A Brief Review of Statistical Models in Budgetary Studies

Chung-Kil Chung, Ph.D.*

〈Contents〉

I. Statistical Methods in Budgetary Studies

Budgetary studies include more than quantitative data analysis. Actually quantitative analysis was at the periphery of budgetary research till a decade ago among political scientists, if not among economists. Therefore, a focus on the technical problems of data analysis in assessing the methodological quality of budgetary studies may be misleading. It is nevertheless important to summarize major deficiencies of data-analysis techniques used in the budgetary studies. And this is what this small paper tries to do.

Among those problems, I picked up the problems in statistical models which were most frequently used: the violation of basic assumptions in those models which were naively applied by researchers.

Quantitative analysis of budgetary data can be traced back to German Cameralists. But statistical approach (that is inferential statistics) began in 1950's when Fabricant's famous study on the determinants of state expenditure first appeared.(1) Another seminal study by H. Brazer(2) in the same period triggered tremendous amount of cross-sectional analysis by economists.

Political scientists who usually treated budgetary studies as one part of Executive-Congress relationship to focus on formal budgetary process began to imitate econonmists' approach from 1960's. But this trend of cross-sectional analysis

* Associate Professor, Graduate School of Public Administration, SNU.

(1) Solomon Fabricant: *The Trend of Government Activity in the United States Since 1900.* New York National Bureau of Economic Research, 1952.

(2) Harvey E. Brazer: *City Expenditures in the United States.* New York: National Bureau of Economic Research, 1959.

(3) O.A. Davis, M.A.H. Dempster and A. Wildawsky: "Theory of the Budgetary Process", *American Political Science Review,* Vol. 60, September 1966.

was set back by Wildawsky's incrementalism which was operationalized, with some other's help, into autoregressive time series analysis.[3] Crecine went beyond this by originating simulation model which digs out the internal process of budgeting and connecting this to the time-series data.[4]

The most striking fact about budgetary data analysis is that simple ordinary least squares (OLS) method has been used with little attention to the probable violation of its assumptions. This is particularly surprising in the studies of economists[5] who have developed quantitative model nearly 30 years ago and who have been familiar with econometrics. Maybe the violation of OLS assumptions is not so serious as a mere guess suggests, but more probably there is no remedy as yet to cure this problem. Let's start from cross-sectional analysis.

II. Cross-Sectional Analysis

Cross-sectional analysis of budgetary data has been very popular among economists. The starting point of economists is to calculate elasticity (such as income) of public expenditure. Since Fabricant's study in 1950's, many factors were included as independent variables, especially political factors from 1960's, but the basic model and methods were not changed.[6]

We can find two distinct models which are basically linear. These are:

$$y=\beta_1+\beta_2x_2+\cdots\cdots\cdots\cdots+\beta_kx_k+u$$

$$y=(\beta_0x_2^{\beta_2}x_3^{\beta_3}\cdots x_k^{\beta_k})u$$

the first of which is usually applied. Their method was almost unexceptionally Ordinary Least Square (OLS).

Their dependent variable was per capita public expenditure. And their independent variables were per capita income, degree of population density, degree of unbanization (which were Fabricant's) plus industrialization, voting turn--out, party competition and so on which are slightly different among analysts and among functions of expenditure in focus. Unit of analysis was state or other local government.

1. Ecological Fallacy?

As dependent variable is a systemic factor which represents the whole part of the unit of analysis (e.g. Michigan's total expenditure on High Ways devided by the population), ecological correlation does not matter here.

(4) John P. Crecine: "A Simulation of Municipal Budgeting: The Impact of Problem Environment" in W. Coplin (ed.), *Simulation in the Study of Politics*. Chicago: Markham, 1968. See also John P. Crecine: *Governmental Problem-Solving*, Chicago: Rand McNally, 1969.

