Barometr Regionalny
Rok 2023 Tom 19 nr 1

Selection of Explanatory Variables for Linear Regression Models Estimated on Regional Panel Data

Mieczysław Kowerski
Academy of Zamość, Poland

Abstract

One of the problems in estimating of a single-equation linear regression model is the selection of explanatory variables. While many methods of selecting variables for models estimated on the basis of time series or cross-sectional data have been developed, there are no such methods of selecting variables for panel data models. The lack of an appropriate method for selecting variables for linear panel data models may lead to incorrect parameter values for some variables, which makes it difficult and sometimes even impossible to interpret the results of estimated model. Methods of selecting variables for panel data models cannot be based on the Pearson linear correlation coefficient. Therefore, a three-step procedure of variable selection for linear panel data models has been proposed, providing the correct parameter signs for all selected variables. The procedure is illustrated with the selection of variables for panel data models with fixed effects of the average annual unemployment rate according to Labor Force Survey (LFS) in Polish voivodships in the years 2010-2021 (balanced panel consisting of 192 observations).

Keywords: variable selection, fixed effects linear panel data models

DOI: 10.56583/br.2141

Introduction

One of the problems in estimating of a single-equation linear regression model is the selection of explanatory variables (specification problem). In an attempt to solve this problem, many methods of selecting variables for models estimated on the basis of time series or cross-sectional data have been developed. Grabiński, Wydymus and Zeliaś (1982) discuss over thirty such methods. A large group of variable selection methods uses the properties of the Pearson linear correlation coefficient. In Polish literature, a special place is occupied here by the Hellwig (1969) method for predictors' selection. Due to the presence in many computer econometric packages, stepwise regression as well as the from general to specific method are quite often used (Charemza and Deadman 1997). However, there are no statistical methods for selecting explanatory variables for linear regression models estimated on the basis of panel data.

Panel data (longitudinal data) is defined as a set of information about the population of uniquely identifiable objects observed over time. It can be said that panel data is a certain number of time series, each of which contains information about the values of the considered variable for a specific object (e.g., region) (Witkowski 2012, 267) or panel data is a certain number of cross-sectional series in subsequent periods, each of which contains information about the values of the considered variable in all examined objects. Thus, panel data is inherently "three-dimensional" and the dimensions are: variables, objects and time. Therefore, panel data have both the features of cross-sectional

E-mail addresses and ORCID digital identifiers of the authors
Mieczysław Kowerski • e-mail: mieczyslaw.kowerski@akademiazamojska.edu.pl • ORCID: 0000-0002-2147-2037

data (describing the set of objects at a single moment) and the features of time series (describing the object in different periods). If the same objects are observed in all periods, we are dealing with a balanced panel, while if for some periods there is no data on all objects, the panel is unbalanced.

The size of the panel is determined by the number of objects in the set (N) and the number of periods (T). In a balanced panel, the number of observations of each variable is $N \cdot T$.

If we have two variables (Y and X) describing the analyzed statistical population, then:

Y_{it} — is the value of the variable Y for the i-th object (region) in the t-th period (e.g., year),

X_{it} — is the value of the variable X for the i-th object (region) in the t-th period,

where: $i = 1,2,...,N$ and $t = 1,2,...,T$.

The main advantage of panel data is a greater amount of information about the same objects compared to, for example, cross-sectional data. Panel data enable simultaneous observation of the diversity of the studied objects and their evolution over time, which allows for better identification of the studied phenomenon. They make it possible to control and/or identify unobservable specific effects in regression models, and thus the use of the panel allows the removal of the estimator bias due to the omission of an important factor (Witkowski 2012, 268–269). Panel data provide a much larger number of observations, which increases the precision of inference and allows estimation of the dynamics of phenomena even when the number of periods is small.

The panel linear regression model is:

$$Y_{it} = \alpha_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \varepsilon_{it}, \tag{1}$$

where:

X_{jit} — value of j-th explanatory variable for the i-th object (region) in the t-th period,

$j = 1,2,...,k$.

