REGRESSION MODELS FOR THE COMBINATION OF TWO "INCOMPLETE" DATA SETS IN SECONDARY ANALYSIS

David Elesh, Temple University
James R. Taylor, University of Wisconsin

1.0. Introduction.

A problem frequently encountered in secondary analysis is that there is no one data file completely adequate to the researcher's needs. To be specific, data files which have detailed and comprehensive information on one topic required for an analysis may have only limited coverage of another. Suppose, for example, that one wanted to investigate the effects of health on labor supply and earnings using existing data resources. Perhaps the best available data on health is the National Health Survey, but it has only limited information on labor supply and earnings. On the other hand, the Michigan Survey of Income Dynamics has excellent coverage of labor supply and earnings and poor health data. How, then, are the files to be combined in an analysis? The traditional answer to this question is simply to analyze the data sets separately and to bridge their inadequacies with a variety of extrapolations, inferences, and "informed judgments" which the data may, to varying degrees, support. The problem with this approach is, of course, the difficulty of assessing the inferences made from it. As often as not, confidence in an author's conclusions comes more from the persuasiveness of his theoretical argument than from the weight of the empirical evidence behind it.

Consequently, we propose a different approach to the problem -- one which attempts to combine the best elements of two or more data sets into a single, analyzable file. For present purposes, we shall assume that such data sets are either samples of the same kind (identical probability, simple random, etc.) from the same population or censuses of the same population. To accomplish this combination of files, we shall also make certain distributional assumptions in the context of which a specific model will be estimated. We want to stress the importance of these assumptions at the outset since the validity of our approach depends directly upon their validity. The use of our approach requires prior investigation of the validity of the assumptions unless there is supporting a priori knowledge.

To clarify the exposition, we shall assume in our description of the approach that we wish to combine only two files, "adding" a variable from one file to the other so that regression or other statistical procedures may be performed which would include that variable. A generalization of the approach will be taken up at a later time.

We shall begin with an examination of a "complete data" model and a "restricted complete data" model, both containing p variables, all of which are jointly observed in the same sample. The information gained from this exercise will then be used to place constraints on an "incomplete data" model in which p-1 variables are jointly observed in one sample and the pth variable is "added" from a second and independent sample from the same population. Given these constraints, we shall then show that parameter estimates for the "incomplete" model (1) heuristically parallel that of the "complete data" model and (2) are maximum likelihood estimates.

1.1. The "Complete Data" Model.

Let us suppose that $\underline{Y}$ is a dependent variable of ultimate interest, $\underline{H}$ is the variable to be "added," $\underline{X}_{11}$ and $\underline{X}_{12}$ are sets of other variables, and $\underline{u}$ is a disturbance vector. The regression of $\underline{Y}$, given $\underline{H}$, on the remaining variable then can be written as

$$(1) \qquad (\underline{Y}|\underline{H}) = (\underline{H}\,|\,\underline{X}_{11}\,|\,\underline{X}_{12}) \begin{pmatrix} \theta_{10} \\ \theta_{11} \\ \theta_{12} \end{pmatrix} + \underline{u}$$

Let us also suppose that $\underline{Y}$ and $\underline{H}$ are jointly observed random variables (the $\underline{X}$'s are fixed) with the following likelihood function obtained from a random sampling of a bivariate normal distribution:

$$(2) \qquad \begin{pmatrix} \underline{Y} \\ \underline{H} \end{pmatrix} \sim BVN\left[\begin{pmatrix} \underline{X}_{11} & 0 & \underline{X}_{12} \\ 0 & \underline{X}_{11} & 0 \end{pmatrix} \begin{pmatrix} \underline{\theta}_{11} & \theta_{10}\underline{\beta} \\ & \underline{\beta} \\ & \theta_{12} \end{pmatrix}; \begin{pmatrix} (\sigma_1^2 + \theta_{10}^2\sigma_2^2)I_n & \theta_{10}\sigma_2^2 I_n \\ \theta_{10}\sigma_2^2 I_n & \sigma_2^2 I_n \end{pmatrix} \right]$$

