Cross-Validated Sequentially Constructed Multiple Regression

Slav Angelov (1), Eugenia Stoimenova (2)

New Bulgarian University (1); Institute of Information and Communication Technologies, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (2), Sofia, Bulgaria

Introduction

Let observe the multiple regression as a tool for prediction. A well-known fact is that under the Gaus-Markov conditions the least squares estimator is the best linear unbiased estimator for multiple regression. Let assume that the G-M conditions hold, the input data are suitable for linear model and that the coefficients are estimated by the least squares estimator. Another important thing is that we are assuming that we have independent model variables. However, usually in practice, we have correlations between the variables, and as a result, the model has multicollinearity problem. Multicollinearity causes some of the coefficients in the model to be estimated with high variance or even to be biased because of the presence of suboptimal solutions. This is leading to unstable model, and the final result is poor predictions. Moreover, if there is a presence of multicollinearity and the model is recalculated by a data set with outliers than the adverse effects will be increased. If we examine the outliers as a separate problem, we should say that they can be influential to the model or not. Usually, even if they are not too influential they have higher variance than the other observations, and their individual or group effect causes the model to be biased which again may lead to poor prediction results for observations out of the learning set. Additionally, poor prediction result may be caused by overfitting the model - there is a small number of observations per model variable, and as a result, the estimated model is misleading. Under our initial assumptions, we can summarize that to obtain a linear model with better prediction performance than a multiple regression we need to take into consideration all of the mentioned factors.

Objective

To simplify the understanding of the aim of this paper, we will define one task. Let us assume that over a chosen learning set we have estimated a regression model with the desired model statistics and prediction results over some test sets. We want to assure that this model will remain with as nice diagnostics and prediction performance as possible after adding more observations to its estimation. The method proposed in this paper is an instrument for such a handling.

Cross-validated sequentially constructed linear model

We are observing the classical multiple regression model :

$$Y = \alpha_0 + \alpha_1.X_1 + \alpha_2.X_2 + \cdots + \alpha_k.X_k + \epsilon\,,$$

where Y is the predicted variable, α_i is the coefficient in front of predictor X_i for $i = 1, \ldots, k$ and ϵ is the error term. The model intercept term is α_0. $X_1, ..., X_k$ are centered.

The technique that we are proposing merges some of the model variables into components reducing in that manner the total number of variables. The way we merge two chosen variables into one component is the core of the method. Each component is forged in a procedure that involves two model variables (or other already obtained components). Let us assume for simplicity that we want to combine variables X_1 and X_2 into one component Z. Moreover, the estimate a_1 of the coefficient in front of X_1 is bigger in absolute value than the estimate a_2 of the coefficient in front of X_2. Then $Z(k) = X_1 + k * X_2$, where we are searching for the optimal $k \in (k_0 - \mu, k_0 + \mu), k_0 = \frac{a_2}{a_1}$, μ is a small number usually less than 1, while minimizing the *Root mean squared error after cross-validation* (RMSECV) for the multiple regression with the rest of the model variables (except the mentioned two) and the new component $Z(k)$. For $k = k_0$ we will attain the initial regression model.

$$\min_k RMSECV(k) = \min_k \sqrt{\frac{\Sigma_{i=1}^{n}(\epsilon_i(k))^2}{n}}, \quad k \in (k_0 - \mu, k_0 + \mu), \mu > 0,$$

where $\epsilon_i(k) = Y(i) - f_i(k)$ is the error from the i-th observation from the model $f_i(k)$. $f_i(k)$ is the multiple regression model with input variables $Z(k), X_3, X_4, ..., X_n$ estimated with the full data set except the i-th observation. Note that the input variables of $f_i(k)$ are centered without considering the i-th observation. The RMSECV(k) function is continuous and positive. Moreover, it is bounded from above. The simplest way to see this, without going into details, is to note that RMSECV(k) is constructed from errors derived from regression models and these errors are expected to be as small as possible. Thus, searching for the global minimum of the RMSECV(k) function is a valid operation.

The proposed method's framework

The proposed method has the following algorithm :

We have a multiple regression model with n variables. The input variables are centered.

1. We choose the model variable which has the worst estimate based on its absolute t-value. Then we find the most correlated with it model variable, and we combine them into one component ;
2. The regression model is estimated with the new component instead of the two chosen variables from Step 1 and all of the rest model variables.
3. Step 1 is repeated with the n-1 model variables ;
4. Step 2 is repeated with the obtained component from Step 3 and so on ;
5. The procedure ends when we have achieved absolute t-values over a chosen threshold for all of the derived model variables or when the model variables are reduced to a predefined percentage from their initial number. We suggest that the procedure should stop when all of the model variables are with absolute t-values over 4.5, or the number of the derived model variables is around 50% of the initial number of variables.

The obtained components are then used instead of the model variables.

The idea behind the approach

First, let us note that we are merging the variable with the highest estimation error for its coefficient with a variable that is most highly correlated with it. Thus, we are reducing the level of multicollinearity for the model, and as a consequence, we hope to achieve better estimated coefficients for the model with the new component. A handful instruments to measure the level of multicollinearity of a model are the *Variance inflation factors* (VIFs). VIFs measure the correlation between a chosen variable with all the other model variables.

