Summary

This thesis consists of four self-contained essays. They are, however, interrelated in several ways. The most basic connection is that they are all concerned with the proper stochastic specification of a model. More specifically, we have tried to integrate the stochastic specification with the rest of the structure by regarding the systematic part of the residuals as additional explanatory variables which happen to be unmeasured. Then we must specify how the observed and unobserved variables are related to each other.

The major precedent for this approach is the work by Marshak and Andrews (1944), Mundlak (1963), and Mundlak and Hoch (1965) on the specification of micro production functions. They regarded the residual in the production function as made up of unmeasured inputs such as the "entrepeneurial capacity" of the firm. Although these inputs are unknown and regarded as random by the econometrician, they may or may not be random to the firm. The answer to that question determines how the unobservable is related to the observed factors of production.

Another connecting thread in these essays is an attempt to identify the structural relationship between an individual's wages and his characteristics. The relationship is

$$Y_{1i} = \underset{\sim}{x}'\underset{\sim}{\beta}_3 + \gamma_{23}S_i + \lambda_3 f_i + \nu_{3i}, \quad i = 1,\ldots, q, \tag{1}$$

where Y_1 is the log of earnings in year one, S is years of schooling, f is the systematic part of the residual, reflecting unobserved characteristics, and ν_3 is transitory income. For the moment we will ignore the other observed

characteristics in $\underset{\sim}{x}$. To complete the stochastic specification we need to model the relationship between f and S. It is approximated by

$$(2) \qquad S_i = \underset{\sim}{x}_i{}'\underset{\sim}{\beta}_2 + \lambda_2 f_i + \nu_{2i} \quad .$$

We suspect that λ_2 is non-negligible because although f is random to the external observer, it is known to the individual and forms the initial conditions that he faces in deciding how much schooling is right for him. ν_2 contains other characteristics that are not relevant for wage determination.

The observable characteristics in $\underset{\sim}{x}$ might include family background measurements such as father's schooling or occupation. We will take $\underset{\sim}{x}$ to be independent of f by construction. This means reinterpreting f as the part of the unobserved characteristics that is not predictable from $\underset{\sim}{x}$. Of course this affects our interpretation of $\underset{\sim}{\beta}$. For example, if x is mother's schooling and f includes genetic ability, then the reinterpreted β reflects both the return to the mother's pre-school investment in the child and the spurious effect of mother's education as a proxy for the initial ability of the child. Of if x is father's income, then even if it has no direct effect on the son's earnings, our reinterpreted β will not be zero.

In order to separate the structural effects of x from the proxy effects, we would have to relate mother's and father's observed characteristics to their unobserved characteristics, f' and f". The we would allow f' and f" to be correlated both with each other (assortative mating) and with f. But this more complicated model is irrelevant if all we want to estimate is the return to the son's schooling. For the γ's are not affected by the way in which we divide up the joint effect of $\underset{\sim}{x}$ and f.

A general setting for these models is provided in Chapter 2. There we study the identification of systems which are triangular but fail to be recursive because the residuals from the different equations contain common omitted variables. The identification problem is approached as the first step in an estimation problem. We want to describe a likelihood function, for example in terms of its mode and some measures of dispersion. But first we would like to know if the maximum of the likelihood corresponds to a unique vector of structural parameters. If not we have multiple peaks, a ridge or a pleateau, and the problem is to describe ML regions for the structural parameters.

Clearly the model in (1.2) is not identified. A plausible source of additional information would be another measurement of earnings:

$$Y_{2i} = \underset{\sim}{x}_i{}'\underset{\sim}{\beta}_4 + \gamma_{24}S_i + \gamma_{34}Y_{1i} + \lambda_4 f_i + \nu_{4i} \quad . \tag{3}$$

But in fact the model remains unidentified no matter how many measurements of this kind we have. And this is true even if γ_{34} equals zero; e.g. if there is enough time between the measurements so that they do not have a transitory piece in common.

More promising would be the availability of an early (pre-school) test score:

$$T_i = \underset{\sim}{x}_i{}'\underset{\sim}{\beta}_1 + \lambda_1 f_i + \nu_{1i} \quad . \tag{4}$$

If T is excluded from all of the other equations, then the model is (in general) identified provided there is one additional restriction besides those implied by the triangular structure. But if there are no other restrictions (i.e. $\gamma_{34} \neq 0$), then the ML estimate is a region. It turns out that we can uniquely solve for the other parameters once we know $\rho = 1 - \sigma_{\nu_1}{}^2/\sigma_T{}^2$, the reliability of T. The ML region for the other parameters is generated by the following ML interval for ρ: $0 \leq \rho \leq R^2_{T.\underset{\sim}{x},S,Y_1,Y_2}$.

A hard question in this model is whether the combination of omitted characteristics that ties together the income and schooling residuals is the same combination that connects the schooling and test residuals. There is a straightforward answer under a narrow measurement error interpretation of f. Then ν_1 is interpreted as a test-retest error that could in principle be eliminated by replicating the test. So it is reasonable to assume that ν_1 is independent of everything else and there is clearly just one f, namely the systematic part of the test (the "true score") that is not captured by $\underset{\sim}{x}'\underset{\sim}{\beta}_1$.

There is, however, an alternative more general interpretation of f. It is that IQ tests are designed to predict academic performance and need not capture (or appropriately weight) the set of characteristics relevant for economic success. This suggests having two distinct but correlated unobservables, f_1 and f_2. f_1 reflects the weighting of the omitted characteristics relevant for predicting economic success and f_2 reflects the weighting appropriate for scholastic achievement. Then f_1 is excluded from the S equation, f_2 is excluded from the Y equations, and neither is excluded from the T equation. Both of these interpretations of f are pursued in our empirical application in Chapter 4. It is based on the 1964 CPS-NORC veteran's data, which has previously been studied by Griliches and Mason (1972) and Duncan (1968), among others.

So one source of identification is the availability of additional relationships which contain the omitted characteristics. A related source is an appropriate grouping device. The use of grouping methods in errors-in-variables contexts goes back to Wald (1940) and to the empirical work of

Friedman (1957) and Eisner (1958). One novelty of our approach is that the unobservable need not be constant within the group. For example, let

$$
\begin{aligned}
(5)\qquad S_{ij} &= \underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_2 && + \lambda_2 a_{ij} + \nu_{2ij} \\
Y_{1ij} &= \underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_3 + \gamma_{23}S_{ij} + && + \lambda_3 a_{ij} \quad \nu_{3ij} \\
Y_{2ij} &= \underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_4 + \gamma_{24}S_{ij} + \gamma_{34}Y_{1ij} && + \lambda_4 a_{ij} + \nu_{4ij}, \quad \begin{matrix} i=1,\dots q \\ j=1,\dots p \end{matrix}
\end{aligned}
$$

where the subscripts refer to the jth individual in the ith group. This grouping will buy us something if the systematic part of the residuals (a_{ij}) has a group structure while the equation specific effects (ν_{ij}) do not. Regarding the a_{ij} as a set of pq "nuisance" parameters makes it clear that any prior information we can apply to them will be very useful. It seems reasonable to use the following representation for our prior $a_{ij}=f_i+g_{ij}$ with f_i randomly distributed across groups and the g_{ij} randomly distributed within groups. So we are connecting the residuals from the different equations via a common systematic factor which has a variance components structure of the sort used by Balestra and Nerlove (1966).

We could, of course, regard each member of a group as a separate equation and return to our earlier framework with p factors wich are themselves correlated via their dependence on one common factor. But the replication case is sufficiently important that we have devoted Chapter 3 to developing it in some generality. For example, it is no longer necessary to have an equation such as T which contains the unobservable but excludes S. In fact (5) is identified provided there is one restriction in addition to those implied by the triangular structure.

This is similar to the identification condition for the model which has an early test score but no replication. In fact a comparison of Theorem 4 in Chapter 2 with Theorem 1 in Chapter 3 shows that the identification problems in the two models are identical. So in the unidentified case we again have a simple description of the ML region. Now $\lambda = \sigma_f^2/(\sigma_f^2 + \sigma_g^2)$ is the key parameter. Given λ the reduced form can be uniquely solved for the other structural parameters. Then the ML region is generated by the following ML interval for $\lambda : 0 \leq \lambda \leq T^2$ where $T^2 = (\psi - \frac{1}{p})/(1 - \frac{1}{p})$ and ψ is the largest squared canonical correlation of the endogenous variables with a set of group indicator dummy variables (If there are x's then the endogenous variables are replaced by an appropriate set of residuals). If there is no group structure then ψ is 1/p. T^2 is the fraction of the unexplained variance which is accounted for by the group structure.

In our empirical application of the non-replication models we are able to reduce some of the ML problems to standard LIML calculations or to Hannan's (1967) extension of LIML. In other versions of the model the likelihood function is relatively intractable and we have followes Jöreskog and Goldgerger (1973) in adapting a numerical minimization program by Jöreskog (1970) to our problems. But we show in Chapter 3 that considerable analytic concentration of the likelihood function is possible in the replication models. Some of our algorithms can be interpreted as a canonical correlation procedure, others as constructing a proxy for the unobservable and including it in a regression. We show how our procedures generalize the more familiar single equation variance components pooling of the within and between group information. In addition we describe the computational and interpretational differences in a fixed vs a random effects treatment of the unobservable.

In Chapter 5 we present an application of these techniques, using data on brothers to control not only for between family parental background differences but also for individual within family differences which may be correlated with achieved schooling levels later on. We also make some attempts to explore the sensitivity of the results to the one factor assumption, obtaining ML regions in the two factor case.

The common focus of our examples and applications on one empirical problem has the advantage of providing these essays with some additional unity. But it has the disadvantage of suggesting, I believe incorrectly that our approach is limited to the stochastic specification of human capital models. So our concluding chapter, in addition to making connections to the literature and suggesting extensions, will sketch an application to a combined time-series cross-section analysis of individual firm production and factor demand relations. Thus the conclusion will link back to the major precedent for our approach.

UNOBSERVABLES IN ECONOMETRIC MODELS

A thesis presented

by

Gary Edward Chamberlain

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

January 1975

7696
1-2

to

Charlotte

Acknowledgements

My thesis committee consisted of Martin Feldstein, Edward Leamer, and Zvi Griliches. Each of them, in different ways and at different times during my undergraduate and graduate years at Harvard, gave freely of their insights and experience. My particular debt for this thesis is to Zvi Griliches. He was kind enough to devote a good deal of his time to specific discussions of the work as it evolved, but perhaps more important was his ability to communicate his enthusiasm for specifying and solving empirical problems.

I would like to thank Claire Kelleher and Debbie Hirsch for what seemed to be an endless job of typing. For financial support, I am indebted to the National Science Foundation (NSF Grant No. GS 2762X and an NSF Fellowship in Economics).

Finally I would like to thank my wife Charlotte for the depths of patience we scarcely knew she had.

Table of Contents

Chapter 1

Introduction

This thesis consists of four self-contained essays. They are, however, interrelated in several ways. The most basic connection is that they are all concerned with the proper stochastic specification of a model. More specifically, we have tried to integrate the stochastic specification with the rest of the structure by regarding the systematic part of the residuals as additional explanatory variables which happen to be unmeasured. Then we must specify how the observed and unobserved variables are related to each other.

The major precedent for this approach is the work by Marshak and Andrews (1944), Mundlak (1963), and Mundlak and Hoch (1965) on the specification of micro production functions. They regarded the residual in the production function as made up of unmeasured inputs such as the "entrepreneurial capacity" of the firm. Although these inputs are unknown and regarded as random by the econometrician, they may or may not be random to the firm. The answer to that question determines how the unobservable is related to the observed factors of production. For example, firm effects representing unmeasured fixed inputs are likely to be taken into account by the entrepreneur in making his factor demand decisions. The firms with more of the fixed inputs (under decreasing returns to the variable factors) use more of the variable inputs, and so part of the production function residual is transmitted to the factor demand equations.

The general model that we work with is

$$(1) \qquad \underset{\sim}{y}_i{}'\underset{\sim}{\Gamma} + \underset{\sim}{x}_i{}'\underset{\sim}{B} = \underset{\sim}{f}_i{}'\underset{\sim}{\Lambda} + \underset{\sim}{\nu}_i, \qquad i=1,\ldots,q,$$

where $\underset{\sim}{y}_1$ is an m x 1 vector of endogenous variables, $\underset{\sim}{x}_i$ is an n x 1 vector of exogenous variables, $\underset{\sim}{\Gamma}$ is an upper triangular matrix of parameters with ones on the diagonal, $\underset{\sim}{B}$ is a parameter matrix, and there are q observations. We have examined the identification and estimation of this model and have applied it in two empirical studies of the structural relationship between an individual's wages and his characteristics.

The residuals in (I.1) are assumed to be independent across observations. If they were also independent across equations then the model would be recursive and readily identifiable. Conversely, if the residuals were freely correlated across equations then the standard Cowles Commission results would apply. Our interest is in the intermediate cases where some but not all of the identification comes from covariance restrictions on the residuals. They are assumed to have a factor analytic structure where $\underset{\sim}{f}_i$ is a vector of latent variables and $\underset{\sim}{\Lambda}$ is a matrix of coefficients (factor loadings). The unobservable $\underset{\sim}{f}_i$ are distributed as a multivariate random sample. $\underset{\sim}{\nu}_i$ is a vector of equation specific effects which are distributed independently of $\underset{\sim}{f}$ as a random sample with covariance matrix $\underset{\sim}{U} = \text{diag}\ \{\sigma_1{}^2, \ldots, \sigma_m{}^2\}$.

This model is useful in a wide variety of micro-econometric applications. Examples include studies of social mobility and the determinants of socioeconomic achievement. The triangular structure arises from making measurements on an individual's characteristics at a particular time. Then the measured variable becomes a characteristic which can determine subsequent measurements. $\underset{\sim}{x}$ and $\underset{\sim}{f}$ are a set of characteristics which potentially affect all subsequent observations. The distinction between them is that $\underset{\sim}{f}$ is unobservable. The assumed independence of $\underset{\sim}{x}$ and $\underset{\sim}{f}$ simply means that we interpret $\underset{\sim}{f}$ to be the part of the unobservable characteristics that is not predictable from $\underset{\sim}{x}$. This of course affects our interpretation of $\underset{\sim}{B}$ and limits the restrictions we can impose on $\underset{\sim}{B}$. For example, x_1 may have no effect on y_k if all other relevant characteristics are included. But if the partial correlation is non-zero (partialling on the other included x's), then with our interpretation of $\underset{\sim}{f}$ we cannot exclude x_1 from that equation. $\underset{\sim}{\Gamma}$, however, is unaffected by the way in which we divide up the joint effect of $\underset{\sim}{x}$ and $\underset{\sim}{f}$.

The identification problem in this model can be approached from at least two points of view. The traditional one is to ask "What are the limits of observational information?" If the reduced form parameters are known with certainty, what aspects of the structure can we uncover? An alternative approach, which I prefer, is to treat the identification problem as one apsect of investigating a likelihood function. We typically start by investigating the mode and then proceed to examine measures of dispersion. But a logically prior question is whether the maximum of the likelihood corresponds to a unique vector of structural parameters. If not, then we have multiple peaks, a ridge, or a plateau, and the problem is to describe ML regions for the structural parameters.

The general treatment of model (1) remains an elusive goal. Chapter 2 is confined to the one factor case, but even then a complete identification analysis is not available except for special cases. We do, however, have some useful necessary conditions, and in addition a set of sufficient conditions which provide a constructive method for obtaining the structural parameters from the reduced form.

In the one factor case ($\underset{\sim}{f}_i{}'\underset{\sim}{\Lambda} = f_i\underset{\sim}{\lambda}'$ where $\underset{\sim}{\lambda}$ is m x 1), it is clear that at least m restrictions are necessary for identification. For example, if all of the m factor loadings are zero then the model is recursive. Our first two Theorems in Chapter 2 place conditions on the way in which zero restrictions on $\underset{\sim}{B}$ and $\underset{\sim}{\Gamma}$ must be allocated, both across the equations (vertically) and across the variables (horizontally). Theorem 1 shows that for each $k \leq m$ there must be at least k restrictions on the last k equations. Theorem 2 shows that for each $k \leq m$ there must be at least k restrictions, each of which excludes an x or one of the following variables from an equation: $y_1, \ldots, y_k$.

The basic idea behind our sufficient condition is to use a proxy for the unobservable f and then solve the resulting errors-in-variables problem by finding a suitable instrument. For consider the following example:

$$\begin{aligned} (2) \qquad y_1 &= \qquad\qquad\qquad\quad \lambda_1 f + \nu_1 \\ y_2 &= \qquad\qquad\qquad\quad \lambda_2 f + \nu_2 \\ y_3 &= \gamma_{13} y_1 + \gamma_{23} y_2 + \lambda_3 f + \nu_3 \\ y_4 & \qquad\qquad \gamma_{24} y_2 + \lambda_4 f + \nu_4 \quad . \end{aligned}$$

We can use y_1 as a proxy for f in the y_3 equation:

$$(3) \qquad y_3 = (\gamma_{13} + \frac{\lambda_3}{\lambda_1}) \, y_1 + \gamma_{23} y_2 + \nu_3 - \frac{\lambda_3}{\lambda_1} \, \nu_1 \quad .$$

This results in a standard errors-in-variables problem due to the "measurement error" in y_1. It can be cured by using y_4 as an instrument for y_1. Similarly y_3 can be used as an instrument for y_1 in the y_4 equation. But complications arise when more than one variable needs an external instrument. For then the instrumental variable (IV) normal equations need not have full rank. Also we must be careful that our choice of proxy does not contaminate the coefficients of interest. For example, in (I.3) γ_{13} is contaminated by the use of y_1 as a proxy. These problems are dealt with in Theorem 3 by giving a sufficient condition for a parameter to be estimable from the IV equations.

Our use of endogenous variables as instruments is similar to Hurwicz's (1946) suggestion to use lagged values of an error-ridden variable as instruments of a time series context. Also Liviatan (1963) used past and future values of consumption and income as instruments for measured income.

A special case for which we do have a general analysis sets $\gamma_{1k} = 0$ for $k > 1$. So y_1 is excluded from all of the other equations. This case is of special interest because it allows us to make a substantive distinction between the concepts of "measurement error" and "unobservable". $y_1 = \lambda_1 f + \nu_1$ in (2) is a typical measurement error equation, particularly if we scale f so that $\lambda_1 = 1$. Then we can interpret f as the "permanent" or "systematic" part of y_1. But if y_1 itself appears in some other equation along with f, then we are including both the measured variable and the "true" variable. Now there are cases in which this may be reasonable, e.g. a measured test score may have a credential or certification effect above and beyond the "true score". But in general if $\lambda_{1k} \neq 0$ we will not want to regard y_1 as measuring the unobservable f subject to error.

This errors-in-variables specialization of (1) is a special case of the Geraci and Goldberger (1971) and Geraci (1974) models in that $\underset{\sim}{\Gamma}$ is triangular. But it is more general in that part of the identification is coming from restrictions on the residual covariance matrix. Geraci and Goldberger assume that ν_1, the measurement error, is independent of everything else, but they allow the other ν's to be freely correlated. So they are confined to using x's as instruments whereas I can potentially use y's as instruments. Their results are similar to mine in that the identification of the different equations is tied together. Theorem 4 in chapter 2 shows that the entire structure is identified by a single zero restriction on $\underset{\sim}{\Gamma}$ or $\underset{\sim}{B}$ provided a rank condition holds. For then with y_1 as a proxy for f, there is some equation with an excluded variable that can be used as an instrument for y_1. This amounts to subtracting off $\sigma_1{}^2$ from $\sigma_{y_1}{}^2$, purging y_1 of the measurement error. But then the purged y_1 can be used in the other equations as an exact proxy for f. The sufficient rank condition is that the exclusion occur in an equation in which f actually appears, and that the excluded variable appear (with a non-zero coefficient) in an equation containing f preceding the one it is excluded from.

Another special feature of this errors-in-variables special case is that we can give a complete answer to <u>both</u> parts of the identification problem. For in addition to necessary and sufficient conditions for the likelihood function to have a unique maximum, we have a simple description of the ML region in the unidentified case. We can uniquely solve for the other parameters once we know $\rho = 1 - \sigma_1{}^2/\sigma_{y_1}{}^2$, the reliability of y_1. It is shown in Chapter 6 that the ML region for the other parameters is generated by the following ML interval for ρ: $0 \leq \rho \leq R^2_{y_1 .\, \underset{\sim}{x},\, y_2, \ldots, y_m}$. We have put this result under "Extensions" since we are just beginning to develope useful bounds of this sort.

If the standard Cowles Commission model without restrictions on the structural residual covariance matrix is not identified, then typically the ML intervals are unbounded and do not contain any useful information. But in our model the use of "unidentified" is somewhat misleading. For we do have identification in the sense of a non-trivial bound. The use of bounds in errors-in-variables models goes back to Frisch (1934) who pointed out that the appropriate weighted regression could be bounded by the elementary regressions. This solid angle bound was proved very laboriously by Reirsol (1945), more directly by Dhondt (1960), and recently quite elegantly by Keller (1973) using the spectral properties of positive matrices. A related bound, which can also be found in Frisch, is used by Harberger (1953).

So one source of identification is the availability of additional relationships which contain the unobservable. A related source is an appropriate grouping device. The use of grouping methods in errors-in-variables models can be found in Wald (1940) and in the empirical work of Friedman (1957) and Eisner (1958). One novelty of our approach is that the unobservable need not be constant within the group. In Chapter 3 we study the identification and estimation of the following replication model:

(4) $\underset{\sim}{y}_{ij}'\underset{\sim}{\Gamma} + \underset{\sim}{x}_{ij}'\underset{\sim}{B} = \underset{\sim}{f}_i'\underset{\sim}{\Lambda} + \underset{\sim}{\nu}_{ij}$, $i=1,\dots,q;\ j=1,\dots, p$,

where the subscripts refer to the jth observation in the ith group. The residuals are assumed to have a multivariate variance components decomposition: $\underset{\sim}{f}_i$ is a vector of random group effects and $\underset{\sim}{\nu}_{ij}$ is a vector of individual effects which are distributed independently of $\underset{\sim}{f}_i$ as a random sample over i and j with covariance matrix $\underset{\sim}{V}$. A variety of cases are considered. The most interesting identification results are for the one factor model ($\underset{\sim}{f}_i'\underset{\sim}{\Lambda} = f_i\underset{\sim}{\lambda}'$) with $\underset{\sim}{V} = \tau\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U}$ where $\underset{\sim}{U}$ is a diagonal matrix of equation specific residual variances. This case arises when we assume that there is a common left out variable a_{ij}. Then we introduce a prior for the a_{ij} which has the following variance components representation: $a_{ij} = f_i + g_{ij}$, where the f_i are distributed as a random sample across groups and the g_{ij} are a random sample within groups with $\tau = \sigma_g^2/\sigma_f^2$.

So we are taking the variance components specification studied by Balestra and Nerlove (1966), Wallace and Hussein (1969), Maddala (1971), Nerlove (1971), and Mazodier (1971) and embedding it in a larger system. A common complaint lodged against the random effects specification relative to a fixed effects approach is that the independence of the random effects from the observable explanatory variables is often implausible. For example the firm effects in a production function are unlikely to be independent of the variable inputs. But part of the variance components specification is quite plausible. The random sample view of the f_i amounts to adding an exchangeable prior to a set of fixed effects dummy variables. The prior is exchangeable if its form is unaffected by permuting the f's, so that the i subscript is just a labeling device with no substantive content (de Finetti, 1937). This is often appropriate at the level of individuals , families, or homogeneous firms. Similarly the g_{ij} are assumed to be exchangeable within the groups. So the problem is to keep the persuasive marginal prior distribution for the a_{ij} without making implausible independence assumptions about the joint distribution of a_{ij} and the observable variables. We accomplish this by building in the dependence by embedding a_{ij} in a simultaneous system.

Our principal result on the identification of this model is contained in Theorems 1 and 2 in Chapter 3. The necessary and sufficient condition for identification from zero restrictions on $\underset{\sim}{\Gamma}$ or $\underset{\sim}{B}$ is that there must be at least one exclusion which occurs in an equation that contains f and for which the excluded variable appears in a preceding equation that contains f. This is very similar to the condition in the errors-in-variables special case of (1) $(\gamma_{1k} = 0$ for $k > 1)$. In fact the identification problems in the two models are formally identical! The availability of replication converts a general unobservables model into the errors-in-variables special case, with its much simpler analysis. So not surprisingly we also have a complete analysis of the "unidentified" case. Now $\lambda = \sigma_f^2/(\sigma_f^2 + \sigma_g^2)$ is the key parameter. Given λ the reduced form can be uniquely solved for the other structural parameters. Then the ML region in generated by the following ML interval for λ: $0 \leq \lambda \leq T^2$ where $T^2 = (\psi - \frac{1}{p})/(1 - \frac{1}{p})$ and ψ is the largest squared canonical correlation of the endogenous variables with a set of group indicator dummy variables (if these are x's then the endogenous variables are replaced by an appropriate set of residuals). If there is no group structure then ψ is $\frac{1}{p}$. So T^2 is the fraction of the unexplained variance that is accounted for by the group structure. It is the appropriate generalized R^2 for this problem.

Our work on estimation has mostly been devoted to ML algorithms for the replication model (4). For example, in the one factor model ($\underset{\sim}{f}_i{}'\underset{\sim}{\Lambda} = f\underset{\sim}{\lambda}'$) with $\underset{\sim}{\Gamma} = \underset{\sim}{I}$ (no simultaneity problem) and with $\underset{\sim}{V}$ unrestricted so that the equation specific effects are freely correlated, the ML estimator of $\underset{\sim}{\lambda}$ conditional on $\underset{\sim}{B}$ can be obtained from a canonical correlation analysis of the residuals and a set of group indicator dummy variables. In fact, is is the same canonical correlation problem that results from regarding the f_i as a set of fixed effects dummy variables which are subject to proportionality constraints across the equations. That is the sort of model considered by Hauser and Goldberger (1971). Writing the model in structural form with one of the y's as a proxy for f lets us obtain the canonical correlation solution as an application of Hannan's (1967) extension of LIML. This rather surprising algebraic identity between the ML fixed effects and random effects estimators has been observed in the simpler factor model without the group structure. In that model Whittle (1953) found that his fixed effects estimator of the factor loadings agreed with the random effects ML algorithm devised by Lawley (1940) (also see the Uppsala Symposium, 1953). The estimation of $\underset{\sim}{B}$, however, differs in the two models. We show how the random effects procedure generalizes the more familiar single equation pooling of within and between group information. The random effects estimator is, in an appropriate metric, "between" the ML fixed effects estimator and the pooled OLS estimator.

Another case in which considerable analytic progress is possible has $\underset{\sim}{\Gamma} = \underset{\sim}{I}$ (no simultaneity), $\underset{\sim}{B} = -\underset{\sim}{\eta}\underset{\sim}{\lambda}'$, and $\underset{\sim}{V}$ unrestricted. The constraint on $\underset{\sim}{B}$ arises from postulating an unobservable h_{ij} which depends on observables $(\underset{\sim}{x}_{ij}'\underset{\sim}{\eta})$ and on an unobservable f_i that is constant across the group: $h_{ij} = \underset{\sim}{x}_{ij}'\underset{\sim}{\eta} + f_i$. This sort of model (without the group structure) is used by Griliches and Mason (1972) and in our own empirical work in Chapter 4. It is also similar to Jöreskog and Goldberger's (1973) MIMIC model. We show that conditional on one parameter (a generalized signal-noise ratio), the ML estimator in this model can be obtained analytically from an eigenvalue problem. So the algorithm reduces to a straightforward one dinensional numerical maximization problem. We have been more successful than Jöreskog and Goldberger because the replication allows us to leave the equation specific effects freely correlated and still have a restriction connecting the slopes with the residual covariance matrix. A more direct counterpart to their model would take $\underset{\sim}{V}$ diagonal in which case the analytic concentration of the likelihood would have to be conditional on $\underset{\sim}{V}$. With $\underset{\sim}{V}$ unrestricted and a fixed effects interpretation of f, we would be back in the Hauser and Goldberger case and the complete ML solution would fall out of a canonical correlation analysis.

Chapters 4 and 5 are empirical studies of the structural relationship between an individual's wages and his characteristics. The relationship is

$$(5) \qquad Y_{1i} = \underset{\sim}{x}'\underset{\sim}{\beta}_3 + \gamma_{23}S_i + \lambda_3 f_i + \nu_{3i}, \quad i = 1,\ldots q,$$

where Y_1 is the log of earnings in year 1, S is years of schooling, f is the systematic part of the residual, reflecting unobserved characteristics, and ν_3 is transitory income. For the moment we will ignore the other observed characteristics in $\underset{\sim}{x}$. To complete the stochastic specification we need to model the relationship between f and S. It is approximated by

$$(6) \qquad S_i = \underset{\sim}{x}_i'\underset{\sim}{\beta}_2 + \lambda_2 f_i + \nu_{2i} \quad .$$

We suspect that λ_2 is non-negligible because although f is random to the external observer, it is known to the individual and forms the initial conditions that he faces in deciding how much schooling is right for him. ν_2 contains other characteristics that are not relevant for wage determination.

The observable characteristics in $\underset{\sim}{x}$ might include family background measurements such as father's schooling or occupation. We will take $\underset{\sim}{x}$ to be independent of f by construction. This means reinterpreting f as the part of the unobserved characteristics that is not predictable from $\underset{\sim}{x}$. Of course this affects our interpretation of $\underset{\sim}{\beta}$. For example, if x is mother's schooling and f includes genetic ability, then the reinterpreted β reflects both the return to the mother's pre-school investment in the child and the spurious effect of mother's education as a proxy for the initial ability of the child. Or if x is father's income, then even if it has no direct effect on the son's earnings, our reinterpretated β will not be zero.

In order to separate the structural effects of x from the proxy effects, we would have to relate mother's and father's observed characteristics to their unobserved characteristics, f' and f''. Then we would allow f' and f'' to be correlated both with each other (assortative mating) and with f. But this more complicated model is irrelevant if all we want to estimate is the return to the son's schooling. For the γ's are not affected by the way in which we divide up the joint effect of $\underset{\sim}{x}$ and f.

Clearly the model in (5, 6) is not identified. A plausible source of additional information would be another measurement on earnings:

(7) $$Y_{2i} = \underset{\sim}{x}_i{}'\underset{\sim}{\beta}_4 + \gamma_{24}S_i + \gamma_{34}Y_{1i} + \lambda_4 f_i + \nu_{4i} \quad .$$

But in fact Theorem 2 in Chapter 2 shows that the model remains unidentified no matter how many measurements of this kind we have. And this is true even if γ_{34} equals zero; e.g. if there is enough time between the measurements so that they do not have a transitory piece in common.

More promising would be the availability of an early (pre-school) test score:

(8) $$T_i = \underset{\sim}{x}_i{}'\beta_1 + \lambda_1 f_i + \nu_{1i} \quad .$$

If T is excluded from all of the other equations then Theorem 4 of Chapter 2 applies. If the λ's are non-zero then one additional restriction is required for identification. In the absence of such a restriction (e.g. $\gamma_{34} \neq 0$), the ML estimate is a region generated by the following ML interval for the reliability of the test ($\rho = 1 - \sigma_1{}^2/\ \sigma_T{}^2$): $0 \leq \rho \leq R^2_{T.\,\underset{\sim}{x},\,S,\,Y_1,\,Y_2}$.

A hard question in this model is whether the combination of omitted characteristics that ties together the income and schooling residuals is the same combination that connects the schooling and test residuals. There is a straightforward answer under a narrow measurement error interpretation of f. Then ν_1 is interpreted as a test-retest error that could in principle be eliminated by replicating the test. So it is reasonable to assume that ν_1 is independent of everything else and there is clearly just one f, namely the systematic part of the test (the "true score") that is not captured by $\underset{\sim}{x}'\underset{\sim}{\beta}_1$.

There is, however, an alternative more general interpretation of f. It is that IQ tests are designed to predict academic performance and need not capture (or appropriately weight) the set of characteristics relevant for economic success. This suggests having two distinct but correlated unobservables, f_1 and f_2. f_1 reflects the weighting of the omitted characteristics relevant for predicting economic success and f_2 reflects the weighting appropriate for scholastic achievement. Then f_1 is excluded from the S equation, f_2 is excluded from the Y equations, and neigher is excluded from the T equation. Both of these interpretations of f are pursued in our empirical application in Chapter 4. It is based on the 1964 CPS-NORC veteran's data, which has previously been studied by Griliches and Mason (1972) and Duncan (1968), among others.

Some of the ML problems in Chapter 4 reduce to standard LIML calculations or to Hannan's (1967) extension of LIML. In other versions of the model the likelihood function is relatively intractable and we have followed Jöreskog and Goldberger (1973) in adapting a numberical minimization program by Jöreskog (1970, 1973) to our problems.

Our Chapter 5 application of the replication model in (4) uses Gorseline's (1932) data on brothers to control not only for between family parental background differences but also for individual within family differences which may be correlated with achieved schooling levels later on. The sort of model we use is

$$
\begin{aligned}
(9)\qquad S_{ij} &= \lambda_1 a_{ij} + \nu_{1ij} \\
Y_{1ij} &= \gamma_{12} S_{ij} + \lambda_2 a_{ij} + \nu_{2ij} \\
Y_{2ij} &= \gamma_{13} S_{ij} + \lambda_3 a_{ij} + \nu_{3ij} \\
a_{ij} &= f_i + g_{ij}, \quad i = 1,\dots,q; \quad j = 1,\dots,p \;,
\end{aligned}
$$

where the subscripts refer to the jth individual in the ith family. This grouping will buy us something if the systematic part of the residuals (a_{ij}) has a group structure while the equation specific effects do not.

This model is identified by the exclusion of Y_1 from the Y_2 equation provided $\lambda_1\lambda_2 \neq 0$ (Theorem 1, Chapter 3). In fact it is just identified and the ML estimates of the structural parameters can be obtained from a canonical correlation analysis of the reduced form.

In our actual application we did not have an additional observation on income but we did have a crude measure of the non-pecuniary income of the individual's occupation. In either case an assumption that merits a sensitivity analysis is the independence of ν_2 and ν_3. Allowing them to be correlated is equivalent to letting Y_2 depend on Y_1, i.e. not constraining $\gamma_{23} = 0$. So the sensitivity analysis is supplied by our bound that $\lambda = \sigma_f^2/(\sigma_f^2 + \sigma_g^2)$ is between zero and a generalized R^2 based on the largest canonical correlation between S, Y_1, Y_2 and a set of family indicator dummy variables. The resulting bound is in fact very tight.

Our empirical work in Chapter 5 also pursues a two factor extension of the model. Although the one factor structural model is just identified, the question of how many factors (with family components) is testable, and amounts to the increment in the generalized R^2 from adding another factor. We find some evidence for a second factor but none for a third. However, even a second factor makes the model highly unidentified. But a rather natural set of prior restrictions is to have one purely family factor, reflecting parental background characteristics and other experiences shared by the brothers, and a second factor with both a family and an individual component. These prior restrictions give a bound which is informative although not particularly sharp.

As in our Chapter 4 application, the hard question in this model is just how much structure to give the residuals. We do have some mild prior beliefs that some aggregation is possible, that a few appropriate indices will do an adequate job of summarizing the countless characteristics that could conceivably be measured. But in models of the size we have been working with, if "few" is more than "one" we quickly reach a point where the likelihood is quite diffuse. The nature of the problem is related to the way in which we have been using unobservables. A good analogy might be a model with a lagged dependent variable and serial correlation. Our primary interest has been in "cleaning up" the cross equation serial correlation so that the triangular structure will yield a truely recursive system. Although Chapter 3 developes proxies for the unobservable as an aid to the interpretation of our algorithms, the main focus of our empirical work is not in constructing indices of "ability". We just want to capture enough of the omitted characteristics to avoid serious bias in the coefficients of interest. As with serial correlation, we just want to clean it up as efficiently as possible, without focusing on the omitted variables that produce it. We would like to leave the form of the serial correlation as an empirical question, and the same is true of the number of factors. With richer data sets this would be possible, but the degrees of freedom is the number of factors and so we have had to impose quite a few prior restrictions. To some extent we are relying on a "half a loaf" justification, hoping that we can "sweep out" the major connections in the residuals.

There is another role for unobservables in econometric models that is not touched upon in my applied work. It would make the measurement of the unobservable the primary objective. This is closer to the spirit of Griliches' (1973) observation that "Substantive unobservables...are variables about which we are willing to make many more a priori assumptions. They are the carriers of some of the content of our theories and we are willing to specify which other variables affect them and are affected by them in turn".

An example would be an attempt to construct a "pure" price index, purged of quality change. In his refinement of Cagan's (1965) use of secondhand prices to measure quality differences, Hall (1969, 1971) specified the following relationship:

(10) $$\log P_{it\tau} = \log \bar{P}_{it} + \log b_{i,\ t-\tau} + \log D_{i\tau} + \nu_{it}$$

where i indexes models, t is calendar time, τ is age, $h = t - \tau$ is vintage, $\bar{P}_t$ is a price index for new capital goods corrected for quality change, D_τ is a depreciation index, and $\nu_{t\tau}$ is a random disturbance. Hall shows that the vintage effects can only be estimated up to an additive constant, and so only departures from an unidentified quality trend are estimable. Hall remedies this by combining the secondhand prices with the hedonic hypothesis, relating the embodied technical change to changes in the observed characteristics of capital goods:

(11) $$\log b_h = \eta_1 \log x_{h1} + \ldots + \eta_m \log x_{hm} .$$

We would want to consider including unmeasured characteristics f_h together with an appropriate grouping device.

A possibility, investigated by Ohta and Griliches (1973), is to group observations by makes or brands, allowing us to pick up changes in omitted characteristics that are common to all models of a given make. An appropriate prior for the f_h might be exchangeable across makes with distributed lag type smoothness restrictions across vintages (e.g. Leamer, 1972 or Schiller, 1973). But the main point that I want to emphasize is that the primary focus would be on measuring the unobservable.

Our concluding Chapter 6 briefly examines some extensions and suggestions for further research. The common focus of our examples and applications on one empirical problem has the advantage of providing these essays with some additional unity. But it has the disadvantage of suggesting, I believe incorrectly, that our approach is limited to the stochastic specification of human capital models. So we will sketch an application to a combined time-series cross-section analysis of individual firm production and factor demand relations. Thus the conclusion will link back to the major precedent for our approach.

Chapter 2

The Identification of Triangular Systems

1. Introduction

The model we consider is

$$\underset{\sim}{y}_i{}'\underset{\sim}{\Gamma} + \underset{\sim}{x}_i{}'\underset{\sim}{B} = \underset{\sim}{f}_i{}'\underset{\sim}{\Lambda} + \underset{\sim}{\nu}_i{}' \quad , \qquad i=1,\dots,q \tag{I.1}$$

where $\underset{\sim}{y}_i$ is an m x 1 vector of endogenous variables, $\underset{\sim}{x}_i$ is an n x 1 vector of exogenous variables, $\underset{\sim}{\Gamma}$ is an upper triangular matrix of parameters with ones on the diagonal, $\underset{\sim}{B}$ is an n x m parameter matrix, and there are q observations. The residuals are assumed to be independent across observations. If they were also independent across equations then the model would be recursive and readily identifiable. Conversely, if the residuals were freely correlated across equations then the standard Cowles Commission results would apply. Our interest is in the intermediate cases where some but not all of the identification comes from covariance restrictions on the residuals. They are assumed to have a factor analytic structure where $\underset{\sim}{f}_i$ is an N x 1 vector of latent variables and $\underset{\sim}{\Lambda}$ is an N x m matrix of coefficients (factor loadings). The unobservable $\underset{\sim}{f}_i$ are distributed as a multivariate random sample with covariance matrix $\underset{\sim}{\Phi}$. $\underset{\sim}{\nu}_i$ is an m x 1 vector of equation specific effects which are distributed independently of $\underset{\sim}{f}_i$ as a random sample with covariance matrix $\underset{\sim}{U} = \text{diag}\,\{\sigma_1{}^2,\dots,\ \sigma_m{}^2\}$.

This model is useful in a wide variety of micro-econometric applications. Examples include studies of social mobility and the determinants of socio-economic achievement. The triangular structure arises from making measurements on an individual's characteristics at a particular time.

Then the variable becomes a characteristic which determines subsequent measurements. $\underset{\sim}{x}$ and $\underset{\sim}{f}$ are a set of characteristics which potentially affect all subsequent observations. The distinction between them is that $\underset{\sim}{f}$ is unobservable. The assumed independence of $\underset{\sim}{x}$ and $\underset{\sim}{f}$ simply means that we interpret $\underset{\sim}{f}$ to be the part of the unobservable characteristics that is not predictable from $\underset{\sim}{x}$. This of course affects our interpretation of $\underset{\sim}{B}$ and limits the restrictions we can impose on $\underset{\sim}{B}$. For example x_1 may have no effect on y_k if all other relevant characteristics are included. But if the partial correlation is non-zero (partialling on the other included x's), then with our interpretation of $\underset{\sim}{f}$ we cannot exclude x_1 from that equation. $\underset{\sim}{\Gamma}$, however, is unaffected by the way in which we divide up the joint effect of $\underset{\sim}{x}$ and $\underset{\sim}{f}$.

Under normality assumptions (or limiting ourselves to second order moments), the distribution of $\underset{\sim}{y}$ conditional on $\underset{\sim}{x}$ is completely characterized by the following reduced form parameters:

$$\text{(I.2)} \qquad \begin{aligned} \underset{\sim}{\Pi} &= -\underset{\sim}{B}\underset{\sim}{\Gamma}^{-1} \\ \underset{\sim}{\Sigma} &= \underset{\sim}{\Gamma}^{-1'}(\underset{\sim}{\Lambda}'\underset{\sim}{\Phi}\underset{\sim}{\Lambda} + \underset{\sim}{U})\underset{\sim}{\Gamma}^{-1} \; . \end{aligned}$$

The identification problem is to recover $\underset{\sim}{\Gamma}$, $\underset{\sim}{B}$, $\underset{\sim}{\Lambda}$, $\underset{\sim}{\Phi}$, and $\underset{\sim}{U}$ from the reduced form.

This problem can be approached from at least two points of view. The traditional one is to ask "What are the limits of observational information?" If the reduced from parameters are known with certainty, what aspects of the structure can we uncover? An alternative approach, which I prefer, is to treat the identification problem as one aspect of investigating a likelihood function. We typically start by investigating the mode and then proceed to examine measures of dispersion. But a logically prior question is whether the maximum of the likelihood corresponds to a unique vector of structural parameters. If not, then we have multiple peaks, a ridge or a plateau, and the problem is to describe ML regions for the structural parameters.