From this construction of the joint distribution, we can see that the marginal distributions of the two variables are

$$(3) \qquad \underline{Y} \sim N\left[(\underline{X}_{11}\,|\,\underline{X}_{12})(\theta_{11} + \theta_{10}\underline{\beta}) \; ; \; (\sigma_1^2 + \theta_{10}^2\sigma_2^2)\, I \right]$$

$$(4) \qquad \underline{H} \sim N[\underline{X}_{11}\,\underline{\beta};\ \sigma_2^2 I]$$

while the conditional distribution for $\underline{Y}$, given $\underline{H}$, is

$$(5) \qquad (\underline{Y}|\underline{H}) \sim N\left[(\underline{H}\,|\,\underline{X}_{11}\,|\,\underline{X}_{12}) \begin{pmatrix} \theta_{10} \\ \theta_{11} \\ \theta_{12} \end{pmatrix};\ \sigma_1^2 I \right]$$

being definable from the ratio of the joint distribution to the marginal distribution for $\underline{H}$ (Graybill, 1961:63).

Now suppose only independent samples of the marginal distributions are observed. This we shall call the "restricted complete data" model. Equation (1) cannot be estimated because knowledge of the conditional distribution is necessary, and the marginal distributions do not contain sufficient information to identify the parameters of the conditional

distribution. To show this, we need only note that model (3) gives the estimate $\widehat{\underline{\theta}_{11} + \theta_{10}\underline{\beta}}$, $\underline{\hat{\theta}}_{12}$, and $\widehat{\sigma_1^2 + \theta_{10}^2\sigma_2^2}$ while model (4) gives estimators $\hat{\underline{\beta}}$ and $\hat{\sigma}_2^2$ -- all of which are clearly inappropriate in the sense that not all the required parameters are estimable since model (3) is over-parameterized.[1]

Our situation is thus analogous to the "incomplete data" model in that only the marginal distributions of our variables are known. Consequently, if a procedure can be found which permits estimation of model (1) where models (2) and (5) are unobservable it may also be applicable to the analysis of the "incomplete data" model.

What is needed are constraints which specify a relationship between the marginal and conditional (or jointly) distributions so that the latter can be identified from the available data in the former. Consider, for example, the following constraints:

$$\underline{C}'\underline{\theta}_{11} = 0; \quad \underline{C}'\underline{\beta} \neq 0 \tag{6}$$

where $\underline{C}$ is a vector of known constants. Reference to models (3) and (4) provides an interpretation for (6): it says that the vectors of variables, $\underline{X}_{11}$, do not have the same effect on $\underline{Y}$ as they have on $\underline{H}$. Since the number of cases in which a set of independent variables has different effects on two different dependent variables probably is larger than the number of cases in which the effects are the same, the constraints are not particularly restrictive.

Now in models (3) and (4) let

$$\underline{\ell}_{11} = \underline{\theta}_{11} + \theta_{10}\underline{\beta}$$

$$\underline{\ell}_{12} = \underline{\theta}_{12}$$

and

$$\eta^2 = \sigma_1^2 + \theta_{10}^2\sigma_2^2 .$$

The marginal equations for $\underline{Y}$ and $\underline{H}$ for two independent samples would then be

$$\underline{Y} = \left(\underline{X}_{11} \mid \underline{X}_{12}\right)\begin{pmatrix}\underline{\ell}_{11}\\ \underline{\ell}_{12}\end{pmatrix} + \underline{v}$$

$$\underline{H} = \underline{X}_{21}\underline{\beta} + \underline{w}$$

where we emphasize the independence of the marginal distribution by changing the first subscript of the X matrix in the equation for $\underline{H}$. The application of ordinary least squares procedures to these equations thus would give the maximum likelihood estimates $\hat{\underline{\ell}}_{11}$, $\hat{\underline{\ell}}_{12}$, $\hat{\eta}^2$, $\hat{\underline{\beta}}$, and $\hat{\sigma}_2^2$.

However, since by (6) $\underline{C}'\underline{\theta}_{11} = 0$, we can obtain estimates for each of the parameters of models (1) and (5) from them (Scheffe, 1959: 16-19):

$$\hat{\theta}_{10} = \underline{C}'\hat{\underline{\ell}}_{11} \,/\, \underline{C}'\hat{\underline{\beta}}$$

$$\underline{\hat{\theta}}_{11} = \hat{\underline{\ell}}_{11} - \hat{\theta}_{10}\hat{\underline{\beta}}$$

$$\underline{\hat{\theta}}_{12} = \hat{\underline{\ell}}_{12}$$

$$\hat{\sigma}_1^2 = \hat{\eta}^2 - \hat{\theta}_{10}^2\hat{\sigma}_2^2$$

by the invariance property of maximum likelihood estimators (Graybill, 1961: 36-37). Thus our constraints have permitted us to estimate all the parameters of the conditional model of $\underline{Y}$, given $\underline{H}$, even though only independent samples from the marginal distributions of the two variables were observed.