(5) See, for example, following two reviews. Roy W. Bahl: "Studies on Determinants of Public Expenditures: A Review" in S.J. Mushkin and J.F. Cotten (eds.), *Functional Federalism: Grants in Aid and PPB Systems*, Washington D.C.: George Washington University Press, 1968. Gail Wilensky: "Determinants of Local Government Expenditures", in J.P. Crecine (ed.), *Financing the Metropolis*, Beverly Hills: Sage Publications, 1970.

(6) The followings are only few among many studies by Political Scientists. Ira Sharkansky: "Economic and Political Correlates of State Government Expenditure", *Midwest Journal of Political Science*, May 1967. Robert C. Fried: "Communism, Urban Budgets and Two Italies: A Case Study in Comparative Urban Budgets", *Journal of Politics*, November 1971. Guy Peters: "Economic and Political Effects on the Development of Social Expenditures in France, Sweden and United Kingdom", *Midwest Journal of Political Science*, May 1972. Andrew T. Cowart: "Partisan Politics and the Budgetary Process in Oslo", *American Journal of Political Science*, November 1975. Harold L. Wilensky: *The Welfare State and Equality*, Berkeley: The University of California Press, 1975. Valerie Bunce: *Changing Leaders and Changing Policies*, Paper delivered at the 1976 Annual Meeting of the American Political Science Association in Chicago.

However independent variables are usually aggregate measures (except some political factors). Moreover, dependent variable which is decided by the activities of a local government as a whole can also be regarded as an aggregate of small parts. Because of these reasons, ecological correlation should be as carefully avoided as data can be datailed. But, to do this requires more than detailed data set: we should know the internal structure of local government to decide whether dependent variable is simply an aggregate or a systemic one. Brazer's analysis shows that within-state variation is bigger than among-states variation in dependent variable.

More crucial problem related to this fact is that aggregate income (or density or urbanization) may not show true parameter (β) which relates expenditure decision to individual income. This problem can be solved partly by introducting Gini coefficient as one of the independent variables.

2. Multi-collinearity

Since industrialization, urbanization and per capita income may go together, correlation among these independent variables will be big. Bahl well summarized problems and implications of this multi-collinearity in budgetary studies.(7) But still many researchers are repeating same mistakes, arguing that their studies show non-significant correlations between dependent variable and some independent variables which were reported as significant by others. The point is, of course, that the variance of certain $\hat{\beta}$ may be big, because

$\mathrm{Var}(\hat{\beta}) = \sigma^2(x'x)^{-1}$, where x represents matrix of x's

and collinearity among Xs increases, the determinant of $(x'x)^{-1}$ becomes bigger.

The crucial problem here is: even though we can prepare mathematical remedy for this by such means as finding out instrumental variable which is related to each variables, we cannot easily pick up such an instrumental variable in budgetary field.

3. Heteroscedastic Disturbances

Many economists as well as political scientists tried OLS in budgetary studies. OLS estimator of parameters are;

$$E(\hat{\beta}) = E[(x'x)^{-1}x'y] = E[(x'x)^{-1}x'(x\beta+u)]$$
$$= E[(x'x)^{-1}x'x\beta] + E[(x'x)^{-1}x'u]$$
$$= \beta + E(x'x)^{-1}x'u \quad \cdots\cdots 1$$

Now, assuming plim $(x'x)^{-1}$ is finite fixed and plim $(x'u)=0$. also $E(u)=0$,

$$E(\hat{\beta}) = \beta + E[(x'x)^{-1}x'u] = \beta \cdots\cdots 2$$

As far as independent variables are uncorrelated with disturbances (u) and $E(u)=0$, OLS estimate of β is unbiased and consistent.

$$\mathrm{Var}(\hat{\beta}) = E(\hat{\beta}-\beta)\ (\hat{\beta}-\beta)' = (x'x)^{-1}x'uu'x\ (x'x)^{-1} \cdots\cdots 3$$

Assuming $E(uu') = \sigma^2 In$ in OLS

$$\mathrm{Var}(\hat{\beta}) = (x'x)^{-1}x'\sigma^2 In\ x(x'x)^{-1}$$
$$= \sigma^2(x'x)^{-1}x'x(x'x)^{-1} = \sigma^2(x'x)^{-1} \cdots 4$$