For the model, should be selected explanatory variables that are the cause or symptom[1] of changes in the value of the analyzed phenomenon (the dependent variable). There is no problem of selection if we want to verify a certain theory that clearly indicates the factors (variables) determining the studied dependent variable. In regional studies, this is, for example, the theory of factor productivity (Dańska-Borsiak 2011; Tokarski, Roszkowska, and Gajewski 2005), which usually analyses the dependence of GDP in individual regions on capital and labour, or the theory of economic convergence of regions (Dańska-Borsiak 2011), which uses an unconditional β-convergence model that makes the value of GDP per capita dependent on its growth rate.[2]

Of course, the knowledge of the researcher about the factors determining the analyzed phenomenon cannot be overestimated here. You can also use the results of previous studies, although there is no certainty that the factors that determined the studied phenomenon in the past still do so.

But in the case of time-space (regional) analyzes, the diversity of many phenomena does not have formulated theories that precisely determine the factors determining these phenomena. Very often we talk about the impact of economic, social, demographic or behavioral factors, which can be measured using very different indicators (variables). Then the question arises, which of them to introduce into the panel data model?

In contrast to single-equation linear regression models built on time or cross-sectional series, in the case of panel data there are no methods for selecting variables for panel data models. On the other hand, it is not excluded that, although the relationship between the dependent variable and the potential explanatory variable is, for example, positive, the inclusion of such a variable in a model with many other variables will cause the parameter on it to turn out to be negative. That is, there will be a phenomenon of the lack of coincidence analogous to the one observed in the case of models estimated on the basis of time or cross-sectional series (Hellwig 1976), caused by the catalysis effect (Hellwig 1977). In the case of panel data, we cannot use the Pearson linear correlation coefficient to determine the relationship between two variables (dependent and explanatory or

1. In econometric modeling, a symptom is a phenomenon that is not the direct cause of the analyzed phenomenon, but "behaves" similarly to it (e.g. has a similar trajectory).

2. In conditional β-convergence model as explanatory variables in addition to GDP growth rate, there are other variables describing the economic situation of the region and it is necessary to decide what variables describe this situation.

two explanatory), although the authors of some analyses do so, (e.g., Borsuk and Kostrzewa 2020; Driver, Grosman, and Scaramozzino 2020; Herman 2019; Karkowska 2019; Pluskota 2020). Some authors leave the calculated matrices of Pearson linear correlation coefficients without comment, while others analyzing the level of significance of correlation coefficients and try to answer the question which variables should be included in panel data models as explanatory variables. In contrast, the Pearson linear correlation coefficient is not a measure of the relationship between variables with a panel structure. It is used to measure the relationship between two one-dimensional variables (time series or cross-sectional series). Therefore, if we have a balanced panel of N units in T periods, then in the formula of the Pearson linear correlation coefficient each of the $N \cdot T$ observations is treated regardless of which object and what period it concerns. The calculated Pearson linear correlation coefficient is therefore interpreted as a measure of the dependence between two variables in one object in $N \cdot T$ periods (e.g., years) or as a measure of dependence in $N \cdot T$ units in one period (e.g., year). For example, if we have a panel of 500 objects in 10 years—i.e., 5,000 observations, then the correlation coefficient can be treated as a measure of the relationship of two variables in 5,000 objects in one year or in one object in 5,000 years (which is obvious nonsense). The Pearson linear correlation coefficient does not take into account the basic advantage of panel data, which is to provide more complete information about the phenomenon being studied. For its calculation, it does not matter that one observation applies to i object in the period t and another to j object in the period $t + 1$, which is essential in the analysis of panel data. Thus, calculating and inferring from the matrix of linear Pearson correlation coefficients about relationships between variables, with a panel data structure is completely misplaced. The Pearson linear correlation coefficient is not a tool for panel analysis (Kowerski and Bielak 2021).

The aim of the study is to propose a method of selecting variables for regional panel data models. Of course, the proposed method is not able to replace the substantive selection of variables, which should precede the proposed statistical method of choice.

1 Proposal of a method for selecting variables for a single-equation static panel data model

Estimated static panel linear regression models make it possible to identify, in addition to explanatory variables affecting the dependent variable in the same way, individual factors (specific effects) that are characteristic for individual objects (regions) belonging to the panel and also affect the dependent variable. And this significantly improves the results of the analyzes carried out. There are two types of static models with specific effects:

- Panel linear model with fixed specific (individual) effects:

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \varepsilon_{it}, \tag{2}$$

  where α_i is fixed in time specific effect for i-th region.
  Fixed specific effects are often interpreted as an individual intercepts in model, different for each region but constant over time.
- Panel linear model with random specific (individual) effects:

$$Y_{it} = \alpha_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \gamma_{it}, \tag{3}$$

  where $\gamma_{it} = \alpha_i + \varepsilon_{it}$ is new disturbance which is the sum of fixed effects and disturbance (Maddala 2006, 645).

