and also preventing multi-collinearity, in one package.

The hyper-parameters δ and ς are important factors in controlling the speed of convergence of the Parameter Selection Algorithm. In Algorithm 1, the smaller the value of δ (similarly, the larger the value of ς), the bigger the space of search in step 5. As a result, the speed of convergence would depend greatly on these two parameters.

Algorithm 1 Parameter Selection Algorithm

1. Inputs to the algorithm: \mathbf{X} , \mathbf{Y} , α , M , δ , ς .
2. Construct the matrix of transformations $\mathbf{Z} = \mathcal{F}(\mathbf{X}; \alpha, M)$.
3. Construct the matrix \mathbf{z}^δ from \mathbf{Z} .
4. Construct the matrix \mathbf{z}^ς from \mathbf{z}^δ .
5. Solve $\underset{\{\beta_i\}_{i \in \tau}}{\text{minimize}} \left\| \mathbf{Y} - \sum_{i \in \tau} \beta_i \mathbf{z}_i^\varsigma \right\|_2$ subject to $VIF \leq 10$.

Candidates for ς and δ . The hyper-parameter δ is straightforward to settle. Most of the contribution of a model comes from variables having a high univariate correlation coefficient with the output. As a result, we could discard variables with smaller contributions. Here, small is measured with respect to the highest absolute value univariate correlation coefficient with the output. Usually, an absolute value univariate correlation coefficient of 0.5, and above, is considered to be high [30]. This is 50% of the maximum allowed absolute correlation of 1. Hence, from a conservative perspective, we could set δ to be half of the maximum highest absolute value univariate correlation coefficient among all variables. We call this the default value of δ .

On the other hand, the hyper-parameter ς can be characterized with the *VIF* concept. Each element of the *VIF* vector can be expressed as

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{13}$$

Here, R_j^2 is the multiple R^2 for the regression of x_j against other inputs. Hence, if we want two inputs to have a small correlation with each other, we must have a possible *VIF* between them to be less than 10. This would impose an $R^2 = 0.9$ between those variables. Hence, a correlation of ~ 0.95 would say if two inputs are highly correlated or not. On the other hand, we know that if we set the independence limit $\varsigma = 0.95$, we would construct a huge dictionary of inputs when transformations are available. To have a balance between the two, our recommendation is $\varsigma = 0.80$.

Synthetic examples

In this section, we provide a few synthetic examples using Parameter Selection Algorithm. In the following examples, we try to show that the algorithm that we have proposed is capable of finding the non-linear transformations in a model.

Example. Taking x_1, x_2, x_3 to be independent uniformly distributed random variables between 0 and 100, we sampled 1000 data points and then created the non-linear functional $y = 120 + 80x_1x_3$. We take the original input matrix \mathbf{X} to be composed of all x_1, x_2 , and x_3 . Using the traditional best subset selection [7], accompanied with a control over *VIF* not to get above 10, we get the results shown in Fig 1. From this figure, it is clear that the best subset selection model is not capable of capturing the correct non-linearity in the model. The heteroscedasticity of the residual plot can be seen in Fig 2. The found best subset of parameters is $\{x_1, x_2, x_3\}$. On the other hand, if Algorithm 1 is used, with a strict choice of $\varsigma = 0.5$ and the default value of δ , the non-linearity is captured completely by our method (See Figs 3 and 4). The subset of parameters found by our method is the model non-linear parameter $\{x_1x_3\}$.

Example. If χ is a uniform random random variable between 0 and 1, we set

$$\begin{aligned} x_1 &= 100\chi, \\ x_2 &= \chi + 0.1, \\ x_3 &= 100\chi. \end{aligned} \tag{14}$$

We sampled 1000 data points of x_1, x_2 , and x_3 and then created the non-linear functional $y = 120 + \frac{1000}{x_2}$. We take the original input matrix \mathbf{X} to be composed of all x_1, x_2 , and x_3 . Using the traditional best subset selection [7], accompanied with a control over *VIF* not to get above 10, we get the results shown in Fig 5. Again, from this figure, it is clear that the best subset selection model is not capable of capturing the correct non-linearity in the model. The heteroscedasticity of the residual plot can be seen in Fig 6. The found subset of parameters is $\{x_1, x_2\}$. On the other hand, if Algorithm 1 is used, the non-linearity is captured completely (See Figs 7 and 8). The subset of parameters found by our proposed method is the non-linear parameter $\{\frac{1}{x_2}\}$. Here, $\varsigma = 0.8$ and the default value of δ was used.

Real data example

The synthetic examples in the previous section showed the capability of our method in capturing the true non-linearity of a dataset. In this section, we show a real data case study.

Cardiovascular Diseases (CVDs) are the major cause of deaths in the United States, killing more than 350,000 people every year [31]. One of the major contributors to CVDs is arterial stiffness [32, 33]. Arterial stiffness can be approximated by Carotid-femoral Pulse Wave Velocity (PWV) [34]. In fact, PWV is one of the most important quantitative index for arterial stiffness [33]. PWV measures the speed of the arterial pressure waves traveling along the blood vessels and higher PWV usually highlights stiffer arteries. Increased aortic stiffness is related to

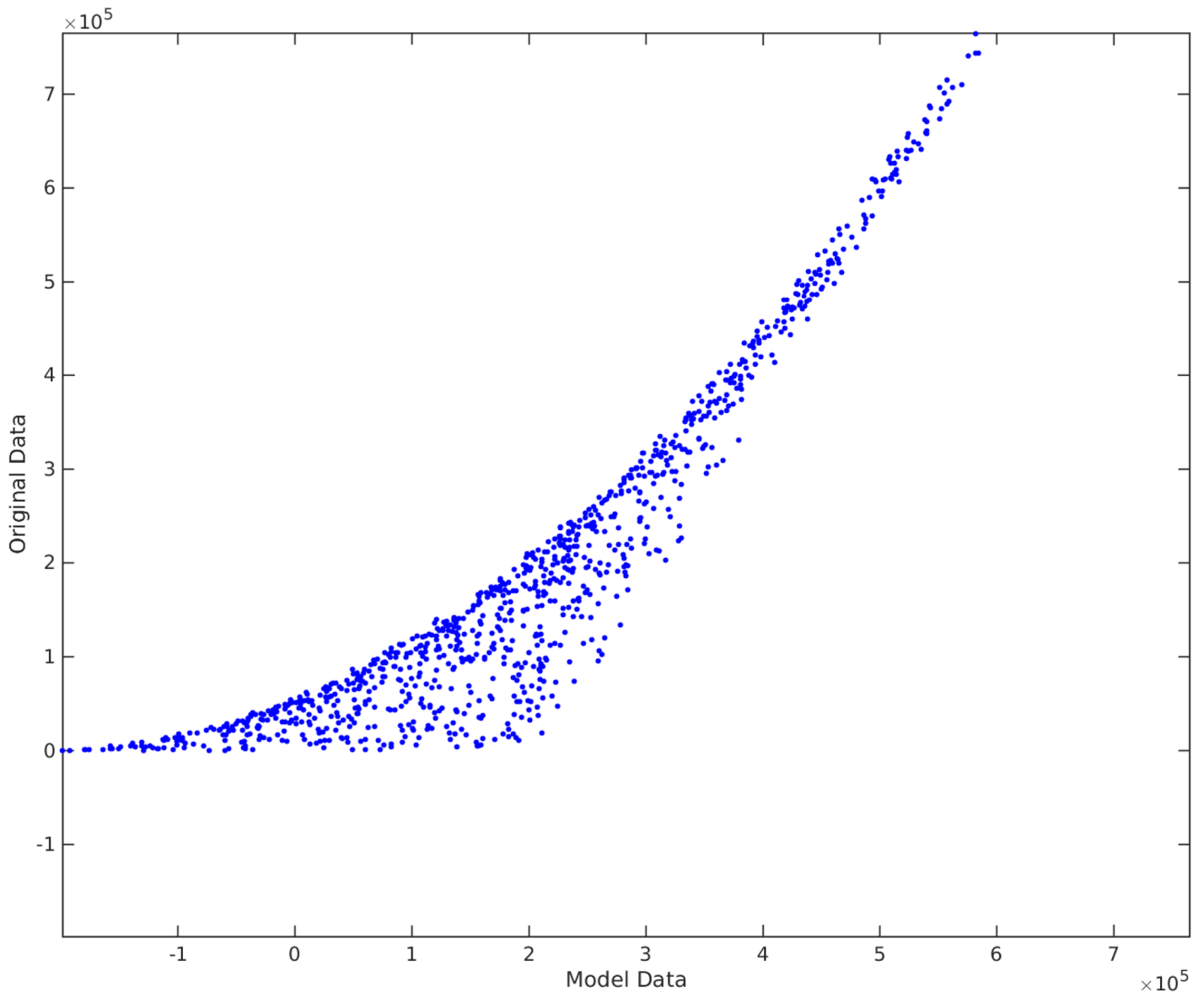


Fig 1. Traditional best subset selection method applied on $y = 120 + 80x_1x_3$. The horizontal axis shows the model found by the best subset selection method. The vertical axis shows the output y .

<https://doi.org/10.1371/journal.pone.0187676.g001>

many clinically adverse cardiovascular outcomes [32]. PWV constitutes an independent and valuable marker for cardiovascular diseases (CVDs) and its use is crucial as a routine tool for clinical patient assessment.

In this section, our aim is not to present the most accurate PWV model. However, our goal is to show that if our technique of model construction is used (see Algorithm 1), we are able to find a more interpretable model.