The general treatment of this model remains an elusive goal. I will examine the case in which replication is available in Chapter 3. Then it is possible to obtain identification conditions which are both necessary and sufficient. This paper is confined to the one factor model. Even then a complete solution is not available except for special cases. We do, however, have some useful necessary conditions, and in addition a set of sufficient conditions which provide a constructive method for obtaining the structural parameters from the reduced form.

II. Identification

We will work with the one factor version of (I.1). So N = 1 and $\underset{\sim}{\Lambda} = \underset{\sim}{\lambda}'$ where $\underset{\sim}{\lambda}$ is m x 1. If there are no restrictions on $\underset{\sim}{B}$, then the problem is to uncover $\underset{\sim}{\Gamma}$, $\underset{\sim}{\lambda}$, and $\underset{\sim}{U}$ from

(II.1) $\underset{\sim}{\Sigma} = \underset{\sim}{d}\underset{\sim}{d}' + \underset{\sim}{V}$

where $\underset{\sim}{V} = \underset{\sim}{\Gamma}^{-1'}\underset{\sim}{U}\underset{\sim}{\Gamma}^{-1}$, $\underset{\sim}{d} = \underset{\sim}{\Gamma}^{-1'}\underset{\sim}{\lambda}$, and we have scaled f so that $\phi = 1$. Now if we knew $\underset{\sim}{d}$ then we could use Gaussian elimination on $\underset{\sim}{\Sigma} - \underset{\sim}{d}\underset{\sim}{d}'$ to uniquely obtain $\underset{\sim}{\Gamma}^{-1}$ and $\underset{\sim}{U}$. Since $\underset{\sim}{d}$ is m x 1 this correctly suggests that we need m restrictions on $\underset{\sim}{\Gamma}$. The first two theorems give necessary conditions on the placement of these restrictions.

II.a Necessary Conditions

Theorem 1: If $\underset{\sim}{B}$ is unrestricted and $\underset{\sim}{\Gamma}$ is only subject to zero restrictions, then identification of the model requires (at least) one exclusion in the m^{th} equation, two exclusions in the last two equations, and in general k exclusions in the last k equations for k = 1,..., m.

So we are insisting that the restrictions be spread out or at least not clustered on the earlier equations.

Proof: Let $\underset{\sim}{C} = \underset{\sim}{\Gamma}^{-1}$. $\underset{\sim}{C}$ is upper triangular with ones on the diagonal and γ_{hk} is a function of C_{ij} where $h \leq i < k$ and $h < j \leq k$. Let $\underset{\sim}{A} = \underset{\sim}{\Sigma} - \underset{\sim}{d}\underset{\sim}{d}' = \underset{\sim}{C}'\underset{\sim}{U}\underset{\sim}{C}$. Then by Gaussian elimination

(II.2) $$A\begin{pmatrix}1 & 2 & \dots & i-1 & j\\ 1 & 2 & \dots & i-1 & i\end{pmatrix} / A\begin{pmatrix}1 & 2 & \dots & i-1\\ 1 & 2 & \dots & i-1\end{pmatrix} = C_{ij}$$

where $A\begin{pmatrix}h_1 & \dots & h_p\\ k_1 & \dots & k_p\end{pmatrix}$ is the minor formed from rows $h_1,\dots,h_p$ and columns $k_1,\dots,k_p$ of $\underset{\sim}{A}$ (Gantmacher [1959] chapter II). Equation (II.2) only depends on $d_i,\dots,d_j$. But $\gamma_{hk} = 0$ can be written in terms of C_{ij} with $j \leq k$. So a restriction on the k^{th} equation gives a constraint that only involves $d_1, \ldots, d_k$. The m restrictions on $\boldsymbol{\Gamma}$ give m such equations which must be solved for $d_1, \ldots, d_m$. So there has to be at least one restriction on the m^{th} equation. And there must be a restriction on the m^{th} or $(m-1)^{th}$ equations in order to catch d_{m-1}. So there must be at least two restrictions on the last two equations. Continuing this argument completes the proof.

Corollary: A necessary condition for identification from zero restrictions on $\mathbf{B}$ and $\boldsymbol{\Gamma}$ is that k of the restrictions must fall on the last k equations.

Proof: We regard $\mathbf{x}$ as having a multivariate distribution with covariance matrix $\boldsymbol{\Psi}$. There is an upper triangular matrix L with ones on the diagonal which will diagonalize $\boldsymbol{\Psi}$: $\mathbf{L}'\boldsymbol{\Psi}\mathbf{L} = \text{diag}\ \{\psi_1, \ldots, \psi_n\}$. So we can rewrite our model as

$$(\text{II.3}) \qquad (\mathbf{x}' \ \mathbf{y}') \begin{bmatrix} \mathbf{L} & \mathbf{B} \\ \mathbf{0} & \boldsymbol{\Gamma} \end{bmatrix} = (\mathbf{0} \ \boldsymbol{\lambda}') + (\mathbf{u}' \ \boldsymbol{\nu}')$$

where $\mathbf{u}$ has a diagonal covariance matrix and is independent of $\boldsymbol{\nu}$. Now apply the Theorem. Instead of $\mathbf{d}$ we have $\begin{pmatrix} \mathbf{0} \\ \mathbf{d} \end{pmatrix}$ but it is still true that solving for d_m requires a restriction on the last equation, solving for d_{m-1} requires a restriction on one of the last two equations, etc.

Theorem 1 gives placement conditions on the way the restrictions are allocated across the equations. Our second result will constrain the placement of restrictions relative to the variables.

Theorem 2: If $\underset{\sim}{B}$ is unrestricted then a necessary condition for identification from zero restrictions on $\underset{\sim}{\Gamma}$ is that for each $k \leq m$ there must be k restrictions, each of which excludes one of the following variables from an equation: $f, y_1, \ldots, y_k$.

Proof: Let $\underset{\sim}{G} = (\underset{\sim}{g}_1 \ldots \underset{\sim}{g}_{m+1})$ where $\underset{\sim}{g}_1 = \underset{\sim}{d}$ and $\underset{\sim}{g}_{i+1}/\sigma_i$ is the i^{th} row of $\underset{\sim}{C} = \underset{\sim}{\Gamma}^{-1}$. Then $\underset{\sim}{\Sigma} = \underset{\sim}{G}\underset{\sim}{G}'$ and we have to recover $\underset{\sim}{G}$ from $\underset{\sim}{\Sigma}$. We have already seen that at least m restrictions are needed to identify the model and so the Theorem is true for $k = m$. If for $k < m$ we were given the coefficients of $y_{k+1}, \ldots, y_{m-1}$ in equation $k + 2$ through m, and if $\sigma^2_{k+1}, \ldots, \sigma^2_m$ were known, then we would know the last m-k columns of $\underset{\sim}{G}$; i.e. with $\underset{\sim}{G} = (\underset{\sim}{G}_1 \ \underset{\sim}{G}_2)$, we would know $\underset{\sim}{G}_2$. Then the problem is to obtain $\underset{\sim}{G}_1$ from $\underset{\sim}{\Sigma} - \underset{\sim}{G}_2\underset{\sim}{G}_2'$. Note that $\underset{\sim}{G}_1$ is unrestricted except for the restrictions implied by the triangularity of $\underset{\sim}{\Gamma}$. For any $\bar{\underset{\sim}{G}}_1$ such that $\bar{\underset{\sim}{G}}_1\bar{\underset{\sim}{G}}_1' = \underset{\sim}{G}_1\underset{\sim}{G}_1'$, there is a $(k+1)x(k+1)$ rotation $\underset{\sim}{R}$ such that $\underset{\sim}{G}_1 = \bar{\underset{\sim}{G}}_1\underset{\sim}{R}$, $\underset{\sim}{R}'\underset{\sim}{R} = I$. So there must be $(k+1)^2-(k+1)(k+2)/2 = k(k+1)/2$ restrictions on $\underset{\sim}{G}_1$ in order to pin down the rotation. The triangular structure imposes $k(k-1)/2$ restrictions and so we need an additional k restrictions on the coefficients of $f, y_1, \ldots, y_k$.

Corollary: A necessary condition for identification from zero restrictions on $\underset{\sim}{B}$ and $\underset{\sim}{\Gamma}$ is that for each $k \leq m$ there must be k restrictions, each of which excludes one of the following variables from an equation:

$f, x_1, \ldots, x_n, y_1, \ldots, y_k$.

Proof: Rewrite the model as in (II.3) and apply the Theorem.

II.b Sufficient Conditions

The basic idea is to use a proxy for the unobservable f and then solve the resulting errors in variables problem by finding a suitable instrument. Say we have

$$
\text{(II.4)}\qquad
\begin{aligned}
y_1 &= \lambda_1 f + \nu_1\\
y_2 &= \lambda_2 f + \nu_2\\
y_3 &= \gamma_{23} y_2 + \lambda_3 f + \nu_3\\
y_4 &= \gamma_{24} y_2 + \lambda_4 f + \nu_4 \quad .
\end{aligned}
$$

We can use y_1 as a proxy for f in the y_3 equation:

$$
\text{(II.5)}\qquad y_3 = \frac{\lambda_3}{\lambda_1} y_1 + \gamma_{23} y_2 + u_3 - \frac{\lambda_3}{\lambda_1} u_1 \quad .
$$

This results in a standard errors in variables problem due to the "measurement error" in y_1. It can be cured by using y_4 as an instrument for y_1. Similarly y_3 can be used as an instrument for y_1 in the y_4 equation.

Complications arise when more than one variable needs an external instrument. Then the instrumental variable (IV) normal equations need not have full rank. For consider the following model:

$$
\text{(II.6)}\qquad
\begin{aligned}
y_1 &= \lambda_1 f + \nu_1\\
y_2 &= \lambda_2 f + \nu_2\\
y_3 &= \gamma_{13} y_1 + \lambda_3 f + \nu_3\\
y_4 &= \gamma_{14} y_1 + \gamma_{24} y_2 + \gamma_{34} y_3 + \lambda_4 f + \nu_4\\
y_5 &= \gamma_{25} y_2 + \lambda_5 f + \nu_5\\
y_6 &= \gamma_{26} y_2 + \lambda_6 f + \nu_6
\end{aligned}
$$

where we are trying to identify γ_{24}. We can use y_1 as a proxy for f in the y_4 equation:

$$
\text{(II.7)}\qquad y_4 = (\gamma_{14} + \frac{\lambda_4}{\lambda_1}) y_1 + \gamma_{24} y_2 + \gamma_{34} y_3 + \nu_4 - \frac{\lambda_4}{\lambda_1} \nu_1 \quad .
$$

Then y_1 and y_3 are correlated with ν_1 and so we use y_5 and y_6 as instruments for them. The IV normal equations are

$$\begin{bmatrix} \sigma_{51} & \sigma_{53} & \sigma_{52} \\ \sigma_{61} & \sigma_{63} & \sigma_{62} \\ \sigma_{21} & \sigma_{23} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \gamma_{14} + \frac{\lambda_4}{\lambda_1} \\ \gamma_{34} \\ \gamma_{24} \end{bmatrix} = \begin{bmatrix} \sigma_{54} \\ \sigma_{64} \\ \sigma_{24} \end{bmatrix}$$

or

(II.8) $\quad \underset{\sim}{P}\,\underset{\sim}{\eta} = \underset{\sim}{p}$.

The first two columns of $\underset{\sim}{P}$ are proportional to $\begin{bmatrix} d_5 \\ d_6 \\ d_2 \end{bmatrix}$ and so

$\underset{\sim}{P}$ is singular. But the third column is

$$d_2 \begin{bmatrix} d_5 \\ d_6 \\ d_2 \end{bmatrix} + \sigma_2^{\,2} \begin{bmatrix} \gamma_{25} \\ \gamma_{26} \\ 1 \end{bmatrix}$$

and this is not in general proportional to the first two. So γ_{24} is an estimable function and hence is identifiable.

A valid criticism of this example is that we could have used y_3 as a proxy for f in the y_4 equation. Then only y_3 would have needed an external instrument, $\underset{\sim}{P}$ would have been non-singular, and the problem of determining the estimable functions would not have arisen. But it is not always possible to find a proxy that will avoid the problem. For consider the following example:

$$\text{(II.9)}\qquad \begin{array}{lllllll}
y_1 = & & & & & & \lambda_1 f + \nu_1 \\
y_2 = & & & & & & \lambda_2 f + \nu_2 \\
y_3 = & \gamma_{13} y_1 & + \gamma_{23} y_2 & & & + & \lambda_3 f + \nu_3 \\
y_4 = & \gamma_{14} y_1 & + \gamma_{24} y_2 & & & + & \lambda_4 f + \nu_4 \\
y_5 = & \gamma_{15} y_1 & + \gamma_{25} y_2 & + \gamma_{35} y_3 & + \gamma_{45} y_4 & + & \lambda_5 f + \nu_5 \\
y_6 = & & \gamma_{26} y_2 & & & + & \lambda_6 f + \nu_6 \\
y_7 = & & \gamma_{27} y_2 & & & + & \lambda_7 f + \nu_7 \\
y_8 = & & \gamma_{28} y_2 & & & + & \lambda_8 f + \nu_8 \quad .
\end{array}$$

We want to identify γ_{25}. The only feasible proxy for f is y_1; any other choice would contaminate the y_2 coefficient. For example using y_3 gives

$$y_5 = (\gamma_{15} - \frac{\lambda_5}{\lambda_3}\gamma_{13})y_1 + (\gamma_{25} - \frac{\lambda_5}{\lambda_3}\gamma_{23})\, y_2$$
$$+ (\gamma_{35} + \frac{\lambda_5}{\lambda_3})\, y_3 + \gamma_{45} y_4 + \nu_5 - \frac{\lambda_5}{\lambda_3}\nu_3 \quad ,$$

and the IV equations can at most identify $\gamma_{25} - \frac{\lambda_5}{\lambda_3}\gamma_{23}$. So with y_1 as the proxy we have

$$\text{(II.10)}\qquad y_5 = (\gamma_{15} + \frac{\lambda_5}{\lambda_1})\, y_1 + \gamma_{25} y_2 + \gamma_{35} y_3 + \gamma_{45} y_4 + \nu_5 - \frac{\lambda_5}{\lambda_4}\nu_1 \quad .$$

External instruments are needed for y_1, y_3, and y_4. The only candidates are y_6, y_7, y_8 . Again we form the IV equations $\underset{\sim\sim}{P\eta} = \underset{\sim}{p}$ with the i,j element of $\underset{\sim}{P}$ equal to σ_{ij} for i=6, 7, 8, 2 and j=1, 3, 4, 2. As before the variables which do not require external instruments are put last.

Now the first three columns of $\underset{\sim}{P}$ are

$$d_1 \begin{bmatrix} d_6 \\ d_7 \\ d_8 \\ d_2 \end{bmatrix} , \quad d_3 \begin{bmatrix} d_6 \\ d_7 \\ d_8 \\ d_2 \end{bmatrix} + \gamma_{23}\, \sigma_2{}^2 \begin{bmatrix} \gamma_{26} \\ \gamma_{27} \\ \gamma_{28} \\ 1 \end{bmatrix} , \quad d_4 \begin{bmatrix} d_6 \\ d_7 \\ d_8 \\ d_2 \end{bmatrix} + \gamma_{24}\, \sigma_2{}^2 \begin{bmatrix} \gamma_{26} \\ \gamma_{27} \\ \gamma_{28} \\ 1 \end{bmatrix} .$$

Let $r = -\gamma_{23}/\gamma_{24}$. Then $\underset{\sim}{P}_2 + r\, \underset{\sim}{P}_3$ is either $\underset{\sim}{0}$ or proportional to the first column of $\underset{\sim}{P}$. So $\underset{\sim}{P}$ is singular but again γ_{25} is in general estimable.

These ideas are systematically developed in Theorem 3, but first we need some definitions.

<u>Definition 1:</u> y_h can be used as a <u>proxy</u> for f in the k^{th} equation provided $\lambda_h \neq 0$ and $\gamma_{kh} = 0$.

<u>Definition 2:</u> Let c_{ij} be the i, j element of $\underset{\sim}{\Gamma}^{-1}$. Then y_j does not <u>depend</u> (either directly or indirectly) on y_i of $c_{ij} = 0$.

<u>Definition 3:</u> With y_h as the proxy for f, we rewrite the k^{th} equation in its <u>proxy form</u>:

(II.11)
$$y_k = \underset{\sim}{x}' \left(\underset{\sim}{\beta}_k - \frac{\lambda_k}{\lambda_h} \underset{\sim}{\beta}_h\right) + \sum_{j=1}^{h-1} \left(\gamma_{jk} - \frac{\lambda_k}{\lambda_h} \gamma_{jh}\right) y_j$$
$$+ \sum_{j=h}^{k+1} \gamma_{jk}\, y_j + \frac{\lambda_k}{\lambda_h}\, y_h + \nu_k - \frac{\lambda_k}{\lambda_h}\, \nu_h$$

if k > h, and

$$y_k = \underset{\sim}{x}'\left(\underset{\sim}{\beta}_k - \frac{\lambda_k}{\lambda_h} \underset{\sim}{\beta}_h\right) + \sum_{j=1}^{k-1} \left(\gamma_{jk} - \frac{\lambda_k}{\lambda_h} \gamma_{jh}\right) y_j$$
$$- \frac{\lambda_k}{\lambda_h} \sum_{j=k}^{h-1} \gamma_{jh} y_j + \frac{\lambda_k}{\lambda_h} y_h + \nu_k - \frac{\lambda_k}{\lambda_h} \nu_h$$

if k < h.

Definition 4: y_j can be used as an *instrument* for the k^{th} equation with y_h as the proxy provided y_j does not depend on y_h or y_k. Any x can be used as an instrument. y_j is an *external instrument* if it does not appear in the proxy form of the k^{th} equation.

Definition 5: γ_{jk} is *contaminated* by the use of y_h as a proxy in the k^{th} equation if $\gamma_{jh} \neq 0$. β_{jk} is contaminated if $\beta_{jh} \neq 0$.

Definition 6: Let J_2 index the set of variables in the proxy form which can be used as instruments. The remaining variables in the proxy form are indexed by J_1. Let I_1 index the external instruments and set $I_2 = J_2$. Then the *instrumental variable (IV) equations* are $\underset{\sim}{P}\underset{\sim}{\eta} = \underset{\sim}{p}$ where

$$\underset{\sim}{P} = (\underset{\sim}{P}_1 \ \underset{\sim}{P}_2) = \begin{bmatrix} \underset{\sim}{P}_{11} & \underset{\sim}{P}_{12} \\ \underset{\sim}{P}_{21} & \underset{\sim}{P}_{22} \end{bmatrix}$$

and $\underset{\sim}{P}_{11} = (\sigma_{ij})$ with i in I_1, j in J_1; $\underset{\sim}{P}_{12} = (\sigma_{ij})$ with i in I_1, j in J_2, etc. $\underset{\sim}{\eta}$ contains the parameters (suitably ordered) in the proxy form and the typical element of $\underset{\sim}{p}$ is σ_{ik} with i in I_1 or I_2.

Theorem 3: Given a proxy y_h and a set of instruments for the k^{th} equation, then γ_{jk} is identified from the IV equations $\underset{\sim}{P}\underset{\sim}{\eta} = \underset{\sim}{p}$ if it is not contaminated and if either a) $\underset{\sim}{P}$ is non-singular or b) y_j is used as an instrument (i.e. $j \in J_2$) and rank $\underset{\sim}{P}_1$ + rank $\underset{\sim}{P}_2$ = rank $\underset{\sim}{P}$.

Proof: We will ignore the x's. It is straightforward to modify the proof as in the Corollaries to Theorems 1 and 2. First it is necessary to check that the IV equations are in fact satisfied by $\underset{\sim}{\Sigma} = \underset{\sim}{\Gamma}^{-1'}(\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U})\underset{\sim}{\Gamma}^{-1}$. We will collect terms in $\sigma_f^2, \sigma_1^2, \ldots, \sigma_m^2$ and examine the IV equations term by term. Collecting terms in σ_f^2 for the i^{th} IV equation gives

$$d_i\lambda_h(d_k - \sum_{j=1}^{k-1} \gamma_{jk}d_j) = d_i\lambda_k(d_h - \sum_{j=1}^{h-1} \gamma_{jh}d_j) \quad .$$

Since $\underset{\sim}{\lambda} = \underset{\sim}{\Gamma}'\underset{\sim}{d}$ we have $\lambda_k = d_k - \sum_{j=1}^{k-1} \gamma_{jk}d_j$ and so the condition reduces to $d_i\lambda_h\lambda_k = d_i\lambda_h\lambda_k$.

Collecting terms in σ_t^2 gives

$$c_{ti}\lambda_h(c_{tk} - \sum_{j=1}^{k-1} \gamma_{jk}c_{tj}) = c_{ti}\lambda_k(c_{th} - \sum_{j=1}^{h-1} \gamma_{jh}c_{tj}).$$

$c_{tk} - \sum_{j=1}^{k-1} \gamma_{jk}c_{tj}$ is the inner product of the t^{th} row of $\underset{\sim}{C} = \underset{\sim}{\Gamma}^{-1}$ and the k^{th} column of $\underset{\sim}{\Gamma}$. This is δ_t^k (= one if t=k and zero otherwise). Thus the condition reduces to $c_{ti}\lambda_h\delta_t^k = c_{ti}\lambda_k\delta_t^h$. This is satisfied if $c_{ki}=c_{hi}=0$ so that the instrument y_i does not depend on y_k or y_h.

Thus the IV equations are valid relationships connecting the structural and reduced form parameters. If the IV equations have full rank then clearly $\underset{\sim}{\eta}$ is identified. If, however, $\underset{\sim}{P}$ is singular, then the key to finding the estimable functions is the non-singularity of $\underset{\sim}{P}_{22}$. For $\underset{\sim}{P}_{22}$ is the variance-covariance matrix of the y_j's with $j \varepsilon J_2$. Our rank condition states that $\underset{\sim}{P}\underset{\sim}{\ell} = \underset{\sim}{0}$ implies $\underset{\sim}{P}_1\underset{\sim}{\ell}_1 = \underset{\sim}{P}_2\underset{\sim}{\ell}_2 = \underset{\sim}{0}$ since the intersection of the column spaces of $\underset{\sim}{P}_1$ and $\underset{\sim}{P}_2$ only contains $\underset{\sim}{0}$. But $\underset{\sim}{P}_2$ has full column rank and so $\underset{\sim}{\ell}_2 = \underset{\sim}{0}$ and $\underset{\sim}{\eta}_2$ is uniquely determined. This completes our proof.

The rank condition will clearly fail if there are fewer instruments than variables in the proxy form of the equation. It will also fail if one of the instruments is an exogenous variable which is uncorrelated with any of the variables appearing in the proxy form. But it is <u>not</u> true that an external instrument must be correlated with at least one of the variables that requires an external instrument. For example, suppose

$$\text{(II.12)} \quad \begin{aligned} y_1 &= && \lambda_1 f + \nu_1 \\ y_2 &= \beta_{12}x && + \lambda_2 f + \nu_2 \\ y_3 &= \quad \gamma_{13}y_1 + \gamma_{23}y_2 && + \lambda_3 f + \nu_3 \end{aligned}$$

and we use y_1 as the proxy to identify γ_{23}. Then y_2 can instrument itself but we have to use x to instrument y_1. This may seem to be a problem since x and y_1 are uncorrelated. But

$$\underset{\sim}{P} = \begin{bmatrix} 0 & \beta_{12}\,\sigma_x^{\,2} \\ \lambda_1\lambda_2 & \beta_{12}^{\;2}\,\sigma_x^{\,2} + \lambda_2^{\,2} + \sigma_2^{\,2} \end{bmatrix}$$

is clearly non-singular as long as y_2 is correlated with x.

Corollary: If γ_{jk} is identified then we can rewrite the kth equation as

$$\text{(II.13)} \quad \tilde{y}_k = (y_k - \gamma_{jk}y_j) = \underset{\sim}{x}'\underset{\sim}{\beta}_k + \sum_{i=j}^{k-1} \gamma_{ik}y_i + \lambda_k f + \nu_k$$

and apply the Theorem to identify the remaining parameters. A similar result holds if β_{jk} is identified.

Proof: It is only necessary to check that the IV equations are valid. But they are the same sort of IV equations that were checked in the Theorem.

For an example of the Corollary, let

$$\text{(II.14)} \quad \begin{aligned} y_1 &= && \lambda_1 f + \nu_1 \\ y_2 &= && \lambda_2 f + \nu_2 \\ y_3 &= \quad \gamma_{23}y_2 && + \lambda_3 f + \nu_3 \\ y_4 &= \quad \gamma_{34}y_3 && + \lambda_4 f + \nu_4 \\ y_5 &= \gamma_{15}y_1 + \gamma_{25}y_2 + \gamma_{35}y_3 && + \lambda_5 f + \nu_5 \end{aligned}$$

and try to identify γ_{15}. Then the proxy must be y_2 or y_3 with y_4 as the external instrument. But y_4 depends on y_2 and y_3 and cannot serve as an instrument. So first we identify γ_{25} and γ_{35} by letting y_1 be the proxy with y_4 as an external instrument. Then

(II.15) $$\tilde{y}_5 = y_5 - \gamma_{25}y_2 - \gamma_{35}y_3 = \gamma_{15}y_1 + \lambda_5 f + \nu_5 \ .$$

Now let y_4 be the proxy:

(II.16) $$\tilde{y}_5 = \gamma_{15}y_1 - \frac{\lambda_5}{\lambda_4}\gamma_{34}y_3 + \frac{\lambda_5}{\lambda_4} y_4 + \nu_5 - \frac{\lambda_5}{\lambda_4}\nu_4 .$$

Then only y_4 needs an external instrument and we can use y_2. Thus γ_{15} is in general identifiable.

II.c A General Treatment of Some Special Cases

Our first special case has $\gamma_{1k} = 0$ for $k > 1$. So y_1 is excluded from all of the other equations.

Theorem 4: In (I.1) with $\gamma_{1k} = 0$ for $k > 1$, a sufficient condition for identification is that a single $\gamma_{st} = 0$, $s > 1$, provided the following rank condition holds: $\sigma_2^{\ 2} > 0, \ldots, \sigma_m^{\ 2} > 0$ and

(II.17) $$\lambda_t \sum_{j=s}^{t-1} \gamma_{sj}\lambda_j/\sigma_j^{\ 2} \neq 0 .$$

The condition is also necessary if we confine ourselves to zero restrictions on $\underset{\sim}{\Gamma}$.

Proof: Let $\underset{\sim}{C} = \underset{\sim}{\Gamma}^{-1}$. Then

$$\underset{\sim}{\Gamma} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Gamma}_2 \end{bmatrix}, \quad \underset{\sim}{C} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Gamma}_2^{-1} \end{bmatrix} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{C}_2 \end{bmatrix} ,$$

and we can partition $\underset{\sim}{\Sigma}$ and $\underset{\sim}{d}$ into

$$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_{11} & \underset{\sim}{\sigma}_{12} \\ \underset{\sim}{\sigma}_{21} & \underset{\sim}{\Sigma}_2 \end{bmatrix} , \quad \underset{\sim}{d} = \begin{bmatrix} d_1 \\ \underset{\sim}{d}_2 \end{bmatrix}$$

with $\mathbf{\Sigma}_2 = \mathbf{d}_2\mathbf{d}_2' + \mathbf{C}_2'\mathbf{V}_2\mathbf{C}_2$ and $\mathbf{V}_2 = \text{diag}\ \{\sigma_2^2,\dots,\sigma_m^2\}$. $\mathbf{d}_2 = \sqrt{\tau}\ \boldsymbol{\sigma}_{21}$ where $\tau = 1/d_1^2$, and so

(II.18) $\mathbf{\Sigma}_2 = \tau\boldsymbol{\sigma}_{21}\boldsymbol{\sigma}_{21}' + \mathbf{C}_2'\mathbf{V}_2\mathbf{C}_2$.

We have to recover τ, $\mathbf{\Gamma}_2$, and $\mathbf{V}_2$ from $\mathbf{\Sigma}_2$ and $\boldsymbol{\sigma}_{21}$ where $\mathbf{\Gamma}_2$ is upper triangular with ones on the diagonal. This identification problem is identical to the one discussed in Chapter 3, Theorem 1. We will simply sketch the proof used there:

$$(\mathbf{\Sigma}_2 - \tau\boldsymbol{\sigma}_{21}\boldsymbol{\sigma}_{21}')^{-1} = \mathbf{\Gamma}_2\mathbf{V}_2\mathbf{\Gamma}_2'$$

and so given τ we can uniquely solve for $\mathbf{\Gamma}_2$ and $\mathbf{V}_2$ by Gaussian elimination if $\mathbf{V}_2$ is positive definite. The notation is simplified if we reverse the order of the equations so that $\mathbf{\Gamma}$ is lower triangular. Then the zero element is γ_{hk} where $h = m - s + 1$, $k = m - t + 1$. Solving for γ_{hk} and setting the result to zero gives an equation for τ which can be simplified to the following linear equation:

(II.19) $\sigma_2^{h,k} - \mathbf{s}_h'\mathbf{S}^{-1}\mathbf{s}_k + \kappa Q = 0$

where $\sigma_2^{h,k}$ is the h, k element of $\mathbf{\Sigma}_2^{-1}$, $\mathbf{S}$ is the k-1 by k-1 principal submatrix of $\mathbf{\Sigma}_2^{-1}$, $\mathbf{s}_i' = (\sigma_2^{i1} \dots \sigma_2^{i,k-1})$, i=h, k, $\tau = \kappa/(1 + \kappa\boldsymbol{\sigma}_{21}'\mathbf{\Sigma}_2^{-1}\boldsymbol{\sigma}_{21})$,

$$Q = (\bar{\mathbf{c}}'\mathbf{S}^{-1}\bar{\mathbf{c}})(\sigma_2^{hk} - \mathbf{s}_h'\mathbf{S}^{-1}\mathbf{s}_k) + c_h c_k + c_h(\bar{\mathbf{c}}'\mathbf{S}^{-1}\mathbf{s}_k) + c_k(\bar{\mathbf{c}}'\mathbf{S}^{-1}\mathbf{s}_h)$$
$$- (\bar{\mathbf{c}}'\mathbf{S}^{-1}\mathbf{s}_h)(\bar{\mathbf{c}}'\mathbf{S}^{-1}\mathbf{s}_k),$$

with $\mathbf{c} = \mathbf{\Sigma}_2^{-1}\ \boldsymbol{\sigma}_{21}$ and $\bar{\mathbf{c}}$ contains the first k-1 elements of $\mathbf{c}$. The rank condition in the Theorem ensures that $Q \neq 0$. If it is satisfied, the τ is identified and hence $\mathbf{\Gamma}_2$ and $\mathbf{V}_2$. $\sigma_1^2 = \sigma_{11} - d_1^2$ and scaling $\sigma_f^2 = 1$ gives $\lambda_1 = d_1$ and $\boldsymbol{\lambda}_2 = \mathbf{\Gamma}_2'\mathbf{d}_2$.

To interpret the rank condition we will say that the s^{th} equation is not connected to the rest of the structure if $\lambda_s = 0$. If an equation is not connected then it factors out of the likelihood function and y_s is actually exogenous. So our condition says that the exclusion must occur in a connected equation and that either the excluded variable is connected or it appears (with a non-zero coefficient) in a connected equation preceding the one it's excluded from. We should note that even if $\gamma_{sj}\lambda_j \neq 0$ for at least one j, it is still possible for the sum in (II.17) to be zero. But this possibility is of interest only in the unlikely event that there is an a priori restriction of that form.

Corollary 1: In (I.1) with $\gamma_{1k} = 0$ for $k > 1$, a sufficient condition for identification is that a single $\beta_{st} = 0$ provided the following rank condition holds: $\sigma_2^2 > 0, \ldots, \sigma_m^2 > 0$ and

$$\text{(II.20)} \quad \lambda_t \sum_{j=1}^{t-1} \beta_{sj} \lambda_j/\sigma_j^2 \neq 0 .$$

Proof: Rewrite the model as in (II.3) and apply the Theorem. The rank condition does not include the coefficients in the x equations since $\lambda_j = 0$ in that case.

The force of the rank condition is that the exclusion must occur in a connected equation and that $\mathbf{x}_s$ must appear in a connected equation preceding the one it is excluded from. So exclusions in y_1 will not identify the structure.

Our second special case is based on imposing proportionality restrictions across the coefficients of $\underset{\sim}{x}$ and f: $\underset{\sim}{B} = -\underset{\sim}{\eta}\underset{\sim}{\lambda}'$. This case arises when the observed and unobserved characteristics are aggregated into a single index $h_i = \underset{\sim}{x}_i'\underset{\sim}{\eta} + f_i$ and only affect the y's via their effect on h:

$$y_k = \gamma_{1k} y_1 + \ldots + \gamma_{k-1,k} y_{k-1} + \lambda_k h + \nu_k \quad .$$

An example in hedonic models of wage determination (e.g. Chapter 4 or Griliches and Mason, 1972) would be aggregating an individual's background characteristics and unobserved initial ability into one index of early human capital which is then the causal variable in determining measures of later achievement. A similar restriction has also been discussed by Jöreskog and Goldberger (1973).

Corollary 2: In model (I.1) with $\underset{\sim}{B} = -\underset{\sim}{\eta}\underset{\sim}{\lambda}'$, a sufficient condition for identification is that a single $\gamma_{st} = 0$ provided the following rank condition holds: $\underset{\sim}{V} = \text{diag}\ \{\sigma_1^{\ 2}, \ldots, \sigma_m^{\ 2}\}$ is positive definite and

$$\text{(II.21)} \qquad \lambda_t \sum_{j=s}^{t-1} \gamma_{sj}\lambda_j/\sigma_j^{\ 2} \neq 0 \quad .$$

If we restrict ourselves to zero restrictions on $\underset{\sim}{\Gamma}$ then the condition is also necessary.

Proof: $\underset{\sim}{\Pi} = \underset{\sim}{\eta}\underset{\sim}{d}'$ lets us solve for $\underset{\sim}{d}$ up to a sign normalization and a scale factor τ: $\underset{\sim}{g} = \underset{\sim}{d}/\sqrt{\tau}$. Then

$$\underset{\sim}{\Sigma} = \tau \underset{\sim}{g}\underset{\sim}{g}' + \underset{\sim}{\Gamma}'^{-1} \underset{\sim}{V} \Gamma^{-1}$$

will let us solve for τ provided a single element of $\underset{\sim}{\Gamma}$ is zero and the rank condition holds. The proof is the same as in the Theorem. Given τ we obtain $\underset{\sim}{\Gamma}^{-1}$ and $\underset{\sim}{V}$ by Gaussian elimination on $\underset{\sim}{\Sigma} - \tau\underset{\sim}{g}\underset{\sim}{g}'$.

The interpretation of the rank condition is the same as in the Theorem.

III An Example: The Structural Relationship Between Wages and Characteristics

Our example is based on the following sequential income generating model:

$$\begin{aligned}
T_1 &= && && \lambda_1 f + \nu_1 \\
S &= && && \lambda_2 f + \nu_2 \\
T_2 &= \gamma_{23} S && && + \lambda_3 f + \nu_3 \\
Y_1 &= \gamma_{24} S && && + \lambda_4 f + \nu_4 \\
Y_2 &= \gamma_{25} S + \gamma_{45} Y_1 && && + \lambda_5 f + \nu_5
\end{aligned}$$

where T_1 is a test score measuring early (pre-school) ability (f), S is years of schooling, T_2 is a measure of post-school ability, which reflects the value added of the schooling, and Y_1 and Y_2 are repeated observations on earnings.

Potential x's for such a model would include a variety of background variables such as father's schooling, income, or family wealth. If these variables are unrestricted then they do not affect the identification and we will surpress them. But note that some reinterpretation of the model may be necessary in order to make these variables exogenous. Any notion of intergenerational stationarity would suggest that father's schooling and income are subject to a similar set of equations, with an f' for father's "ability". Presumably f' and f are correlated, both for genetic and other reasons. So the background variables are not exogenous unless we reinterpret f to be the part of son's ability that is not predictable from the father's characteristics. This will alter the background coefficients but will not affect the γ's.

A direct application of Theorem 1 shows that without an early test score the model is not identified. For then S $(=y_1)$ is not excluded from any of the other equations (and we are implicitly assuming that none of the λ's are zero). Although the Theorem refers to the model as a whole, its proof shows that none of the schooling coefficients are individually identified, since the rotation indeterminancy will confound each γ_{2k} with λ_k and the preceding λ's and γ's. Note that the identification condition fails even if Y_2 excludes Y_1 (e.g. if there is enough time between the measurements so that they do not have a transitory piece in common). Also adding additional income measurements of this kind does not solve the problem.

But if there is an early test score then Theorem 4 applies. The exclusion of T_2 from Y_1 is sufficient for identification provided neither λ_3 nor λ_4 is zero. In the absence of Y_2 the model is just identified and given ML estimates of $\underset{\sim}{\Pi}$ and $\underset{\sim}{\Sigma}$ we obtain the ML estimates of $\underset{\sim}{\Gamma}$, $\underset{\sim}{V}$, and $\underset{\sim}{B}$, by solving a set of recursive linear equations. If we do have another observation on earnings then the model is overidentified. Chapter 4 indicates how a program by Jöreskog (1970, 1973) can be adapted to impose the constraints.

Next assume that there is a common measurement error in the two tests so that ν_1 and ν_3 are correlated. This particular sort of two factor model, with the second factor only connecting a pair of the equations, can be put into our one factor framework by rewriting the T_2 equation as

$$T_2 = \gamma_{13}T_1 + \gamma_{23}S + \lambda_3'f + \nu_3'$$

with $\gamma_{13} = E(\nu_1\nu_3)/\sigma_1^{2}$ and ν_3' independent of ν_1. Note that the independence of the seond factor from f is simply a reinterpretation of it. To the extent that the second factor is correlated with f it is not affecting our estimates of the structural γ's, although the λ's are affected by the reinterpretation.

Without Y_2 the model as a whole is not identified (since with $\gamma_{13} = 0$ it is just identified by Theorem 4). But Theorem 3 shows that γ_{23} is identified. For we can use T_1 as a proxy for f in the T_2 equation:

$$T_2 = (\gamma_{13} + \frac{\lambda_3'}{\lambda_1})\, T_1 + \gamma_{23}S + \nu_3' - \frac{\lambda_3'}{\lambda_1}\nu_1 \quad .$$

Now use Y_1 and S as instruments. The rank condition has

$$\underset{\sim}{P} = \begin{bmatrix} \sigma_{14} & \sigma_{24} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$= \begin{bmatrix} (\gamma_{24}\lambda_2 + \lambda_4)\lambda_1 & (\gamma_{24}\lambda_2 + \lambda_4)\lambda_2 + \gamma_{24}\sigma_2^{2} \\ \lambda_2\lambda_1 & \lambda_2^{2} + \sigma_2^{2} \end{bmatrix}$$

This is non-singular if $\lambda_1\lambda_4\sigma_2^{2} \neq 0$.

Our last example assumes that ν_1 and ν_2 (and perhaps ν_3) are correlated. This second factor could reflect the part of scholastic ability or "testwiseness" that is not correlated with f. Then the correlation between ν_1 and ν_2 can be captured by calling S the first equation and rewriting the T_1 equation (= y_2):

$$S = \lambda_2 f + \nu_2$$

$$T_1 = \gamma_{12} S + \lambda_1{}' f + \nu_1{}'$$

with $\gamma_{12} = E(\nu_1 \nu_2)/\sigma_2{}^2$. Now ν_2 is independent of $\nu_1{}'$ and we see from Theorem 1 that the model is not identified since S (=y_1) is never excluded. The problem persists no matter how many additional indicators we add, so long as they all include S. Allowing for ν_3 to be correlated with ν_1 and ν_2 raises problems with staying in a one factor framework. But clearly this only makes the model even less identified.

Chapter 3

Unobservables with a Variance Components Structure

1. The General Model

Consider the following model:

$$(I.1)\qquad \underset{\sim}{y}_{ij}'\,\underset{\sim}{\Gamma} + \underset{\sim}{x}_{ij}'\,\underset{\sim}{B} = \underset{\sim}{f}_i'\,\underset{\sim}{\Lambda} + \underset{\sim}{\nu}_{ij}'\ , \qquad i = 1,\ldots,q \qquad j = 1,\ldots,p,$$

where $\underset{\sim}{y}_{ij}$ is an m x 1 vector of endogenous variables, x_{ij} is an n x 1 vector of exogenous variables, $\underset{\sim}{\Gamma}$ is an upper triangular matrix of parameters with ones on the diagonal, $\underset{\sim}{B}$ is an n x m parameter matrix, and the subscripts refer to the j^{th} observation in the i^{th} group. The novelty of the paper lies in the structure of the residuals. They are assumed to have a multivariate variance components decomposition: $\underset{\sim}{f}_i$ is an N x 1 vector of group effects which are distributed as a random sample with covariance matrix $\underset{\sim}{\Phi}$, $\underset{\sim}{\Lambda}$ is an N x m coefficient matrix, and $\underset{\sim}{\nu}_{ij}$ is an m x 1 vector of individual effects which are distributed independently of $\underset{\sim}{f}_i$ as a random sample over i and j with covariance matrix $\underset{\sim}{V}$.

Then under normality assumptions (or limiting ourselves to second order moments), the distribution of $\underset{\sim}{y}$ is completely characterized by the following reduced form parameters:

$$(I.2)\qquad \begin{aligned} \underset{\sim}{\Pi} &= -\,\underset{\sim}{B}\underset{\sim}{\Gamma}^{-1} \\ \underset{\sim}{\Theta} &= (\underset{\sim}{\Lambda}\underset{\sim}{\Gamma}^{-1})'\,\underset{\sim}{\Phi}\,(\underset{\sim}{\Lambda}\underset{\sim}{\Gamma}^{-1}) \\ \underset{\sim}{\Sigma} &= \underset{\sim}{\Gamma}^{-1}{}'\underset{\sim}{V}\underset{\sim}{\Gamma}^{-1}. \end{aligned}$$

The identification problem in this model is to recover $\underset{\sim}{\Gamma}$, $\underset{\sim}{B}$, $\underset{\sim}{\Lambda}$, $\underset{\sim}{\Phi}$, and $\underset{\sim}{V}$ from the reduced form. If the residuals had no group structure and were uncorrelated across equations then the model would be recursive and readily identifiable. But without any covariance restrictions the model is not identified without restrictions on $\underset{\sim}{B}$ and $\underset{\sim}{\Gamma}$. Our approach considers intermediate cases which combine restrictions on $\underset{\sim}{B}$ and $\underset{\sim}{\Gamma}$ with factor analytic restrictions on the covariances. In particular we have developed the following four models:

Model 1: $\underset{\sim}{\Gamma} = \underset{\sim}{I}$

Model 2: $\underset{\sim}{V} = \text{diag}\ \{v_1, \ldots, v_m\}$

Model 3: $N = 1$, $\underset{\sim}{\Lambda} = \underset{\sim}{\lambda}'$ where $\underset{\sim}{\lambda}$ is m x 1, $\underset{\sim}{\Gamma} = \underset{\sim}{I}$, $\underset{\sim}{B} = -\underset{\sim}{\eta}\underset{\sim}{\lambda}'$ where $\underset{\sim}{\eta}$ is n x 1.