Each region is assigned a random variable, the implementation of which is responsible for the specific effect at a specific moment. In this model, specific effects are different from period to period. Specific effects in a model with random effects are often interpreted as specific random components (Biørn 2017, 65–66).

To estimate a model with fixed effects, the least squares dummy variables method—LSDV (Maddala 2006, 644) is used. A general least squares method—GLS is used to estimate the parameters of a model with random effects (Hsiao 2014, 39–40). Both methods are included in the

GRETL program (Kufel 2011) and will be used in this work. The estimated models are verified by means of appropriate statistical tests. For static models, the presented method uses two tests:

- Calculated for a fixed-effects model, the Welch test for differences in intercepts between regions. The null hypothesis is that all regions have a common intercept. The rejection of the null hypothesis means that each region has its own intercept, which in turn means that the on the variability of the dependent variable, in addition to the specified explanatory variables, influence specific (individual) effects characteristic only for each of the analyzed regions. The rejection of the null hypothesis means that a panel data model with fixed effects is appropriate to estimate the dependent variable.
- Calculated for model with random effects Hausman test for consistency of GLS estimator. The null hypothesis is: the GLS estimator is consistent. Fail to reject the null hypothesis means that a panel data model with random effects is appropriate for estimating the dependent variable.

In the presented method, we assume that a substantive selection of variables was made earlier, consisting in the selection of potential explanatory variables that may causally or symptomatically affect the dependent variable. The proposed method, on the other hand, allows to select from among the potential explanatory variables influencing on the dependent variable those that will provide the construction of the best linear regression model due to the quality of the estimated parameters and their interpretative properties. In other words, we choose "primus inter pares" variables.

The method consists of three stages and is carried out separately for the fixed-effects model and the random-effects model:

- Stage 1

  Panel data models with fixed and random effects of the dependent variable with each potential explanatory variable separately are estimated. Then:
  - It shall be checked whether the parameter on the explanatory variable is statistically significant at 0.05. If the parameter is not significant, this potential explanatory variable shall be excluded from further testing.
  - Tests shall be carried out:
    - Welch test for differences in intercepts between regions. The rejection of the null hypothesis indicates that a model with fixed specific effects is correct. Fail to reject of the null hypothesis means that the potential explanatory variable is excluded from further research.
    - Hausman test — fail to reject the null hypothesis indicates a model with random specific effects as correct. Rejection of the null hypothesis causes the potential explanatory variable to be excluded from further research.
  - For further calculations, those variables are assumed where the parameters are significant and at the same time the appropriate test indicates that the model is correct.

  It may happen that for some potential explanatory variables, both a model with fixed effects and a model with random effects are appropriate. Then such a variable is taken into account both in the model with fixed effects and in the model with random effects.
- Stage 2

  Panel data model of the analyzed dependent variable with respect to not excluded in the first stage potential explanatory variables is estimated. And the coincidence of the sign of parameter on respective variable in the estimated model with sign of parameter on the same variable in the estimated on the first stage model with one variable is checked. Variables with parameters that do not meet the coincidence are excluded from further testing. This procedure is performed separately for models with fixed and random effects.
- Stage 3

  For the potential explanatory variables remaining after two stages, the method of selection from general to specific is used (Charemza and Deadman 1997). This method assumes that first a model taking into account all potential explanatory variables is built. Then the variable with the highest value of the empirical significance level p (but greater than 0.05) is selected. This variable is removed and new models are evaluated in subsequent steps until all significance level p are less than 0.05.

2 Application of the proposed method — an example

Due to the fact that spatial disparities in unemployment figures across regions are persistent in many economies (Antczak, Gałecka-Burdziak, and Pater 2018, 25), the proposed method is illustrated by the example of the selection of variables for panel data models with fixed effects of the average annual unemployment rate according to the Labor Force Survey (%) in Polish voivodships in the years 2010–2021 (LSUR). For this purpose, a balanced panel consisting of 192 observations was constructed.