The data we present is collected from 5444 Framingham Heart Study (FHS) participants [35]. Each participant had undergone an arterial tonometry data collection. The participants

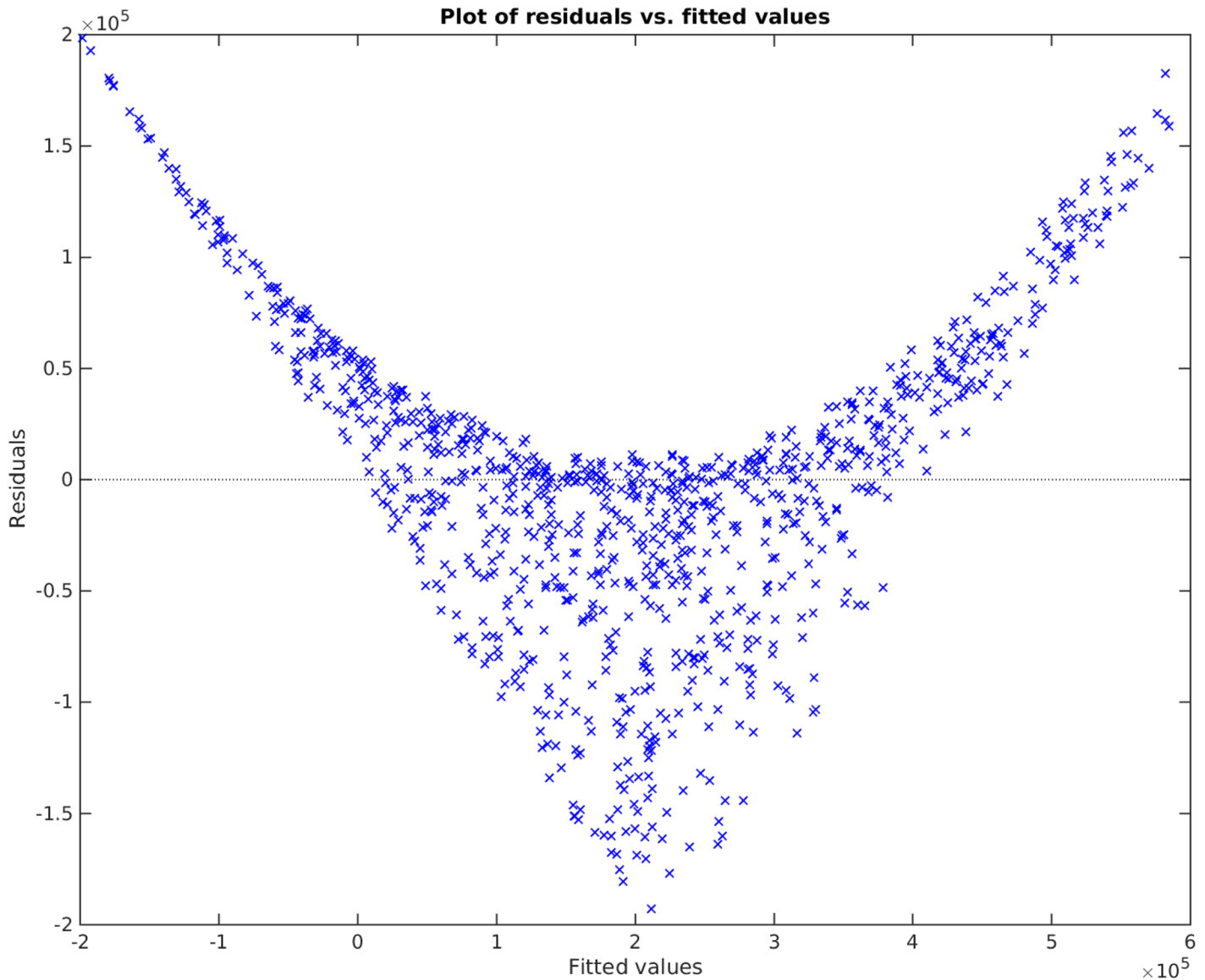


Fig 2. Residual plot of the best subset selection method applied on $y = 120 + 80x_1x_3$.

<https://doi.org/10.1371/journal.pone.0187676.g002>

were part of FHS Cohorts Gen 3 Exam 1 [36], Offspring Exam 7 [37], and Original Exam 26 [38]. The California Institute of Technology and Boston University Medical Center Institutional Review Boards approved the protocol and all participants gave written informed consent. Here, we try to find models for PWV based on the following inputs: Age (A), Pulse Duration (D), Weight (W), Height (H), and Body Mass Index (BMI).

One model is based on the traditional best subset selection method monitored for $VIF < 10$, and the other based on the Parameter Selection Algorithm method (Algorithm 1). The participant characteristics are shown in Table 1.

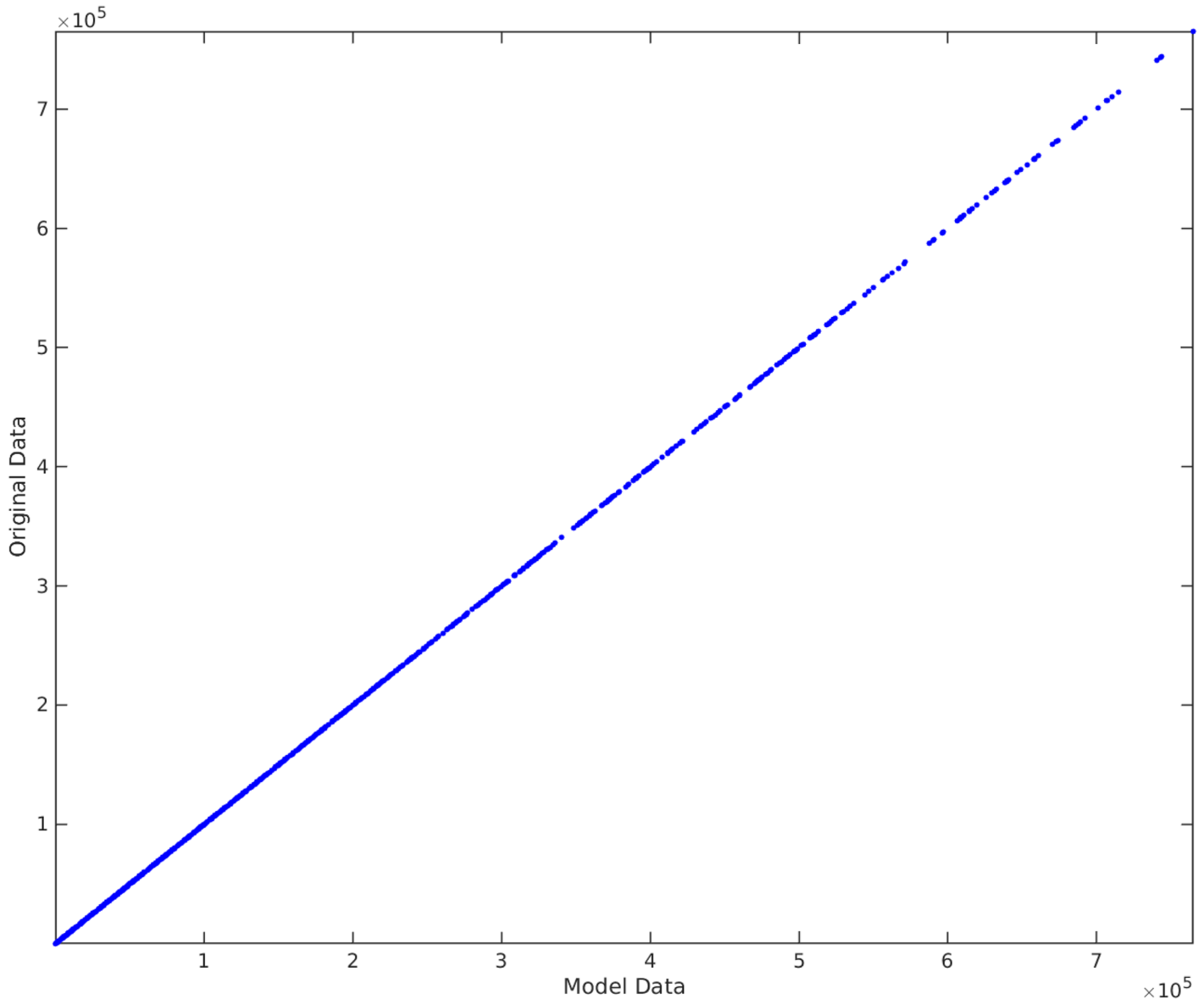


Fig 3. Algorithm 1 applied on $y = 120 + 80x_1x_3$. The horizontal axis shows the model found by our proposed method. The vertical axis shows the output y .
<https://doi.org/10.1371/journal.pone.0187676.g003>

Best subset selection model results

Fig 9 shows the traditional best subset selection method applied on PWV data. As seen in the plot, the best subset selection model cannot capture the non-linearity in the data set and completely misses the PWV values above 15. The heteroscedasticity of the residual can be seen from the Bland-Altman plot in Fig 10 and residual plot in Fig 11. The R^2_{adj} of this model is 0.56737. The found subset of parameters is $\{D, A, BMI, H\}$. The p-value of these parameters are 3×10^{-45} , $0, 2 \times 10^{-14}$, and 1×10^{-11} , receptively.

