SELECTION OF VARIABLES IN THE LINEAR REGRESSION WITH THE RESTRICTED RRQR ALGORITHM

Dr. Lobna Eid AL-Tayeb

Associate Prof. of statistic
Faculty of Commerce, Al-Azhar University (Girls' Branch), Egypt
lobnaalatyeb@azhar.edu.eg

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.4, No.1, Part 1., January 2023

APA Citation:

AL-Tayeb, L. E. (2023) Selection of Variables in the Linear Regression with the Restricted PRQR Algorithm, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 4(1)1, 985-999.

Website: https://cfdj.journals.ekb.eg/

SELECTION OF VARIABLES IN THE LINEAR REGRESSION WITH THE RESTRICTED RRQR ALGORITHM

Dr. Lobna Eid AL-Tayeb

Abstract

Variable selection is a contentious issue that has spawned a variety of methods for finding the optimum regression equation with the fewest parameters. There are better linear independence features in the matrices of systems that are indeterminately compatible when using the RRQR (Rank-Revealing QR factorization) algorithm. An advantage of the RRQR technique is that it can be used to select variables with higher linear independence when determining the rank of a matrix. The RRQR decomposition with restricted pivot and an empirical model selection criterion such as Mallows' Cp are described in this paper. This procedure's benefits can be shown in two different scenarios, 'QR' and 'RRQR' decompositions.

Keywords: RRQR Algorithm, Mallows Cp criterion, regression equation.

1. INTRODUCTION

The problem of variable selection in linear regression continues to be the focus of many specialists [3]. Its general purpose is to establish a linear regression equation for a response variable Y in terms of certain predictor variables $Z_1, Z_2, \dots Z_K$ (or their functions), trying to reconcile two opposing criteria: on the one hand, include as many variables Z_s as possible so that the equation is useful for predictive purposes, and on the other hand, include the fewest possible number of variables Z_s to reduce the costs of obtaining information. In this way, the problem becomes "to find a balance between simplicity and fit" and it is precisely what we refer to when talking about the problem of "selecting the best regression equation".

When trying to solve this problem, there are two risks: on the one hand, that of including irrelevant variables, and on the other, that of omitting some relevant ones. It is necessary to take into account the essential difficulty that constitutes the ignorance of the variance of the random observations, which implies the need for subjective judgments. Thus, it can be said that there is no single procedure to select the best regression equation and that research continues in order to provide new procedures for solving this problem. In recent years, ideas have emerged regarding the inclusion of numerical criteria in the fit [5] that could contribute in some way to selecting the best regression equation in specific problems. The first problem has to do with the rank of the matrix. How I know studied in [4], [5] there are basically three types of algorithms for the computation of the rank of a matrix. These algorithms are: singular value decomposition [SVD] is certainly a decomposition that reveals the numerical rank, URV decomposition and the RRQR. Of the three algorithms, the one that offers a lower computational cost is the RRQR. This algorithm links to the second question, since the problem is equivalent to the variable selection problem in linear regression continues, studied in detail in [9]. We pursue, in this work, three fundamental objectives. The first refers to the evaluation of the RRQR algorithm with restricted pivoting for its application in linear regression. The second, to the study of the main statistical procedures provided in the literature for the selection of the best regression equation. And the third, to propose a new procedure for the selection of variables in the regression, combining numerical aspects and empirical criteria for the selection of models. We will deal with QR and RRQR decompositions in the initial stages, in the next stage we will establish the link with regression and summarize the main model selection procedures that the literature provides and that are most commonly used. A new procedure is explained in the upcoming section and finally examples are provided that allow a comparison of the proposed procedure with those used in practice.

2. THE 'QR' AND 'RRQR' DECOMPOSITIONS

The linear regression equation

$$y = z\beta + \varepsilon,$$

With design matrix Z_{mm}, vector of observations y_{mzl} and vector of random errorsε_{mzl}, it can be solved approximately by the method of least squares using the normal equations $Z'Zb = Z'y$. these allow calculating the estimates b of β that minimizes

$$|| y - Z\beta ||_2,$$

Which can be done as long as Z is full range and $Z'Z$ is well conditioned. In practice, between the variables that make up the columns of Z there may be quasi collinearity, and the vector of parameters to be calculated will be highly affected by error.

