Volume 54 • Number 5 • 2025
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: www.tandfonline.com/journals/lssp20

A general procedure for evaluating models and ensemble Support Vector Regression

Guoyi Zhang & Yulei He

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2024.2384569

A general procedure for evaluating models and ensemble Support Vector Regression

Guoyi Zhang[a,b] and Yulei He[a]

[a]National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland, USA; [b]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

ABSTRACT

In practice, we may want to discover if there is a relationship between a response variable as a function of the predictor variables. Multiple linear regression (MLR) is a popular tool for such purpose. When the relationship is nonlinear, nonparametric regression methods such as local linear regression, smoothing splines, and support vector regression (SVR) provide flexible alternatives to MLR. How do we compare the performance of these methods and choose an appropriate one for use? In this research, we propose a general procedure to evaluate the performance of different regression methods for use in large data. We also propose an ensemble SVR for regression analysis. The proposed methods are applied to address research questions using the Research and Development Survey, conducted by the National Center for Health Statistics.

ARTICLE HISTORY
Received 5 July 2023
Accepted 21 July 2024

KEYWORDS
Ensemble support vector regression (ESVR); Explained variance; Mean squared test set error; Monte Carlo simulation; Multiple linear regression; Randomized complete block design

1. Introduction

With the advancement of numerous statistical and machine learning regression techniques, such as Multiple Linear Regression (MLR), nonparametric regression, Random Forests (RF) (Breiman 2001), and Support Vector Regression (SVR) (Cortes and Vapnik 1995), it is important to compare the performance of these various methods and to determine the most appropriate one for use. To address this issue, our research aims to provide clarity by posing two critical questions: "How can we effectively compare the performance of these methods and select the most suitable one for use?" and "What methodologies are best suited for nonlinear and nonparametric relationships for a given application?"

Historically, researchers have focused on assessing goodness of fit (GOF) (Fisher 1922) and lack of fit tests for linear models (Utts 1982; Su and Yang 2006). Eubank and Spiegelman (1990) proposed nonparametric regression methodology to test the adequacy of parametric linear models. More research on the application of nonparametric regression and smoothing to lack of fit tests can be found in Hart (1997). In addition to conventional goodness of fit (GOF) and lack of fit tests for model evaluation, the mean squared test set error or so called prediction error (PE) is a widely uesd metric. This approach involves randomly partitioning data into two segments: a training set, which is utilized for model development, and a test set, which serves to assess the models' performance based on PE. However, from a statistical standpoint, relying solely on PE derived from a single test set can lead to biased results and may not provide a comprehensive assessment of model fitness. To mitigate this bias, we propose employing multiple random data splits to generate multiple training and test sets. By calculating the differences in PEs between

CONTACT Guoyi Zhang gzhang123@gmail.com National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782, USA

two models as the average of the multiple PE differences obtained from the same test set, we aim to enhance the reliability and accuracy of model performance comparisons.

Among these regression techniques, SVR stands out as a class of Support Vector Machines (SVM) (Vapnik 1991) specifically designed for regression problems. SVR employs an ϵ-insensitive loss function to penalize data points exceeding a predefined threshold ϵ (Vapnik 1995). Its competitiveness with auto regressive moving average (ARMA) models has been demonstrated by Thissen et al. (2003). SVR shares similarities with other regularization approaches like ridge regression, cubic smoothing splines, and thin plate splines, as they all draw upon the theory of reproducing kernel Hilbert spaces (Wahba 1990).

Ensemble methods, including bagging (Breiman 1996), stacking (Hastie, Tibshirani, and Friedman 2001), and boosting, have proven effective in enhancing predictive performance by combining predictions from multiple models. The concept of ensemble methods has also been applied in nonparametric regression. For instance, Lee (2004) combined smoothing splines with different smoothing parameters to improve estimators. Drawing inspiration from SVR and ensemble methods, this research proposes two modified support vector regression models for regression analysis.

The paper is structured as follows: In Sec. 2, we present the Monte Carlo simulation procedure, hypothesis tests, variance estimation, and multiple comparisons. In Sec. 3, we detail the two proposed modified SVR models. In Sec. 4, we apply these methods to analyze problems using data from the Research and Development Survey (RANDS) conducted by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention. Lastly, Sec. 5 provides concluding remarks.

2. A Proposed evaluation procedure

This section proposes an evaluation procedure for model comparisons. First, we present a general algorithm for multiple random splits. After introducing the PE and explained variance, we extend the binomial test and paired t-test to the comparison procedure and derive asymptotic properties. We then propose methods of variance estimation and multiple comparisons by utilizing a completely randomized block design.

