Selection of Proxy Variables

Kiyoshi FUJIKAWA*
Hikaru HASEGAWA**

Abstract

The problem which we treat in this paper is to select the most suitable variable among some which all are considered to be plausible proxy variables for an unobservable independent variable. It intuitively seems that the more correlative proxy variable to the true unobservable variable tends to be selected. However, it is shown that there exists the case where the less correlative proxy variable should be selected, if the criterion of selection is the minimum coefficient mean squared error.

1. Introduction

Suppose an econometrician encounters an unobservable independent variable in estimating his/her econometric model. First of all, he/she faces an alternative whether this unobservable variable would be omitted (Choice of Omitted variable model) or a substitutive proxy variable would be used (Choice of Proxy variable model). So far, many researchers have made a comparison between Omitted variable model and Proxy variable model. For instance, McCallum (1972) and Wickens (1972), which treated proxy vari-

* Faculty of Economics, Osaka University of Economics, 2–2–8 Ohsumi Higashi-Yodogawa-ku Osaka 533, Japan.
** School of Business and Economics, Kinki University, 3–4–1 Kowakae Higashi-Osaka, Osaka 577, Japan.

ables as one example of errors-in-variables, showed that Proxy variable model is better under a criterion of minimum asymptotic coefficient bias even if the proxy variable is poor. Aigner (1974) also concluded, based on the mean squared error (MSE) criterion, that inclusion of the proxy variable is recommended over a broad range of empirical situation although it is not necessarily a superior strategy unequivocally.

Contrary to them, Frost (1979) showed that using proxy variable indiscriminately can be risky from the viewpoint of the squared bias and variance. Ohtani (1981) also showed the almost same conclusion under the criterion of MSE of prediction. After all, it is not easy to conclude on whether proxy variables should be used or unobservable variables shound be omitted.

However, once an econometrician decided to introduce a proxy variable, the second problem to be solved is what kind of proxy variables should be selected. So far the study treating this problem is very scanty. Therefore, we here try to answer this problem following Frost and Ohtani's model.

It intuitively seems that the more correlative proxy variable to the true unobservable variable (say, rich proxy variable) should be selected. However, there exists the case where the less correlative proxy variable (say, poor proxy variable) should be selected, if the criterion of selection is the minimum coefficient MSE.

We first, in section 2, present the model, and calculate the MSE's of the ordinary least squares (OLS) estimators of coefficient vectors of rich and poor proxy variable models. We next, in the following sections 3 and 4, compare these MSE's. In section 3, the criterion of richness of the proxy variable is simple correlation coefficients, and in section 4 it is partial correlation coefficients.

2. Model and MSE's of OLS estimators of coefficient parameters

Let us assume that the following linear regression model as a true one:

(2.1) $y=x\delta+z\gamma+u=X\beta+u,$

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 whose elements have independently identical distribution with zero mean and variance σ^2, and δ, γ and σ^2 are unknown scalar parameters. $X=[x, z]$ and $\beta'=[\delta, \gamma]$. It is also assumed that the means of elements of x *and* z are zero and that their norms are 1.

Since z is unobservable, we introduce the following two different proxy variable models:

(2.2) $y=x\delta_1+z_1\gamma_1+u_1=X_1\beta_1+u_1$ and

(2.3) $y=x\delta_2+z_2\gamma_2+u_2=X_2\beta_2+u_2,$

where z_1 and z_2 are $n\times 1$ vectors of proxy variables for the true unobservable independent vector z. $X_i=[x,z_i]$ and $\beta_i=[\delta_i, \gamma_i]$ for $i=1$ and 2. It is also assumed that the means of elements of z_1 and z_2 are zero and that their norms are 1.

Next, we consider the estimators of coefficient parameters and calculate their MSE's. In the models (2.2) and (2.3), we obtain the following estimators:

(2.4) $$\hat{\beta}_i=\begin{bmatrix}\hat{\delta}_i\\ \hat{\gamma}_i\end{bmatrix}=\frac{1}{1-(x'z_i)^2}\begin{bmatrix}(x'y)-(x'z_i)(z_i'y)\\ (z_i'y)-(x'z_i)(x'y)\end{bmatrix}$$

By using these estimators, we can calculate their MSE's. For $i=1$

and 2,

$$(2.5)\quad MSE(\hat{\beta}_i)=E[(\hat{\beta}_i-\beta)'(\hat{\beta}_i-\beta)] =E(\hat{\delta}_i-\delta)^2+E(\hat{\gamma}_i-\gamma)^2 =MSE(\hat{\delta}_i)+MSE(\hat{\gamma}_i),$$

where

$$MSE(\hat{\delta}_i)=[\{(x'z)-(x'z_i)(z'z_i)\}^2\gamma^2+(1-(x'z_i)^2)\sigma^2]/(1-(x'z_i)^2)^2 \text{ and}$$

$$MSE(\hat{\gamma}_i)=\frac{[\{(x'z_i)^2-(x'z_i)(x'z)+(z'z_i)-1\}^2\gamma^2+(1-(x'z_i)^2)\sigma^2]}{(1-(x'z_i)^2)^2}$$

