A Technique to Improve the Fit of Linear Regression Models for Successive Sets of Data

Sung H. Park*

1. Introduction

In empirical study for fitting a multiple linear regression model for successive cross-sections data observed on the same set of independent variables over several time periods, one often faces the problem of poor R^2, the multiple coefficient of determination, which provides a standard measure of how good a specified regression line fits the sample data.

For a sample of size n for each cross-section, a full representation of the model may be written as

$$\boldsymbol{Y}_t=\boldsymbol{X}_t\boldsymbol{B}+\boldsymbol{u}_t \quad t=1, 2, \cdots, T \tag{1}$$

where $\boldsymbol{Y}_t$ and $\boldsymbol{u}_t$ are $n\times 1$ and $\boldsymbol{X}_t$ is $n\times(k+1)$, assuming there are an intercept term and k independent variables. The assumptions of the model (1) are, for $t=1, 2, \cdots, T$,

$$E(\boldsymbol{u}_t)=0 \tag{2.1}$$

$$E(\boldsymbol{u}_t\boldsymbol{u}_t')=\sigma^2\boldsymbol{I} \tag{2.2}$$

$$\text{rank}(\boldsymbol{X})=k+1 \tag{2.3}$$

$$\boldsymbol{X} \text{ is fixed.} \tag{2.4}$$

One additional assumption that must be made for the model (1) is that there

* Assistant Professor, Mississippi State University. The author would like to thank Professor Ronald R. Hocking for his constructive comments on earlier drafts of this paper.

is possible correlation among the disturbances of any two time periods, t and t', that is,

$$E(u_t u_t') = \rho\sigma^2 I \tag{3}$$

with a possible strong correlation between two consecutive time periods, $t-1$ and t.

In practice, one often has sample data in the form of such cross sections. For instance, in business and economics, situations arise where we wish to consider the behavior of a dependent variable to a set of independent variables over a number of time periods. An example can be found in Chung [1] where the market price of a common stock is the dependent variable and the independent variables are expected earnings per share, growth of market price per share, and various indices of risk. His data were collected on the same variables over a 10 year period.

An obvious approach to estimate the coefficients B_t in equation (1) is the application of the least squares method to each individual equation. Another approach is to estimate all T equations in the model simultaneously. However, because of the Assumptions (2. 1 to 2. 4), it can be verified that the equation-by-equation least squares will give the same estimates as do the least squares applied to the entire set of equations simultaneously. A proof of this identity may be found in Huang [5].

In empirical model-fitting by equation-by-equation least squares, we often face a poor R^2, which means that the percentages of the variation in the dependent variable explained by the independent variables are low. Such a low R^2 discourages the use of the model to explain or predict the behavior of the dependent variable from the independent variables. It is, therefore, desirable to obtain as high an R^2 as possible.

A reason of poor R^2 may be from a model misspecification, i.e., omission of some independent variables which should be specified in the model. The problem of model misspecification has received considerable attention by many researchers in the literature. Some of the authors are Griliches [2],

Toro-Vizcarrondo and Weeks [10], Rao [6], Rosenberg and Levy [7], and Hocking [4]. However, they are mainly interested in characteristics and consequences of misspecified multiple linear regression models. The purpose of this paper is to discuss a technique to improve the fit of misspecified models for successive sets of data. The technique, in essence, consists in taking the residuals of a misspecified model fitted to the set of data of time period $t-1$, and incorporating them as an additional independent variable in the model fitting the set of data of time period t.

The idea of the technique was originally suggested by Chung [1], and supported by Heiser and Dewan [3] through a simulation study. However, neither of them reported any theoretical discussions about the technique. This paper attempts to set up theoretical background of the methods, and, hopefully, serves as theoretical guide lines in empirical research. The model with the residuals as an additional variable will be called the "modified model". We shall mainly concern ourselves with following three areas of interest.

(a) How are the R^2 values of the model improved?

(b) What is the resulting error sum of squares of the modified model?

(c) What inferences can be made about the regression coefficient of the additional variable of the modified model?

2. Theoretical Considerations

Suppose that we have two sets of data Y_{t-1} and Y_t over two consecutive time periods, which were observed on the same variables. For brevity of expression we will use Y_1 and Y_2 in place of Y_{t-1} and Y_t from now on. We assume that the adequate regression model is, for $i=1$ and 2,

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}\boldsymbol{B}_i + \boldsymbol{u}_i \\ &= \boldsymbol{X}_p\boldsymbol{B}_{pi} + \boldsymbol{X}_k\boldsymbol{B}_{ki} + \boldsymbol{u}_i \end{aligned} \tag{4}$$

where $\boldsymbol{X}_p=(\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{k-1})$ and $\boldsymbol{B}_{pi}'=(B_{0i}, B_{1i}, B_{2i}, \cdots, B_{(k-1)i})$. In accordance with

the Assumptions (2. 1) to (2. 4) we assume that $E(u_1u_2')=\rho\sigma^2I$, X is fixed with full rank, and $E(u_i)=0$ and $E(u_iu_i')=\sigma^2I$ for $i=1$ and 2.