Another way to calculate b by the least squares' method is to make an orthogonal decomposition of the matrix$Z$, for example the one known asQR, with Q orthogonal and R triangular superior. In this way, it is minimized

$$|| Q' (y - Z\beta ||_2 = || y - Z\beta ||_2,$$

and it is more precise because it uses orthogonal transformations and does not construct the $Z'Z$ matrix, which in general is badly conditioned. The usual triangulation process in the QR algorithm is carried out by choosing as the pivot column of Z in each step the one with the highest Euclidean norm. In the restricted pivoting that we propose, the first columns of Z are those that correspond to a small number of variables, considered essential according to the specific application, which will not be eligible to carry out the permutations required by the transformation. This can be represented in a matrix as $ZP = QR$, where P represents the permutation of the columns.

This transformation is valid also when there is collinearity, being then the rank r of Z less than the number n of columns. In this case, we can consider R partitioned into blocks:

$$R \approx \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}$$

and if $||R_{22}||_2 < \varepsilon$ then

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

with R_{11} triangular superior rzr, of the same rank as Z. It is known that in this case, between the singular values of Z and $||R_{22}||_2$ there exists the following relationship

$\sigma_1 (Z) \geq \cdots . \geq \sigma_r (Z) \geq \varepsilon > \| R_{22} \| > \sigma_r + 1 (Z) \geq \cdots . \geq \sigma_n (Z)$,

Which theoretically defines the range r associated with tolerance ε. But since σr + 1 (Z) < ε, it is not always true that $||R_{22}||_2 < \varepsilon$ (Chan, 1987), which implies that the QR decomposition does not always reveal the range. Hence, the need to introduce a correction in the QR decomposition aimed at making the matrix R reveal the range, which constitutes the RRQR algorithm of Chan and Foster. This algorithm then provides in any case the ε-numerical rank of the matrix Z, as well as the permutation P that gives the order of importance of the columns, that is, of the variables for the regression.

3. LINK WITH REGRESSION

In the selection of the variables for the regression, two complementary types of analysis must be taken into account: that of the matrix Z and that of its relationship with the vector$y$. In analysis of the Z it is the one carried

out from the RRQR algorithm, which provides us with the information of how many are the most linearly independent variables of Z according to the numerical range, and which these variables according to the permutation obtained are.

For the analysis relative to the vector y, what is important to determine is which are the columns of Z that best describe y, that is, which are the variables most correlated with y, which at the same time are not highly correlated with each other. This analysis constitutes precisely the general objective of the statistical procedures for the selection of variables. Below we present a summary table of the main procedures provided in the literature.

Table 1: The main procedures selection of variables

PROCEDURES	GENERAL DESCRIPTION
All Possible Regressions	• Calculate all possible regression sets.
	• Evaluate each regression equation according to some empirical model selection criterion, such as the coefficient of determination, the adjusted coefficient of determination, the Mallows Cp or any other.
Step-by-Step Regression Upward inclusion Downward elimination PRESS	• Employ stop rules based on significance levels for hypothesis testing or other criteria.
	• Combines all regressions, residual analysis and validation technique.

Regression with dormant roots	• It is an extension of Principal Components Regression to examine alternative equations and eliminate predictive variables.
Staged regression	• Fit the equation with the Z most correlated with y, calculate the residuals and consider the residuals as the response variable of an equation with the Z (of the rest) most correlated with this new answer. Continue until you reach a desired state.

None of the above procedures gives a completely satisfactory solution. Thus, in the case of "All possible regressions" and even with some of its variants such as the one known as "Best Regression Subset" there is the opinion that it takes too much computation time and too much effort to examine all the regression equations. In popular computational procedures such as Step-by-Step Regression, Upward Inclusion, Downward Elimination, and different variations of these ideas, the criticism is made that the data analyst's ability to make decisions is lacking. In PRESS there is also a huge amount of computing and no precise rules are given to choose the best model. Arbitrarily biased estimates are used in Latent Root Regression, in addition to the fact that the search for latent roots is computationally costly. In Staged Regression there is no least squares solution and this does not satisfy many users of regression.

In the following section we present a new procedure that combines numerical criteria with empirical criteria for the selection of models.

4. NEW PROCEDURE FOR THE SELECTION OF VARIABLES IN THE REGRESSION: SELVAR

This procedure first uses the restricted RRQR algorithm [9] proposed by us to determine the rank of the matrix Z and the order of importance of the variables and then calculates the different regression equations by groups of p variables applying the Mallows Cp criterion [10], using the provides the information that allows the choice of the best model. Here p takes the values $n+1, n+2, \ldots, r$ where N is the number m of prioritized variables and r is the range. The total number T of models that must be analyzed is $T = 2^{r-n} - 1$, where $r - n$ is the number of groups. Thus, the computational complexity turns out to be much lower than that of other procedures for the selection of variables in the regression.

