



Pooling Cross Section and Time Series Data

Author(s): Nicholas N. N. Nuamah

Source: *The Statistician*, Vol. 35, No. 3 (1986), pp. 345-351

Published by: Blackwell Publishing for the Royal Statistical Society

Stable URL: <http://www.jstor.org/stable/2987750>

Accessed: 13/05/2009 16:42

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=black>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.



Royal Statistical Society and Blackwell Publishing are collaborating with JSTOR to digitize, preserve and extend access to *The Statistician*.

<http://www.jstor.org>

Pooling cross section and time series data

NICHOLAS N. N. NUAMAH

Department of Statistics, Institute of National Economy, Cherniahovskova 22/1, Odessa 56, U.S.S.R.

Abstract. Pooling cross section and time series data has become necessary in recent times and this has occupied the minds of many statisticians. Different methods such as the Error Component Model, the Least Squares, the Covariance Analysis have been used in the pooling and estimation of all observations.

Krastin (1981) analysed the properties of the regression estimates of a simple equation on cross section data over a period of 2 years by one of the Ordinary Least Square methods and came to a very important conclusion that although the estimate is biased it is not due to the influence of autocorrelation and drift correlation but due to the covariance of means.

Here we consider only one of the Ordinary Least Square (OLS) methods and the covariance analysis. We analyse the circumstances under which the covariance of means can result in the biasness of the regression coefficients obtained by the OLS method.

1 Introduction

The bread of any statistician is data. Depending on the basic information available he defines the aim of the model to be built, the method of building it and the analytical use of it. One of the major requirements of building a good model is that observations should be sufficiently large. Some statisticians suggest that in building a cross section model the number of individual units, say firms or household, should be 6–8 times greater than the number of the independent variables, others suggest 5–6 times.

In practice, in economics, such conditions are not often satisfied. So very often it is exceptionally important to make the most efficient use of the few observations that are available for different individuals over time to estimate that part of the relationship containing variables. Statisticians have recently tried to search for a method of combining cross section and time series data in economic statistical modelling. The process of this combination is known as pooling.

In building model based on pooled data we try to solve two major problems in correlation and regression analysis: firstly, the expansion of the investigated population with the aim of reducing the effect of stochastic fluctuations thus raising the stability of the postulated models and their reliability in solving different economic problems. The use of pooled data with the aim of increasing sampling data is justified if, and only if, the character of the relationship under study does not change from time to time; secondly, the expansion of the area of application of the obtained statistical model.

However, the use of such data adds a new dimension of difficulty to the problem of model specification because the disturbance term is likely to consist of time-series-related disturbances, cross-section disturbances and a combination of both.

2 The model

Supposing data is available on N firms over T periods of time. Let there be a stochastic

relationship between some dependent variable Y_{ij} and some independent variable X_{gj} and unobserved random variables ε_{gij} ($i=1,2,\dots,N_j; j=1,2,\dots,T; g=1,2,\dots,m$). A mathematical model may be written as:

$$Y = X\beta + \varepsilon \tag{2.1}$$

where Y is a $N_j \times 1$ vector,

$X - N_j \times m$ matrix of full rank,

$\beta - m \times 1$ vector and,

$\varepsilon - N_j \times 1$ vector,

$N = N_j T$.

Today there are several methods of building models based on the data presented in (2.1) differing amongst themselves both in their widespread use and in their scientific substance. The basis for their classification may be on the method of pooling the data into one cell. The pooling could be done on different stages of transformation of the normal equation and its solution: in one case; before applying the least squares method on the relationship under study; in another case while applying the least squares method; and, in the third case, after the application of the least squares method.

In considering the method to use in pooling the data for modelling, some questions arise and have to be borne in mind.

Firstly, to what extent does the result of the model of the pooled data,

$$\bar{y} = \bar{b}_0 + \bar{b}_1 X_1 + \dots + \bar{b}_m \bar{X}_m$$

corresponds to the true models, that is the ones which would have been got for each year,

$$Y_T = b_{0T} + b_{1T} X_1 + \dots + b_{mT} X_m$$

Secondly, which method enables us to determine the existence of structural changes and also enables us to consider these changes in the model?

Thirdly, which method could be applied to the pooled data so as to obtain a non-biased estimate of the general model?

Fourthly, which statistical paradox arises in those methods which don't provide non-biased estimates. Is the paradox the result of stochastic variable or is it the characteristics of a given process?

Fifthly, if the statistical paradox is the characteristics of a given process then in which conditions do they arise and in which they do not?

Lastly, but not the least, how best can the parameters of the model be estimated so that the model could be used for statistical analysis and forecasting?

The economic interpretation and the area of application of any method used in pooling data depends on some assumptions.