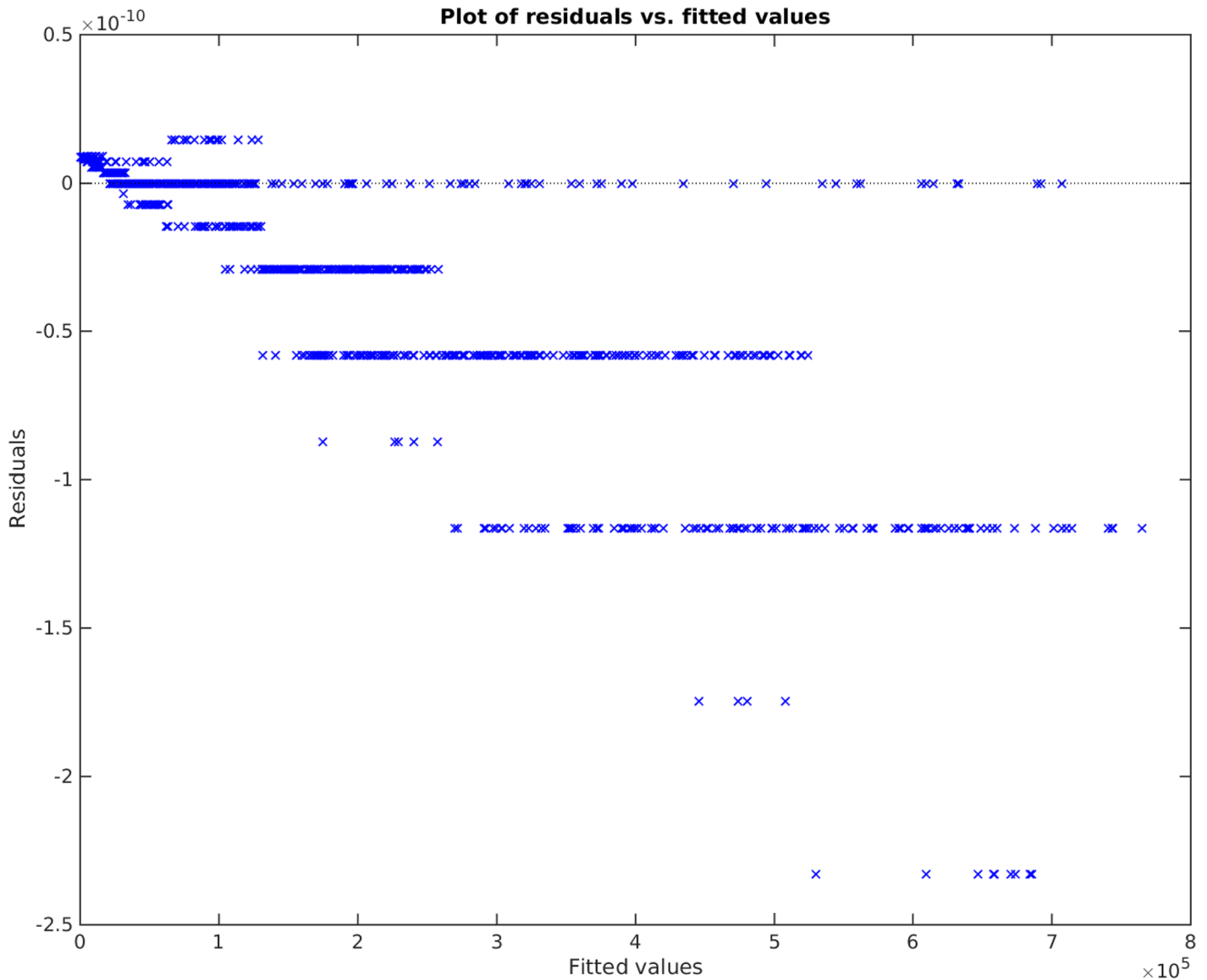


Fig 4. Residual plot of our proposed method applied on $y = 120 + 80x_1x_3$. Note that the vertical axis is of the order 10^{-10} . The error perceived here is due to floating point and rounding error.

<https://doi.org/10.1371/journal.pone.0187676.g004>

Parameter selection algorithm results

Fig 12 shows the Parameter Selection Algorithm method applied on PWV data. Here, δ was the default value and $\zeta = 0.8$. As seen on the plot, Parameter Selection Algorithm can fairly capture the non-linearity in the data set. The residuals can be seen in the Bland-Altman plot in Fig 13 and residual plot in Fig 14. The R^2_{adj} of the model is 0.63052 (The correlation coefficient

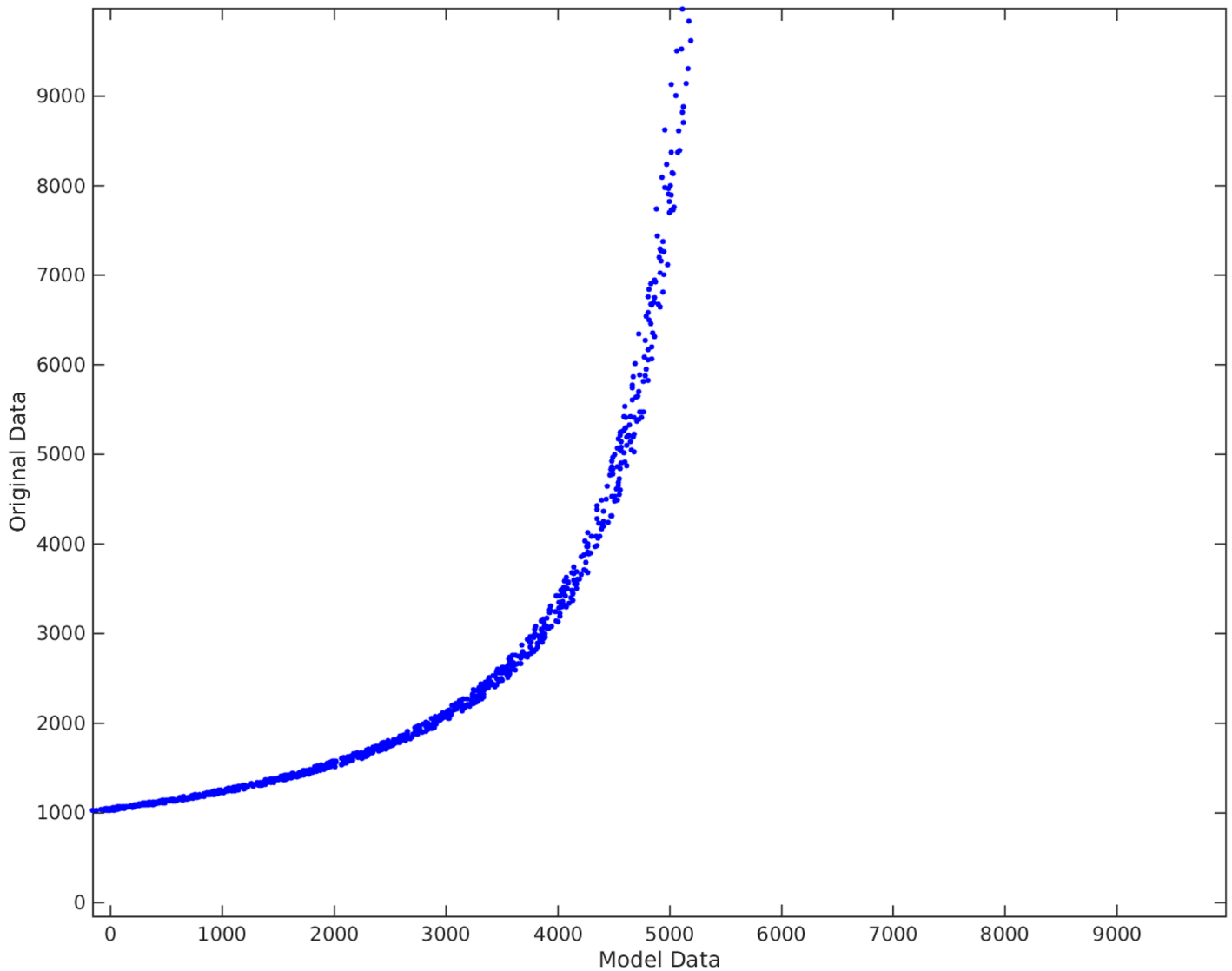


Fig 5. Traditional best subset selection method applied on $y = 120 + \frac{1000}{x^2}$. The horizontal axis shows the model found by the best subset selection method. The vertical axis shows the output y .

<https://doi.org/10.1371/journal.pone.0187676.g005>

is 0.79). The found subset of parameters is

$$\left\{ DA, \frac{A^2}{\sqrt{D}}, \frac{(\log(A))^2}{(\log(W))^2}, \frac{(\log(BMI))^2}{A} \right\}. \tag{15}$$

The p-value of these parameters are 6×10^{-4} , $0, 2 \times 10^{-21}$, and 8×10^{-46} , receptively.