Lastly we would like to introduce an abbreviated expression of the simple correlation coefficient. For any vector v and w,

$$R(vw)=(v-\bar{v}c)'(w-\bar{w}c)/[\{(v-\bar{v}c)'(v-\bar{v}c)\}\{(w-\bar{w}c)'(w-\bar{w}c)\}]^{\frac{1}{2}},$$

where $\bar{v}$ and $\bar{w}$ are sample means of vector v and w, and c is a vector whose elements are 1. Note that $R(vw)=v'w$ if means and norms of both vectors v and w are zero and 1, respectively, and that $v'w$ represents the angle between the vectors v and w.

Using this abbreviation, we employ the following notations in the following sections for $i=1$ and 2.

$$h=R(xz)\geqq 0,\ f_i=R(xz_i)>0 \text{ and } g_i=R(zz_i)>0.$$

The inequalities are assumed just for simplicity.

By these notations, MSE($\hat{\delta}_1$) and MSE($\hat{\gamma}_1$) are rewritten as follows:

$$(2.6)\quad MSE(\hat{\delta}_i)=[(h-f_ig_i)^2\gamma^2+(1-f_i)^2\sigma^2]/(1-f_i^2)^2,$$

$$(2.7)\quad MSE(\hat{\gamma}_i)=[\{f_i(f_i-h)+g_i-1\}^2\gamma^2+(1-f_i^2)^2\sigma^2]/(1-f_i^2)^2$$

3. Comparison of two proxy variable models based on the simple correlation coefficients

The criterion of richness of proxy variables in this section is to have larger simple correlation coefficient to the true variable.

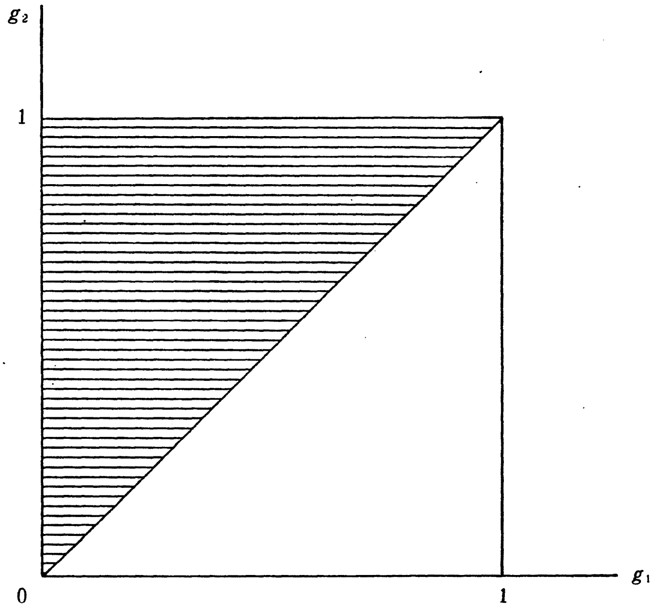

Figure 3.1
The part of horizontal lines shows the region where z_2 is richer than z_1.

Suppose the second proxy (z_2) is richer than the first one (z_1). This assumption means $g_2 \geqq g_1$ in this criterion. The part of horizontal lines in Figure 3.1 is the region where z_2 is richer than z_1.

Since our model, equation (2.1), contains two coefficient parameters δ and γ. We focus on these parameters separately in sections 3.1 and 3.2 and consider them at the same time in section 3.3.

3.1. For MSE $(\hat{\delta}_i)$

Here we focus on MSE $(\hat{\delta}_i)$ and assume that $\hat{\delta}_1$ dominates $\hat{\delta}_2$, *i. e.*,