Let X_k be designated as the aforementioned misspecified variable. Since X_k is "missing," the equation-by-equation least squares fit is

$$\hat{Y}_1=X_p\hat{B}_{p1} \text{ and } \hat{Y}_2=X_p\hat{B}_{p2}$$

where $\hat{B}_{pi}$ is the least squares estimate of B_{pi} of model (4) when X_k is misspecified. Because of the misspecified variable, it is supposed that the R^2 values are low. Let

$$r^2=(e_1'e_2)^2/(e_1'e_1e_2'e_2) \tag{5}$$

where $e_1=Y_1-\hat{Y}_1$ and $e_2=Y_2-\hat{Y}_2$. Taking e_1 and incorporating them as an additional independent variable, the modified model

$$\tilde{Y}_2=X_pb_p+e_1b_k$$

is fitted. Note that b_p and b_k are used to denote the least squares estimates of B_{p2} and B_{k2} of model (4) when e_1 replaces X_k in the model for the second period, i.e., time t. The estimates b_p and b_k are

$$\begin{aligned}\begin{bmatrix}b_p\\b_k\end{bmatrix}&=\left[(X_p,e_1)'(X_p,e_1)\right]^{-1}(X_p,e_1)'Y_2\\&=\begin{bmatrix}X_p'X_p, & X_p'e_1\\e_1'X_p, & e_1'e_1\end{bmatrix}^{-1}\begin{bmatrix}X_p'Y_2\\e_1'Y_2\end{bmatrix}\\&=\begin{bmatrix}(X_p'X_p)^{-1}, & 0\\0, & (e_1'e_1)^{-1}\end{bmatrix}\begin{bmatrix}X_p'Y_2\\e_1'Y_2\end{bmatrix}\\&=\begin{bmatrix}(X_p'X_p)^{-1}X_p'Y_2\\(e_1'e_1)^{-1}e_1'Y_2.\end{bmatrix}.\end{aligned} \tag{6}$$

Letting $R_p=X_p(X_p'X_p)^{-1}X_p'$ and taking into consideration that $e_1=(I-R_p)Y_1$ and $e_2=(I-R_p)Y_2$, it is easy to find that

$$b_k=\frac{Y_1'(I-R_p)Y_2}{Y_1'(I-R_p)Y_1} \text{ or } \frac{r^2Y_2'(I-R_p)Y_2}{Y_1'(I-R_p)Y_2} \tag{7}$$

since $r^2=(e_1'e_2)^2/(e_1'e_1e_2'e_2)$. Thus, the sum of squares due to regression of the modified model is

$$\mathrm{SSR}=(b_p',b_k)\begin{bmatrix}X_p'Y_2\\ e_1'Y_2\end{bmatrix}=r^2Y_2'Y_2+(1-r^2)Y_2'R_pY_2.$$

Accordingly, the coefficient of determination, R_m^2, for the modified model is

$$R_m^2=\frac{\mathrm{SSR}-n\bar{Y}_2^2}{Y_2'Y_2-n\bar{Y}_2^2}=\frac{Y_2'[Ir^2+(1-r^2)R_p-J/n]Y_2}{Y_2'(I-J/n)Y_2}$$

where $J=(1, 1, \cdots, 1)'(1, 1, \cdots, 1)$.

Now it is of interest to determine the functional relationship between R_m^2 and r^2. Let $A=Ir^2-(1-r^2)R_p-J/n$, $B=I-J/n$ and $r^2=r_0^2$, a value of r^2 obtained from equation (5) for given consecutive sets of data. The relationships

$$AB=A,$$
$$AA=Ir_0^4+(1-r_0^4)R_p-J/n, \text{ and}$$
$$BB=B$$

indicate that R_m^2 is the ratio of two dependent quadratic forms, which is not in the form of a F-distribution. Thus, letting $q_1=Y_1'AY_2$ and $q_2=Y_2'BY_2$, R_m^2 may be expressed by a Taylor series expansion in such a way that

$$R_m^2=q_1/q_2=E(q_1)/E(q_2)+[(q_1)-E(q_1)]/E(q_2)-E(q_1)[(q_2)-E(q_2)]/[E(q_2)]^2$$
$$-[q_1-E(q_1)][q_2-E(q_2)]/[E(q_2)]^2+E(q_1)[q_2-E(q_2)]/[E(q_2)]^3+\cdots \qquad (9)$$

