SELECTION OF PROXY VARIABLES:
IF THERE ARE TWO CANDIDATES FOR THE PROXY VARIABLE,
WHICH SHOULD WE SELECT?

Kiyoshi FUJIKAWA* and Hikaru HASEGAWA**

代理変数の選択：条件無し平均二乗予測誤差
に基づく二つの代理変数モデルの比較

藤川清史*・長谷川 光**

Received September 19, 1988

【要 約】 我々計量経済学者がモデルのなかで観測不可能な変数に出会う。その変数の代理変数の
候補が二つ以上あるとしよう。この場合我々はその内のどれかを選択しなければならない。本稿の目
的は、代理変数の選択を特に条件無し平均二乗予測誤差に基づいて行うときの一つの指標を考慮する
ことである。直感的には真の観測不可能な変数と相関が高い（これは経済理論的に意味があると言
い替えてもよい）代理変数が選ばれるべきであると考えられるが、以下では次のよう逆説的な場合
が存在することを示す。つまり、相関の強さを単純相関係数で計り、二つの代理変数モデルの条件無
し（あるいは条件付でも良いのだ）平均二乗予測誤差を比較すれば、真正数との相関係数の低い代
理変数がよりふさわしい場合が存在するのである。

1. Introduction

When econometricians face to an unobservable variable in their econometric models,
they usually introduce a proxy variable. Some authors have made a comparison between

*Faculty of Business Administration and Information
経営情報学部

** School of Business and Economics, Kinki University
近畿大学 商経学部
the omitted-variable-model and the proxy-variable-model (e.g., McCallum (1972), Wickens (1972), Aigner (1974), Frost (1979) and Ohtani (1981)). Since Frost’s and Ohtani’s models are easy to handle, our discussion will be developed following their models as we show later. Ohtani (1981) has made a comparison especially conditional mean square prediction errors (CMSPE) and has concluded that indiscriminate usage of proxy variable is risky. However, these authors did not consider the case there exist plural candidates for a proxy variable for the unobservable true independent variable. In other words, they did not refer to the problem what kind of proxy variable is worth while to use. While our discussion follows Frost’s and Ohtani’s method, we treat following two problems, that is to say, “which is better, the omitted-variable-model or the proxy-variable-model?” and “which proxy-variable-model is better between two?”, based on the unconditional mean square prediction errors (UMSPE) in addition to CMSPE.

In Section 2, we shall present the models and notations, and calculate the UMSPE. In Section 3, comparing between UMSPE derived from the omitted-variable-model and that from the proxy-variable-model, we shall lead the same conclusion led by Ohtani (1981). In Section 4, we make a comparison of CMSPEs or UMSPEs between two different kinds of proxy-variable-models based on a simple correlation coefficient criterion.

We derive two main conclusions. The first one is that the conclusion derived by Ohtani (1981) based on CMSPE also holds when UMSPE is used to compare the omitted-variable-model and the proxy-variable-model. The second one is that (even if it intuitively seems natural to select the proxy variable highly correlated to the true variable,) there exists a case where the proxy variable whose correlation is lower is more suitable for a proxy variable if we make a comparison of CMSPEs or UMSPEs through a simple correlation coefficient criterion.

2. Model

Assume the following model to be the true regression model:

\[ y = x^T \delta + z^T \gamma + u \quad ; \quad u \sim N(0, \sigma^2 I_n), \quad (2-1) \]

where \( y \) and \( x \) are \( n \times 1 \) observable vectors, \( z \) is an \( n \times 1 \) unobservable vector, \( u \) is an \( n \times 1 \) vector of normally distributed stochastic variables with mean zero and covariance matrix \( \sigma^2 I_n \), and \( \delta \), \( \gamma \) and \( \sigma^2 \) are unknown scalar parameters. We assume that

\[ x^T x = z^T z = 1 \quad \text{and} \quad \bar{y} = \bar{x} = \bar{z} = 0, \]

where \( \bar{y} \), \( \bar{x} \) and \( \bar{z} \) are sample means of \( y \), \( x \) and \( z \), respectively. Since \( z \) is unobservable,
we can consider the following alternative models:

\[
\text{the omitted-variable-model} \quad y = x\delta + u \\
\text{and} \quad y = x\delta + z_i\gamma + u_i, (i = 1, 2),
\]

where \( z_i \) and \( z_{i*} \) represent vectors of two different proxy variables which substitute the unobservable true independent variable \( z \). It is assumed that \( z_i'z_i = 1, (i = 1, 2) \) and \( \bar{z}_i = 0, \) where \( \bar{z}_i \) are the sample means of \( z_i \).

