Dhaka Univ. J. Sci. 60(2): 141-145, 2012 (July)

Forward Selection Procedure for Linear Model Building Using Spearman's Rank Correlation

Md Siddiqur Rahman[1] and Jafar A. Khan[2]

[1]*Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh*

[2]*Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1000, Bangladesh*

Received on 25. 09. 2010. Accepted for Publication on 23. 01. 2012.

Abstract

Forward selection (FS) is a step-by-step model-building algorithm for linear regression. The FS algorithm was expressed in terms of sample correlations where Pearson's product-moment correlation was used. The FS yields poor results when the data contain contaminations. In this article, we propose the use of Spearman's rank correlation in FS. The proposed method is called FSr. We conduct an extensive simulation study to compare the performance of FSr with FS. The proposed FSr performs better than the FS algorithm in the contaminated data. We also demonstrate a real data application of FSr.

Key words: Forward selection; Product-moment correlation; Spearman's rank correlation.

I. Introduction

When the number d of candidate covariates is small, one can choose a linear prediction model by computing a reasonable criterion (e.g., Mallows C_P, AIC, FPE or cross-validation error) for all possible subsets of the predictors. However, as d increases, the computational burden of this approach (sometimes referred to as all possible subsets regression) increases very quickly. Typically, when we have a large collection of possible covariates, we hope to select a parsimonious set from the large collection for the efficient prediction of a response variable. This is one of the reasons why step-by-step model-building algorithms like Forward selection ("FS") or Stepwise ("SW") (Furnival and Wilson 1974, Weisberg 1985, Gatu and Kontoghiorghes 2006, Huo and Ni 2007, and Das and Kempe 2008) is popular. Khan *et al.* (2007) expressed the FS algorithm in terms of sample correlations. They used Pearson's product-moment correlation (r) in FS in variable selection. Since Spearman's rank correlation (ρ) is a standard estimate of association that is invariant to monotone transformation of the data, we propose the use of Spearman's ρ in FS. We replace the Pearson's product-moment correlations in FS with Spearman's ρ, and call the proposed algorithm FSr.

It should be emphasized that with our approach we consider the problem of "selecting" a list of important predictors, but we do not yet "fit" the selected model. The final model resulting from the selection procedure usually contains only a small number of predictors compared to the initial dimension d, when d is large. Note that we always use models with intercepts.

The rest of the article is organized as follows. In section II, we review the FS procedure and present FSr. In section III, we present the results of a simulation study comparing the performance of FSr and FS. Section IV contains a real-data application. Section V contains the limitations of FSr, and section VI is the conclusion.

II. Forward selection (FS) and the use of Spearman's ρ in FS

Let $X_1, X_2, \cdots, X_d$ be *n* dimensional vectors representing the covariates, and Y the *n*-dimensional vector representing the response. By location and scale transformations we can always assume that the variables have been standardized to have mean zero and unit length. The FS procedure selects the covariate (X_1, say) that has the maximum absolute correlation $|r_{1y}|$ with Y, and calculates the residual vector $Y - r_{1y}X_1$. All other covariates are then 'adjusted for X_1' and entered into competition. That is, each X_j is regressed on X_1, and the corresponding residual vector $Z_{j.1}$ (which is orthogonal to X_1) is obtained. The correlations of these $Z_{j.1}$ with the residual vector $Y - r_{1y}X_1$, called partial correlations between X_j and Y adjusted for X_1, decide the next variable to enter the regression model, and so on. We need $(d-1)$ steps to get the ordering of all d predictors.

Let r_{jy} denote the correlation between X_j and Y, and R_X the correlation matrix of the covariates $X_1, X_2, \cdots, X_d$. Suppose without loss of generality that X_1 has the maximum absolute correlation with Y. Then

X_1 is the first entered variable in the regression model. The predictors in the current regression model are *active* predictors. The remaining candidate predictors are *inactive* predictors. The partial correlations between X_j $(j \neq 1)$ and Y adjusted for X_1 are denoted by $r_{jy.1}$. The second covariate X_2 (say) that enters the regression model is then the covariate that has the maximum absolute partial correlation $r_{jy.1}$ with Y.