2.1. A general algorithm for model comparisons

The main idea of the algorithm is to randomly choose multiple data splits, therefore creating multiple training sets and test sets. Suppose we are interested in comparing r models. The proposed Monte Carlo simulation algorithm for comparison is as follows:

Algorithm 1: General Monte Carlo simulation procedure

```
Randomly create $K$ splits and save the splits to $K$ different files
for each split do
    read the split from the data file;
    for each method $(1,2,\dots,r)$ do
        fit the model with the training data;
        use the fitted model to do prediction on the test data;
        compute the measures of prediction performance;
    record the results to the output file;
Use statistical tools (binomial test and paired t-test) to analyze the results in the output file;
```

Notes:

1. Two sampling methods can be considered for random splits, simple random sampling (SRS) and stratified sampling (Lohr 2021, chapters 2 and 3). For SRS, all possible splits are equally

likely to be selected. For stratified sampling, a SRS is selected within each predetermined stratum. For example, a SRS is selected from each of the 50 states (and D.C.) in the U.S., where state is defined as a stratum.

2. Algorithm 1 is called repeated learning-testing (RLT) and was first introduced by Breiman et al. (1984). RLT can be considered as an approximation of the leave-p-out method (Shao 1993) which means every possible set of p data points are successively "left out" from the sample and used for validation. Leave-p-out is an exhaustive cross-validation method. Leave-p-out with $p = 1$ is the well-known special case, leave-one-out.
3. Algorithm 1 can be carried out by parallel computing. We can, for example, request $K/10$ processors and let each processor do the computations for 10 splits simultaneously.

2.2. Measurements of the proposed methods

For each of the splits, suppose s% of the observations (the selected size should be rounded to the nearest integer) are randomly selected in the training set, and the rest are in the test set. To measure prediction precision, define PE for the kth test set as

$$\mathrm{PE}_k = \frac{\sum_{i \in \text{test set k}} (y_i - \hat{y}_i)^2}{n * (1 - s/100)},$$

where y_i is the response, $\hat{y}_i$ is the estimator from a model, n is the sample size and $n * (1 - s/100)$ is the size of the test set.

Another commonly used measure of prediction quality is explained variance (EV) defined as the proportion of the variance in the response that is explained by the model. Define $\overline{y_k} = \sum y_i/(n * (1 - s/100))$ for i in test set k, $\mathrm{Var}(y_k) = \sum_{i \in \text{test set k}} (y_i - \bar{y}_k)^2/(n * (1 - s/100))$, and EV for the kth test set as

$$\mathrm{EV}_k = 1 - \frac{\mathrm{PE}_k}{\mathrm{Var}(y_k)}.$$

EV compares a suggested model to the simplest model using mean prediction for the response y (mean model). If $\hat{y}_i = \bar{y}_k$, PE_k is equal to $\mathrm{Var}(y_k)$, so that $\mathrm{EV}_k = 0\%$. Any improvement in prediction performance will lead to a decrease in PE_k, thus increasing EV up to its possible maximum value of 100%. If $\mathrm{EV}_k < 0\%$, the suggested model performs even worse than the mean model. This could be due to a lack of fit, severe over-fitting to the training set, or that the test set is not representative of the data before splitting occurred.

2.3. Hypothesis testing

From a computational view, PE and EV can be used to measure performance of the models. From a statistical view, we may want to know if there are significant differences among the models or not. Let y_i be the response, $\boldsymbol{\beta}$ be a vector of parameters, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{id})$ be the covariates, d be the number of covariates and $f(\mathbf{x}_i)$ be a smooth function. Suppose we are interested in comparing the following two models,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \tag{1}$$

and

$$y_i = f(\mathbf{x}_i) + \epsilon_i \tag{2}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, for $i = 1, 2, ..., n$. Model (1) is a linear model and model (2) is a nonparametric regression model, for example, using SVR.

2.3.1. Binomial test

For each split, build models (1) and (2) based on the training set, and calculate PEs using the test set. Let L be the number of all possible splits for the data. For $i = 1, 2, \ldots, L$, define

$$r_i = \begin{cases} 1 & \text{PE of model (1) is smaller than that of model (2) for split } i \\ 0 & \text{otherwise.} \end{cases}$$

Let p be the proportion of all possible splits where the PE of model (1) is smaller than that of model (2), i.e. $p = \sum_{i=1}^{L} r_i/L$. We now compare the proportion p to $\frac{1}{2}$,

$$H_0 : p \geq 1/2 \text{ versus } H_\alpha : p < 1/2. \tag{3}$$

The hypothesis test can be rewritten as follows:

$$H_0 : p = 1/2 \text{ versus } H_\alpha : p < 1/2. \tag{4}$$

If H_0 is rejected, there is evidence that the linearity assumption is not appropriate. Therefore, model (2) is preferred. This is a different view from testing linearity of the model.