Since B is a positive definite matrix, $E(q_2)=E(Y_2'BY_2)$ is positive. Therefore, the denominators do not vanish in equation (9). The conditional expectation of R_m^2 with respect to r_0^2 in equation (9) is approximately

$$E(R_m^2)=E(q_1/q_2)\fallingdotseq E(q_1)/E(q_2)-\mathrm{Cov}[q_1,q_2]/[E(q_2)]^2+E(q_1)\mathrm{Var}(q_2)/[E(q_2)]^3,$$

where $\fallingdotseq$ is used to indicate 'is approximately equal to.' Let $M=E(Y_2)$. From equation (4), $M=XB_2=X_pY_{p2}+X_kB_{k2}$. Recalling that $\mathrm{Var}(Y_2)=\sigma^2I$, $AB=A$ and $BB=B$, we can obtain that

$$E(q_1)=E(Y_2'AY_2)=tr(A\sigma)^2+M'AM$$

$$=\sigma^2[r_0^2(n-k-1)+k]+\boldsymbol{M}'\boldsymbol{A}\boldsymbol{M},$$

$$E(q_2)=E(\boldsymbol{Y}_2'\boldsymbol{B}\boldsymbol{Y}_2)=tr(B\sigma)^2+\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M}$$

$$=(n-1)\sigma^2+\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M},$$

$$\mathrm{Cov}(q_1,q_2)=\mathrm{Cov}(\boldsymbol{Y}_2'\boldsymbol{A}\boldsymbol{Y}_2,\boldsymbol{Y}_2'\boldsymbol{B}\boldsymbol{Y}_2)=2\,tr(AB\sigma^4)+4\sigma^2\boldsymbol{M}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{M}$$

$$=2\sigma^4[r_0^2(n-k-1)+k]+4\sigma^2\boldsymbol{M}'\boldsymbol{A}\boldsymbol{M},$$

$$\mathrm{Var}(q_2)=\mathrm{Var}(\boldsymbol{Y}_2'\boldsymbol{B}\boldsymbol{Y}_2)=2\,tr(B\sigma^2)^2+4\sigma^2\boldsymbol{M}'\boldsymbol{B}\boldsymbol{B}\boldsymbol{M}$$

$$=2(n-1)\sigma^4+4\sigma^2\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M}.$$

Therefore, the approximate conditional expectation of R_m^2 can be written as

$$E(R_m^2)\fallingdotseq\frac{r_0^2[a\sigma^6+b\sigma^4+c\sigma^2+d]+e}{[\sigma^2(n-1)+\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}]^3} \tag{10}$$

where

$$a=(n-k-1)(n-1)^2,$$

$$b=(n-1)(n-3)\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{R}_p)\boldsymbol{M}+2n(n-k-1)\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}$$

$$c=[(n-k-1)\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}+2(n-1)\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{R}_p)]\boldsymbol{M}[\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}]$$

$$d=[\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{R}_p)\boldsymbol{M}][\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}]^2$$

$$e=\{[k\sigma^2+\boldsymbol{M}(\boldsymbol{R}_p-\boldsymbol{J}/n)\boldsymbol{M}][\sigma^2(n-1)+\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}]^2$$

$$+2\sigma^4k\boldsymbol{M}'(\boldsymbol{I}-\boldsymbol{J}/n)\boldsymbol{M}-2(n-1)\sigma^4\boldsymbol{M}'(\boldsymbol{R}_p-\boldsymbol{J}/n)\boldsymbol{M}\}.$$

Note that equation (10) expresses $E(R_m^2)$ as a linear function of r_0^2, with a constant positive slope for $n>k+1$, which implies that as r_0^2 increases, $E(R_m^2)$ increases with a constant rate.

The error sum of squares, SSE, of the modified model can be immediately obtained from equation (8).

$$\mathrm{SSE}=\boldsymbol{Y}_2'\boldsymbol{Y}_2-\mathrm{SSR}$$

$$=\boldsymbol{Y}_2'\boldsymbol{Y}_2-[r^2\boldsymbol{Y}_2'\boldsymbol{Y}_2+(1-r^2)\boldsymbol{Y}_2'\boldsymbol{R}_p\boldsymbol{Y}_2]$$

$$=(1-r^2)\boldsymbol{Y}_2'(\boldsymbol{I}-\boldsymbol{R}_p)\boldsymbol{Y}_2. \tag{11}$$