To carry out the SELVAR procedure, we have developed a MATLAB program [7] that consists of the following steps:

- Step 1: Center and scale the data and find the correlation matrix. Solve the entire model (if range near deficiency allows) to determine the estimator of the variance of the random observations. If not, determine it in step 4.
- Step 2: Ask the user to declare how many and which variables he considers essential; by default, take the one most correlated with the dependent variable y.
- Step 3: Apply the RRQR algorithm with restricted pivoting to the matrix Z, taking into account the prioritization defined in the previous step, and discard the variables corresponding to the deficiency in rank:
- $Z \rightarrow [Zq, Zk]$, where Zq represents the columns corresponding to the q prioritized variables and Zk, those corresponding to the k variables that complete the range r.

- Step 4: Calculate the regression equation with $n + k = r$ variables, that is, $y = f$ (Zq, Zk), and calculate the residual sum of squares RSS and the Mallows statistic Cp. If the complete model could not be solved, determine the variance estimator here.
- Step 5: Obtain the regression equations of the group of $p = q + 1$ variables, increasing the first of the non-essential ones, and calculate the corresponding Cp.
- Step 6: Decide which is the best equation for group p, according to the criterion Cp.
- Step 7: Repeat steps 5 and 6 for $p = n + j, j = 2,3, \ldots, k - 1$.
- Step 8: Choose the best equation of all.

5. RESULTS

Examples: We will analyze two examples. One, the well-known Hald data [2] on cement cooling, which has been used by countless researchers [8]. Another, the data obtained by the Egypt third national communication under the United Nations framework convention on climate change, with the purpose of establishing yield prediction equations hydrocarbons processing under processing conditions [6].

Example 1. The variables in the cement cooling data are four calcium compounds, and the response variable$y$, the heat released per gram of cement. In the application of the RRQR algorithm with restricted pivoting, as it does not have a prioritization criterion for the variables, the SELVAR procedure sets Z_4 as it is the one with the highest correlation with$y$. Rank three and permutation (4; 3, 1, 2) are obtained, which leads to discarding the variable Z_2.

The groups of equations considered were:

Two models of two variables (Z_4, Z_3 and Z_4, Z_1) and one of three (Z_4, Z_3, Z_1). The following table shows the results obtained:

Table 2: The results Two models of two variables (Z_4, Z_3 and Z_4, Z_1) and one of three (Z_4, Z_3, Z_1).

p	Equation	RSS	Cp	|Cp - p|	Best by group
2	$Y_{est} = f$ (Z_4; Z_3)	174.6	21.3	18.5	
	$Y_{est} = f$ (Z4; Z1)	73.6	3.6	1.6	*
3	$Y_{est} = f$ (Z_4; Z_3; Z_1)	51.7	1.6	1.6	*

As we have already said, this example has been the subject of extensive treatment in the statistical literature [1], so we can summarize the results reported on the application of other different model selection procedures.

Table 3: The different model selection procedures.

	Best regression equation	
PROCEDURES	**1st. place**	**2nd . place**
R^2	$f(Z_1, Z_4)$	f (Z_1, Z_2)
All equations S^2	f (Z_1, Z_2)	f (Z_1, Z_4)
Cp	f (Z_1, Z_2)	
Downward elimination	f (Z_1, Z_2)	
Step by Step (Back and forth)	f (Z_1, Z_2)	
PRESS	f (Z_1, Z_2)	
Dormant roots	f (Z_1, Z_2, Z_4)	
Staged regression	f (Z_1, Z_4)	
Restricted RRQR Cp	f (Z_4, Z_1)	f (Z_4, Z_3, Z_1)

It can be seen that according to our procedure, for simplicity, $Y_{est} = f\ (Z_4;\ Z_1)$ would be chosen as the best regression equation, and secondly, $Y_{est} = f\ (Z_4;\ Z_3, Z_1)$, which does not include the variable Z_2.

This is because Z_2 is highly correlated with Z_4, and was discarded, with Z_4 being the prioritized variable due to its maximum correlation with y. Hence, we consider of higher quality an equation with variables Z_4 and Z_1 as the best model, and not with Z_2 and Z_1.

Example 2. In the problem of predicting the hydrocarbon yield, the predictor variables are: number of processed components (Z_1), height of the plant (Z_2), dry matter (Z_3), content of nitrogen, sulphur and Oxygen ($Z_4$to Z_9) and the response variable y, the yield in gallons.

We consider two variants, a first, with prioritization of the variables Z_2 and Z_3 taking into account the criteria of the specialist combined with the correlation with and, and a second, where Z_7 (nitrogen content) is automatically prioritized because it is the most correlated.