Model 4: $N = 1$, $\underset{\sim}{\Lambda} = \underset{\sim}{\lambda}'$, $\underset{\sim}{B} = -\underset{\sim}{\eta}\underset{\sim}{\lambda}'$, $\underset{\sim}{V} = \tau\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U}$ where τ is a positive scalar and $\underset{\sim}{U} = \text{diag}\ \{\sigma_1^2, \ldots, \sigma_m^2\}$.

Our interest in these models stems from the work by Chamberlain and Griliches (1974). They used data on brothers to estimate the return to education in the presence of an unobserved ability variable. In their model the group is a family and $\underset{\sim}{y}$ could include years of schooling, test scores, income, occupational status, etc.; $\underset{\sim}{x}$ could include age and family background characteristics such as father's income and schooling. The residual covariances are generated by a common omitted "ability" variable with a variance components structure: $a_{ij} = f_i + g_{ij}$. The restrictions on $\underset{\sim}{B}$ in models 3 and 4 arise when the background variables are combined with the unobservable to form a "human capital" variable $\underset{\sim}{x}'_{ij}\underset{\sim}{\eta} + a_{ij}$ which appears with coefficient λ_k in the k^{th} equation. This sort of restriction was also used by Griliches and Mason (1972) and by Jöreskog and Goldberger (1973).

The plan of the paper is as follows: Section II provides an identification analysis of these models; Section III derives maximum likelihood estimators and Section IV provides an interpretation of them. Section V developes an example based on the causes and consequences of permanent income.

II. Identification

Model 1: Here the only problem is to recover $\underset{\sim}{\Lambda}$ and $\underset{\sim}{\Phi}$ from $\underset{\sim}{\Theta} = \underset{\sim}{\Lambda}'\underset{\sim}{\Phi}\underset{\sim}{\Lambda}$. If $\underset{\sim}{\Phi}$ is restricted to an identity matrix then this is a standard rotation problem in factor analysis. With $\underset{\sim}{\Phi}$ non-diagonal it is an identification problem with oblique factors. Some results are reported in Reiersol (1950), Howe (1955), and Anderson and Rubin (1956). For example, in the oblique case with $\underset{\sim}{\Phi}$ unrestricted except for scale normalizations that fix the diagonal elements, it is sufficient that each row of $\underset{\sim}{\Lambda}$ have at least $N - 1$ fixed elements provided the following rank condition holds. Let $\underset{\sim}{\Lambda}$ be any solution satisfying the restrictions and let $\underset{\sim}{\Lambda}^s$ be the submatrix of $\underset{\sim}{\Lambda}$ consisting of those columns that have fixed elements in the s^{th} row. Then $\underset{\sim}{\Lambda}$ is unique if for all $s = 1,\ldots,N$ we have rank $(\underset{\sim}{\Lambda}^s)$ equal to the smallest of the number m_s and N, where m_s is the number of fixed elements in the s^{th} row of $\underset{\sim}{\Lambda}$.

There is an example in Section V of a multi-factor model with enough restrictions to uniquely solve for $\underset{\sim}{\Lambda}$ and $\underset{\sim}{\Phi}$.

Model 2: If $\underset{\sim}{V} = \text{diag}\ \{v_1,\ldots,v_m\}$ has positive diagonal elements then $\underset{\sim}{\Sigma}$ is positive definite and thus has a unique factorization by Gaussian elimination into $\underset{\sim}{A}'\underset{\sim}{C}\underset{\sim}{A}$ where $\underset{\sim}{A}$ is an upper triangular matrix with ones on the diagonal and $\underset{\sim}{C}$ is a diagonal matrix of positive elements (e.g. Gantmacher

(1959), chapter II). Then simply identify Γ^{-1} with A and V with C. Given Γ we recover B from Π and $\Gamma'\Theta\Gamma = \Lambda'\Phi\Lambda$ leaves us with the same identification problem as Model 1.

Model 3: Since there is only one factor we just need a scale normalization. So setting $\Phi = 1$ we can use $\Theta = \lambda\lambda'$ to solve for λ (up to a sign normalization) and $\Pi = -\eta\lambda'$ lets us solve for η.

Model 4: This presents the most interesting identification problem. Clearly if we knew τ (and with the group effects scaled so that $\Phi = 1$), we could recover Γ^{-1} and U from $\Sigma - \tau\Theta$ by Gaussian elimination. So we need one additional piece of information. First we will consider zero restrictions on Γ: $\gamma_{st} = 0$, without restricting B.

Theorem 1: In the one factor model with $V = \tau\lambda\lambda' + U$, a sufficient condition for identification is that a single $\gamma_{st} = 0$ provided the following rank condition holds: $U = \text{diag}\ \{\sigma_1^2,\dots,\sigma_m^2\}$ is positive definite and

$$\text{(II.1)} \qquad \lambda_t \sum_{j=s}^{t-1} \gamma_{sj}\ \lambda_j/\sigma_j^2 \neq 0.$$

The condition is also necessary if we confine ourselves to zero restrictions on Γ.

Proof: Let $d = \Gamma'^{-1}\lambda$. Then with the group effects scaled so that $\Phi = 1$, we have $\Theta = dd'$ and $(\Sigma - \tau dd')^{-1} = \Sigma^{-1} + \kappa cc' = \Gamma U^{-1}\Gamma'$ where $\kappa = \tau/(1 - \tau d'\Sigma^{-1}d)$ and $c = \Sigma^{-1}d$. It will simplify the notation to reverse the order of the equations so that, for this proof only, Γ is lower triangular. So now the zero element is γ_{hk} where $h = m - s + 1$, $k = m - t + 1$. Then let $A = \Sigma^{-1} - \kappa cc'$ and use Gaussian elimination to solve for γ_{hk}:

$$\text{(II.2)}\qquad \gamma_{hk} = A\binom{1\ 2\ \dots\ k-1\ h}{1\ 2\ \dots\ k-1\ k} \Big/ A\binom{1\ 2\ \dots\ k-1}{1\ 2\ \dots\ k-1}$$
$$= 0,$$

where $A\binom{h_1\ \cdots\ h_p}{k_1\ \cdots\ k_p}$ is the minor formed from rows $h_1,\dots,h_p$ and columns $k_1,\dots,k_p$ of $\underset{\sim}{A}$ (Gantmacher (1959), chapter II).

Expanding the bordered determinant gives

$$A\binom{1\ 2\ \dots\ k-1\ h}{1\ 2\ \dots\ k-1\ k} = A\binom{1\ 2\ \dots\ k-1}{1\ 2\ \dots\ k-1}\,(a_{hk} - \underset{\sim}{a}_h{}'\,\bar{\underset{\sim}{A}}^{-1}\underset{\sim}{a}_k)$$

where $\underset{\sim}{a}_i{}' = (a_{i1}\dots a_{i,k-1})$, $i = h,k$, and $\bar{\underset{\sim}{A}}$ is the (k-1) by (k-1) principal submatrix of $\underset{\sim}{A}$. So we must solve for τ from

$$a_{hk} - \underset{\sim}{a}_h{}'\,\bar{\underset{\sim}{A}}^{-1}\underset{\sim}{a}_k = 0.$$

Let $\bar{\underset{\sim}{A}} = \underset{\sim}{S} + \kappa\bar{\underset{\sim}{c}}\bar{\underset{\sim}{c}}'$ where $\underset{\sim}{S}$ is the (k-1) by (k-1) principal submatrix of $\underset{\sim}{\Sigma}^{-1}$ and $\bar{\underset{\sim}{c}}$ contains the first k-1 elements of $\underset{\sim}{c}$. Then $\underset{\sim}{a}_i = \underset{\sim}{s}_i + \kappa c_i\bar{\underset{\sim}{c}}$ with $\underset{\sim}{s}_i{}' = (\sigma^{i1}\dots\sigma^{i,k-1})$, $i = h, k$ and we can write the restriction as

$$\text{(II.3)}\qquad \sigma^{hk} + \kappa c_h c_k$$
$$- (\underset{\sim}{s}_h + c_h\bar{\underset{\sim}{c}})'[\underset{\sim}{S}^{-1} - \kappa(\underset{\sim}{S}^{-1}\bar{\underset{\sim}{c}}\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1})/(1 + \kappa\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\bar{\underset{\sim}{c}})](\underset{\sim}{s}_k + \kappa c_k\bar{\underset{\sim}{c}}) = 0.$$

Fortunately this can be simplified to the following linear equation in κ:

$$\text{(II.4)}\qquad \sigma^{hk} - \underset{\sim}{s}_h{}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_k + \kappa Q = 0$$

where

$$Q = (\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\bar{\underset{\sim}{c}})(\sigma^{hk} - \underset{\sim}{s}_h{}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_k) + c_h c_k + c_h(\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_k) + c_k(\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_h)$$
$$- (\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_h)(\bar{\underset{\sim}{c}}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_k).$$

So $\tau = \kappa/(1 + \kappa\underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d})$ is uniquely, globally, identifiable iff $Q \neq 0$. But if $Q = 0$ then $\gamma_{hk} = 0$ implies that $\sigma^{hk} - \underset{\sim}{s}_h{}'\underset{\sim}{S}^{-1}\underset{\sim}{s}_k = 0$. If this holds for $Q \neq 0$ then $\kappa = \tau = 0$, a possibility we will exclude since τ is a variance ratio. So our rank condition can be written

$$\text{(II.5)}\qquad \Sigma^{-1}\binom{1\ 2\ \dots\ k-1\ h}{1\ 2\ \dots\ k-1\ k} \neq 0.$$

In order to write this in terms of the structural paramters, we will apply the Cauchy-Binet formula (Gantmacher, chapt. 1) to $\underset{\sim}{\Sigma}^{-1} = \underset{\sim}{\Gamma}(\tau\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U})^{-1}\underset{\sim}{\Gamma}'$:

(II.6) $$\Sigma^{-1}\begin{pmatrix}1 & 2 & \dots & k-1 & h\\ 1 & 2 & \dots & k-1 & k\end{pmatrix}$$

$$= \sum_{p_i,q_i} \{\Gamma\begin{pmatrix}1 & 2 & \dots & k-1 & h\\ p_1 & p_2 & \dots & p_{k-1} & p_k\end{pmatrix}\}\{(\tau\lambda\lambda' + U)^{-1}\begin{pmatrix}p_1 & \dots p_k\\ q_1 & \dots q_k\end{pmatrix}\}\{\Gamma\begin{pmatrix}1 & 2 & \dots & k-1 & k\\ q_1 & \dots & q_{k-1} & q_k\end{pmatrix}\}$$

$$\neq 0$$

where $1 \leq p_1 < \dots < p_k \leq m$ and $1 \leq q_1 < \dots < q_k \leq m$. Since $\underset{\sim}{\Gamma}$ is lower triangular with ones on the diagonal this can be simplified to

(II.7) $$\sum_{j=k}^{h} \gamma_{hj}\{(\tau\lambda\lambda' + U)^{-1}\begin{pmatrix}1 & 2 & \dots & k-1 & j\\ 1 & 2 & \dots & k-1 & k\end{pmatrix}\} \neq 0.$$

So our rank condition can be written as

(II.8) $$\lambda_k \sum_{j=k+1}^{h} \gamma_{hj}\ \lambda_j/\sigma_j^{\,2} \neq 0\ .$$

Reordering the equations so that $\underset{\sim}{\Gamma}$ is upper triangular gives the condition in the Theorem and completes our proof.

To interpret the rank condition we will say that the s^{th} equation is not connected to the rest of the structure if $\lambda_s = 0$. If an equation is not connected then it factors out of the likelihood function and y_s is actually exogenous. So our condition says that the exclusion must appear in a connected equation and that either the excluded variable is connected or it appears (with a non-zero coefficient) in a connected equation preceding the one it's excluded from. In this latter case y_s is exogenous and we are insisting that it be correlated with one of the endogenous variables that can appear in the t^{th} equation. This condition is similar to the rank condition in Theorem 2 for identification by excluding exogenous variables.

We should note that even if $\gamma_{sj}\ \lambda_j \neq 0$ for at least one j, it is still possible for the sum in (II.1) to be zero. But this possibility is uninteresting and extremely unlikely.

Corollary 1 (Model 4): In the one factor model, with $\underset{\sim}{V} = \tau\underset{\sim\sim}{\lambda\lambda}' + \underset{\sim}{U}$ and $\underset{\sim}{B} = -\underset{\sim\sim}{\eta\lambda}'$, the necessary and sufficient condition for identification (within the class of zero restrictions on $\underset{\sim}{\Gamma}$) is the one given in Theorem 1.

Proof: Let $\underset{\sim}{d} = \underset{\sim}{\Gamma}'^{-1}\underset{\sim}{\lambda}$ and scale $\Phi = 1$ so that $\underset{\sim}{\Theta} = \underset{\sim\sim}{dd}'$. Then $\underset{\sim}{\Pi} = -\underset{\sim\sim}{\eta d}'$ lets us solve for $\underset{\sim}{\eta}$ but otherwise contains no information that is not in $\underset{\sim}{\Theta}$.

To interpret this result we note that treating the f_i as fixed effects would lead to a set of group dummy variables whose coefficients would be constrained as in the corollary. So adding more variables (i.e. $\underset{\sim}{X}$) constrained in this way does not affect the identification analysis.

Corollary 2: If the proportionality constraint of corollary 1 holds only across a subset of the equations, e.g. $\underset{\sim}{B} = (\underset{\sim}{B}_1 \vdots -\underset{\sim\sim}{\eta\bar{\lambda}}')$ where $\underset{\sim}{\bar{\lambda}}' = (\lambda_t\ , \ldots, \lambda_m)$, then it is still true that the necessary and sufficient condition for identification is the one given in Theorem 1.

Proof: We will do the case in which only the first equation is unconstrained and leave the extension to the reader. Write the first column of $\underset{\sim}{B}$ as $-(\lambda_1\underset{\sim}{\eta} + \underset{\sim}{\zeta})$ so that $\underset{\sim}{B} = -\underset{\sim\sim}{\eta\lambda}' - (\underset{\sim}{\zeta}\ \underset{\sim}{0})$. Then $\underset{\sim}{\Pi} = -(\underset{\sim\sim}{\eta d}' + \underset{\sim\sim}{\zeta a}')$ where $\underset{\sim}{a}'$ is the first row of $\underset{\sim}{\Gamma}^{-1}$. Now the first row of $\underset{\sim}{\Gamma}'^{-1}\underset{\sim\sim}{U\Gamma}^{-1}$ is $\sigma_1{}^2\underset{\sim}{a}'$ and so

$$\underset{\sim}{\Pi} = -[\underset{\sim\sim}{\eta d}' + \underset{\sim}{\zeta}(\underset{\sim}{s}_1' - \tau d_1\underset{\sim}{d}')/\sigma_1{}^2]$$

where $\underset{\sim}{s}_1'$ is the first row of $\underset{\sim}{\Sigma} = \tau\underset{\sim\sim}{dd} + \underset{\sim}{\Gamma}'^{-1}\underset{\sim\sim}{U\Gamma}^{-1}$.

Form an m x m non-singular matrix $\underset{\sim}{T} = (\underset{\sim}{t}_1 \ldots \underset{\sim}{t}_m)$ such that $\underset{\sim}{t}_1'\underset{\sim}{d} = 0$, $\underset{\sim}{t}_2'\underset{\sim}{s}_1 = 0$, and $\underset{\sim}{t}_i$ is orthogonal to $\underset{\sim}{d}$ and $\underset{\sim}{s}_1$ for i = 3, ..., m (amendments are straightforward in the unlikely event that $\underset{\sim}{d} \propto \underset{\sim}{s}_1$). Then the information in $\underset{\sim}{\Pi}$ is equivalent to the information in

$$\underset{\sim\sim}{\Pi T} = -[\underset{\sim}{\zeta}(\underset{\sim}{s}_1'\underset{\sim}{t}_1)/\sigma_1^2 \vdots (\underset{\sim}{\eta} - \frac{\tau d_1}{\sigma_1^2} \underset{\sim}{\zeta})(\underset{\sim}{d}'\underset{\sim}{t}_2) \vdots \underset{\sim}{0}]$$

So we can solve for $\underset{\sim}{\zeta}/\sigma_1^2$ and for $\underset{\sim}{\eta} - \tau d_1/\sigma_1^2 \underset{\sim}{\zeta}$. Given any τ we can solve for $\underset{\sim}{\eta}$ and a triangular factorization of $\underset{\sim}{\Sigma} - \tau \underset{\sim\sim}{dd}'$ will give σ_1^2 and hence $\underset{\sim}{\zeta}$. Thus there is no additional information on τ and its identification must come from $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$ as in Theorem 1.

A natural extension of the proportionality restrictions would be to impose them across some of the γ's in addition to $\underset{\sim}{B}$ and $\underset{\sim}{\lambda}$. Consider the following example:

$$\text{(II.9)}\quad \begin{aligned} y_1 &= (\underset{\sim}{x}'\underset{\sim}{\eta} + f)\lambda_1 + \nu_1 \\ y_2 &= (\underset{\sim}{x}'\underset{\sim}{\eta} + f)\lambda_2 + \gamma_{12}y_1 + \nu_2 \\ y_k &= [(\underset{\sim}{x}'\underset{\sim}{\eta} + f)\lambda_2 + \gamma_{12}y_1]\ \lambda_3/\lambda_2 + \gamma_{2k}y_2 + \ldots + \gamma_{k-1,k}y_{k-1} + \nu_k,\ k=3,\ldots,m, \end{aligned}$$

where the ν's have the model 4 covariance matrix $V = \tau\underset{\sim\sim}{\lambda\lambda}' + \underset{\sim}{U}$.

Corollary 3: The necessary and sufficient condition for identification of (II.9) is the one given in Theorem 1.

Proof: Start with a three equation model. Instead of having a $\gamma_{st} = 0$ we have the non-linear restriction that $\gamma_{12}/\gamma_{13} = \lambda_2/\lambda_3$. So in terms of counting restrictions and unknowns it would appear that we are identified without the zero restriction of $\underset{\sim}{\Gamma}$. But writing the restriction in terms of reduced form parameters gives $(\sigma_{12} - \tau d_1 d_2)/(\sigma_{13} - \tau d_1 d_3) = d_1/d_3$ and unfortunately τ cancels out. Instead of being able to solve for τ we have the reduced form constraint $\sigma_{12}/\sigma_{13} = d_2/d_3$. Thus the number of reduced form degrees of freedom is effectively reduced by one. Adding more equations gives more restrictions of the form $\gamma_{12}/\gamma_{1k} = \lambda_2/\lambda_k$ but they translate directly into the reduced form restrictions $\sigma_{12}/\sigma_{1k} = d_2/d_k$ without letting us solve for τ.

Next we will consider arbitrary linear restrictions on the endogenous variables in a given equation. Such a restriction on the t^{th} equation can be written $\gamma_{gt} + \ell_{g+1}\,\gamma_{g+1,t} + \ldots + \ell_{t-1}\,\gamma_{t-1,t} + \ell_t = 0$ where the first non-zero element in $\underset{\sim}{\ell}$ was ℓ_g and we have divided through by it.

Corollary 4: The restriction that $\gamma_{gt} + \ell_{g+1}\,\gamma_{g+1,t} + \ldots + \ell_{t-1}\,\gamma_{t-1,t} + \ell_t = 0$ is sufficient for identification provided the following rank condition holds:

(II.10) $$\lambda_t \sum_{j=g}^{t-1} \tilde{\gamma}_{gj}\, \lambda_j/\sigma_j^2 \neq 0$$

where
$$\tilde{\gamma}_{gj} = \gamma_{gj} + \sum_{h=g+1}^{j} \ell_h\, \gamma_{h,j} \quad .$$

Proof: Let $\underset{\sim}{P}$ be an upper triangular matrix which only differs from an identity matrix in that $p_{gj} = \ell_j$, $j = g + 1, \ldots, t$. Then we can rewrite our model as $(\underset{\sim}{y}'\underset{\sim}{P}^{-1})(\underset{\sim}{P}\underset{\sim}{\Gamma}) + \underset{\sim}{x}'\underset{\sim}{B} = \underset{\sim}{f}'\underset{\sim}{\Lambda} + \underset{\sim}{\nu}'$ where $\underset{\sim}{\tilde{\Gamma}} = \underset{\sim}{P}\underset{\sim}{\Gamma}$ is still upper triangular with ones on the diagonal. Now the restriction is $\tilde{\gamma}_{gt} = 0$ and we apply Theorem 1.

The rank condition requires that the constrained equation be connected to the rest of the structure. In addition we must have y_g connected or a $\tilde{\gamma}_{gj} \neq 0$ for a connected equation between y_g and y_t (i.e. $g < j < t$). $\tilde{\gamma}_{gj} \neq 0$ requires that y_j is included in the restriction ($\ell_j \neq 0$) or that the j^{th} equation includes y_g or a later variable which is included in the restriction.

Theorem 2: Consider the one factor model with $\underset{\sim}{V} = \tau\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U}$ and $\underset{\sim}{B}$ unrestricted except for zero restrictions. Then the necessary and sufficient condition for identification by excluding exogenous variables is that a single $\beta_{st} = 0$ together with the following rank condition:
$U = \text{diag}\ \{\sigma_1^2, \ldots, \sigma_m^2\}$ is positive definite and

(II.11)
$$\lambda_t \sum_{j=1}^{t-1} \tilde{\gamma}_j \lambda_j/\sigma_j^2 \neq 0$$

where $\tilde{\gamma}_j = \sum_{h=1}^{j} \pi_{sh}\gamma_{hj}$ and π_{sh} is the (s,h) element of $\underset{\sim}{\Pi}$.

An important implication of this rank condition is that X_s must actually appear in some equation preceding the one it is excluded from,

Proof: $\underset{\sim}{B} = -\underset{\sim}{\Pi}\underset{\sim}{\Gamma}$ and so the restriction that $\beta_{st} = 0$ implies $\sum_{h=1}^{t} \pi_{sh}\gamma_{ht} = 0$.

Now apply corallary 4. By locating the first non-zero element in row s of $\underset{\sim}{\Pi}$ we can write (II.11) in the form used in corollary 4. The condition that some $\beta_{sj} \neq 0$ for $j < t$ is necessary to ensure that some $\pi_{sh} \neq 0$ for $h < t$, since $\underset{\sim}{\Pi} = -\underset{\sim}{B}\,\underset{\sim}{\Gamma}^{-1}$ and $\underset{\sim}{\Gamma}^{-1}$ is upper triangular. Other implications of the rank condition follow from our discussion of corollary 4. For example the exclusion must occur in a connected equation.

III. Estimation

We will describe maximum likelihood (ML) algorithms under normality assumptions. In most cases it is not possible to give a complete analytic solution. Then our aim is maximum analytic concentration of the likelihood function before turning to numerical techniques. The derivations are given in an Appendix. Interpretations of our algorithms will be given in the next section.

Models 1 and 2: In both models the reduced form $\underset{\sim}{\Pi}$, $\underset{\sim}{\Theta}$, and $\underset{\sim}{\Sigma}$ are unconstrained except for the rank restriction on $\underset{\sim}{\Theta}$. First we will derive the ML estimator of $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$ conditional on $\underset{\sim}{\Pi}$. Arrange the observations so that the first p are from group 1, the second p are from group 2, etc. Then let

$$\underset{\sim}{Y} = \begin{bmatrix} \underset{\sim}{y}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{y}'_{pq} \end{bmatrix} , \qquad \underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{x}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{x}'_{pq} \end{bmatrix}$$

($\underset{\sim}{Y}$ is pq x m, $\underset{\sim}{x}$ is pq x n) and form the matrix of reduced form residuals $\underset{\sim}{E} = \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\Pi}$. Let $\underset{\sim}{J} = I_{\sim q} \otimes \ell_{\sim p}$ be a set of group indicator dummy variables where $\ell_{\sim p}$ is a p x 1 vector of ones. $\underset{\sim}{R} = \underset{\sim}{E}'\underset{\sim}{E}/pq$ is the sample covariance matrix of the residuals and $\underset{\sim}{\bar{R}} = \underset{\sim}{E}'\underset{\sim}{J}\underset{\sim}{J}'\underset{\sim}{E}/qp^2$ is formed by first averaging the residuals over each group and then forming their sample covariance matrix.

Then solve the eigenvalue problem

(III.1) $\underset{\sim}{\bar{R}}\underset{\sim}{G} = \underset{\sim}{R}\underset{\sim}{G}\underset{\sim}{K}$

where $\underset{\sim}{K} = \text{diag}\,\{\rho_1, \ldots, \rho_N\}$ contains the N largest eigenvalues and $\underset{\sim}{G}$ contains the eigenvectors scaled so that

$$\underset{\sim}{G}'\underset{\sim}{\bar{R}}\underset{\sim}{G} = (\rho\underset{\sim}{K} - \underset{\sim}{I})(\underset{\sim}{I} - \underset{\sim}{K})^{-1} / p^2 .$$

Θ is constructed from

(III.2) $\Theta = p^2/p\text{-}1 \;\; RGK(I - K)G'R$

and

(III.3) $\Sigma = R - \Theta$.

The ML estimator of Π given Θ and Σ is generalized least square (GLS). We arrange the columns of Π into a single stacked mn x 1 vector $\delta = \text{vec}(\Pi)$. The computations are simplified by analytically inverting the disturbance covariance matrix to obtain the following formula for the GLS estimator of δ:

(III.4) $\delta^* = (H_W + H_B)^{-1}(H_W\hat{\delta}_W + H_B\hat{\delta}_B)$

where $\hat{\delta}_W$ is the least squares estimate just using the within family moments and $\hat{\delta}_B$ just uses the between family moments:

$$\hat{\delta}_{wk} = W_x^{-1} W_{xy_k}$$

$$\hat{\delta}_{Bk} = B_x^{-1} B_{xy_k} \quad , \quad k = 1,\ldots,m,$$

with $T_x = X'X$, $B_x = X'JJ'X/p$, $W_x = T_x - B_x$ and similar expressions for W_{xy_k} and B_{xy_k}. H_W and H_B are the precision matrices for $\hat{\delta}_W$ and $\hat{\delta}_B$:

(III.5) $H_W = [E(\hat{\delta}_W - \delta)(\delta_W - \delta)']^{-1} = \Sigma^{-1} \otimes W_x$

$$H_B = [E(\hat{\delta}_B - \delta)(\hat{\delta}_B - \delta)']^{-1} = 1/p \;(\Theta + 1/p\;\Sigma)^{-1} \otimes B_x \quad .$$

So we pool the "within" and "between" OLS estimators, weighting by their precision matrices.

If the x's differ across equations then the ML estimator of δ based on just the within group deviations is not $\hat{\delta}_W$ but rather (conditional on Σ) the Zellner "seemingly unrelated" GLS estimator:

$$\underset{\sim}{\delta}_{W}^{GLS} = \begin{bmatrix} \sigma^{11}W_{\underset{\sim}{x}_1x_1} & \cdots & \sigma^{1m}W_{\underset{\sim}{x}_1x_m} \\ \vdots & & \vdots \\ \sigma^{m1}W_{\underset{\sim}{x}_mx_1} & \cdots & \sigma^{mm}W_{\underset{\sim}{x}_mx_m} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^{m} \sigma^{1k} W_{\underset{\sim}{x}_1y_k} \\ \vdots \\ \sum_{k=1}^{m} \sigma^{mk} W_{\underset{\sim}{x}_my_k} \end{bmatrix} .$$

where $\underset{\sim}{x}_k$ contains the exogenous variables actually included in the k^{th} equation. There is a similar estimator $\underset{\sim}{\delta}_B^{GLS}$ (using just the between group variation) which replaces $\underset{\sim}{\Sigma}^{-1}$ by $1/p\ (\underset{\sim}{\Theta} + 1/p\ \underset{\sim}{\Sigma})^{-1}$. The ML estimator of $\underset{\sim}{\delta}$ is a matrix weighted average of these between and within group GLS estimators, weighting by their precision matrices:

(III.6) $\quad \underset{\sim}{\delta}^* = (\underset{\sim}{H}_W + \underset{\sim}{H}_B)^{-1}(\underset{\sim}{H}_W\underset{\sim}{\delta}_W^{GLS} + \underset{\sim}{H}_B\underset{\sim}{\delta}_B^{GLS})$

with

(III.7) $\quad \underset{\sim}{H}_W = [E(\underset{\sim}{\delta}_W^{GLS} - \underset{\sim}{\delta})(\underset{\sim}{\delta}_W^{GLS} - \underset{\sim}{\delta})']^{-1} = \underset{\sim}{\Sigma}^{-1} * \underset{\sim}{\bar{W}}$

$\quad \underset{\sim}{H}_B = [E(\underset{\sim}{\delta}_B^{GLS} - \underset{\sim}{\delta})(\underset{\sim}{\delta}_B^{GLS} - \underset{\sim}{\delta})']^{-1} = 1/p\ (\underset{\sim}{\Theta} + 1/p\ \underset{\sim}{\Sigma})^{-1} * \underset{\sim}{\bar{B}},$

where the k, k' block of $\underset{\sim}{\bar{W}}$ is $W_{\underset{\sim}{x}_kx_{k'}}$ and "*" is a generalized Hadamard product which sets the k, k' block of $\underset{\sim}{\Sigma}^{-1} * \underset{\sim}{\bar{W}}$ equal to $\sigma^{kk'}W_{\underset{\sim}{x}_kx_{k'}}$ with a similar expression for the k, k' block of $\underset{\sim}{H}_B$.

The GLS procedure can be simplified by concentrating the intercepts out of the likelihood function. This is possible since the ML estimates of the hyperplanes corresponding to each of the equations pass through the sample means. Thus if we partition $\underset{\sim}{\delta}_k$ into the intercept δ_{1k} and the slope coefficients $\underset{\sim}{\delta}_{2k}$, then conditional on $\underset{\sim}{\delta}_{2k}$ the GLS estimate of δ_{1k} is OLS:

$$\delta_{1k} = \bar{y}_k - \underset{\sim}{\bar{x}}'\underset{\sim}{\delta}_{2k}, \quad k = 1,\ldots,m,$$

where $\bar{y}_k$ is the grand mean of y_k and $\underset{\sim}{\bar{x}}$ is the row vector of grand means for the exogenous variables (other than the intercept). So the δ_{1k} can be concentrated out of the likelihood function simply by replacing each variable by its deviation from the overall sample mean and proceeding without intercepts.

Then the joint maximum for $\underset{\sim}{\Pi}$, $\underset{\sim}{\Theta}$, and $\underset{\sim}{\Sigma}$ can be obtained by iterating on the ML equations for $\underset{\sim}{\Pi}$ (given $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$) and the ML equations for $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$ (given $\underset{\sim}{\Pi}$).

It is fairly straightforward to modify the algorithm to deal with unbalanced samples. For example in the one factor case with $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ we let α index the different group sizes with p_α individuals in each of q_α groups. The total number of groups is $q = \sum_\alpha q_\alpha$ and with $\bar{p} = (\sum_\alpha p_\alpha q_\alpha)/q$ there are $\bar{p}q$ observations in the total sample. In order to aggregate over groups of different sizes we have to condition on $\psi = \underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$:

$$\text{(III.8)} \quad \bar{\underset{\sim}{R}} = \sum_\alpha p_\alpha q_\alpha [p_\alpha\psi/(1 + p_\alpha\psi)]\bar{\underset{\sim}{R}}_\alpha/\bar{p}q$$

$$\underset{\sim}{R} = \sum_\alpha p_\alpha q_\alpha \underset{\sim}{R}_\alpha/\bar{p}q \quad .$$

We obtain $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$ from the eigenvalue decomposition of $\bar{\underset{\sim}{R}}$ in the metric of $\underset{\sim}{R}$. Then the concentrated likelihood function which is derived in the Appendix just depends on ψ, leaving us with a straightforward one-dimensional maximization problem.

<u>Model 3</u>: Here the structural and reduced forms are identical. It will be convenient to use the reduced form notation with the unobservable scaled so that $\Phi = 1$ and $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$. We will display explicit ML estimators for $\underset{\sim}{d}$, $\underset{\sim}{\Sigma}$, and $\underset{\sim}{\eta}$ conditional on $\psi = \underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$. Then the likelihood can be concentrated to a function just of ψ, leaving a simple scalar maximization problem.

The ML estimator of $\underset{\sim}{d}$ is obtianed from the following eigenvalue problem:

$$\text{(III.9)} \quad \underset{\sim}{Q}^{-1}\underset{\sim}{d} = 1/\rho \; \underset{\sim}{T}_Y{}^{-1}\underset{\sim}{d}$$

where $1/\rho$ is the smallest root and

$$\underset{\sim}{Q} = \underset{\sim}{B}_Y + \frac{1}{1-\zeta}\ \underset{\sim}{H}_{YX}\underset{\sim}{H}_X{}^{-1}\underset{\sim}{H}_{XY}$$

$$\psi = d'\Sigma^{-1}d, \quad \zeta = 1/(1 + p\psi)$$

$$\underset{\sim}{H}_{YX} = \underset{\sim}{W}_{YX} + \zeta\underset{\sim}{B}_{YX}\ , \quad \underset{\sim}{H}_{XY} = \underset{\sim}{H}'_{YX}$$

$$\underset{\sim}{H}_X = \underset{\sim}{W}_X + \zeta\underset{\sim}{B}_X$$

with $\underset{\sim}{T}_Y = \underset{\sim}{Y}'\underset{\sim}{Y}$, $\underset{\sim}{B}_Y = \underset{\sim}{Y}'\underset{\sim}{J}\underset{\sim}{J}'\underset{\sim}{Y}/p$, $\underset{\sim}{W}_Y = \underset{\sim}{T}_Y - \underset{\sim}{B}_Y$, and similar expressions for $\underset{\sim}{W}_{YX}$, $\underset{\sim}{B}_{YX}$, $\underset{\sim}{W}_X$ and $\underset{\sim}{B}_X$.

The ML estimator of $\underset{\sim}{\Sigma}$ is

(III.10) $\quad \underset{\sim}{\Sigma} = \underset{\sim}{T}_Y/pq - \xi\zeta\underset{\sim}{d}\underset{\sim}{d}'$

where ξ is a simple function of ψ and ρ (see Appendix).

Once we have computed the ML estimators of $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ (conditional on ψ), we form $\underset{\sim}{r} = \underset{\sim}{Y}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}/\psi$ and obtain the following ML estimator of $\underset{\sim}{\eta}$:

(III.11) $\quad \underset{\sim}{\eta} = (\underset{\sim}{W}_x + \zeta/1\text{-}\zeta\ \underset{\sim}{T}_x)^{-1}(\underset{\sim}{W}_{xr} + \zeta/1\text{-}\zeta\ \underset{\sim}{T}_{xr})$.

Model 4: If the model is just identified, e.g. only a single $\gamma_{st} = 0$, then the reduced form $\underset{\sim}{\Sigma}$ is unconstrained and the model 3 algorithm can be used. Otherwise we first condition on $\underset{\sim}{\Gamma}$, $\underset{\sim}{U} = \text{diag}\ \{\sigma_1{}^2,\ldots,\ \sigma_m{}^2\}$, τ and $\mu = \underset{\sim}{\lambda}'\underset{\sim}{U}^{-1}\underset{\sim}{\lambda}$ to obtain ML estimators of $\underset{\sim}{\lambda}$ and $\underset{\sim}{\eta}$:

(III.12) $\quad \underset{\sim}{Q}^{-1}\underset{\sim}{\lambda} = 1/\rho\ \underset{\sim}{U}^{-1}\underset{\sim}{\lambda}$

where $1/\rho$ is the smallest root and

$$Q = \{\ \frac{\tau\mu}{\tau\mu+1}\ \underset{\sim}{T}_A + \frac{1}{\tau\mu+1}\ [(1 - \zeta)\underset{\sim}{B}_A + \underset{\sim}{H}_{AX}\underset{\sim}{H}_X{}^{-1}\underset{\sim}{H}_{XA}]\}\ /\ pq$$

with $\mu = \underset{\sim}{\lambda}'\underset{\sim}{U}^{-1}\underset{\sim}{\lambda}$, $\zeta = 1/[1 + p(\mu/\tau\mu + 1)]$ and $\underset{\sim}{T}_A$, $\underset{\sim}{B}_A$, $\underset{\sim}{H}_{AX}$ are defined as in (III.9) with $\underset{\sim}{A} = \underset{\sim}{Y}\underset{\sim}{\Gamma}$ replacing $\underset{\sim}{Y}$.

The ML estimator of $\underset{\sim}{\eta}$ is

$$(\text{III.13}) \quad \underset{\sim}{\eta} = (\underset{\sim}{W}_x + \zeta/1-\zeta\ \underset{\sim}{T}_x)^{-1}(\underset{\sim}{W}_{xr} + \zeta/1-\zeta\ \underset{\sim}{T}_{xr})$$

where $\underset{\sim}{r} = \underset{\sim}{A}\underset{\sim}{U}^{-1}\underset{\sim}{\lambda}/\underset{\sim}{\mu}$. Then the likelihood can be concentrated to a function of τ, μ, σ_k^2, $k = 1,\ldots,m$, and maximized numerically, still conditioning on $\underset{\sim}{\Gamma}$ (see Appendix for details).

The ML estimator of Γ is obtained from the GLS formula in (III.6) where the seemingly unrelated equations are

$$y_k - \underset{\sim}{x}'\underset{\sim}{\eta}\lambda_k = \gamma_{1k}y_1 + \ldots + \gamma_{k-1,k}y_{k-1} + \nu_k, k=1,\ldots,m$$

and we replace $\underset{\sim}{\Theta}$ by $\underset{\sim}{\lambda}\underset{\sim}{\lambda}'$ and replace $\underset{\sim}{\Sigma}$ by $\tau\underset{\sim}{\lambda}\underset{\sim}{\lambda}' + \underset{\sim}{U}$ (note we are using $\Gamma_{ij} = -\gamma_{ij}$ for $i \neq j$). Then the joint maximum for $\underset{\sim}{\Gamma}$, $\underset{\sim}{\eta}$, $\underset{\sim}{\lambda}$, τ and $\underset{\sim}{U}$ can be obtained by iterating on the ML equations for $\underset{\sim}{\Gamma}$ (given $\underset{\sim}{\eta}$, $\underset{\sim}{\lambda}$, τ, and $\underset{\sim}{U}$) and the ML equations for $\underset{\sim}{\eta}$, $\underset{\sim}{\lambda}$, τ, $\underset{\sim}{U}$ (given $\underset{\sim}{\Gamma}$). Some other methods for dealing with Model 4 (in particular the treatment of $\underset{\sim}{\Gamma}$) are discussed in the Appendix.

IV. Interpretation

First we will review some properties of the single equation variance components model. It was introduced into the econometric literature by Balestra and Nerlove (1966):

$$\text{(IV.1)} \quad y_{ij} = \alpha + \underline{x}'_{ij}\underline{\beta} + f_i + \nu_{ij} \ , \qquad \begin{array}{l} i = 1,\ldots,q \\ j = 1,\ldots,p \ , \end{array}$$

where the f_i are a random sample from a distribution with mean zero and variance σ_f^2 and the ν_{ij} are independently distributed across groups (i) and within groups (j) with mean zero and variance σ^2. There is an alternative "fixed effects" model (e.g., Kuh (1959)):

$$\text{(IV.2)} \quad y_{ij} = \alpha_i + \underline{x}'_{ij}\underline{\beta} + \nu_{ij}$$

which allows a separate intercept for each group. The interpretation and relative merits of these two models have given rise to some confusion. My preference is to regard (IV.2) as the "true" model and arrive at (IV.1) by adding uncertain prior information. Then the choice between the two models will rest on the persuasiveness of the prior.

In (IV.1) the problem is non-spherical disturbances and $\underline{\beta}$ is estimated by generalized least squares (GLS). This can be simplified to:

$$\text{(IV.3)} \quad \underline{\beta}^* = (\underline{W}_x + \zeta\underline{B}_x)^{-1}(\underline{W}_x\underline{b}_w + \zeta\underline{B}_x\underline{b}_B)$$

(Maddala (1971)), where $\underline{b}_w = \underline{W}_x^{-1}\underline{W}_{xy}$, $\underline{b}_B = \underline{B}_x^{-1}\underline{B}_{xy}$; $\underline{W}_x$, $\underline{W}_{xy}$, etc. are defined in (III.4); and $\zeta = 1/(1 + p\psi)$ with $\psi = \sigma_f^2/\sigma^2$.

So $\underline{\beta}^*$ is a matrix weighted average of the within and between group least squares estimators. This is the natural way to pool two independent estimators; $\underline{W}_x$ and $\underline{B}_x$ are proportional to the precision matrices.

Using analysis of covariance identities we also have

$$\text{(IV.4)} \quad \underline{\beta}^* = (\underline{W}_x + \frac{\zeta}{1-\zeta}\,\underline{T}_x)^{-1}(\underline{W}_x\underline{b}_w + \frac{\zeta}{1-\zeta}\,\underline{T}_x\underline{b}_T)$$

where $\underline{b}_T = \underline{T}_x^{-1}\underline{T}_{xy}$. Note that here $\underline{b}_w$ and $\underline{b}_T$ are attainable endpoints corresponding to the variance ratio $\psi = \sigma_f^2/\sigma^2$ taking on its extreme values of infinity and zero. In (IV.3) $\underline{b}_B$ is never reached since $0 \leq \zeta \leq 1$; i.e., the between group least squares estimator does not have a life of its own. It is, however, a very useful estimator in the presence of measurement error. One of our objectives is to give it an independent role.

The estimator in (IV.3) and (IV.4) can be obtained from the fixed effects model (IV.2) by adding an exchangeable prior for the α_i. The prior is exchangeable if its form is unaffected by permuting the α's, so that the i subscripts are just a labeling device with no substantive content. Then the prior must be a mixture of independent and identical distributions (de Finetti (1964), Hewitt and Savage (1955)). Assuming normality we have

$$\text{(IV.5)} \quad \alpha_i \sim \text{i.i.d. } N(\overline{\alpha}, \sigma_\alpha^2)$$

where $\overline{\alpha}$ and σ_α^2 are called hyperparameters (Good (1965)); their prior distribution generates the mixture. With a "flat" prior for $\overline{\alpha}$ and conditional on $\psi = \sigma_f^2/\sigma^2$, the posterior mean for β

is the GLS estimator in (IV.3). So ψ measures the strength of our prior belief that the α_i are all equal. As ψ varies from zero (certainty) to infinity (diffuse), the posterior mean goes from $\underline{b}_T$ to $\underline{b}_W$.