Lemma 1. *Let L be the number of all possible splits for the data, K be the number of random splits selected, and R be the number of random splits where model (1) has a smaller PE than model (2). Then*

$$E\left[\hat{p} = \frac{R}{K}\right] = p \tag{5}$$

Proof. For $i = 1, 2, \ldots, L$, define

$$u_i = \begin{cases} 1 & \text{split } i \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\hat{p} = \frac{R}{K} = \frac{\sum_{i=1}^{K} r_i}{K} = \frac{\sum_{i=1}^{L} u_i * r_i}{K}.$$

Since all the possible training sets are equally likely to be selected, we have $E(u_i) = K/L$, so that

$$E[\hat{p}] = \frac{\sum_{i=1}^{L} E(u_i) * r_i}{K} = \frac{\sum_{i=1}^{L} \frac{K}{L} * r_i}{K} = p.$$

□

Theorem 1. *For $i, j = 1, 2, \ldots, K$, and $i \neq j$, The following statement is true,*

$$P(r_i = 1 | H_0, r_j = 1) = P(r_i = 0 | H_0, r_j = 1) = 0.5.$$

In other words, the outcome variables r_i and r_j are independent and R has a binomial distribution with parameter K and probability of success of 1/2.

Proof. Under the null hypothesis $H_0 : p = 1/2$, r_i and r_j have the same Bernoulli distribution. □

Reject H_0 when $R < R_\alpha^*$, where R_α^* can be found from the cumulative distribution function of the binomial distribution. For example, randomly split the data 100 times, i.e. $K = 100$. At the level of $\alpha = 0.05$, reject H_0 when $R < R_\alpha^* = 42$. The confidence interval of p can be constructed by the normal approximation for binomial distributions. For example, if $\hat{p} = 0.8$, then the margin of error is $1.96 * \sqrt{0.8 * (1 - 0.8)/100}$ and a 95% confidence interval can be found as $[0.7216, 0.8784]$. We may increase the number of random splits K to reduce the margin of error.

2.3.2. Paired t-test

Another test is to compare the PEs and EVs using a paired t-test. Let $PE_i^{(1)}$ and $PE_i^{(2)}$ be the prediction error from model (1) and model (2) for the ith test set, respectively. Define the PE improvement as

$$\delta_i = PE_i^{(1)} - PE_i^{(2)}, \tag{6}$$

which is a measure of the ith split. We will reject H_0 if the confidence interval for the true PE improvement $D = E(\delta_i)$ is positive.

Theorem 2. *Let $\delta_i, i = 1, 2, \ldots, K$ be the measures of differences between the PEs or EVs of the K random splits defined in (6). Let L be the number of all the possible splits, and $D = \frac{1}{L}\sum_{i=1}^{L}\delta_i, i = 1, 2, \ldots, L$. Then*

$$\bar{\delta} = \frac{1}{K}\sum_{i=1}^{K}\delta_i \tag{7}$$

is a design unbiased estimator of D, and the sample variance

$$s^2(\bar{\delta}) = \frac{L-K}{L*K} * \frac{1}{K-1}\sum_{i=1}^{K}(\delta_i - \bar{\delta})^2 \tag{8}$$

is a design unbiased estimator of $Var(\bar{\delta}) = \left(1 - \frac{K}{L}\right) * \frac{S^2}{K}$ where $S^2 = \frac{1}{L-1}\sum_{i=1}^{L}(\delta_i - D)^2, i = 1, 2, \ldots, L$. In other words, $E[\bar{\delta}] = D$ and $E[s^2(\bar{\delta})] = Var(\bar{\delta})$. Furthermore, the sampling distribution of

$$\frac{\bar{\delta} - D}{\sqrt{(1 - K/L)}(S/\sqrt{K})}$$

is approximately normal with mean 0 and variance 1.