Each inactive covariate X_j should be regressed on X_1 to obtain the residual vector $Z_{j.1}$ as follows

$$Z_{j.1} = X_j - \beta_{j1} X_1, \quad (1)$$

where

$$\beta_{j1} = \frac{1}{n} X_1^t X_j = r_{j1}. \quad (2)$$

Moreover,

$$\frac{1}{n} Z_{j.1}^t Y = \frac{1}{n} (X_j - \beta_{j1} X_1)^t Y = r_{jy} - r_{j1} r_{1y}, \quad (3)$$

and

$$\frac{1}{n} Z_{j.1}^t Z_{j.1} = \frac{1}{n} (X_j - \beta_{j1} X_1)^t (X_j - \beta_{j1} X_1) = 1 - r_{j1}^2. \quad (4)$$

The partial correlation $r_{jy.1}$ is given by

$$r_{jy.1} = \frac{Z_{j.1}^t (Y - \beta_{y1} X_1)/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}\, SD(Y - \beta_{y1} X_1)}. \quad (5)$$

Note that the factor $SD(Y - \beta_{y1} X_1)$ in the denominator of (5) is independent of the covariates X_j $(j = 2, 3, \cdots, d)$ being considered. Hence, when selecting the covariate X_j that maximizes the partial correlation $r_{jy.1}$, this constant factor can be ignored. This reduces computations and therefore is more efficient. It thus suffices to calculate

$$\tilde{r}_{jy.1} = \frac{Z_{j.1}^t (Y - \beta_{y1} X_1)/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}}, \quad (6)$$

where $\tilde{r}_{jy.1}$ is proportional to the actual partial correlation. Since $Z_{j.1}$ and X_1 are orthogonal and by using (3) and (4), $\tilde{r}_{jy.1}$ can be written as follows

$$\tilde{r}_{jy.1} = \frac{Z_{j.1}^t Y/n}{\sqrt{Z_{j.1}^t Z_{j.1}/n}} = \frac{r_{jy} - r_{j1} r_{1y}}{\sqrt{1 - r_{j1}^2}}. \quad (7)$$

Now, suppose without loss of generality that X_2 (or, equivalently $Z_{2.1}$) is the new active covariate, because it maximizes $\tilde{r}_{jy.1}$ (and thus also the partial correlation $r_{jy.1}$). All the inactive covariates should now be orthogonalized with respect to $Z_{2.1}$.

Orthogonalization of $Z_{j.1}$ with respect to $Z_{2.1}$: Each inactive vector $Z_{j.1}$ should be regressed on $Z_{2.1}$ to obtain the residual vector $Z_{j.12}$ as follows

$$Z_{j.12} = Z_{j.1} - \beta_{j2.1} Z_{2.1}.$$

Here,

$$\beta_{j2.1} = \frac{Z_{2.1}^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n}$$

$$= \frac{X_2^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n} \text{ [because of orthogonality]} \quad (8)$$

$$= \frac{X_2^t (X_j - r_{j1} X_1)/n}{Z_{2.1}^t Z_{2.1}/n} \text{ [Using (1) and (2)}$$

$$= \frac{r_{2j} - r_{21} r_{j1}}{1 - r_{21}^2} \text{ [Using (squared) denominator of (7) for } j = 2].$$

Thus, $\tilde{r}_{jy.1}$ and $\beta_{j2.1}$ are expressed in terms of original correlations. By mathematical induction, $\tilde{r}_{jy.1\cdots k}$ can be expressed as

$$\tilde{r}_{jy\cdots k} = \frac{Z_{j.1\cdots k}^t Y/n}{\sqrt{Z_{j.1\cdots k}^t Z_{j.1\cdots k}/n}},$$

for $k = 2, 3, \cdots, (d-1)$; j inactive,

and

$$\beta_{jh.1\cdots(h-1)} = \frac{Z_{h.1\cdots(h-1)}^t Z_{j.1\cdots(h-1)}/n}{Z_{h.1\cdots.(h-1)}^t Z_{h.1\cdots(h-1)}/n},$$

for $h = 2, 3, \cdots, k$; j inactive,

where,

$$Z_{j.1\cdots k} = Z_{j.1\cdots(k-1)} - \beta_{jk.1\cdots(k-1)} Z_{k.1\cdots(k-1)}.$$

Now, FS algorithm is summarized as follows (Khan *et al.* 2007):

1. To select the first covariate X_{m1}, determine $m_1 = \text{argmax} \left| r_j \right|$.

2. To select the kth covariate $k = 2, 3, \cdots)$, calculate $\tilde{r}_{jy.m_1 \cdots m_{(k-1)}}$, which is proportional to the partial correlation between X_j and Y adjusted for $X_{m1} \cdots, X_{m(k-1)}$ and then determine $m_k = \arg\max |\tilde{r}_{jy.m_1 \cdots m_{(k-1)}}|$.

The FSr algorithm: The Spearman's rank correlation ρ is just the Pearson's product moment correlation r applied to the rank ordered data. Since Spearman's ρ is a more reliable estimate of association in the presence of contaminations in the data, we propose to use this in FS. That is, we rank the values of each covariate, and then consider these ranks as the original values to apply the FS algorithm.