The natural proxy for the unobservable group effects is the posterior mean for the α_i. Conditional on $\underline{\beta}$ and the variance ratio ψ, the posterior mean is

$$\text{(IV.6)} \quad \alpha_i^* = \frac{p\psi}{1+p\psi}\,\hat{\alpha}_i + \frac{1}{1+ p\psi}\,\hat{\alpha}$$

where $\hat{\alpha}_i = \frac{1}{p}\sum_{j=1}^{p} (y_{ij} - \underline{x}'_{ij}\underline{\beta}) = \bar{y}_i - \bar{\underline{x}}'_i\,\underline{\beta}$

and $\hat{\alpha} = \bar{y} - \bar{\underline{x}}'\underline{\beta} = \frac{1}{q}\sum_{i=1}^{q} \hat{\alpha}_i$. So we take the fixed effects $\hat{\alpha}_i$ obtained by forcing the hyperplane through the group means, and shrink them towards the pooled OLS estimator $\hat{\alpha}$. Note that the shrinkage factor approaches zero if there are a large number of observations on each group or if there is a strong group structure.

Models 1 and 2: We can interpret the eigenvalue problem in (III.1) as a canonical correlation analysis of the residuals $\underline{E}$ and the set of group indicator dummy variables $\underline{J}$. We find the linear combination of residuals from the m equations that is most highly correlated with the group structure, subject to the restriction of being uncorrelated with the first index, and so on. The eigenvectors are the canonical weights for constructing these indices and the squared canonical correlations are the eigenvalues.

Our estimator of $\underline{\Theta}$ can be interpreted as solving a minimum norm approximation problem. For it is easy to show that conditional on $\underline{B}$ the corrected between group moments, $\underline{\overline{R}}_c = \frac{p}{p-1}(\underline{\overline{R}} - \frac{1}{p}\underline{R})$, give an unbiased estimate of $\underline{\Theta}$:

$$\text{(IV.7)} \qquad \underline{\Theta} = E(\underline{\overline{R}}_c) = E[\frac{p}{p-1} (\underline{\overline{R}} - \frac{1}{p}\underline{R})].$$

So it is reasonable to estimate $\underline{\Theta}$ by finding a matrix of rank N such that

$$\text{(IV.8)} \qquad || \underline{\overline{R}}_c - \underline{\Theta} ||_Q$$

is a minimum, where $|| \; ||_Q$ denotes the matrix norm in the metric of Q:

$$||\underline{A}||_Q = \text{tr.}\,\underline{QAQA}.$$

A natural choice of metric is $\underline{R}^{-1}$. For then the equations with poor fit are given less weight in the approximation error (and more generally the linear combinations of equations with poor fit are given less weight). It is easy to show that the $\underline{\Theta}$ in (III.2) solves this problem.

Another interpretation of our estimator of $\underline{\Theta}$ can be based on constructing proxies for the unobservables and using them in a regression. For the one factor model the fixed effects analog to $\hat{\alpha}_i$ in (IV.6) is

$$\text{(IV.9)} \qquad \hat{f}_i = (\underline{\overline{y}}_i' - \underline{\overline{x}}_i' \, \underline{\Pi})\underline{g} = \underline{\overline{E}}_i\underline{g}$$

where $\underline{R}\; \underline{g} = \rho\underline{\overline{R}}\; \underline{g}$.

The scale of $\hat{f}_i$ is arbitrary since it can't be separated from the scale of the coefficients. We resolve this by setting $\sigma_f^2 = 1$. Then $\underline{\Theta} = \underline{d}\,\underline{d}'$ and $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$ is a generalized variance ratio analogous to σ_f^2/σ^2. Then the posterior mean proxy corresponding to (IV.6) is

$$\text{(IV.10)} \qquad f_i^* = \frac{p\psi}{1+p\psi}\,\hat{f}_i + \frac{1}{1+p\psi}\,\hat{\bar{f}}$$

where

$$\hat{\bar{f}} = \frac{1}{q}\sum_{i=1}^{q}\hat{f}_i\ .$$

So again the exchangeable prior induces a shrinking towards an averaged estimator; the shrinking can be substantial if the number of observations in each group is small and if the signal-noise ratio $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$ is not too large.

We can interpret the ML estimator of $\underline{d}$ as regressing the residuals on the proxy $\underline{f}^*$:

$$\text{(IV.11)} \quad \underline{f}^* = \begin{pmatrix} f_1^* \\ \vdots \\ f_q^* \end{pmatrix} \otimes \underline{\ell}_\rho = \underline{J\bar{E}g}$$

$$\underline{d} \propto \underline{E}'\underline{f}^* \propto \underline{\bar{R}g}.$$

So $\underline{d}$ satisfies the dual of (III.1).

$$\text{(IV.12)} \quad \underline{\bar{R}g} = \rho\,\underline{Rg}$$

$$\underline{\bar{R}}^{-1}\underline{d} = \frac{1}{\rho}\,\underline{R}^{-1}\underline{d}.$$

If we scale so that $\underline{d}'\underline{\bar{R}}^{-1}\underline{d} = \frac{1}{p-1}(p-1/\rho)$, then $\underline{dd}'$ gives the (one factor) $\underline{\Theta}$ in (III.2).

Note that the regression on f_i^* is proportional to the regression on $\hat{f}_i$ provided the residuals sum to zero. If the equations include constant terms, then the GLS estimate of the hyperplane corresponding to each of the equations passes through the overall sample means. In that case we have the surprising result that the estimate of $\underline{d}$ is unaffected by adding the exchangeable prior to the fixed effects model. A similar result has been observed in the simpler factor model without the group structure. In that model Whittle (1953) found that his fixed effects estimator of the factor loadings agreed with the random effects ML algorithm devised by Lawley (1940) (also see the Uppsala Symposium (1953)).

In the fixed effects model the problem is to impose proportionality restrictions across the coefficients of the dummies in the different equations. The solution as a canonical correlation is given in Hauser and Goldberger (1971). Actually their model is a special case of Hannan's (1967) application of ML to a subsystem. A set of m equations form a subsystem if there are m-1 zero restrictions on each equation. Hannan showed that limited information maximum likelihood (LIML) applied to a subsystem can be reduced to a canonical correlation problem of the Hauser-Goldberger type. We can see that the fixed effects model fits Hannan's framework by rewriting

$$\underline{y}_k = \underline{X}\underline{\delta}_k + \underline{J}\underline{f}d_k + \underline{\varepsilon}_k , \quad k = 1,\dots,m$$

$$(\underline{J} = \underline{I}_q \otimes \underline{\ell}_p)$$

as

$$\text{(IV.13)} \quad \underline{y}_k = \underline{X}(\underline{\delta}_k - \underline{\delta}_m d_k) + \underline{y}_m d_k + (\underline{\varepsilon}_k - \underline{\varepsilon}_m d_k) \qquad k = 1,\dots,m-1$$

$$y_m = \underline{X\delta}_m + \underline{Jf} + \underline{\varepsilon}_m$$

(assuming that $d_m \neq 0$, we normalize so that $d_m = 1$). There are m-2 restrictions on each of the first m-1 equations since only one of the m-1 variables $y_1, \ldots y_{m-1}$ appears in any of these equations. Thus they form a subsystem. Furthermore the m^{th} equation is just identified which implies that LIML is in fact FIML.

In the multi-factor version we regress $\underline{E}$ on $\underline{J\bar{E}G}$ to obtain

$$\text{(IV.14)} \quad \underline{\tilde{D}} = \underline{\bar{R}GH}$$

where $\underline{H}$ is a diagonal scaling matrix. If the columns of $\underline{\tilde{D}}$ are properly scaled then $\underline{\tilde{D}\tilde{D}}'$ will give the $\underline{\Theta}$ in (III.2). But the decomposition of $\underline{\Theta}$ into $\underline{D\Phi D}'$ is not identified without further restrictions, nor are separate proxies for the different factors. We can only specify the space spanned by the factors (the column space of $\underline{J\bar{E}G}$).

The GLS estimator of $\underline{\delta}$ = vec ($\underline{B}$) is given in (III.4). It is a generalization of the single equation variance components pooling in (IV.3). Again we are taking a matrix weighted average of the within and between group estimators, weighting by their precision matrices. The correspondence with the single equation case is even closer when we compare the fixed and random effects estimators of $\underline{\delta}$. In the fixed effects case, we simply form the proxy for $\hat{f}_i$ in (IV.9) and regress $\underline{y}_k$ on $\underline{X}$ and this proxy to obtain $\hat{\underline{\delta}}_{fk}$, $k = 1, \ldots, m$. Some straightforward but tedious algebra will demonstrate the following relationship between the fixed and random effects estimators:

$$\text{(IV.15)} \qquad \underline{\delta}_k^* = (\underline{W}_x + \frac{\zeta}{1-\zeta}\underline{T}_x)^{-1}(\underline{W}_x\hat{\underline{\delta}}_{fk} + \frac{\zeta}{1-\zeta}\underline{T}_x\hat{\underline{\delta}}_{Tk}),\ k=1,\ldots,m,$$

where $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$, $\zeta = 1/(1 + p\psi)$, $\underline{W}_x$ and $\underline{T}_x$ are defined in (III.4), and $\underline{\delta}_{Tk} = \underline{T}_x^{-1}\underline{T}_{xy_k}$ is the pooled OLS estimator.

If there is only a single variable in $\underline{X}$ then (IV.15) is a simple weighted average and the random effects $\underline{\delta}^*$ is in between the fixed effects estimator and the pooled OLS estimator. But with several x's we have a matrix weighted average like (IV.4). In fact (IV.15) is identical to the single equation pooling formula (IV.4) with the fixed effects estimator replacing the unconstrained within group OLS $\underline{b}_w$, and with $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$ replacing the single equation variance ratio. So we can use $\hat{\underline{\delta}}_f$ to reduce the formula for $\underline{\delta}^*$ in (III.4) from a matrix weighted average that runs over equations and variables to one which just runs over the variables in $\underline{X}$, pooling each equation separately from the others.

Model 3: The ML estimator of $\underline{\eta}$ in (III.11) is

$$\underline{\eta} = (\underline{W}_x + \frac{\zeta}{1-\zeta}\underline{T}_x)^{-1}(\underline{W}_{xr} + \frac{\zeta}{1-\zeta}\underline{T}_{xr})$$

where $\zeta = 1/(1+p\psi)$ and $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$. This is identical to the single equation GLS estimator in (IV.4) if we replace σ_f^2/σ^2 by the generalized variance ratio $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$ and aggregate the y_k's into a single index $\underline{r} = \underline{Y}\underline{\Sigma}^{-1}\underline{d}/\psi$. Then $r_{ij} = \underline{x}_{ij}'\underline{\eta} + f_i + \varepsilon_{ij}$ is treated like a single equation components model with $\sigma_f^2/\sigma^2 = \psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$. The weights $\underline{\Sigma}^{-1}\underline{d}/\psi$ do not seem to come from a canonical correlation problem. But as $\psi \to \infty$ our estimator reduces to ML for the fixed effects model. There we do a canonical correlation analysis on $\underline{Y}$ and $(\underline{X}\ \underline{J})$ where $\underline{J} = \underline{I}_q \otimes \underline{\ell}_p$ is a set of dummy variables. The payoff from the exchangeable prior on the f_i is that unlike the fixed effects estimator, our estimator uses some of the between group variation in estimating $\underline{\eta}$.

As in (IV.11) our estimators of $\underline{d}$ and $\underline{\Sigma}$ can be interpreted as regression statistics in a model based on a proxy for the unobservable. Conditional on $\underline{\eta}$, $\underline{d}$, and $\underline{\Sigma}$, the posterior mean for f_i is

(IV.16) $$f_i^* = \frac{p\psi}{1+p\psi}\hat{f}_i + \frac{1}{1+p\psi}\hat{f}$$

where

$$\hat{f}_i = \underline{\bar{Y}}_i\underline{\Sigma}^{-1}\underline{d}/\psi - \underline{\bar{X}}_i\underline{\eta}, \qquad \hat{f} = \frac{1}{q}\sum_{i=1}^{q}\hat{f}_i .$$

This is analogous to (IV.10) except that now we take the canonical index of the averaged y's and subtract off the averaged contribution $\underline{\bar{X}}_i\underline{\eta}$ of the observed characteristics. The weights $p\psi/(1+p\psi)$ and $1/(1+p\psi$ are the same as (IV.10) with more shrinking if the groups are small and if the group effects have relatively small variance. Then the $\hat{f}_i$ are not very estimable individually and so we do more smoothing towards their average $\hat{f}$.

Now we can use initial estimates of $\underline{d}$ and $\underline{\Sigma}$ to form the composite proxy $\underline{X\eta}^* + f^*$, and then run the multivariate regression of $\underline{Y}$ on $X\underline{\eta}^* + f^*$ to obtain new estimates of $\underline{d}$ and $\underline{\Sigma}$. Then they can be used to reform $\underline{\eta}^*$ and $\underline{f}^*$ to repeat the process. This iterative scheme is actually a powering method for solving the eigenvalue problem in (III.9). Of course much faster techniques are available, but this helps our intuitive appreciation of the algorithm. The sequence of regressions will also reproduce our estimator of $\underline{\Sigma}$ in (III.10)

Model 4: The ML estimator of $\underline{\eta}$ is the same as in Model 3 except that now $\underline{r} = \underline{YU}^{-1}\underline{\lambda}/\mu$ and $\zeta = 1/[1 + p(\frac{\mu}{1+\tau\mu})]$ with $\mu = \underline{\lambda}'\underline{U}^{-1}\underline{\lambda}$. To see that $\mu/(1+\tau\mu)$ is the appropriate variance ratio, we consider the single equation version with $(f_i + g_{ij})\lambda_k + \varepsilon_{ijk}$ replaced by $f_i + g_{ij} + \varepsilon_{ij}$. Then $\mu = \sigma_f^2/\sigma^2$ and

$$\frac{\mu}{1+\tau\mu} = \frac{\sigma_f^2/\sigma^2}{1+\sigma_g^2/\sigma^2} = \frac{\sigma_f^2}{\sigma_g^2 + \sigma^2}$$

is the appropriate ratio of between group variance to within group variance.

The posterior mean proxy for f is

$$\text{(IV.17)} \qquad f_i^* = (1-\zeta)\hat{f}_i + \zeta\hat{f}$$

where $\hat{f}_i = \underline{\overline{Y}}_i\underline{U}^{-1}\underline{\lambda}/\mu - \underline{\overline{X}}_i\underline{\eta}$, $\hat{f} = \frac{1}{q}\sum_{i=1}^{q}\hat{f}_i$ and $\zeta = 1/[1 + p(\frac{\mu}{1+\tau\mu})]$. We have already seen that $\mu/(1+\tau\mu)$ is the appropriate ratio of between to within group variance for this problem. Note that we do less shrinking if $\sigma_a^2 = \sigma_f^2(1+\tau)$ is large, but for a given σ_a^2 we shrink to the mean more fcrcefully as $\tau = \sigma_g^2/\sigma_f^2$ increases. The proxy for the within family deviations $\underline{g}_i' = (g_{i1}\cdots g_{ip})$ is

$$\text{(IV.18)} \qquad \underline{g}_i^* = \frac{\tau\mu}{\tau\mu+1} [(\underline{Y}_i \underline{U}^{-1} \underline{\lambda}/\mu - \underline{X}_i \underline{\eta}) - f_i^* \underline{\ell}_p].$$

So given the canonical index of the y's, we subtract off the effects of the observed characteristics, $\underline{X\eta}$, and we also subtract off the unobserved family effects f*. The shrinkage factor is analagous to $p\psi/(1+p\psi)$ in (IV.6) because here p=1, only a single individual per group, and $\tau = \sigma_g^2/\sigma_f^2$ converts $\mu = \underline{\lambda}'\underline{U}^{-1}\underline{\lambda}$ from a family variance ratio (recall $\sigma_f^2 = 1$) to an individual variance ratio which would be σ_g^2/σ^2 in the single equation case.

V. The Causes and Consequences of Permanent Income

An example of Model 1 is a system of Engel curves based on components of permanent income. The model develops Friedman's observation that the horizon relevant for forming income expectations depends on the variability of the income series. Thus self-employed businessmen and wage earners form their expectations in different ways. This can be formalized by observing that optimal (e.g., minimum mean square error) forecasts of a stochastic process depend on the underlying autoregressive structure of the process. So when we can identify separate income streams for the same individual, it is a natural step to treat them separately in forming permanent income proxies. Holbrook and Stafford (1971) estimated this sort of model from a three year panel of consumers. For the time being I will specialize their model by assuming that the different components of permanent income are constant over the three years.

Then extending the model to several consumption goods gives

$$\text{(V.1)} \qquad C_{itk} = \sum_{h=1}^{N} \lambda_{hk} \tilde{Y}_{ih} + u_{itk}$$

$$Y_{ith} = \tilde{Y}_{ih} + v_{ith} \quad , \qquad \begin{array}{l} i = 1,\ldots,q \\ t = 1,\ldots,T \end{array}$$

We assume that the permanent components of income $\tilde{Y}_h$ (corresponding to the f_h) are independently distributed across individuals as a multivariate $N(0,\underline{\Phi})$. The observed income component Y_{ith} is assumed to be an unbiased estimate of the permanent component. The transitory components of consumption (u_{itk}) and income (v_{itk})

are assumed to be serially uncorrelated but freely contemporaneously correlated both within and across consumption and income categories. Allowing for non-zero correlation between transitory consumption and income is important if we only have independent observations on income and savings. For then consumption is generated as a residual and errors in income reporting will be transmitted to the consumption data and will induce a correlation between the transitory components.

Let $\underline{L} = (\lambda_{hk})$ be the matrix of marginal propensities (or elasticities in the logarithmic version). Then we have $D' = (\underline{L}\ \underline{I})$ and

$$\text{(V.2)} \qquad \underline{\Theta} = \underline{D\Phi D'} = \begin{bmatrix} \underline{L}'\underline{\Phi}\underline{L} & \underline{L}'\underline{\Phi} \\ \underline{\Phi}\underline{L} & \underline{\Phi} \end{bmatrix}$$

Thus $\underline{L}$ and $\underline{\Phi}$ can be recovered from $\underline{\Theta}$ and if $\underline{\Theta}$ has been constrained to have rank = N the relationship is uniquely given by

$$\text{(V.3)} \qquad \underline{L} = \underline{\Theta}_{22}^{-1}\,\underline{\Theta}_{21}$$

$$\underline{\Phi} = \underline{\Theta}_{22}\ .$$

Given our interpretation of $\underline{\Theta}$ as a rank N approximation to $\bar{R}_c = \frac{p}{p-1}\ (\underline{\bar{R}} - \frac{1}{p}\underline{R})$,we can interpret our estimator $\underline{L} = \underline{\Theta}_{22}^{-1}\underline{\Theta}_{21}$ as a set of corrected and smoothed between group regressions. Simply regressing on time averages would give $\underline{\bar{R}}_{22}^{-1}\underline{\bar{R}}_{21}$. Our estimator differs from this in two ways. First we correct $\underline{\bar{R}}$ for incomplete averaging of the transitory effects by subtracting off $\frac{1}{T}\ \underline{R}$. This correction would be negligible for a long time series or if the grouping were done by cities, but it could be crucial for a three year panel. Then the corrected $\underline{\bar{R}}$, i.e., $\underline{\bar{R}}_c$, is approximated (smoothed) by a matrix of lower rank. This conforms to Friedman's symmetric view of the problem; for once we have $\underline{\bar{R}}_c$ we can either regress Y on C or run C on Y and take the reciprocal. Imposing the rank constraint guarantees that we get the same answer either way.

The adding up property $\underline{R} = \underline{\Theta} + \underline{\Sigma}$ that is used to estimate $\underline{\Sigma}$ gives a decomposition of the total variance $\underline{R}$ into permanent ($\underline{\Theta}$) and transitory ($\underline{\Sigma}$) components.

In the general multi-factor model it is not possible to assign separate proxies to the different factors. This corresponds to our inability to separate $\underline{D}$ and $\underline{\Phi}$ in $\underline{\Theta} = \underline{D\Phi D'}$. But in this example there are enough restrictions. The restrictions are that the multiple regression of y_k on the proxies $\tilde{y}_1, \ldots, \tilde{y}_N$ should give zero coefficients except for the coefficient of y_k which should be one:

$$\text{(V.4)} \quad b_{y_k \tilde{y}_h \cdot \tilde{y}_1 \cdots \tilde{y}_N} = \begin{cases} 1 \text{ if } k=h \\ 0 \text{ otherwise .} \end{cases}$$

This in turn implies that

$$\text{(V.5)} \quad b_{y_k \tilde{y}_h \cdot \tilde{y}_k} = 0 \text{ if } k \neq h.$$

(V.5) is a natural condition for an efficient proxy. For if the partial correlation were not zero, we could exploit it to improve our specification of $\tilde{y}_k$. The formula for the proxies is

$$\text{(V.6)} \quad (\tilde{y}_1 \ \ldots \ \tilde{y}_N) = \underline{J\bar{E}G\Phi}^{1/2} \quad .$$

In the random effects model we would shrink towards the mean as in (IV.10).

Holbrook and Stafford relaxed the constancy of permanent income by using a set of exogenously given growth rates. They grouped people on the basis of observed characteristics (occupation, sex, race, education, age) and assigned growth rates from national averages. In our framework this would be $\tilde{Y}_{it} = a_{it}\tilde{Y}_i$ where the a_{it} are growth rates subject to an arbitrary normalization. With $a_{it} = 1$, we just have to estimate $\tilde{Y}_i$, the individual's permanent income in year one. The generalization of our algorithm is straightforward. We use weighted averages to form $\bar{R}$, weighting by growth rates. The extension is similar to the unbalanced sample algorithm in that $\psi = \underline{d}'\underline{\Sigma}^{-1}\underline{d}$ affects the weighting scheme. So we end up with a concentrated likelihood function that just depends on ψ. Details are given in the Appendix.

Up to this point we have been modeling the consequences of permanent income in terms of its effect on observed consumer behavior. Now we will construct a Model 3 example by looking at the causes of permanent income. A common suggestion is to construct a proxy based on individual characteristics such as age, education, race, etc. This would extend our model (with one type of income) to

$$\text{(V.7)} \quad \begin{aligned} C_{itk} &= \lambda_k \tilde{Y}_{it} + u_{itk} \\ Y_{it} &= \tilde{Y}_{it} + v_{it} \\ \tilde{Y}_{it} &= \underline{x}'_{it} \underline{\eta} + f_i \end{aligned}$$

where $\underline{x}$ is a set of observed characteristics and f_i picks up omitted characteristics that do not vary over the sample period. So we are specifying a richer prior for permanent income.

In the permanent income proxy

$$\tilde{Y} = \underline{x}'\underline{\eta} + f \ ,$$

$\underline{x}$ would include characteristics that are known both to the individual and to the econometrician. They are causal variables used to project future income. But there are additional variables known to the individual, e.g., various dimensions of ability, which are unobservable to the external observor (witness the poor explanatory power of cross sectional income generating functions). These make up f_i and have to be inferred by observing their consequences; i.e., using average consumption (in addition to average income) as a proxy for permanent income. Using value of home as a proxy has this flavor as does Liviatan's (1963) suggestion to use past and future consumption as instruments for measured income.

Finally, we take the $\hat{f}_i$'s in (IV.16) and pull them towards their mean. This is like using average community income as an indicator of an individual's permanent income and forms the basis of Friedman's (1957) reinterpretation of Duesenberry's (1949) relative income hypothesis.

A more careful look at our interpretation of $\underline{\eta}$ brings us to Models 2 and 4. The problem is that some of the observed characteristics, such as schooling (S), may be correlated with the unobserved characteristics (f), e.g., "ability." So we have the simultaneity problem captured by Models 2 and 4.

To be specific, consider estimating the returns to schooling in the presence of an unobserved ability variable:

$$\text{(V.8)}\qquad \begin{aligned} S_i &= \lambda_1 f_i + w_i \\ \tilde{Y}_i &= \gamma S_i + f_i \\ Y_{it} &= \tilde{Y}_i + v_{it} \\ C_{it} &= \lambda_2 \tilde{Y}_i + u_{it}, \qquad \begin{matrix} i = 1,\ldots,q \\ t = 1,\ldots,T \end{matrix} \quad . \end{aligned}$$

C and Y are the logs of consumption and income and our interest centers on γ, the rate of return. This could fit our model 2 framework except that there is no within group variation on S and so $\underline{V}$ is not positive definite. In fact $\underline{\Sigma}$ does not depend on γ and the identification must be based entirely on $\underline{\Theta}$. So the independence of the transitory u and v is irrelevant for the identification of γ.

The between group information is

$$\underline{\Theta} = \underline{dd}' + \underline{hh}' = \underline{DD}'$$

with

$$\underline{d} = \begin{bmatrix} \lambda_1 \\ 1+\lambda_1\gamma \\ \lambda_2(1+\lambda_1\gamma) \end{bmatrix} \sigma_f, \quad \underline{h} = \begin{bmatrix} 1 \\ \gamma \\ \lambda_2\gamma \end{bmatrix} \sigma_w$$

For any $\underline{\overline{D}}$ such that $\underline{\overline{DD}}' = \underline{\Theta}$ there is a rotation $\underline{P}$ such that $\underline{D} = \underline{\overline{D}P}$, $\underline{P}'\underline{P} = \underline{I}$. So set

$$\underline{P} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

and try to obtain α from the restriction that $d_3/d_2 = \lambda_2 = h_3/h_2$. This gives

$$\text{(V.9)} \quad (\overline{d}_3\cos\alpha - \overline{h}_3\sin\alpha)/(\overline{d}_2\cos\alpha - \overline{h}_2\sin\alpha)$$

$$= (\overline{d}_3\sin\alpha + \overline{h}_3\cos\alpha)/(\overline{d}_2\sin\alpha + \overline{h}_2\cos\alpha).$$

Unfortunately (V.9) reduces to $\overline{d}_3/\overline{d}_2 = \overline{h}_3/\overline{h}_2$ independently of α. It results in a reduced form restriction without shedding any light on the rotation angle. We should note, however, that λ_2 is identified. Just reinterpret f to be that part of "ability" that is uncorrelated with S. The problem is that then γ looses its structural interpretation.

The basic difficulty is that we cannot separate w_i from f_i. A solution is to have an indicator that intervenes between f and S, e.g., an early test score: $T_i = \lambda_1 f_i + e_i$ (an adult score would have the disadvantage of being dependent on S). This is a powerful piece of information; now we can identify γ with just a replicated income series:

$$\text{(V.10)}\qquad \begin{aligned} T_i &= \lambda_1 f_i + e_i \\ S_i &= \lambda_2 f_i + w_i \\ \tilde{Y}_i &= \gamma\, S_i + f_i \\ Y_{it} &= \tilde{Y}_i + v_{it} \quad . \end{aligned}$$

Assuming that v_i and e_i are independent, we have

$$\underline{\theta} = \underline{dd}' + \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_w^2 & \gamma\sigma_w^2 \\ 0 & \gamma\sigma_w^2 & \gamma^2\sigma_w^2 \end{bmatrix}$$

with $\underline{d}' = (\lambda_1 \quad \lambda_2 \quad 1 + \gamma\lambda_2)\sigma_f$.

So $\underline{\bar{d}} = \begin{pmatrix} \theta_{12} \\ \theta_{13} \end{pmatrix}$ gives us $\begin{pmatrix} d_2 \\ d_3 \end{pmatrix}$ up to scale factor $r = 1/d_1{}^2$:

$$\begin{bmatrix} \theta_{22} & \theta_{23} \\ \theta_{32} & \theta_{33} \end{bmatrix} = r\underline{\bar{d}\bar{d}}' + \begin{bmatrix} 1 & \gamma \\ \gamma & \gamma^2 \end{bmatrix} \sigma_w^2 .$$

We can solve for r by equating the different estimators of γ:

$$\text{(V.11)}\qquad (\theta_{33} - r\bar{d}_3^2)/(\theta_{23} - r\bar{d}_2\bar{d}_3) = \gamma$$

$$= (\theta_{23} - r\bar{d}_2\bar{d}_3)/(\theta_{22} - r\bar{d}_2^2)$$

giving

$$\text{(V.12)} \qquad r = (\theta_{22}\theta_{23} - \theta_{23}^2)/(\theta_{33}\bar{d}_2^2 + \theta_{22}\bar{d}_3^2 - 2\theta_{23}\bar{d}_2\bar{d}_3).$$

Then we estimate γ from either of the formulas in (V.11). This amounts to taking the between group covariance $\underline{\theta}$, subtracting off the effects of the common ability variable ($\underline{dd}'$), and computing either the regression of Y on S or the reciprocal of S on Y.

Our model 4 example is based on the Chamberlain-Griliches (1974) reanalysis of Gorseline's (1934) data on brothers. In their model the group is a family and an attempt is made to allow not just for omitted family effects but also for variation at the individual level. This is accomplished via a prior for the unobservable (a_{ij}) which invokes exchangeability at two levels, both within and across families:

$$\text{(V.13)} \qquad \begin{aligned} a_{ij} &= f_i + g_{ij}, \qquad \tau = \sigma_g^2/\sigma_f^2 \\ S_{ij} &= \lambda_1 a_{ij} + w_{ij} \\ \tilde{Y}_{ij} &= \gamma\ S_{ij} + a_{ij} \\ Y_{ij} &= \tilde{Y}_{ij} + v_{ij} \\ C_{ij} &= \lambda_2 \tilde{Y}_{ij} + u_{ij} \quad . \end{aligned}$$

So we are taking another pass at the model in (V.9). We will be more successful this time because a_{ij} and w_{ij} have different group structures: a_{ij} has a family component f_i but by assumption w_{ij} does not. So we can separate them by appropriate grouping without bringing in an intervening indicator between a_{ij} and S_{ij}. This gives us a more unified model, avoiding the (V.11) assumption that an IQ test and measured income are "parallel" measurements on the same underlying dimension. The cost of this unification is more stringent assumptions on the equation specific errors u, v, and w. For in (V.9 - V.13) we just used the between group $\underline{\theta}$; but here the within group $\underline{\Sigma}$ plays a crucial role, requiring independence assumptions for u_{ij}, v_{ij}, and w_{ij}.

Without the proportionality restriction across S and "a" in the C and Y equations we would have

$$\begin{aligned} S_{ij} &= \lambda_1 a_{ij} + w_{ij} \\ Y_{ij} &= \gamma_1 S_{ij} + \lambda_2 a_{ij} + v_{ij} \\ C_{ij} &= \gamma_2 S_{ij} + \lambda_3 a_{ij} + u_{ij} \end{aligned}$$

Then by Theorem 1 the exclusion of Y from the C equation is sufficient to identify the model provided $\lambda_2\lambda_3 \neq 0$ and $(\sigma_u^2, \sigma_v^2, \sigma_w^2) > 0$. Corollary 3 shows that the proportionality restriction $\gamma_1/\gamma_2 = \lambda_2/\lambda_3$ does not alter the identification condition. The (II.4) solution for $\tau = \sigma_g^2/\sigma_f^2$ is

$$\text{(V.14)} \quad \tau = (\sigma_{13}\sigma_{12} - \sigma_{23}\sigma_{11})/(\sigma_{13}d_1d_2 + \sigma_{12}d_1d_3 - \sigma_{23}d_1^2 - \sigma_{11}d_2d_3).$$

Then given τ we take the within family covariance Σ, subtract off the individual effects of the common ability variable ($\tau\underline{dd}'$), and then estimate the γ's by regressing Y and C on S, using these corrected within family moments.

Appendix A

Maximum Likelihood Estimation of the Reduced Form

The reduced form of our model is

$$(A.1) \quad y_{ijk} = \underset{\sim}{X}_{ij}\,\underset{\sim}{\delta}_k + \underset{\sim}{F}_i\underset{\sim}{d}_k + v_{ijk}$$

$$= \underset{\sim}{X}_{ij}\,\underset{\sim}{\delta}_k + \varepsilon_{ijk} \qquad \begin{array}{l} i = 1,\ldots,p \\ j = 1,\ldots,q \\ k = 1,\ldots,r \end{array}$$

where i indexes families or groups, j runs over individuals within a family, and k indexes the equations. We assume that the $n \leq r$ family factors $\underset{\sim}{F}_i = (f_{i1}\ldots f_{in})$ are distributed independently of v as a random sample (over families) from a multivariate $N(\underset{\sim}{0},\underset{\sim}{\Phi})$:

$$(A.2) \quad Ef_{ih}f_{i'h'} = \begin{cases} \Phi_{hh'}, \text{if } i = i' \\ 0 \text{ otherwise} \end{cases} .$$

The v's are assumed to be a random sample (over individuals) from a multivariate $N(0,\underset{\sim}{\Sigma})$:

$$(A.3) \quad Ev_{ijk}v_{i'j'k'} = \begin{cases} \sigma_{kk'} \text{ if } i=i' \text{ and } j=j' \\ 0 \text{ otherwise} \end{cases} .$$

Since the ε's corresponding to different families are independent, it is convenient to group the observations by family. Within a family we have the p individual observations on the first equation followed by p observations on the second equation, etc.:

(A.4)

$$\underset{\sim}{y}' = (y_{111},\ldots,y_{1p1},y_{112},\ldots,y_{1p2},\ldots, y_{q11},\ldots,y_{qpr}) .$$

Then letting $\underset{\sim}{y}_i$, $\underset{\sim}{X}_i$, and $\underset{\sim}{\varepsilon}_i$ denote the ith family blocks:

$$\underset{\sim}{y}_i = \begin{pmatrix} y_{i11} \\ \cdot \\ \cdot \\ \cdot \\ y_{ipr} \end{pmatrix} , \quad \underset{\sim}{X}_i = \begin{pmatrix} \underset{\sim}{x}_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{x}_{ip} \end{pmatrix} , \quad \underset{\sim}{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i11} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{ipr} \end{pmatrix}$$

we have

$$\underset{\sim}{y}_i = (\underset{\sim}{I}_r \otimes \underset{\sim}{X}_i)\underset{\sim}{\delta} + \underset{\sim}{\varepsilon}_i, \; i = 1,\ldots,q$$

or

$$\text{(A.5)} \qquad \underset{\sim}{y} = \underset{\sim}{Z}\underset{\sim}{\delta} + \underset{\sim}{\varepsilon}$$

where

$$\text{(A.6)} \qquad \underset{\sim}{Z} = (\underset{\sim}{\ell}_q \otimes \underset{\sim}{I}_r) \odot \underset{\sim}{X}$$

($\underset{\sim}{\ell}_q$ is a qx1 vector consisting entirely of ones; $\odot$ is the Khatri-Rao product:

$$\text{if } \underset{\sim}{A} = \begin{pmatrix} \underset{\sim}{A}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{A}_q \end{pmatrix} \quad \text{and } \underset{\sim}{B} = \begin{pmatrix} \underset{\sim}{B}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{B}_q \end{pmatrix} \quad \text{then } \underset{\sim}{A} \odot \underset{\sim}{B} = \begin{pmatrix} \underset{\sim}{A}_1 \otimes \underset{\sim}{B}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{A}_q \otimes \underset{\sim}{B}_q \end{pmatrix},$$

Rao and Mitra (1971)).

Let $\underset{\sim}{D} = (\underset{\sim}{d}_1 \ldots \underset{\sim}{d}_r)'$ be the coefficient matrix of the family effects and let $\underset{\sim}{\Theta} = \underset{\sim}{D}\underset{\sim}{\Phi}\underset{\sim}{D}'$. Then given our ordering of the data the disturbance covariance matrix is block diagonal:

$$\text{(A.7)} \qquad E\, \underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}' = \underset{\sim}{I}_q \otimes \underset{\sim}{\Omega}$$

with

$$\text{(A.8)} \qquad \underset{\sim}{\Omega} = \underset{\sim}{\Theta} \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' + \underset{\sim}{\Sigma} \otimes \underset{\sim}{I}_p .$$

So the log likelihood function is (apart from an irrelevant constant):

$$\text{(A.9)} \qquad L(\underset{\sim}{y} | \underset{\sim}{Z}, \underset{\sim}{\delta}, \underset{\sim}{\Theta}, \underset{\sim}{\Sigma}) =$$

$$- \frac{q}{2} \ln |\underset{\sim}{\Omega}| - \frac{1}{2} (\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta})' (\underset{\sim}{I}_q \otimes \underset{\sim}{\Omega}^{-1}) (\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta}) .$$

The first task is to simplify $\underset{\sim}{\Omega}^{-1}$ and $|\underset{\sim}{\Omega}|$. We let the columns of $\underset{\sim}{S} = (\underset{\sim}{s}_1 \ldots \underset{\sim}{s}_r)$ be a set of linearly independent common conjugate axes of $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$:

(A.10) $\underset{\sim}{S}'\underset{\sim}{\Sigma}\,\underset{\sim}{S} = \underset{\sim}{I},\ \underset{\sim}{S}'\underset{\sim}{\Theta}\,\underset{\sim}{S} = \{\Psi_1, \ldots, \Psi_n, \underset{\sim}{0}\}$

(brackets denote a diagonal or block diagonal matrix:

$$\{\Psi_1, \ldots, \Psi_n, \underset{\sim}{0}\} = \begin{pmatrix} \Psi_1 & & \underset{\sim}{0} \\ & \ddots \quad \Psi_n & \\ \underset{\sim}{0} & & \underset{\sim}{0} \end{pmatrix} \quad).$$

Similarly we choose $\underset{\sim}{T} = (\underset{\sim}{t}_1 \ldots \underset{\sim}{t}_p)$ so that

(A.11) $\underset{\sim}{T}'\underset{\sim}{T} = \underset{\sim}{I}_p, \underset{\sim}{T}'\underset{\sim}{\ell}\underset{\sim}{\ell}'\underset{\sim}{T} = \{p, \underset{\sim}{0}\}$

($\underset{\sim}{t}_1 = \underset{\sim}{\ell}_p / \sqrt{p}$). Now $\underset{\sim}{S} \otimes \underset{\sim}{T}$ can be used to diagonalize $\underset{\sim}{\Omega}$ and factor $\underset{\sim}{\Omega}^{-1}$:

(A.12) $(\underset{\sim}{S} \otimes \underset{\sim}{T})'\ \underset{\sim}{\Omega}(\underset{\sim}{S} \otimes \underset{\sim}{T}) =$

$$\underset{\sim}{I}_r \otimes \underset{\sim}{I}_p + \{\Psi_1, \ldots, \Psi_n, \underset{\sim}{0}\} \otimes \{p, \underset{\sim}{0}\}$$

$$\underset{\sim}{\Omega}^{-1} = (\underset{\sim}{S} \otimes \underset{\sim}{T})\{(1 + \Psi_1 p)^{-1}, 1, \ldots, (1 + \Psi_2 p)^{-1}, 1, \ldots, 1\}(\underset{\sim}{S} \otimes \underset{\sim}{T})'.$$

Let $m_h = 1 - (1 + \Psi_h p)^{-1}$, h=1,...,n so that

(A.13)

$$\underset{\sim}{\Omega}^{-1} = (\underset{\sim}{S} \otimes \underset{\sim}{T})(\underset{\sim}{I}_r \otimes \underset{\sim}{I}_p - \{m_1,\dots,m_n, \underset{\sim}{\bigcirc}\} \otimes \{1, \underset{\sim}{\bigcirc}\})(\underset{\sim}{S} \otimes \underset{\sim}{T})'$$

$$= \underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_p - \frac{1}{p}\underset{\sim}{S}\{m_1,\dots,m_n, \underset{\sim}{\bigcirc}\}\underset{\sim}{S}' \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' .$$

Then with $\underset{\sim}{c}_h = \sqrt{m_h/p}\ \underset{\sim}{s}_h$ and $\underset{\sim}{C} = (\underset{\sim}{c}_1 \dots \underset{\sim}{c}_n)$ we have the following decomposition of $\underset{\sim}{\Omega}^{-1}$:

(A.14) $$\underset{\sim}{\Omega}^{-1} = \underset{\sim}{\Sigma} \otimes \underset{\sim}{I}_p - \underset{\sim}{C}\underset{\sim}{C}' \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' .$$

The determinant of $\underset{\sim}{\Omega}$ can be obtained from (A.12):

$$|\underset{\sim}{S}\underset{\sim}{S}' \otimes \underset{\sim}{T}\underset{\sim}{T}'|\ |\underset{\sim}{\Omega}| = \prod_{h=1}^{n} (1 + p\Psi_h)$$

$$|\underset{\sim}{S}\underset{\sim}{S}' \otimes \underset{\sim}{T}\underset{\sim}{T}'| = |\underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_p| = |\underset{\sim}{\Sigma}|^{-p}$$

and so

(A.15)

$$|\underset{\sim}{\Omega}| = |\underset{\sim}{\Sigma}|^{p} \prod_{h=1}^{n} (1 + p\Psi_h).$$

This can be expressed in terms of $\underset{\sim}{C}$ and $\underset{\sim}{\Sigma}$ by letting

$$\underset{\sim}{M} = \begin{pmatrix} \underset{\sim}{M}_1 \\ \underset{\sim}{\bigcirc} \end{pmatrix} \quad , \underset{\sim}{M}_1 = \{\sqrt{m_1/p} \;,\ldots,\; \sqrt{m_n/p}\}$$

so that

$$\text{(A.16)} \quad \underset{\sim}{C} = \underset{\sim}{S}\underset{\sim}{M}$$

$$\underset{\sim}{C}'\underset{\sim}{\Sigma}\ \underset{\sim}{C} = \underset{\sim}{M}'\underset{\sim}{S}'\underset{\sim}{\Sigma}\ \underset{\sim}{S}\ \underset{\sim}{M} = \{m_1/p,\ldots,\ m_n/p\}$$

$$=\{\Psi_1/(1+p\Psi_1),\ldots,\Psi_n/(1+p\Psi_n)\}.$$

Then we have

$$\text{(A.17)} \quad |\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\ \underset{\sim}{C}| = \prod_{h=1}^{n} (1 - m_h) = \prod_{h=1}^{n} (1 + p\Psi_h)^{-1}$$

and

$$\text{(A.18)} \quad |\underset{\sim}{\Omega}| = |\underset{\sim}{\Sigma}|^p\ |\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}|^{-1}.$$

