The Combined Method of Forecasting the Investments within the Framework of Panel Data Models

Lyudmila O. Babeshko, Maria Y. Mikhaleva* and Irina V. Orlova

Financial University under the Government of the Russian Federation, Moscow, Russia

Abstract: The article is devoted to the panel data modeling of the firm's investments depending on its market value and the size of fixed assets. The Grunfeld's investment data as provided in R package were used as the initial data. The data frame contains annual observations for 11 firms over 20 years. The main econometric models for panel data (pooled model, fixed effects model, random effects model) were estimated. To make choice the most effective specification of the model the character of effects was tested. The heterogeneity of firms was explained by individual random factors. The comparative analysis of parameters' estimates was performed using the basic panel data models and their optimal combination in the framework of combined assessment (forecasting). Weight coefficients of hybrid forecasts are assigned as directed by the combined model list in accordance with standard optimality requirements. It was shown that the results of the combined assessment coincided with the estimates of the random effects model.

Keywords: Panel data, combined model, random effects model, fixed effects model, specification test, combined forecast, weight coefficients.

INTRODUCTION

Panel Data Models: Specification and Estimation Methods

There are three main types of econometric models for panel data in econometrics. The first one is pooled regression model. It does not consider individual features of panels:

$$y_{it} = \mu_i + x_{it} \cdot \beta + \varepsilon_{it},\ \mu_i = \underset{i}{const} = \mu, \quad (1)$$

where y_{it} is a dependent variable for panel *i* at time *t*, x_{it} is a 1×*k* vector of regressors (independent variables); ε_{it} is a term of random disturbance, $i = 1,...,n, t = 1,\ldots,T$, $\beta = (\beta_1, \beta_2, ..., \beta_k)'$ is a *k*×1 vector of the slope coefficients; μ is intercept which is common for panels in any period *t*.

The second type of panel data models is fixed effects (FE) model with the individual-specific effects μ_i, $i = 1,...,n$,

$$y_{it} = \mu_i + x_{it} \cdot \beta + \varepsilon_{it},\ \mu_i \neq \underset{i}{const}; \quad (2)$$

And the next type is random effects (RE) model. The RE model assumes that the individual-specific effects m_i are distributed independently of the regressors, $i = 1,\ldots,n$:

$$y_{it} = \mu_i + x_{it} \cdot \beta + \varepsilon_{it},\ \mu_i \neq \underset{i}{const},\ \mu_i = \mu + m_i,$$

$$y_{it} = \mu + x_{it} \cdot \beta + m_i + \varepsilon_{it} = \mu + x_{it} \cdot \beta + v_{it}, \quad (3)$$

$$E\{m_i\} = 0,\ Var\{m_i\} = \sigma_m^2,\ Cov\{m_i, \varepsilon_{it}\} = 0 \quad \forall i, j, t,$$

$$Cov\{\varepsilon_{it}, x_{is}\} = 0 \quad \forall i, j, t, s,\ Cov\{m_i, x_{jt}\} = 0 \quad \forall i, j, t$$

The individual random effects are uncorrelated with the regressors. If this premise is violated, the RE model's estimates are biased and untenable. The estimators for panel data models differ based on whether they consider the between or within variation in the data. These estimation methods are based on the exclusion of individual and general means. Fixed effects estimator uses the within variation and the individual-specific deviations of variables from their time-averaged values [Green 2012]:

$$Y^* = X^* \cdot \beta_w + \varepsilon^*, \quad (4)$$

where $Y^* = (I - P_w)Y$ — vector of deviations of endogenous variable from it time-averaged values,

$X^* = (I - P_w)X$ — matrix of individual-specific deviations of regressors from their time-averaged values,

$I - P_w = I - DD'/T$ — within transformation operator,

$\underset{nT,n}{D} = I_{n,n} \otimes I_T$ — matrix of dummy variables to control for the inter-individual heterogeneity of panel data in fixed effects models,

$$\hat{\beta}_w = \left(X^{*'} X^*\right)^{-1} X^{*'} Y^*, \quad (5)$$

*Address correspondence to this author at the Financial University under the Government of the Russian Federation, Moscow, Russia;
E-mail: MMikhaleva@fa.ru

— the vector of ordinary least squares (OLS) estimates of slope coefficients, which are used to obtain estimate of intercept:

$$\hat{\mu} = A_D Y - A_D X \hat{\beta}_w = \overline{Y} - \overline{X}\hat{\beta}_w , \quad (6)$$

where

$$A_D = \left(D'D\right)^{-1} D' = \frac{1}{T} I_{n,n} \cdot D' = \frac{1}{T} D'$$

- the operator to form the vector of individual means. The procedure (5), (6) is for ensuring consistency of parameter estimates if data include many panels and for explaining the differences within the panels.