From (15) one can interpret that Age (A) is a dominant factor in PWV. Furthermore, The Age adjustments with Heart Rate $HR = \frac{60}{D}$ is of great importance. The other interpretable factors are the adjusted values of slenderness (body mass index BMI , and the weight W) with

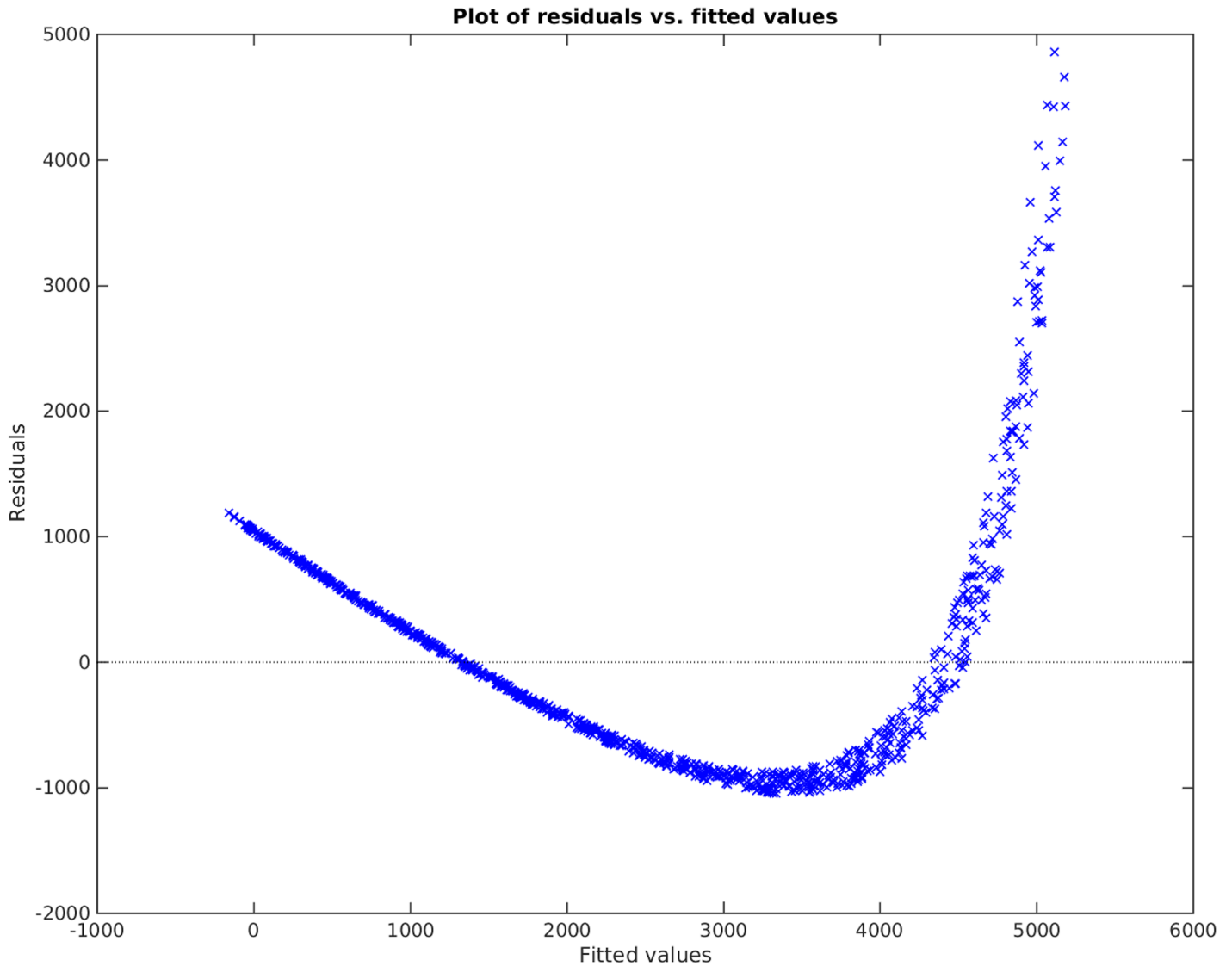


Fig 6. Residual plot of the best subset selection method applied on $y = 120 + \frac{1000}{x^2}$.

<https://doi.org/10.1371/journal.pone.0187676.g006>

respect to Age (A). Height (H) is not a factor of importance at all. As we can observe, the Parameter Selection Algorithm can provide an interpretable non-linear model of this critical physiological parameter.

Comparison and results discussion

Comparing Figs 9 and 12, it is clear that the Parameter Selection Algorithm method is superior to the best subset selection method. The R_{adj}^2 of the Parameter Selection Algorithm model is almost %11 better than the best subset selection method. Both methods suffer in capturing all the variation and non-linearity in data (compare Fig 10 to Fig 13 and Fig 11 to Fig 14).

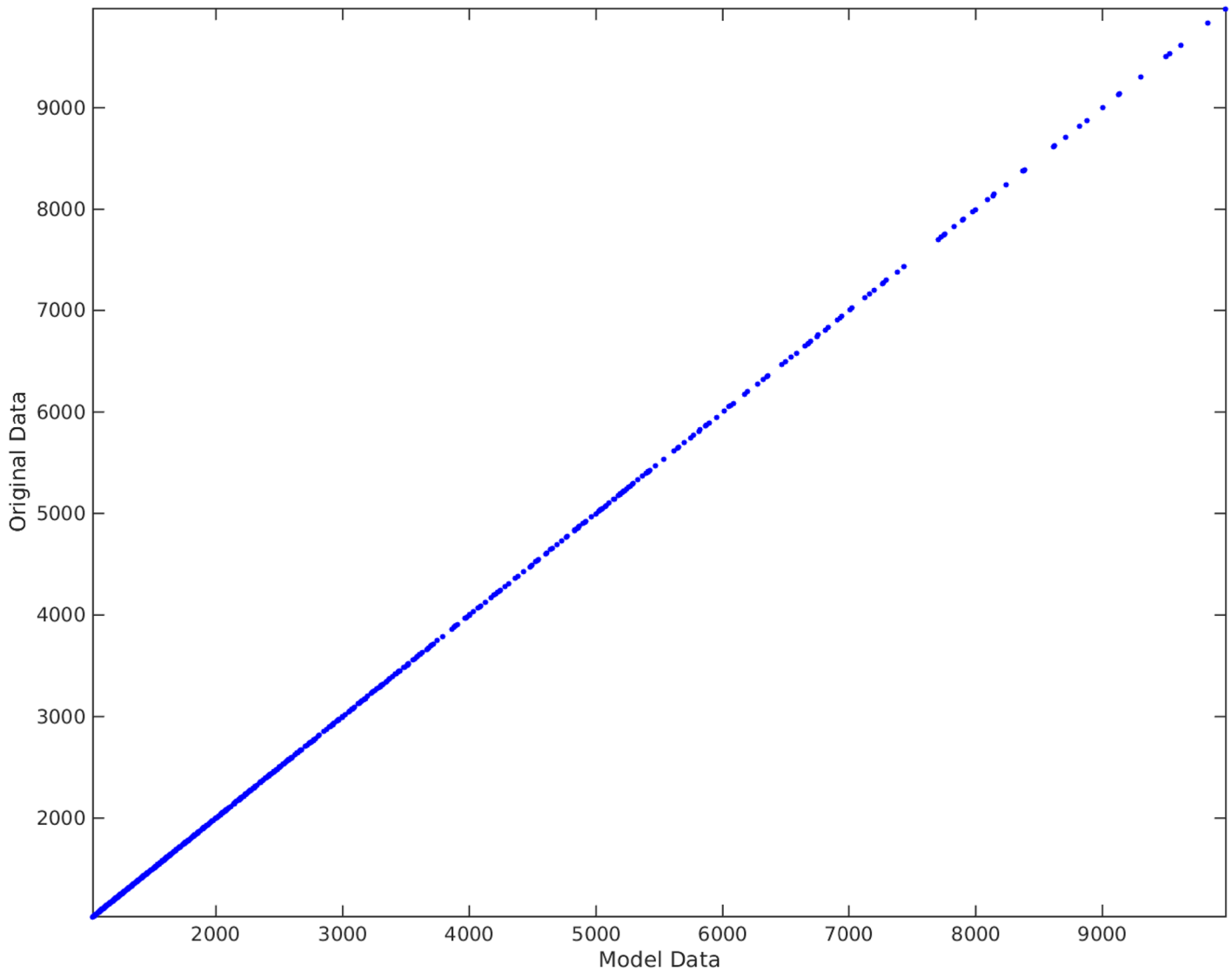


Fig 7. Algorithm 1 applied on $y = 120 + \frac{1000}{x^2}$. The horizontal axis shows the model found by our proposed method. The vertical axis shows the output y .

<https://doi.org/10.1371/journal.pone.0187676.g007>

However, Parameter Selection Algorithm is better in this respect. The heteroscedasticity of the best subset selection method is worse than that of the Parameter Selection Algorithm method (compare Fig 11 to Fig 14). The Bland-Altman limits of agreement of the Parameter Selection Algorithm method is also better than those of the best subset selection method (compare Fig 10 to Fig 13). The latter shows that the Parameter Selection Algorithm method is a more precise method than the best subset selection method.