Consider the following hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_T = \beta \tag{2.2}$$

In this case let us rewrite equation (2.1) in general form as

$$y = \beta_0 + \beta_1 X_{ij} + a_j + \gamma_j X_{ij} + \varepsilon_{ij}$$

or

$$y = (\beta_0 + a_j) + (\beta_1 + \gamma_j) X_{ij} + \varepsilon_{ij}$$

where a_j are the time effects.

A test of whether such a change in β_1 is significant is provided by a test of the null hypothesis that the coefficient γ_1 is 0. The above assumption states that the coefficient

vectors are fixed and the period under study are homogeneous in so far as the β_j 's are concerned.

Note that this assumption is very important in that it is the basis upon which pooling is justified. If it is not true, then the data on all periods cannot be pooled to estimate a single relationship between variables and we might as well fit T different lines:

$$\begin{aligned} y_1 &= \beta_{01} + \beta_{11}X_1 + e_1 \\ y_2 &= \beta_{02} + \beta_{12}X_2 + e_2 \\ &\vdots \\ y_j &= \beta_{0j} + \beta_{1j}X_j + e_j \\ &\vdots \\ y_T &= \beta_{0T} + \beta_{1T}X_T + e_T \end{aligned}$$

where β_{0T} is the intercept for T th year.

3 The methods of estimation

Let's assume that the hypothesis (2.2) holds then

$$y = (\beta_0 + a_j) + \beta_1 X_j + \varepsilon_j \tag{3.1}$$

That is to say, that the disturbing elements affect only the level and not the form of the functional relationship so that the constant terms may vary from time to time.

In equation (3.1), a test of whether such a change is statistically significant is provided by a test of the null hypothesis that $a=0$ or a test for the hypothesis

$$H_0^*: a_1 = a_2 = \dots = a_T = a \tag{3.2}$$

Case I

If the hypothesis (3.2) is true, then this case corresponds to the situation in which both the slope and the intercept are allowed to remain constant within the period under study.

In this case least-square analysis leads to the use of the usual pooled estimate of β from the j th groups.

We may rewrite equation (2.1) as

$$Y = X\beta + \varepsilon \tag{3.3}$$

where:

Y is a $N \times 1$ vector, X - $N \times m$ matrix and ε a $N \times 1$ vector. Equation (3.3) practically means that breaking the data into j groups (i.e. treating the data for different years separately) is insignificant, that is to say that the variation of the variable Y is efficiently explained by the variations of the variable X . The normal equation $X'X\beta = X'Y$ is the usual form of regression by the OLS with N observations and $N - m - 1$ degrees of freedom.

Without any loss of generality let's assume that X is a vector of $N \times 1$ and that the data is balanced (i.e. $N_1 = N_2 = \dots = N_T$) and also that all variables are measured as deviations from their respective overall means.

The postulated model becomes

$$y_{.j} = \hat{\mu}_{..} + \beta(X_{.j} - \bar{X}_{..}) + \varepsilon_{.j} \tag{3.4}$$

where

$$\hat{\mu}_{..} = \bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{N_j} \sum_{j=1}^T y_{ij}; \quad \bar{X}_{..} = \frac{1}{N} \sum_{i=1}^{N_j} \sum_{j=1}^T X_{ij}$$

The residual of the OLS for model (3.4) will be

$$e_{.j} = (y_{.j} - \bar{y}_{..}) - b(X_{.j} - \bar{X}_{..}) \tag{3.5}$$

and the sum of the square error equals

$$T^* = \sum \sum e_{.j}^2 = \sum_{j=1}^T \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_{..})^2 - 2b \sum_{j=1}^T \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_{..})(X_{ij} - \bar{X}_{..}) + b^2 \sum_{j=1}^T \sum_{i=1}^{N_j} (X_{ij} - \bar{X}_{..})^2$$

Taking the partial derivatives of T^* with respect to the parameter estimate b and setting equal to zero, we get

$$b = T^*_{xy} / T^*_{xx} \tag{3.6}$$

where

$$T^*_{xy} = \sum_{i=1}^N (y_{.j} - \bar{y}_{..})(X_{.j} - \bar{X}_{..}), \text{ and } T^*_{xx} = \sum_{i=1}^N (X_{.j} - \bar{X}_{..})^2.$$

Regression coefficient b in equation (3.6) is simply the gradient of the regression line of Y against X for the whole of the pooled data. It is therefore sometimes called the common regression slope of Y on X .

Case II

We consider the case where the hypothesis (3.2) is not true and that

$$a_j = a + \delta_j \tag{3.7}$$

but let's assume that δ_j is not correlated with $\varepsilon_{.j}$ i.e.

$$E(\varepsilon_{.j}, \delta_j) = 0.$$

If we had used the OLS on the data, we would have got the regression coefficient as in (3.6).