If $E(uu') = \sigma^2\Omega$, then OLS estimate of

$$\mathrm{Var}\ (\hat{\beta}) = (x'x)^{-1}x'uu'x(x'x)^{-1}$$
$$= \sigma^2(x'x)^{-1}x'\Omega x(x'x)^{-1} \cdots\cdots 5$$

GLS starts from different expectation of (uu') from OLS. Here in GLS, $E(uu') = \sigma^2\Omega \cdots\cdots 6$

If we get p^{-1} such that $p'p = \Omega$ and $p^{-1\prime}p^{-1} = \Omega^{-1}$ and premultiply the model of $Y = xb + u$ by p^{-1} such that $p^{-1}y = y_*, \cdots\cdots$, and run OLS to this model of $y_* = x_*b + u_*$, then we have GLS estimate of parmeters as follows:

$$E(b) = E[x'\Omega^{-1}x)^{-1}x'\Omega^{-1}y]$$
$$= E[(x'\Omega^{-1}x)^{-1}x'\Omega^{-1}(x\beta+u)$$
$$= \beta + E[(x'\Omega^{-1}x)^{-1}x'\Omega^{-1}u] \cdots\cdots 7$$
$$= \beta \text{ (since } X \text{ and } U \text{ are uncorrelated)}$$

$$E(b-\beta)\ (b-\beta)' = \mathrm{Var}(b)$$
$$= \sigma^2(x'\Omega^{-1}x)^{-1} \cdots\cdots 8$$

Here we also assume $E(u)=0$, $E(x'u)=0$

As is well known, GLS estimate of β (that

(7) Bahl: *op. cit.*, pp. 186-189.

is b) is the Best Linear Unbiased Estimator (BLUE), since it is unbiased, consistent and efficient. Efficiency of b over $\hat{\beta}$ is well shown by Johnston.[8]

Now, we go back to budgetary studies where we may imagine at least two reasons why $E(uu') \neq \sigma^2 I_n$. First one is that the variance of disturbance may be positively related to one of independent variables (per capita income) or combination of two. Bahl and Saunders[9] reported that while grant-in-aid from the federal government is a very significant factor which determines the per capita expenditure for all states but not significant among 15 high income-high density states. Standard error was too great for the 15 states. Though we can explain this as some interaction effect between the federal aid and income-density variable on the per capita expenditure (that is: relationship between grant-in-aid and per capita expenditure is not linear) as the authors did, we can also imagine that high income states show big disturbances, which will lead to big σ^2 in those states. GLS would have shown us which is correct, but authors did not try.

The other reason why $E(uu') \neq \sigma^2 In$ is that there may be some regional difference among states such as Deep South, New England or Mountain Area. This means that some random shock or other unique factors in one state of the Deep South may influence per capita expenditure of other Deep South States as well as of its own state: $E(u_i u_j) \neq 0$. This problem may be solved by introducing another independent variable (region) or by GLS.

As the OLS estimate of $\hat{\beta}$ is unbiased and consistent, we can use OLS to calculate residuals and then see the relationship between squared residuals and each of independent variables for the first case of our discussion and if there is a relationship, then figure out σ_i^2 for X_i. For the second case, we can correlate residuals of each state and find σ_{ij} for off-diagonal elements of Ω matrix, but this is practically very difficult.

Though I did not do any data analysis for checking this, I must complain one thing more: R^2 to be used as an indicator for independent variable to explain the variation in per capita expenditure. OLS measure of R^2 is crippled when the assumption of $E(uu') = \sigma^2 In$ is violated.