It will also be useful to have an expression for $\underset{\sim}{\Theta}$ in terms of $\underset{\sim}{C}$ and $\underset{\sim}{\Sigma}$ (thus demonstrating that our reparameterization is one-to-one). For this we use (A.10):

$$\underset{\sim}{\Theta}\ \underset{\sim}{S} = \underset{\sim}{\Sigma}\ \underset{\sim}{S}\{\Psi_1,\ldots,\Psi_n,\ \underset{\sim}{\bigcirc}\ \}$$

$$\underset{\sim}{\Theta} = \underset{\sim}{\Sigma}\ \underset{\sim}{S}\{\Psi_1,\ldots,\Psi_n,\ \underset{\sim}{\bigcirc}\ \}\underset{\sim}{S}'\underset{\sim}{\Sigma}$$

$$= \underset{\sim}{\Sigma}\ \underset{\sim}{C}\{\ \frac{p\Psi_1}{m_1}\ ,\ldots,\ \frac{p\Psi_n}{m_n}\ \}\ \underset{\sim}{C}'\underset{\sim}{\Sigma},$$

and so

(A.19) $\underset{\sim}{\Theta} = \underset{\sim}{\Sigma}\underset{\sim}{C}\{1 + p\Psi_1,\ldots,1 + p\Psi_n\}\underset{\sim}{C}'\underset{\sim}{\Sigma}.$

The reparameterized log likelihood function reduces to

(A.20)

$$L(\underset{\sim}{y}|\underset{\sim}{Z},\underset{\sim}{\delta},\underset{\sim}{C},\underset{\sim}{\Sigma}) = -\frac{pq}{2}\,\ell n\,|\underset{\sim}{\Sigma}|$$

$$+\ \frac{q}{2}\,\ell n|\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}|$$

$$-\ \frac{1}{2}\,(\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta})'(\underset{\sim}{I}_q \otimes (\underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_p - \underset{\sim}{C}\underset{\sim}{C}' \otimes \underset{\sim}{\ell}_p\underset{\sim}{\ell}_p'))(\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta}).$$

The problem now is to simplify the exponent term. We let $\underset{\sim}{e} = \underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta}$ be the vector of reduced form residuals. Then with

$$\underset{\sim}{e}_{ik} = \begin{pmatrix} e_{i1k} \\ \cdot \\ \cdot \\ \cdot \\ e_{ipk} \end{pmatrix} \quad \text{and} \quad \underset{\sim}{e}_i = \begin{pmatrix} \underset{\sim}{e}_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{e}_{ir} \end{pmatrix}, \qquad \begin{matrix} i = 1,\ldots,q \\ \\ k = 1,\ldots,r, \end{matrix}$$

the first term in the exponent is:

(A.21)

$$\underset{\sim}{e}'(\underset{\sim}{I}_q \otimes (\underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_p))\underset{\sim}{e} = \sum_{k,k'=1}^{r}\ \sum_{i=1}^{q} \sigma^{kk'}\, e_{ik}'e_{ik'}$$

$$= pq \ \mathrm{tr} \ \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R},$$

where $\underset{\sim}{R}$ is the covariance matrix of the reduced form residuals:

$$\underset{\sim}{E} = \begin{pmatrix} \underset{\sim}{e}_{11} & & \underset{\sim}{e}_{1r} \\ \cdot & & \cdot \\ \cdot & \cdot \ \cdot \ \cdot & \cdot \\ \cdot & & \cdot \\ \underset{\sim}{e}_{q1} & & \underset{\sim}{e}_{qr} \end{pmatrix} , \qquad \underset{\sim}{R} = \frac{1}{pq} \underset{\sim}{E}'\underset{\sim}{E}.$$

The remaining term in the exponent is

$$\text{(A.22)} \quad \underset{\sim}{e}'(\underset{\sim}{I}_q \otimes (\underset{\sim}{C}\underset{\sim}{C}' \otimes \underset{\sim}{\ell}_p\underset{\sim}{\ell}_p'))\underset{\sim}{e}$$

$$= \sum_{h=1}^{n} \sum_{i=1}^{q} \underset{\sim}{e}_i' (\underset{\sim}{c}_h\underset{\sim}{c}_h' \otimes \underset{\sim}{\ell}_p\underset{\sim}{\ell}_p')\underset{\sim}{e}_i$$

$$= \sum_{h=1}^{n} \sum_{k,k'=1}^{r} \sum_{i=1}^{q} c_{hk}c_{hk'}\underset{\sim}{e}_{ik}' \underset{\sim}{\ell}_p\underset{\sim}{\ell}_p' \underset{\sim}{e}_{ik'}$$

$$= \sum_{h=1}^{n} \sum_{k,k'=1}^{r} \sum_{i=1}^{q} p^2 \ c_{hk}c_{hk'} \ \bar{e}_{ik}\bar{e}_{ik'}$$

$$= p^2 \ q \ \mathrm{tr} \ \underset{\sim}{C}' \ \underset{\sim}{\bar{R}} \ \underset{\sim}{C},$$

where $\underset{\sim}{\bar{R}}$ is the covariance matrix of the average residuals, averaged over each family:

$$\bar{e}_{ik} = \frac{1}{p} \underset{\sim}{\ell}_p' \underset{\sim}{e}_{ik} \qquad \underset{\sim}{\bar{E}} = \begin{pmatrix} \bar{e}_{11} & & \bar{e}_{1r} \\ \cdot & & \cdot \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ \cdot & & \cdot \\ \bar{e}_{q1} & & \bar{e}_{qr} \end{pmatrix} ,$$

and

$$\underset{\sim}{\bar{R}} = \frac{1}{q} \underset{\sim}{\bar{E}}' \underset{\sim}{\bar{E}} .$$

Thus our canonical form for the likelihood function is

(A.23) $$L(\underset{\sim}{y}|\underset{\sim}{Z},\underset{\sim}{\delta},\underset{\sim}{C},\underset{\sim}{\Sigma}) = -\frac{pq}{2} \ln|\underset{\sim}{\Sigma}|$$

$$+ \frac{q}{2} \ell n|\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}|$$

$$- \frac{pq}{2} \operatorname{tr} \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} + \frac{p^2q}{2} \operatorname{tr} \underset{\sim}{C}'\underset{\sim}{\bar{R}}\ \underset{\sim}{C}.$$

Now we are ready to differentiate L and solve for $\underset{\sim}{\Sigma}$ and $\underset{\sim}{C}$. Since $\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}$ is a diagonal matrix and

(A.24) $$\partial \underset{\sim}{c}_h'\underset{\sim}{\Sigma}\underset{\sim}{c}_h/\partial\underset{\sim}{\Sigma}^{-1} = -\underset{\sim}{\Sigma}\underset{\sim}{c}_h\underset{\sim}{c}_h'\underset{\sim}{\Sigma},$$

we have

(A.25) $$\partial \ell n|\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}|/\partial\underset{\sim}{\Sigma}^{-1} = \sum_{h=1}^{n} p\ \underset{\sim}{\Sigma}\underset{\sim}{c}_h\underset{\sim}{c}_h'\ \underset{\sim}{\Sigma}/(1-p\underset{\sim}{c}_h'\ \underset{\sim}{\Sigma}\underset{\sim}{c}_h)$$

$$= p\underset{\sim}{\Sigma}\underset{\sim}{C}\{1+p\Psi_1,\ldots,1+p\Psi_n\}\underset{\sim}{C}'\underset{\sim}{\Sigma}$$

$$= p\underset{\sim}{\Theta}\ (A.19).$$

So setting $\partial L/\partial \underset{\sim}{\Sigma}^{-1} = (\underset{\sim}{0})$ implies that

$$\text{(A.26)}\quad \underset{\sim}{\Sigma} = \underset{\sim}{R} - \underset{\sim}{\Theta}.$$

The first order conditions for $\underset{\sim}{c}_h$ are

$$\text{(A.27)}\quad \partial L/\partial \underset{\sim}{c}_h = -pq\underset{\sim}{\Sigma}\underset{\sim}{c}_h/(1-p\underset{\sim}{c}_h' \underset{\sim}{\Sigma}\, \underset{\sim}{c}_h) + p^2 q\, \underset{\sim}{\bar{R}}\underset{\sim}{c}_h$$

$$= (\underset{\sim}{0}),\ h=1,\ldots,n,$$

and so

$$\text{(A.28)}\quad \underset{\sim}{\bar{R}}\underset{\sim}{C} = \frac{1}{p}\underset{\sim}{\Sigma}\, \underset{\sim}{C}\{1+p\Psi_1,\ldots,1+p\Psi_n\}.$$

We can eliminate $\underset{\sim}{\Sigma}$ from this expression by using (A.26),(A.19), and (A.16):

$$\text{(A.29)}\quad \underset{\sim}{\Sigma}\underset{\sim}{C} = \underset{\sim}{R}\underset{\sim}{C} - \underset{\sim}{\Theta}\underset{\sim}{C}$$

$$= \underset{\sim}{R}\underset{\sim}{C} - \underset{\sim}{\Sigma}\underset{\sim}{C}\{1 + p\Psi_1,\ldots,1+p\Psi_n\}\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}$$

$$= \underset{\sim}{R}\underset{\sim}{C} - \underset{\sim}{\Sigma}\underset{\sim}{C}\{\Psi_1,\ldots,\Psi_n\},$$

and so

(A.30) $$\underset{\sim}{\Sigma}\underset{\sim}{C} = \underset{\sim}{R}\underset{\sim}{C}\{1/(1+\Psi_1),\ldots,1/(1+\Psi_n)\}.$$

Then with

(A.31) $$\rho_h = \frac{1}{p}(1+p\Psi_h)/(1+\Psi_h) \text{ and } \underset{\sim}{\Lambda} = \{\rho_1,\ldots,\rho_n\}$$

we substitute (A.30) into (A.28) to obtain

(A.32) $$\underset{\sim}{\bar{R}}\underset{\sim}{C} = \underset{\sim}{R}\underset{\sim}{C}\underset{\sim}{\Lambda}.$$

So the columns of $\underset{\sim}{C}$ are eigenvectors of $\underset{\sim}{\bar{R}}$ in the metric of $\underset{\sim}{R}$. The eigenvectors corresponding to the n largest roots should be chosen since we will show that L is an increasing function of the ρ_h. The scale of the $\underset{\sim}{c}_h$ can be obtained from (A.28) and (A.16):

(A.33) $$\underset{\sim}{C}'\underset{\sim}{\bar{R}}\underset{\sim}{C} = \frac{1}{p}\{\Psi_1,\ldots,\Psi_n\}$$

$$\underset{\sim}{c}_h'\underset{\sim}{\bar{R}}\underset{\sim}{c}_h = \Psi_h/p = \frac{1}{p^2}(p\rho_h - 1)/(1-\rho_h),\ h = 1,\ldots,n.$$

Finally we can use (A.19),(A.30) and (A.31) to derive the M.L. estimate of $\underset{\sim}{\Theta}$:

$$\text{(A.34)} \qquad \underset{\sim}{\Theta} = \underset{\sim}{\Sigma}\underset{\sim}{C}\{1+p\Psi_1,\ldots,1+p\Psi_n\}\underset{\sim}{C}'\underset{\sim}{\Sigma}$$

$$= \frac{p^2}{p-1}\underset{\sim}{R}\underset{\sim}{C}\ \{\rho_1(1-\rho_1),\ldots,\rho_n(1-\rho_n)\}\underset{\sim}{C}'\underset{\sim}{R}.$$

Tests of our model can be obtained by evaluating L at the maximizing values of the parameters. So we need to simplify the following four terms from (A.23): $^{1)}|\underset{\sim}{\Sigma}|$, $^{2)}|\underset{\sim}{I}_n - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}|$, $^{3)}\mathrm{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{R}$, and $^{4)}\mathrm{tr}\underset{\sim}{C}'\underset{\sim}{\overline{R}}\underset{\sim}{C}$.

$$1) \qquad |\underset{\sim}{\Sigma}| = |\underset{\sim}{R} - \underset{\sim}{\Theta}| = |\underset{\sim}{R}|\ |\underset{\sim}{I}_r - \underset{\sim}{R}^{-1}\underset{\sim}{\Theta}|\ .$$

Let $\underset{\sim}{H} = \underset{\sim}{C}\{1+p\Psi_1,\ldots,1+p\Psi_n\}\ \underset{\sim}{C}'\underset{\sim}{\Sigma}$ so that

$$\underset{\sim}{\Theta} = \underset{\sim}{\Sigma}\underset{\sim}{H} \quad \text{(A.19)}$$

$$\underset{\sim}{R}^{-1}\underset{\sim}{\Theta} = \underset{\sim}{R}^{-1}(\underset{\sim}{R} - \underset{\sim}{\Theta})\underset{\sim}{H} = (\underset{\sim}{I} - \underset{\sim}{R}^{-1}\ \underset{\sim}{\Theta})\underset{\sim}{H}.$$

Note that the (non-zero) roots of $\underset{\sim}{H}$ coincide with those of

$$\{1+p\Psi_1,\ldots,1+p\Psi_n\}\ \underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C} = \{\Psi_1,\ldots,\Psi_n\} \quad \text{(A.16)}\ .$$

Thus $\underset{\sim}{H}$ has non-negative roots, $\underset{\sim}{I} + \underset{\sim}{H}$ is non-singular, and

$$\underset{\sim}{R}^{-1}\underset{\sim}{\Theta} = \underset{\sim}{H}(\underset{\sim}{I}+\underset{\sim}{H})^{-1}.$$

Since the roots of $\underset{\sim}{H}$ are Ψ_h the (non-zero) roots of $\underset{\sim}{R}^{-1}\underset{\sim}{\Theta}$ are $\Psi_h/(1+\Psi_h) = p(1-\rho_h)/(p-1)$, and so

(A.35) $$|\underset{\sim}{\Sigma}| = |\underset{\sim}{R}| \prod_{h=1}^{n} p(1-\rho_h)/(p-1).$$

(A.36) 2) $$|\underset{\sim}{I} - p\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}| = \prod_{h=1}^{n} \left(1- \frac{p\Psi_h}{1+p\Psi_h}\right) = \prod_{h=1}^{n} \left(\frac{1}{\rho_h} -1\right)/(p-1)$$

(A.16,A.31).

(A.37) 3) $$\mathrm{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} = \mathrm{tr}\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\Sigma}+\underset{\sim}{\Theta}) = r + \sum_{h=1}^{n} \Psi_h \quad \text{(A.10)}.$$

(A.38) 4) $$\mathrm{tr}\underset{\sim}{C}'\bar{\underset{\sim}{R}}\underset{\wedge}{C} = \frac{1}{p}\,\mathrm{tr}\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}\{1+p\Psi_1,\dots,1+p\Psi_n\} \quad \text{(A.28)}$$

$$= \frac{1}{p} \sum_{h=1}^{n} \Psi_h \quad \text{(A.16)}.$$

So the exponent terms cancel in (A.23) and apart from an irrelevant constant

(A.39) $$L_n^* = -\frac{pq}{2}\{\ln|\underset{\sim}{R}| + \ln \prod_{h=1}^{n} p(1-\rho_h)/(p-1)\}$$

$$+ \frac{q}{2} \ln \prod_{h=1}^{n} \left(\frac{1}{\rho_h} -1\right)/(p-1).$$

A likelihood ratio test for n factors vs. n+1 factors can be based on the large sample χ^2 distribution of $-2(L_n^* - L_{n+1}^*)$. $\utilde{R}$ and $\utilde{\overline{R}}$ are computed using the M.L. estimate of $\utilde{\delta}$ (see (A.44)). To determine the degrees of freedom, note that constraining $\utilde{\theta}$ to have rank n lets us determine all the elements of $\utilde{\theta}$ from the first n columns. Since $\utilde{\theta}$ is symmetric there are $rn - \frac{n(n-1)}{2}$ free elements and so restricting the model to n factors instead of n+1 imposes

$$r(n+1) - \frac{(n+1)n}{2} - [rn - \frac{n(n-1)}{2}] = r-n$$

constraints. Thus

$$\text{(A.40)} \qquad -2(L_n^* - L_{n+1}^*) \sim \chi^2(r-n).$$

The M.L. estimator of $\utilde{\delta}$(given $\utilde{\Omega}$) is the GLS estimator

$$\utilde{\delta}^* = [\utilde{Z}'(\utilde{I}_q \otimes \utilde{\Omega}^{-1})\utilde{Z}]^{-1}\utilde{Z}'(\utilde{I}_q \otimes \utilde{\Omega}^{-1})\utilde{y}.$$

If the same X's appear in each equation then we have

$$\text{(A.41)} \qquad \utilde{Z}'(\utilde{I}_q \otimes \utilde{\Omega}^{-1})\utilde{Z} =$$

$$\sum_{i=1}^{q}(\utilde{I}_r \otimes \utilde{X}_i')(\utilde{\Sigma}^{-1} \otimes \utilde{I}_p - \utilde{C}\utilde{C}' \otimes \utilde{\ell}_p\utilde{\ell}_p')(\utilde{I}_r \otimes \utilde{X}_i)$$

$$= \utilde{\Sigma}^{-1} \otimes \utilde{T}_{xx} - p\utilde{C}\utilde{C}' \otimes \utilde{B}_{xx}$$

$$= \boldsymbol{\Sigma}^{-1} \otimes \mathbf{W}_{xx} + (\boldsymbol{\Sigma}^{-1} - p\mathbf{C}\mathbf{C}') \otimes \mathbf{B}_{xx}$$

where

$$\mathbf{T}_{xx} = \sum_{i=1}^{q} \mathbf{X}_i'\mathbf{X}_i, \quad \mathbf{B}_{xx} = \frac{1}{p} \sum_{i=1}^{q} \mathbf{X}_i'\boldsymbol{\ell}_p\boldsymbol{\ell}_p'\mathbf{X}_i$$

$$\mathbf{W}_{xx} = \mathbf{T}_{xx} - \mathbf{B}_{xx}$$

and we'll be using similar expressions for $\mathbf{W}_{xy_k}$ and $\mathbf{B}_{xy_k}$.

Then using

$$(\boldsymbol{\Sigma}^{-1} - p\mathbf{C}\mathbf{C}')^{-1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{C}(\mathbf{C}'\boldsymbol{\Sigma}\,\mathbf{C} - \frac{1}{p}\,\mathbf{I})^{-1}\mathbf{C}'\boldsymbol{\Sigma}$$

together with

$$\mathbf{C}'\boldsymbol{\Sigma}\,\mathbf{C} = \{\Psi_1/(1+p\Psi_1) \;,\ldots,\; \Psi n/(1+p\Psi_n)\} \qquad \text{(A.16)}$$

and

$$\boldsymbol{\theta} = \boldsymbol{\Sigma}\,\mathbf{C}\{\, 1+p\Psi_1,\ldots,\, 1+p\Psi_n\}\,\mathbf{C}'\boldsymbol{\Sigma} \qquad \text{(A.19)}$$

gives

(A.42) $\quad (\,\boldsymbol{\Sigma}^{-1} - p\mathbf{C}\mathbf{C}'\,)^{-1} = \;\boldsymbol{\Sigma} + p\,\boldsymbol{\theta}.$

The remaining term in $\boldsymbol{\delta}^*$ is

$$\mathbf{Z}'(\mathbf{I}_q \otimes \boldsymbol{\Omega}^{-1})\mathbf{y} = \sum_{i=1}^{q} (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}_i' - \mathbf{C}\mathbf{C}' \otimes \mathbf{X}_i'\boldsymbol{\ell}_p\boldsymbol{\ell}_p')\mathbf{y}_i .$$

The $\sum_i (\Sigma^{-1} \otimes X_i')y_i$ term can be partitioned into r blocks, each with as many rows as there are exogenous variables. The kth block is

$$\sum_{k'=1}^{r} \sigma^{kk'} T_{xy_{k'}} = \sum_{k'=1}^{r} \sigma^{kk'} T_{xx} \hat{\delta}_{Tk'}$$

where $\hat{\delta}_{Tk} = T_{xx}^{-1} T_{xy_k}$. Thus the whole term can be written as $(\Sigma^{-1} \otimes T_{xx}) \hat{\delta}_T$ with $\hat{\delta}_T' = (\hat{\delta}_{T1}', \ldots, \hat{\delta}_{Tr}')$.

Similarly the second term is $p(CC' \otimes B_{xx})\hat{\delta}_B$ with $\hat{\delta}_{Bk} = B_{xx}^{-1} B_{xy_k}$ and $\hat{\delta}_B' = (\hat{\delta}_{B1}', \ldots, \hat{\delta}_{Br}')$. Then using the identity

$$(I_r \otimes T_{xx}) \hat{\delta}_T = (I_r \otimes W_{xx})\hat{\delta}_W + (I_r \otimes B_{xx})\hat{\delta}_B$$

we have

$$\text{(A.43)} \quad Z'(I_q \otimes \Omega^{-1})y = (\Sigma^{-1} \otimes W_{xx})\hat{\delta}_W + [(\Sigma^{-1} - pCC') \otimes B_{xx}]\hat{\delta}_B .$$

Combining (A.41), (A.42), and (A.43) shows that the GLS estimator of δ pools the "within" and "between" OLS estimators, weighting by their precision matrices;

$$\text{(A.44)} \qquad \delta^* = (H_W + H_B)^{-1}(H_W\hat{\delta}_W + H_B\hat{\delta}_B)$$

where

$$H_W^{-1} = E(\hat{\delta}_W - \delta)(\hat{\delta}_W - \delta)' = \Sigma \otimes W_{xx}^{-1}$$

$$H_B^{-1} = E(\hat{\delta}_B - \delta)(\hat{\delta}_B - \delta)' = p(\theta + \frac{1}{p}\Sigma) \otimes B_{xx}^{-1} .$$

If the X's differ across equations then the i^{th} block of Z is no longer $I_r \otimes X_i$ but rather the block diagonal matrix $\{X_{i1},\dots,X_{ir}\}$ where

$$X_k = \begin{pmatrix} X_{1k} \\ \vdots \\ X_{qk} \end{pmatrix}$$ is the set of exogenous variables appearing

in the k^{th} equation. Now the k,k' block of the Σ part of $Z'(I_q \otimes \Omega^{-1})Z$ is not $\sigma^{kk'} T_{xx}$ but $\sigma^{kk'} T_{x_k x_{k'}}$ with $T_{x_k x_{k'}} =$ $= \sum_{i=1}^{q} X'_{ik} X_{ik'}$. This can be written as $\Sigma^{-1} * \bar{T}_{xx}$ where the k,k' block of $\bar{T}_{xx}$ is $T_{x_k x_{k'}}$ and $*$ is a generalized Hadamard product (Rao, 1973). Similarly the "between" term of $Z'(I_q \otimes \Omega^{-1})Z$ is $pCC' * \bar{B}_{xx}$. Then using the analysis of covariance identity $\bar{T}_{xx} = \bar{W}_{xx} + \bar{B}_{xx}$ we have

$$\text{(A.45)} \qquad Z'(I_q \otimes \Omega^{-1})Z = \Sigma^{-1} * \bar{W}_{xx} + (\Sigma^{-1} - pCC') * \bar{B}_{xx} .$$

The $Z'(I_q \otimes \Omega^{-1})y$ term can be partitioned into r blocks of which the k^{th} is

$$\sum_{k'=1}^{r} [\sigma^{kk'} \underset{\sim}{W}_{x_k y_{k'}} + (\sigma^{kk'} - p\,\mu_{kk'}) \underset{\sim}{B}_{x_k y_{k'}}]$$

where $\mu_{kk'}$ is the k,k' element of $\underset{\sim}{C}\underset{\sim}{C}'$. With unequal X's the M.L. estimate of $\underset{\sim}{\delta}$ based on just the within group deviations is not $\hat{\underset{\sim}{\delta}}_W$ but rather (conditional on $\underset{\sim}{\Sigma}$) the Zellner "seemingly unrelated" GLS estimator:

$$\underset{\sim}{\delta}_W^{GLS} = \begin{bmatrix} \sigma^{11} \underset{\sim}{W}_{x_1 x_1} & \cdots & \sigma^{1r} \underset{\sim}{W}_{x_1 x_r} \\ \vdots & & \vdots \\ \sigma^{r1} \underset{\sim}{W}_{x_r x_1} & \cdots & \sigma^{rr} \underset{\sim}{W}_{x_r x_r} \end{bmatrix} \begin{bmatrix} \sum_{k=1}^{r} \sigma^{1k} \underset{\sim}{W}_{x_1 y_k} \\ \vdots \\ \sum_{k=1}^{r} \sigma^{rk} \underset{\sim}{W}_{x_r y_k} \end{bmatrix} .$$

There is a similar estimator $\underset{\sim}{\delta}_B^{GLS}$ (using just the between group variation) which replaces $\underset{\sim}{\Sigma}^{-1}$ by $\frac{1}{p}(\underset{\sim}{\theta} + \frac{1}{p}\underset{\sim}{\Sigma})^{-1} = \underset{\sim}{\Sigma}^{-1} - p\underset{\sim}{C}\underset{\sim}{C}'$ (from (A.42)). Thus

$$\text{(A.46)} \quad \underset{\sim}{Z}'(\underset{\sim}{I}_q \otimes \underset{\sim}{\Omega}^{-1})\underset{\sim}{y} = (\underset{\sim}{\Sigma}^{-1} * \underset{\sim}{\overline{W}})\underset{\sim}{\delta}_W^{GLS} + \frac{1}{p}[(\underset{\sim}{\theta} + \frac{1}{p}\underset{\sim}{\Sigma})^{-1} * \underset{\sim}{\overline{B}}]\underset{\sim}{\delta}_B^{GLS} .$$

The M.L. estimate of $\underset{\sim}{\delta}$ (given $\underset{\sim}{\Omega}$) is a matrix weighted average of the within and between group GLS estimators, weighting by their precision matrices:

$$\text{(A.47)} \quad \underset{\sim}{\delta}^{*} = (\underset{\sim}{H}_W + \underset{\sim}{H}_B)^{-1}(\underset{\sim}{H}_W \underset{\sim}{\delta}_W^{GLS} + \underset{\sim}{H}_B \underset{\sim}{\delta}_B^{GLS})$$

with

$$H_w = [E(\delta_w^{GLS} - \delta)(\delta_w^{GLS} - \delta)']^{-1} = (\Sigma^{-1} * \overline{W})$$

$$H_B = [E(\delta_B^{GLS} - \delta)(\delta_B^{GLS} - \delta)']^{-1} = [\frac{1}{p}(\Theta + \frac{1}{p}\Sigma)^{-1} * \overline{B}] \quad .$$

The joint maximum for δ and Ω is obtained by iterating on the M.L. equations for Σ and C (given δ) and the M.L. equation for δ (given Σ and C).

The GLS procedure can be simplified by concentrating the intercepts out of the likelihood function. This is possible since the M.L. estimates of the hyperplanes corresponding to each of the equations pass through the sample means. For if g_k is an eigenvector of $\Sigma + p\,\theta$ then $\ell_q \otimes g_k \otimes \ell_p$ is an eigenvector of $I_q \otimes \Omega$. So as long as $\Sigma + p\,\theta$ has full rank, then r of the eigenvectors of $I_q \otimes \Omega$ span the column space of $\ell_q \otimes I_r \otimes \ell_p$ which spans the r intercept variables. Thus if we partition δ_k into the intercept δ_{1k} and the slope coefficients δ_{2k}, then conditional on δ_{2k} the GLS estimate of δ_{1k} is OLS (e.g., Rao and Mitra (1971), chap. 8):

$$\delta_{1k} = \overline{y}_k - \overline{X}_k \, \delta_{2k},$$

where $\overline{y}_k$ is the grand mean of y_k and $\overline{X}_k$ is the row vector of grand means for the exogenous variables (other than the intercept) in the k^{th} equation. So the δ_{1k} can be concentrated out of the likelihood function simply by replacinq each variable by its deviation from the overall sample mean and proceedinq without intercepts.

We conclude this Appendix by displaying the asymptotic information matrices for the one factor version of our models. Stacking the parameters into a vector, $\underset{\sim}{\xi}$, we let $\underset{\sim}{\Xi} = -q\operatorname*{plim}_{q\to\infty}(\frac{1}{q}\ \partial^2 L/\partial\xi\partial\xi')$. Then we can approximate the variance of the M.L. estimate of $\underset{\sim}{\xi}$ by $V(\underset{\sim}{\xi}) = \underset{\sim}{\Xi}^{-1}$. It is straightforward but rather tedious to show that

$$\text{(A.48)}\qquad \underset{\sim}{\Xi}_{\delta\delta'} \neq q \text{ plim } (\underset{\sim}{H}_W + \underset{\sim}{H}_B)/q \quad \text{as given in (A.47)}$$

$$\underset{\sim}{\Xi}_{dd'} = \eta_1\underset{\sim}{\Sigma}^{-1}+\eta_2\underset{\sim}{c}\underset{\sim}{c}'$$

$$\underset{\sim}{\Xi}_{\sigma\sigma'} = (pq/2)\underset{\sim}{J}'[(\underset{\sim}{\Sigma}^{-1}- \underset{\sim}{c}\underset{\sim}{c}')\otimes(\underset{\sim}{\Sigma}^{-1}- \underset{\sim}{c}\underset{\sim}{c}')+(p-1)\underset{\sim}{c}\underset{\sim}{c}' \otimes \underset{\sim}{c}\underset{\sim}{c}']\underset{\sim}{J}$$

$$\underset{\sim}{\Xi}_{\delta d'} = (\underset{\sim}{0})$$

$$\underset{\sim}{\Xi}_{\delta\sigma'} = (\underset{\sim}{0})$$

$$\underset{\sim}{\Xi}_{d\sigma'} = \eta_3[\underset{\sim}{c}' \otimes (\underset{\sim}{\Sigma}^{-1} - p\underset{\sim}{c}\underset{\sim}{c}')]\underset{\sim}{J}$$

where $\eta_1 = p^2q\Psi/(1+\Psi)$, $\eta_2 = p^2q(1-p\Psi)/(1+p\Psi)$, $\eta_3 = pq(1+p\Psi)^{-½}$.

The $r(r+1)/2$ distinct elements from the upper triangle of Σ are contained in $\underset{\sim}{\sigma}' = (\sigma_{11}\sigma_{12}\sigma_{22}\cdots\sigma_{1r}\cdots\sigma_{rr})$. Let $\underset{\sim}{\bar{\sigma}}$ be the r^2 by 1 vector obtained by stacking the columns of $\underset{\sim}{\Sigma}$: $\underset{\sim}{\bar{\sigma}} = \text{vec } \underset{\sim}{\Sigma}$. Then J is the r^2 by $r(r+1)/2$ matrix with $J_{ij} = 1$ if $\bar{\sigma}_i = \sigma_j$ and $J_{ij} = 0$ otherwise.

Since the M.L. estimate of $\underset{\sim}{\Sigma}$ is $\underset{\sim}{R}-\underset{\sim}{d}\underset{\sim}{d}'$ we can write the concentrated log likelihood $\tilde{L}$ in terms of $\underset{\sim}{\delta}$ and $\underset{\sim}{d}$. Then we have

(A.49) $$\underset{\sim}{\tilde{\Xi}}_{dd'} = -q\ \mathrm{plim}(\tfrac{1}{q}\ \partial^2L/\partial d\partial d') = \zeta_1\ \underset{\sim}{\Sigma}^{-1} + \zeta_2\ \underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}\underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}$$

with $$\zeta_1 = pq(p-1)\Psi(1+\Psi)/(1+p\Psi)$$

$$\zeta_2 = pq[(p^2-p-4)\Psi^2 - (p^2-4p+11)\Psi + (p-1)]/(1+p\Psi)^2.$$

The information matrix $\tilde{\Xi}$ is block diagonal in $\underset{\sim}{\delta}$ and $\underset{\sim}{d}$ and so

(A.50) $$V(\underset{\sim}{d}) = (\underset{\sim}{\tilde{\Xi}}_{dd'})^{-1} = \nu_1\underset{\sim}{\Sigma} + \nu_2\underset{\sim}{d}\underset{\sim}{d}'$$

with $$\nu_1 = -1/\zeta_1 \text{ and } \nu_2 = \zeta_2/(\zeta_1 + \zeta_1\zeta_2\Psi).$$

For the structural form of our model we replace $\underset{\sim}{d}$ by $\underset{\sim}{\gamma}$ and let $\underset{\sim}{\Sigma} = \tau\underset{\sim}{\gamma}\underset{\sim}{\gamma}' + \underset{\sim}{\Delta}$ where $\underset{\sim}{\Delta}$ is the diagonal matrix $\{\sigma^2_{u_1},\ldots,\sigma^2_{u_K}, \sigma^2_w\}$. Also let $\mu = \underset{\sim}{\gamma}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{\gamma}$, $\underset{\sim}{c} = \underset{\sim}{\Sigma}^{-1}\underset{\sim}{\gamma}/(1+p\mu)^{1/2}$, and reinterpret X_k to include $\underset{\sim}{y}_s$ if $k \leq K$. We partition $\underset{\sim}{\delta}_k = \begin{pmatrix}\alpha_k\\ \beta_k\end{pmatrix}$ for $k \leq K$ with $\delta_{k+1} = \alpha_s$, and also set up

$$\underset{\sim}{\alpha} = \begin{pmatrix}\underset{\sim}{\alpha}_1\\ \vdots\\ \underset{\sim}{\alpha}_K\\ \underset{\sim}{\alpha}_s\end{pmatrix} \qquad \underset{\sim}{\beta} = \begin{pmatrix}\beta_1\\ \vdots\\ \beta_K\end{pmatrix} \qquad \underset{\sim}{\gamma} = \begin{pmatrix}\gamma_1\\ \vdots\\ \gamma_K\end{pmatrix}, \text{ and } \underset{\sim}{\sigma} = \begin{pmatrix}\sigma^2_{u_1}\\ \vdots\\ \sigma^2_{u_K}\\ \sigma^2_w\end{pmatrix}.$$

Then we have

(A.51) $$\underset{\sim}{\Xi}_{\delta\delta'} = q\ \mathrm{plim}(\underset{\sim}{H}_W + \underset{\sim}{H}_B)/q \qquad \text{as in (A.47)}$$

$$\underset{\sim}{\Xi}_{\gamma\gamma'} = \kappa_1\underset{\sim}{\Sigma}^{-1} + \kappa_2\underset{\sim}{c}\underset{\sim}{c}'$$

$$\Xi_{\tau\tau} = \kappa_3$$

$$\underset{\sim}{\Xi}_{\sigma\sigma'} = \frac{pq}{2}\ \underset{\sim}{\Sigma}^{-1} * (\underset{\sim}{\Sigma}^{-1} - 2\underset{\sim}{c}\underset{\sim}{c}') + \frac{p^2q}{2}\ (\underset{\sim}{c}\underset{\sim}{c}')*(\underset{\sim}{c}\underset{\sim}{c}')$$

$$\underset{\sim}{\Xi}_{\gamma\tau} = \kappa_4 \underset{\sim}{c}$$

$$\underset{\sim}{\Xi}_{\gamma\sigma'} = \kappa_5 \underset{\sim}{\Sigma}^{-1}\{\underset{\sim}{c}\} + \kappa_6 \underset{\sim}{c}(\underset{\sim}{c}*\underset{\sim}{c})'$$

$$\underset{\sim}{\Xi}_{\tau\sigma'} = \kappa_7(\underset{\sim}{c}*\underset{\sim}{c})'$$

$$\underset{\sim}{\Xi}_{\alpha(\gamma'\tau\sigma')} = (\underset{\sim}{0})$$

$$\underset{\sim}{\Xi}_{\beta(\gamma'\tau\sigma')} = \gamma_s \underset{\sim}{\Xi}_{\underline{\gamma}(\gamma'\tau\sigma')}$$

where $\{\underset{\sim}{c}\}$ is the diagonal matrix with the elements of $\underset{\sim}{c}$ on the diagonal and

$$\kappa_1 = pq\mu\ [(p-1)\tau^2\mu + (\tau^2 + 2\tau + p)]/(1 + p\mu)$$

$$\kappa_2 = pq[p(p-1)\tau^2\mu^2 + ((2p-3)\tau^2 - 2p\tau - p^2)\mu + (\tau^2+2\tau+p)]/(1+p\mu)$$

$$\kappa_3 = \frac{pq}{2}\ \mu^2\ [(p-1)\mu(p\mu+2) + 1]/(1+p\mu)^2$$

$$\kappa_4 = pq\mu\ [p(p-1)\ \tau\mu^2 + 2(p-1)\tau\mu + (\tau+1)]/(1+p\mu)^{3/2}$$

$$\kappa_5 = pq[(p-1)\tau\mu + (\tau+1)]/(1+p\mu)^{1/2}$$

$$\kappa_6 = -pq(\tau+p)/(1+p\mu)^{1/2}$$

$$\kappa_7 = \frac{pq}{2}\ [(p-1)\mu\ (p\mu + 2) + 1]/(1+p\mu).$$

Appendix B

Unbalanced Groups

We will work with a single unobservable. The generalization to several factors is non-trivial. Let $\alpha = 1, \ldots, N$ index the different family sizes. There are p_α individuals in each of q_α families or groups. The total number of families is $q = \sum_{\alpha=1}^{N} q_\alpha$ and $\bar{p}q = \sum_{\alpha} p_\alpha q_\alpha$ is the total number of individuals in the sample. The observable f is scaled so that $\phi = \sigma_f^2 = 1$ and $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$. For families of a given size the observations are arranged as in (A.4). Then the families with p_1 members are followed by those with p_2 members, etc. With this arrangement of the data the covariance matrix of the reduced form disturbance vector $\underset{\sim}{\varepsilon}$ is the block diagonal matrix

$$\text{(B.1)} \qquad E\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}' = \{\underset{\sim}{I}_{q_1} \otimes \underset{\sim}{\Omega}_1, \ldots, \underset{\sim}{I}_{q_N} \otimes \underset{\sim}{\Omega}_N\}$$

$$\underset{\sim}{\Omega}_\alpha = \underset{\sim}{d}\underset{\sim}{d}' \otimes \underset{\sim}{\ell}_{p_\alpha}\underset{\sim}{\ell}_{p_\alpha}' + \underset{\sim}{\Sigma} \otimes \underset{\sim}{I}_{p_\alpha} \quad , \; \alpha = 1,\ldots,N.$$

As in (A.13) we have the following decomposition of $\underset{\sim}{\Omega}_\alpha^{-1}$:

$$\text{(B.2)} \qquad \underset{\sim}{\Omega}_\alpha^{-1} = \underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_{p_\alpha} - \frac{m_\alpha}{p_\alpha}\underset{\sim}{s}\underset{\sim}{s}' \otimes \underset{\sim}{\ell}_{p_\alpha}\underset{\sim}{\ell}_{p_\alpha}'$$

where

$$\text{(B.3)} \qquad (\underset{\sim}{d}\underset{\sim}{d}')\underset{\sim}{s} = \psi\underset{\sim}{\Sigma}\underset{\sim}{s}, \quad \underset{\sim}{s}'\underset{\sim}{\Sigma}\underset{\sim}{s} = 1$$

$$\underset{\sim}{s} = \underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}/(\underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d})^{1/2} \quad , \; \psi = \underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$$

$$m_\alpha = \frac{p_\alpha\psi}{1 + p_\alpha\psi} \quad , \; \alpha = 1,\ldots,N \, .$$

But now we cannot rescale s to absorb m_α/p_α . Instead we have to keep $\psi = \underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$, the generalized signal-noise ratio, as a separate parameter.