As a result of substantive selection, 23 potential explanatory variables characterizing the economic situation, labor market, demographic situation, religiousness of residents and material con-

Table 1. Potential explanatory variables proposed in the substantive selection process

Variable	Acronym
Economic situation	
GDP per capita. Constant prices 2021 (thousand PLN)	PKB
Investments per capita. Constant prices 2021 (thousand PLN)	INV
Gross value of fixed assets in the national economy per capita. Constant prices 2021 (thousand PLN)	FAssets
Expenditures of budgets of communes and cities with powiat status per capita. Constant prices 2021 (thousand PLN)	MuniEx
EU funds per capita. Constant prices 2021 (thousand PLN)	UE
Annual inflation rate (%)	CPI
Labor market	
Yearly average labor force participation rate according to LFS (%)	LFSAR
Balance of jobs created and eliminated per 1000 employees	C_L
Percentage in total employment in Section A: agriculture, forestry, hunting, fishing (%)	SecA
Religiousness of residents	
Share of religious marriages (%)	RELIG
Demographic situation	
Population density (persons per km^2)	DENSI
Urbanization rate — share of urban population (%)	URB
Fertility rate	FERTIL
Marriages per 1000 population	MARR
Average life expectancy of a male newborn (years)	LEM0
Average life expectancy of a female newborn (years)	LEW0
Dependency ratio. Non-working age population per 100 people of working age	DEPEND
Material condition of households	
Average monthly expenses per person. Constant prices 2021 (thousand PLN)	EXP
Average monthly disposable income per person. Constant prices 2021 (thousand PLN)	INCOMES
Percentage of households equipped with a dishwasher (%)	DISHW
Percentage of households owning a car (%)	CAR
Percentage of households with a home cinema system (%)	CINEMA
Apartments per 1000 inhabitants	DWEL

Note: Apart from GDP per capita, other value data have been converted into 2021 prices using the consumer price index (CPI). The GDP values in 2021 prices were determined on the basis of the GDP growth rate in constant prices in subsequent years in individual voivodships and GDP per capita in voivodships in current prices presented by the Central Statistical Office

dition of households in the voivodships forming the panel in the analyzed period were specified. Thus, these are both causal variables and symptomatic variables[3].

Six variables were used to describe the economic situation, which in this method are treated equally: each of them is the cause of differentiation of unemployment rates in regions and over time. You can guess that they are related to each other, which can usually lead to a lack of coincidence in the case of introducing all or part of them into the model. The proposed method will "indicate" which variables will meet the statistical properties of the model.

Similar situations occur for the other proposed potential explanatory variables.

In the first stage, 18 potential explanatory variables were selected for the fixed-effects model, where the parameters in single-variable models proved to be significant and the Welch test indicated the existence of fixed specific effects. Among the variables "qualified" for the second stage, there were 5 variables describing the economic situation of voivodships; the higher the GDP, the level of investment, the value of fixed assets and local government expenditures, the lower the unemployment rate. By contrast, the unemployment rate is higher in regions with higher levels of EU subsidies, which may be due to more support for regions at risk of permanent unemployment. The

Table 2. Results of estimation of panel data models with fixed effects of the unemployment rate on individual potential explanatory variables and a model with all variables selected in stage 1