R^2 is based on the calculation of error variance and total variance in Y (dependent variable). Since R^2 is regression variance divided by total variance in Y, smaller error variance will increase R^2. OLS estimate of error variance is under-estimated. In OLS we have, for bivariate case,[10]

$$e'e = \sum e_i^2 = (N-2)\sigma^2, \text{ in case } E(uu') = \sigma^2 In$$

$$e'e = \sum e_i^2 = (N-1)\sigma^2 - \sum x_i^2 \sigma_i^2 / \sum x_i^2,$$

$$\text{in case } E(uu') = \sigma_i^2 \quad \cdots\cdots 9$$

where $x_i = X_i - \bar{X}$,

compared to GLS estimator when[11] we have

$$E(uu') = \sigma^2 \Omega = \sigma_i^2 In, \text{ and}$$

$$e'\Omega^{-1}e = (N-2)\sigma^2 \cdots\cdots 10$$

Now it is clear that OLS's underestimate of disturbance variance depends on the number of case and some other factor in 9. In cross-sectional budget analysis, many used 48 American states: clearly some bias is there. More serious is the case of time-series data analysis where only around 20 years were covered.

(8) J. Johnston: *Econometric Methods*, New York: McGraw-Hill, p. 217.

(9) Roy W. Bahl and R.J. Saunders: "Determinants of Changes in State and Local Government Expenditures", *National Tax Journal*, Vol. 18 (March 1965), pp. 50-57.

(10) See Jan Kmenta: *Elements of Econometrics*. New York: Macmillan Publishing Co., pp. 254-255. Also see J. Johnston: *op. cit.*, p. 24.

(11) The proof of the equation 10 can be acquired from the author. It is too long to be printed.

II. Time-series Analysis

1. Three Basic Models

$Y_t=\beta+\beta X_t+U_t$ ·····11

$Y_t=\delta-\phi Y_{t-1}+U_t$ ·····12

$Y_t=\alpha+\phi Y_{t-1}+\beta X_t+U_t$ ·····13

where X is independent variable.

These three simple models can generate various other models according to the different assumptions on the causal link between X_t and Y_t (e.g. lagged model) and on the nature of U_t (so, relationship between U_t and U_{t-1} or U_t itself). And lastly we can also use pooled time-series and cross-sectional data set. But in budgetary studies, three basic models are largely used and no more than those. So I will treat these three models in that order.

2. Autoregressive Disturbances

$Y_t=\beta_1+\beta X_t+U_t$ ·····11 repeated

This model was pretty popular during 1960's in budgetary studies. Unit of analysis was either city government or state government.

Problem with this model is that U_t may be correlated to U_{t-1} of U_{t-1}, which will make OLS estimate of parameters $(\hat{\beta})$ less efficient: $\hat{\beta}$ in OLS is not the best linear unbiased estimate (BLUE).

A) But $\hat{\beta}$ still is unbiased and consistent, even when U_t is correlated to U_{t-1}, because

$$\begin{aligned} E(\hat{\beta}) &= E[(x'x)^{-1}x'y] \\ &= E[(x'x)^{-1}x'(x\beta+u)] \\ &= E[(x'x)^{-1}x'x\beta]+E[(x'x)^{-1}x'u] \\ &= \beta \end{aligned}$$

since we assume $E(U_t)=0$ and X_t is no correlated to U_t

B) Problem arises when we calculate confidence interval for $\hat{\beta}$, which needs var $(\hat{\beta})$, in case of autoregressive disturbances.

$$\begin{aligned} \text{Var}(\hat{\beta}) \text{ in OLS} &= [(x'x)^{-1} x'uu'x(x'x)^{-1}] \\ &= \sigma^2 In(x'x)^{-1} \text{ only when } E(uu')=\sigma^2 In \end{aligned}$$

Suppose $u_t=\rho u_{t-1}+e_t$ (where e_t is white noise), then

$$E(uu')=\sigma^2\Omega=\sigma^2\begin{pmatrix} 1 & \rho & \rho^2\cdots\rho_{n-1} \\ \rho & 1 & \\ \vdots & & 1\ddots \\ \rho_{n-1} & & \ddots 1 \end{pmatrix}$$