Proof. Since all the possible training sets are equally likely to be selected, we have $E(u_i) = K/L$, so that

$$E(\bar{\delta}) = E\left(\frac{1}{K}\sum_{i=1}^{K}\delta_i\right) = E\left(\frac{1}{K}\sum_{i=1}^{L}u_i * \delta_i\right) = \frac{1}{K}\sum_{i=1}^{L}E(u_i) * \delta_i = \frac{1}{K}\sum_{i=1}^{L}\frac{K}{L} * \delta_i = D.$$

To show $E[s^2(\bar{\delta})] = Var(\bar{\delta})$, it suffices to show that $E\left(\frac{1}{K-1}\sum_{i=1}^{K}(\delta_i - \bar{\delta})^2\right) = S^2$.

$$\begin{aligned}
E\left(\frac{1}{K-1}\sum_{i=1}^{K}(\delta_i - \bar{\delta})^2\right) &= \frac{1}{K-1}E\left(\sum_{i=1}^{K}((\delta_i - D) - (\bar{\delta} - D))^2\right) \\
&= \frac{1}{K-1}\left[E\left(\sum_{i=1}^{K}(\delta_i - D)^2\right) - 2 * E(K(\bar{\delta} - D)^2) + E(K(\bar{\delta} - D)^2)\right] \\
&= \frac{1}{K-1}\left[E\left(\sum_{i=1}^{K}(\delta_i - D)^2\right) - E(K(\bar{\delta} - D)^2)\right] \\
&= \frac{1}{K-1}\left[E\left(\sum_{i=1}^{L}u_i * (\delta_i - D)^2\right) - K \cdot \mathrm{Var}(\bar{\delta})\right] \\
&= \frac{1}{K-1}\left[\sum_{i=1}^{L}\frac{K}{L} * (\delta_i - D)^2 - \left(1 - \frac{K}{L}\right) * S^2\right] \\
&= \frac{1}{K-1}\left[\frac{K}{L} * (L-1) * S^2 - \left(1 - \frac{K}{L}\right) * S^2\right] = S^2.
\end{aligned}$$

The asymptotic distribution of $(\bar{\delta} - D)/(\sqrt{(1 - K/L)}(S/\sqrt{K}))$ follows from the central limit theorem for SRS directly (Hajek 1960). □

Hence, a two-sided confidence interval for D is commonly approximated as follows:

$$\left[\bar{\delta} - t_{1-\alpha/2,K-1} * \sqrt{s^2(\bar{\delta})},\ \bar{\delta} + t_{1-\alpha/2,K-1} * \sqrt{s^2(\bar{\delta})}\right], \tag{9}$$

where $t_{1-\alpha/2,K-1}$ is the critical value from the t distribution with significance level of α and with $K-1$ degrees of freedom.

2.4. Variance estimation and multiple comparisons

In a randomized complete block design (RCBD), each treatment is included once in each block. The purpose of blocking is to sort experimental units into groups within each of which the elements are homogeneous with respect to the response variable; differences between groups are as great as possible. In this research, each split is considered as a block, and each model is considered as a treatment. A mixed effect RCBD model is as follows:

$$Y_{ij} = \mu.. + \rho_i + \tau_j + \epsilon_{ij}, i = 1, 2, ..., K, j = 1, 2, ..., r, \tag{10}$$

where i and j are the indices of the splits and models, respectively; ρ_i is the block effect for the K blocks and τ_j is the treatment effect for the r models. The block effect is not of interest and therefore is treated as a random variable with $\rho_i \overset{\text{iid}}{\sim} N(0, \sigma_\rho^2)$. The error term $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is independent from ρ_i, and the treatment effects have the constraint of $\sum_{j=1}^{r} \tau_j = 0$. We can calculate the percentage of variability explained by splits, and perform multiple comparisons among the different models (Kutner et al. 2004, chapter 27.6).

3. Two modified SVR models

In this section, first we present a quick review of SVR. Next, we propose two modified versions of SVR called h-step reduced SVR (RSVR) and h-step ensemble SVR (ESVR).

3.1. A review of SVR

Let $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ be a set of n observations where $\mathbf{x}_i \in R^d$ and $y_i \in R$ for all i. Define $\phi : R^d \rightarrow \mathcal{F}$ be a mapping from R^d to the feature space $\mathcal{F}$, and

$$y = f(\mathbf{x}) = \mathbf{w}^t \phi(\mathbf{x}) + b$$

where $\mathbf{w}$ is a vector of coefficients in the feature space $\mathcal{F}$ and b is the bias. SVR attempts to find an optimal set of parameters $\mathbf{w}$ and b in order to ensure the flatness/smoothness property, which leads to the following optimization criterion (Alpaydin 2014)

$$\underset{w,b}{\text{MIN}}\ \frac{1}{2}||\mathbf{w}||^2, \text{ subject to } -\epsilon \le y_i - \mathbf{w}^t \phi(\mathbf{x}) - b \le \epsilon.$$