OLS estimators of the coefficient parameters in the models (2-2) and (2-3) can be written as follows, respectively.

\[
\hat{\delta}_0 = x'y \\
\text{and} \quad \hat{\delta}_i = \frac{1}{a_i} \left[ \begin{array}{c} (x'y) - (x'z_i)(z_i'y) \\ a_i (z_i'y) - (x'z_i)(x'y) \end{array} \right],
\]

where \( a_i = 1 - (x'z_i)^2, (i = 1, 2) \).

On the other hand, we consider the models for the prediction period, letting \( y^*, x^*, z^* \) and \( z_{i*} \) be the scalar observations covering the prediction period on \( y, x, z \) and \( z_i \).

\[
y^* = x^*\delta + z^*\gamma + u^* \\
y^* = x^*\delta + u^* \\
\text{and} \quad y^* = x^*\delta + z_{i*}\gamma + u_{i*}, (i = 1, 2).
\]

We assume \( \text{E}(u^*) = 0, \text{E}(u^*u^*) = \sigma^2 \) and \( \text{E}(u^*u_j) = 0, \) where \( u_i \) is the \( j \)-th element of vector \( u \) \((j = 1, 2, \cdots, n)\). According to these models, the following predictors for \( y^* \) are considered.

\[
\hat{y}_0^* = x^*\hat{\delta}_0 \\
\text{and} \quad \hat{y}_i^* = x^*\hat{\delta}_i + z_{i*}\hat{\gamma}_i, (i = 1, 2).
\]

Then, we can calculate UMSPE as

\[
\begin{align*}
\text{MSE}(\hat{y}^*_0) &= \text{E}(y^* - \hat{y}^*_0)^2 \\
&= [z^* - x^*(x'z)]y^2 + (1 + x^*x^2)\sigma^2 \\
\end{align*}
\]

and
MSE \((\tilde{y}_t^*) = E(y^* - \hat{y}_t^*)^2\) 
\[= [a_i z^* - x^*[(x'z) - (x'z)(z'z)]] - z^*[(z'z) - (x'z)(x'z)]\] 
\[+ [1 + (x^* + z^*)^2 - 2(x'z)(x'z)]/a_i] \sigma^2 \quad (i = 1, 2).\]

Our interest is to compare these UMSPES. In order to simplify our discussion, letting \(s^* = [x^*, z^*, z_t^*]\) and \(s' = [x, z, z_t]\), we assume

\[s^*s^* = ss'/n\]
\[
\begin{bmatrix}
1/n & (x'z)/n & (x'z_t)/n \\
(x'z)/n & 1/n & (z'z_t)/n \\
(x'z_t)/n & (z'z_t)/n & 1/n
\end{bmatrix}
\]

(see, Ameniya (1980, p.334)). Under this assumption,

\[
MSE(\tilde{y}_t) = (1 - h^2)\sigma^2/n + (1 + 1/n)\sigma^2
\]  
(2-4)

and

\[
MSE(\hat{y}_t) = (a_i - h^2 - g_i^2 + 2f_i g_i)\sigma^2/na_i + (1 + 2/n)\sigma^2,
\]  
(2-5)

where \(h = x'z, f_i = x'z_i\) and \(g_i = z'z_i\) (\(i = 1, 2\)). Note that \(h, f_i\) and \(g_i\) denote the simple correlation coefficients among \(x, z\) and \(z_t\), since they are normalized.

3. A comparison between the omitted-variable-model and the proxy-variable-model

To compare the MSEs, we shall take the difference between (2-5) and (2-4),

\[
MSE(\hat{y}_t^*) - MSE(\tilde{y}_t^*) = -(g_i - hf_i)^2/na_i + \sigma^2/n.
\]  
(3-1)

So far as vector \(x\) is assumed to be given, the right hand side of the equation (3-1) can be rewritten as follows by introducing the partial correlation coefficient of \(z\) and \(z_t\) which is denoted by \(R(zz_t|x)\) and the coefficient \(r^2 = (1 - h^2)/(\sigma^2)\) as Ohtani (1981) proposed,

\[
MSE(\hat{y}_t^*) - MSE(\tilde{y}_t^*) = [-R^2(zz_t|x)t^2 + 1] \sigma^2/n.
\]  
(3-2)