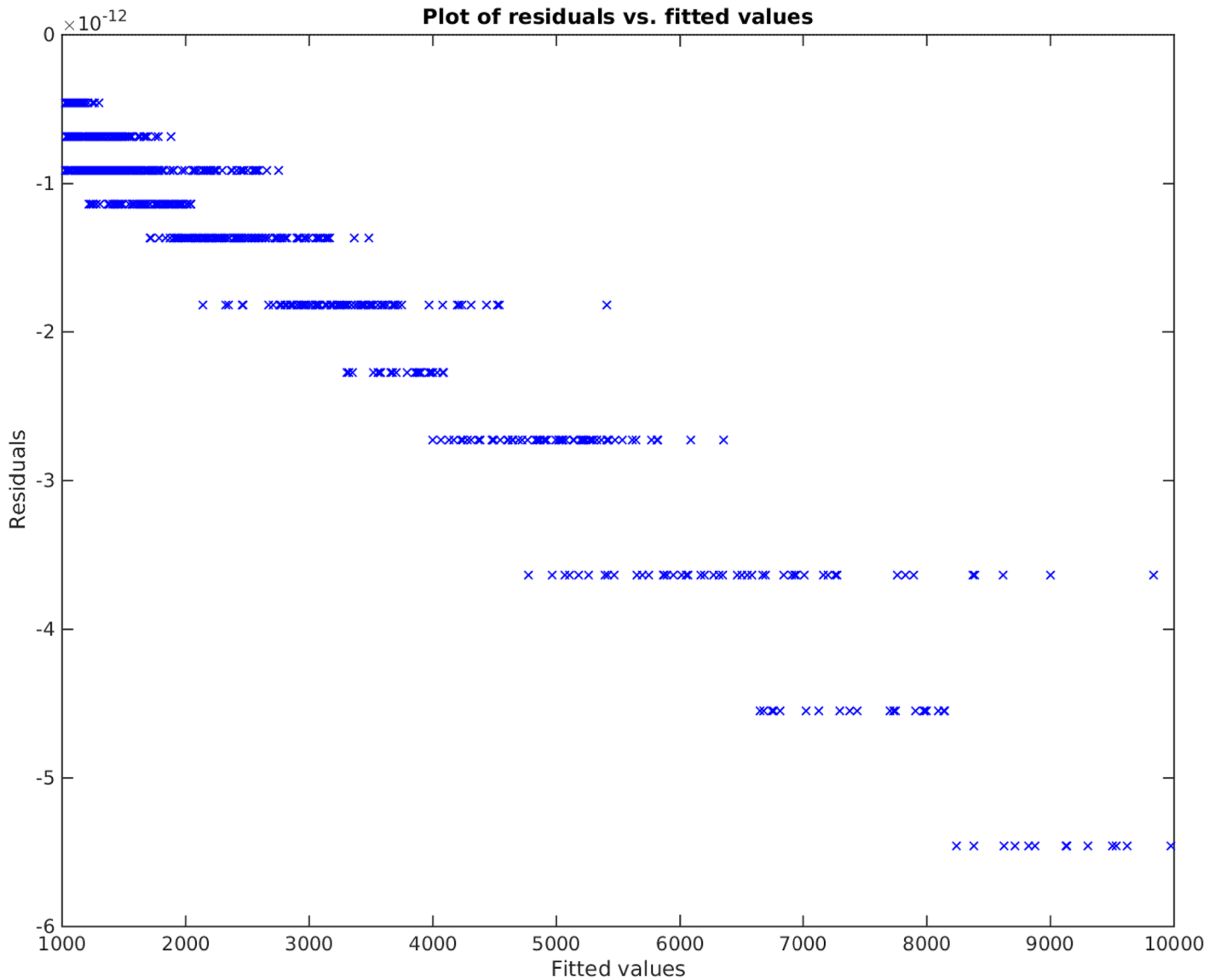


Fig 8. Residual plot of our proposed method applied on $y = 120 + \frac{1000}{x^2}$. Note that the vertical axis is of the order 10^{-12} . The error perceived here is due to floating point and rounding error.

<https://doi.org/10.1371/journal.pone.0187676.g008>

Table 1. Participant characteristics.

| | Range | Median |
|----------|----------------|--------|
| Duration | 0.58 to 1.77 | 0.98 |
| Age | 19 to 99 | 46 |
| Weight | 83 to 339 | 165 |
| Height | 54.00 to 78.75 | 66.50 |
| BMI | 15.47 to 51.47 | 25.89 |
| PWV | 3.5 to 29.6 | 7.4 |

<https://doi.org/10.1371/journal.pone.0187676.t001>

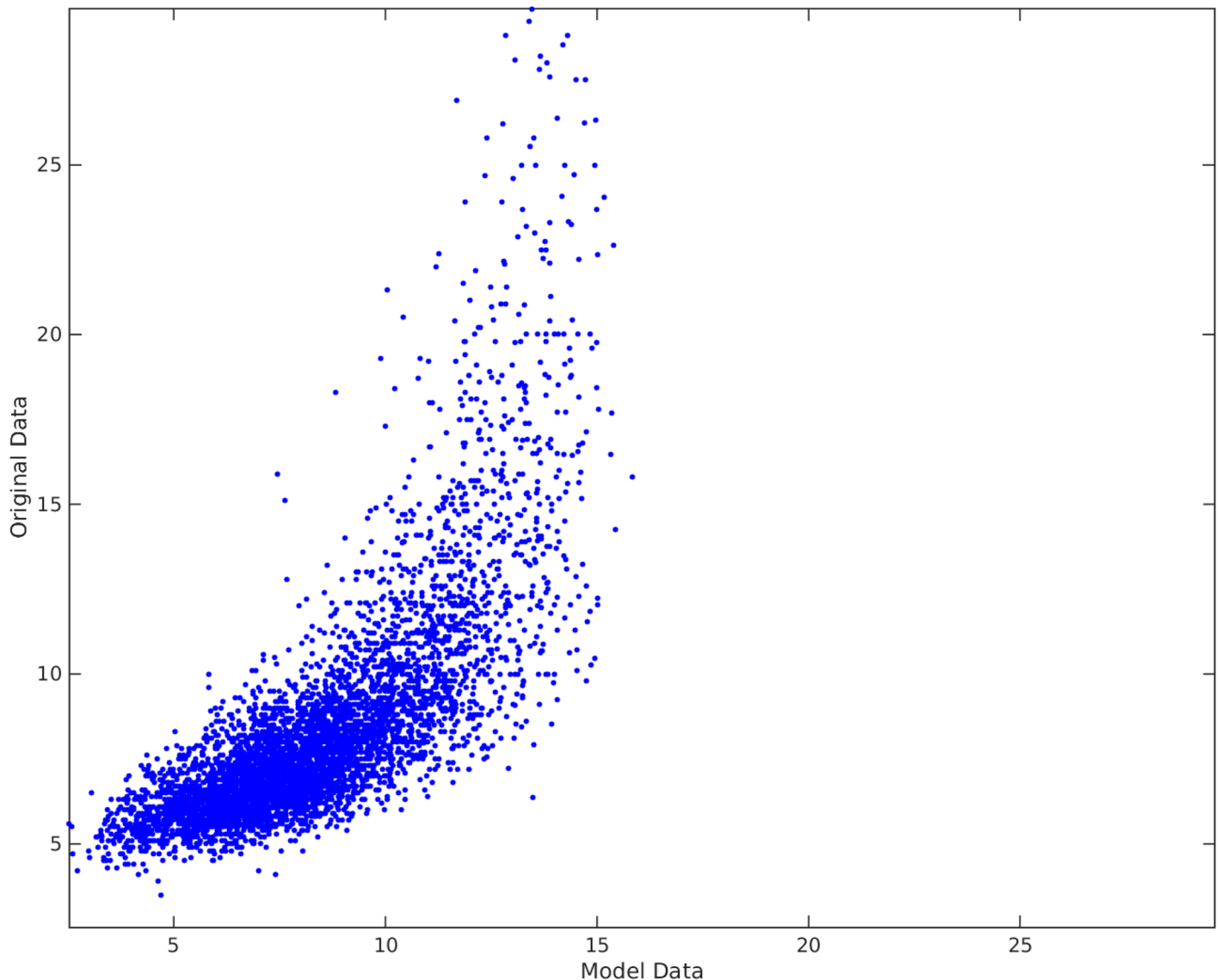


Fig 9. Traditional best subset selection method applied on PWV data. The horizontal axis shows the model found by the best subset selection method. The vertical axis shows the recorded PWV data. The R^2_{adj} of the model is 0.56737.

<https://doi.org/10.1371/journal.pone.0187676.g009>

Comparison with neural networks

Although our purpose, in this article, is not to compete with state of the art statistical learning algorithms, we decided to compare our results with ANN. We provided the same input $\{D, A, W, BMI, H\}$ to a neural network with five nodes. The output estimate of the neural network had a 0.81 correlation with the true values. Our method has a correlation coefficient of 0.79. Although the Parameter Selection Algorithm is designed mainly to act as an interpretable data mining method, it has a relatively acceptable accuracy. The 0.02 drop in correlation coefficient could possibly be neglected with the fact that, compared to neural networks, the non-linear