Next the slack variables $\zeta > 0$ and $\zeta^* > 0$ are introduced to penalize points from the ϵ-insensitive band, where ζ and ζ^* correspond to upper and lower deviations, respectively. The parameter C is introduced to control the tradeoff between the empirical risk and regularization terms or the tradeoff between the complexity of the function and magnitude of deviations from the ϵ-insensitive band,

$$\underset{\mathbf{w},b}{\text{MIN}}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\zeta + \zeta^*), \text{ subject to } -\epsilon - \zeta^* \le y_i - \mathbf{w}^t \phi(\mathbf{x}) - b \le \epsilon + \zeta$$

Optimal values of C and ϵ are found by cross-validation. The name of support vector machines comes from the fact that the solution depends only on a subset of observations, which are called support vectors. In the case of SVR, support vectors are observations on the boundary and outside of the ϵ-insensitive band. Increasing ϵ means a larger insensitive band with less support vectors in general. Using the Lagrangian and corresponding optimal conditions, the solution of SVR has the following form (Alpaydin 2014)

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + \hat{b}$$

where α_i and α_i^* are nonzero Lagrange multipliers and $K(.,.)$ is the kernel function. In this research, the following radial basis function is used as kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2)$$

where γ is the parameter of the kernel function, and its optimal value is found by cross-validation.

3.2. h-step reduced SVR (RSVR) and h-step ensemble SVR (ESVR)

The idea of RSVR originates from using a reduced model to alleviate over-fitting and to improve the model prediction ability. Given a set of predictors $\mathbf{x} \in R^d$, define an h-step sub-covariate of $\mathbf{x}$ as $\mathbf{z} = (z_1, z_2, ..., z_{(d-h)})$ for $h = 1, 2, ..., d-1$. For example, the covariates corresponding to 1-step ($h = 1$) are $\mathbf{z} = (z_1, z_2, ..., z_{(d-1)})$, and there are d combinations of the $d-1$ predictors. Let E_m be the set of all the possible subsets of predictors for the mth stage reduced models for $m = 1, 2, ..., h$, and let M_m be the total number of all the possible subsets at the mth stage. For example, for 2-step ($h = 2$), $E_m = E_1$ or E_2. There are d possible sub-covariates in E_1 corresponding to $h = 1$, and d choose $(d-2)$, i.e. $d * (d-1)/2$ combinations of the $d-2$ predictors in E_2. Therefore, $M_2 = d + d * (d-1)/2$. Using cross-validation, the best reduced model can be found for each m, and for all the possible subsets. The algorithm for the proposed h-step reduced SVR (RSVR) procedure is as follows:

Algorithm 2: h-step reduced SVR procedure (RSVR)

```
for each model $m$ do
   for each subset in $E_m$ do
      fit SVR with the training data;
      use 10 fold cross-validation (CV) to find optimal parameter estimates ($C$, $\epsilon$ and $\gamma$);
      record the performance result (from CV) and optimal parameter estimates and the corresponding subset;
Find the best-performing model based on the CV out of all the models; Use the best-performing model to do prediction on the testing data.
```

The other proposed method is h-step ensemble SVR (ESVR) which combines all the reduced models in E_m for $m = 1, 2, ..., h$. In general, ensemble methods produce more accurate predictions than a single model would and also reduce the spread or dispersion of the predictions (Hastie, Tibshirani, and Friedman 2001). The proposed h-step ensemble SVR (ESVR) procedure is as follows:

Algorithm 3: h-step ensemble SVR procedure (ESVR)

```
for each model $m$ do
   for each subset in $E_m$ do
      fit SVR with the training data;
      use 10 fold cross-validation to find optimal parameter estimates ($C$, $\epsilon$, and $\gamma$);
      use the model with optimal parameter estimates to do prediction on the testing data;
      record the result;
average all the results.
```

4. Simulations

We did a small simulation study to investigate the power of the binomial test. Throughout the paper, we used a significance level of 0.05. The following function from Hastie, Tibshirani, and Friedman (2001) was used:

