A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

*Jacek Stelmach**

ON ESTIMATION OF A QUANTITY OF BASE MODELS WITH PARAMETRIC AND PERMUTATION TESTS

Abstract. One of the crucial problems in multiple-model approach of the regression is estimation of optimal number of base models. If the quantity is too low – it increases the prediction error whereas too high number of models increases time and complication of calculations. Unfortunately, the estimation of the quantity of base models based on the analysis of prediction error can lead to its overestimation. This paper proposes a formal approach where the predictions obtained with the models aggregated from different number of base models are compared. In this approach both: parametric and permutation tests were applied with the empirical data from petroleum industry.

Key words: permutation tests, aggregation models, regression methods.

I. INTRODUCTION

Parametric regression methods need to fulfill restricted assumptions. One of the solutions is using nonparametric methods, for instance multi-model approach. In this method, one of the most important factors is an optimal number of base models that gives small prediction error, reasonable time and complexity of the calculations. The methods of estimation of optimal number based on the analysis of prediction error can lead to an overestimation so alternative ways are proposed. The method proposed in this article bases on the analysis of prediction results. It is an adaptation of the formal method proposed by Latinne *et al.* (2002). That method was prepared for the classification purposes, the adaptation expands it into regression models. Additionally, different tests parametric and permutation are proposed what allows using this method also for the data with non-normal distribution.

II. PROBLEM DESCRIPTION

The aggregation of base models in parallel architecture means that the prediction results are a certain function of base models predictions (Garnar, 2008, p. 63):

* Ph.D. student, Department of Statistics, University of Economics, Katowice.

$$\widehat{D}^*(x) = \Psi\left(\widehat{D}_1(x), \widehat{D}_2(x), ..., \widehat{D}_N(x)\right) \quad (1)$$

where: $\widehat{D}_i(x)$ - prediction results of *i-th* base model.

The prediction error of aggregate model (when aggregation is carried out as an average of prediction results) can be decomposed into:

$$e(D^* \mid X) = N(X) + O(X) + S^2(X) \quad (2)$$

where:
O(X) - systematic effect (influence of a bias)
N(X) - noise error (irreducible)
$S^2(X)$ - variance effect (influence of a variance)

In a majority of regression methods, aggregation decreases the prediction error if only systematic effect does not dominate and the base models vary as much as possible. In such case at least two crucial problems appear: a way of creating independent and diverse base models and an optimal number of these models. The lower number of base models increases prediction error, the higher does not guarantee smaller error, increases calculation time and – if base model quantity is really very big – could lead to over- learning of the final model. The first issue is solved by a great number of different ways that assure the diversity of base models like:

- sampled learning dataset,
- selection of predictors to base models,
- change of dependent variable,
- change of parameters of regression method,
- different regression methods (Gatnar, 2008, pp. 103-106).

The second issue: the inference basing only on the analysis of the prediction error can lead in practice to the overestimation (Gatnar, 2008, p. 80). Breiman (1996) said that more than 25 base models for regression purposes should give the same misclassification rate (Breiman, 1996, p. 135). Different number of base models (but in classification purposes) proposed Opitz and Maclin (1999) for neural networks and/or decision trees – plateau of an error was observed at 10-15 base models (Opitz, Maclin, 1999, p. 182). However, no formal method that estimates the optimal number of base models in regression models was found in the literature.

III. PROPOSED METHOD

Proposed method bases on the investigations described by Latinne *et al.* (2002). Original method was prepared for classification purposes. The idea bases on the difference between prediction results for multi-models with different number of base models. If there are two models with K and N number of base models ($K < N$) and K_{min} is the smallest value for which the difference between results for both models is not significant – it means that K_{min} is searched value. Latinne carried out McNemar test to verify the null hypothesis about the lack of differences between prediction results. In this case – for regression models, other tests, described below were used. The following sequence of actions is proposed:

1. The aggregated model D_K^* of K base models is created, where K=1, 2, 3,…,50.
2. The aggregated models D_N^* of N base models are created, connected with each D_S^* model, where N= K+1, K+2,…100.
3. The prediction is calculated for D_K^* and D_N^* models.
4. If there are statistically significant differences between the prediction results of D_K^* and D_N^* models – K value shall be increased. In the opposite case, K value is an optimal value. It means that a base of the decision is verifying of null hypothesis: there are no significant differences between prediction results of both models.

