Regression Analysis of Time Series vs Cross Section Data

Vasyl Gorbachuk[1], Aydin Gasanov[2]
[1]Glushkov Cybernetics Institute, National Academy of Sciences of Ukraine, Kyiv, Ukraine
[2]Open Internatioanl University "Ukraine", Kyiv, Ukraine
[1]GorbachukVasyl@netscape.net, [2]AydinG@rambler.ru

***Abstract* — Some important classical regression properties for time series data are more restrictive than similar properties for cross section data.**

Keywords — regression; time series; cross section; data processing; testing; estimation.

I. INTRODUCTION

Contrary to cross section data, time series observations are ordered in time: the value of employment (minimum wage, inflation) in a given country at the year t can depend on that indicator at the previous years $t-1$, $t-2,\ldots$, 1 [1–3]. The statistical properties of OLS estimators (estimators of ordinary least squares, OLS) as random variables are based on the assumption that samples are randomly drawn from the appropriate population. Because different random samples contain generally different values of dependent and independent variables (income, wage, education level, work experience), the OLS estimators, computed on various samples, would generally differ [4–6].

As the value of Dow Jones Industrial Average at the end of trading day on April 12, 2013, or the value of gross domestic product of Ukraine in 2013 are not foreknown, then those variables may be viewed as random variables. A sequence of random variables ordered in time is called a stochastic (random) process or a time series process. The dataset of time series is the realization (a possible outcome) of random process [7].

II. THE REQUIRED PROPERTIES

The standard OLS conditions for cross section data, revised for finite (small) samples of time series (TS), assume the following properties [7]:

TS1) model linearity in parameters;

TS2) zero conditional mean of error variable;

TS3) no perfect collinearity among variables;

TS4) homoskedasticity of errors;

TS5) no serial correlation among errors;

TS6) normality of errors.

The properties TS1)–TS3) give unbiasedness of OLS estimators of parameters; the properties TS4) and TS5) are necessary for computing the variances of OLS estimators; the property TS6) is necessary for statistical inference. The properties TS1)–TS5) are called the Gauss–Markov assumptions.

The property TS1) means the stochastic process $\{x_{t1}, x_{2t}, \ldots, x_{tk}, y_t\}_{t=1}^{n}$ is described by some model linear in $\{\beta_i\}_{i=0}^{k}$

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t , \quad (1)$$

where: n is the number of observations (time periods); y_t is the value of explained (dependent) variable (regressand) at (time) period t; $\{u_t\}_{t=1}^{n}$ is the sequence of measurement errors (disturbances); x_{ti} is the value of explanatory (independent) variable (regressor) x_i at period t, $i = 1,2,\ldots,k$; β_i is estimated parameter; $i = 0,1,\ldots,k$.

The model, accounting for the history

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t , \quad (2)$$

where y_t is the general fertility rate (number of newborn per 1000 women of childbearing age) at period t, z_t is the real monetary value (of the right ensured by law) of personal tax exemption (the tax value of giving a child) at period t, takes into account behavioral motivations and biological circumstances of decisions to have a child. The model (2) us reduced to the model (1) if $x_{ti} = z_{t-i}$, $i = 0,1,2$.

The property TS2) requires that at any period the average value of error does not depend on the independent variables:

$$E(u_t \mid X) = 0 \quad \forall t = 1,2,\ldots,n , \quad (3)$$

where X is the array consisting of n rows $(x_{\tau 1}, x_{\tau 2}, \ldots, x_{\tau k})$ and corresponding k columns, $\tau = 1,2,\ldots,n$. This property means the error u_t at period t is uncorrelated with each explanatory variable x_i, $i = 1,2,\ldots,k$, at every period $\tau = 1,2,\ldots,n$. If u_t does not depend on X and $E(u_t) = 0$, then the property TS2) holds automatically.

At $\tau = t$ the relationship (3) implies the known for cross section data condition

$$E(u_t \mid x_{t1}, x_{t2}, \ldots, x_{tk}) = 0 .$$

If this condition does not hold, then each variable x_{ti} is called contemporaneously exogenous, $i = 1, 2, \ldots, k$. This condition implies that u_t is uncorrelated with each explanatory variable x_{ti} :

$$Corr(u_t, x_{ti}) = 0 , \; i = 1, 2, \ldots, k ; \; t = 1, 2, \ldots, n . \qquad (4)$$

The property (3) requires more than the equalities (4):

$$Corr(u_t, x_{si}) = 0 ;$$

$$i = 1, 2, \ldots, k ; \; t = 1, 2, \ldots, n ; \; s = 1, 2, \ldots, n . \qquad (5)$$

If the condition (5) does not hold, then the explanatory variables x_i are called strictly exogenous, $i = 1, 2, \ldots, k$. The condition (4) is sufficient for proving consistency of the OLS estimators. The condition (5) is used for proving unbiasedness of the OLS estimators.