$$y_i = 1/2 * (1 + x_i)^3 + \epsilon_i, \tag{11}$$

where $i = 1, 2, ..., n$, ϵ_i's are with zero mean and constant variance σ^2, and x_i's are from a uniform distribution with domain of $[-1, 1]$. Equation (11) has three segments that cannot be closely fit by linear or quadratic functions. The following four factors were considered in the experimental design: (1) The total sample size ($n = 2000$ or 20,000), (2) The noise level ($\sigma = 0.0625$ or 0.2500), (3) Split ratio, i.e. the fraction of the size of the training set (s% = 50% or 80%) and it's complement, the test set, and (4) The number of splits ($K = 100$ or 1000). The simulation results, the power of the binomial tests for the various combinations, are in Table 1. The fit of a cubic model is compared to that of a fourth-order power fit. The power of the test represents the probability of finding evidence that the cubic model is better than the fourth-order power fit. This is a standard four factorial design and is fit using an analysis of variance (ANOVA) model. The results are shown in Table 2. The power of the binomial test (two-sided) is not significantly associated with the total sample size and the noise level. However, the number of splits and the split ratio have a significant association with the power of the binomial test. Comparing the 4th order power fit with the cubic power fit (the true model), there is an 8.5% reduction in power when comparing an 80% split ratio to a 50% split ratio. The larger the size of testing set, the larger the power of the test. Compared with 100 splits, the power from using 1000 splits has about 23.2% increase. The larger the number of splits, the larger the power of the test. Users are recommended to use a minimum of 100 splits when applying this approach to achieve acceptable power. The split ratio (s%) and number of splits (K) both have effects on the variation of the binomial test, hence the power of the test. It is interesting to see that adding noise has no significant effect.

Table 1. Simulation results for the binomial test.

Sample size	Noise level(σ)	Split ratio	Number of splits(K)	Power(quart/cubic)
2000	0.25	0.5	100	0.758
2000	0.25	0.5	1000	0.92
2000	0.25	0.8	100	0.607
2000	0.25	0.8	1000	0.893
2000	0.0625	0.5	100	0.727
2000	0.0625	0.5	1000	0.915
2000	0.0625	0.8	100	0.6
2000	0.0625	0.8	1000	0.872
20000	0.25	0.5	100	0.736
20000	0.25	0.5	1000	0.917
20000	0.25	0.8	100	0.588
20000	0.25	0.8	1000	0.873
20000	0.0625	0.5	100	0.726
20000	0.0625	0.5	1000	0.926
20000	0.0625	0.8	100	0.614
20000	0.0625	0.8	1000	0.898

Table 2. Four factorial ANOVA model for the binomial test.

	Estimate	std.Error	t value	*p*-value
intercept	0.712	0.018	39.531	0.0000
sample size	−0.0017	0.016	−0.109	0.9154
split ratio	−0.085	0.016	−5.276	0.0003
noise level(σ)	0.0017	0.016	0.109	0.9154
number of splits(K)	0.2322	0.016	14.417	0.0000

5. Applications

5.1. Data description and preparation

The Research and Development Survey (RANDS) (https://www.cdc.gov/nchs/rands/) consists of a series of primarily recruited probability-sampled commercial panel surveys. RANDS started in 2015 (ongoing) and is conducted by the National Center for Health Statistics (NCHS) part of the Centers for Disease Control and Prevention (CDC); external vendors are used for data collection. These surveys largely utilize the web mode for data collection. Each RANDS survey has its own focus with a general interest on a range of health-related topics including chronic conditions, healthcare access and utilization, opioid use, and COVID-19 (He et al. 2020; Irimata and Scanlon 2022).

The public-use version of RANDS 1 (https://www.cdc.gov/nchs/rands/data.htm) is used to illustrate the proposed research. RANDS 1 data was collected during the fall of 2015 by Gallup, Inc. using the web mode. It had 2,304 completed interviews with a completion rate of 24%. Questionnaire topics included demographics, healthcare access and utilization, chronic conditions, food security, general health, health insurance, physical activity, psychological distress, and alcohol/tobacco use.

For the purpose of demonstration, body mass index (BMI) is a continuous dependent (outcome) variable. The independent variables (features) include:

a. Demographic variables (e.g., age, sex, race/ethnicity, region, education, income, marital status, employment status, housing status)
b. Health conditions (e.g., chronic conditions such as diabetes, hypertension, asthma, emphysema/chronic bronchitis or chronic obstructive pulmonary disease (COPD), taking medications for some conditions, and self-rated health status)
c. Access to healthcare (e.g., health insurance coverage, delayed or can't afford healthcare, using the internet for health information and appointments)
d. Health behaviors (e.g., tobacco use, alcohol use, physical activity)
e. Psychological distress variables (e.g., How often feeling worried, nervous, or anxious.)
f. Sampling design variables (e.g., survey weight).