IV. DATA DESCRIPTION

The experiment was carried out for the dataset with dependent variable "Average month price of paraffin wax melt point 56-58C" [EUR/t] from the period Jan 2003 to Aug 2012, named *WAX*. This data is published by ICIS (International Chemical International Service): http://www.icisprising.com. Last four cases (May 2012 to Aug 2012) were cut from the dataset, for ex-post analysis purposes – calculations of prediction results. Predictors were chosen – according to authors experience as a representative of upstream variables and variables that represent downstream - exchange rates of the examples of companies that consume paraffin wax and industry waxes [USD]:

- price of crude oil - Brent barrel [USD/barrel] – *OIL,*
- Caterpillar – *CAT,*
- Goodyear – *GT,*
- Freeport-Mcmoran Copper&Gold – *FCX,*
- United States Steel Co – *X,*

additionally:

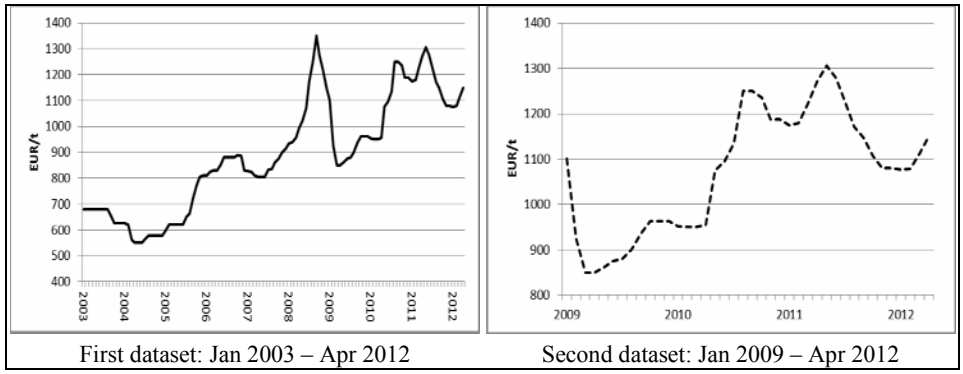
- EURO to USD exchange rate – *E2U,*
- lagged dependent variable – *WAX.*

All the predictors were lagged – according to maximum correlation with dependent variable.

Table 1. Lag values of the predictors according to max correlation with dependent variable.

Lag values of the predictors							
Predictor name	OIL	CAT	GT	FCX	X	E2U	WAX
Lag value	2, 3	2	6	3	3	5, 6	1, 2, 3

It was decided to examine two different datasets to achieve diversity of base models. The first one includes the original period: Jan 2003 to Apr 2012, the second – a period after big fluctuations of the economic crisis of Jan 2009 to Apr 2012. A graph of both cases is presented in Figure 1.



First dataset: Jan 2003 – Apr 2012 Second dataset: Jan 2009 – Apr 2012

Fig. 1. A graph of dependent variable for two datasets used in the experiment.

V. BASE MODELS

There were chosen a set of regression methods for creation base models purposes –recommended to achieve sufficient diversity and independence of the models:

- projection pursuit regression (PPR),
- neural network (multilayer perceptron MLP),
- regression tree,
- random forests.

Additionally, bagging (bootstrap aggregating) sampling was used to prepare data subset with: 30 and 50 cases. As a result, 240 base models were created for the experiment purposes and average of base model results was selected as the aggregation function.