First variant: According to the pre-set tolerance, the RRQR algorithm gives rank 5 for the matrix Z and permutation (2, 3; 1, 6, 5, 8, 4, 9, 7), which leads to discard the variables Z_8, Z_4, Z_9, and Z_7. Here it should be clarified that Z_1 was not prioritized, despite the fact that together with Z_2 and Z_3 they are easily measured without the need for chemical analysis, because their correlation with and was low. On the other hand, Z_7 was discarded by appearing in the last position of the permutation, and this is what leads to the analysis of a second variant.

The groups of equations considered were:

- Three models of three variables (Z_2, Z_3. Z_1; Z_2, Z_3. Z_6; Z_2, Z_3. Z_1)
- Three models of four variables (Z_2, Z_3, Z_1, Z_6; Z_2, Z_3, Z_1, Z_5; Z_2, Z_3, Z_6, Z_5)
- A model of five variables (Z_2, Z_3, Z_1, Z_6, Z_5).

The models are in this case the seven mentioned above since $r = 5$ and $n = 2$. If we compare it with the $2^{n-1} = 511$ for n = 9, which would require the analysis of all possible models, the effort reduction is undoubted computational that is obtained.

We summarize the results of the procedure in the following table:

Table 4: The results of the procedure for example 2.

p	Equation	*RSS*	*Cp*	$\lvert Cp - p \rvert$	Best by group
3	Y_{est} = f(Z2; Z3; Z1)	1.9	36.8	33.8	
	Y_{est} =f(Z2, Z3; Z6)	1.2	22.1	19.1	*
	Y_{est} =f(Z2, Z3; Z5)	1.8	34.6	31.6	
4	Y_{est} =f(Z2; Z3; Z1;Z6)	1.2	23.2	19.2	
	Y_{est} =f(Z2; Z3; Z1; Z5)	1.8	36.6	32.6	
	Y_{est} =f(Z2; Z3; Z6; Z5)	1.1	21.7	17.7	*
5	Y_{est} =f(Z_2, Z_3; Z_1; Z_6, Z_5)	1.1	23.3	18.3	*

Second variant:

By not defining essential variables, the procedure automatically prioritizes Z_7, which is the one with the highest correlation with y. By applying the RRQR, rank 5 is obtained for Z and permutation (7; 3, 2, 1, 5, 4, 6, 9, 8), then the variables Z_4, Z_6, Z_9 and Z_8 are discarded.

The groups of equations considered were:

- Four two-variable models (Z_7; Z_3; Z_7, Z_2; Z_7, Z_1 and Z_7, Z_5)
- Six models of three (Z_7, Z_3, Z_2; Z_7, Z_3, Z_1; Z_7, Z_3, Z_5; Z_7, Z_2, Z_1; Z_7, Z_2, Z_5 and Z_7, Z_1, Z_5)
- Four models of four (Z_7, Z_3, Z_2, Z_1; Z_7, Z_3, Z_2, Z_5; Z_7, Z_3, Z_1, Z_5 and Z_7, Z_2, Z_1, Z_5).
- A model of five (Z_7, Z_3, Z_2, Z_1, Z_5).

That is, in total N = 2^{5-1} - 1 = 15 models out of 511 possible. Below we summarize the results obtained for the best model of each group when applying the procedure.

Table 5: the results for the best model of each group when applying the procedure.

p	Equation	RSS	Cp	$\|Cp - p\|$
2	$Y_{est} = f(Z_7; Z_2)$	0.91	15	13
3	$Y_{est} = f(Z_7; Z_3, Z_2)$	0.62	8.4	5.4
4	$Y_{est} = f(Z_7; Z_3, Z_2, Z_1)$	0.43	5.6	1.6
5	$Y_{est} = f(Z_7; Z_3, Z_2, Z_1, Z_5)$	0.36	6.8	1.8

Taking into account this table and the one provided by the first variant, the specialist has the necessary information to make the final choice of the most convenient model, depending on whether or not it is possible to perform laboratory analysis in the cultivation region in question.

6. CONCLUSIONS

The new SELVAR procedure that we present reduces considerably the number of regressions to be carried out to choose the best equation, by setting the maximum size given by the range calculated using the restricted RRQR algorithm, and the variables corresponding to said range taking into account criteria of the specialist. It also allows the intervention of the human factor in decisions. Experimentation with the use of other model selection criteria such as robust Cp, as well as simulation studies that allow the comparison of criteria will serve to deepen the possibilities to select the best regression equation.