Considering a cross section data over a period of 2 years and using algebraic transformation it can be seen that formula (3.6) becomes

$$b = \frac{\text{Cov}_{x_1y_1} + \text{Cov}_{x_2y_2} + 0, 5 \text{ Covav}_{x_2y_2}}{S^2_{x_1} + S^2_{x_2} + 0, 5 \text{ Covav}_{x_2}^2} \tag{3.8}$$

where

$$\text{Cov}_{x_jy_j} = \frac{\sum_{i=1}^{N_j} (X_{ij} - \bar{X}_{.j})(y_{ij} - \bar{y}_{.j})}{N_j}; \quad S^2_{x_j} = \frac{\sum_{i=1}^{N_j} (X_{ij} - \bar{X}_{.j})^2}{N_j}$$

$$\text{Covav}_{x_2y_2} = \frac{\sum_{i=1}^{N_j} (\bar{X}_{.1} - \bar{X}_{.2})(\bar{y}_{.1} - \bar{y}_{.2})}{N_j} = (\bar{X}_{.1} - \bar{X}_{.2})(\bar{y}_{.1} - \bar{y}_{.2})$$

$$\text{Covav}_{x_2}^2 = \frac{\sum_{i=1}^{N_j} (\bar{X}_{.1} - \bar{X}_{.2})^2}{N_j}$$

and

$$\bar{X}_{.j} = \frac{\sum_{i=1}^{N_j} X_{ij}}{N_j}; \bar{y}_{.j} = \frac{\sum_{i=1}^{N_j} y_{ij}}{N_j}$$

It is easy to notice from formula (3.8) that the regression coefficient depends not only on the nature of the relationship of the variables in question for each year

(1) $\text{Cov}_{x_jy_j}; S_{x_j}^2$

but also on the changes in time, of the means of the interrelated variables, that is, on the covariance means

$$\text{Covav}_{x_1y_1}; \text{Covav}_{x_2}^2$$

(2) If $\bar{y}_{.1} = \bar{y}_{.2}$ but $\bar{x}_{.1} \neq \bar{x}_{.2}$ then from formula (3.8) $\text{Covav}_{x_1y_1}$ falls out and the regression coefficient will be biased in a direction of sharp decrease.

(3) If $\bar{y}_{.1} \neq \bar{y}_{.2}$ but $\bar{x}_{.1} = \bar{x}_{.2}$ then the values of $\text{Covav}_{x_1y_1}$ and $\text{Covav}_{x_2}^2$ vanish and formula (3.8) reduces to the analysis of covariance

(4) If $\bar{y}_{.1} = \bar{y}_{.2}$ and $\bar{x}_{.1} = \bar{x}_{.2}$.

The result is as in (4) above

(5) When $\bar{y}_{.1} \neq \bar{y}_{.2}$ and $\bar{x}_{.1} \neq \bar{x}_{.2}$, then we can distinguish 4 cases.

(a) $\bar{x}_{.1} - \bar{x}_{.2} > 0$, but $\bar{y}_{.1} - \bar{y}_{.2} < 0$

(b) $\bar{x}_{.1} - \bar{x}_{.2} < 0$, but $\bar{y}_{.1} - \bar{y}_{.2} > 0$

(c) $\bar{x}_{.1} - \bar{x}_{.2} > 0$, and $\bar{y}_{.1} - \bar{y}_{.2} > 0$

(d) $\bar{x}_{.1} - \bar{x}_{.2} < 0$ and $\bar{y}_{.1} - \bar{y}_{.2} < 0$

In cases (a) and (b) the means of the variables change in an opposite direction. The regression coefficient will be biased in the direction of decrease. In case (c) and (d) the means of the variables change in the same direction. The regression coefficient will be biased in the direction of an increase if the change of the dependent variable is greater than that of the independent variable. If, on the other hand, the change of the dependent variable is smaller than the change of the independent variable bias is expected to be in the direction of a decrease.

One of the most important methods of pooling cross section data on all periods to

give more efficient estimates of a single relationship between variables in the case II is by using the analysis of covariance (ANCOVA).

Cochran (1957), in discussing the principal uses of ANCOVA considered its use in fitting regressions in multiple classifications (i) fit a separate regression of Y on X within each class (ii) test whether the slope or positions of the lines differ from class to class, and (iii) if advisable, make a combined estimate of a common slope.

By its logical properties ANCOVA combines the properties of both variance and regression analysis. The analysis of variance as we know, is meant for the study of the influence of one or more qualitative (attributive) factor; regression analysis—for studying the influence of quantitative factors on Y . With the help of ANCOVA we try to study the joint influence of both the qualitative and quantitative factors.