In RE models it is assumed that the endogenous variable in each panel is influenced by specific independent and equally distributed random factors, in addition to the general ones. Therefore, the autocovariation matrix of random disturbances of the RE model has a specific structure. Taking it into account it is possible to increase the efficiency of parameter estimates by using generalized least squares method (GLS) [Verbick 2008]:

$$C_{vv} = T \cdot \sigma_b^2 \cdot P_w + \sigma_\varepsilon^2 \left(I - P_w\right),$$

where σ_b^2 is the variance of the disturbances between-group regression; σ_ε^2 is the variance of the disturbances of within-group regression. GLS estimates of the RE model are equivalent to the OLS estimates of the model with transformed variables

$$\underset{nT,1}{Y^{**}} = \underset{nT,n+k}{X^{**}} \cdot \underset{n+k,1}{\beta} + \underset{nT,1}{v^{**}} . \quad (7)$$

In (7) it is used within-group transformation:

$$Y^{**} = \frac{1}{\sigma_\varepsilon}\left(Y - \theta\overline{Y}\right),\ X^{**} = \frac{1}{\sigma_\varepsilon}\left(X - \theta\overline{X}\right), \quad (8)$$

where the individual averages $\overline{Y}$, $\overline{X}$ are used with a certain fixed weight:

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T} \cdot \sigma_b} . \quad (9)$$

To estimate the transformation parameter θ, it is necessary to estimate the within-group and between-group regressions. These regressions' residuals are used for estimating the standard deviations in the formula (9). According to the transformation formulas (8) if $\theta = 1$ the model (7) corresponds to the between regression (4), if $\theta = 0$ model (7) corresponds to the pooled model (1), therefore, the RE model can be considered as some combination of them.

Combined Method of Forecasting

It is interesting to compare the estimates of the endogenous variable obtained by the RE model and the estimates obtained in the framework of the combined assessment (forecasting), choosing as its components the estimates of the pooled models and the FE model. The evaluation of weight coefficients for individual forecasts significantly affects the accuracy of the resulting forecast [Bates and Granger 1969; Granger 1989]. A combined forecast can be represented as a linear combination:

$$\hat{Y}_c = \sum_{i=1}^{m} g_i \hat{Y}_i = g_1\hat{Y}_1 + g_2\hat{Y}_2 + ... + g_m\hat{Y}_m , \quad (10)$$

where $g = \left(g_1, g_2, ..., g_m\right)^T$ — a *m*×1 vector of weight coefficients, $\hat{Y}_i$ — forecast *n*×1 vector of basic model *i*:

$$\hat{Y}_i = Y + e_i ,\ i = 1,...,m , \quad (11)$$

where Y is *n*×1 vector of the true values of the endogenous variable; e_i is *n*×1 vector of forecast errors within the framework of the model *i* of the combined model list,

$$E\left\{e_i\right\} = 0 , \quad (12)$$

m — number of individual forecasts.

In [Babeshko and Byvshev 2017] the weight coefficients are determined under standard optimality requirements: unbiased errors of the combined forecast

$$E\left\{e_c\right\} = E\left\{\hat{Y}_c - Y\right\} = 0 \quad (13)$$

and minimality of their dispersion $Var\left\{e_c\right\}$. The requirement of unbiasedness (13) is reduced to the condition of normalization of weight coefficients

$$\sum_{i=1}^{m} g_i = 1 . \quad (14)$$

Because

$$E\left\{e_c\right\} = E\left\{\hat{Y}_c - Y\right\} = E\left\{\left(\sum_{i=1}^{m} g_i - 1\right)Y + \sum_{i=1}^{m} g_i e_i\right\} =$$

$$= \left(\sum_{i=1}^{m} g_i - 1\right)E\left\{Y\right\} + \sum_{i=1}^{m} g_i E\left\{e_i\right\} = \left(\sum_{i=1}^{m} g_i - 1\right)E\left\{Y\right\} = 0 ,$$

only if conditions (14) and (12) are true. The vector of the unbiased error of the combined forecast $e_c = e \cdot g$ has the variance

$$Var\{e_c\} = g^T C_{ee} g,$$

where *e* is $(n \times m)$- matrix with covariance C_{ee}. Its columns are vectors of forecasts error $e_i, i = 1,...,m$, obtained in the framework of the base set of models.

Thus, the conditions for optimality of the weight coefficients $g = (g_1, g_2, ..., g_m)^T$ can be formalized in the form of the Lagrange function:

$$L(g,\lambda) = g^T C_{ee} g - 2\lambda(g^T I - 1), \quad (15)$$

where λ is Lagrange multiplier, *I* is a unit column vector. The first order necessary conditions for an extremum of the function (15) lead to a system of equations:

$$\begin{cases} \frac{\partial L}{\partial g} = 2C_{ee}g - 2\lambda = 0 \\ \frac{\partial L}{\partial \lambda} = 2\left(g^T I - 1\right) = 0 \end{cases}.$$

The solution of this system allows to obtain the vector of weight coefficients [Babeshko and Yasakova 2017]:

$$g = \left(I^T C_{ee}^{-1} I\right)^{-1} C_{ee}^{-1} I, \quad (16)$$

where C_{ee} is an empirically estimated covariance matrix.