(3.1) $MSE(\hat{\delta}_2) - MSE(\hat{\delta}_1) \geqq 0.$

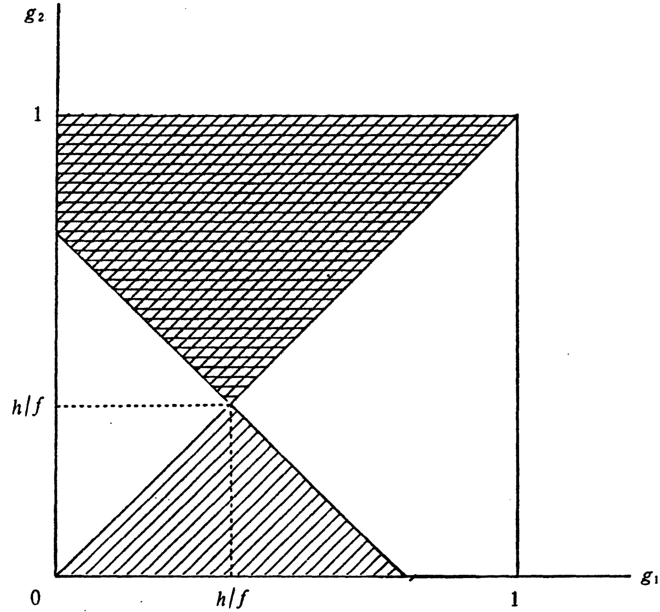

Figure 3.2

We assume $0<h/f<1$. The part of oblique lines shows the region where (3.3) holds, *i. e.*, $\hat{\delta}_1$ dominates $\hat{\delta}_2$. The part of horizontal lines shows the paradoxical region where (3.4) holds.

This inequality is rewritten as follows using (2.6):

$$(3.2)\quad [(1-f_1^2)^2(h-f_2g_2)^2-(1-f_2^2)^2(h-f_1g_1)^2]\gamma^2 + (1-f_1^2)(1-f_2^2)(-f_1^2+f_2^2)\sigma^2 \geqq 0.$$

We treat in this paper the following two special cases, but this is for convenience and does not change our conclusion.

3. 1. 1. The case of $f_1=f_2=f$ and $g_1 \neq g_2$

First of all, let us condider the case of $f_1=f_2=f$ and $g_1 \neq g_2$. In this case, (3.2) is reduced to

$$(3.3)\quad (g_2-g_1)(g_2+g_1-2h/f) \geqq 0.$$

Figure 3.2 shows the region where (3.3) holds. Of course, h/f may

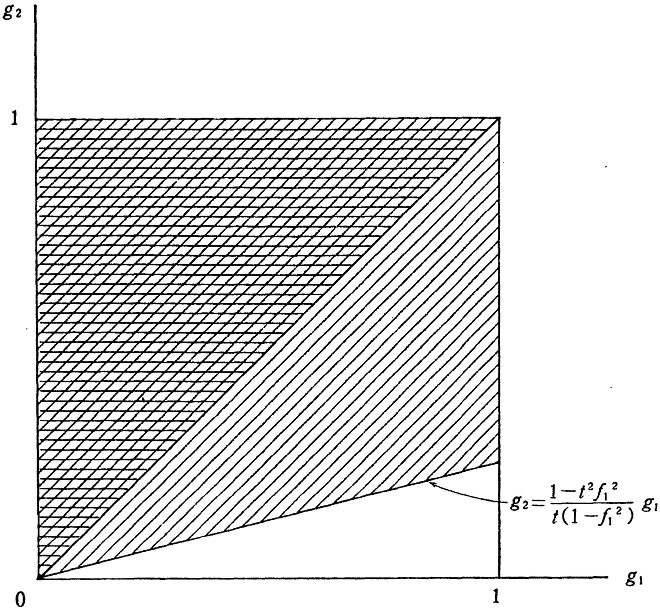

Figure 3.3
The part of oblique lines shows the region where (3.9) holds, *i. e.*, $\hat{\delta}_1$ dominates $\hat{\delta}_2$. The part of horizontal lines shows the paradoxical region where (3.4) holds.

lie outside of the square in Figure 3.2, but it is possible that there exists the case where this figure shows. In this figure, we can find a paradoxical region, that is to say, the region such as

(3.4) $g_2 \geqq g_1$ and $MSE(\hat{\delta}_2) \geqq MSE(\hat{\delta}_1)$.

That is to say, though the inequality of $g_2 \geqq g_1$ holds, which means the second proxy (z_2) is richer than the first one (z_1) in a comparison of the simple correlation coefficients, it is possible that $\hat{\delta}_1$ dominates $\hat{\delta}_2$ in the context of MSE.

However, even if the existence of (3.4) is confirmed, readers may suspect that this paradox is because of seemingly high correlation between z and z_2, and may insist that the richness of proxy