As in (A.15) we have

$$\text{(B.4)} \qquad |\underset{\sim}{\Omega}_\alpha| = |\underset{\sim}{\Sigma}|^{p_\alpha} \, (1 + p_\alpha\psi) \, .$$

For families of size p_α the exponent term is evaluated as in (A.21) and (A.22) to give

(B.5) $$\frac{-p_\alpha q_\alpha}{2} \operatorname{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{R}_\alpha + \frac{p_\alpha q_\alpha}{2} \; m_\alpha \, \underset{\sim}{s}'\underset{\sim}{\bar{R}}_\alpha\underset{\sim}{s} \; .$$

So we define

(B.6) $$\underset{\sim}{R} = \frac{1}{\bar{p}\, q} \sum_{\alpha=1}^{N} p_\alpha q_\alpha \, \underset{\sim}{R}_\alpha$$

(B.7) $$\underset{\sim}{\bar{R}} = \frac{1}{\bar{p}\, q} \sum_{\alpha=1}^{N} p_\alpha q_\alpha m_\alpha \underset{\sim}{\bar{R}}_\alpha = \frac{1}{\bar{p}\, q} \sum_{\alpha=1}^{N} p_\alpha q_\alpha \frac{p_\alpha \psi}{1+p_\alpha \psi} \underset{\sim}{R}_\alpha \; ,$$

and our canonical form for the likelihood function is

(B.8) $$\begin{aligned} & L(\underset{\sim}{y}|\underset{\sim}{Z}, \underset{\sim}{\delta}, \underset{\sim}{s}, \underset{\sim}{\Sigma}, \psi) = \\ & -\frac{1}{2}\bar{p}q \, \ell n|\underset{\sim}{\Sigma}| - \frac{1}{2}\sum_{\alpha=1}^{N} q_\alpha \, \ell n(1 + p_\alpha\psi) \\ & -\frac{\bar{p}q}{2} \operatorname{tr} \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} + \frac{\bar{p}q}{2} \; \underset{\sim}{s}'\underset{\sim}{\bar{R}}\underset{\sim}{s} \; . \end{aligned}$$

The constraint $\underset{\sim}{s}'\underset{\sim}{\Sigma}\underset{\sim}{s} = 1$ is imposed by constructing the Lagrangian

(B.9) $$\Delta = L - \frac{\bar{p}q}{2} \tau \; \underset{\sim}{s}'\underset{\sim}{\Sigma}\underset{\sim}{s}$$

(we should note that in (A.23) there is the implicit constraint that $\underset{\sim}{C}'\underset{\sim}{\Sigma}\underset{\sim}{C}$ is a diagonal matrix; but it does not have to be imposed since the unconstrained ML solution satisfies the constraint). Setting $\partial\Delta/\partial\Sigma^{-1} = \underset{\sim}{(0)}$ gives

(B.10) $$\underset{\sim}{\Sigma} = \underset{\sim}{R} - \tau \, \underset{\sim}{\Sigma} \, \underset{\sim}{s}\underset{\sim}{s}' \, \underset{\sim}{\Sigma}.$$

The first order condition for $\underset{\sim}{s}$ is

(B.11) $$\begin{aligned} \partial\Delta/\partial s &= \bar{p}q \, \underset{\sim}{\bar{R}} \, \underset{\sim}{s} - \bar{p}q \, \tau \, \underset{\sim}{\Sigma} \, \underset{\sim}{s} = \underset{\sim}{(0)} \\ \underset{\sim}{\bar{R}}\underset{\sim}{s} &= \tau \, \underset{\sim}{\Sigma} \, \underset{\sim}{s} \; . \end{aligned}$$

Since (B.10) implies

(B.12) $$\begin{aligned} \underset{\sim}{\Sigma}\underset{\sim}{s} &= \underset{\sim}{R}\underset{\sim}{s} - \tau \, \underset{\sim}{\Sigma} \, \underset{\sim}{s} \\ \underset{\sim}{\Sigma}\underset{\sim}{s} &= \frac{1}{1+\tau} \, \underset{\sim}{R}\underset{\sim}{s}, \end{aligned}$$

we can eliminate $\underset{\sim}{\Sigma}\underset{\sim}{s}$ from (B.11) and (B.12) to obtain

$$\text{(B.13)} \qquad \underset{\sim}{\bar{R}}\underset{\sim}{s} = \lambda\, \underset{\sim}{R}\, \underset{\sim}{s} \;, \quad \lambda = \frac{\tau}{1+\tau} \quad .$$

Thus $\underset{\sim}{s}$ is an eigenvector of $\underset{\sim}{\bar{R}}$ in the metric of $\underset{\sim}{R}$. We will see that the $\underset{\sim}{s}$ corresponding to the largest eigenvalue should be chosen. The scale of $\underset{\sim}{s}$ is determined from (B.10):

$$\text{(B.14)} \qquad \underset{\sim}{s}'\underset{\sim}{R}\underset{\sim}{s} = 1 + \tau = \frac{1}{1-\lambda} \quad .$$

Then from (B.3) we obtain

$$\text{(B.15)} \qquad \underset{\sim}{d} = \sqrt{\psi}\, \underset{\sim}{\Sigma}\underset{\sim}{s} = \frac{\sqrt{\psi}}{\tau}\, \underset{\sim}{\bar{R}}\, \underset{\sim}{s} \quad .$$

So $\underset{\sim}{d}$ could be obtained from the dual of (B.13):

$$\text{(B.16)} \qquad \underset{\sim}{\bar{R}}^{-1}\underset{\sim}{d} = \frac{1}{\lambda}\, \underset{\sim}{R}^{-1}\underset{\sim}{d} \; .$$

$\underset{\sim}{\Sigma}$ can be obtained from (B.10) and (B.15):

$$\text{(B.17)} \qquad \underset{\sim}{\Sigma} = \underset{\sim}{R} - \frac{\tau}{\psi}\, \underset{\sim}{d}\underset{\sim}{d}' \quad .$$

The above analysis is all conditional on the signal-noise ratio ψ. The concentrated likelihood function $L(\psi)$ is formed by evaluating L at the maximizing values of $\underset{\sim}{\Sigma}$ and $\underset{\sim}{s}$ for a given ψ. Then ψ is chosen to maximize $L(\psi)$. So we have to evaluate 1) $|\underset{\sim}{\Sigma}|$, 2) $\operatorname{tr} \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R}$, and 3) $\underset{\sim}{s}'\underset{\sim}{\bar{R}}\underset{\sim}{s}$:

$$\text{(B.18)} \qquad 1) \quad \underset{\sim}{\Sigma} = \underset{\sim}{R} - \frac{\tau}{(1+\tau)^2}\, \underset{\sim}{R}\underset{\sim}{s}\underset{\sim}{s}'\underset{\sim}{R}$$

$$= \underset{\sim}{R}(\underset{\sim}{I} - \frac{\tau}{(1+\tau)^2}\, \underset{\sim}{s}\underset{\sim}{s}'\underset{\sim}{R})$$

$$\text{(B.19)} \qquad |\underset{\sim}{\Sigma}| = |\underset{\sim}{R}|\; (1 - \frac{\tau}{(1+\tau)^2}\; \underset{\sim}{s}'\underset{\sim}{R}\underset{\sim}{s})$$

$$= |\underset{\sim}{R}|\, \frac{1}{1+\tau} = |\underset{\sim}{R}|\; (1 - \lambda) \quad .$$

$$\text{(B.20)} \qquad 2) \quad \underset{\sim}{\Sigma}^{-1} = \underset{\sim}{R}^{-1} + \frac{\tau}{1+\tau}\, \underset{\sim}{s}\, \underset{\sim}{s}' \qquad\qquad \text{(B.18)}$$

$$\text{(B.21)} \qquad \operatorname{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} = r + \frac{\tau}{1+\tau}\; \underset{\sim}{s}'\underset{\sim}{R}\underset{\sim}{s} = r + \tau \; .$$

(B.22) 3)

$$\underset{\sim}{s}'\bar{\underset{\sim}{R}}\underset{\sim}{s} = \lambda(1 + \tau) = \tau$$

So the exponent terms cancel and

$$L(\psi) = -\frac{\bar{p}q}{2}\,\ell n\,|\underset{\sim}{R}| - \frac{\bar{p}q}{2}\,\ell n\,(1 - \lambda)$$

$$-\frac{1}{2}\sum_{\alpha=1}^{N} q_\alpha \ell n(1 + p_\alpha\psi) \quad .$$

This is an increasing function of λ for $\lambda < 1$ and so the largest root sould be chosen in (B.13).

Appendix C

ML Estimation of Model 3

Our starting point is equation (A.23), specialized to one factor:

(C.1) $$\begin{aligned} & L(\underset{\sim}{y}|\underset{\sim}{Z}, \underset{\sim}{\eta}, \underset{\sim}{c}, \underset{\sim}{\Sigma}) \\ & = -\frac{pq}{2}\ln|\underset{\sim}{\Sigma}| + \frac{q}{2}\ln(1 - p\underset{\sim}{c}'\underset{\sim}{\Sigma}\underset{\sim}{c}) \\ & -\frac{pq}{2}\ \mathrm{tr}\ \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} + \frac{p^2q}{2}\ \underset{\sim}{c}'\underset{\sim}{\bar{R}}\underset{\sim}{c} \end{aligned}$$

where now

$$\begin{aligned} \underset{\sim}{R} &= (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\eta}\underset{\sim}{d}')'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\eta}\underset{\sim}{d}')/pq \\ \underset{\sim}{\bar{R}} &= (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\eta}\underset{\sim}{d}')'\underset{\sim}{J}\underset{\sim}{J}'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\eta}\underset{\sim}{d}')/qp^2 \\ &= (\underset{\sim}{\bar{Y}} - \underset{\sim}{X}\underset{\sim}{\eta}\underset{\sim}{d}')'(\underset{\sim}{\bar{Y}} - \underset{\sim}{\bar{X}}\underset{\sim}{\eta}\underset{\sim}{d}')/q \end{aligned}$$

and $\underset{\sim}{J} = \underset{\sim}{I}_q \otimes \underset{\sim}{\ell}_p$ is a set of group indicator dummy variables. Then $\partial L/\partial\eta = 0$ can be simplified to

(C.2) $$\underset{\sim}{\eta} = (\underset{\sim}{W}_X + \zeta\underset{\sim}{B}_X)^{-1}\ (\underset{\sim}{W}_{XY} + \zeta\underset{\sim}{B}_{XY})\underset{\sim}{\Sigma}^{-1}\ \underset{\sim}{d}/\psi$$

where $\zeta = 1/(1 + p\psi)$.

Concentrating $\underset{\sim}{\eta}$ out of the likelihood function gives

(C.3) $$\begin{aligned} L(\underset{\sim}{y}|\underset{\sim}{Z}, \underset{\sim}{c}, \underset{\sim}{\Sigma}) = & -\frac{pq}{2}\ln|\underset{\sim}{\Sigma}| + \frac{q}{2}\ln\zeta \\ & -\frac{1}{2}\ \mathrm{tr}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{T}_Y + \frac{p}{2}\underset{\sim}{c}'(\underset{\sim}{B}_Y + \frac{1}{p\psi\zeta}\ \underset{\sim}{H}_{YX}\underset{\sim}{H}_X^{-1}\underset{\sim}{H}_{XY})\ \underset{\sim}{c} \quad . \end{aligned}$$

We will proceed conditional on ψ, and so we have the constraint that $\underset{\sim}{c}'\underset{\sim}{\Sigma}\underset{\sim}{c} = 1/(1 + p\psi)$. This is imposed by forming the Lagrangian:

(C.4) $$\Delta = L - \frac{pq\xi}{2}\ \underset{\sim}{c}'\underset{\sim}{\Sigma}\underset{\sim}{c} \quad .$$

Then $\partial\Delta/\partial\Sigma^{-1} = 0$ gives

(C.5) $$\underset{\sim}{\Sigma} = \frac{1}{pq}\ T_Y - \xi\zeta dd'$$

and $\partial\Delta/\partial c = 0$ gives

(C.6) $$\underset{\sim}{Q}\underset{\sim}{c} = q\ \xi\ \underset{\sim}{\Sigma}\ \underset{\sim}{c}$$

where $\underset{\sim}{Q} = \underset{\sim}{B}_Y + \frac{1}{1-\zeta}\ \underset{\sim}{H}_{YX}\underset{\sim}{H}_X^{-1}\underset{\sim}{H}_{XY}$.

Combining this with (C.5) we have

(C.7) $$\underset{\sim}{Q}\underset{\sim}{c} = \frac{1}{p}\ \frac{\zeta}{1+\zeta\psi\xi}\ \underset{\sim}{T}_Y\ \underset{\sim}{c}$$

which can be rewritten as

(C.8) $$\underset{\sim}{Q}^{-1}\underset{\sim}{d} = \frac{1}{\rho}\ \underset{\sim}{T}_Y^{-1}\ \underset{\sim}{d}$$

$$\frac{1}{\rho} = \frac{1}{p}\ \xi/(1 + \xi\psi\zeta) \quad .$$

Examination of the concentrated likelihood function in (C.10) shows that the smallest root should be chosen. Then given ρ we can solve for ξ from

(C.9) $$\xi = p/(\ \frac{1}{\rho} - 1 + \zeta) \quad .$$

The scale normalization for c follows from (C.6): $\underset{\sim}{c}'\underset{\sim}{Q}\underset{\sim}{c} = q\ \zeta\ \xi$. Equivalently, the normalization for $\underset{\sim}{d}$ is

(C.10) $$\underset{\sim}{d}'\underset{\sim}{T}_Y^{-1}\underset{\sim}{d} = \frac{\psi}{pq(1 + \zeta\ \psi\ \xi)} \quad .$$

Finally, we use (C.5) and (C.7) to write the concentrated likelihood function as a function just of ψ:

(C.11) $$L(\psi) = -\frac{pq}{2}\ \ell n(1 - \rho + \rho\zeta) + \frac{q}{2}\ \ell n\zeta \ .$$

So our algorithm reduces to a one dimensional maximization problem.

Appendix D

ML Estimation of Model 4

We can apply (C.2) to obtain

(D.1) $\utilde{\eta} = (\utilde{W}_X + \zeta \utilde{B}_X)^{-1}(\utilde{W}_{XA} + \zeta \utilde{B}_{XA})\ \utilde{\Sigma}^{-1}\utilde{\lambda}/\psi$

where $\utilde{A} = \utilde{Y}\utilde{\Gamma}$. But now $\utilde{\Sigma}$ is constrained by $\utilde{\Sigma} = \tau\utilde{\lambda}\utilde{\lambda}' + \utilde{U}$, and so we have $\psi = \utilde{\lambda}'\utilde{\Sigma}^{-1}\utilde{\lambda} = \mu/(\tau\mu + 1)$, $\mu = \utilde{\lambda}'U^{-1}\utilde{\lambda}$, and

(D.2) $\utilde{\Sigma}^{-1}\utilde{\lambda}/\psi = \utilde{U}^{-1}\utilde{\lambda}/\mu$.

Similarly (C.3) can be simplified to

(D.3) $$L(\utilde{y}|\utilde{Z},\ \utilde{c},\ \utilde{\tau},\ \utilde{U},\ \utilde{\Gamma}) = -\frac{pq}{2}\,\ell n|\utilde{U}| - \frac{pq}{2}\,\ell n(1 + \tau\mu) + \frac{q}{2}\,\ell n\zeta$$
$$-\frac{1}{2}\,\mathrm{tr}\utilde{U}^{-1}\utilde{T}_Y + \frac{p}{2}\,(1 + \tau\mu)\utilde{c}'\utilde{Q}\utilde{c}/(1 - \zeta)$$

with

$$Q = \frac{\tau\mu}{1+\tau\mu}\ \utilde{T}_A + \frac{1}{1+\tau\mu}\,[(1 - \zeta)\utilde{B}_A + \utilde{H}_{AX}\utilde{H}_X^{-1}\utilde{H}_{XA}] \quad .$$

$\utilde{c}'\utilde{U}\utilde{c}$ is fixed conditional on μ and τ, and so we maximize the Lagrangian

(D.4) $\Delta = L - \dfrac{pq\xi}{2}\ \utilde{c}'\utilde{U}\utilde{c}$.

$\partial\Delta/\partial c = 0$ gives

(D.5) $$\utilde{Q}^{-1}\utilde{\lambda} = \frac{1}{\rho}\,\utilde{U}^{-1}\utilde{\lambda}$$
$$\rho = q\xi(1 - \zeta)/(1 + \tau\mu)$$

where the smallest root should be chosen. The scale of $\utilde{\lambda}$ is given by $\utilde{\lambda}'\utilde{Q}^{-1}\utilde{\lambda} = \mu/\rho$.

Then the concentrated likelihood function is

(D.6) $$L(\mu,\ \tau,\ \utilde{U},\ \utilde{\Gamma}) = -\frac{pq}{2}\,\ell n|\utilde{U}| - \frac{pq}{2}\,\ell n(1 + \tau\mu)$$
$$+ \frac{q}{2}\,\ell n(1 + \tau\mu)/(1 + (p + \tau\mu)) - \frac{1}{2}\,\mathrm{tr}\utilde{U}^{-1}\utilde{T}_A + \frac{1}{2}\,\rho.$$

This must be maximized numerically as a function of μ, τ, and $\utilde{U}$. The gradient and hessian of L require the evaluation of first and second order derivatives of the eigenvalue with respect to elements of the quadratic forms in (D.5). Expressions for such derivatives are given in Wilkinson (1965), Jennrich and Robinson (1969), and Jöreskog and Goldberger (1973).

If $\underset{\sim}{\Gamma} = \underset{\sim}{I}$ then the above analysis reduces the ML problem to the level of difficulty of a first order factor model. Jöreskog (1967), Jennrich and Robinson (1969), and Jöreskog and Goldberger (1973) have had considerable success in the numerical maximization of first order factor model likelihood functions similar to (D.6). However, if $\underset{\sim}{\Gamma}$ is unknown, then performing such a maximization on the residuals from each iteration of the GLS estimator of $\underset{\sim}{\Gamma}$ may be quite costly. One alternative would be to include $\underset{\sim}{\Gamma}$ in the hill-climbing algorithm, so that only one sequence of variable metric iterations would be used on (D.6). Another alternative, which seems attractive if there is a large number of unknown parameters in $\underset{\sim}{\Gamma}$ relative to $\underset{\sim}{\lambda}$, is to concentrate $\underset{\sim}{\Gamma}$ and $\underset{\sim}{\eta}$ out of the likelihood function via GLS, and then use some modification of a gradient method to maximize over λ, $\underset{\sim}{U}$, and τ. This would be similar to Jöreskog's (1970, 1973) treatment of the second order factor model.

Chapter 4

Education, Income, and Ability Revisited[1]

Introduction

This paper reanalyzes the 1964 CPS-NORC veteran's data. A description of the sample and the data is contained in Griliches and Mason (1972); we have reproduced part of their table 1, summarizing some of the major characteristics of the sample. Our interest centers on the schooling coefficient in a semi-logarithmic income generating function with the log of income (LINC) as the dependent variable. We want to know how much of the observed coefficient is due to a selectivity bias, simply reflecting the correlation of schooling with ability instead of a value added by the schooling itself.

This question was examined in some detail by Griliches and Mason. They introduced a variety of background variables and a test score (AFQT) in an attempt to control for the individual's initial ability. Some of their results are reproduced in table 2. We will follow them in devoting most of our attention to the schooling increment variable (SI). It is the part of total schooling (ST) incurred during or after military service. Since the test is administered prior to entering the service, it can be regarded as a measure of "early" ability relative to the schooling increment. As shown in section V this is quite crucial to our approach.

We see in table 2 that introducing the background characteristics and the test score (equation 1 vs. equation 4) produces a

Table 1: Means and Standard Deviations of Variables: Veteran's Age 21-34 in 1964 CPS Subsample

Variable	Mean or Fraction in Sample	SD	Symbol in Subsequent Tables	Group Name
Personal background:				
Age (years)	29.0	3.5	Age	
Color (white)	0.96	*	C	
Schooling before Service (years)	11.5	2.3	SB	
Total schooling (years)	12.3	2.5	ST	
Schooling increment (years)	0.8	1.4	SI	
AFQT (percentile)	54.6	24.8	AFQT	
Length of active military service (months)	30.7	16.9	AMS	
Father's schooling (years)	8.7	3.2	FS	Fa. stat.
Father's occupational SES	29.0	20.6	FO	
Grew up in South	0.29	*	ROS	Reg. bef.
Grew up in large city	0.22	*	POC	
Grew up in suburb of large city	0.05	*	POS	
Log current occupational SES	3.47	0.68	LOSES	
Actual income (weekly dollars)	122.5	52.4	...	
Log actual income	4.73	0.40	LINC	

NOTE: N = 1,454 for this and subsequent tables based on the 1964 CPS. Fa. stat. = father's status; reg. bef. = region before.

*The standard deviation for a dummy variable is equal to f(1-f), where f is the fraction in the sample having the requesite characteristic. Thus, it is computable from the numbers given in the first column.

Table 2: Regression Equations with Log Income as Dependent Variable

	Coefficient (Standard Error) Of						
Regression No.	Color	SB	SI	ST	AFQT	Other Variables in Equation*	R^2
1	.2548 (.0472)	.0502 (.0042)	.0528 (.00702)	...	...	Age, AMS	.1666
2	.2225 (.0479)	.0418 (.0049)	.0475 (.0072)	...	.00154 (.00045)	Age, AMS	.1732
3	.1904 (.0473)	.0379 (.0045)	.0496 (.0070)	...	...	Age, AMS, fa. stat., reg. bef.	.2129
4	.1714 (.0479)	.0328 (.0050)	.0462 (.0071)	...	.00105 (.00045)	Age, AMS, fa. stat., reg. bef.	.2159
5	.2544 (.0471)	...	...	.0508 (.0039)	...	Age, AMS	.1665
6	.22245 (.04793)	...	...	.0433 (.0044)	.00150 (.00045)	Age, AMS	
7	.1907 (.0473)	...	...	.0408 (.0041)	...	Age, AMS, fa. stat., reg. bef.	.2115
8	.1732 (.0479)	...	...	.0365 (.0046)	.00097 (.00044)	Age, AMS, fa. stat., reg. bef.	.2141
9	.1335 (.0487)	...	...	....	.00252 (.00041)	Age, AMS, fa. stat., reg. bef.	.1794
10	.1742 (.0488)	...	...	...	...	Age, AMS, fa. stat., reg. bef.	.1578

NOTE: See table 1 for definitions.

*Variable groups are denoted as follows: fa. stat. = fa. occ. and fa. schooling; reg. bef. = ROS, POC, POS.

decline in the SI coefficient from .053 to .046, which is only 12%. Our analysis takes off from equation 4, asking whether there are important dimensions of ability, unaccounted for by the available variables, which seriously bias the SI coefficient.

Section II trys to obtain identification from the residual covariance matrix. An argument very similar to the one in Chamberlain and Griliches (1974) can be used, with the availability of a test score substituting for the within family replication on brothers. It turns out, however, that the results are very sensitive to some of the more questionable assumptions of the model, and we conclude that by itself this approach is not very informative. In section III we structure the background coefficients in the income and test equations by imposing proportionality restrictions derived from an aggregation assumption. This analysis, standing by itself, is also inconclusive. But by meshing the two approaches we obtain in section IV a plausible model which is quite informative about the SI coefficient. Our substantive finding is that there is little evidence of bias from the omission of important dimensions of initial ability. Section V asks whether a similar result holds for total schooling (ST) or for schooling before service (SB). Working just with ST we find that the AFQT cannot be used as a measure of early ability relative to ST. This is because SB does have a value added in determining the test score. But regarding the test as a measure of late, post-school ability results in an unidentified model. So we turn to a more careful examination of the SB-SI split, trying to identify the bias in the return to SB. Our estimate is that it is quite small once we have controlled for the available background variables. There is a brief concluding section.

II. Structuring the Residual Covariance Matrix

We will work with the following model:

$$\text{(II.1)}\qquad Y = \text{LINC} = \underset{\sim}{X}\underset{\sim}{\xi}_1 + SI\beta_1 + H\gamma_1 + u$$

$$O = \text{LOSES} = \underset{\sim}{X}\underset{\sim}{\xi}_2 + SI\beta_2 + H\gamma_2 + v$$

$$SI = \underset{\sim}{X}\underset{\sim}{\xi}_3 + H\gamma_3 + w,$$

$$T = \text{AFQT} = \underset{\sim}{X}\underset{\sim}{\xi}_4 + H\gamma_4 + t$$

where X includes COLOR, AGE, AMS, and the background characteristics POC, POS, FO, FS, SB, ROS. H is a combination of unobserved characteristics such as genetic ability and parental wealth. Although it is presumably correlated with the observed background characteristics, we can transform the model to make H and $\underset{\sim}{X}$ uncorrelated.

Let $\underset{\sim}{b}_{H,X} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{H}$ and rewrite

$$\text{(II.2)}\qquad \underset{\sim}{X}\underset{\sim}{\xi}_k + H\gamma_k = \underset{\sim}{X}(\underset{\sim}{\xi}_k + \underset{\sim}{b}_{H,X}\gamma_k) + (H - \underset{\sim}{X}\underset{\sim}{b}_{H,X})\gamma_k$$

$$= \underset{\sim}{X}\underset{\sim}{\tilde{\xi}}_k + \tilde{H}\gamma_k\ , \qquad k = 1,\ldots,4.$$

Now $\underset{\sim}{X}$ is orthogonal to $\tilde{H}$ by construction, and we can treat $\underset{\sim}{X}$ as exogenous. The point is that to the extent H is correlated with $\underset{\sim}{X}$, it does not bias the estimates of β_1 and β_2 in a regression that includes $\underset{\sim}{X}$. So we reinterpret H as that part of initial ability (after SB but before SI) that is uncorrelated with $\underset{\sim}{X}$. Then we must also reinterpret the ξ's to include not only the direct effect of $\underset{\sim}{X}$ but also the indirect effect via its correlation with the originial H. The possibility of decomposing the SB coefficient into its direct and indirect components will be considered in section V.

Then surpressing the slope coefficients, which are unconstrained and hence do not help to identify the β's, we can substitute T for H in the Y and O equations:

$$\text{(II.3)}\quad \begin{aligned} Y &= SI\beta_1 + T\,\gamma_1/\gamma_4 + u - \gamma_1/\gamma_4\,t \\ O &= SI\beta_2 + T\,\gamma_2/\gamma_4 + v - \gamma_2/\gamma_4\,t \end{aligned} \quad .$$

Now we have an errors in variables problem caused by the measurement error in T. Define ρ_N, the net reliability of T, as the fraction of the variance of T which is due to the systematic influence of H: $\rho_N = \gamma_4{}^2\sigma_H{}^2 / (\gamma_4{}^2\sigma_H{}^2 + \sigma_t{}^2)$, and let $\alpha_1 = \gamma_1/\gamma_4$. Then we have the following bias formulas (e.g. Griliches and Ringstad [1974]):

$$\text{(II.4)}\quad \begin{aligned} \text{plim}\ \hat{\alpha}_1 &= \alpha_1 - (1 - \rho_N)\ \alpha / (1 - r_{T, SI}^2) \\ \text{plim}\ \hat{\beta}_1 &= \beta_1 + (1 - \rho_N)\ \alpha\ b_{T,SI} / (1 - r_{T, SI}^2) \end{aligned}$$

where $\hat{\alpha}_1 = b_{Y,T.\,SI}$, $\hat{\beta}_1 = b_{Y,SI.\,T}$

and all of the variables have been replaced by their residuals from a regression on $\underset{\sim}{X}$. Solving for β_1 and simplifying gives

$$\text{(II.5)}\quad \beta_1 = \text{plim}\ (b_{Y,SI} - (1/\rho_N)\, b_{Y,T}\, b_{T,SI}) / (1 - (1/\rho_N)\, r_{SI,T}^2),$$

and there is a similar formula for β_2.

So we can obtain estimates of β_1 and β_2 conditional on ρ_N. Whether or not there is a useful prior bound on ρ_N depends crucially on our interpretation of H. One interpretation is that H (or $\underset{\sim}{X}\underset{\sim}{\xi}_4 + \gamma_4 H$) is the "true score". Then the test adequately measures the relevant initial characteristics except for an error (t) which could be eliminated by replicating the test. In this case it's reasonable to assume that t is uncorrelated with everything else. Furthermore, to the extent that the AFQT test is comparable to civilian IQ tests, we can bound the reliability ($\rho = 1 - \sigma_t^2/\sigma_T^2$) at say $\rho \geq .6$. Then using $(1 - \rho) = (1 - R^2_{T.X})(1 - \rho_N)$ we have the bounds on β_1 and β_2 given in table 3. We see that over this range of reliabilities there is not much bias in the schooling coefficients.

Table 3: Estimates of (II.3) Conditional on the Reliability

ρ*	ρ_N**	β_1	β_2
.60	.45	.042	.108
.70	.59	.044	.112
.80	.72	.045	.115
.90	.86	.046	.117
.95	.93	.046	.118

*$\rho = 1 - \sigma_t^2 / \sigma_T^2$ is the reliability of T.

**$\rho_N = \gamma_4^2 / (\gamma_4^2 + \sigma_t^2)$ is the reliability of T net of $\underset{\sim}{X}(\sigma_H^2 = 1)$.

An alternative, more general interpretation of H is that IQ tests are designed to predict academic performance and need not capture (or appropriately weight) the set of characteristics relevant for economic success. Under this interpretation the test is only capturing a piece of the relevant initial conditions. Since it is being used outside the context it was designed for, fewer prior restrictions can be imposed. We cannot restrict the reliability and considerable care is required in making independence assumptions about t. For example a low reliability means that much of the test distribution is being assigned to the residual t. But if the test is a reasonable predictor of academic success and if H is not capturing that, then t and the schooling residual w will be correlated.

So we will try to estimate the reliability. The reduced form is

$$\text{(II.6)}\qquad \begin{aligned} Y &= Hd_1 + u + \beta_1 w \\ O &= Hd_2 + v + \beta_2 w \\ SI &= Hd_3 + w \\ T &= Hd_4 + t \end{aligned}$$

where

$$\underset{\sim}{d} = \begin{bmatrix} \gamma_1 + \beta_1\gamma_3 \\ \gamma_2 + \beta_2\gamma_3 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}$$

The reduced form residual covariance matrix is $\underset{\sim}{I}_N \otimes \underset{\sim}{\Omega}$ with

$$\text{(II.7)}\quad \underset{\sim}{\Omega} = (\omega_{ij}) = \underset{\sim}{d}\underset{\sim}{d}' + \underset{\sim}{T}$$

$$\underset{\sim}{T} = (\upsilon_{ij}) = \begin{bmatrix} \sigma_u^2 + \beta_1^2\sigma_w^2 & \beta_1\beta_2\sigma_w^2 & \beta_1\sigma_w^2 & 0 \\ & \sigma_v^2 + \beta_2^2\sigma_w^2 & \beta^2\sigma_w^2 & 0 \\ & & \sigma_w^2 & 0 \\ & & & \sigma_t^2 \end{bmatrix}$$

where we have scaled H so that $\sigma_H{}^2 = 1$ and we have assumed that u, v, t, and w are independent. So $\underset{\sim}{\bar{d}}_I = \begin{bmatrix} \omega_{14} \\ \omega_{24} \\ \omega_{34} \end{bmatrix}$ gives us $\underset{\sim}{d}_I = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$ up to a scale factor γ_4. Then let $\tau = 1/\gamma_4{}^2 = 1/(\omega_{44}\rho_N)$ and write the upper left 3x3 corner of $\underset{\sim}{\Omega}$ as

$$\text{(II.8)}\quad \underset{\sim}{\Omega}_I = \tau\underset{\sim}{\bar{d}}_I\underset{\sim}{\bar{d}}_I' + \underset{\sim}{T}_I.$$

So given τ we can solve for

$$\text{(II.9)}\quad \sigma_w^2 = \omega_{33} - \tau\bar{d}_3{}^2$$

$$\beta_1 = (\omega_{13} - \tau\bar{d}_1\bar{d}_3)/\sigma_w^2$$

$$\beta_2 = (\omega_{23} - \tau\bar{d}_2\bar{d}_3)/\sigma_w^2 .$$

But we can also solve for

$$\beta_1\beta_2 = (\omega_{12} - \tau\bar{d}_1\bar{d}_2)/\sigma_w^2 .$$

τ is determined by making the separate solutions for β_1 and β_2 agree with the solution for their product. This yields:

$$(\text{II},10)\quad \tau = (\omega_{13}\omega_{23} - \omega_{12}\omega_{33})/(\omega_{13}\bar{d}_2\bar{d}_3 + \omega_{23}\bar{d}_1\bar{d}_3 - \omega_{12}\bar{d}_3^{\,2} - \omega_{33}\bar{d}_1\bar{d}_2).$$

So τ is identified and we can use (II.9) to solve for β_1 and β_2. This is equivalent to substituting $\rho_N = 1/(\omega_{44}\tau^2)$ into the unscrambled errors-in-variables formula (II.5).

An alternative interpretation of this procedure is based on instrumental variables. When we use T as a proxy for H in the Y equation (II.3); the problem is to find an instrument for T. But O is uncorrelated with u and t and is correlated with T because they both depend on H . Similarly Y can be used as a instrument for T in the O equation.

Unfortunately we cannot relax the independence assumptions on u, v, t, and w without making the model unidentified. But if u consists largely of luck which results in a higher income than an individual's schooling and ability would have predicted, then he is likely to also have a higher occupational status, implying a positive correlation between u and v. On the other hand, if u and v reflect the individual's preferences for income vs. status, and if, given his schooling and ability, he can trade off one for the other, then the correlation could be negative. So we want to relax the no correlation assumption and try to obtain identification in the sense of a non-trivial bound. In the Chamberlain-Griliches (1974) model the results were not sensitive to this assumption and a sharp bound was obtained. We can either allow a non-zero correlation between u and v, or alternatively (and equivalently) rewrite the Y equation to include O and keep the E(uv) = 0 assumption;

(II.11) $Y = SI\beta_1 + O\lambda + H\gamma_1 + u.$

Then in the reduced form we have $\omega_{12} = d_1 d_2 + \beta_1\beta_2\sigma_w^2 + \lambda\sigma_v^2$. As in the Chamberlain-Griliches model, conditioning on λ will identify the rest of the model and the non-negativity constraints on the variances will generate a bound on λ .

We will also attempt a sensitivity analysis of the covariance between t and w. To do this we structure the residual covariances in terms of two distinct but correlated kinds of ability, economic (H_1) and scholastic (H_2). Then we have

(II.12)
$$\begin{aligned} Y &= SI\beta_1 + H_1\gamma_1 + u \\ O &= SI\beta_2 + H_1\gamma_2 + v \\ SI &= H_2\gamma_3 + w' \\ T &= H_1\kappa_1 + H_2\kappa_2 + t'. \end{aligned}$$

The test is assumed to measure a combination of both kinds of ability. The simultaneity problem results from the correlation between H_1 and H_2, which we express in terms of a shared set of characteristics H:

(II. 13)
$$\begin{aligned} H_1 &= H\psi_1 + e_1 \\ H_2 &= H\psi_2 + e_2, \end{aligned}$$

with e_1 independent of e_2 by construction, and as above H, e_1, and e_2 are orthogonal to X by construction. Then we can rewrite

$$\text{(II.14)}\quad \begin{aligned} SI &= H\gamma_3 + w \\ T &= H\gamma_4 + t \end{aligned}$$

where t and w are uncorrelated with H but now t is correlated with w.

This model is not identified although again there is the possibility of useful bounds. The reduced form Ω is now

$$\text{(II.15)}\quad \underset{\sim}{\Omega} = \underset{\sim}{d}\underset{\sim}{d}' + \begin{bmatrix} \sigma_u^2 + \beta_1^2\sigma_w^2 & \beta_1\beta_2\sigma_w^2 & \beta_1\sigma_w^2 & \beta_1\sigma_{tw} \\ & \sigma_v^2 + \beta_2^2\sigma_w^2 & \beta_2\sigma_w^2 & \beta_2\sigma_{tw} \\ & & \sigma_w^2 & \sigma_{tw} \\ & & & \sigma_t^2 \end{bmatrix} .$$

So conditioning on σ_{tw} we can choose initial values for β_1 and β_2 and take

$$\begin{bmatrix} \omega_{14} \\ \omega_{24} \\ \omega_{34} \end{bmatrix} - \begin{bmatrix} \beta_1 \\ \beta_2 \\ 1 \end{bmatrix} \sigma_{tw}$$

to get $\underset{\sim}{\overline{d}}_I$ which, as shown above, lets us solve for new values for β_1 and β_2 which can be used to repeat the process to convergence.

The results of applying these models to the veterans data are shown in table 4. We see that the model with both λ and

σ_{tw} constrained to zero gives highly implausible schooling coefficients. So it is not surprising that quite small changes from zero in λ or σ_{tw} imply very substantial changes in the β's. That a 10% increase in occupational prestige would be associated with either a .6% or a .9% increase in income, for given background, schooling and ability, is not implausible. But as λ varies over this range the schooling coefficient β_1 varies from .02 to .05. There is a bound in that higher values of λ than indicated would imply $\sigma_w^2 < 0$. But lower values are not ruled out and so the bound is not useful over the controversial range for β_1 from zero or .02 to .05.

Note that the table is quite informative on the test's reliability $\rho = 1 - \sigma_t^2/\sigma_T^2$. The low $\hat{\beta}$'s for $\lambda = 0$ arise because the test is estimated to be very unreliable and so $b_{Y,SI\cdot X,T}$ is given a large downward adjustment (with a corresponding upward adjustment to $b_{Y,T\cdot X,SI}$). Even with $\lambda \neq 0$ we can bound ρ at .7 in the sense that higher values would imply restrictions on the reduced form likelihood that would be testable. Now these low reliabilities suggest that the common omitted variable we are picking up is not IQ, at least if the high reliabilities quoted for standard intelligence tests can be applied to the AFQT. Thus the two factor model with $\sigma_{tw} \neq 0$ is quite relevant.

Table 4: Residual Covariance Estimates of (II.1), (II.11) and (II.12)

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$*	$\hat{\rho}_N$**	λ	r_{tw}***
-.043	-.089	.33	.08	.0	.0
.019	.053	.38	.15	.063	.0
.031	.080	.43	.22	.073	.0
.044	.109	.70	.58	.083	.0
.020	.054	.32	.06	.0	.133
.031	.081	.31	.05	.0	.164
.042	.104	.30	.04	.0	.189
.062	.150	.29	.02	.0	.231
.086	.204	.28	.01	.0	.262

Note: The residual covariances are based on OLS regressions of LINC, LOSES, AFQT, SI on AGE, AMS, POC, POS, FO, FS, SB, ROS. All rows of the table are equally likely, giving different interpretations of the M.L. reduced form Ω by making different assumptions about λ and σ_{tw}.

*$\rho = 1 - \sigma_t^2/\sigma_T^2$ is the reliability of T.

**$\rho_N = \gamma_4^2 / (\gamma_4^2 + \sigma_t^2)$ is the reliability of T net of X$(\sigma_H^2 = 1)$.

***r_{tw} is the correlation between t and w in (II.14).

In that model we find that again rather small departures from $\sigma_{tw} = 0$ imply substantial changes in the β's. Since any t,w correlation between zero and .2 or .25 is not implausible, there is little direct information here on the schooling coefficients. There is an upper bound on the β's; higher values than shown would imply negative values for ρ_T.

So our analysis of the residual covariances has not been very informative. Most important is the negative inference that given our prior range of plausible values for β_1, we cannot accept the restricted model with $\lambda = 0$ and $\sigma_{tw} = 0$. But to get more out of these residual covariances we have to put more in. The next section looks at imposing more structure on the background coefficients.

III. Structuring the Background Coefficients

The proportionality restrictions we will use are based on aggregating the individual's characteristics into a single factor (G), "human capital" or "capacity." Then we can write the structural form as

$$
\text{(III.1)}\quad
\begin{aligned}
Y &= \underset{\sim}{X}_1\underset{\sim}{\xi}_1 + && SI\beta_1 + Gr_1 + \upsilon_1 \\
T &= \underset{\sim}{X}_1\underset{\sim}{\xi}_3 + && \qquad\;\; G \;\; + \upsilon_3 \\
G &= \qquad\quad \underset{\sim}{M}\underset{\sim}{\pi} && \\
SI &= \underset{\sim}{X}_1\underset{\sim}{\xi}_4 + \underset{\sim}{M}\underset{\sim}{\zeta} && \qquad\qquad\;\; + \upsilon_4
\end{aligned}
$$

where $\underset{\sim}{X}_1$ includes AGE, AMS, POC, POS and M includes the observable background characteristics FO, FS, SB, ROS.[2] The υ's are allowed to be freely correlated across the equations and so there are no restrictions on the unobservable characteristics.

Then surpressing $\underset{\sim}{X}_1$, the reduced form is

$$
\text{(III.2)}\quad
\begin{aligned}
Y &= \underset{\sim}{M}(\underset{\sim}{\pi}r_1 + \underset{\sim}{\zeta}\beta_1) + \varepsilon_1 = \underset{\sim}{M}\underset{\sim}{\delta}_1 + \varepsilon_1 \\
T &= \underset{\sim}{M}\underset{\sim}{\delta}_3 + \varepsilon_3 \\
SI &= \underset{\sim}{M}\underset{\sim}{\delta}_4 + \varepsilon_4 \, ,
\end{aligned}
$$

with $\underset{\sim}{\delta}_1 = \underset{\sim}{\delta}_3 r_1 + \underset{\sim}{\delta}_4\beta_1$. Letting $\underset{\sim}{\Delta} = (\underset{\sim}{\delta}_1\underset{\sim}{\delta}_3\underset{\sim}{\delta}_4)$ and $\underset{\sim}{\mu} = \begin{pmatrix} 1 \\ -r_1 \\ -\beta_1 \end{pmatrix}$ lets us write the constraint as

$$
\text{(III.3)}\quad \underset{\sim}{\Delta}\underset{\sim}{\mu} = (\underset{\sim}{0}).
$$

This will uniquely determine $\underset{\sim}{\mu}$ if and only if rank $\underset{\sim}{\Delta} = 2$. So there is a necessary order condition that $\underset{\sim}{M}$ contain at least two background variables. If there are just two, we simply do OLS and solve for β_1 and r_1 from the unconstrained reduced form, i.e., indirect least squares. With more variables in $\underset{\sim}{M}$ the restrictions can be imposed via limited information single equation maximum likelihood (LISE).[3] For we can substitute T for G in the Y equation:

(III.4) $Y = SI\beta_1 + Tr_1 + \nu_1'$,

thereby freeing up the background variables to be used as instruments for SI and T. Applying LISE to this equation is in fact full information maximum likelihood (FIML) since the other two equations in the system are just identified.

Adding the occupational SES equation gives

(III.5) $O = SI\beta_2 + Gr_2 + \upsilon_2$,

and the reduced form is

(III.6) $O = \underset{\sim}{M}(\underset{\sim}{\pi}r_2 + \underset{\sim}{\zeta}\beta_2) + \varepsilon_2 = \underset{\sim}{M}\underset{\sim}{\delta}_2 + \varepsilon_2$

with $\underset{\sim}{\delta}_2 = \underset{\sim}{\delta}_3 r_2 + \underset{\sim}{\delta}_4\beta_2$. So now $\underset{\sim}{\Delta} = (\underset{\sim}{\delta}_1\underset{\sim}{\delta}_2\underset{\sim}{\delta}_3\underset{\sim}{\delta}_4)$ is subject to two constraints:

(III.7) $\underset{\sim}{\Delta}(\underset{\sim}{\mu}_1 \ \underset{\sim}{\mu}_2) = (\underset{\sim}{0})$

with $\underset{\sim}{\mu}_1 = \begin{bmatrix} 1 \\ 0 \\ -r_1 \\ -\beta_1 \end{bmatrix}$, $\underset{\sim}{\mu}_2 = \begin{bmatrix} 0 \\ 1 \\ -r_2 \\ -\beta_2 \end{bmatrix}$.

The necessary and sufficient condition for identification of the subspace spanned by $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ is that rank $\underset{\sim}{\Delta} = 2$. Given that subspace we can recover the β's and r's by excluding O from the Y equation and vice versa.

The overidentifying restrictions in this model can be imposed by a straightforward extension of LISE. For as in (III.4) we can rewrite the Y and O equations as

$$\text{(III.8)} \quad Y = SI\beta_1 + Tr_1 + \nu_1'$$

$$O = SI\beta_2 + Tr_2 + \nu_2' \ .$$

So these two equations are just identified relative to each other: Y excludes O and O excludes Y. Hannan (1967) showed that for such a subsystem, limited information maximum likelihood (LIML) can be obtained from a canonical correlation analysis which is a straightforward extension of the LISE eigenvalue problem. Since the T and SI equations are just identified, LIML is FIML.

Applying the Y-T-SI model to the veterans data gives $\hat{\beta}_1 = .063$ with an (asymptotic) standard error of .041. The concentrated likelihood function in table 4 confirms the imprecision of this point estimate. Adding the O equation as in (III.8) gives $\hat{\beta}_1 = .028$ but again the concentrated likelihood function is quite flat. The next section attempts a more informative analysis by combining the proportionality restrictions with the residual covariance structure of section II.

Table 5: Concentrated Likelihood Function for β_1

Y - T - SI, (III.1)							
β_1 :	.00	.02	.03	.04	.05	.06	.07
L.R.:	.27	.56	.70	.81	.93	1.00	.93
χ^2 :	2.59	1.15	.72	.43	.14	.00	.14
Y - O - T - SI, (III.1) + (III.5)							
β_1 :	.00	.02	.03	.04	.05	.06	.07
L.R.:	.80	.98	.99	.96	.86	.74	.59
χ^2 :	.45	.04	.01	.09	.30	.61	1.04

Note: L.R. = Likelihood ratio; $\chi^2 = -2 \text{ Log (L.R.)}$ is approximately distributed as $\chi^2(1)$.