III. Simulations

To investigate the behavior of our FSr proposals, we consider a simulation setting similar to that used by Frank and Friedman (1993). We first create a linear model,

$$Y = L_1 + L_2 + \cdots + L_k + \sigma\varepsilon_i, \qquad \text{(i)}$$

with k latent variables, where $L_1, L_2, \cdots, L_k$ and ε are independent standard normal variables. The value of σ is chosen so that the single-to-noise ratio is 3. A set of d candidate predictors is created as follows. Let $e_1, e_2, \cdots, e_d$ be independent standard normal variables and let

$$X_i = L_i + \tau\, e_i, \quad i = 1, 2, \cdots, k,$$

$$X_{k+1} = L_1 + \delta\, e_{k+1},$$

$$X_{k+2} = L_1 + \delta\, e_{k+2},$$

$$X_{k+3} = L_2 + \delta\, e_{k+3},$$

$$\vdots$$

$$X_{3k-1} = L_k + \delta\, e_{3k-1},$$

$$X_{3k} = L_k + \delta\, e_{3k},$$

and

$$X_i = e_i, \quad i = 3k+1, 3k+2, \cdots, d.$$

The constants $\delta = \sqrt{5}$ and $\tau = \sqrt{0.5}$ are chosen so that $\text{corr}(X_1, X_{k+1}) = \text{corr}(X_1, X_{k+2}) =$

$$\text{corr}(X_2, X_{k+3}) = \cdots = \text{corr}(X_k, X_{3k}) = \frac{1}{3}.$$

Note that covariates $X_1, X_2, \cdots, X_k$ are low noise perturbations of the latent variables and constitute our target covariates. Variables $X_{3k+1}, X_{3k+2}, \cdots, X_d$ are independent noise covariates and variables $X_{k+1}, X_{k+2}, \cdots, X_{3k}$ are noise covariates that are correlated with the target covariates.

To allow for a fraction $\in$ of outliers, we consider the following sampling distributions, listed in increasing order of difficulty:

(1) $\varepsilon \sim N(0,1)$, no contamination

(2) $\varepsilon \sim (1-\in)N(0,1) + \in N(0,1)/uniform(0,\ 1)$, symmetric, slash contamination

(3) Same as (2), except that contaminated cases come along with high leverage X-values

(normal random variables with mean 50 and variance 1 in our simulation)

(4) $\varepsilon \sim (1-\in)N(0,1) + \in N(50,1)$, asymmetric, shifted normal contamination.

To compare our FSr with the FS, we generate 1000 independent samples of size $n = 150$ from the four simulation designs just described, with $k = 10$ latent variables and $d = 50$ candidate covariates. For each simulated data set, we sequence the variables using FSr and FS.

To summarize the simulation results, for each sequence we determine the number t_m of target variables included in the first m sequenced variables, with m ranging between 1 and 20. Fig. 1 shows the average (over 1000 data sets) of t_m for each of the methods and sampling situations. We display here the results for the case where $\in = .15$.

From Fig. 1(a), we see that the two procedures perform well in the uncontaminated case. But the performance of FSr is slightly better than the FS. Figs. 1(b)–1(d) show that, as expected, the performance of FS deteriorates considerably under contamination, but the FSr procedure is much less affected by contamination. In the design with high leverage but asymmetric, shifted normal contamination, FSr shows slightly better performance than FS [Fig. 1(d)]. Generally, all the figures show that FSr procedure is much less affected than FS in the contaminated data.