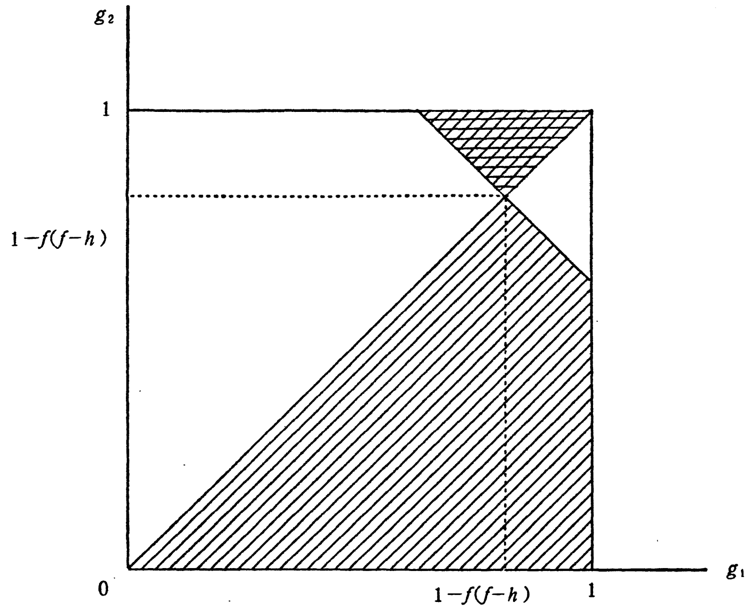

Figure 3.4

We assume $0<1-f(f-h)<1$. The part of oblique lines shows the region where (3.12) holds, *i. e.*, $\hat{\gamma}_1$ dominates $\hat{\gamma}_2$. The part of horizontal lines shows the paradoxical region where (3.13) holds.

variables should be discussed based on the partial correlation coefficients. This problem will be treated in section 4.

3.1.2. The case of $h=0$, $f_1 \neq f_2$ and $g_1 \neq g_2$

Next, we condider the case of $h=0$, $f_2 \neq f_2$ and $g_1 \neq g_2$. The assumption $h=0$ is for convenience to draw a figure. In this case, the inequality (3.2) $(MSE(\hat{\delta}_2)-MSE(\hat{\delta}_1) \geqq 0)$ can be written as

$$(3.6)\quad [(1-f_1^2)^2(f_2g_2)^2-(1-f_2^2)^2(f_1g_1)^2]\gamma^2 + (1-f_1^2)(1-f_2^2)(f_2^2-f_1^2)\sigma^2 \geqq 0.$$

The following two inequalities give a sufficient condition which

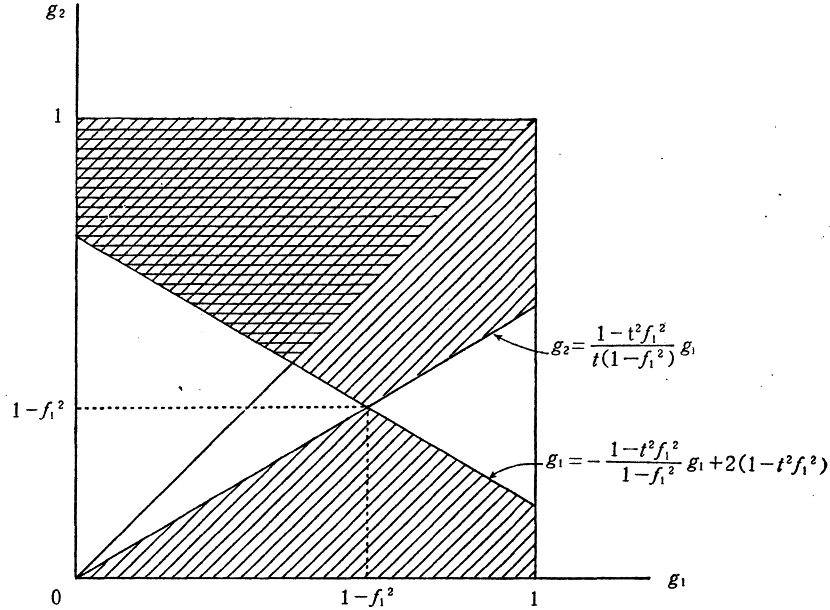

Figure 3.5

The part of oblique lines shows the region where (3.16) holds, *i.e.*, $\hat{\gamma}_1$ dominates $\hat{\gamma}_2$. The part of horizontal lines shows the paradoxial region where (3.13) holds.

satisfies (3.6).

(3.7) $(1-f_1^2)^2(f_2g_2)^2-(1-f_2^2)^2(f_1g_1)^2 \geqq 0$ and

(3.8) $f_2^2-f_1^2 \geqq 0$.

We assume that $0<f_1 \leqq f_2$ and $f_2=tf_1$ $(t \geqq 1)$ so that (3.8) may hold. Then, (3.7) is reduced to

(3.9) $g_2 \geqq [(1-t^2f^2)/\{t(1-f_1^2)\}]g_1$,

where $0<[(1-t^2f_1^2)/\}t(1-f_1^2)\}]<1$, since $0<f_1 \leqq f_2$ and $t \geqq 1$. Figure 3.3 shows the region where (3.9) holds. Here also, we can find the paradoxical region where (3.4) ($g_2 \geqq g_1$ and $MSE(\hat{\delta}_2) \geqq MSE(\hat{\delta}_1)$) holds.