IV. The Joint Treatment: Meshing the Two Approaches

The basic idea behind our joint treatment of covariance and slope restrictions is to extend the proportionality assumption across both the observed and unobserved characteristics. So G, the "human capital" variable, is expanded to

$$\text{(IV.1)} \qquad G = \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3$$

where, as in section II, H is the part of initial ability that is uncorrelated with the observed background characteristics $\underset{\sim}{M}$. Then surpressing the exogenous variables that appear in all of the equations ($\underset{\sim}{X}_1$) we have

$$\text{(IV.2)} \qquad \begin{aligned} Y &= SI\beta_1 + Gr_1 + u \\ T &= G + t \\ G &= \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3 \\ SI &= \underset{\sim}{M}\underset{\sim}{\zeta} + H\gamma_4 + w. \end{aligned}$$

So the coefficient of H in the Y equation is constrained to be $\gamma_1 = r_1\gamma_3$. This model is similar to the one in section IV of the Griliches-Mason paper, except they excluded H from the SI equation. We will refer to that model as Y1, and the model without the $\gamma_4 = 0$ constraint as Y2.[4/] Both of these models assume that u, v, t, and w are uncorrelated with each other. Following section II we will also consider the model Y3 in which $\sigma_{tw} \neq 0$.

The interpretation of Y3 needs additional comment. It is a hybrid combination of the two factor model of section II and the one factor structure for the background coefficients introduced in section III. As in (II.12) we specify

$$\text{(IV.3)}\quad \begin{aligned} Y &= SI\beta_1 + G_1 r_1 + u \\ T &= G_1\kappa_1 + G_2\kappa_2 + t' \\ SI &= \underset{\sim}{M}\underset{\sim}{\zeta}' + G_2\gamma_4' + w'. \end{aligned}$$

This disaggregates the human capital variable into the bundle of characteristics relevant for economic success, G_1, and for scholastic success, G_2. The correlation between G_1 and G_2 is represented via their common dependence on a shared set of attributes G:

$$\text{(IV.4)}\quad \begin{aligned} G_1 &= G\psi_1 + e_1 \\ G_2 &= G\psi_2 + e_2 , \end{aligned}$$

where e_1 and e_2 are independent of G and of each other. In section II this was a completely general way of specifying the correlation, but now the model is completed with a more detailed prior for G:

$$\text{(IV.5)}\quad G = \underset{\sim}{M}\underset{\sim}{\pi}' + H\gamma_3'.$$

Thus $\underset{\sim}{M}$ affects Y and T in a constrained way, working only through the general ability factor G. Then we can rewrite this model so that it is identical to (IV.2) except now t and w are correlated.

There is, of course, the less constrained model:

$$\text{(IV.6)}\quad \begin{aligned} G_1 &= \underset{\sim}{M}\pi_1 + H\,\mu_1 + e_1 \\ G_2 &= \underset{\sim}{M}\pi_2 + H\,\mu_2 + e_2 \quad . \end{aligned}$$

But this model is not particularly estimable from our data; it essentially takes us back to section II.

Note that the SI equation is not subject to the proportionality restriction. Even if the constraint were reasonable for ST, which is unlikely, there is no reason to constrain the way ST splits into SB and SI. This point is quite important. For if the proportionality restriction did hold across the SI equation, then the rank condition for identification would fail identically.

As for estimation, the two stage procedure used by Griliches and Mason is quite reasonable for model Y1. They constructed a $\hat{T}$ from a first stage regression of T on $\underset{\sim}{M}$ and used the fitted values to get $\hat{\beta}_1 = b_{Y,SI\cdot\hat{T}}$. [5/] In model Y2, which does not exclude H from SI, there is again a reasonable two stage procedure. But now we must include SI as well as $\underset{\sim}{M}$ in the first stage $\hat{T}$ regression. For in general all of the included exogenous variables must be used in the first stage of a two stage least squares procedure (see, e.g., Brundy and Jorgenson [1974]). It may seem odd to use SI to construct $\hat{T}$ since the schooling increment is obtained after the test. But provided $\gamma_4 \neq 0$, SI can serve as a proxy for H. To clarify this we write the system as

$$\text{(IV.7)}\qquad \begin{aligned} Y &= SI\beta_1 + Tr_1 + (u - r_1 t) \\ T &= \underset{\sim}{M}(\underset{\sim}{\pi} - \underset{\sim}{\zeta}\gamma_3/\gamma_4) + SI\gamma_3/\gamma_4 + (t - \frac{\gamma_3}{\gamma_4}w) \\ SI &= \underset{\sim}{M}\underset{\sim}{\zeta} \qquad\qquad\qquad\qquad\qquad\quad + H\gamma_4 + w. \end{aligned}$$

Then rewrite the T equation so that its residual (t') is orthogonal to SI:

$$\text{(IV. 8)}\qquad T = \underset{\sim}{M}(\underset{\sim}{\pi} - \underset{\sim}{\zeta}\gamma_3/\gamma_4) + SI[1 - \sigma_w^2/(\sigma_w^2 + \gamma_4^2\sigma_H^2)]\gamma_3/\gamma_4 + t'.$$

So we regard SI as measuring H subject to error, and thus the SI coefficient is proportionately reduced by $\gamma_4^2\sigma_H^2/(\gamma_4^2\sigma_H^2 + \sigma_w^2)$, the ratio of "signal" to "total" variance (net of M).

Now H and w are independent of t' (by construction), and also of u and t. So again SI factors out of the likelihood function; i.e., it's exogenous. The T equation in (IV.8) contains all of the exogenous variables and its residual is freely correlated with the Y residual. So LISE applied to the Y equation in (IV. 7) is FIML.

In the Y3 model, SI becomes endogenous and must be instrumented along with T. For we have

$$\text{(IV. 9)}\qquad \begin{aligned} Y &= SI\beta_1 + Tr_1 + u - r_1 t \\ T &= M\pi \qquad\qquad + H\gamma_3 + t \\ SI &= M\zeta \qquad\qquad + H\gamma_4 + w, \end{aligned}$$

and so SI is correlated with t if $\sigma_{tw} \neq 0$. The residual covariance matrix $\underset{\sim}{\Sigma}$ is

$$(\text{IV.10}) \quad \underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_u^2 + r_1^2\sigma_t^2 & -r_1\sigma_t^2 & -r_1\sigma_{tw} \\ & \sigma_t^2 + \gamma_3^2 & \sigma_{tw} + \gamma_3\gamma_4 \\ & & \sigma_w^2 + \gamma_4^2 \end{bmatrix}$$

(recall $\sigma_H^2 = 1$). Since this is unconstrained (except for inequality constraints) and since the T and SI equations are just identified, we can obtain FIML by applying LISE to the Y equation with SI and T endogenous. In fact this is just the estimator given in section III.

Our estimates for the first two models are shown in table 6. As expected, model Y1 gives a $\hat{\beta}_1$ close to the $b_{Y,SI \cdot M}$ estimate in table 2. But the test coefficient has increased by a factor of 9.4 over $b_{Y,T.M,SI}$ and by a factor of 3.2 over $b_{Y,T.SI}$. This reflects the low reliabilities: $\rho = .35$ and $\rho_N = .10$.

Table 6: Models Y1 and Y2.

	Coefficient (standard error of)	
	SI	G
Y1	.047 (.007)	.010 (.0008)
Y2	.020 (.008)	.0094 (.0009)

Table 7: Concentrated Likelihood for β_1 in Model YO3

β_1	.00	.02	.03	.04	.05	.071
L.R.	.01	.10	.23	.42	.66	1.00
χ^2	8.62	4.53	2.96	1.72	.82	.00

Note: L.R. = Likelihood ratio; χ^2 = -2LOG(L.R.) is approximately distributed as $\chi^2(1)$.

Table 8: M.L. Estimates for Model YO3

Dependent		Coefficient of			
Variable	COLOR	SI	G	H	
Y = LINC	.237	.071	.0099	...	σ_u^2 = .121
O = LOSES	.091	.183	.025	...	σ_v^2 = .313
T = AFQT	18.00	...	1.0	...	σ_t^2 = 419.7
SI	-.10	...	...	-.737	σ_w^2 = 1.209

Note: G = .091FO + .386FS + 4.523SB - 4.346ROS + 5.344H

$\hat{r}_{uv} = .136$, $\hat{r}_{tw} = .429$, $\hat{\rho} = .32$

POC, POS, AGE, AMS appear in all of the equations; FO, FS, SB, ROS enter SI unconstrained. H is normalized so that $\sigma_H^2 = 1$. The estimate of λ in (II.8) is .083.

Table 9: M.L. Estimates for Model YO3 with $\gamma_4 = 0$.

Dependent		Coefficient of			
Variable	COLOR	SI	G	H	
Y = LINC	.234	.049	.0096	...	σ_u^2 = .121
O = LOSES	.085	.129	.024	...	σ_v^2 = .313
T = AFQT	17.97	...	1.0	...	σ_t^2 = 405.0
SI	-.112	...	...	.00	σ_w^2 = 1.754

Note: G = .100FO + .458FS + 4.444SB - 4.256ROS + 6.528H

$\hat{r}_{uv} = .114$, $\hat{r}_{tw} = .215$, $\hat{\rho} = .34$

POC, POS, AGE, AMS appear in all of the equations; FO , FS, SB, ROS enter SI unconstrained. H is normalized so that $\sigma_H^2 = 1$. The estimate of λ in (II.8) is .071.

In model Y2 we have very similar results for r_1 but the β_1 estimate drops to .020 with a rather small standard error of .008. So a likelihood ratio (L.R.) test for $\gamma_4 = 0$ gives a very significant $\chi^2(1) = 60.6$. But the low reliabilities for T imply that much of the T distribution is being assigned to the residual t and calls into question the independence of t and w. Allowing for a σ_{tw} covariance leads to model Y3, which has already been given in table 5. There we have a rather high $\hat{\beta}_1 = .062$ and the t,w correlation is quite substantial: $\hat{r}_{tw} = .32$. But the concentrated likelihood function is quite flat and a L.R. test of $\sigma_{tw} = 0$ gives an insignificant $\chi^2(1) = 1.2$. Discriminating between the two models will require more information.

So we add the status equation:

$$\text{(IV.11)} \quad \begin{aligned} Y &= SI\beta_1 + Tr_1 + u - r_1 t \\ O &= SI\beta_2 + Tr_2 + v - r_2 t \\ T &= \underset{\sim}{M}\underset{\sim}{\pi} \quad + H\gamma_3 + t \\ SI &= \underset{\sim}{M}\underset{\sim}{\zeta} \quad + H\gamma_4 + w. \end{aligned}$$

Assuming that t and w are independent gives model YO2 and dropping that assumption gives YO3. In both models σ_{uv} is left unconstrained.

If the residual covariance matrix ($\underset{\sim}{\Sigma}$) were unconstrained, then FIML for YO2 could be obtained via Hannan's extension of LISE, treating SI as included exogenous, M as excluded exogenous, and T endogenous. For YO3 we would take T and SI as endogenous, obtaining the estimates given at the end of section III. But in fact $\underset{\sim}{\Sigma}$ is constrained:

(IV.12)
$$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_u^2 + r_1^2\sigma_t^2 & \sigma_{uv} + r_1 r_2 \sigma_t^2 & -r_1\sigma_t^2 & -r_1\sigma_{tw} \\ & \sigma_v^2 + r_2^2\sigma_t^2 & -r_2\sigma_t^2 & -r_2\sigma_{tw} \\ & & \sigma_t^2 + \gamma_3^2 & \sigma_{tw} + \gamma_3\gamma_4 \\ & & & \sigma_w^2 + \gamma_4^2 \end{bmatrix} .$$

So the upper right hand corner is constrained with

(IV.13) $\sigma_{13}/\sigma_{23} = r_1/r_2$ and $\sigma_{14}/\sigma_{24} = r_1/r_2$.

r_1 and r_2 can be obtained from the slopes, as in section III; thus we have two constraints in model YO3 and with $\sigma_{tw} = 0$ we have

(IV.14) $\sigma_{13}/\sigma_{23} = r_1/r_2$, $\sigma_{14} = \sigma_{24} = 0$

for three constraints.

The unrestricted $\underset{\sim}{\hat{\Sigma}}$ gives

(IV.15)
$$\begin{bmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \end{bmatrix} = \begin{bmatrix} -3.57 & -.013 \\ -11.20 & -.348 \end{bmatrix}.$$

with $r_1 = .009$, $r_2 = .026$ and $r_1/r_2 = .359$. So $\sigma_{13}/\sigma_{23} = .319$ is quite good but $\sigma_{14}/\sigma_{24} = .037$ seems terrible. In fitting the

Y-T-SI version (Y3), however, instead of $\sigma_{14} = -.013$ we obtained -.082, which is much closer to satisfying the constraint. This instability suggests that σ_{14} is not being estimated very precisely. To check this we constrained it at different values and found we could get to $\sigma_{14} = -.12$ with little decline in the likelihood and little change in σ_{24}. So it is not surprising that imposing both of the restrictions in (IV.13) gives an insignificant $\chi^2(2) = 2.16$. These constraints cannot be imposed with simple analytic techniques, and so we have used a general numerical minimization procedure adapted for such problems by K. Jöreskog. It is important to have reasonable starting values for the algorithm; fortunately our previous results provide very good ones. Details are given in the Appendix.

Model YO2 calls for $\sigma_{tw} = 0$. But the unconstrained $\sigma_{24} = -r_2\sigma_{tw} = -.348$, and it is quite stable for different values of σ_{14}. Imposing the restriction, while conditioning on the two restrictions in (IV.13), gives $\chi^2(1) = 4.88$ which is very surprising if σ_{tw} is really zero. So the non-zero correlation between t and w in YO3 ($\hat{r}_{tw} = .43$) is being estimated quite precisely. The concentrated likelihood function for β_1 in the YO3 model is given in table 7. At last we have reasonably strong information over the critical range from $\beta_1 = .01$ to .05. The M.L. estimate is .07 with $\gamma_4 = -.74$. But there is little evidence that γ_4 is in fact negative; constraining $\gamma_4 = 0$ gives $\chi^2(1) = .97$ and $\hat{\beta}_1 = .049$. Values of β_1 as low as .02 or .03, however, are quite strongly ruled out. Since model YO2 gives $\hat{\beta}_1 = .021$ (similar to Y2), it can be rejected.

There is an additional aggregation possibility, namely combining SI and G into a measure of late (post school) human capital. This would imply a proportionality restriction across SI and G in the Y and O equations: $\beta_1/\beta_2 = r_1/r_2$. With YO3 we get $\hat{\beta}_1/\hat{\beta}_2 = .38$, $\hat{r}_1/\hat{r}_2 = .37$ and with $\gamma_4 = 0$: $\hat{\beta}_1/\hat{\beta}_2 = .38$ $\hat{r}_1/\hat{r}_2 = .40$. The decline in likelihood from imposing the restriction is barely perceptable with $\chi^2(1) = .13$ and for γ_4 constrained to zero, $\chi^2(1) = .25$.[6/] It is shown in the next section that such proportionality constraints across late indicators cannot by themselves identify the model; but they do indicate the fruitfulness of the aggregation approach we've been following.

V. The Returns to SB and ST

This section examines the return to schooling before service (SB) and to total schooling (ST). We are interested in seeing whether our results could be obtained without the SB-SI split, in order to make comparisons with other samples which do not have this information. The approach directly parallel to ours would replace SI by ST and remove SB from M, leaving just the other background variables $\underset{\sim}{B}$ = (FO, FS, ROS). Using the Y1 model in this way gives essentially $\hat{\beta}_1 = b_{Y,ST.B}$, as we would expect. But both the Y2 and Y3 models give significantly <u>negative</u> estimates for the schooling coefficient! The reason for this striking conflict with the SI results is that we can no longer use T as a measure of "early" ability. For in the model

$$\text{(V.1)} \quad Y = ST\beta_1 + Gr_1 + u$$

$$T = \underset{\sim}{B}\underset{\sim}{\Theta} + H\gamma_3 + t$$

we must assume that SB does not affect T, given B and H. We did obtain estimates of an SB coefficient (η) in the previous section, but that was after reinterpreting H to be orthogonal to $\underset{\sim}{M} = (\underset{\sim}{B},SB)$. Thus, we were estimating $\eta + \gamma_3\ b_{H,SB.B}$, which could be positive even if $\eta = 0$. But now we have direct evidence that $\eta \neq 0$; for to reconcile the SI results with the peculiar ST results, we must assume that SB <u>does</u> have a value added in increasing T, so that T cannot be regarded as a measure of preschool ability.

In fact, it is better to regard T as a measure of post-school ability, although this is not strictly correct since SI intervenes between T and Y. Then we have

$$\text{(V.2)}\quad \begin{aligned} Y &= Gr_1 + u \\ T &= G \quad + t \\ G &= \underset{\sim}{B}\underset{\sim}{\theta} + ST\eta + H\gamma_3 \\ ST &= \underset{\sim}{B}\underset{\sim}{\mu} \qquad\quad + H\gamma_4 + w. \end{aligned}$$

We can estimate r_1 by substituting T for G in the Y equation and using $\underset{\sim}{B}$ and ST as instruments for T. This gives $\hat{r}_1$=.011, quite close to our previous estimates. But η is still not identified and neither is $\beta_1 = r_1\eta$. We conclude that late indicators alone cannot identify the model, at least not without replication within families.

So our methodology does not generalize to samples which do not specify the part of schooling received after the test. But we can still ask whether our conclusions generalize. In particular we find that $b_{Y,SI.B,SB}$ is not seriously biased upwards. Is this also true of $b_{Y,SB.B,SI}$, and hence of the average return $b_{Y,ST.B}$?

We will summarize the selection bias in $b_{Y,SB.B,SI}$ by adding the following equation to our SI models (eg., (IV.2)):

$$\text{(V.3)}\quad SB = \underset{\sim}{B}\underset{\sim}{\mu} + H\gamma_5 + w'.$$

This equation is also of interest because it suggests we can obtain more efficiency by using SB as an additional indicator for H. To check this, we could solve SB out of the T and SI equations to obtain a more fully "reduced" form; then allowing for correlation between t,w, and w' we could try and determine what parameters are identified and what, if any, is the efficiency gain. But there is a much simpler answer. For we have shown that the model can be transformed so that $\tilde{H}$ is uncorrelated with SB, and hence SB can be regarded as exogenous. So the H in (V.3) is uncorrelated with $\tilde{H}$, and the SB equation factors out of the likelihood function, without affecting our inferences on the other parameters.

Thus estimating γ_5 requires more information. If we assume that the return to SB is the same as for SI, then given the section IV estimates of β_1 and r_1, we can estimate η from $\beta_1 = \eta r_1$. Comparing this estimate with $\tilde{\eta} = \eta + \gamma_3 b_{H,SB.B}$ lets us obtain $\hat{\gamma}_5 = b_{SB,H.B}$. Using the YO3 model with γ_4 restricted to zero gives $\hat{\beta}_1 = .049$,and $r_1\hat{\tilde{\eta}} = .043$, implying a slight <u>downward</u> bias in the return to SB. Corresponding to this we find that normalizing $\sigma^{*}_{H.B} = 1$ implies a negative $\hat{\gamma}_5 = -.41$, but the point estimate is quite imprecise. Allowing for a declining marginal return to schooling, i.e., $r_1\eta > \beta_1$, would only make γ_5 more negative, as would using the YO3 estimates with γ_4 unrestricted. We conclude that given the measured background **variables that are** available, there is little evidence that important unmeasured characteristics are producing an upward bias in the SI <u>or</u> SB coefficients.

VI Summary and Extensions

This paper has tried to assess the value of some new methodology by applying it to a substantive empirical problem: the bias in income - schooling regressions caused by the omission of an early "ability" variable. A straightforward approach is to hold constant as many observable initial conditions as possible, and in our data there are several. But this can be inadequate for two reasons: the proxies may be measured with error and they may not include all of the relevant variables. We have used the test score as an example of each of these cases. In the first case we assume that the test adequately measures the initial conditions except for an error which could in principle be eliminated by replicating the test. This suggests bounds on the reliability of the test and within those bounds we find little bias in the schooling coefficient.

In the second case the test is only assumed to capture a part of the relevant initial conditions. Then we are trying to estimate the reliability of the test outside the context it was designed for and so fewer prior restrictions can be imposed. For example if much of the test distribution is assigned to the error, then the independence of that error and the schooling residual is implausible. So we have a negative prior covariance between the reliability of the test and that residual covariance.

In fact all of our models produce low reliabilities and so we try to obtain identification without constraining the schooling and test residuals to be uncorrelated. This is accomplished by meshing our covariance structure with the background coefficient restrictions suggested by Griliches and Mason. The resulting estimates give fairly strong evidence against a substantial bias in the schooling coefficient.

The models we use are extensions of the work by Zellner (1970) and Goldberger (1972) to a simultaneous equations context. Our general framework is a triangular structural model with factor analytic covariance restrictions. Many of the estimation problems can be handled by standard simultaneous equation techniques. However, our favored model (Y03) has restrictions across the slopes and residual covariances which cannot be imposed analytically. The restrictions are similar to those in the Jöreskog-Goldberger (1974) MIMIC model and the Appendix shows how to fit them into Jöreskog's (1970, 1973) class of covariance structures.

A general identification analysis of our class of models is given in Chapters 2 and 3. It is a specialization to triangular structures of the work by Geraci-Goldberger (1971) and Geraci (1974), but an extension in that part of the identification is coming from covariance restrictions. So it includes the Chamberlain-Griliches (1974) model, which did not have a test score but did have replication within families. The general analysis shows that the identification problem with that sort of replication is identical to having an additional indicator (e.g. a test) which is connected to the rest of the structure only via its dependence on the unobservable. Also we would like to know if we can allow the test and schooling errors (t and w) to be freely correlated, drop the restrictions on the background coefficients, and still obtain identification by having additional indicators which depend on schooling and the unobserved "ability". The answer in "no" (Chapter 2, Section III), but is is not obvious from a bare-hands inspection of the reduced form.

Appendix

This appendix gives some computational details on our use of Jöreskog's (1970, 1973) program ACOVSM. The general model assumes an N by p data matrix $\underset{\sim}{Z}$ with N observations on p variables and assumes that the rows of $\underset{\sim}{Z}$ are independently distributed, each having a multivariate normal distribution with the same variance-covariance matrix $\underset{\sim}{\Sigma}$. It is assumed that

$$E(\underset{\sim}{Z}) = \underset{\sim}{A}\underset{\sim}{\Xi}\underset{\sim}{P}$$

where $\underset{\sim}{A}$ and $\underset{\sim}{P}$ are known matrices and $\underset{\sim}{\Xi}$ is a matrix of parameters. $\underset{\sim}{\Sigma}$ has the form

$$\underset{\sim}{\Sigma} = \underset{\sim}{B}(\underset{\sim}{\Lambda}\underset{\sim}{\Phi}\underset{\sim}{\Lambda}' + \underset{\sim}{\Psi}^2)\underset{\sim}{B}' + \underset{\sim}{\Theta}^2,$$

where $\underset{\sim}{B}$, $\underset{\sim}{\Lambda}$, the symmetric matrix $\underset{\sim}{\Phi}$ and the diagonal matrices $\underset{\sim}{\Psi}$ and $\underset{\sim}{\Theta}$ are parameter matrices. Parameters can be assigned fixed values and groups of parameters can be constrained to be equal. However, parameters in $\underset{\sim}{\Xi}$ cannot be equated to parameters in $\underset{\sim}{\Sigma}$, a point we will return to below.

We have put our YO models in this form by first writing the SI equation as

$$SI = \underset{\sim}{X}_1 \underset{\sim}{\xi}_4 + \underset{\sim}{M}\underset{\sim}{\zeta}^* + G\gamma^*_4 + w$$

where $\gamma^*_4 = \gamma_4/\gamma_3$ and $\zeta^* = \zeta - \pi\gamma^*_4$. Then set $\underset{\sim}{P} = \underset{\sim}{I}$ and absorb $\underset{\sim}{X}_1 \underset{\sim}{\xi}_i$, $i = 1, \ldots 4$, $SI\beta_1$, $SI\beta_2$ and $\underset{\sim}{M}\underset{\sim}{\zeta}$ into $\underset{\sim}{A}\underset{\sim}{\Xi}$. This leaves

$$Y = Gr_1 + u$$

$$O = Gr_2 + v$$

$$T = G + t$$

$$SI = G\gamma^*_4 + w$$

$$G = \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3.$$

Then following Jöreskog and Goldberger (1974) we can write this as a second order factor model:

$$\underset{\sim}{Z}_i = \begin{bmatrix} \underset{\sim}{M}' \\ Y \\ O \\ T \\ SI \end{bmatrix}_i = \begin{bmatrix} \underset{\sim}{I}_4 & \underset{\sim}{o} & \underset{\sim}{O} \\ & r_1 & \\ \underset{\sim}{O} & r_2 & \underset{\sim}{I}_4 \\ & 1 & \\ & \gamma_4^* & \end{bmatrix} \begin{bmatrix} \underset{\sim}{M}' \\ G \\ u \\ v \\ t \\ w \end{bmatrix}_i$$

$$= \underset{\sim}{B}\underset{\sim}{f}_{1\,i}$$

and

$$\underset{\sim}{f}_{1\,i} = \begin{bmatrix} \underset{\sim}{M}' \\ G \\ u \\ v \\ t \\ w \end{bmatrix}_i = \begin{bmatrix} \underset{\sim}{I}_4 & o & \underset{\sim}{O} \\ \underset{\sim}{\pi}' & \gamma_3 & \\ \underset{\sim}{O} & & \underset{\sim}{I}_4 \end{bmatrix} \begin{bmatrix} \underset{\sim}{M}' \\ H \\ u \\ v \\ t \\ w \end{bmatrix}_i$$

$$= \underset{\sim}{\Lambda}\underset{\sim}{f}_{2i},\ i = 1, \ldots, N.$$

This defines $\underset{\sim}{B}$ and $\underset{\sim}{\Lambda}$, and we set $\underset{\sim}{\Psi} = \underset{\sim}{\Theta} = (\underset{\sim}{O})$ and

$$\underset{\sim}{\Phi} = E(\underset{\sim}{f}_{2i}\ \underset{\sim}{f}'_{2i})$$

$$= \begin{bmatrix} \underset{\sim}{\Phi}_m & & & & & \\ & 1 & & & & \\ & & \sigma_u^2 & & & \\ & & \sigma_{uv} & \sigma_v^2 & & \\ & \underset{\sim}{O} & & & \sigma_t^2 & \\ & & & & \sigma_{tw} & \sigma_w^2 \end{bmatrix}$$

where $\underset{\sim}{\Phi}_m$ is constrained to equal the sample covariance matrix of $\underset{\sim}{M}$. For YO2 we constrain $\sigma_{tw} = 0$; for YO3 we leave it free.

It may seem odd to include SI in the design matrix $\underset{\sim}{A}$ since SI is endogenous. But since the model has a triangular structure there is no Jacobian and the program is maximizing the correct likelihood function. However, the information matrix approximation is not correct. It is block diagonal in $\underset{\sim}{\Xi}$ and $\underset{\sim}{\Sigma}$ when in fact the ML estimates of the β's and $\underset{\sim}{\Sigma}$ are correlated. The necessary correction to the information matrix is analagous to the difference between the reduced and structural form information matrices in the appendix to Chamberlain and Griliches (1974).

The proportionality constraint across SI and G at the end of section IV can be imposed as follows: absorb $\underset{\sim}{X}_1\underset{\sim}{\xi}_i$, i = 1, ...4, and $\underset{\sim}{M}\underset{\sim}{\zeta}$ into $\underset{\sim}{A}\underset{\sim}{\Xi}$, leaving

$$\begin{aligned} Y &= (G + SI\beta)r_1 + u \\ O &= (G + SI\beta)r_2 + v \\ T &= G + t \\ G &= \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3 \\ SI &= H\gamma_4 + w. \end{aligned}$$

This can be written as the following second order factor model:

$$\underset{\sim}{Z}_i = \begin{bmatrix} \underset{\sim}{M}' \\ Y \\ O \\ T \\ SI \end{bmatrix}_i = \begin{bmatrix} \underset{\sim}{I}_4 & & \underset{\sim}{0} & \\ & r_1 & r_1 & \\ \underset{\sim}{0} & r_2 & r_2 & \underset{\sim}{I}_3 \\ & 1 & 0 & \\ & 0 & 1/\beta & \underset{\sim}{0} \end{bmatrix} \begin{bmatrix} \underset{\sim}{M}' \\ G \\ SI\beta \\ u \\ v \\ t \end{bmatrix}_i$$

$$= \underset{\sim}{B}\underset{\sim}{f}_{1i},$$

and

$$\underset{\sim}{f}_i = \begin{bmatrix} \underset{\sim}{M}' \\ G \\ SI\beta \\ u \\ v \\ t \end{bmatrix}_i = \begin{bmatrix} \underset{\sim}{I}_4 & & \underset{\sim}{0} & \\ \underset{\sim}{\pi}' & \gamma_3 & & \\ & \gamma_4\beta & 0 & 1 \\ \underset{\sim}{0} & & I_3 & \underset{\sim}{0} \end{bmatrix} \begin{bmatrix} \underset{\sim}{M}' \\ H \\ u \\ v \\ t \\ w\beta \end{bmatrix}_i$$

$$= \underset{\sim}{\Lambda}\underset{\sim}{f}_{2i},\ i = 1, \ldots, N.$$

Then set $\underset{\sim}{\Psi} = \underset{\sim}{\Theta} = (\underset{\sim}{0})$ and

$$\Phi = E\,(\underset{\sim}{f}_{2i}\ \underset{\sim}{f}'_{2i})$$

$$= \begin{bmatrix} \underset{\sim}{\Phi}_m & & & & & \\ & 1 & & & & \\ & & \sigma_u^2 & & & \\ & & \sigma_{uv} & \sigma_v^2 & & \\ & \underset{\sim}{0} & & & \sigma_t^2 & \\ & & & & \beta\sigma_{tw} & \beta^2\sigma_w^2 \end{bmatrix} .$$

So by setting $\beta^* = 1/\beta$, $\sigma^*_{tw} = \beta\sigma_{tw}$, and $\sigma^{*2}_w = \beta^2\sigma^2_w$, we can write the model in the (A.2) form. There are two equality constraints on $\underset{\sim}{B}$ and $\underset{\sim}{\Phi}_m$ is constrained to be the sample covariance matrix of $\underset{\sim}{M}$.

Our experience with the program has suggested two modifications. First is the need for a more accurate matrix inversion routine. The program could not invert the information matrix at the maximum because the triangular factorization routine produced a negative diagonal element due to round-off error. But direct inspection of the concentrated likelihood function in table 6 shows that at least β_1 is being estimated quite precisely. So perhaps the program should just produce the variance-covariance matrix of the (numerically) estimable functions. A related problem is the choice of an initial approximation to the inverse of the Hessian matrix. When the information matrix cannot be inverted at the initial parameter values, the program substitutes an identity matrix. This results in a much costlier problem since the Fletcher-Powell iterations have to build up the inverse of the Hessian from scratch. It would seem preferable to keep as much of the information matrix as possible, say by setting negative diagonal elements in the triangular factorization to a small positive number.

The second modification would allow constraints across $\underset{\sim}{\Xi}$ and $\underset{\sim}{\Sigma}$. This would be useful because the likelihood function is maximized analytically with respect to $\underset{\sim}{\Xi}$ conditional on $\underset{\sim}{\Sigma}$. Then the numerical problem is to maximize the concentrated likelihood function over $\underset{\sim}{\Sigma}$. So it is desirable to put as many of the parameters as possible into $\underset{\sim}{\Xi}$. Consider, for example, the Jöreskog-Goldberger MIMIC model:

$$y_1 = \beta_1 y^* + u_1$$
$$\vdots$$
$$y_m = \beta_m y^* + u_m$$
$$y^* = \underset{\sim}{\alpha}'\underset{\sim}{x} + \varepsilon \quad .$$

If $\beta_m \neq 0$ we can scale y^* so that $\beta_m = 1$ and write the model as

$$\begin{aligned} y_1 &= \beta_1 y_m + u_1 - \beta_1 u_m \\ &\cdot \\ &\cdot \\ &\cdot \\ y_{m-1} &= \beta_{m-1} y_m + u_{m-1} - \beta_{m-1} u_m \\ y_m &= \underset{\sim}{\alpha}'\underset{\sim}{x} + \varepsilon + u_m . \end{aligned}$$

Given the triangular structure we can absorb $\beta_i y_m$, $i = 1, \ldots, m - 1$ and $\underset{\sim}{\alpha}'\underset{\sim}{x}$ into $\underset{\sim}{A}\underset{\sim}{\Xi}$ leaving

$$\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix}_i = \begin{bmatrix} \underset{\sim}{I}_{m-1} & \begin{matrix} -\beta_1 \\ \cdot \\ \cdot \\ \cdot \\ -\beta_{m-1} \end{matrix} & \underset{\sim}{O} \\ \underset{\sim}{O} & 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_{m-1} \\ u_m \\ \varepsilon \end{bmatrix}_i$$

$$= \underset{\sim}{\Lambda}\underset{\sim}{f}_i$$

with

$$\underset{\sim}{\Phi} = E(\underset{\sim}{f}_i \underset{\sim}{f}_i') = \begin{bmatrix} \sigma_1^2 & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \sigma_m^2 & \\ & \underset{\sim}{O} & & & \sigma_\varepsilon^2 \end{bmatrix}$$

and $\underset{\sim}{B} = \underset{\sim}{I}$, $\underset{\sim}{\Psi} = \underset{\sim}{\Theta} = (\underset{\sim}{O})$.

Thus a first order factor structure will do, provided we can impose equality constraints across $\underset{\sim}{\Lambda}$ and $\underset{\sim}{\Xi}$. The advantage in this formulation of the problem is that now $\underset{\sim}{\alpha}$ can be analytically concentrated out of the likelihood function.

The analagous way of writing our YO model is given in (IV.16) and (IV.17). This is the easiest way to display the nature of the constraints. It is also a natural way to estimate the model, since $SI\beta_i$, Tr_i, i = 1, 2, $\underset{\sim}{\pi}$, $\underset{\sim}{\zeta}$, and $\underset{\sim}{\xi_i}$, i = 1, ..., 4, can all be put in $\underset{\sim}{\Xi}$. Then $\underset{\sim}{\Sigma}$ can be modeled as

$$\begin{bmatrix} Y \\ O \\ T \\ SI \end{bmatrix}_i = \begin{bmatrix} 1 & 0 & -r_1 & \underset{\sim}{0} & \\ 0 & 1 & -r_2 & & \\ & & 1 & 0 & \gamma_3 \\ & \underset{\sim}{0} & 0 & 1 & \gamma_4 \end{bmatrix} \begin{bmatrix} u \\ v \\ t \\ w \\ H \end{bmatrix}_i$$

$$= \underset{\sim}{\Lambda}\underset{\sim}{f_i}$$

with

$$\underset{\sim}{\Phi} = \begin{bmatrix} \sigma_u^2 & & & & \\ \sigma_{uv} & \sigma_v^2 & & & \\ & & \sigma_t^2 & & \\ & \underset{\sim}{0} & \sigma_{tw} & \sigma_t^2 & \\ & & 0 & 0 & 1 \end{bmatrix} .$$

So we can set $\underset{\sim}{B} = \underset{\sim}{I}$, $\underset{\sim}{\Psi} = \underset{\sim}{\Theta} = (\underset{\sim}{0})$, and just use a first order factor structure. There are, however, equality constraints across $\underset{\sim}{\Lambda}$ and $\underset{\sim}{\Xi}$. The advantage is that $\underset{\sim}{\pi}$ is analytically concentrated out of the likelihood function.

Footnotes

1/ I am indebted to Bronwyn Hall for computational assistance and to Zvi Griliches and Edward Leamer for helpful comments.

2/ Including POC and POS in the set of constrained background characteristics has little effect on the results.

3/ Since the test score is a percentile the assumption that T is normally distributed (conditional on $\underset{\sim}{X}$) is questionable. However, rescaling the test scores to have a normal distribution did not affect the results.

4/ This model has been considered by S. Cardell and M. Hopkins (unpublished manuscript, Harvard University).

5/ This method does not, however, fully utilize the sample information in Y when constructing $\hat{T}$. The reduced form

$$\begin{aligned} Y &= SI\beta_1 + \underset{\sim}{M}\underset{\sim}{\pi} r_1 + H\gamma_3 r_1 + u \\ T &= \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3 + t \end{aligned}$$

makes it clear that both T and Y contain information on $\underset{\sim}{\pi}$ (provided $\underset{\sim}{M}$ contains more than one variable). We can impose this proportionality restriction by doing a one dimensional search. Constrain $\beta_1 = \beta_1^0$ and let $Y^0 = Y - SI\beta_1^0$. Then rewrite the system as

$$\begin{aligned} Y^0 &= Tr_1 + u - r_1 t \\ T &= \underset{\sim}{M}\underset{\sim}{\pi} + H\gamma_3 + t \\ SI &= \underset{\sim}{M}\underset{\sim}{\zeta} + w. \end{aligned}$$

Note that the SI equation factors out of the likelihood function and that the Y^0 and T residuals are freely correlated. So LISE on the Y^0 equation is FIML and varying β_1^0 lets us plot a concentrated likelihood function for β_1.

6/ Details on writing the restriction in Jöreskog's framework are given in the Appendix.

Chapter 5

Returns to Schooling of Brothers and Ability As an Unobservable Variance Component*

I. Introduction

In earlier papers Griliches (1970 and 1972) investigated the bias in estimates of returns to schooling due to the omission of an ability measure from the estimating relation. Another controversial source of bias is the possible direct influence of parental background (economic, social class, and ethnic) on subsequent economic achievement (income and occupation), above and beyond its indirect effect via schooling. One way to hold both parental background and some of the ability differences constant is to analyze the economic experience of brothers. Brothers have largely similar family economic and motivational backgrounds and also differ less in native ability. It is the purpose of this paper to report on a reanalysis of a rather old set of such data and to develop a somewhat novel methodology for the analysis of this kind of problem.

*This Chapter was written jointly with Zvi Griliches. An abridged version appears in Chamberlain and Griliches (1974). We are indebted to Ruth Helpman for research assistance and to NSF Grant No. GS 2762X for financial support.

The next section of the paper outlines the content and source of our data and presents the results of a straightforward covariance analysis of them. In the third section we develop a more explicit model in which ability (and parental background) is a left out variable having a differential within and between family (variance-components) structure. We discuss the question of identification in such a model and outline a maximum likelihood estimation procedure for this model. The final sections of the paper present the results of applying this model to our data, discuss tests of the model, and suggest some extensions. The estimation procedure is presented in greater detail in Chapter 3, Appendix A.

2. A Reanalysis of the Gorseline Data.

One of the first consistent and detailed analyses of the "ability bias" issue can be found in Gorseline's (1932) book, written in the late 1920's. He set out to solve the ability-schooling conundrum through the collection of data on income, schooling, and other characteristics of brothers. He managed to collect such data for about 172 sets of brothers or 368 individuals. Using rather primitive but reasonable methods of analysis (comparing the mean income of brothers with more schooling to the mean of those with less) he concluded that indeed schooling did pay, even holding family background constant. He did not use, however, his data to estimate how much the usual measure of return (not holding parental background constant) is biased upward. The major facts about his sample are presented in Table 1 and the derivation of the variables is described more fully in Appendix B.

Since he published almost all of his data, we decided to reanalyze them with the above question in mind. The procedure used was first to estimate an income-schooling relationship across all individuals in the sample ignoring the familial information and then compare it with estimates in which each brother's characteristic (his income, schooling, age, etc.) are measured around his own family's mean. This procedure eliminates from the relationship both the common influence of parental background and the common part of their genetically inherited "abilities". It holds constant, as well as it could ever be done, the "parental background" or "social class" effects in such relationships. The results of this reanalysis, limited in this paper to the sub-sample of 156 pairs of brothers, are summarized in Table 2. They show clearly that at least in 1927, in Indiana, differences in parental background were not an important source of bias in the estimated returns to (the coefficient of) schooling.[1] This does not mean that parental background does not account for a significant fraction of the total variance in income. In fact,

[1] Additional analyses of the data using the rate at which schooling was completed as a measure of ability and allowing for the birth-order of brothers did not change this conclusion significantly.

Table 1: Characteristics of the Gorseline Sample

Brothers	Number of sets
2	156
3	9
4	6
5	1
Total in sample	368

Variables	Means	Standard Deviations Total	Within Families
S - Schooling (Grade attained)	11.64	3.47	2.14
YL - Log Income, 1927	7.53	.688	.386
OL - Log Occupation SES	3.63	.699	.500
AGE -	36.45	10.8	3.7
EXP - Experience (Age-Age stopped school)	17.02	12.1	5.0

Source: D.E. Gorseline, The Effect of Schooling Upon Income, Indiana University, 1932). Occupation scored according to Duncan's SES scale. N = 368; sets of brothers = 172.

Table 2: Gorseline Data Regressions

Dependent Variable	Coefficients of							
	S	EXP	$(EXP)^2$	Age	$(Age)^2$	OSESL	R^2	SEE
YL	.120 (.012)	.050 (.009)	-.0006 (.0002)				.260	.603
	.077 (.013)	.052 (.009)	-.0007 (.0002)			.373 (.053)	.363	.560
	.082 (.010)			.088 (.021)	-.0008 (.0003)		.259	.604
	.044 (.011)			.083 (.020)	-.0008 (.0003)	.367 (.052)	.361	.561
YLD								
	.109 (.010)	.018 (.009)	.0001 (.0002)				.277	.317
	.085 (.012)	.018 (.009)	.0000 (.0002)			.155 (.045)	.304	.311
	.080 (.008)			.061 (.022)	-.0004 (.0003)		.326	.306
	.059 (.010)			.052 (.021)	-.0003 (.0003)	.155 (.043)	.353	.300
OSESL								
	.104 (.010)			.005 (.003)			.250	.608
OSESLD								
	.135 (.010)			.011 (.006)			.352	.400

YL - Log Income, 1927
S - Schooling, Grade attained
Exp - (Age-Age stopped school)
OSESL - Log current (1928) Occup. SES

Variables with D suffix and all the variables in regressions with YLD or OSESLD as dependent variables are measured around family means. N = 312; pairs of brothers = 156.

the total variance in the logarithms of income is reduced from .47 in the sample at large to .15 between brothers only. This reduction, however, is due not only to the elimination of parental background, but also to the elimination of all other characteristics, such as rural versus urban location or age, which are common to pairs of brothers. In any case, the estimate of the marginal effect of schooling does not appear to be biased when such effects are ignored. This rather surprising result lead us to reconsider whether our expectation that holding family background constant should have reduced the estimated schooling coefficient is indeed warranted. To do so we have to spell out the underlying model in some detail. Let the true income relationship be

(1) $y_{ij} = \beta S_{ij} + \gamma A_{ij} + u_{ij}$

where y is the logarithm of income, S is the highest grade of schooling attained and A is an unobserved measure of an individual's background such as his social class and IQ. The index i stands for families, while j runs over individuals within a family; u_{ij} is a random variable unrelated to either S or A; and all variables are measured around their total sample means, obviating the necessity of writing down constants in the various equations of the model.