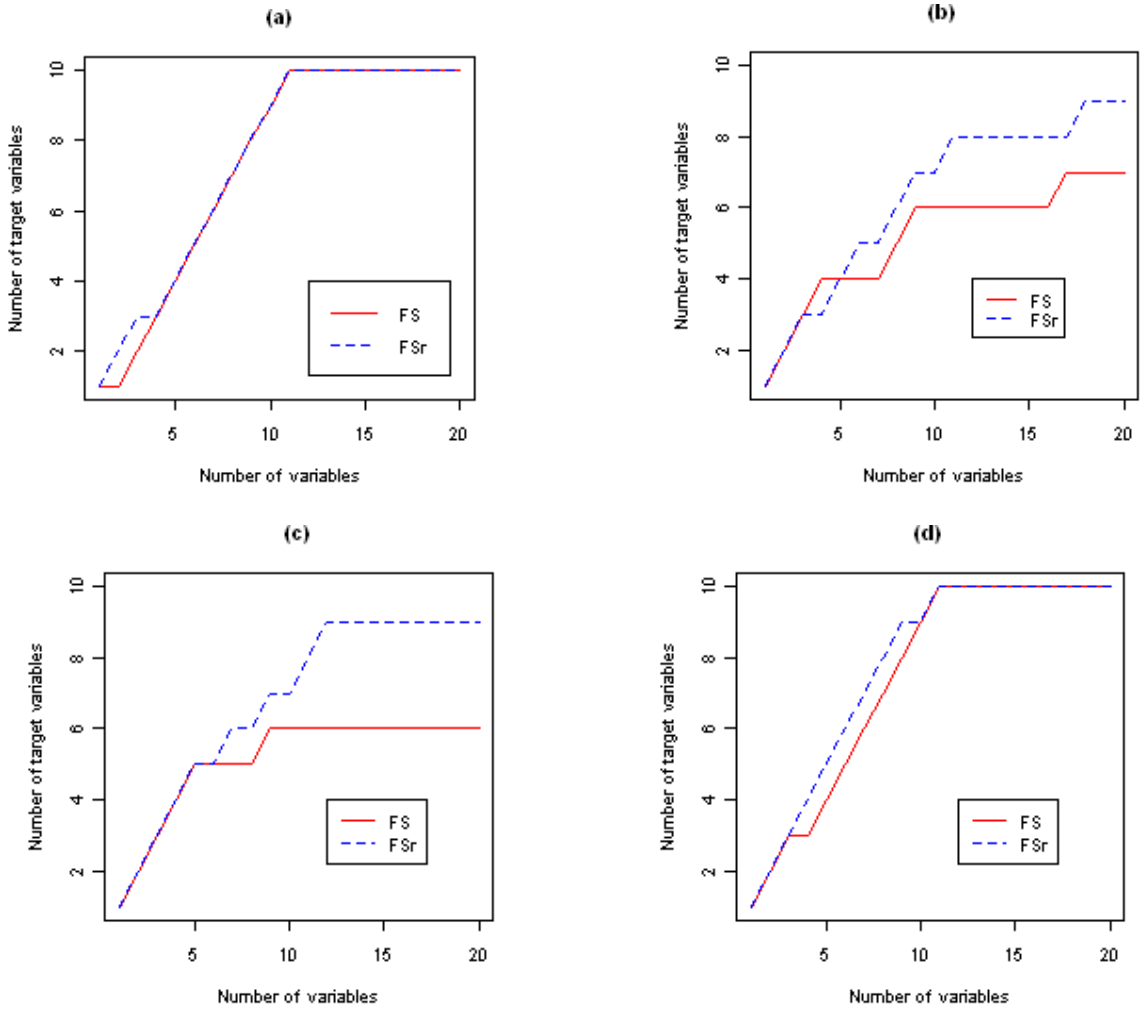

Fig. 1. Averages of the number of target variables t_m versus m for each of the methods and sampling situations considered. (a) No contamination; (b) slash contamination; (c) slash contamination/high leverage; (d) normal contamination/high leverage. We generated data sets of $d = 50$ predictors, $k = 10$ latent variables and 15% of contamination ($\epsilon = 0.15$) (— FS; --- FSr).

IV. Example

In this section, we use a real data set to further illustrate the performance of FSr compared to FS. A part of the data set considered in Table L of Draper and Smith (1998) is considered for this purpose. The response variable is the overall grade. We consider the 6 continuous covariates, which are numbered from 1 to 6.

In practice, we do not often know the number of covariates that are needed in the model. Thus, a graphical tool to select the size of the reduced set would be useful. For this purpose, Figure 2 was used. Starting with the first variable in the sequence, we increase the number of variables (along the sequence) and each time fit a regression model to compute the coefficient of multiple determination $\left(R^2\right)$. We then plot these R^2 values against the number of variables in the model to obtain a learning curve (see also, Croux *et al*. 2003). The size of the reduced set, m, can be selected as the point at which the learning curve no longer has a considerable slope.

FSr sequenced the covariates in the following order: (3,1, 6, 5, 4, 2). Figure 2(a) shows the learning curve for this data set. This plot suggests a reduced set of size 2, which includes covariates (3, 1). For comparison, the following different sequence was obtained by the FS: (2, 3, 1, 4, 5, 6). Fig. 2(b) shows the learning curve for this data set. The plot for FS suggests a reduced set of size 3, which includes covariates (2, 3, 1). That is, FS includes an additional covariate to get the same performance as FSr.