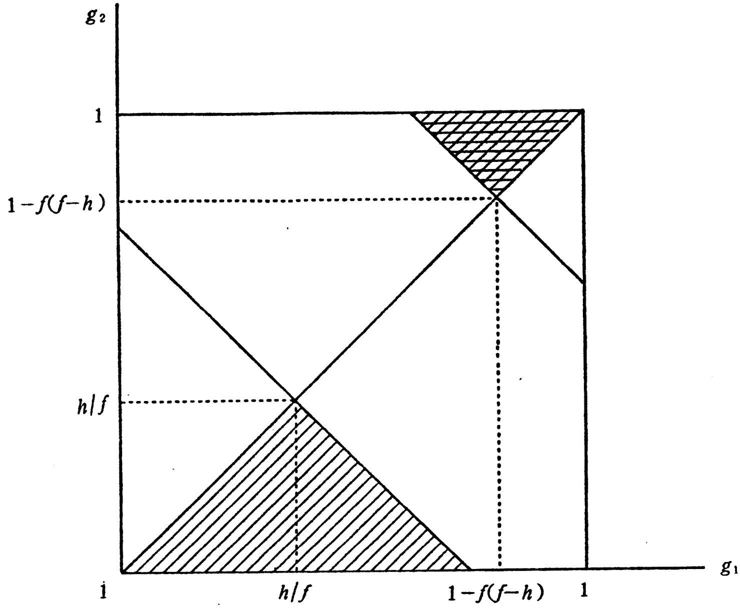

Figure 3.6

We assume $0<1-f(f-h)<1$. The part of oblique lines shows the r egion where (3.3) and (3.12) hold, *i. e.*, $\hat{\beta}_1$ dominates $\hat{\beta}_2$. The part of horizontal lines shows the paradoxical region where (3.17)' holds.

3.2. For MSE $(\hat{\gamma}_1)$

We here consider only MSE$(\hat{\gamma}_1)$ and assume that $\hat{\gamma}_1$ dominates $\hat{\gamma}_2$, *i. e.*,

(3.10) $MSE(\hat{\gamma}_2)-MSE(\hat{\gamma}_1)\geqq 0.$

This inequality is rewritten as follows using (2.7):

(3.11) $(1-f_1^2)^2\{f_2(f_2-h)+g_2-1\}^2-(1-f_2^2)^2\{f_1(f_1-h)+g_1-1\}^2]\gamma^2$
$+(1-f_1^2)(1-f_2^2)(-f_1^2+f_2^2)\sigma^2\geqq 0.$

3.2.1. The case of $f_1=f_2=f$ and $g_1\neq g_2$

We first examine the case of $f_1=f_2=f$ and $g_1\neq g_2$, In this case, (3.11) is reduced to

(3.12) $(g_2-g_1)[g_2+g_1-2\{1-f(f-h)\}\geqq 0.$

Figure 3.4 shows the region where (3.12) holds. This figure assumes $1-f(f-h)<1$. We can find that there also exists a paradoxical region such as

(3.13) $g_2\geqq g_1$ and $MSE(\hat{\gamma}_2)\geqq MSE(\hat{\gamma}_1)$.

3.2.2. The case of $h=0$, $f_1\neq f_2$ and $g_1\neq g_2$

Next, we investigate the case of $h=0$, $f_1\neq f_2$ and $g_1\neq g_2$. Then a sufficient condition for (3.11)($MSE(\hat{\gamma}_2)-MSE(\hat{\gamma}_1)\geqq 0$) is that the following inequalities hold good.

(3.14) $(1-f_1^2)\,|f_2^2+g_2-1|\geqq(1-f_2^2)\,|f_1^2+g_1-1|$ and

(3.15) $f_2^2-f_1^2\geqq 0$.

if we assume that $f_1\leqq f_2$ and $f_2=tf_1$ $(t\geqq 1)$ so that (3.15) may holds, (3.14) is reduced to

(3.16) $(1-f_1^2)\,|g_2-(1-t^2f_1^2)\,|\geqq(1-t^2f_1^2)\,|g_1-(1-f_1^2)\,|$

Figure 3.5 shows the case where (3.16) holds. Here we can also find a paradoxical region where (3.13) holds.

3.3. For MSE $(\hat{\beta}_i)$

Since MSE $(\hat{\beta}_i)$ is total of MSE $(\hat{\delta}_i)$ and MSE $(\hat{\gamma}_i)$, in order to find a sufficient condition of the region where $\hat{\beta}_1$ dominates $\hat{\beta}_2$, we have only to put the figure of MSE $(\hat{\delta}_i)$ upon that of $MSE(\hat{\gamma}_i)$. One example is shown in Figure 3.6, which treats the case of $f_1=f_2=f$ and $g_1\neq g_2$. It is clear that we here also can find a paradoxical region such as

(3.17) $g_2\geqq g_1$ and $MSE(\hat{\beta}_2)\geqq MSE(\hat{\beta}_1)$,

which is shown as the part of horizontal lines in Figure 3.6.