Moreover, since our estimators in (7) are maximum likelihood estimators, they will have the properties of consistency and asymptotic normality. They may, however, be biased. In particular, $\hat{\theta}_{10}$ is a "ratio estimator," and "ratio estimators" are rarely unbiased (Donahue, 1964). But since $\hat{\theta}_{10} = \underline{C}'\hat{\underline{\theta}}_{11} \,/\, \underline{C}'\hat{\underline{\beta}}$ for some a priori constraint vector $\underline{C}$, there may exist an optimal choice for $\underline{C}$. For example, one might choose $\underline{C}$ so that the mean square error for $\hat{\theta}_{10}$ is at a minimum. This would minimize the variance plus the square of the bias of $\hat{\theta}_{10}$, thus having the desirable effect of maximizing the predictive power of $\underline{H}$. However, further research on the optimal choice of $\underline{C}$ is necessary.

I.2. The "Incomplete Data" Model.

As noted earlier, the "incomplete data" model is analogous to the "restricted complete data" model in that only independent samples from the marginal distributions of $\underline{Y}$ and $\underline{H}$ are known. However, the former differs from the latter in that the sampling frames, sampling procedures, and administrative procedures may differ for the two data files to be combined whereas, in the latter case, where there is only a single parent population, these differences do not exist. Consequently, comparison of the two data files to be combined in these terms is a necessary part of validating the "incomplete data" model.

At the same time, it should be said that, despite these differences, the procedures of the preceding section could be used to solve models based on two data files ("incomplete data" models). However, it is instructive to examine an alternative procedure for solving "incomplete data" models. We shall show that, given our assumptions and constraints, this alternative procedure produces results identical to those in the preceding section.

We begin with a regression model for $\underline{Y}$, given $\underline{H}$, which has a somewhat different form from model (5) due to the fact that $\underline{Y}$ is observed in one sample and $\underline{H}$ in another:

$$(9)\qquad \left(\underset{\sim}{Y}^{n_1x1} \,|\, \underset{\sim}{H}_2^{n_2x1}\right) =$$

$$\left(\hat{\underset{\sim}{H}}_1^{n_1x1} \,;\, \underset{\sim}{X}_{11}^{n_1xp_1} \,;\, \underset{\sim}{X}_{12}^{n_1xp_2}\right) \begin{bmatrix} \theta_{10}^{1x1} \\ \underset{\sim}{\theta}_{11}^{p_1x1} \\ \underset{\sim}{\theta}_{12}^{p_2x1} \end{bmatrix} + \underset{\sim}{u}_1^{n_1x1}$$

where $\underset{\sim}{Y}$ is again the dependent variable, a column vector of n_1 observations; $\underset{\sim}{H}_2$ is the variable to be "added" observed in sample 2, a column vector of n_2 observations. $\hat{\underset{\sim}{H}}_1$ is the "added" measure, a column vector of n_1 observations; $\underset{\sim}{X}_{11}$ is a submatrix of independent variables n_1xp_1; $\underset{\sim}{X}_{12}$ is another submatrix of independent variables n_1xp_2; θ_{10} is the coefficient for $\hat{\underset{\sim}{H}}_1$; θ_{11} is the column vector (p_1x1) of coefficients for $\underset{\sim}{X}_{11}$; θ_{12} is the column vector (p_2x1) of coefficients for $\underset{\sim}{X}_{12}$; and $\underset{\sim}{u}_1$ is the column vector (n_1x1) of disturbances. The initial subscript indexes the sample in question; the second distinguishes subsets of independent variables. Since, except for $\hat{\underset{\sim}{H}}_1$, equation (9) represents a conventional regression model estimable by ordinary least squares, the next step is to determine $\hat{\underset{\sim}{H}}_1$ and its implications for the analysis.