For cross section demographic data, an explicit relation between the error u_t of person a and the explanatory variable of another person b in the sample is not stated because that error may refer to the person a only and automatically does not depend on the explanatory variables of other persons, under random sampling (the standard OLS condition for cross section data). For time series data, random sampling is not required while the property TS2) is needed.

The property TS2) is not satisfied when the unobservables (omitted variables and measurement errors in any regressors) correlate with some regressors at a certain period. In the simple static regression model

$$y_t = \beta_0 + \beta_1 z_t + u_t \qquad (6)$$

the property TS2) postulates zero correlation of u_t with past and future values of independent variable z :

$$Corr(u_t, z_s) = 0 ; \; t = 1, 2, \ldots, n ; \; s = 1, 2, \ldots, n .$$

Therefore the assumption TS2) rules out any lagged influence from z to y (if such an influence exists, then a distributed lag model should be estimated). Besides, the assumption TS2) excludes a possibility of influence from current changes in error u to future changes in explanatory variable z , which in fact forbids any feedback from contemporary value of y to future values of z .

If in the model (6) y_t is the murder rate (the number of murders per 10000 people) in a given city at the year t , z_t is the number of policemen per capita in that city at the year t , then it may be supposed that u_t does not correlated with z_t , $z_{t-1}, \ldots, z_1$. At the same time it may be assumed the city changes the values of z , based upon the past values of y . As larger error u_t is associated with larger level of y_t , then u_t may correlate with z_{t+1} , violating the condition TS2). For a distributed lag model, generalized the model (6), the similar considerations on violation the condition TS2) are remained. Contrary to possible correlation between u_t and z_{t+1} , $z_{t+2}, \ldots, z_n$, a possible correlation between u_t and z_{t-1} , $z_{t-2}, \ldots, z_1$ is under control.

The strictly exogenous explanatory variable z does not react to the past of explained variable y : for instance, the amount of rainfall at any future year is not associated with the crops at current or past years. At the same time the labor input is chosen by the farmer who may account for the crop at previous year. Such policy variables, as expenditures on public welfare, growth in money supply, highway speed limits, are often influenced by the past values of explained variable.

If explanatory variables are nonrandom or fixed in repeated samples, then the property TS2) holds true. However, in the time series observations, explanatory variables should be random.

The property TS3) assumes in the sample (and therefore in the underlying time series process) no independent variable is constant or a perfect linear combination of other independent variables. As the similar property for cross section data, the property TS3) allows correlation of explanatory variables but not perfect correlation in the sample.

Theorem 1. If the conditions TS1), TS2), TS3) hold true, then the OLS estimators $\hat{\beta}_i$ are unbiased:

$$E(\hat{\beta}_i \mid X) = E(\hat{\beta}_i) = \beta_i , \; i = 0, 1, \ldots, k .$$

The proof of theorem 1 reproduces the proof of corresponding theorem for cross section data. The analysis of biasedness due to omitted variables reproduces the corresponding analysis for cross section data as well.

The property TS3) in the truncated model (2) with finite distributed lags

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t \qquad (7)$$

and independent variables $x_{t1} = z_t$, $x_{t2} = z_{t-1}$ rules out in the sample x_{t1} is constant (the values of z_1 , $z_2, \ldots, z_n$ are the same) or x_{t2} is constant (the values of z_0 , $z_1, \ldots, z_{n-1}$ are the same) and also excludes in the sample x_{t1} is a perfect linear combination of x_{t2} (if $z_t = a + bt$, then $z_{t-1} = a + b(t-1) = a + bt - b = z_t - b$ is an exact linear function of z_t).

The property TS4) posits the variance $Var(u_t \mid X)$ does not depend on X and equals to $Var(u_t)$, and $Var(u_t)$ does not depend on $t = 1, 2, \ldots, n$ and equals to σ^2

(homoskedasticity of errors). If the property TS4) does not hold, then errors u_t are heteroskedastic. Under some conditions, testing for heteroskedasticity for cross section data is transferred to time series data.

The property TS4) for the truncated model (1)

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t , \quad (8)$$

where y_t is the average rate of 3-month treasury bills (at the year t in the USA), x_{1t} is inflation level (%), x_{2t} is the ratio (%) of budget deficit to gross domestic product (GDP), stands for the unobservables, affecting y_t, have a constant variance over time. As the changes of economic policy variables influence on the variability of interest rates, then the property TS4) may be violated. The property TS4) will be violated if the interest rate volatility depends on x_{1t} or x_{2t}.