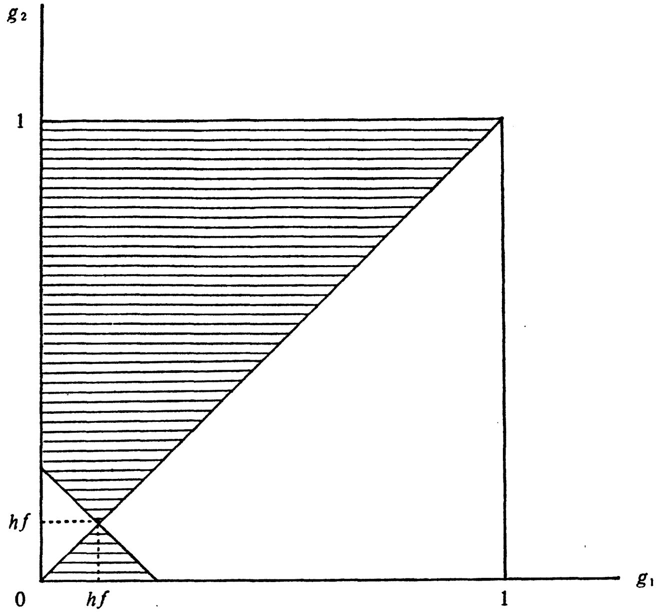

Figure 4.1 The part of horizontal lines shows the region where z_2 is richer than z_1.

4. Comparison of two proxy variabe models based on the partial correlatilon coefficients

As we mentioned in the previous section, we here regard the partial correlation coefficients as a criterion of richness of proxy variables instead of the simple correlation coefficients.

Before entering a discussion on a comparison of two proxy variable models, we rewrite the partial correlation coefficient according to our notations. It is well known that the partial correlation coefficient of two variables (say z and z_1) given a variable (say, x) is defined as the simple correlation coefficient of residuals after a regression by the common variable x. Therefore, the partial cor-

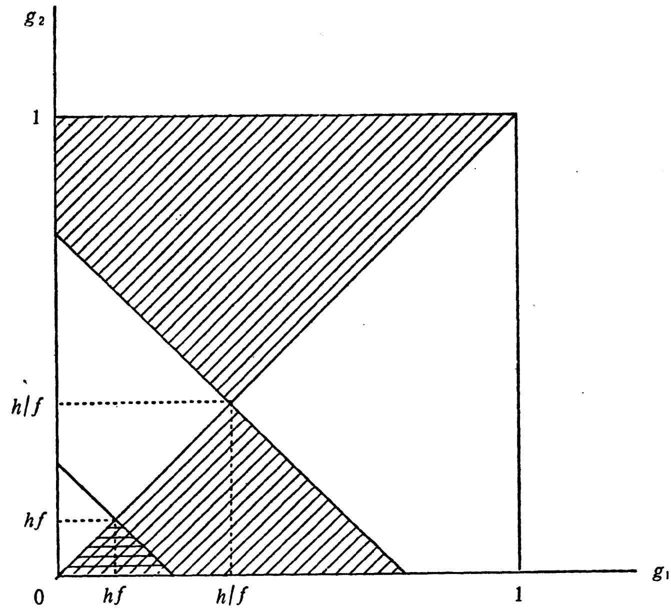

Figure 4. 2
We assume $0<h/f<1$. The part of oblique lines shows the region where $\hat{\delta}_1$ dominates $\hat{\delta}_2$. The part of horizontal lines shows the paradoxical region.

relation coefficient of z and z_1 given x, $R(zz_1 | x)$, can be written as follows according to our notations:

(4. 1) $R(zz_i | x)=(g_i-hf_i)/[(1-h^2)(1-f_i^2)]^{\frac{1}{2}}$.

4.1. Comparison of two proxy variable models as an extension of Frost and Ohtani's model

First of all, we check Ohtani's conclusion. Especially, Ohtani (1981) used the minimum *MSE* of conditional prediction as the criterion of selection of proxy variable, whereas the criterion of richness of proxy variables is to have greater partial correlation coefficients like ours.

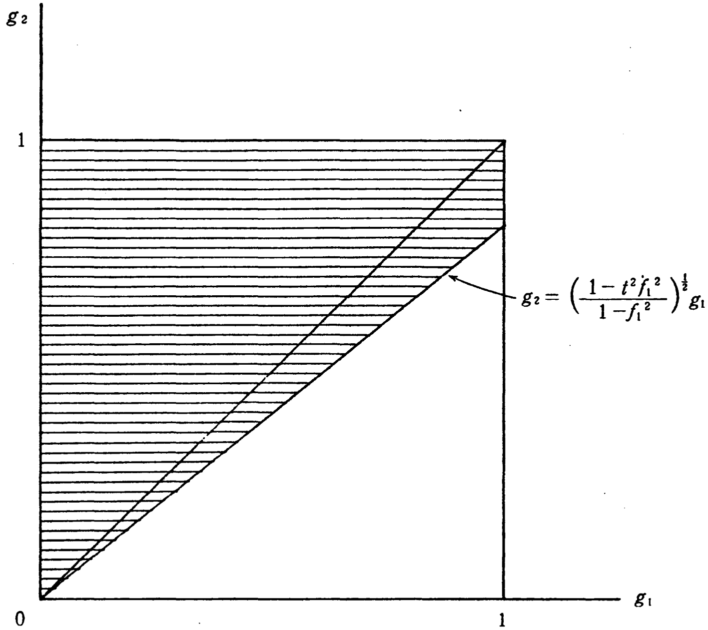

Figure 4.3
The part of horizontal lines shows the region where z_2 is richer than z_1.