Now in this case, Var $(\hat{\beta})$ in OLS will be

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E[(x'x)^{-1}\ x'uu'x\ (x'x)^{-1}] \\ &= \sigma^2 E[(x'x)^{-1}\ x'\Omega x\ (x'x)^{-1}] \end{aligned}$$

which turns out to be for $\hat{\beta}_2$ (the regression coefficient of X_2 in bivarivate case)[12]

$$\text{Var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum x^2_t}\left(1+2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2}+\cdots\cdots +2\rho^{n-1}\frac{\sum x_t x_{t-n}}{\sum x_t^2}\right) \cdots\cdots 14$$

where $x_t=X_t-\bar{X}_t=X_t-E(X_t)$

and $t=1\sim N$

Now $\sum x_t\cdot x_{t-1}/\sum x_t^2$ is correlation between x_t and x_{t-1}

when $E(x_t)=E(x_{t-1})$. Let us denote γ for this correlation. Then

$$\begin{aligned} \text{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_t^2}(1+2\rho\gamma+2\rho^2\gamma^2+\cdots\cdots +2\rho^{n-1}\gamma^{n-1}) \\ &= \frac{\sigma^2}{\sum x_t^2}\left(\frac{1+\rho\gamma}{1-\rho\gamma}\right) \cdots\cdots 15 \end{aligned}$$

Since $|\rho|<1$

$|\gamma|<1$

GLS estimate of var (b) was fully explained by Kmenta,[13] which turns out to be for bivariate case,

$$\text{Var}(b_2)=\sigma^2(x'\Omega^{-1}x)^{-1} \text{ where } E(uu')=\sigma_t^2$$

$$=\frac{\sigma^2}{\sum x_t^2}\left[\frac{1-\rho^2}{1-2\rho\gamma+\rho^2}\right] \cdots\cdots 16$$

If $E(uu')=\sigma^2 In$, then, of course, 16 and 15 all will turn into $\sigma^2/\sum x_t^2$ which is from $\sigma^2(x'x)^{-1}$. Thus as far as $E(uu')=\sigma^2 In$, then it will

(12) See J. Johnston: *op. cit.*, p. 247.

(13) J. Kmenta: *op. cit.*, pp. 275-277.

be quite different from 15 as well as from 16. Johnston[14] fully developed the comparison of var ($\hat{\beta}$) based on $\sigma^2(x'x)^{-1}$ with 15 when ρ is not zero. Secondly, Kmenta[15] showed that var (b_2) in 16 is always smaller than var ($\hat{\beta}_2$) in 15.

Practical problem in budgetary studies arised, because researchers unexceptionally used OLS. Moreover, they used $\sigma^2(x'x)^{-1}$ based on the assumption of $E(uu')=\sigma^2 I_n$, which is not the actual case. This year's random shock will leave some part of its impact on the next year's expenditure, net of the influence of exogenous variables (X_s). This means that $u_t=\rho u_{t-1}+e_t$, where e_t is a white noise. Therefore we can think that in the model of $y_t=\beta_1+\beta x_t+u_t$, ρ (auto-correlation) will be positive.

Many exogenous variables can also be regarded as serially correlated. GNP per capita, Urbanization and Industrialization will be time-wise correlated positively, in State level budget analysis, i not in municipal government. This means that γ in 15 and 16 will be positive.

Now going back to 16 where

$$\text{Var}(b_2)=\frac{\sigma^2}{\sum x_t^2}\left(\frac{1-\rho^2}{1-2\rho\gamma+\rho^2}\right) \quad 16 \text{ repeated}$$

we can see that the right side parenthesis can become bigger or smaller than unity in accordance with $0>\rho>\gamma$ or $0<\rho<\gamma$ respectively. This means that the var ($\hat{\beta}_2$) based on the assumption of $E(uu)=\sigma^2 I_n$ will be smaller than var (b_2) calculated by GLS when $0>\rho>\gamma$; confidence interval for $\hat{\beta}_2$ will be narrower and picking up insignificant variables as significiant. If $0<\rho<\gamma$, then opposite result will come out: claiming significant variables as insignificant.

Because of naive framework of this model about time series mechanism, I did not check the above argument for federal level. And state level data are not available.