The property TS5) claims no conditional serial correlation between errors at any different time periods:

$$Corr(u_t, u_s \mid X) = 0 \quad \forall t \neq s .$$

If the array X is believed to be nonrandom, then the property TS5) has a form

$$Corr(u_t, u_s) = 0 \quad \forall t \neq s . \quad (9)$$

When the relationship (9) is not satisfied, then errors u_t undergo serial correlation (autocorrelation). If the error u_t is positive (the value of y_t in model (8) is high) and the average of next period error u_{t+1} is also positive (the value of y_{t+1} is also high), then $Corr(u_t, u_{t+1}) > 0$ and the relationship (9) is not satisfied.

The property TS5) does not eliminate temporal correlation for independent variable x_{1t} or independent variable x_{2t}.

For cross section data the condition on absence of conditional serial correlation between errors u_t, u_h at any different time periods t, h, because the random sampling assumption implies independence of errors u_t, u_h.

Sometimes the Gauss–Markov properties TS1)–TS5) are satisfied for cross section applications, for which the random sampling assumption does not hold, if sample sizes are large in comparison with the population size. The property TS5) is not satisfied if t, s in the relationship (8) are interpreted as cities instead of time periods.

Theorem 2. Under the Gauss–Markov assumptions TS1)–TS5), the following equality takes place:

$$Var(\hat{\beta}_i \mid X) = \frac{\sigma^2}{SST_i (1 - R_i^2)} , \; i = 1, 2, \ldots, k ,$$

where: SST_i is the total sum of squares $\sum_{t=1}^{n} (x_{ti} - \hat{x}_{ti})^2$ at explanatory variable x_i; R_i^2 is the R-squared from regression of x_i on the other independent variables x_1, $x_2, \ldots$, x_{i-1}, $x_{i+1}, \ldots$, x_k.

The proof of theorem 2 reproduces the corresponding proof for cross section data. Time series data and cross section data have the same the reasons of large variances, including multicollinearity between independent variables.

In the model (7) with finite distributed lags, the multicollinearity between explanatory variables z_t, z_{t-1} might be a result of the nature of those variables: if $\{z_t\}$ is a sequence of unemployment levels, then those levels are changing slowly.

Theorem 3. Under the Gauss–Markov assumptions TS1)–TS5), the estimator

$$\hat{\sigma}^2 = \frac{SSR}{df}$$

is an unbiased estimator of σ^2, where the degress of freedom $df = n - k - 1$.

Theorem 4. Under the Gauss–Markov assumptions TS1)–TS5), the OLS estimators are the best linear unbiased estimators (BLUEs) conditional on X.

The theorems 2–4 transfer the desirable properties of multiple linear regression from cross section data to finite samples of time series data.

The property TS6) supposes that errors u_t are independent of X, are independently and identically distributed with mean 0 and variance σ^2 (belong to the class $N(0, \sigma^2)$).

The property TS6) implies the properties TS3)–TS5).

Theorem 5. Under the assumptions TS1), TS2), TS6), the condition central limit theorem, the OLS-estimators are normally distributed random variables conditional on X.For the OLS-estimators, each t-statistic has a t-distribution, and each F-statistic has F-distribution. Hence, the usual for cross section data construction of confidence intervals are valid for time series data.

According to the theorem 5, estimating and hypothesis testing for cross section regressions are directly applied to time series regressions: t-statistic may be used for testing of statistical significance of individual explanatory variables, and F-statistic – for testing of their joint significance. Meanwhile, the assumptions TS2) and TS5) of classical linear model for time series data are more restrictive than the similar assumptions for cross section data.

REFERENCES

[1] V. M. Gorbachuk, Econometric Programming of TSP and Eviews, Preprint 96-14, Kyiv: Glushkov Cybernetics Institute, Academy of Sciences of Ukraine, 1996. (In Ukrainian).

[2] V. Gorbachuk, Macroeconomic Methods, Kyiv: Alterpress, 1999. (In Ukrainian).

[3] V. M. Gorbachuk, Methods of Industrial Organization, Kyiv: A. S. K., 2010. (In Ukrainian).

[4] V. M. Gorbachuk, "Regression analysis of time series and Granger causality" , The 14-th International Scientific Kravchuk Conference, vol. 3, Kyiv: NTUU "Kyiiv Polytechnic Institute", 2012, p. 11.

[5] V. Gorbachuk, "Causality in time series analysis", Nonlinear analysis and applications, Kyiv: NTUU "Kyiv Polytechnic Institute", 2012, p. 30.

[6] V. M. Gorbachuk, Y. G. Krivonos, "The features of regression analysis for time series", Computer Mathematics, 2012, № 2, p. 3–12. (In Russian).

[7] J. M. Wooldridge. Introductory Econometrics: a Modern Approach. 4-th ed., Mason, OH: Cencage Learning, 2009.