According to Ohtani (1981) the MSE of the conditional prediction is written as follows for $i=1$ and 2:

(4.2) $MSE(\hat{y}_i)=\sigma^2[t^2_\gamma(1-R^2(zz_1\,|\,x))+2]$,

where $\hat{y}_1$ is a pcedictor such as $\hat{y}_1=x\hat{\delta}_1+z_1\hat{\gamma}_1$, t_γ is the ratio of the true γ to the standard error of its estimator, (i. e., $t_\gamma{}^2=\gamma^2(1-h^2)/\sigma^2$) and $R(zz_i\,|\,x)$ is the partial correlation coefficient for z and z_i given x. We can obtain from (4.2) the following relation between $R(zz_i\,|\,x)$ and $MSE(\hat{y}_i)$:

(4.3) $|R(zz_2\,|\,x)|\geqq|R(zz_1\,|\,x)|\Longleftrightarrow MSE(\hat{y}_2)\leqq MSE(\hat{y}_1)$.

(4.3) means that if the partial correlation coefficient is greater, the *MSE* of prediction is smaller, and this relation is reasonable and suits with our intution.

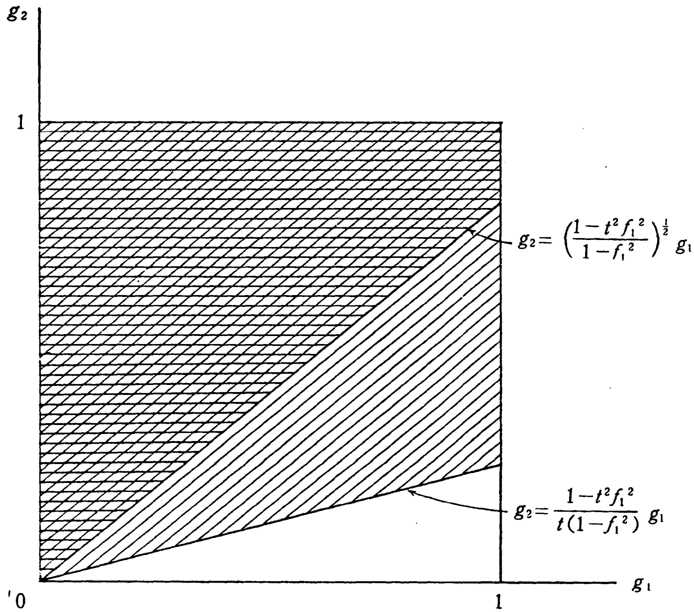

Figure 4.4
The part of oblique lines shows the region where $\hat{\delta}_1$ dominates $\hat{\delta}_2$. The part of horizontal lines shows the paradoxical region.

However, as we will show in the following sections, using the *MSE* of coefficients the opposite relation may exist.

4.2. For MSE($\hat{\delta}_i$)

4.2.1. The case of $f_1=f_2=f$ and $g_1 \neq g_2$

We consider the case of $f_1=f_2=f$ and $g_1 \neq g_2$ and assume z_2 is richer than z_1 ($R_2(zz_2 | x) \geqq R^2(zz_1 | x)$). This region can be defined as (4.4).

(4.4) $(g_2-g_1)(g_2+g_1-2hf) \geqq 0.$

Figure 4.1 shows the region where (4.4) holds. By the way, We already showed the region whese $\hat{\delta}_1$ dominites $\hat{\delta}_2$ in Figure 3.2.

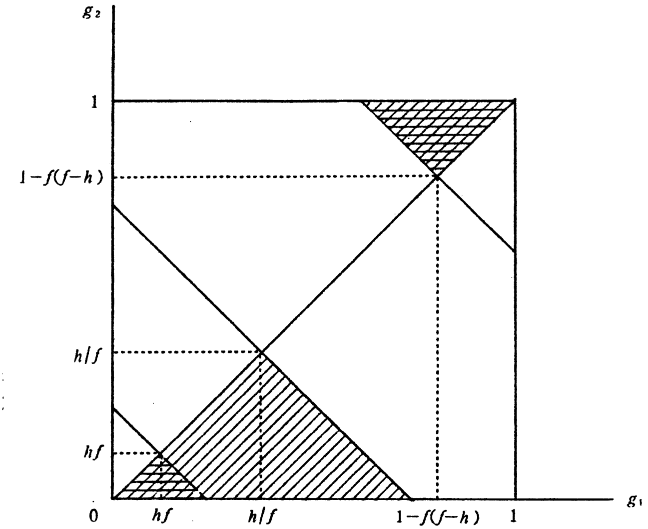

Figure 4.5
We assume $0<1-f(f-h)<1$. The part of oblique lines shows the region where $\hat{\beta}_1$ dominates $\hat{\beta}_2$. The part of horizontal lines shows the paradoxical region.