3. Auto-regressive Model

$$y_t=\delta+\phi y_{t-1}+u_t \cdots\cdots\cdots\cdots\cdots\cdots 12 \text{ repeated}$$

Lindblom-Wildawsky's incrementalism triggered quite different form of time-series analysis of budgetary data[15]

This model shows great power in explaining variation in y_t (R^2 is almost always greater than .80), compared to those of cross-sectional studies or time-series studies using $y_t=f(x_t)$.

However, this model in this simple form has many serious statistical problems as well as analytical problems in operationalizing incrementalism. I will focus on statistical problems, briefly mentioning the second problem so far as this is related to the first problems.

Statistical problems have two distinct dimensions: stationarity problem and autoregressive disturbances. Let me start with the second one.

A) Assumption of Normal Random Shock

In order to make a clear and simple argument, I will treat three possible models in time series:

$$y_t=\delta+\phi y_{t-1}+u_t \cdots\cdots\cdots\cdots\cdots\cdots 17$$

$$y_t=\delta+\phi y_{t-1}-\theta u_{t-1}+u_t \cdots\cdots\cdots\cdots\cdots\cdots 18$$

$$y_t=\delta+\phi y_{t-1}+\rho e_{t-1}+u_t \cdots\cdots\cdots\cdots\cdots\cdots 19$$

where u_t is a white noise

and $\rho e_{t-1}+u_t=e_t$

Here, 17 is the so-called AR(1) and 18 is ARMA (1, 1).

Which of above three alternatives best describes the actual budgetary time series is an empircal question. Indentification of AR (p) or ARMA ($p.q$) is well described in Box-Jenkin's framework. However, which of these three best operationalized the incremental budgetary decision-making is a theoretical question. Actually

(14) J. Johnston: *op. cit.*, pp. 246-249.
(15) J. Kmenta: *op. cit.*, p. 278.
(16) See footnotes 3 and 4.

Wildawsky et. al. proposes[17] the model which is $y_t=\phi y_{t-1}+u_t$. But in applying this model, they regressed y_t on y_{t-1}. So the result is either the same as 17 or the intercept is regarded as one part of residuals. In the latter case, this model means that this year's expenditure is the last year's budget multiplied by some constant and the rest is a random shock.

As random shock Ut is a white noise, it should have an equal possibility of becoming minus or plus. The case when random shock adds some expenditure is not difficult to imagine: a big disaster, unoccasional events, etc. But the case when Ut has a minus sign is very complicated here in 17. In this model, we can easily recognize that Ut is given equal weight (ϕ) as the other part of y_{t-1} in constructing y_t. If we think that the random shock of the last year will not demand the same amount of expenditure for this year's budget as it did last year, then the subtracted part will be regarded as one part of this year's minus Ut. In addition, other events like severe revenue reduction (in case of total budget) or need for unexpected increase in other expenditures (in case of one function or agency) are likely to reduce this year's budget, which can be regarded as one part of minus Ut.

As a whole, assumption of random shock (Ut) in 17 is plausible in this sense: more plausible than in 18. In 18, random shock of last year has impact on y_t by the amount of $(\phi-\theta)$ u_{t-1}, so the possibility of u_t's being minus is greatly reduced. Now the problem is: which of 17 and 19 is the best theoretical model for operationalizing incremental idea. If 19 is theoretically more plausible, then we have autocorrelated disturbances and OLS will have serious problems not only in the estimate of var (ϕ) but also in E (ϕ) itself.

(17) Davis et al. in footnote 3.

$$\begin{array}{ccc} y_{t-1}\leftarrow u_{t-1} & y_{t-1}\leftarrow u_{t-1} & \begin{array}{c}\downarrow\\ y_{t-1}\leftarrow e_{t-1}\leftarrow u_{t-1}\end{array} \\ \downarrow\phi & \downarrow\phi\swarrow-\theta & \downarrow\phi \quad \downarrow\rho \\ y_t\longleftarrow u_t & y_t\longleftarrow u_t & y_t\longleftarrow e_t\longleftarrow u_t \\ (17) & (18) & (19) \end{array}$$

Excluding intercepts, model 17, 18 and 19 are shown above. Now in 19, u_{t-1} influences this year's budget (y_t) by the amount of $(\phi+\rho)$ u_{t-1}. Here ρ can be either minus or plus, so the result is the same as 18 in terms of u_{t-1}'s impact on y_t. The difference between 18 and 19 is that random shock in 18 is incorporated into the core part of y_t after two years but in 19 it continuosly work as exogenous part as well as endogenous part (core part) of y_t till it dies out ($\rho^n\to 0$) when $|\rho|<1$.