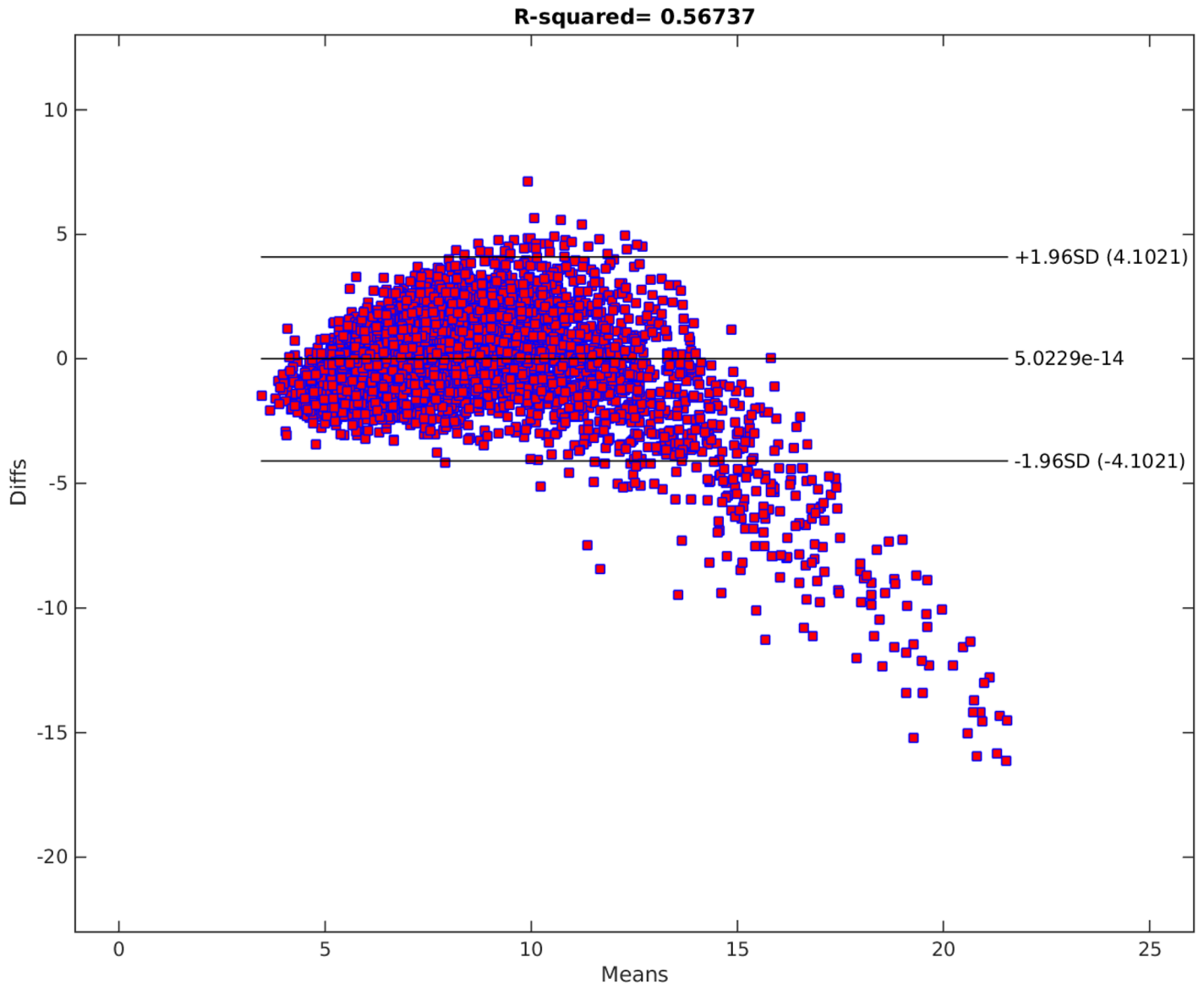


Fig 10. Bland-Altman of the traditional best subset selection model. The horizontal axis shows the means of the fitted and original PWV values. The Vertical axis shows the differences between the fitted and original PWV values.

<https://doi.org/10.1371/journal.pone.0187676.g010>

output of the Parameter Selection Algorithm is interpretable and also behaves as a dimensionality reduction algorithm.

Finally, we again mention that our goal is not to show the best possible model for PWV, but rather to show the capabilities of our presented method.

Other applications

The interpretability of Algorithm 1 would be an advantage in analyzing physiological data. In other words, complex biomedical and bioengineering databases could be appropriate fits for this method.

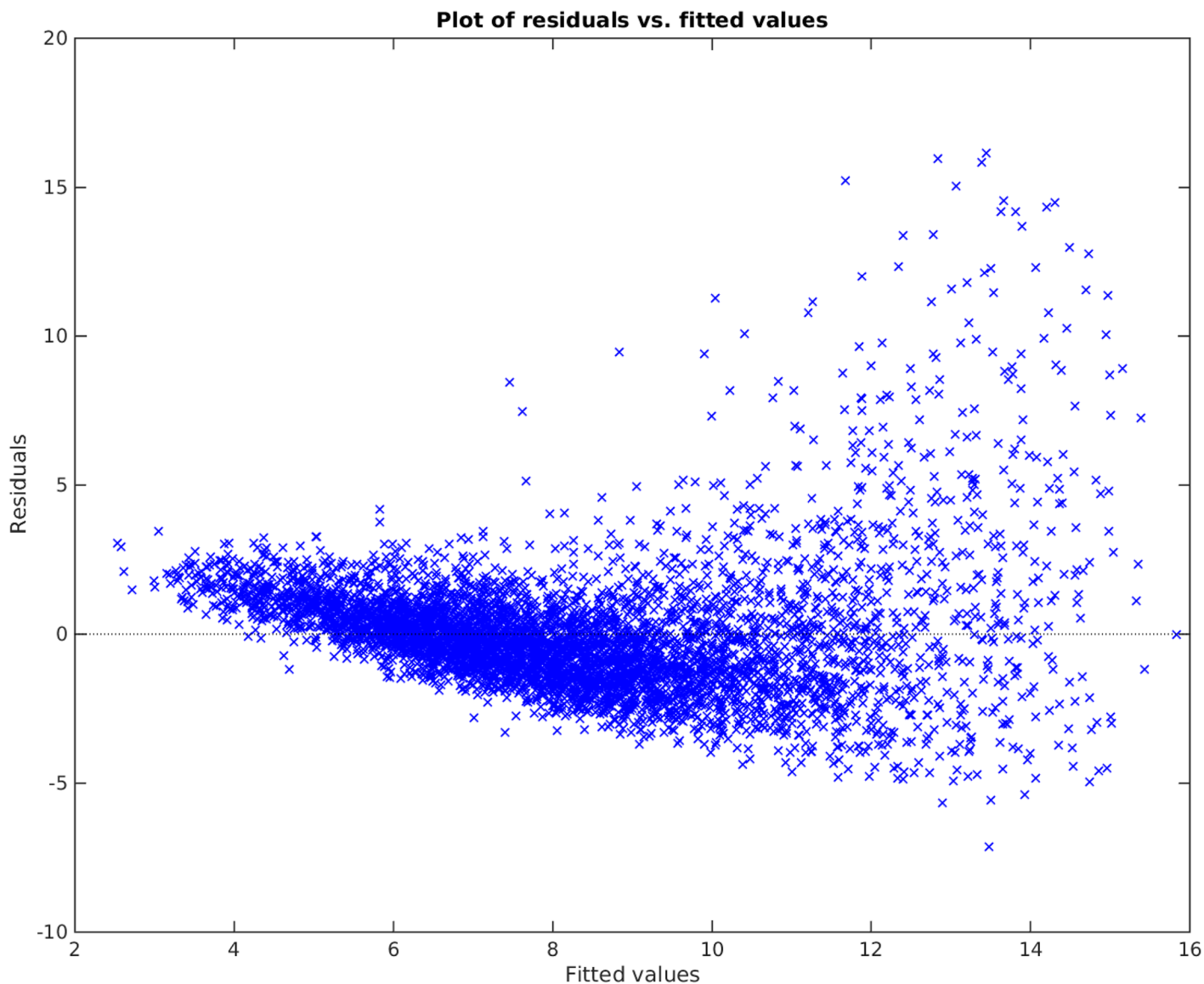


Fig 11. Residual plot of best subset selection method.

<https://doi.org/10.1371/journal.pone.0187676.g011>

In previous section, we expressed the application of our algorithm to PWV data. This suggests that any continuous physiological variable can be treated the same way. For example, important biomedical continuous variables such as Cardiac Output (CO) [39], Ejection Fraction (EF) [40], Stroke Volume (SV) [39], Blood Pressure (BP), and Homeostatic Model Assessment (HOMA) [41] can all be estimated and interpreted using Algorithm 1. This list variables can be extended beyond the mentioned cases.