According to a sequence described in chapter III, prediction results of D_K^* and D_N^* models shall be tested in order to accept or reject null hypothesis. The first calculation gave a picture of problems:

- a sample size is extremely low (4 cases – a period May 2013 to Aug 2013 was taken into consideration),
- although for around 80% of cases hypothesis about normal distribution was not rejected, low sample size tends to be cautious.

Therefore, both parametric tests that need to comply with the assumption of normal distribution and permutation tests (do not have such requirements) were carried out. As a parametric test the well known t-Student test was chosen. Test F, verifying the difference between the variances did not give significant results.

VI. PERMUTATION TEST DESCRIPTION

The idea of permutation test was proposed by R. A. Fisher. This test does not require any knowledge of the distribution since, instead of using any theoretical distribution, *ASL* (Achieved Significance Level) is estimated from empirical permutation distribution. And the power of permutation test is similar to parametric test, see Good P. I. (1994). The test used in the investigations verifies the hypothesis H_0: there are no differences between *A* and *B* populations, represented by the samples *a* and *b*. The sequence of actions is as follows:

1. Calculate the value of test statistics T^* for tested samples *a* and *b*:

$$T^* = \frac{\left|\overline{a} - \overline{b}\right|}{\overline{a} + \overline{b}} + \frac{\left|s_a^2 - s_b^2\right|}{s_a^2 + s_b^2} \tag{3}$$

where:

$\overline{a}, \overline{b}$ - mean estimator from samples: *a* and *b,*

s_a^2, s_b^2 - variance estimator from samples: *a* and *b.*

2. Perform a permutation (*M* times, usually it is recommended to be $M>1000$)[1] of dataset, it destroys existing dependencies of dataset.

3. Calculate the value of tests statistics for these permutations T_i, where $i=1, 2,\ldots,M$.

[1] Hesterberg T. *et al* (2003)*, The practice of business statistics, Companion chapter 18 – Bootstrap methods and permutation tests*, W. H. Freeman and Company, New York 2003, p. 45.

4. Locate calculated value of T^* in T_i distribution and estimate *p-value* as *ASL*:

$$ASL \approx \frac{card\{T_i : T^* \geq T_i\}}{M} \tag{4}$$

5. If received *ASL* value is less than assumed value of α level (for one-sided rejected region), the null hypothesis cannot be rejected.

VII. EXPERIMENT RESULTS

The experiment was carried out with Monte Carlo sampling for two tests cases:

- all the regression methods described in chapter V were used,
- only regression tree and random forests methods were used.

The results presented in Figure 2 allow estimating the optimal value as 25-30 base models. And bigger number of base models does not improve the prediction. Additionally, the prediction results (aggregated model with 25 base models) compared with multiple regression model and real data are presented in Table 2 and Figure 3.