The equation (3-2) is synonymous with the equation (8) in Ohtani (1981, p. 627), which is

\[
[MSE(\tilde{y}_t) - MSE(\tilde{y}_t^*)]/\sigma^2 = -R^2(zz_t|x)t^2 + 1,
\]  
(3-2)
where $\hat{y}_i = x\delta_i + z, \hat{y}_s$, $(i=1, 2)$ and $\hat{y}_o = x\delta_o$, and here MSE is the conditional mean square prediction error. Therefore, in spite of using UMSPE, we can derive the same result as Ohtani (1981) which employs CMSPE. After all, using proxy variables indiscriminately is also risky from the view point of a comparison between the omitted-variable-model and the proxy-variable-model based on UMSPE.

Consider the region such that

$$\text{MSE}(\hat{y}_s^*) - \text{MSE}(\hat{y}_o^*) \geq 0,$$

which is reduced to

$$\frac{(g_i - h_i)^2}{(\sigma/\gamma)^2 + f_i^2} \geq 1.$$

(3-4)

The region where the inequality (3-4) is outside of an oval with a rotation term on $f_i - g_i$ plane.

To let the discussion easier, we assume $h = 0$ as the first step. Substituting $h = 0$ to (3-4), we get the following inequality,

$$\frac{g_i^2}{(\sigma/\gamma)^2 + f_i^2} \geq 1.$$

(3-5)

Figures 1 and 2 show the region where (3-5) holds. From these figures, we can see that

Fig. 1  $0 < \sigma/\gamma < 1$.

The part of oblique lines shows the region where (3-5) holds.

Fig. 2  $1 < \sigma/\gamma$

The part of oblique lines shows the region where (3-5) holds.
the region where (3-5) holds varies according to the values of the parameter $\sigma$ and $\gamma$. These figures tell the following two points. First of all, when $\sigma$ is small, we had better use the proxy variable model, and the region where using the proxy-variable-model is suggested becomes narrower as $\sigma$ increases. Namely, if the true regression model is assumed effective (i.e., small $\sigma$), the proxy-variable-model works effectively. Secondary, as to $\gamma$, when $\gamma$ is large, we had better use the proxy-variable-model. That is to say, in case the effect of true unobservable independent variable $z$ is considered to be important (i.e., $\gamma$ is large), efficiency of the proxy-variable-model increases.

As the second step, we consider the region where (3-4) holds in the case $h \neq 0$. Note that $-1 \leq h \leq 1$, since $h$ is a simple correlation coefficient between $x$ and $z$. We introduce the following $F$ and $G$ to eliminate the rotation term in the inequality (3-4):

$$
\begin{bmatrix}
F \\
G
\end{bmatrix} =
\begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}
\begin{bmatrix}
f_i \\
g_i
\end{bmatrix}.
$$

(3-6)

We can get the following from (3-6) easily,

$$
\begin{bmatrix}
f_i \\
g_i
\end{bmatrix} =
\begin{bmatrix}
F \cos \theta + G \sin \theta \\
-F \sin \theta + G \cos \theta
\end{bmatrix}.
$$

(3-7)

Letting the inequality (3-4) be the following form

$$(\gamma^2 h^2 + \sigma^2)f_i^2 - 2\gamma^2 h f_i g_i + \gamma^2 g_i^2 \geq \sigma^2,$$

and substituting (3-7) to (3-4)', we obtain

$$
[[(\gamma^2 h^2 + \sigma^2)\cos^2 \theta + \gamma^2 \sin^2 \theta + 2\gamma^2 \sin \theta \cos \theta]F^2 + 2[(\gamma^2 h^2 + \sigma^2 - \gamma^2)\sin \theta \cos \theta - \gamma^2 \sin \theta (\cos^2 \theta - \sin^2 \theta)]FG + [(\gamma^2 h^2 + \sigma^2)\sin^2 \theta + \gamma^2 \cos^2 \theta - 2\gamma^2 \sin \theta \cos \theta]G^2 \geq \sigma^2.
$$

(3-8)

Letting the rotation term in (3-8) be equal to zero:

$$(\gamma^2 h^2 + \sigma^2 - \gamma^2)\sin \theta \cos \theta - \gamma^2 \sin \theta (\cos^2 \theta - \sin^2 \theta) = 0.
$$

(3-9)

Calculating the above,

$$\tan 2\theta = 2\gamma^2 h / [\sigma^2 - \gamma^2(1-h^2)] = B / A,$$

where $A = \sigma^2 - \gamma^2(1-h^2)$ and $B = 2\gamma^2 h$. 