	Single variable models			Model with all variables selected in stage 1		
Variable	**Estimated parameter**	***p*-value of parameter**	***p*-value of Welch test**	**Estimated parameter**	***p*-value of parameter**	**Coincidence of the sign of parameter**
PKB	−0.0580	< 0.0001	< 0.0001	0.0651	0.0140	No
INV	−0.3760	< 0.0001	< 0.0001	0.1812	0.0647	No
FAssets	−0.0442	< 0.0001	< 0.0001	−0.0290	0.0167	Yes
MuniEx	−1.2384	< 0.0001	0.0071	1.3554	0.0039	No
UE	0.0161	0.0336	< 0.0001	−0.0019	0.7672	No
LFSAR	−0.4006	< 0.0001	< 0.0001	−0.2597	0.0022	Yes
C_L	−0.0778	0.0007	< 0.0001	−0.0146	0.2009	Yes
SecA	0.1054	< 0.0001	< 0.0001	0.0633	0.0113	Yes
RELIG	0.0820	< 0.0001	< 0.0001	−0.0161	0.5796	No
DENSI	−0.0042	< 0.0001	< 0.0001	−0.0039	0.0645	Yes
URB	−0.0927	< 0.0001	< 0.0001	−0.0180	0.5194	Yes
FERTIL	−7.2269	< 0.0001	< 0.0001	−4.6763	0.0316	Yes
LEW0	0.6950	0.0009	< 0.0001	−0.5084	0.0157	No
EXP	−6.1704	< 0.0001	< 0.0001	2.1034	0.1433	No
INCOMES	−5.7912	< 0.0001	0.0002	−4.0733	< 0.0001	Yes
DISHW	−0.1760	< 0.0001	< 0.0001	−0.0912	0.0501	Yes
CINEMA	−0.1909	0.0001	< 0.0001	0.0390	0.2121	No
DWEL	−0.0299	< 0.0001	< 0.0001	−0.0182	0.1001	Yes
Variables eliminated in the stage 1						
CPI	−0.1623	0.6462	< 0.0001			
MARR	0.4170	0.3521	< 0.0001			
LEM0	0.1765	0.2162	< 0.0001			
DEPEND	−0.0605	0.2248	0.0139			
CAR	0.0436	0.1579	< 0.0001			

3. Symptomatic variables are variables that are not the direct cause of the phenomenon being studied (the dependent variable) but they are "behave" like dependent variable (Nowak 2002, 9).

Table 3. Second stage of variable selection

		Step 1		Step 2		Step 3		Step 4		Step 5		Step 6		Step 7		Step 8		Step 9		Step 10		Step 11		Step 12	
Variable	Sign[a]	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p	sign	p
PKB	−	+	0.0140	+	0.0134	+	0.0202	+	0.0200	+	0.0121	+	0.0025	+	0.0003	+	0.0005	+	0.0003						
INV	−	+	0.0647	+	0.0655	+	0.0746	+	0.1081	+	0.1304														
FAssets	−	−	0.0167	−	0.0109	−	0.0281	−	0.0278	−	0.0428	−	0.0797	−	0.1112	−	0.1807	−	0.1654	+	0.0029	+	0.0028		
MuniEx	−	+	0.0039	+	0.0078	+	0.0068	+	0.0042	+	0.0044	+	0.0023												
UE	+	−	0.7672																						
LFSAR	−	−	0.0022	−	0.0027	−	0.0006	−	0.0009	−	0.0006	−	0.7082	−	<0.0001	−	0.0002	−	0.0002	−	0.0189	−	0.0202	−	0.0365
C_L	−	−	0.2009	−	0.1947	−	0.2085	−	0.2187	−	0.3835	−	0.1392	−	0.2244	−	0.2529	−	0.2478	−	0.3115	−	0.3223	−	0.2749
SecA	+	+	0.0113	+	0.0136	+	0.0675	+	0.0906	+	0.1038	+	0.0055	+	0.2199	+	0.6784	+	0.6061	−	0.6113				
RELIG	+	−	0.5796	−	0.6234																				
DENSI	−	−	0.0645	−	0.0711	−	0.014	−	0.0144	−	0.0118	−	0.0055	−	0.0002	−	0.0008	−	0.0006	−	0.5416	−	0.5485	−	0.7011
URB	−	−	0.5194	−	0.5796	−	0.6256	−	0.6588	−	0.4675	−	0.4295	−	0.6625	−	0.5521	−	0.591	−	0.0142	−	0.0197	−	0.0059
FERTIL	−	−	0.0316	−	0.0411	−	0.0258	−	0.0289	−	0.0164	−	0.012	−	0.1677	−	0.1193	−	0.0592	−	0.1905	−	0.2553	−	0.2187
LEW0	+	−	0.0157	−	0.0185	−	0.0036	−	0.0015	−	0.0014	−	0.0015	−	0.0224										
EXP	−	+	0.1433	+	0.152	+	0.1307	+	0.1401																
INCOMES	−	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	< 0.0001	−	0.1663
DISHW	−	−	0.0501	−	0.0493	−	0.0432	−	0.0359	−	0.0442	−	0.0539	−	0.0347	−	0.0102	−	0.0073	−	0.0016	−	< 0.0001	−	< 0.0001
CINEMA	−	+	0.2121	+	0.2187	+	0.2427																		
DWEL	−	−	0.1001	−	0.0884	−	0.0759	−	0.0782	−	0.1094	−	0.1575	−	0.859	+	0.6760								