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the Pearson's correlation coefficient R^2 that we can calculate if we regress x_j against all the other model variables.

VIFs over 4 are signs of a moderate multicollinearity problem, VIFs over 10 show extreme multicollinearity.

Second, each component is produced while minimizing RMCECV. RMSECV is the RMSE (see the example) while performing leave-one-out cross-validation [Mevik and Cederkvist, 2004]. This type of cross-validation is the one with the smallest variance of the error compared to the other types of cross-validation. Its goal is to see how well a regression model is capable of predicting out of sample observations. While minimizing RMSECV(k), we are searching for a component $Z(k)$ that allows the model that uses it to be as robust to changes in the test set as possible concerning the out of sample prediction performance.

An example

The real example is based on accounting and macroeconomic information from the firms in the Bulgarian gas distribution sector in the period 2007-14. The goal is the predicting of the *Return on assets* financial ratio for the next observed period using the input data from the current period. The full data set consists of 116 observations (three of these observations are omitted to improve the model). The model has seven variables one of which is a macroeconomic one. The description of the input will be skipped.

$$ROA(t) = \alpha_1 * PTA(t-1) + \alpha_2 * DC(t-1) + \alpha_3 * FL(t-1) + \alpha_4 * PTL(t-1) + \\ + \alpha_5 * LAOR(t-1) + \alpha_6 * I_{gas}(t-1) + \alpha_7 * Firm.size(t-1).$$

The proposed method is applied to the training set A_1 which contains the observations from years 2007-2010 (45 observations). The derived components by the method are three (43% of the primary variables). Then the regression model $F_{A_1^+}$ with the derived components is estimated from the set A_1^+ (64 observations) which contains one additional year 2011. The model over set A_1^+ is tested for the set B which contains the observations from years 2012-14 (49 observations). We want to compare $F_{A_1^+}$ with the multiple regression model F over the set A_1^+. The criteria for the comparison is the RMSE error, for more information see [Chai and Draxler, 2014].

$$RMSECV_f(A) = \sqrt{\frac{\Sigma_{i \in A}[\epsilon_f(i)]^2}{length(A)}},$$

where A is the test set, f is the regression model, $\epsilon_f(i)$ is the error of the i-th observation for model f.

$RMSE_{F_{A_1^+}}(A_1^+)$	$RMSE_F(A_1^+)$
0.01557	0.01564

We can see from the table that $F_{A_1^+}$ has lower prediction error. The results are close to each other, but the $F_{A_1^+}$ is with only three variables compared to the F model which is with 7, and these three variables are derived from 43 observations, that is 38% of the full data set.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.035989 -0.009293 -0.000917  0.008052  0.034613

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.035272   0.001783  19.783  < 2e-16 ***
Comp1        0.702125   0.058112  12.082  < 2e-16 ***
Comp2        0.051699   0.005210   9.922 2.87e-14 ***
Comp3       -0.770587   0.100550  -7.664 1.84e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.01426 on 60 degrees of freedom
Multiple R-squared:  0.8123,    Adjusted R-squared:  0.8029
F-statistic: 86.53 on 3 and 60 DF,  p-value: < 2.2e-16
```

We can see that the three components are well estimated while more data is added. The summary statistics are obtained using the *lm()* function in R language [R, 2017].

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.032800 -0.008674 -0.000790  0.007295  0.035164

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.035272   0.001807  19.523  < 2e-16 ***
PTA          0.649109   0.072873   8.907 2.55e-12 ***
DC          -0.654386   0.129761  -5.043 5.13e-06 ***
LAOR         0.012968   0.004448   2.916  0.00509 **
FL           0.052943   0.011205   4.725 1.59e-05 ***
PTL         -0.027365   0.008457  -3.236  0.00204 **
GP.index    -0.028285   0.012805  -2.209  0.03129 *
Firm.size    0.007876   0.003433   2.294  0.02554 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.01445 on 56 degrees of freedom
Multiple R-squared:  0.8201,    Adjusted R-squared:  0.7976
F-statistic: 36.46 on 7 and 56 DF,  p-value: < 2.2e-16
```

The summary statistics of the multiple regression over the set A_1^+.

About the multicollinearity issue it will be only mentioned that the highest VIF for the components in $F_{A_1^+}$ is 1.11 compared to 4.34 for F.

Acknowledgement. The authors acknowledge funding by the Bulgarian fund for scientific investigations Project I02/19.

References

[Chai and Draxler, 2014] Chai,T. and Draxler,R.R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE) ? Arguments against avoiding RMSE in the literature*. Geosci. Model Dev., Volume 7 : 1247Ű1250.

[R, 2017] R Core Team (2017). *R : A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https ://www.R-project.org/.

[Mevik and Cederkvist, 2004] Mevik, B. and Cederkvist,H.R. (2004). *Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR)*. Journal of Chemometrics, Volume 18 : 422-429.