Generally the mathematical model for ANCOVA is

$$y_{ij} = \mu_{..} + \tau_i + a_j + \lambda_{ij} + \sum_{g=1}^m b_g(X_{gij} - \bar{X}_{g..}) + \varepsilon_{ij} \quad (3.9)$$

where τ_i cross section effect; a_j —time effect and λ_{ij} -interaction effect of both τ_i and a_j . This is the nonadditive case.

If we make covariance analysis assumption that each source of variation (treatments and blocks for the random block design) introduces an additive effect for the covariate as well as the response Y then λ_{ij} drops from equation (3.9).

Sometimes, depending on the type of data available for analysis we consider the covariance analysis with only either τ_i or a_j included in the postulated model.

From the foregoing we can see how ANCOVA models permit us to detect explicitly the principally important components—the time effect, the firm effect and the interaction effect—which the regression analysis (OLS) is not capable of.

4 An example

We present here models based on an artificial data and obtained by the covariance analysis and regression analysis presented in the previous section.

The number of the individual units were taken to be 10 and time—to be 5 years where Y is made to be influenced by three factors X_1, X_2, X_3 . Let's consider here the case where only time effect is included in the covariance model.

The models for the years are:

$$\hat{y}_1 = 0,50802 + 0,83799x_1 + 2,91539x_2 + 0,54332x_3 \quad (4.1)$$

$$\hat{y}_2 = 0,64911 + 0,83219x_1 + 0,98442x_2 + 1,79970x_3 \quad (4.2)$$

$$\hat{y}_3 = 4,20456 - 2,07390x_1 + 1,83675x_2 - 0,58856x_3 \quad (4.3)$$

$$\hat{y}_4 = 0,89414 - 1,34709x_1 + 7,82014x_2 + 0,29846x_3 \quad (4.4)$$

$$\hat{y}_5 = 2,38612 - 1,54600x_1 + 3,97959x_2 + 0,89098x_3 \quad (4.5)$$

When the yearly regression coefficients are averaged (excluding the constant term)' the general model becomes

$$\hat{y} = 1,54446 - 0,65285x_1 + 3,50726x_2 + 0,58878x_3 \quad (4.6)$$

The model obtained by applying OLS (the regression analysis) is

$$\hat{Y}_{(OLS)} = -0,16458 + 0,53469x_1 + 4,38508x_2 + 0,85310x_3 \quad (4.7)$$

and the ANCOVA:

$$\hat{y} = 1,55289 - 0,50689x_1 + 3,37488x_2 + 0,48897x_3 \quad (4.8)$$

As we can see, the model (4.7) is highly biased. When b_{gj} for the last 3 years of equations (4.1)–(4.5) are negative, in equation (4.7) it is positive. As for the ANCOVA it corresponds well to the annual models and to the mean of the respective coefficients (compare (4.6) and (4.8)). If we look at the intercept in equation (4.7) and compare it with the intercepts of the five equations for each year we establish a statistical paradox. Though there is no single negative intercept in (4.1)–(4.5), the intercept in (4.7) is negative.

This example confirms that the OLS may not be an effective method of pooling, estimating and analysing panel data especially when the character of the data is not checked, that is when we are not convinced that there is no time effect.

Conclusion

In a situation when $E(\bar{x}_j) = E(\bar{x}_T)$ the covariance adjustment is not needed to remove bias due to X because there is no such bias. The OLS could be used to give an efficient estimate of the parameter.

The OLS method is simpler than the ANCOVA in pooling cross section and the time series data when its conditions are satisfied but care must be taken when it is being applied.

When $E(\bar{x}_j) \neq E(\bar{x}_T)$, the only alternative to the above method is the ANCOVA. It could be used even when conditions for OLS are satisfied because it might still improve the precision of the model.

The models obtained by using the ANCOVA have more stable, unbiased and economic interpreted coefficient than the model calculated by the OLS.

References

- BRIGMANE, A. (1977) Srvneniye metodik pastroyeniya funktsii urayanosti pa mnogoletnik dannix, *Matematicheskiye metodi v economickye*, 18, pp. 55–74 (Riga, Zinatne).
 COCHRAN, W.G. (1957) Analysis of covariance: its nature and uses, *Biometrics*, 13, pp. 261–281.
 KRASTIN, O.P. (1981) Isucheniye Statisticheskix Zavisimostei pa mnogeletnix dannix (Moskova, Finansy i Statistika).
 NERLOVE, M. (1971) A note on Error Components Models, *Econometrica*, 39, pp. 383–396.

NOTE

1. The constant is calculated as follows:

$$b_0 = \bar{y}_{..} - b_1\bar{x}_{1..} - \dots - b_m\bar{x}_{m..}$$

though sometimes it is calculated as an arithmetic mean of b_{gj}