To do this, let us suppose that $\underset{\sim}{H}_2$ is a measure available in the second sample but not in the first. Now suppose that there are a number of other variables which are common to both surveys. From this common list, we want to find the subset that will predict $\underset{\sim}{H}_2$ as well as possible. We may write this prediction equation as

$$(10)\qquad \underset{\sim}{H}_2^{n_1x1} = \underset{\sim}{X}_{21}^{n_2xp_1}\ \underset{\sim}{\beta}^{p_1x1} + \underset{\sim}{u}_2^{n_2x1}$$

where $\underset{\sim}{H}_2$ is the "added" variable in the second sample, a column vector n_2x1; $\underset{\sim}{X}_{21}$ is the subset of independent variables described in $\underset{\sim}{X}_{11}$ but measured in the second sample, a submatric n_2xp_1; $\underset{\sim}{\beta}$ is the column vector of coefficients, and $\underset{\sim}{u}_2$ is the disturbance vector (n_2x1); and where the disturbances, $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$, have a bivariate normal distribution with zero means, finite variances, and zero covariances.

$$\begin{pmatrix} \underset{\sim}{u}_1 \\ \underset{\sim}{u}_2 \end{pmatrix} \sim \text{BVN}\left[\begin{pmatrix} \underset{\sim}{0}^{n_1x1} \\ \underset{\sim}{0}^{n_2x1} \end{pmatrix} \begin{pmatrix} (\sigma_1^2 + \theta_{10}^2\sigma_2^2) I_{n_1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \sigma_2^2 I_{n_2} \end{pmatrix}\right]$$

Thus $\hat{\underset{\sim}{H}}_1$ can be computed as

$$(11)\qquad \hat{\underset{\sim}{H}}_1 = \underset{\sim}{X}_{11}\,(\underset{\sim}{X}_{21}'\underset{\sim}{X}_{21})^{-1}\underset{\sim}{X}_{21}'\underset{\sim}{H}_2 \equiv \underset{\sim}{X}_{11}\hat{\underset{\sim}{\beta}}\ .$$

Substituting equation (11) into equation (9) we get

$$(12)\qquad (\underset{\sim}{Y}|\underset{\sim}{H}_2) = [\underset{\sim}{X}_{11}\hat{\underset{\sim}{\beta}}\,;\underset{\sim}{X}_{11}\,;\underset{\sim}{X}_{12}] \begin{bmatrix} \theta_{10} \\ \underset{\sim}{\theta}_{11} \\ \underset{\sim}{\theta}_{12} \end{bmatrix} + \underset{\sim}{u}_1$$

$$= [\underset{\sim}{X}_{11}\,;\underset{\sim}{X}_{12}] \begin{bmatrix} \underset{\sim}{\theta}_{11} + \theta_{10}\hat{\underset{\sim}{\beta}} \\ \underset{\sim}{\theta}_{12} \end{bmatrix} + \underset{\sim}{u}_1$$

In words, what we have done has been to predict a variable, $\underset{\sim}{H}$, found in the second sample but not in the first from a list of variables common to both surveys; then using the coefficients from the equation run with the second sample data and the scores of the appropriate variables (the $\underset{\sim}{X}_{11}$) in the first survey, we produced a set of predicted values for the $\underset{\sim}{H}$ variable for the first sample. The result is equation (9).

However, equation (9) cannot be estimated as it stands because its coefficient matrix is singular as shown below:

$$[\underset{\sim}{X}_{11}\,(\underset{\sim}{X}_{21}'\underset{\sim}{X}_{21})^{-1}\underset{\sim}{X}_{21}'\underset{\sim}{H}_2\,;\underset{\sim}{X}_{11}\,;\underset{\sim}{X}_{12}]\ \cdot$$

$$\left[\begin{array}{c|c|c} 1^{1x1} & \underset{\sim}{0}^{1xp_1} & \underset{\sim}{0}^{1xp_2} \\ \hline -(\underset{\sim}{X}_{21}'\underset{\sim}{X}_{21})\bar{\underset{\sim}{X}}_{21}^{1}\underset{\sim}{H}_2 & I_{p_1} & \underset{\sim}{0}^{p_1xp_2} \\ \hline \underset{\sim}{0}^{p_2x1} & \underset{\sim}{0}^{p_2xp_1} & I_{p_2} \end{array}\right]$$

$$= [\underset{\sim}{X}_{11}\,(\underset{\sim}{X}_{21}'\underset{\sim}{X}_{21})^{-1}\underset{\sim}{X}_{21}'\underset{\sim}{H}_2 - \underset{\sim}{X}_{11}\,(\underset{\sim}{X}_{21}'\underset{\sim}{X}_{21})^{-1}\cdot$$

$$\underset{\sim}{X}_{21}'\underset{\sim}{H}_2\,;\underset{\sim}{X}_{11}\,;\underset{\sim}{X}_{12}] = \left[\underset{\sim}{0}^{n_1x1}\,;\underset{\sim}{X}_{11}^{n_1xp_1}\,;\underset{\sim}{X}_{12}^{n_1xp_2}\right].$$

Consequently, as was done with the "complete data" model, it will be necessary to place constraints on the coefficients.