[a]Sign of parameter in single variable model.

discussed dependencies are causal. The dependence of the unemployment rate on the inflation rate turned out to be insignificant. The unemployment rate in the analyzed period also depended on the situation on regional labor markets. The significant negative relationship between the unemployment rate and labor force participation rate of which the unemployment rate is a component is fully understandable. The unemployment rate was lower in regions where the number of jobs created exceeded the number of jobs lost, but it was higher the higher the share of employment in Section A companies in a given region. These are also causal dependencies. The unemployment rate was lower in more urbanized and densely populated regions, but also in regions with higher fertility rates, which may be symptomatic relation. Similarly, the positive dependence of the unemployment rate and the average life expectancy of a female newborn, as well as the share of religious marriages, may be symptomatic.

A significant dependency was found between the situation of households and the unemployment rate. In regions where households have higher incomes and expenses and are better equipped, the unemployment rate is lower. Although in this case it is difficult to say that the situation of households is the cause of a lower unemployment rate — rather it is the low unemployment rate that positively affects the situation of households. This requires a solution to the problem of endogeneity at the stage of model construction.

The second stage of the procedure for selecting variables for the fixed-effects model began with its estimation based on all 18 variables selected in the first stage. Next, the coincidence of parameter signs on individual explanatory variables in the estimated model with parameter signs on the same variables estimated in the first stage single-variable models was checked. In 8 cases, the signs did not match. Such variables should be eliminated, but elimination was made starting from the variable on which the parameter was characterized by the highest p-value. In the course of this procedure, further variables appeared on which the signs did not match the signs in the corresponding single-variable models. The EU variable was eliminated first, followed by RELIG, CINEMA, EXP, INV, MuniEx, LEW0, DWEL, GDP, SecA and FAssets. In total, 11 variables were eliminated in the second stage. There were 7 variables left (LFSAR, URB, DISHW, C_L, DENSI, FERTIL and INCOMES), with the parameters on the last four being insignificant.

In the third stage, variables where parameters were insignificant were eliminated, starting with the largest p-value (DENSI). In the next steps, C_L and FERTIL were eliminated, obtaining an optimal set of 4 variables: LFSAR, URB, DISHW, INCOMES.

In the model estimated on the basis of the finally selected variables, the parameters for all variables are negative and significant at the level of 0.05. The unemployment rate is lower in re-

Table 4. Results of estimation of panel data model with fixed effects of the unemployment rate (LFSUR) on the finally selected variables

	Parameter	***p*-value**
Variable		
Intercept	22.8642	< 0.0001
LFSAR	−0.1671	0.0007
URB	−0.0327	0.0038
INCOMES	−1.0170	0.0325
DISHW	−0.1013	0.0002
Coefficient		
LSDV R-squared	0.9039	
Within R-squared	0.5704	
Welch test for differing intercepts in regions; test statistics $F(15, 66)$	4.4317	< 0.0001
Wooldridge test for autocorrelation in panel data. H_0: No first-order autocorrelation; test statistics $F(1, 15)$	0.0739	0.7895

gions with higher labor force activity, more urbanized ones, where households have higher incomes and are better equipped with durable goods — here represented by dishwashers. The estimated model better explains the differences in the values of the dependent variable between regions than the changes in the dependent variable in each region during the analyzed time (LSDV R-squared higher then within R-squared). The Welch test confirms the correct use of a model with fixed effects, and the Wooldridge test for autocorrelation in panel data confirms the absence of first-order autocorrelation.