REFERENCES

1. Antil, H., Chen, D., & Field, S. (2018). A note on qr-based model reduction: Algorithm, software, and gravitational wave applications. *Computing in Science & Engineering*, *20*(4), 10-25.
2. Brook, R. J., & Arnold, G. C. (2018). Applied regression analysis and experimental design. CRC Press.
3. Bischof, C. H., & Quintana-Ortí, G. (1998). Algorithm 782: Codes for rank-revealing QR factorizations of dense matrices. ACM Transactions on Mathematical Software (TOMS), 24(2), 254-257.
4. Cardoso, M. A., Durlofsky, L. J., & Sarma, P. (2009). Development and application of reduced-order modeling procedures for subsurface flow simulation. International journal for numerical methods in engineering, 77(9), 1322-1350.
5. Deviation Maximization for Rank-Revealing QR Factorizations. Numerical Algorithms https://doi.org/10.1007/s11075-022-01291-1
6. EGYPT THIRD NATIONAL COMMUNICATION (2016) Under the United Nations Framework Convention on Climate Change Egyptian Environmental Affairs Agency (EEAA) 30 Misr Helwan Road, Maadi, Cairo, Egypt
7. Hunt, B. R., Lipsman, R. L., & Rosenberg, J. M. (2014). A guide to MATLAB: for beginners and experienced users. Cambridge university press.
8. Jessie, J. A., & Santhi, A. S. (2019). Effect of Temperature on Compressive Strength of Steel Fibre Reinforced Concrete. Journal of Applied Science and Engineering, 22(2), 233-238.
9. Mederos, M. V., Linares, G., & Estrada, J. L. (2013). SELECCION DE VARIABLES EN LA REGRESION LINEAL CON EL ALGORITMO RRQR RESTRINGIDO. Investigación Operacional, 21(3), 203-209.
10. Ronchetti, E., & Staudte, R. G. (1994). A robust version of Mallows's Cp. Journal of the American Statistical Association, 89(426), 550-559.

إختيار المتغيرات في الانحدار الخطي باستخدام المقننة RRQRالخوارزمية

د. لبني عيد الطيب

تعد مشكلة الاختيار للمتغير في الانحدار الخطي محل تركيز العديد من المتخصصين. حيث الغرض العام منه هو إنشاء معادلة انحدار خطي لمتغير استجابة Y من المتغيرات Z_1 ، Z_2 ،...، Z_K في محاولة للتوفيق بين معيارين متعارضين: من ناحية ، تضمين العديد من المتغيرات Z_s قدر الإمكان حتى تكون المعادلة مفيدة للأغراض التنبؤية ، ومن ناحية أخرى ، بتضمين أقل عدد ممكن من المتغيرات Z_s لتقليل تكاليف الحصول على المعلومات. وبهذه الطريقة تصبح المشكلة "إيجاد توازن بين البساطة والملاءمة" وهذا بالضبط ما نشير إليه عند الحديث عن مشكلة "اختيار أفضل معادلة انحدار."

عند محاولة حل هذه المشكلة، هناك نوعان من المخاطر: من ناحية، مخاطر تضمين المتغيرات غير ذات الصلة، ومن ناحية أخرى، مخاطر حذف بعض المتغيرات ذات الصلة. من الضروري مراعاة الصعوبة الأساسية التي تشكل عدم المعرفة بالتباين في الملاحظات العشوائية، مما يعني الحاجة إلى أحكام ذاتية. وبالتالي، يمكن القول إنه لا يوجد إجراء واحد لاختيار أفضل معادلة انحدار وأن البحث مستمر من أجل توفير إجراءات جديدة لحل هذه المشكلة.

نسعى في هذا البحث إلى ثلاثة أهداف أساسية. يشير الأول إلى تقييم خوارزمية RRQR مع التمحور المقنن لتطبيقها في الانحدار الخطي. والثاني يتعلق بدراسة الإجراءات الإحصائية الرئيسية الواردة في الأدبيات لاختيار أفضل معادلة انحدار. والثالث، اقتراح إجراء جديد لاختيار المتغيرات في الانحدار، والجمع بين الجوانب العددية والمعايير التجريبية لاختيار النماذج.

سنتعامل مع التحليلات QR و RRQRفي المراحل الأولية، في المرحلة التالية سننشئ الرابط مع الانحدار ونلخص إجراءات اختيار النموذج الرئيسية التي توفرها الأدبيات والتي هي الأكثر استخدامًا. وفي النهاية يتم عرض أمثلة تسمح بمقارنة الإجراء المقترحة مع تلك المستخدمة في البحث.

الكلمات المفتاحية: خوارزمية RRQR ، معيار Mallows Cp ، معادلة الانحدار .