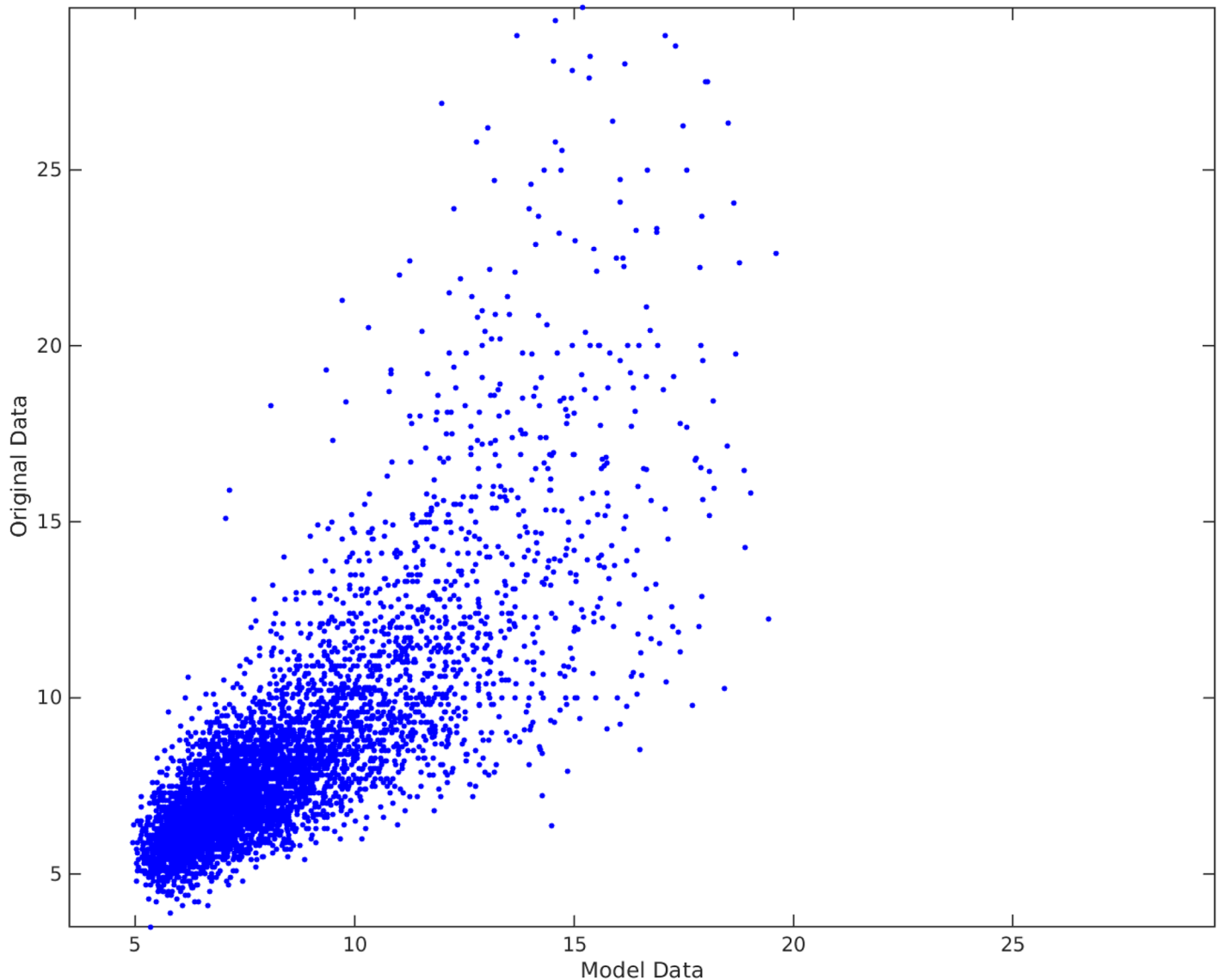


Fig 12. Parameter selection algorithm method applied on PWV data. The horizontal axis shows the model found by the best subset selection method. The vertical axis shows the recorded PWV data. The R^2_{adj} of the model is 0.63052.

<https://doi.org/10.1371/journal.pone.0187676.g012>

Conclusion and future works

In this paper, we have introduced the Parameter Selection Algorithm (Algorithm 1) by which one can simultaneously capture some of the non-linearities of the data into the model, introduce automatic interpretable interaction and transformation among predictions, and also pick the best model. This approach minimizes the efforts done by an analyst and is virtually automatic. So far, up to the best of our knowledge, no other algorithm or method is able to perform these tasks at the same time automatically.

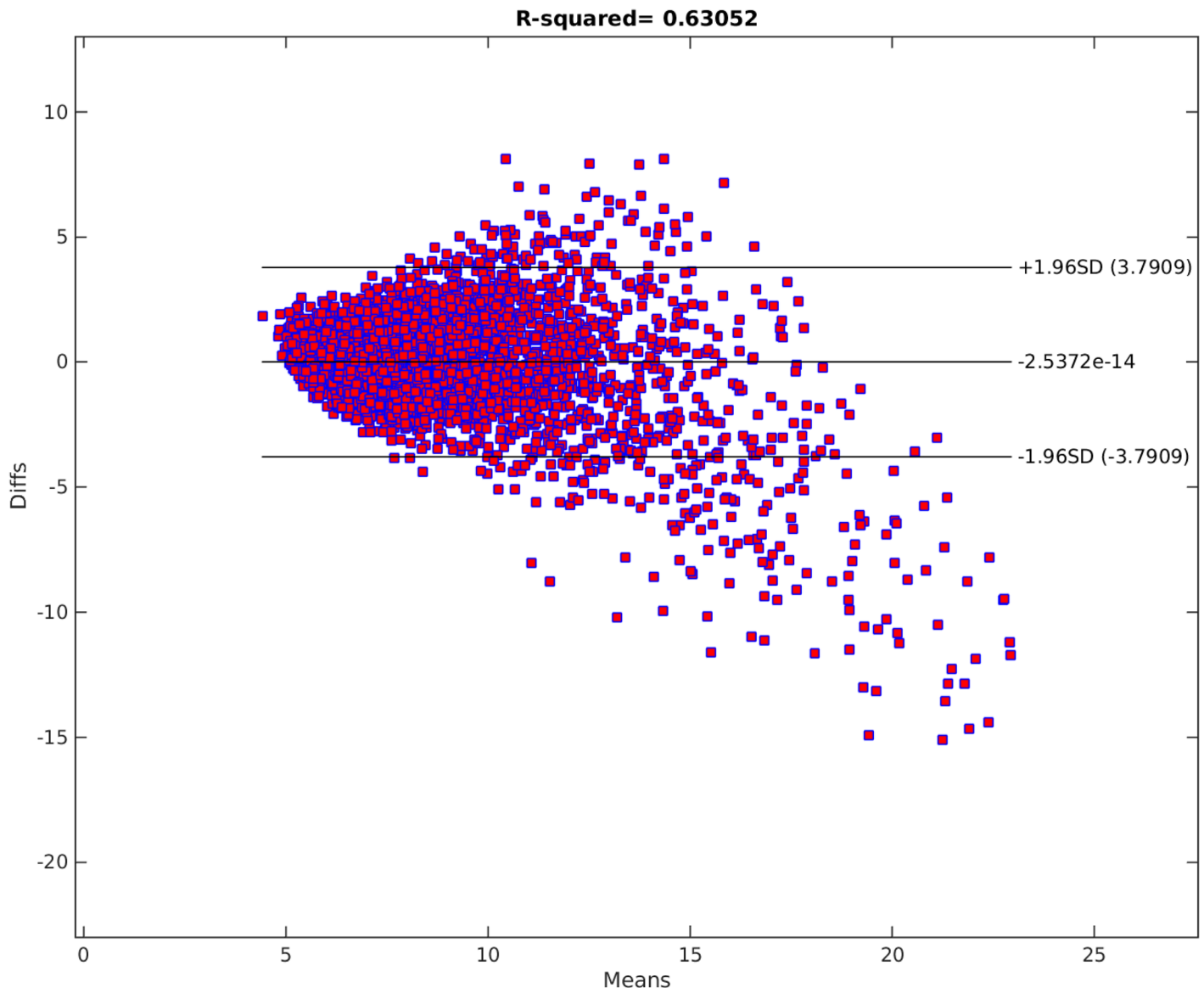


Fig 13. Bland-Altman of the parameter selection algorithm model. The horizontal axis shows the means of the fitted and original PWV values. The Vertical axis shows the differences between the fitted and original PWV values.

<https://doi.org/10.1371/journal.pone.0187676.g013>

Here, our purpose has not been to introduce a competing statistical learning method, but to furnish a data mining tool. Despite this, we have shown that our model is almost as good as the state of the art statistical learning algorithms.

This data mining approach provides an interpretable dimensionality reduction model that faithfully models the data. We believe, the Parameter Selection Algorithm could have versatile applications in biostatistics as shown by one of the examples in this manuscript.

The hyper-parameters ζ and δ , presented in this article, are analyzed and set heuristically. In a future work, we intend to perform a more detailed analysis to possibly quantify optimum values for them. Furthermore, instead of just solving step 5 in Algorithm 1, we could also add