EMPIRICAL RESULTS

The comparison of the results of panel modeling and combined forecasts is performed for the model of dependence of the firm's investments (I_{it}) on its market value (F_{it}) and the value of fixed assets (C_{it}) [Green 2012]:

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it}, \quad (17)$$

i is the number of the company, $i = 1,...,n$, $t = 1,...,T$, according to the built-in software environment R "Grunfeld data", which includes annual observations for 11 large manufacturing firms over twenty years.

The equations (17) can be considered separately for each firm and evaluated by the OLS. However, if you combine observations of individual firms in panel data, it is possible to increase sample data, improve the efficiency of estimates, and study of individual characteristics of firms. To evaluate models for panel data in the software environment R, the plm function is used. It supports the estimations methods: pooled OLS (model = "pooling"), fixed effects (model = "within"), random effects (model = "random") [Kleiber and Zeileis 2008].

The following are the results of evaluation and testing of the model (1) according to panel data for four companies: "General Electric", "IBM", "Chrysler", "General Motors" for 20 years:

Pooled model:

$$\underset{(t)}{\underset{(s)}{\hat{I}}} = \underset{(-3,812)}{\underset{(17,549)}{-66,897}} + \underset{(10,735)}{\underset{(0,009)}{0,097}} F + \underset{(8,641)}{\underset{(0,036)}{0,315}} C, \quad \overline{R}^2 = 0,862,$$

$$F = 246,965,\ RSS_{pooled} = 823800,\ \hat{\sigma}^2_{pool} = 10698,7. \quad (18)$$

FE model:

$$\underset{(t)}{\underset{(s)}{\hat{I}}} = \underset{(-1,158)}{\underset{(58,205)}{-67,402}} i_{GM} \underset{(-8,428)}{\underset{(28,246)}{-238,061}} i_{GE} \underset{(-1,772)}{\underset{(15,770)}{-27,951}} i_{Ch} -$$

$$\underset{(-1,734)}{\underset{(14,022)}{-24,312}} i_{IBM} + \underset{(7,450)}{\underset{(0,014)}{0,104}} F + \underset{(16,494)}{\underset{(0,021)}{0,345}} C_{GM} \quad R^2 = 0,863,$$

$$F = 250,344,\ RSS_{FE} = 247700,\ \hat{\sigma}^2_w = 3347,257. \quad (19)$$

RE model:

$$\underset{(t)}{\underset{(s)}{\hat{I}}} = \underset{(-2,930)}{\underset{(25,757)}{-75,464}} + \underset{(9,294)}{\underset{(0,010)}{0,097}} F + \underset{(14,455)}{\underset{(0,024)}{0,342}} C,\ \overline{R}^2 = 0,865,\ F = 253,21,$$

$$RSS_{RE} = 338280,\ \theta = 0,609. \quad (20)$$

The choice of the best model from the set of basic (1)-(3) ones is based on tests that consider their hierarchical structure. The *F*-test tests the hypothesis $H_0 : \mu_i = \mu_j$ (pooled model vs. FE model, function *pFtest*()):

$$F = 57,375,\ p - value < 2.2e - 16,$$

shows that the null hypothesis should be rejected at any reasonable level of significance.

Lagrange multiplier test tests hypotheses $H_0 : \sigma_m^2 = 0$ (combined model vs. RE model, function *plmtest*()):

$$chisq = 340,25,\ p - value < 2.2e - 16,$$

rejects the null hypothesis in favor of the model *RE*.

As a result of the Hausman test: $H_0 : Cov\{\mu_i, x_{jt}\} = 0$ (the RE model vs. FE model, function *phtest*()):

$$chisq = 0,478,\ p - value = 0,787,$$

the null hypothesis is not rejected and that is why the RE model is chosen.

We estimate the endogenous variable of the model (17) in the framework of the combined approach: we choose the combined model and the FE model as the basic models. The parameters estimate of the model (18) - (20) are used to calculate the estimates of the endogenous variable, residuals and the covariance matrix:

$$C_{ee} = \begin{pmatrix} 10698,701 & 3135,405 \\ 3135,405 & 3347,257 \end{pmatrix}.$$

The weights for the combined forecast are determined by the formula (16):

$$g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \left(I^T C_{ee}^{-1} I\right)^{-1} C_{ee}^{-1} I = \begin{pmatrix} 0,3191 \\ 0,6809 \end{pmatrix}. \quad (21)$$

The result of combined forecasting of the endogenous variable on the base of models with weights (21) is estimates (forecasts) of the combined model (Comb):

$$\hat{Y}_c = g_1 \cdot \hat{Y}_{pooled} + g_2 \cdot \hat{Y}_{FE}.$$

Residiual sum of squares $RSS_c = 306351$ is close to $RSS_{RE} = 338280$. Figure **1** shows estimates of investment volumes in the framework of models *Pooled* (points superimposed on the line), *FE* (points superimposed on a bold line) *RE* (bold line), *Comb* (line). For descriptive reasons the estimates are of the one (first) panel of data since the volume of investments of the companies included in the panel data is significantly different, and it affects the scale.