Notice that if there exists a perfect residual correlation, *i.e.*, $r^2=1$, then SSE is equal to zero, which is most desirable in regression analysis. The relationship between SSE and R_m^2 is linear, that is

$$R_m^2=\frac{\mathrm{SSR}-n\bar{Y}_2^2}{Y_2'Y_2-n\bar{Y}_2^2}=\frac{(Y_2'Y_2-\mathrm{SSE})-n\bar{Y}_2^2}{Y_2'Y_2-n\bar{Y}_2^2}=1-\frac{\mathrm{SSE}}{Y_2'Y_2-n\bar{Y}_2^2} \tag{12}$$

Substituting equation (11) into equation (12), we can obtain that

$$R_m^2=1-(1-r^2)\frac{Y_2'(I-R_p)Y_2}{Y_2'Y_2-n\bar{Y}_2^2} \tag{13}$$

From equations (11) and (13), it can be claimed that, when r^2 is significantly different from zero for two sets of cross-sections data, this technique is strongly recommended to reduce SSE, *i.e.*, to increase R^2 for the modified model. Figures 1 and 2 below show the changes in SSE and R_m^2 as r^2 changes.

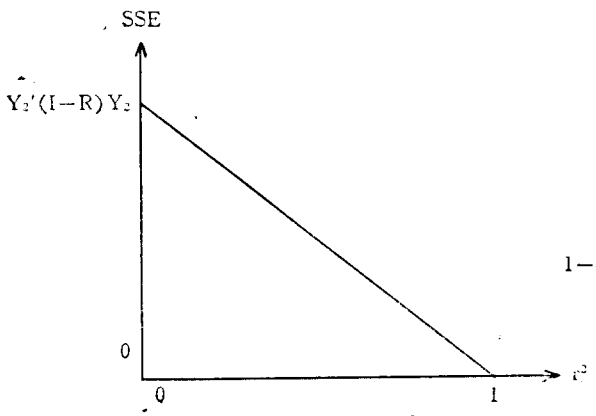

Figure 1: SSE vs. r^2

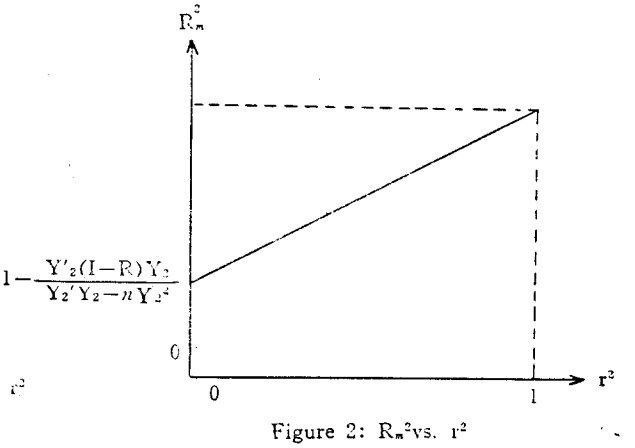

Figure 2: R_m^2vs. r^2

Now it is in order to discuss the regression coefficient of the additional variable of the modified model. Let $M_i=X_pB_{pi}+X_kB_{ki}$. From equation (4), we may write

$$Y_1=X_pB_{p1}+X_kB_{k1}+u_1=M_1+u_1$$
$$Y_2=X_pB_{p2}+X_kB_{k2}+u_2=M_2+u_2.$$

From the assumption $E(u_1u_2')=\rho\sigma^2I$,

$$\begin{aligned}\mathrm{Cov}(e_1,e_2)&=E[e_1-E(e_1)][e_2-E(e_2)]\\&=E[(I-R_p)Y_1-(I-R_p)X_kB_{k1}][(I-R_p)Y_2-(I-R_p)X_kB_{k2}]'\\&=(I-R_p)\rho\sigma^2.\end{aligned}$$

since $(I-R_p)X_p=0$. In equation (7), we obtained b_k, the estimate of the

coefficient of the additional variable in the modified model. If we let $Y_1'(I-R_p)Y_2=v_1$ and $Y_1'(I-R_p)Y_1=v_2$, $E(b_k)$ can be approximated similar to that of $E(R_m^2)$.

$$E(b_k)=E(v_1/v_2)$$

$$\doteqdot \frac{E(v_1)}{E(v_2)}-\frac{\mathrm{Cov}[v_1,v_2]}{[E(v_2)]^2}+\frac{E(v_1)\mathrm{Var}(v_2)}{[E(v_2)]^3} \tag{14}$$