Therefore, if we look at these figures at the same time we can find a paradoxical region where $\hat{\delta}_1$ dominates $\hat{\delta}_2$ whereas z_2 is richer than z_1 (see Figure 4.2).

4.2.2. The case of $h=0$, $f_1 \neq f_2$ and $g_1 \neq g_2$

Here we also assume z_2 is richer than z_1 ($R^2(zz_2|x) \geqq R^2(zz_1|x)$). Let us define this region, assuming $0<f_1 \leqq f_2$ and $f_2 = tf_1$ ($t \geqq 1$)

(4.5) $g_2 \geqq [(1-t^2 f_1^2)/(1-f_1^2)]^{1/2} g_1$.

Figure 4.3 shows the region where (4.5) holds. We already showed the region where $\hat{\delta}_1$ dominates $\hat{\delta}_2$ in Figure 3.3. Therefore, we can also find a paradoxical region where $\hat{\delta}_1$ dominates $\hat{\delta}_2$ whereas

z_2 is richer than z_1 like the previous case (see Figure 4.4).

4.3. For MSE($\hat{\gamma}_1$) and MSE ($\hat{\beta}_1$)

In the case of $f_1=f_2=f$ and $g_1 \neq g_2$, the region where $\hat{\gamma}_1$ dominates $\hat{\gamma}_2$ is shown in Figure 3.4. and the region where z_2 is richer than z_1 is shown in Figure 4.1. Therefore, obviously there exists a paradoxical region.

In the case of $f_1 \neq f_2$ and $g_1 \neq g_2$, the region where $\hat{\gamma}_1$ dominates $\hat{\gamma}_2$ is shown in Figure 3.5. and the region where z_2 is richer than z_1 is shown in Figure 4.3. In this case, we also can find a paradoxical region.

As we have mentioned in Section 3.3, $MSE(\hat{\beta}_i)$ is total of MSE ($\hat{\delta}_i$) and MSE ($\hat{\gamma}_i$). Thus, in order to investigate a sufficient condition of the region where $\hat{\beta}_1$ dominates $\hat{\beta}_2$, we have only to put the figure of $MSE(\hat{\delta}_i)$ upon that of $MSE(\hat{\gamma}_i)$. One example of this region is shown in Figure 4.5. In this figure, we consider the case of $f_1=f_2=f$ and $g_1 \neq g_2$ as we showed in section 3.3. In this case, we also find the paradoxical case such that

(4.6) $R^2(zz_2 | x) \geqq R^2(zz_1 | x)$ and $MSE(\hat{\beta}_2) \geqq MSE(\hat{\beta}_1)$.

5. Concluding remarks

In this paper, we considered the problem of selecting proxy variables when there are two candidates for the true unobservable independent variable.

We found that we shouldselect a proxy variable depending on our purpose. That is to say, there may exist the region where the poorer proxy variable should be selected if our interest is the minimum MSE of the coefficient parmeters. Therefore, in this context, we should be very careful to use a proxy variable when we face a model which contains a unobservable variable, even if the proxy seems to be closely correlative to the true variable. Horever, when our interest is to get a better prediction using the estimated equation, the richer proxy (which is close to the true one) gives a better prediction.

Moreover, we might choose the proxy variable whose t-value is greater than the other's, no matter what our interest may be. This procedure, however, cheks only the correlation between the proxy variable and the dependent variable in the model, and has nothing to do with a richness of proxy variables itself.

Acknowledgements

The authors thank Professors T. Toyoda and K. Ohtani of Kobe University for their valuable comments.

References

Aigner, D. J., 1974, MSE dominance of least squares with errors-of-observation, *Journal of Econometrics* 2, 365-372.

Frost, P. A., 1979, Proxy variables and specifications bias, *The Review of Economics and Statistics 61*, 323-325.

McCallum, B. T., 1972, Relative asymptotic bias from errors of omission and measurement, *Econometrica 40*, 757-758.

Ohtani, K., 1981, On the use of a proxy variable in prediction: An *MSE* comparison, *The Review of Economics and Statistics 63*, 627-628.

Wickens, M. R., 1972, A note on the use of proxy variables, *Econometrica 40*, 759-761.