In summary, 68 survey questions are included as independent variables. The categorical variables are coded by 120 indicator variables. In the training data, we set four outliers/extreme values in the dependent variable (e.g., $BMI < 10$ or $BMI > 60$) to missing. We imputed the missing values by model-based multiple imputation using the R mice package, and randomly selected one imputed data set for model training and validation. We also observed that some continuous independent variables were highly skewed. Therefore, we applied transformations and standardization to address the non-normal problem. The survey weight was treated as an independent variable.

5.2. Results

The R package e1071 was used for SVR, and all other statistical modeling and computation were done by parallel computing in R.

We used simple random sampling to generate $K = 100$ random splits, and set up $s\% = 80\%$ or 90% observations as the training set and 20% or 10% as the test set for each split, respectively. An additive linear model with all predictors was defined as the full model, a linear model selected by the Akaike information criterion (AIC) was defined as the reduced model, and SVR based on AIC selected features was defined as the SVR model. For RSVR and ESVR, we consider a 1-step ($h = 1$) model to illustrate the proposed methods.

Table 3. The number of splits (out of 100) where a model (column) had a smaller PE than the reference (row) for the binomial test (80% split ratio).

	Full model	Reduced model	SVR	RSVR	ESVR
Full model	X	95	92	92	97
Reduced model	X	X	70	67	81
SVR	X	X	X	31	84
RSVR	X	X	X	X	82
ESVR	X	X	X	X	X

Table 4. The mean and standard deviation of the PE improvements, δ_i, paired t-test (80% split ratio).

	Full model	Reduced model	SVR	RSVR	ESVR
Full model	X	0.7302(0.4453)	0.9883(0.7018)	0.9132(0.7015)	1.0938(0.6913)
Reduced model	X	X	0.2581(0.5102)	0.1829(0.5340)	0.3636(0.4958)
SVR	X	X	X	−0.0751(0.2264)	0.1055(0.1336)
RSVR	X	X	X	X	0.1806(0.2117)
ESVR	X	X	X	X	X

First, for split ratio 80%, Table 3 presents the number of times where a model (column) had a smaller PE than the reference (row) for the binomial test. For a 95% confidence level, counts greater than 58 provide evidence that one model (column) has a lower PE than the other (row). The reduced model, SVR, RSVR and ESVR significantly outperform the full model. RSVR and SVR both significantly outperform the reduced model. SVR and RSVR have smaller PEs than the reduced model in approximately 70% of the splits. ESVR has significant lower PEs than all the other models.

Table 4 presents the mean (equation (7)) and standard deviation (equation (8)) of the PE improvements δ_i. It can be shown that the margin of errors $1.96 * SE * \sqrt{1/100}$ are all smaller than the means. Therefore, all the confidence intervals of the tests calculated by equation (9) do not include zero, i.e. all paired t-tests are significant. In other words, the reduced model, SVR, RSVR and ESVR all perform significantly better than the full model. SVR, RSVR and ESVR perform significantly better than the reduced model. Furthermore, ESVR performs significantly better than SVR and RSVR.

Using a RCBD mixed model (equation (10)), we find the random splits variance (σ_ρ^2) is 4.4193 with a standard deviation of 2.1022. The within block variance (σ^2) is 0.1283 with a standard deviation of 0.3582. Random splits account for about $97.18\%(=4.4193/(4.4193 + 0.1283))$ of the total variability, which indicates the proposed multiple splits procedure is successful. Multiple comparisons also show that SVR performs better than the reduced model, and the reduced model performs better than the full model.

Tables 5 and 6 show a similar pattern under a split ratio of 90% for both tests. However, the variation among the testing sets increases as the size of the test set goes down. Comparing the standard errors in Table 4 to those in Table 6, the standard errors increase for all cells. Using a RCBD mixed model, we find the random splits variance is 10.2199 with standard deviation 3.1969 and the within block variance is 0.2338 with standard deviation 0.4836. The variance among the random splits increases significantly. Random splits account for about $97.76\%(=10.2199/(10.2199 + 0.2338))$ of the total variability, which further indicates the importance of multiple splits instead of one split.

6. Conclusion

In conclusion, this research has made contribution in establishing a comprehensive framework for comparing various regression methods by leveraging the concept of multiple splits. The

Table 5. The number of splits (out of 100) where a model (column) had a smaller PE than the reference (row) for the binomial test (90% split ratio).

	Full model	Reduced model	SVR	RSVR	ESVR
Full model	X	87	87	82	87
Reduced model	X	X	61	60	62
SVR	X	X	X	47	76
RSVR	X	X	X	X	65
ESVR	X	X	X	X	X

Table 6. The mean and standard deviation of the PE improvements, δ_i, paired t-test (90% split ratio).