- 18 -
SELECTION OF PROXY VARIABLES

Here, we assume $A > 0$, i.e., $\sigma / \gamma \geq 1$ and $0 < \theta < \pi / 4$ for simplicity. By these assumptions, we obtain

$$\cos^2 \theta = \frac{(A + \sqrt{A^2 + B^2})}{2\sqrt{A^2 + B^2}}$$

(3-10)

and

$$\sin^2 \theta = \frac{(-A + \sqrt{A^2 + B^2})}{2\sqrt{A^2 + B^2}}.$$  

(3-11)

Substituting (3-10) and (3-11) into (3-8), we can get

$$(\sqrt{A^2 + B^2} + A + 2\gamma^2)G^2 / \sigma^2 + (-\sqrt{A^2 + B^2} + A + 2\gamma^2)H^2 / \sigma^2 \geq 1.$$  

(3-12)

This inequality is another expression of (3-4) in the case $h \neq 0$. Figure 3 shows the region where the inequality (3-12) holds.\(^{(12)}\)

![Diagram](image)

Fig. 3  $A = \sigma^2 - \gamma^2 (1 - h^2) > 0$, $\sigma / \gamma > 1$ and $0 < \theta < \pi / 4$.

The part of oblique lines shows the region where (3-12) holds.

4. A comparison between two proxy-variable-models

As we have mentioned in Section 3, the result of the comparison of UMSPEs is synonymous with that of CMSPEs discussed by Ohtani (1981). That is to say, from the equations (3-2) and (3-2)', the next proposition holds.

$$\text{MSE}(\hat{y}_2^*) - \text{MSE}(\hat{y}_1^*) \geq 0$$

(4-1)

$$\iff |R(zz_1|x)| \leq |R(zz_2|x)|$$

$$\iff \text{MSE}(\hat{y}_2) - \text{MSE}(\hat{y}_1) \geq 0.$$  

---

19 ---
The above relation fits our intuition that the proxy variable highly correlated to the unobservable true variable should be selected. Note, however, that the proposition (4-1) is based on the partial correlation coefficients between \( z \) and \( z_i \) \((i = 1, 2)\) under given vector \( x \).

On the other, when we use the simple correlation coefficients, it will be easily found out that there exists such a paradoxical case as

- if \( |g_2| \geq |g_1| \), then \( \text{MSE} \left( \hat{y}_2' \right) \geq \text{MSE} \left( \hat{y}_1' \right) \)
- or
- if \( |g_2| \geq |g_1| \), then \( \text{MSE} \left( \hat{y}_2 \right) \geq \text{MSE} \left( \hat{y}_1 \right) \).

This implication of this case can be shown more explicitly as follows. Assuming \( h > 0 \) and \( f_i > 0 \) \((i = 1, 2)\) for simplicity, consider the region defined such that

\[
\text{MSE} \left( \hat{y}_2' \right) - \text{MSE} \left( \hat{y}_1' \right) \geq 0.
\]

This region is equivalent to

\[
a_2(g_1 - hf_1)^2 - a_1(g_2 - hf_2)^2 \geq 0. \tag{4-2}
\]

#### 4.1 The case of \( f_1 = f_2 = f \) and \( g_1 = g_2 \)

Since \( f_1 = f_2 = f \), it holds that \( a_1 = a_2 \). Therefore, (4-2) can be reduced to

\[
(g_2 - g_1)(2hf - g_1 - g_2) \geq 0. \tag{4-3}
\]

Figure 4 shows the region where the inequality (4-3) holds. We can confirm that there

![Fig. 4](image-url)  
Fig. 4 The part of oblique lines shows the region where (4-3) holds. The part of horizontal lines shows the paradoxical case.
obviously exists the paradoxical region where the following holds:

\[
\text{if } |g_2| \geq |g_1|, \text{ then } \text{MSE} (\hat{y}_2^*) \geq \text{MSE} (\hat{y}_1^*) \tag{4-4}
\]