3 Discussion of the results

The discussion of the obtained results will start with the assumption that the analyzer of factors determining changes in the unemployment rate in Polish voivodships in the years 2010–2021 does not use any method of selecting variables and relies only on substantive knowledge about the studied phenomenon and as explanatory variables it assumes GDP per capita (PKB), as a proxy of various economic development effects, activity rate (LFSAR), population density (DENSITY) and household income per capita (INCOME). The first two variables are undoubtedly causal variables, and the next two are rather symptomatic. The estimated unemployment rate model (LFSUR 2) meets the basic assumptions — all parameters are statistically significant, there is no autocorrelation of disturbances, the Wald test indicates the diversity of intercepts in regions, so the model with fixed specific effects is an appropriate estimation tool. And now let's try to interpret the results obtained. The parameters on the GDP and DENSITY variables are positive — the higher the level of economic development of the voivodship and the more populated the region, the higher the unemployment rate. This contradicts both economic theory and practice. Voivodships with high GDP and population, which is also a symptom of economic development, are usually characterized by a lower unemployment rate. Such a result becomes understandable only when we apply the proposed method of selecting variables.

Among selected with proposed method variables best describing changes in unemployment rates in the analyzed period, there are no variables describing the economic situation of regions, although,

Table 5. Results of estimation of panel data models with fixed effects describing the influence of GDP on the unemployment rate (LFSUR)

	Model PKB per capita		Model LFSUR 1		Model LFSUR 2	
	Parameter	***p*-value**	**Parameter**	***p*-value**	**Parameter**	***p*-value**
Variable						
Intercept	−127.4350	< 0.0001		< 0.0001	35.8724	< 0.0001
LFSAR	1.7523	0.0002			−0.3941	< 0.0001
INCOMES	51.7180	< 0.0001			−5.9829	< 0.0001
PKB			−0.0232	< 0.0001	0.0657	0.0007
DISHW			−0.1487	< 0.0001		
DENSITY					0.0060	< 0.0001
Coefficient						
LSDV R-squared	0.7714		0.8828		0.8866	
Within R-squared	0.7531		0.4763		0.4932	
Welch test for differing intercepts in regions; test statistics $F(15, 66)$	2.1852	0.0156	7.3196	< 0.0001	1.9416	0.0341
Wooldridge test for autocorrelation in panel data. H_0: No first-order autocorrelation; test statistics $F(1, 15)$	0.0876	0.7713	0.0046	0.9467	3.8385	0.0689

as shown in the first stage, 6 variables from this sphere significantly influenced the unemployment rate. The sign of parameter on PKB was negative which seems to agree with theory concerning the functioning of the labor market in short term — higher level of economic development — lower unemployment rate. However, all these variables were eliminated in the second stage. This was because they were strongly related to other variables that were eventually selected. This strong connection caused that the LFSUR2 model had a phenomenon similar to that described by Hellwig (1977) — the catalysis effect, which caused the sign of parameters to "reverse" on variables describing the economic situation. For example, it is quite common to believe that GDP per capita has a positive effect on the unemployment rate (the higher the GDP, the lower the unemployment rate). In the first stage, this belief was confirmed by the results of estimating the model of the unemployment rate in relation to GDP per capita. But in the second stage, GDP was eliminated. The reason was the strong positive dependence of GDP on LFSAR and INCOMES, whose presence as explanatory variables "reversed" the parameter sign on GDP.

Replacing the variables LFSAR and INCOMES with GDP in the best model makes the parameter sign on GDP in line with expectations. A similar situation occurs in the case of other variables describing economic development (INV, FAssets) Thus, although the proposed method is based solely on statistical relationships between individual potential explanatory variables, it also facilitates the tracking of causal relationships.

Conclusions

The number of analyses conducted on the basis of panel linear regression models is growing, which results from their advantage over models estimated on the basis of time series or cross-sectional series (Dańska-Borsiak 2011, 13). However, even the best estimation methods will not allow satisfactory results to be obtained if we introduce the wrong explanatory variables into the models. That is why the selection of variables for models is so important.

The article presents the procedure for selecting variables for panel static models (with fixed and random specific effects). A similar procedure can be developed for dynamic models, but this will require the use of appropriate estimation methods. Currently, the most commonly used is the General Moments Method of Arellano and Bond (1991) and its proper tests (Sargan test for exogeneity of instruments, F test for weak/irrelevant instruments, the Arellano-Bond test for autocorrelations of the disturbances of the first and second order). GMM solves the problem of endogeneity, which quite often occurs in complex economic phenomena.