As mentioned before, the nature of u_t (which should equally be plus or minus) makes 17 theoretically more palusible than 18 and 19, which means we do not have to worry about the atocorrelated distuabances as in 19. This is quite different from the case of 11 which is $y_t=\beta_0+\beta x_t+u_t$

B) Stationarity Assumption in Autoregressive Model

Now let me go back to Wildawsky's model.

$y_t=\phi y_{t-1}+u_t$ ·····20

This model is a hybridized form of 17. Intercept in 17 is regarded as one part of u_t in 20: δ in 17 is the original u_0 The problem with 17 as well as 20 in budgetary data analysis is that ϕ is greater than unity.

As is well known autoregressive scheme is based on the stationarity assumption. Stationarity assumption requires $|\phi|<1$. This assumption is crucial for calculating ϕ itself and its variance.

To make argument more simple, we see again $AR(1)$ model.

$AR(1)$·····$y_t=\delta+\phi y_{t-1}+u_t$ ·····17repeated

If stationarity assumption is valid, then

$E(y_t)=E(y_{t+1})$

$E(y_t \cdot y_{t+l}) = E(y_{t-1} \cdot y_{t+l-1})$

because $E(y_t) = E(y_{t+l}) = \frac{\delta}{1+\phi}$

and u_t and u_{t+l} are white noise

Similarly in $y_t = \phi y_{t-1} + u_t$ ···20 repeated

$E(y_t) = E(u_t + \phi u_{t-1} + \cdots \phi^n y_{t-n})$

If $|\phi| > 1$ then $E(y_t)$ does not converge into some constant, rather it will be explosive as time goes by, which means $E(y_t) \neq E(y_{t+l})$, as can be easily imagined.

In autoregressive scheme, autocorrelation (in this case, which is equal to ϕ) is calculated by

$$\rho = \frac{\mathrm{cov}(y_t \cdot y_{t-1})}{\mathrm{var}(y_{t-1})}$$

$$= \frac{\sum [y_t - E(y_t)][y_{t-1} - E(y_{t-1})]}{\sum [y_{t-1} - E(y_{t-1})]^2}$$

which can not be calculated because $E(y_t) \neq E(y_{t+l})$. Many used autorregressive model which is based on $E(y_t) = E(y_{t+l}) = E(y_{t-1})$. Their method is

$$\rho = \frac{\sum [y_t - \sum y_t/N)(y_{t-1} - \sum y_{t-1}/N)}{\sum (y_{t-1} - y_{t-1}/N)^2}$$

where $\sum y_{t-1}/N$ is not $E(y_{t-1})$ if $|\phi| > 1$.

One of the basic faults in this kind of pseudo AR model is the ever-changing $E(y_t)$ through time, which does not permit us to regard the whole series of data as some part of samples from the finite population. This means, in familiar matrix notations, that stochastic regressors can not be defined as plim $(x'x)^{-1}$ with some fixed finite And, of course, we can not define sampling distribution of dependent variable.

Another way of looking at this pseudo AR model is that we can regard ϕ simply as y_t/y_{t-1}. If random shock has the same size, then its effect on y_t is becoming completely negligible as y_t becomes bigger: No wonder why R^2 in this model is much bigger for big agencies than small agencies or for 1960's than for 1940's.

It is, therefore, quite easy to understand why some other exogenous variables incoporated into this pseudo AR model do not show big impact at all Last year's budget, if it is large, is decisive for this year's budget. Random shock which is very important in earlier stage, now has only a negligible impact.