Figure **1** shows that the RE model estimates practically coincide with the Comb model estimates. The same result can be demonstrated on other panels of data. Table **1** shows RSS for panel data containing observations for firms listed in column 2.

As can be seen from Table **1**, the combined method provides a more accurate result if the model with random effects is adequate. For the third panel data set (IBM, Atlantic Refining, Diamond Match, American Steel) the tests specified the model as a model with fixed effects, therefore, it makes no sense to combine the optimal model estimate *FE* ($RSS_{FE} = 2890$) with the Pooled model estimate *Pooled* ($RSS_{pooled} = 5218$). Although in this case, the combined method increases the accuracy compared to the individual evaluation of the Pooled model.

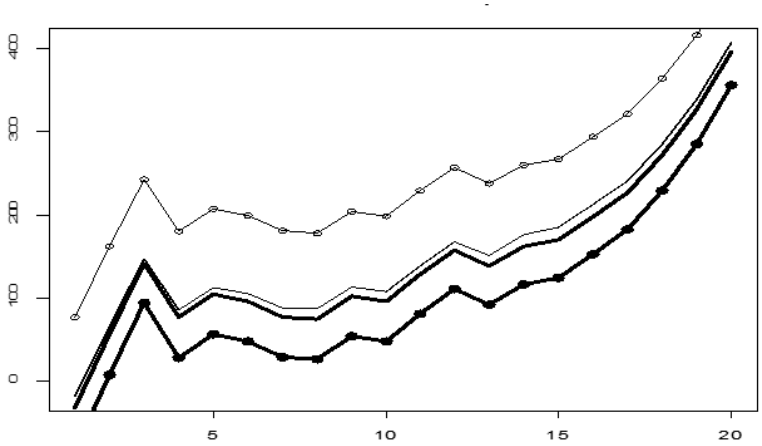

Figure 1: Estimates of the endogenous variable (values of investments).

Table 1:

№	Firms included in panel data	RSS_{pooled}	RSS_{FE}	RSS_{RE}	RSS_c
1	2	3	4	5	6
1	General Electric, IBM, Chrysler, General Motors	823800	247700	338280	306351
2	US Steel, Atlantic Refining, Union Oil, Westinghouse	296500	216520	233616	232540
3	IBM, Atlantic Refining, Diamond Match, American Steel	5218	2890	5218	3301

CONCLUSION

In this paper, to assess the dependence of the firm's investments on its market value and the value of fixed assets, the panel data models were used. The best model selection from a set of candidate models was performed using formal tests. It was shown that for Grunfeld's investment data the optimal model is RE model. Its GLS estimates are weighted average of estimates of within-group and between-group regressions. These weights depend on the ratio of the variances (adjustment parameter θ) of these models. Particular cases of the RE model are Pooled model (with $\theta = 0$, no heterogeneity effects) and FE (with $\theta = 1$, differences between panels being the main source of variation). The estimation results showed, that the combined evaluation with the weight coefficients (16), practically coincides with the results of the RE model provided that its specification was confirmed by the Hausman test.

REFERENCES

Babeshko L.O., Byvshev V.A. 2017. Forecasting of financial and economic indicators on heterogeneous data. Moscow, Russia: Rusains.

Babeshko L.O., Yasakova A.M. 2017. "Hybrid and selective prediction of financial indexes in the framework of randomized collocation," Economics. Taxes. Law, Vol. 20, No 2, pp. 51-57.

Bates J.M., Granger C.W.J. 1969. "The combination of forecasts," Operation Research Quarterly, Vol. 20, No 4, pp. 451–468.

Granger C.W.J. 1989. "Invited review: combining forecasts – twenty years later," Journal of Forecasting, No 8, pp. 167–173.

Green William H. 2012. Econometric Analysis (7th ed.). N.Y.

Kleiber C., Zeileis A. 2008. Applied Econometrics with R. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-77318-6

Verbick M. 2008. A guide to modern econometrics, edited by S.A. Aivazyan. Moscow, Russia: Scientific book.

Received on 08-06-2018 Accepted on 25-09-2018 Published on 12-11-2018

DOI: https://doi.org/10.6000/1929-7092.2018.07.67