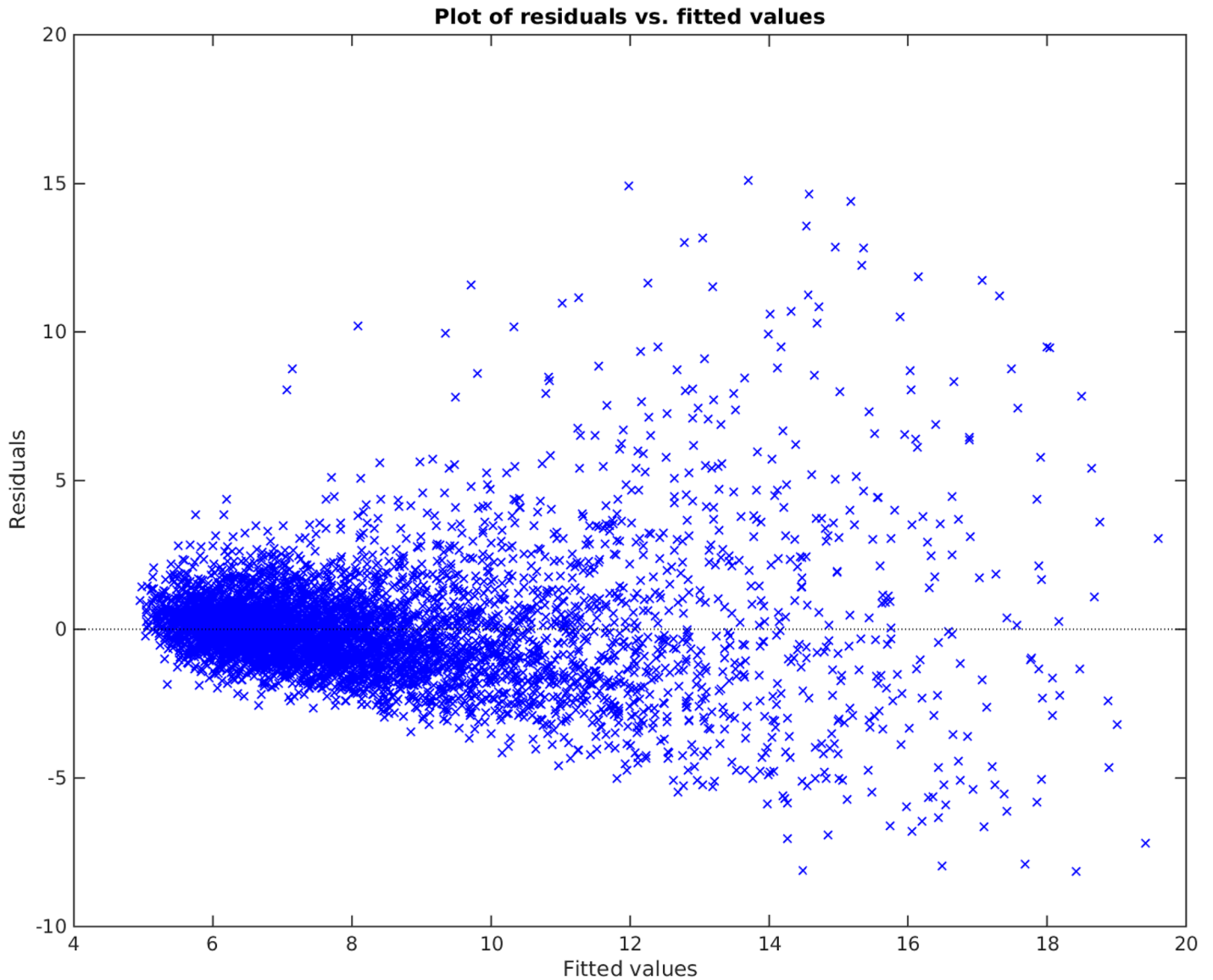


Fig 14. Residual plot of parameter selection algorithm.

<https://doi.org/10.1371/journal.pone.0187676.g014>

the constraint that parameters with high p-values, in a model, should be discarded to provide an even sparser result.

All in all, we see this article as a proof of concept which needs further investigation to analyze the involved hyper-parameters and also tweaks to its optimization core.

Supporting information

S1 Dataset. Example datasets. This file includes all synthetic data examples in this manuscript.

(ZIP)

Acknowledgments

We would like to thank Dr. Niema M. Pahlevan and Prof. Morteza Gharib for giving us the permission to use the Framingham Heart Study data in this paper. The Framingham Heart Study is conducted and supported by the National Heart Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01- HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or conclusions of the Framingham Heart Study or the NHLBI.

Author Contributions

Conceptualization: Peyman Tavallali, Sean Brady.

Formal analysis: Peyman Tavallali, Marianne Razavi.

Investigation: Peyman Tavallali, Marianne Razavi.

Methodology: Peyman Tavallali.

Software: Peyman Tavallali.

Supervision: Peyman Tavallali.

Validation: Peyman Tavallali, Marianne Razavi.

Writing – original draft: Peyman Tavallali.

Writing – review & editing: Peyman Tavallali, Marianne Razavi, Sean Brady.

References

1. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2015.
2. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics Springer, Berlin; 2001.
3. Furnival GM. All possible regressions with less computation. *Technometrics*. 1971; 13(2):403–408. <https://doi.org/10.1080/00401706.1971.10488794>
4. Garside M. The best sub-set in multiple regression analysis. *Applied Statistics*. 1965;p. 196–200.
5. Morgan J, Tatar J. Calculation of the residual sum of squares for all possible regressions. *Technometrics*. 1972; 14(2):317–325. <https://doi.org/10.1080/00401706.1972.10488918>
6. Schatzoff M, Tsao R, Fienberg S. Efficient calculation of all possible regressions. *Technometrics*. 1968; 10(4):769–779. <https://doi.org/10.2307/1267458>
7. Furnival GM, Wilson RW. Regressions by leaps and bounds. *Technometrics*. 2000; 42(1):69–79. <https://doi.org/10.1080/00401706.2000.10485982>
8. Efroymson M. Multiple regression analysis. *Mathematical methods for digital computers*. 1960; 1:191–203.
9. Mallows C. Choosing variables in a linear regression: A graphical aid. In: Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas. vol. 5; 1964.
10. Mallows CL. More comments on Cp. *Technometrics*. 1995; 37(4):362–372. <https://doi.org/10.2307/1269729>
11. Mallows CL. Some comments on Cp. *Technometrics*. 1973; 15(4):661–675. <https://doi.org/10.1080/00401706.1973.10489103>
12. Mallows CL. Choosing a subset regression. In: *TECHNOMETRICS*. vol. 9. AMER STATISTICAL ASSOC 1429 DUKE ST, ALEXANDRIA, VA 22314; 1967. p. 190.
13. Olusegun AM, Dikko HG, Gulumbe SU. Identifying the Limitation of Stepwise Selection for Variable Selection in Regression Analysis. *American Journal of Theoretical and Applied Statistics*. 2015; 4 (5):414–419. <https://doi.org/10.11648/j.ajtas.20150405.22>
14. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970; 12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>

15. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. *Technometrics*. 1970; 12(1):69–82. <https://doi.org/10.1080/00401706.1970.10488634>
16. García C, García J, López Martín M, Salmerón R. Collinearity: Revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*. 2015; 42(3):648–661. <https://doi.org/10.1080/02664763.2014.980789>
17. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*. 1998; 20(1):33–61. <https://doi.org/10.1137/S1064827596304010>
18. Candes EJ, Tao T. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*. 2006; 52(12):5406–5425. <https://doi.org/10.1109/TIT.2006.885507>
19. Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. *The Annals of statistics*. 2004; 32(2):407–499. <https://doi.org/10.1214/009053604000000067>
20. Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*. 2007;p. 2313–2351.
21. Barber RF, Candès EJ, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*. 2015; 43(5):2055–2085. <https://doi.org/10.1214/15-AOS1337>
22. Stone M, Brooks RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society Series B (Methodological)*. 1990;p. 237–269.
23. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *Journal of computational and graphical statistics*. 2006; 15(2):265–286. <https://doi.org/10.1198/106186006X113430>
24. Trefethen LN, Bau III D. *Numerical linear algebra*. vol. 50. Siam; 1997.
25. Box GE, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*. 1964;p. 211–252.
26. Box GE, Tidwell PW. Transformation of the independent variables. *Technometrics*. 1962; 4(4):531–550. <https://doi.org/10.1080/00401706.1962.10490038>
27. MacKay DJ. *Information theory, inference and learning algorithms*. Cambridge university press; 2003.
28. Hinton G, Deng L, Yu D, Dahl GE, Mohamed Ar, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*. 2012; 29(6):82–97. <https://doi.org/10.1109/MSP.2012.2205597>
29. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*. 1970; 12(3):591–612. <https://doi.org/10.1080/00401706.1970.10488699>
30. Cohen J. A power primer. *Psychological bulletin*. 1992; 112(1):155. <https://doi.org/10.1037/0033-2909.112.1.155> PMID: 19565683
31. Association AH, et al. *Heart Disease and Stroke Statistics—At-a-Glance*; 2015.
32. Mitchell GF, Hwang SJ, Vasan RS, Larson MG, Pencina MJ, Hamburg NM, et al. Arterial stiffness and cardiovascular events the Framingham Heart Study. *Circulation*. 2010; 121(4):505–511. <https://doi.org/10.1161/CIRCULATIONAHA.109.886655> PMID: 20083680
33. Mitchell GF, Parise H, Benjamin EJ, Larson MG, Keyes MJ, Vita JA, et al. Changes in arterial stiffness and wave reflection with advancing age in healthy men and women the Framingham Heart Study. *Hypertension*. 2004; 43(6):1239–1245. <https://doi.org/10.1161/01.HYP.0000128420.01881.aa> PMID: 15123572
34. Safar ME, London GM, et al. Therapeutic studies and arterial stiffness in hypertension: recommendations of the European Society of Hypertension. *Journal of hypertension*. 2000; 18(11):1527–1535. <https://doi.org/10.1097/00004872-200018110-00001> PMID: 11081763
35. Framingham Heart Study;. Accessed: 2016-07-14. <https://www.framinghamheartstudy.org/>
36. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, et al. The third generation cohort of the National Heart, Lung, and Blood Institute’s Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology*. 2007; 165(11):1328–1335. <https://doi.org/10.1093/aje/kwm021> PMID: 17372189
37. Kannel WB, Feinleib M, McNAMARA PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families The Framingham offspring study. *American journal of epidemiology*. 1979; 110(3):281–290. <https://doi.org/10.1093/oxfordjournals.aje.a112813> PMID: 474565
38. Dawber TR, Meadors GF, Moore FE Jr. *Epidemiological Approaches to Heart Disease: The Framingham Study**. *American Journal of Public Health and the Nations Health*. 1951; 41(3):279–286. <https://doi.org/10.2105/AJPH.41.3.279>

39. Geerts BF, Aarts LP, Jansen JR. Methods in pharmacology: measurement of cardiac output. *British journal of clinical pharmacology*. 2011; 71(3):316–330. <https://doi.org/10.1111/j.1365-2125.2010.03798.x> PMID: 21284692
40. Greupner J, Zimmermann E, Grohmann A, Dübel HP, Althoff T, Borges AC, et al. Head-to-head comparison of left ventricular function assessment with 64-row computed tomography, biplane left cineventriculography, and both 2-and 3-dimensional transthoracic echocardiography: comparison with magnetic resonance imaging as the reference standard. *Journal of the American College of Cardiology*. 2012; 59(21):1897–1907. PMID: 22595410
41. Matthews D, Hosker J, Rudenski A, Naylor B, Treacher D, Turner R. Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*. 1985; 28(7):412–419. <https://doi.org/10.1007/BF00280883> PMID: 3899825