This model is good so far as we are interested in predicting the approximate size of expenditure for next year from this year. But this model never makes clear why this year's budget predicts next year's budget so closely. Simple snow-ball effect may explain it but when we go down into small program level this model is almost useless. Maybe ever-increasing revenue permitted everincreasing expenditure. The point here is that we must construct more valuable statistical model to explain the trend of expenditure. For this purpose let's turn to another model.

C) Model Alternative to Pseudo Autoregressive Model

To make budgetary trend stationary, we need integration (that is: taking difference). Now the model may be expressed as

$\Delta y_t = \delta + \phi \Delta y_{t-1} + \Delta u_t$ ·····21

where $\Delta y_t = y_t - y_{t-1}$

$\Delta y_{t-1} = y_{t-1} - y_{t-2}$

$\Delta u_t = u_t - u_{t-1}$, therefore

$y_t = \delta + (1+\phi) y_{t-1} - \phi y_{t-2} + u_t - u_{t-1}$ ·····22

This model is theoretically plausible for budgetary studies. Increase in expenditure this year is last year's increase multiplied by ϕ and plus some constant and difference in random shock. As u_t and u_{t-1} are assumed normal random shock, $u_t - u_{t-1}$ will be also normal.

Here it is clear that ϕ is internal parameter and δ is external parameter. Let's see the nature of ϕ first. In 22, last year's expenditure has $(1+\phi)$ impact on this year's budget: so ϕ represents, in this case, that fraction of last year's budget which is increased this year. The expenditure of the year before last year (y_{t-2}) will influence this year's budget by $(1+\phi)$ $(1+\phi)$ for this year. But there is negative impact $(-\phi y_{t-2})$ also. This means that some internal

(within the system) repercussion acts against that increased part of y_{t-2} included in y_{t-1}: within-system feedback which is different from out-system feedback represented by $u_t - u_{t-1}$. This is plausible because only after one year the feedback mechanism works in the budgetary process.

As a whole, ϕ represents the impact of expansionst budgeteers while δ represents ever-increasing demand for public expenditure which will be incorporated into internal scheme right after it enters y_t. Therefore, whether δ is significantly bigger than zero depends on whether external demand is ever-increasing. Suppose this model is correct. Then we may have some clue to assess the impact of external factors on the increase in expenditure. The simple way to do this may be, after checking whether $(u_t - u_{t-1})$ is random white noise, see $(u_t - u_{t-1} + \delta = f(\Delta x_t))$

Now it is clear that this model 21 or 22 is theoretically superior to model 17 or 20. And it is statistically superior to those models, since

$$E(\Delta y_t) = E(\Delta y_{t+t}) = \frac{\delta}{1-\phi} \text{ in } \Delta y_t = \delta + \phi \Delta y_{t-1} - \Delta u_t$$

because $|\phi|$ can be smaller than 1.

D) Reappraisal and Empirical Test.

Whether simple AR (1) or $y_t = \phi y_{t-1} + u_t$ is worse than model 21, which is ARIMA (1, 1, 0), is completely depends on the size of parameter ϕ. If ϕ is greater than unity in $AR(1)$, then there is ever-increasing public expenditure. We know very well that price was continuously increasing over last two decades. If we grasp the real expenditure trend net of price change, then we can test each of the above models. Roughly speaking, many newly emerged functions of federal government have been showing ever-increasing demand. Social Welfare expenditure is typical[18]. In this case, model 21 fits nearly perfectly ($R^2 = .994$), while AR (1) shows $R^2 = .875$. However in Treasurey, Post Office and others, the result was almost the same in 17 as in 21. Now, it is obvious that choice among competing models depends on empirical test. However, if AR (1), model 20 and ARIMA (1, 1, 0) explain the equal size of variation in budget expenditure, we should use ARIMA (1, 1, 0) because of its theoretical and statistical superiority.

(18) The following statistics are calculated by the author, using the U.S. federal government expenditures during the period of 1954~1974.