4.2 The case of \( h=0, f_1 \neq f_2 \) and \( g_1 = g_2 \)

The inequality (4-2) will easily be reduced to the following inequality,

\[
a_2g_1^2 - a_1g_2^2 \geq 0 \tag{4-5}
\]

Therefore,

\[
(g_2 - \sqrt{(a_2/a_1)g_1})(g_2 + \sqrt{(a_2/a_1)g_1}) \leq 0. \tag{4-5}'
\]

In the case \( 0 < f_1 \leq f_2 \) and \( f_2 = tf_1 \ (t \geq 1) \), the absolute value of the slopes of two lines in (4-5)', which is denoted by \( p(t) \), can be written

\[
p(t) = \sqrt{(1 - t^2f_1^2)/(1 - f_1^2)}.
\]

Since \( dp(t)/dt < 0 \), \( p(t) \) decreases as \( t \) increases. In this case, there does not exist the paradoxical case as Figure 5 shows.

However, in the case \( 0 < f_2 \leq f_1 \) and \( f_1 = tf_2 \ (t \geq 1) \), the absolute value of the slope of two lines in (4-5)', which is denoted by \( q(t) \), is

Fig. 5  \( 0 < f_1 < f_2 \) and \( f_2 = tf_1 (t \geq 1) \).

The part of oblique lines shows the region where (4-5) holds.
Fig. 6  \(0 < f_2 < f_1\) and \(f_1 = t f_2 (t \geq 1)\).

The part of oblique lines shows the region where (4-5) holds.

The part of horizontal lines shows the paradoxical case.

\[ q(t) = \sqrt{(1 - f_2^2) / (1 - t^2 f_2^2)}. \]

Since \(dq(t)/dt > 0\), \(q(t)\) gets large as \(t\) increases. In this case, the paradoxical case appears and the region of paradoxical case broadens as the difference of simple correlation coefficients between two proxy variables \((z_1 \text{ and } z_2)\) and \(x\) gets large. Figure 6 shows this interesting fact.

5. Conclusion

As we discussed in Section 3, we have found that the result of the comparison of the unconditional mean square prediction errors (UMSPEs) between the omitted-variable-model and the proxy-variable-model is synonymous with Ohtani (1981). That is to say, the conclusion which we obtained coincides Ohtani's one "using the proxy variable indiscriminately is risky". We also found that the values of \(\sigma\) and \(\tau\) play an important role regarding the efficiency of proxy variable in this comparison.

Another finding is as follows. If we have two candidates for a proxy variable for the unobservable true independent variable in the model, we have to be careful to the paradoxical case in selecting proxy variables. That is to say, though a comparison of MSEs based on the partial correlation criterion derives an intuitively persuasive conclusion that the higher the correlation coefficient is, the smaller the MSE becomes, a comparison of MSEs based on the simple correlation criterion contains the paradoxical case (see, e.g., the equation (4-4)) where we should select the proxy variable whose
correlation coefficient to the true variable is lower.

Acknowledgements

The authors thank professors F. Okazaki of Setsunan University, and T. Toyoda and K. Ohtani of Kobe University for their valuable comments.

Notes

(1) Fujikawa and Hasegawa (1986) have made a comparison between two proxy-variable-models by using a criterion of minimum coefficient MSE.

(2) In the following equation, which is obtained by substituting '=' in place of '>' in the inequality (3.4),
\[ (g_i - hf_i)^2 \gamma^2 + \sigma^2 f_i^4 = \sigma^2, \]
\[ g_i = \frac{a}{\gamma} \text{ when } f_i = 0, f_i = \pm \frac{d}{\sqrt{h^2 \gamma^2 + \sigma^2}} \text{ when } g_i = 0, g_i = h \text{ when } f_i = 1, \text{ and } g_i = -h \text{ when } f_i = -1. \]

(3) In this case, when x is orthogonal to the true variable z, i.e., \( x'z = 0 \), there is no room for the paradoxical case to appear.

References


Summary

Econometricians are often enforced to select a proper proxy variable for an unobservable true variable from a set of alternatives. The purpose of this paper is to consider criteria for selecting a proper proxy variable attaching importance
principally to the unconditional mean square prediction errors. Although it seems that the proxy variable highly correlated (which sometimes means economically significant) to the unobservable true variable should be selected, we show that there exists a case where the proxy variable whose statistical correlation to the true variable is lower is more suitable, when we make a comparison of mean square prediction errors of two different kinds of proxy-variable-models.