From the theory of distribution of quadratic or bilinear forms (for a reference see Searle[8]), we can obtain

$$\begin{aligned}
E(v_1)&=E[Y_1'(I-R_p)Y_2]\\
&=tr[(I-R_p)\rho\sigma^2 I]+M_1'(I-R_p)M_2\\
&=\rho\sigma^2(n-k-1)+M_1'(I-R_p)M_2\\
E(v_2)&=E[Y_1'(I-R_p)Y_1]\\
&=tr(I-R_p)\sigma^2+M_1'(I-R_p)M_1,\\
&=\sigma^2(n-k-1)+M_1'(I-R_p)M_1\\
\mathrm{Cov}(v_1,v_2)&=\mathrm{Cov}[Y_1'(I-R_p)Y_2, Y_1'(I-R_p)Y_1]\\
&=2\rho\sigma^4(n-k-1)+2\rho\sigma^2M_1'(I-R_p)M_1+2\sigma^2M_1'(I-R_p)M_2,\\
\mathrm{Var}(v_2)&=\mathrm{Var}[Y_1'(I-R_p)Y_1]\\
&=2\sigma^4(n-k-1)+4\sigma^2M_1'(I-R_p)M_1.
\end{aligned}$$

Therefore, $E(b_k)$ in equation (14) may be expressed as

$$E(b_k)=\frac{\rho\sigma^2(n-k-1)+M_1'(I-R_p)M_2}{\sigma^2(n-k-1)+M_1'(I-R_p)M_1}$$

$$+\frac{2\sigma^2M_1'(I-R_p)M_2[M_1'(I-R_p)M_2-\rho M_1'(I-R_p)M_1]}{[\sigma^2(n-k-1)+M_1'(I-R_p)M_1]^3}$$

An interesting observation is that $\dfrac{dE(b_k)}{d\rho}$ is a positive constant if $n-k-1\geq 2$, and

$$\begin{aligned}
E(b_k)&=1, \text{ if } \rho=1 \text{ and } M_1=M_2,\\
&>0, \text{ if } \rho>0.
\end{aligned}$$

Therefore, one can easily conjecture the general relationship between $E(b_k)$

and ρ, that is, as ρ increases, $E(b_k)$ increases with a constant rate and the range of $E(b_k)$ for positive ρ_s' is possibly the interval $(0, d)$ where d is close to 1 depending on the values of M_1 and M_2. The relationship between b_k and r^2 is given in equation(7), which is self-explanatory.

3. Summary and Conclusions

The theoretical justification has been pursued for the technique which improves the fit of successive cross-sections data observed on the same set of independent variables over a number of time periods. If we briefly summarize the procedures of the technique, they are as follows.

Suppose, because of some reasons, an important variable is misspecified in the regression equation of time t, which is considered a main source of poor R^2. To improve the fit, take the residuals of a model fitted to the previous set of data (*i.e.*, the data of time $t-1$), and incorporate them as an additional independent variable in the model of time t. It has been observed that the addition of the new variable significantly increases the R^2 value for the modified model of time t, and the increase of R^2 value is highly related to the degree of correlation between the residuals at time $t-1$ and t.

REFERENCES

[1] Chung, P. S., "An Investigation of the Firm Effects Influence on Earnings to Price Ratios of Common Stocks," Paper Presented at the 48th Annual Meetings of the Western Economic Association, August 1973, Claremont California.

[2] Griliches, Z., "Specification Bias in Estimates of Production Function," *Journal of Farm Economics*, February 1957, 8-20.

[3] Heiser, R. F. and Dewan, P. B., "Using a Lagged Residual to Improve the Fit of a Multiple Linear Regression Model:a Monte Carlo Study," 1974

Proceedings of the Business and Economics Statistics Section, American Statistical Association, August 1974, 407-411.

[4] Hocking, R. R.,"Misspecification in Regression," *The American Statistician*, Vol. 28, 1974, 39-40.

[5] Huang, D. S., *Regression and Econometric Methods*, John Wiley & Sons, Inc. New York, 1970.

[6] Rao, P., "Some Notes on Misspecification in Multiple Regression," *The American Statistician*, Vol. 25, 1971, 37-39.

[7] Rosenberg, S. H. and Levy, P. S. "A Characterization on Misspecification in the General Linear Regression Model," *Biometrics*, Vol. 28, 1972, 1129-1132.

[8] Searle, S. R., *Linear Models*, John Wiley & Sons, Inc., New York, 1971.

[9] Toro-Vizcarrondo, C. and Wallace, T. D., "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," *Journal of the American Statistical Association*, Vol. 63, 1968, 558-572.

[10] Walls, R. E. and Weeks, D. L., "A Note on the Variance of a Predicted Response in Regression," *The American Statistician*, Vol. 23, 1969, 24-26.