The proposed method is based solely on statistical procedures and does not take into account the substantive characteristics of individual variables. Therefore, it must be preceded by a substantive choice based on a very good knowledge of the analyzed phenomenon. The proposed method, on the other hand, allows to select from among the potential explanatory variables affecting the dependent variable those that will provide the construction of the best linear regression model due to the quality of the estimated parameters and their interpretative properties.

References

Antczak, E., E. Gałecka-Burdziak, and R. Pater. 2018. "Unemployment and Vacancy Flows in Spatial Labour Market Matching at the Regional Level. The Case of a Transition Country." *Journal of Applied Economics* 21 (1):25–43. doi: 10.1080/15140326.2018.1526874.

Arellano, M., and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58 (2):277–297. doi: 10.2307/2297968.

Biørn, E. 2017. *Econometrics of Panel Data. Methods and Applications*. Oxford: Oxford University Press.

Borsuk, M., and K. Kostrzewa. 2020. "Miary ryzyka systemowego dla Polski. Jak ryzyko systemowe wpływa na akcję kredytową banków?" *Bank i Kredyt* 51 (3):211–238.

Charemza, W., and D. Deadman. 1997. *Nowa ekonometria*. Translated by E.M. Syczewska. Warszawa: Polskie Wydawnictwo Ekonomiczne.

Dańska-Borsiak, B. 2011. *Dynamiczne modele panelowe w badaniach ekonomicznych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Driver, C., A. Grosman, and P. Scaramozzino. 2020. "Dividend Policy and Investor Pressure." *Economic Modelling* 89:559–576. doi: 10.1016/j.econmod.2019.11.016.

Grabiński, T., S. Wydymus, and A. Zeliaś. 1982. *Metody doboru zmiennych w modelach ekonometrycznych*. Warszawa: Państwowe Wydawnictwo Naukowe.

Hellwig, Z. 1969. "Problem optymalnego wyboru predykant." *Przegląd Statystyczny* 16 (3/4): 221–237.

Hellwig, Z. 1976. "Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne." *Przegląd Statystyczny* 23 (1):3–20.

Hellwig, Z. 1977. "Efekt katalizy w modelu ekonometrycznym, jego wykrywanie i usuwanie." *Przegląd Statystyczny* 24 (2):179–191.

Herman, S. 2019. "Impact of Joint-Stock Companies' Financial Condition on Real Activities Manipulation to Manage Earnings." *Wiadomości Statystyczne* 64 (10):36–52.

Hsiao, C. 2014. *Analysis of Panel Data*. 3rd ed. Econometric society monographs. New York, NY: Cambridge University Press.

Karkowska, R. 2019. "Systemic Risk Affected by Country Level Development. The Case of the European Banking Sector." *Argumenta Oeconomica* 2 (43):255–282. doi: 10.15611/aoe.2019.2.11.

Kowerski, M., and J. Bielak. 2021. "Kilka uwag na temat pomiaru zależności pomiędzy zmiennymi o panelowej strukturze danych." *Wiadomości Statystyczne* 66 (5):7–25. doi: 10.5604/01.3001.0014.8781.

Kufel, T. 2011. *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*. 3rd ed. Warszawa: Wydawnictwo Naukowe PWN.

Maddala, G.S. 2006. *Ekonometria*. Translated by M. Gruszczyński, E. Tomczyk and B. Witkowski. Warszawa: Wydawnictwo Naukowe PWN.

Nowak, E. 2002. Zarys *metod ekonometrii. Zbiór zadań*. 3rd ed. Warszawa: Wydawnictwo Naukowe PWN.

Pluskota, A. 2020. "The Impact of Corruption on Economic Growth and Innovation in an Economy in Developed European Countries." *Annales Universitatis Mariae Curie-Skłodowska. Sectio H, Oeconomia* 54 (2):77–87. doi: 10.17951/h.2020.54.2.77-87.

Tokarski, T., S. Roszkowska, and P. Gajewski. 2005. "Regionalne zróżnicowanie łącznej produktywności czynników produkcji w Polsce." *Ekonomista* (2):215–244.

Witkowski, B. 2012. "Modele danych panelowych." In *Mikroekonometria. Modele i metody analizy danych indywidualnych*, edited by M. Gruszczyński, 267–308. Warszawa: Wolters Kluwer Polska.