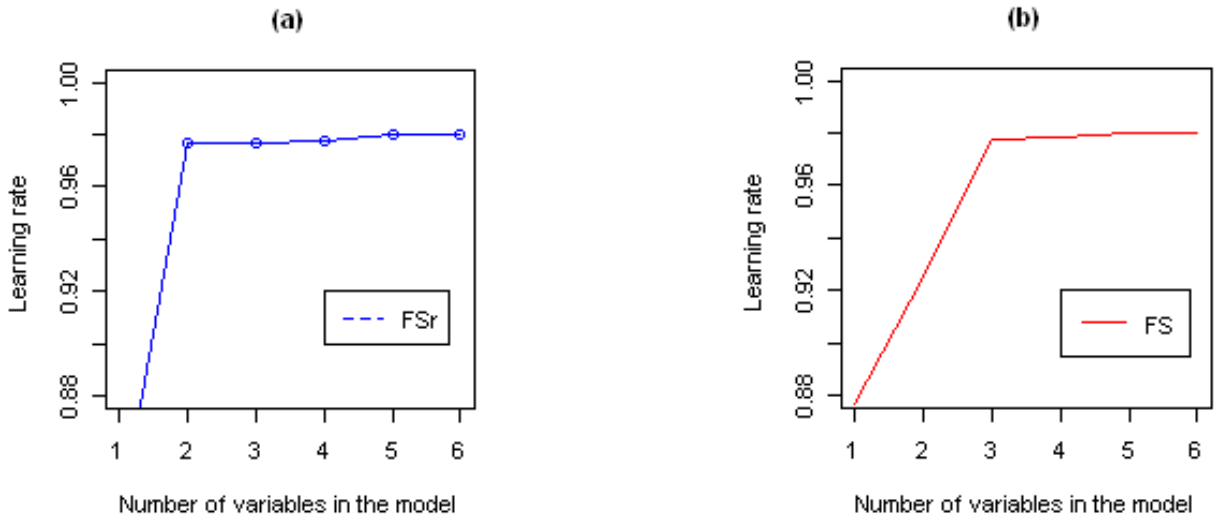

Fig. 2. Learning curve for the data set of Students Questionnaire Averages for 12 instructors

V. Limitations of FSr

If we sequence all d covariates, the FS procedure requires $O(nd^2)$ time. However, when applied with a stopping criterion, the complexity of FS depends on the number of covariates selected in the model. Assuming that the model size will not exceed a certain number $m < d$, the complexity of FS is less than or equal to $O(ndm)$. Since computational complexity of Spearman's ρ is $O(n \log n)$, therefore, the maximum complexity of FSr is $O((n \log n)dm)$, which is slightly larger than the FS.

VI. Conclusion

FS is a popular and computationally suitable algorithm for building linear prediction models. Khan *et al.* (2007) expressed FS in terms of Pearson's product moment correlations. The FS is very sensitive when the data contain contaminations. Since Spearman's rank correlation is a more reliable estimate of association in the presence of contaminations in the data, we have introduced this in FS. That is, we have ranked the values of each covariate, and then considered these ranks as the original values to apply the FS algorithm, and obtained FSr algorithm.

Our proposed FSr method has much better performance compared to the FS algorithm when the data contain contaminations. That is, FSr is more resistant than FS to the contaminated data.

---

1. Croux, C., P. Filzmoser, G. Pison, and P. J. Rousseeuw, 2003. Fitting Multiplicative Models by Robust Alternating Regressions. *Statistics and Computing*, 13, 23-36.
2. Das and Kempe, 2008. Algorithms for subset selection in linear regression. *Proceedings of the 40th annual ACM symposium on Theory of computing. ACM*, New York, NY, USA.
3. Draper, N., and H. Smith, 1998. Applied Regression Analysis. 3rd ed.. New York: John Wiley and Sons. 365 pp.
4. Frank, I., and J. H. Friedman, 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*. 35: 109-148. New York: Springer-Verlag.
5. Furnival, G. and R. Wilson, 1974. Regression by Leaps and Bounds. *Technometrics*. 16: 499-511.
6. Gatu, C. and E.J. Kontoghiorghes, 2006. Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics*. 15: 139-156.
7. Huo and Ni, 2007. When Do Stepwise Algorithms Meet Subset Selection Criteria? The Annals of Statistics 2007, Vol. 35, No. 2: 807-887.
8.. Khan, J. A., S. Van Aelst, R. H. Zamar, 2007. Building a Robust Linear Model with Forward Selection and Stepwise Procedures. *Computational Statistics and Data Analysis* (*CSDA*). 52(1): 239-248.
9. Weisberg, S., 1985. Applied Linear Regression (2nd ed.), Wiley, New York.