A variety of constraints are, of course, possible. For example, we could set $\theta_{10} = 0$, but this has the rather ridiculous effect of asserting that $\underset{\sim}{Y}$ and $\underset{\sim}{H}$ are unrelated. Alternatively, we may employ the general constraint:[2]

$$a_1\theta_{10} + \sum_{j=1}^{p_1} c_j\theta_{11j} + \sum_{j=1}^{p_2} d_j\theta_{12j} = 0$$

where the a_1, c_j, and d_j are all known real numbers such that not all are zero. The specification of the above constraint will depend on the particular model being estimated and the reasonableness of any specific constraint in that context. For example, one

might choose

$$\sum_{j=1}^{p_1} \theta_{11j} = 0$$

$$\sum_{j=1}^{p_2} \theta 12_j = 0 .$$

In our particular case, we shall use neither or these but instead assume that $\underline{C}'\underline{\theta}_{11} = 0$. If we define the first column of the $\underline{X}_{11}$ matrix as a vector containing only unity for elements, then letting

$$\theta_{111} = 0$$

implies that

$$\underline{C}' = (1,0,0,\ldots,0)$$

and that $\underline{C}'\underline{\theta}_{11} = 0$ and $\underline{C}'\underline{\beta} \neq 0$. In other words, the intercept for the $\underline{X}_{11}$ matrix is set to zero. Model (9) can now be solved. The coefficient for $\hat{\underline{H}}_1$ is in fact $\hat{\theta}_{10}$; the last p-1 columns of $\underline{X}_{11}$ give the estimates, $\hat{\underline{\sigma}}_1^2$; and the coefficients of $\underline{X}_{12}$ are the $\hat{\underline{\theta}}_{12}$. The estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be computed from the residual sum of squares for equations (9) and (10).

Comparison of equations (9), (10), (12), with (3) and (4) show that our estimates for θ_{10}, $\underline{\theta}_{11}$, and $\underline{\theta}_{12}$ are the same for both the restricted "complete data" and "incomplete data" models. Furthermore, comparing the "incomplete data" model (9) with the "complete data" model (5), it can be seen that (9) is an approximation of (5) in the sense that $\hat{\underline{H}}_1$ has been used as a substitute for $\underline{H}_1$. The constraint $\underline{C}'\underline{\theta}_{11} = \theta_{111} = 0$ was here chosen for illustrative purposes only, and may not be optimal in terms of the mean square error of $\hat{\theta}_{10}$ criterion previously mentioned. However, reference to models (1), (3) and (4) provides an easy interpretation of the constraint $\theta_{111} = 0$. That is to say $\theta_{111} = 0$ implies that the intercept term of model (3) is equal to the product of θ_{10} and the intercept term of model (4), and that the intercept term of model (1) is zero. Because of our distributional assumptions, the estimates are also maximum likelihood estimates. We should note, however, that we do not know how sensitive these estimates are to violations of the assumptions of distributional normality made here. Testing these assumptions will be one of our first research tasks.

FOOTNOTES

[1] Specifically, model (3) contains one too many parameters for unique estimators to exist unless an additional constraint is imposed. Since model (3) is induced from the bivariate model (2), and models (2) and (4) lead to model (5); it is necessary that any such constraint on model (3) be compatible with models (2) and (5). An appropriate constraint is given below.

[2] This alternative constraint derives from a comparison of models (3) and (12). The original constraint was applied to model (3) is identical to the final form of the parameter matrix of (12). The alternative constraint is applied to the initial parameter matrix of model (12) and is thereby equivalent to original constraint.

REFERENCES

Donahue, James D., Products and Quotients of Random Variables and their Applications, Washington, D.C.: Office of Technical Services, U.S. Department of Commerce, Aerospace Research Laboratories Technical Report, ARL 64-115.(1964).

Graybill, Franklin A., An Introduction to Linear Statistical Models, New York: McGraw-Hill, vol. 1. (1961).

Scheffe, Henry, The Analysis of Variance, New York: Wiley. (1959).