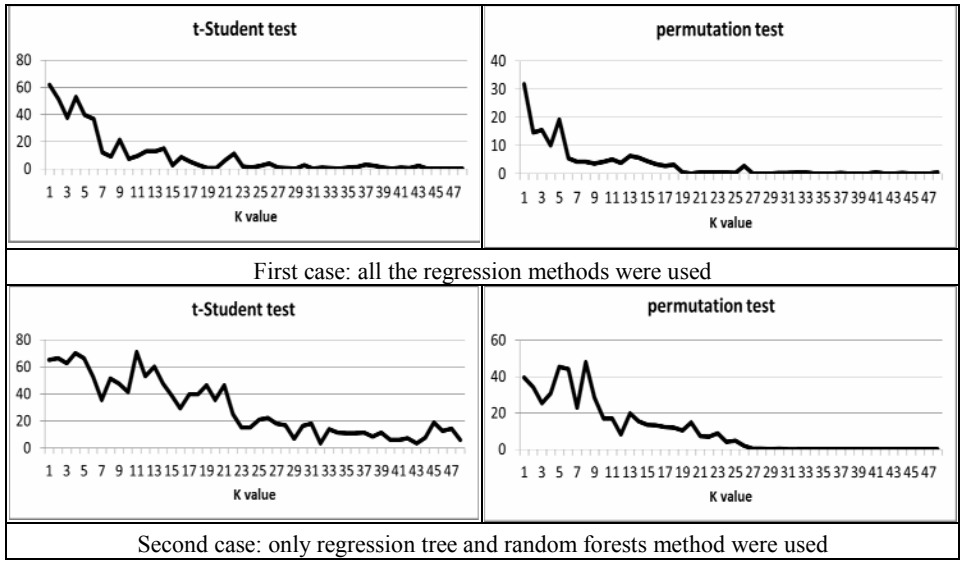


Fig. 2. The average percentage of null hypothesis rejection as a function of K parameter

Table 2. Prediction results of models compared with real data

Prediction results			
Month	real d.	multiple r.	aggregated
May	1170	1184	1172
Jun	1162	1169	1172
Jul	1130	1143	1149
Aug	1130	1086	1126

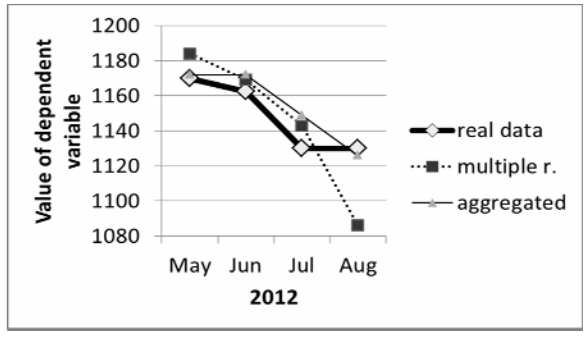


Fig. 3. Prediction results of multiple regression models, aggregated model compared with real data

VIII. CONCLUSIONS

1. The proposed method estimates the optimal number of base models as 25 models. Increasing this quantity has no impact on prediction error, it increases calculation time only.

2. Both tests: parametric and permutation give similar values, however permutation test does not have any requirements about the distribution.

3. Calculated results are similar to estimation proposed by Breiman (1996) and Opitz and Maclin (1999). The prediction of aggregated model with optimal number of base model is much closer to real data than calculated with multiple regression method.

REFERENCES

Breiman L. (1996), *Bagging Predictors,* Machine Learning, 26(2), pp. 123-140.
Gatnar E. (2008), *Podejście wielomodelowe,* Wydawnictwo Naukowe PWN, Warszawa.

Good P. I. (1994) *Permutation Tests: A practical guide for testing Hypotheses*, Springer-Verlag, N. York.
Hesterberg T. *et al* (2003), *The practice of business statistics, Companion chapter 18 – Bootstrap methods and permutation tests*, W. H. Freeman and Company, New York.
Latinne P. *et al*. (2002), *Combining different methods and numbers of weak decision trees*, "Pattern Analysis ans Applications", 5(2), pp. 201-209.
Opitz D., Maclin R. (1999), *Popular Ensemble Methods: An Empirical Study,* Journal of Artificial Intelligence Research 11, pp. 169-198.

Jacek Stelmach

O SZACOWANIU LICZBY MODELI BAZOWYCH ZA POMOCĄ TESTÓW PARAMETRYCZNYCH I PERMUTACYJNYCH

Jednym z kluczowych problemów w wielomodelowym podejściu do zagadnienia regresji jest estymacja optymalnej ilości modeli bazowych. Jeśli ich ilość jest zbyt mała – rośnie błąd predykcji, zbyt duża ilość powiększa czas i komplikację obliczeń. Niestety estymacja tej ilości na podstawie analizy błędu predykcji może prowadzić do jej przeszacowania.

W artykule proponuje się formalne podejście, w którym porównywane są wyniki prognoz otrzymanych z modeli zagregowanych z różnej liczby modeli bazowych. W tym przypadku wykorzystane zostały zarówno testy parametryczne jak i testy permutacyjne, a jako dane testowe: dane empiryczne wykorzystywane w przemyśle rafineryjnym.