	Full model	Reduced model	SVR	RSVR	ESVR
Full model	X	0.6722(0.6414)	0.8776(0.9355)	0.8527(0.9571)	0.9711(0.9258)
Reduced model	X	X	0.2054(0.6867)	0.1805(0.7011)	0.2989(0.6581)
SVR	X	X	X	−0.0248(0.3291)	0.0934(0.1689)
RSVR	X	X	X	X	0.1183(0.2887)
ESVR	X	X	X	X	X

adoption of binomial tests and paired t-tests within this framework enables a robust and systematic evaluation of different methodologies.

Two modified versions of SVR, RSVR, and ESVR, were proposed and compared with SVR. The ESVR method emerged as the top performer among the compared techniques. The remarkable performance of ESVR can be attributed to its integration of the ensemble idea, which enhances its adaptability and accuracy in handling test sets, particularly when dealing with noisy data and nonlinear relationships.

In the RANDS example, the analysis of variance from the RCBD model revealed that the variation resulting from random splits accounts for approximately 97% of the total variation. This finding underscores the success of the proposed multiple splits procedure in effectively mitigating potential biases and ensuring a reliable comparison of regression methods.

In light of these findings, ESVR emerges as a promising candidate for regression analysis, especially in scenarios where data are noisy and the relationship between variables is nonlinear. The robust general framework presented in this research not only contributes to the advancement of comparative methodological studies but also offers valuable insights for practitioners seeking to select the most appropriate regression technique for their specific applications.

The versatility of the general framework developed in this study extends beyond regression problems. It can be readily applied to compare and evaluate models in other domains, showcasing its broad applicability and potential impact across various disciplines.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by Centers for Disease Control and Prevention Fellowship.

References

Alpaydin, E. 2014. *Introduction to machine learning*. 3rd ed. Cambridge, MA: MIT Press.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24 (2):123–40. doi: 10.1007/BF00058655.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi: 10.1023/A:1010933404324.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth Advanced Books and Software.

Cortes, C., and V. N. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3):273–97. doi: 10.1007/BF00994018.

Eubank, R. L., and C. H. Spiegelman. 1990. Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association* 85 (410):387–92. doi: 10.1080/01621459.1990.10476211.

Fisher, R. A. S. 1922. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society* 85 (4):597. doi: 10.2307/2341124.

Hajek, J. 1960. Limiting distributions in simple random sampling from a finite population. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* 3:361–74.

Hart, J. 1997. *Nonparametric smoothing and lack-of-fit tests*. New York: Springer.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning. *Springer Series in Statistics*. New York: Springer.

He, Y., B. Cai, H.-C. Shin, V. Beresovsky, V. Parsons, K. Irimata, P. Scanlon, and J. Parker. 2020. The national center for health statistics' 2015 and 2016 research and development surveys. *Vital and Health Statistics. Ser. 1, Programs and Collection Procedures* 59:1–60.

Irimata, K., and P. Scanlon. 2022. The research and development survey (rands) during covid-19. *Statistical Journal of the IAOS* 38 (1):13–21. doi: 10.3233/SJI-210880.

Kutner, M., C. Nachtsheim, J. Neter, and W. Li. 2004. Applied linear statistical models. 5th ed. New York: McGraw-Hill/Irwin.

Lee, T. C. 2004. Improved smoothing spline regression by combining estimates of different smoothness. *Statistics & Probability Letters* 67 (2):133–40. doi: 10.1016/j.spl.2004.01.003.

Lohr, S. 2021. *Sampling: Design and analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Shao, J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* 88 (422):486–94. doi: 10.1080/01621459.1993.10476299.

Su, Z., and S.-S. Yang. 2006. A note on lack-of-fit tests for linear models without replication. *Journal of the American Statistical Association* 101 (473):205–10. doi: 10.1198/016214505000000709.

Thissen, U., R. van Brakel, A. P. de Weijer, W. J. Melssen, and L. M. C. Buydens. 2003. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems* 69 (1–2):35–49. doi: 10.1016/S0169-7439(03)00111-4.

Utts, J. M. 1982. The rainbow test for lack of fit in regression. *Communications in Statistics - Theory and Methods* 11 (24):2801–15. doi: 10.1080/03610928208828423.

Vapnik, V. 1991. *Principles of risk minimization for learning theory*, eds. J. Moody, S. Hanson, and R. Lippmann, vol. 4. Denver, CO: Morgan-Kaufmann.

Vapnik, V. 1995. *The nature of statistical learning theory*. Berlin, Heidelberg: Springer-Verlag.

Wahba, G. 1990. *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.