Now, the reason why there may be a bias arises from the assumed positive correlation between A and S. Let that correlation be summarized by equation 2.

$$(2)\quad S_{ij} = \eta A_{ij} + w_{ij}$$

where w_{ij} is assumed to be distributed independently of A. To complete the model we specify a variance-components structure for the "ability" variable:

$$(3)\quad A_{ij} = F_i + G_{ij}$$

where F_i is the common family component and G_{ij} is independent of F_i by construction.

We could and do add another set of variables, X's to these equations, but unless they impose additional constraints on the data via additional exclusion restrictions, we just interpret S and Y as deviations from regressions including these X's and proceed as above, ignoring them for purposes of this analysis.

The basic assumptions up to this point are (a) that the left-out determinants of schooling, A, have also an additional direct effect on y (as against w which has only an indirect one) and (b) that these effects have a family (variance-components) structure: $A_{ij} = F_i + G_{ij}$.

To get explicit and simple formulae for the bias in the simple least squares regression coefficient of y on S as an estimator of β, we shall consider large samples both in the i and j dimension, so that we can identify sample moments with the underlying population parameters.[2] It is also convenient to write down the "reduced form" equation for y, by substituting (3) and (2) into (1):

$$(4)\quad y = (\beta\eta + \gamma)(F + G) + \beta w + u$$

The least squares coefficient of y on S is given by

$$\text{plim } b \;= \text{plim } \frac{\text{Cov } yS}{\text{Var } S} = \beta + \gamma \frac{\text{Cov } AS}{\text{Var } S} = \beta + \gamma \frac{\eta \text{ Var } A}{\eta^2 \text{Var } A + \text{Var } w} .$$

Similarly, consider the deviations based estimator b_{yDSD}, with family effects taken (swept) out from the data

$$\text{plim } b_{yDSD} = \beta + \gamma \frac{\text{Cov } G \text{ SD}}{\text{Var SD}} = \beta + \gamma \frac{\eta \text{Var } G}{\eta^2 \text{Var } G + \text{Var } w} .$$

Now, define Var G/Var A = $1-\lambda$ and Var w/η^2Var A = $(1-R^2)/R^2=U$, and concentrate attention on the bias = (plim b − β) of these coefficients:

$$\text{bias } b_{yDSD} = \frac{\gamma}{\eta}\,\frac{1-\lambda}{1-\lambda+U} = \frac{\gamma}{\eta}\,\frac{1}{1 + U/(1-\lambda)}$$

versus

$$\text{bias } b_{yS} = \frac{\gamma}{\eta}\,\frac{1}{1 + U} .$$

Since $0 < \lambda < 1$, the absolute bias in the coefficient estimated from deviations from family means (D) will be smaller

[2] Having a large sample over j implies in our case a large number of brothers per family. This is unnecessary but it simplifies the notation of this section.

than in the coefficient based on the whole set of data. The bias would be nil if γ were zero, i.e. no direct effect of ability or family background on income, and need not be zero but would remain essentially unchanged by the transformation of the data to deviations form if either λ is zero (i.e. there is no family structure to the ability variable) or U is zero, there is no exogeneous component to the schooling variable and hence no distinction can be made between the effects of A and S. Both of the latter possibilities are unlikely.

While we don't know the absolute size of the bias, the expected relative reduction in its size from going to deviations is given by

$$\frac{\text{bias } b_{yDSD}}{\text{bias } b_{yS}} = \frac{(1-\lambda)(1+U)}{1-\lambda + U} = \frac{1+U}{1+U/(1-\lambda)} \cdot$$

It depends on both λ and U. The larger is λ, i.e. the larger is the "family" component in the total variance of ability, and the larger is U, the less is the role of "ability" in the total variance of schooling, the larger will be the reduction in the bias as we move to within family data. But for reasonable values of U and λ this reduction is not that large. λ is the ratio of the variance of family components to the total family background and ability variance.

Its maximum value is probably 0.8 and it is unlikely to fall much below 0.5.[3] At the same time U, the relative ratio of the independent (of family background and market rewarded ability) variance component of schooling is unlikely to exceed unity (implying that half of the variance of schooling is independent of family and individual ability components) or fall much below a third (at least a quarter of the variance of schooling is likely to be unrelated to both socio-economic background or IQ). Putting these two ranges together, implies a bias ratio between .5 and .8. Considering the a **priori** reasonable values of $\lambda = .6$ and $U = 1$, yields a bias ratio of about .6. That is, going from b_{yS} to b_{yDSD} will reduee the "ability" bias by only 40 percent. Since the actual coefficients change in Table 2 only from .120 to .109, (for the version with experience and experience squared) the total bias could be on the order of .028, or about 23 percent of the originally estimated coefficient, which would be consistent with other studies of this subject. Using age instead of experience in the equation produces a much smaller estimate of this bias, since the estimated schooling coefficients change only from .082 to .080.

[3] For any finite set of data, the within variances will not equal their population values even approximately, but rather (p-1)/p times that value, where p is the number of family members per family. In our case, with most of the data being on pairs of brothers, p=2, and the estimated within variances are too small by a half. But in the formula discussed in the text, this cancels out, since taking out the family "mean" effects affects both the numerator and denominator of b_{yDSD} equally (alternatively, the estimated σ_w^2 is too small by the same proportion as σ_G^2.)

Actually we observe in Table 2, occasionally an <u>increase</u> in the estimated coefficient of schooling as we move to the within-families data set (particularly for the occupation dependent form, which we haven't discussed yet). Since our model predicts a <u>decline</u> in the <u>absolute</u> value of the bias, this may be an indication that we originally <u>under</u> estimated rather than <u>over</u> estimated β, implying that γ is not only small but <u>negative</u> (we can always set $\eta = 1$ since the units in which A is measured are to some extent arbitrary). This may not be as surprising as it appears at first sight. It is conceivable that family wealth and "learning" ability lead to an over-investment in schooling and to a negative return to such an "ability" when the attained schooling level is held constant. We shall return to this below.

The results of this section are quite unsatisfactory. Limiting ourselves to within-families data resulted in little change in our conclusions and a realization that not much could be said, in fact, on the basis of such an analysis. The model as written down is not adequately identified. We got some qualitative conclusions by adding the prior information that λ, η, and U are all larger than zero and imposing some bounds on the likely values of λ and U. But to identify the coefficient of interest (β) further and to get explicit estimates of some of the other parameters, we have to expand the model and bring in additional variables, relations, and restrictions.

3. Ability as Unobservable

While the calculations reported above "take care" of parental background differences, even though inefficiently (they ignore the between families information in the sample), they do not correct for possible bias from the individual (within family) genetic differences which may be correlated with achieved schooling levels later on. To take this explicitly into account would require the availability of direct measures of such ability, which are not available for this set of data. But even in their absence, if the missing variable (such as ability) affects more than one dependent variable, a bootstrap operation may be possible. The basic idea for the new approach comes from the realization that such a left out variable must cause similar biases (proportional to each other) in different equations and that taking advantage of that fact may allow one to achieve identification of most of the coefficients of interest.

A general version of our model is given by:

$$y_k = \underset{\sim}{X}\, \underset{\sim}{\alpha}_k + \beta_k y_s + \gamma_k a + u_k$$

$$y_s = \underset{\sim}{X}\, \underset{\sim}{\alpha}_s \qquad + \gamma_s a + w$$

$$a_{ij} = f_i + g_{ij}$$

Where there are K dependent variables (indicators) which all depend on schooling (y_s), independent variables $\underset{\sim}{X}$ (which may differ from equation to equation), and on a left-out random ability variable (a) which affects both y_s and the y_k's, making y_s endogeneous, and has a peculiar structure (a = g + f) which converts this into a variance-components problem, observations being available for p members (index j) in each of q (index i) families. Without the "a" variable, or if $\gamma_k = 0$, and given our assumptions about the independence of u_k from w and a, this would just be a simple recursive system which could be estimated by applying least squares separately to each equation. The simultaneity problem arises when we admit the possibility that $\gamma_k \neq 0$. In general, if there were enough exogeneous variables in the schooling equation which did not appear again in the y_k equations, the endogeneity of y_s problem could be solved using two-stage least squares or other standard simultaneous equations estimation procedures. In our problem, however, the y_s equation will in general not contain enough distinct $\underset{\sim}{X}$'s for the identification of the β's. Instead, we shall have to rely on restrictions that the model imposes on the variance-covariance matrix of the residuals from the reduced form equations. These equations can be written as follows:

$$y_k = \underset{\sim}{X}(\underset{\sim}{\alpha}_k + \beta_k \underset{\sim}{\alpha}_s) + [(\gamma_k + \beta_k \gamma_s)(f + g) + u_k + \beta_k w]$$

$$y_s = \underset{\sim}{X}\, \underset{\sim}{\alpha}_s + [\gamma_s (f + g) + w]$$

where for a particular k, say k = 2, α_2 and α_s are vectors while β_2 is a scalar. The bracketed terms are the reduced form disturbances. More concisely, we can stack the observations and equations and relable the whole system as one multivariate regression:

$$\underset{\sim}{y} = \underset{\sim}{Z}\underset{\sim}{\delta} + \underset{\sim}{\epsilon}$$

where $\underset{\sim}{y}$ runs over all the dependent variables and families and family members and $\underset{\sim}{Z}$ includes all the $\underset{\sim}{X}$'s in all the equations. The variance-covariance matrix of the reduced form disturbances is $E\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}' = \underset{\sim}{I}_q \otimes \underset{\sim}{\Omega}$.

It is clear, that the model together with the assumptions of no correlation among a's, u_k's and w's imposes a number of constraints on the variance-covariance matrix of computed residuals from the regressions of y_k and y_s on $\underset{\sim}{Z}$. It can be shown that $\underset{\sim}{\Omega}$ equals

$$\underset{\sim}{\Omega} = \underset{\sim}{d}\underset{\sim}{d}' \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' + \underset{\sim}{\Sigma} \otimes \underset{\sim}{I}_p$$

where, specializing to the case of K = 2, with y_2 equaling an index of occupational achievement (the logarithm of Duncan's SES occupational score), y_1 = log earnings, and y_s = S = highest grade of schooling attained, we have:

$$\underset{\sim}{d} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \sigma_f \begin{pmatrix} \gamma_1 + \beta_1\gamma_3 \\ \gamma_2 + \beta_2\gamma_3 \\ \gamma_3 \end{pmatrix} \qquad \underset{\sim}{\ell}_p = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\underset{\sim}{\Sigma} = \tau\, \underset{\sim}{d}\underset{\sim}{d}' + \underset{\sim}{V}$$

$$\tau = \sigma_g^2 / \sigma_f^2$$

where p is the number of observations across index j within each i, i.e. in our case the number of family members per family, and

$$\underset{\sim}{V} = \begin{pmatrix} \beta_1^2\sigma_w^2 + \sigma_{u_1}^2 & \beta_1\beta_2\sigma_w^2 & \beta_1\sigma_w^2 \\ & \beta_2^2\sigma_w^2 + \sigma_{u_2}^2 & \beta_2\sigma_w^2 \\ & & \sigma_w^2 \end{pmatrix} .$$

Thus, $\underset{\sim}{d}\underset{\sim}{d}'$ represents the contribution of the family component of the unobservable to $\underset{\sim}{\Omega}$, $\tau\underset{\sim}{d}\underset{\sim}{d}'$ the contribution of the individual (within family) component of the unobservable, while $\underset{\sim}{V}$ encompasses the rest of the within-family individual effects, the u's and w. It is clear that if we could estimate $\underset{\sim}{V}$ directly, we could easily identify the β's since, for example, $\beta_1 = v_{13}/v_{33}$. We cannot do that, but we can estimate $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ and it turns out that in the two indicators case we can solve uniquely for τ and β_1 and β_2. The γ's however, can be estimated only up to a scale factor, since the latter cannot be separated from the arbitrary scale of the a's themselves. A model with more than two indicator variables will in general be overidentified. Some of these overidentifying restrictions could be traded off for relaxing some of the more stringent other assumptions, such as the no correlation assumption between u_1 and u_2. We can show that knowing $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ identifies the structural parameters as follows:

For a given value of $\tau = \sigma_g^2/\sigma_\varepsilon^2$ we can solve for

$$\sigma_w^2 = \sigma_{33} - \tau d_3^2$$

$$\beta_1 = (\sigma_{13} - \tau d_1 d_3)/\sigma_w^2$$

$$\beta_2 = (\sigma_{23} - \tau d_2 d_3)/\sigma_w^2 \quad .$$

But we can also solve for

$$\beta_1\beta_2 = (\sigma_{12} - \tau d_1 d_2)/\sigma_w^2 .$$

τ is determined by making the separate solutions for β_1 and β_2 agree with the solution for their product. This yields:

$$\tau = (\sigma_{13}\,\sigma_{23} - \sigma_{12}\,\sigma_{33})/(\sigma_{13}d_2d_3 + \sigma_{23}d_1d_3 - \sigma_{12}d_3^2 - \sigma_{33}d_1d_2) .$$

So $\tau = \sigma_g^2/\sigma_f^2$ is identified and hence also β_1 and β_2.

The problem then becomes one of estimating $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$. That $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ are in fact estimable, albeit inefficiently, can be seen most quickly by considering estimates based on the "method of moments". Let $\underset{\sim}{R}$ be the matrix of the variances and covariances of the residuals from the reduced form equations estimated by ordinary least squares, and let $\underset{\sim}{\bar{R}}$ be the matrix of variances and covariances of _average_ residuals, averaged separately over each family and variable within family. It is obvious then that

$$\text{plim } \underset{\sim}{R} = \underset{\sim}{\Sigma} + \underset{\sim}{d}\underset{\sim}{d}' = \underset{\sim}{\Sigma} + \underset{\sim}{\Theta}$$

$$\text{plim } \underset{\sim}{\bar{R}} = \frac{1}{p}\underset{\sim}{\Sigma} + \underset{\sim}{d}\underset{\sim}{d}' = \frac{1}{p}\underset{\sim}{\Sigma} + \underset{\sim}{\Theta}$$

where $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ and p is the number of individuals within each family. (We are assuming, for simplicity of exposition, that it is the same across families). It is obvious, that if p where large, $\underset{\sim}{\bar{R}}$ would be a direct estimate of $\underset{\sim}{\Theta}$. Since in our sample p is quite small, mostly p = 2, we get estimates of $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$ as follows:

$$\hat{\underset{\sim}{\Sigma}} = [p/(p-1)]\underset{\sim}{W}$$

$$\hat{\underset{\sim}{\Theta}} = \frac{p}{p-1}\,(\underset{\sim}{\bar{R}} - \frac{1}{p}\,\underset{\sim}{R})$$

where $\underset{\sim}{W} = \underset{\sim}{R} - \underset{\sim}{\bar{R}}$ is the "within" families variance-covariance matrix of the sample residuals. Thus both $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ and $\underset{\sim}{\Sigma}$ are estimable from the sample. But now, when we substitute these expressions in the earlier formulae for β_1, β_2, and $\beta_1\beta_2$, the formula for τ does not simplify as easily, but rather leads to a quadratic equation:

$$\tau^2(\Theta_{13}\,\Theta_{23} - \Theta_{12}\,\Theta_{33}) + \tau(\Theta_{12}\,\sigma_{33} + \Theta_{33}\sigma_{12} - \Theta_{13}\sigma_{23} - \Theta_{23}\sigma_{13})$$

$$+\ \sigma_{13}\,\sigma_{23} - \sigma_{12}\,\sigma_{33} = 0\,.$$

The quadratic term doesn't vanish, since we haven't imposed the condition $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ which implies

$$\Theta_{13}\Theta_{23} - \Theta_{12}\Theta_{33} = d_1d_3d_2d_3 - d_1d_2d_3^3 = 0.$$

Rewriting this equation in terms of observables [substituting $\frac{p}{p-1}\,\underset{\sim}{W}$ for $\underset{\sim}{\Sigma}$ and $\frac{p}{p-1}\,(\underset{\sim}{\bar{R}} - \frac{1}{p}\underset{\sim}{R})$ for $\underset{\sim}{\Theta}$] and reparameterizing it in terms of $\lambda = \frac{p-1}{p}\,\sigma_g^2/\sigma_a^2 = \frac{p-1}{p}\;\frac{\tau}{\tau+1}$, leads to:

$$\lambda^2(R_{13}R_{23} - R_{12}R_{33}) + \lambda(R_{12}W_{33} + R_{33}W_{12} - R_{13}W_{23} - R_{23}W_{13})$$

$$+ W_{13}W_{23} - W_{12}W_{33} = 0$$

and two solutions (roots) for λ (or τ). Since $0 < \lambda < \frac{p-1}{p}$ hopefully one of these roots is inside the relevant interval. We can also show that if the population restrictions on Θ were to hold in the sample, $(p-1)/p$ is a root of this equation. But this implies $\sigma_f^2 = 0$. Hence we should pick the smaller root, if both roots fall into the relevant interval. Given our estimate of λ, we have immediately an estimate of τ and can derive an estimate of β, and of the other parameters of interest.

The above estimation procedure, while inefficient, was outlined to indicate where the basic information for estimation was going to come from and how the different parts are related to each other.

The procedure is inefficient for two reasons: $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ is of rank 1. The estimator of $\underset{\sim}{\Theta}$ used above: $\hat{\underset{\sim}{\Theta}} = \frac{p}{p-1}(\overline{\underset{\sim}{R}} - \frac{1}{p}\underset{\sim}{R})$ was not constrained, however, to have rank = 1. Moreover, $\underset{\sim}{R}$ and $\overline{\underset{\sim}{R}}$ have been derived from OLS residuals of $\underset{\sim}{\mathbf{y}}$ on $\underset{\sim}{Z}$. But we know that $\underset{\sim}{\Omega}$, their variance covariance matrix, is not proportional to an identity matrix. Having an estimate of $\underset{\sim}{\Omega}^{-1}$, we could transform the original variables and get more efficient estimates of the reduced form coefficients and hence also a better set of residuals and improved estimates of $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$.

The problem of estimation is then (a) how best to impose the rank $\underset{\sim}{\Theta} = 1$ condition on our estimates, (b) how to use the estimated $\underset{\sim}{\hat{d}}$ and $\underset{\sim}{\hat{\Sigma}}$ to derive GLS estimates of $\underset{\sim}{\delta}$(the reduced form parameters of the various $\underset{\sim}{X}$'s), and (c) whether and how to iterate between the $\underset{\sim}{\hat{\delta}}$'s and associated $\underset{\sim}{\hat{\varepsilon}}$'s (the reduced form residuals) and the estimate of their (the $\underset{\sim}{\varepsilon}$'s) variance-covariance matrix $\underset{\sim}{\Omega}$.

4. Estimation[4]

Under normality assumptions for f_i, g_{ij} and the disturbances (u_1,u_2 and w)[5], the log likelihood function is (in terms of the stacked model $\underset{\sim}{y} = \underset{\sim}{Z}\underset{\sim}{\delta} + \underset{\sim}{\varepsilon}$, where $\underset{\sim}{y}' = [\underset{\sim}{y}_1', \ldots, \underset{\sim}{y}_K', \underset{\sim}{y}_S']$)

$$\ell n L(\underset{\sim}{y} \mid \underset{\sim}{\delta}, \underset{\sim}{\Omega})$$

$$= \frac{q}{2}\, \ell n |\underset{\sim}{\Omega}^{-1}| - \frac{1}{2}(\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta})' (\underset{\sim}{I}_q \otimes \underset{\sim}{\Omega})^{-1} (\underset{\sim}{y} - \underset{\sim}{Z}\underset{\sim}{\delta}).$$

To simplify estimation we obtain (in the Appendix) a factorization of $\underset{\sim}{\Omega}^{-1}$ into

$$\underset{\sim}{\Omega}^{-1} = \underset{\sim}{\Sigma}^{-1} \otimes \underset{\sim}{I}_p - \underset{\sim}{c}\underset{\sim}{c}' \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p'$$

[4] This section and the associated Appendix is largely due to Gary Chamberlain.

[5] Note that this implies a random ability effects interpretation of the model. A fixed effects version is discussed in Chapter 3.

where $\underset{\sim}{c}$ is related one-to-one to $\underset{\sim}{d}$ ($\underset{\sim}{c}$ is proportional to $\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$) and will be provided with an interpretation below. The likelihood function is then reparameterized in terms of $\underset{\sim}{\Sigma}^{-1}$ and $\underset{\sim}{c}$. The function is further simplified by evaluating $|\underset{\sim}{\Omega}^{-1}|$ explicitly. We also show that the reduced form residuals enter the L.F. only via the sufficient statistics $\underset{\sim}{R}$ and $\underset{\sim}{\overline{R}}$. $\underset{\sim}{R}$ is the matrix of the sums of squares and cross-products of these residuals divided by the total number of observations while $\underset{\sim}{\overline{R}}$ is computed by averaging the residuals over each family and then forming the matrix of weighted (in the case the p_i's differ) sums of squares and cross-products of these residuals divided by the total number of families.[6] The reparameterized and simplified L.F. can be written as:

$$\ell n\, L(\underset{\sim}{y}|\underset{\sim}{\delta}\ ,\ \underset{\sim}{\Sigma}^{-1},\underset{\sim}{c}) = \frac{qp}{2}\,\ell n\,|\underset{\sim}{\Sigma}^{-1}| + \frac{q}{2}\,\ell n(1 - p\underset{\sim}{c}'\underset{\sim}{\Sigma}\underset{\sim}{c})$$

$$-\frac{1}{2}\,pq\ \mathrm{tr}\ \underset{\sim}{\Sigma}^{-1}\underset{\sim}{R} + \frac{1}{2}\,p^2q\ \ \underset{\sim}{c}'\underset{\sim}{\overline{R}}\,\underset{\sim}{c}.$$

The maximization of this function is based on the following iterative algorithm: We start by estimating the reduced form slope coefficients $\underset{\sim}{\delta}$ consistently by ordinary least squares. Conditional on these $\hat{\underset{\sim}{\delta}}$, we proceed to get M.L. estimates of $\underset{\sim}{\Sigma}$ and $\underset{\sim}{d}$ by first calculating the reduced form residuals $\hat{\underset{\sim}{\varepsilon}} = \underset{\sim}{y} - \underset{\sim}{Z}\ \hat{\underset{\sim}{\delta}}$ and arranging them in a pqx(K+1) matrix

[6] When the p_i's differ, the "unbalanced" case, these weights depend on the unknown signal-noise ratio $\underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$. See Chapter 3 for an extension of the estimation procedure to this more complex case.

$$\mathbf{E} = (\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2 \ldots \hat{\boldsymbol{\varepsilon}}_{K+1}).$$

Then we find that linear combination of residuals from the K+1 equations which is most highly correlated with family structure; i.e., letting

$$\mathbf{F} = \mathbf{I}_q \otimes \boldsymbol{\ell}_p$$

be a set of family indicator dummy variables, we choose $\mathbf{c}$ and $\mathbf{f}$ to maximize the correlation T between $\mathbf{E}\mathbf{c}$ and $\mathbf{F}\mathbf{f}$. It can be seen, then, that $\mathbf{c}$ is a set of canonical weights combining the three residual series into one index. For a given $\mathbf{c}$ we obtain $\mathbf{f}$ by regressing $\mathbf{E}\mathbf{c}$ on the family indicators. Since $\mathbf{F}'\mathbf{F} = p\mathbf{I}_q$ and $\mathbf{F}'\mathbf{E} = p\bar{\mathbf{E}}$ where $\bar{\mathbf{E}}$ is the q x (K+1) matrix of residuals averaged over the families, we have

$$T^2 = \frac{p\mathbf{c}'\bar{\mathbf{E}}'\bar{\mathbf{E}}\mathbf{c}}{\mathbf{c}'\mathbf{E}'\mathbf{E}\,\mathbf{c}} = \frac{\mathbf{c}'\bar{\mathbf{R}}\,\mathbf{c}}{\mathbf{c}'\mathbf{R}\,\mathbf{c}} .$$

T^2 is maximized by letting $\mathbf{c}$ be the eigenvector of $\bar{\mathbf{R}}$ in the metric of $\mathbf{R}$ corresponding to the largest eigenvalue ρ:

$$\bar{\mathbf{R}}\,\mathbf{c} = \rho\mathbf{R}\,\mathbf{c}.$$

Note that

$$\rho = \frac{\mathbf{c}'\bar{\mathbf{R}}\,\mathbf{c}}{\mathbf{c}'\mathbf{R}\,\mathbf{c}} = T^2$$

is the square of the maximal canonical correlation coefficient between E and F. An index of family ability is then formed from the fitted values in the regression of Ec on F:

$$\hat{a} = F \bar{E}_c \cdot$$

So the family component of ability for the i th family is estimated by weighting the averaged residuals for the i th family by the canonical weights c. The reduced form ability coefficients d are then obtained by regressing the residuals from each equation on $\hat{a}$:

$$d' = \hat{a}'E = c'\bar{R}.$$

Thus d can also be characterized by the dual relationship

$$\bar{R}^{-1} d = \frac{1}{\rho} R^{-1} d,$$

with the scale of d determined from

$$d' \bar{R}^{-1} d = \frac{1}{p-1} \left(p - \frac{1}{\rho}\right).$$

The M.L. estimate of Σ satisfies the adding up property

$$\Sigma = R - dd'.$$

The M.L. estimate of δ given Σ and d is generalized least squares. The computations are simplified by analytically inverting the disturbance covariance matrix to obtain the following formula for the GLS estimator of δ

$$\underset{\sim}{\delta} = (\underset{\sim}{H}_W + \underset{\sim}{H}_B)^{-1}(\underset{\sim}{H}_W \hat{\underset{\sim}{\delta}}_W + \underset{\sim}{H}_B \hat{\underset{\sim}{\delta}}_B)$$

where $\hat{\underset{\sim}{\delta}}_W$ is the least squares estimate just using the within family moments and $\hat{\underset{\sim}{\delta}}_B$ just uses the between family moments:

$$\hat{\underset{\sim}{\delta}}_{Wk} = \underset{\sim}{W}_{xx}^{-1} \; \underset{\sim}{W}_{xy_k}$$

$$\hat{\underset{\sim}{\delta}}_{Bk} = \underset{\sim}{B}_{xx}^{-1} \underset{\sim}{B}_{xy_k}, \; k = 1,2, \ldots K+1$$

with

$$\underset{\sim}{T}_{xx} = \sum_{i=1}^{q} \underset{\sim}{X}_i' \; \underset{\sim}{X}_i$$

$$\underset{\sim}{B}_{xx} = \frac{1}{p} \sum_{i=1}^{q} \underset{\sim}{X}_i' \; \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' \underset{\sim}{X}_i$$

$$\underset{\sim}{W}_{xx} = \underset{\sim}{T}_{xx} - \underset{\sim}{B}_{xx}$$

with similar expressions for $\underset{\sim}{W}_{xy_k}$ and $\underset{\sim}{B}_{xy_k}$. $\underset{\sim}{H}_w$ and $\underset{\sim}{H}_B$ are the precision matrices for

$\hat{\delta}_W$ and $\hat{\underset{\sim}{\delta}}_B$:

$$\underset{\sim}{H}_W^{-1} = E(\hat{\delta}_W - \underset{\sim}{\delta})(\hat{\underset{\sim}{\delta}}_W - \underset{\sim}{\delta})' = \underset{\sim}{\Sigma} \otimes \underset{\sim}{W}_{xx}^{-1}$$

$$\underset{\sim}{H}_B^{-1} = E(\hat{\underset{\sim}{\delta}}_B - \underset{\sim}{\delta})(\hat{\underset{\sim}{\delta}}_B - \underset{\sim}{\delta})' = p(\underset{\sim}{d}\underset{\sim}{d}' + \frac{1}{p} \underset{\sim}{\Sigma}) \otimes \underset{\sim}{B}_{xx}^{-1}$$

(The GLS procedure when the $\underset{\sim}{X}$'s differ across equations is described in the appendix).

The joint M.L. estimates of $\underset{\sim}{\delta}, \underset{\sim}{\Sigma}$, and $\underset{\sim}{d}$ can be obtained by iterating on these equations. Given an initial consistent estimate of $\underset{\sim}{\delta}$ and the associated reduced form residuals we obtain $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ from the canonical correlation analysis outlined above. Then we form $\underset{\sim}{H}_W$ and $\underset{\sim}{H}_B$ and obtain a new estimate of $\underset{\sim}{\delta}$ by pooling the within and between family estimates. This estimate of $\underset{\sim}{\delta}$ has the asymptotic ($q \rightarrow \infty$) efficiency properties as do the estimates of $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$ based on its residuals. Further iteration is, however, probably desirable.

5. The Main Results

Tables 3 and 4 present the M.L. estimates for our model together with the intermediate calculations. The results are quite consistent with the covariance analysis described in Section 2. Now the schooling coefficients in both the income and occupation equations have increased relative to the OLS values indicating that going to the within family deviations was only a partial cure. Corresponding to the uniformly higher schooling coefficients we obtain negative coefficients for the "ability" variable in both the income and occupation equations. However, the relative magnitudes of the coefficients indicate that the unobserved variable primarily affects income and occupation with only a negligible effect on schooling. For a person who is one standard deviation above the mean of the distribution of the unobservable would be only .03 standard deviations above the mean on the schooling distribution (net of age) but his income would be 41% lower than someone with average "ability". Also the contribution of A to the fit of the equation is much more pronounced for income than for schooling. The signal noise ratio $\gamma_1^2\sigma_a^2 / \left(\gamma_1^2\sigma_a^2 + \sigma_{u_1}^2\right)$ is 72% for Y but $\gamma_3^2\sigma_a^2 / \left(\gamma_3^2\ \sigma_a^2 + \sigma_w^2\right)$ is only .1% for S. So our prior expectation that A would be an important

Table 3: Parameter Estimates: Income-Occupation-Schooling Model. 156 Pairs of Brothers, 1928, Indiana, U.S.A.

Original Data from Gorseline (1932).

	Method			
Coefficients of the structural equations	Biased least squares		Maximum likelihood systems estimates	
	Total sample	Within families	unre- stricted	recursive model
Age in the				
income eq.	.088 (.021)	.061 (.031)	.080 (.020)	.080 (.020)
occupation eq.	.005 (.003)	.011 (.009)	.006 (.003)	.006 (.003)
schooling eq.	-.066 (.019)	.029 (.049)	-.067 (.019)	-.066 (.019)
Age squared in the				
income eq.	-.001 (.0003)	-.000 (.0004)	-.0007 (.0002)	-.0007 (.0002)
Schooling in the				
income eq. β_1	.082 (.010)	.080 (.011)	.088 (.009)	.084 (.009)
occupation eq. β_2	.104 (.010)	.135 (.015)	.107 (.010)	.105 (.010)
"Ability" in the				
income eq. γ_1			.416 (.038)	.417 (.038)
occupation eq. γ_2			.214 (.046)	.210 (.046)
schooling eq. γ_3			-.092 (.178)	.0

The γ coefficients are scaled by assuming that $\sigma_f^2 = 1$ and $\gamma > 0$. The numbers in parenthesis are the computed standard errors. For the M.L. estimates they are based on the structural information matrix Ξ given in (A.51). In the restricted model we delete the row and column of Ξ corresponding to γ_3.

Table 4: Gorseline (1932) Brothers: Intermediate Data and Calculations

Unrestricted model: (based on M.L. reduced form residuals)

$$\underset{\sim}{R} = \begin{bmatrix} .437 & .231 & .928 \\ & .488 & 1.168 \\ & & 11.193 \end{bmatrix} , \quad \underset{\sim}{\overline{R}} = \begin{bmatrix} .313 & .157 & .551 \\ & .243 & .532 \\ & & 6.512 \end{bmatrix}$$

Sample size N = pq = 312

Canonical weights: $\underset{\sim}{c}' = (1.76 \quad .300 \quad -.191)$

Squared canonical correlation coefficients: $\rho_1=.75$, $\rho_2=.63$, $\rho_3=.45$

$$\underset{\sim}{d} = \begin{bmatrix} .408 \\ .204 \\ -.092 \end{bmatrix} , \quad \underset{\sim}{\Sigma} = \begin{bmatrix} .270 & .148 & .966 \\ & .446 & 1.187 \\ & & 11.184 \end{bmatrix}$$

$\sigma^2_{u_1} = .098$, $\sigma^2_{u_2} = .297$, $\sigma^2_w = 11.180$

$\sigma^2_f/\sigma^2_a = .66$

$$\text{plim } \underset{\sim}{\overline{R}} = \underset{\sim}{d}\underset{\sim}{d}' + \frac{1}{p}\underset{\sim}{\Sigma} = \begin{bmatrix} .302 & .158 & .445 \\ & .265 & .575 \\ & & 5.600 \end{bmatrix}$$

cont.

Table 4 (Cont.)

Recursive model: (based on M.L. structural form residuals)

$$\underset{\sim}{R} = \begin{bmatrix} .360 & .135 \\ & .366 \end{bmatrix} , \qquad \underset{\sim}{\overline{R}} = \begin{bmatrix} .267 & .112 \\ & .203 \end{bmatrix}$$

Canonical weights $\underset{\sim}{c}' = (1.776 \quad .279)$.

Squared canonical correlation coefficients: ρ_1=.746, ρ_2=.493

$$\underset{\sim}{\overline{\gamma}} = \begin{bmatrix} .417 \\ .210 \end{bmatrix} , \qquad \underset{\sim}{\Delta} = \begin{bmatrix} .186 & .047 \\ & .322 \end{bmatrix}$$

$\sigma^2_{u_1} = .093$, $\sigma^2_{u_2} = .298$, $\sigma^2_{w} = 11.193$

$\sigma^2_f/\sigma^2_a = .65$

determinant of S, such as IQ or family wealth, (which led us to normalize $\gamma_3 > 0$), is not born out in the data.

Rather we appear to have a recursive model in which the omitted variable affects only income and occupation, i.e. $\gamma_3 = 0$. Since this is equivalent to assuming that $d_3 = 0$, it is a testable restriction on the reduced form equations of our more general model. In the restricted model it is more tractable to work with the structural form of the likelihood function since there is now no correlation between the residuals from the schooling equation and the other equations of the model and so the S equation factors out of the structural likelihood and can be estimated by OLS. The structural covariance matrix for the y_k indicators (Y and O) has diagonal blocks of the form

$$\overline{\underset{\sim}{\gamma}\underset{\sim}{\gamma}}' \otimes \underset{\sim}{\ell}_p \underset{\sim}{\ell}_p' + \underset{\sim}{\Delta} \otimes \underset{\sim}{I}_p$$

where

$$\underset{\sim}{\overline{\gamma}} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \quad \underset{\sim}{\Delta} = \tau \underset{\sim}{\overline{\gamma}}\, \underset{\sim}{\overline{\gamma}}' + \begin{bmatrix} \sigma^2_{u_1} & 0 \\ 0 & \sigma^2_{u_2} \end{bmatrix}$$

and we have normalized $\sigma^2_f = 1$. So $\underset{\sim}{\Delta}$ is unrestricted and identifying $\underset{\sim}{\overline{\gamma}}$ with $\underset{\sim}{d}$ and $\underset{\sim}{\Delta}$ with $\underset{\sim}{\Sigma}$ we can apply our reduced form M.L. algorithm directly. Then given $\underset{\sim}{\overline{\gamma}}$ and $\underset{\sim}{\Delta}$ we can solve for τ, $\sigma^2_{u_1}$, and $\sigma^2_{u_2}$. This part of the log likelihood function evaluated at the maximum ($L^*_{d_3=0}$) is given by (A.39) in the Chapter 3 Appendix where n = 1, $|\underset{\sim}{R}|$ equals the generalized variance of the structural residuals from the first two equations, and ρ_1 is the squared canonical correlation between these residuals and a set of family indicator dummy variables. The second part of $L^*_{d_3=0}$ is simply -pq times the standard error of the OLS estimated schooling equation.[6]

Comparing $L^*_{d_3=0}$ with L^* for the unrestricted model gives a likelihood ratio (L.R.) of .87 and $-2\log(\text{L.R.}) = .27 \sim X^2(1)$ which is entirely consistent with a recursive model. The structural estimates for the restricted model and the intermediate calculations are in tables 3 and 4. Note that we have renormalized so that γ_1,(and hence γ_2) is positive, interpreting A as a joint luck or economic, but not scholastic, "ability" variable. The estimate of β_1 is .084, almost identical to the OLS estimate (.082).[7] Although our estimator was carefully designed to detect omitted variables connecting and biasing the income and schooling relationships we haven't found any. But before accepting OLS we will take a closer look at the results and the assumptions they are based on.

[6] There is also a term $-\frac{K}{2}\ \ell n\ 2\pi - \frac{pq}{2}K - \frac{1}{2}\ell n\ 2\pi - \frac{pq}{2}$

$$= -\frac{K+1}{2}\ \ell n\ 2\pi - \frac{pq(K+1)}{2}$$

which cancels with an identical term in the unrestricted reduced form likelihood.

[7] The departure results from the joint estimation of the Y and O equations together with the variance components mixing of the total and within family OLS estimates.

6. Extensions

The identifiability of our model rests on two key assumptions: that u_1 and u_2, the disturbances in the income and occupation equations, are uncorrelated; and that there is a single common unobservable variable connecting all the residuals. The first assumption is not too plausible. If u_1 consists largely of luck which results in a higher income than an individual's schooling and "ability" would have predicted, then he is likely to also have a higher occupational status, implying a positive correlation between u_1 and u_2. But if u_1 and u_2 reflect the individual's preferences for income vs. status and if, given his schooling and ability, he can trade off one for the other, then the correlation could be negative. So we expand the model to allow for a correlation between u_1 and u_2 or alternatively (and equivalently) rewrite the y_1 equation to include y_2:

$$y_1 = \underset{\sim}{X}\underset{\sim}{\alpha}_1 + \beta_1 y_s + \eta y_2 + \gamma_1 a + u_1$$

and keep the E $u_1 u_2 = 0$ assumption.

Expanding the model in this way has no effect on the reduced form. For we have not added family factors which would break the restrictions on $\underset{\sim}{\Theta}$. We have only altered $\underset{\sim}{\Sigma}$, replacing $v_{12} = \beta_1\beta_2\sigma_w^2$ by $\beta_1\beta_2\sigma_w^2 + \eta\sigma_{u_2}^2$.

Since $\underset{\sim}{\Sigma}$ was unconstrained to begin with, the reduced form is unchanged and our test for $d_3 = \gamma_3 = 0$ remains valid. But we can no longer solve for $\tau= \sigma_g^2/\sigma_f^2$ by making the solutions for β_1 and β_2(for a given τ) agree with the solution for their product; i.e., the structural parameters are not identified.[8] So we have to introduce additional prior information, e.g. about η or about λ. Experience with other data sets would suggest $\eta \geqslant 0$ and on the order of .05 to .15 (see Table 5 in Griliches and Mason (1972) where η is estimated in the presence of a direct ability measure). Alternatively we can compute the β's and η for a given value of the variance ratio $\lambda =\sigma_f^2/\sigma_a^2$. A pure genetic heredity model would predict a ratio of .5 to .6 (see Jencks (1972), Appendix B). Adding common financial wealth to the interpretation of the unobservable suggests the range $.5 \leq \lambda < 1.0$.

We had initially planned to use this prior to see what the resulting range for the β's would be. But in fact the feasible range is not much wider than this. Although we are not identified in the usual sense we do

[8] In the restricted (d_3=0) model everything is still identified except for η and τ.

have two sources of bounds: $0 \leq \lambda \leq 1$ and the implied correlation:

$$r_{12} = \eta\sigma_{u_2} \Big/ \sqrt{\sigma^2_{u_1} + \eta^2\sigma^2_{u_2}}$$

between $u_1 + \eta u_2$ and u_2 in the semi-reduced form (with y_2 but not y_s solved out) must be less than one in absolute value. Putting these bounds together results in bounds on the other parameters of the model as shown in Table 5. Each row of the table is equally likely for they are all based on the same M.L. reduced form estimates of $\underset{\sim}{d}$ and $\underset{\sim}{\Sigma}$. They just represent different ways of allocating σ_{12} between $\tau = \sigma^2_g / \sigma^2_f$ and η. The whole table has the same status as a point estimate. To extend the table to values of $\lambda \leq .49$ (corresponding to $r_{12} = -1.0$) would require restrictions on the reduced form likelihood which would be testable. So we have identification in the sense of obtaining a non-trivial bound. In fact for our case the bound is extremely tight. With $\lambda = .66$ we have $\eta = 0$ and the other parameters take on the previously reported M.L. values. For higher values of λ there is a very slight decline in $\beta_1 + \eta\beta_2$ (the total effect of S including its effect via O), and η increases up to a maximum value of .14; lower values of λ imply $\eta < 0$ and a slight increase in $\beta_1 + \eta\beta_2$. The ratio of the ability coefficients in the income and schooling equations remains unchanged, still reflecting a negligible effect of "ability" on schooling. So our estimates are very robust against the structural no correlation assumption.

There remains the possibility that there may be more than one common unobservable (factor). We have lumped both the family's socio-economic

Table 5: Conditional Estimates in the Expanded Model

	Implied						
Conditional on σ_f^2/σ_a^2 equaling	$\beta_1+\eta\beta_2$	β_1	β_2	η	$\gamma_1+\eta\gamma_2$	γ_3	$r_{u_1u_2}$
.50	.090	.107	.108	-.159	-.416	.092	-.716
.60	.089	.094	.107	-.048	-.416	"	-.096
.70	.088	.085	.107	.024	-.416	"	.040
.75	.087	.082	.107	.052	-.416	"	.080
.80	.087	.079	.106	.075	-.416	"	.111
.85	.087	.077	.106	.096	-.416	"	.135
.90	.087	.075	.106	.113	-.416	"	.155
.99	.086	.071	.106	.138	-.416	"	.184

Expanded model:

$$y_1 = \underset{\sim}{X}_1 \underset{\sim}{\alpha}_1 + \beta_1 y_s + \gamma_1 a + \eta y_2 + u_1$$

or alternatively $\eta = 0$ but $Eu_1u_2 \neq 0$

status and the children's native intelligence into one measure A. But these two different kinds of "inheritance" may not have the same coefficients in the different equations. Moreover there may be more than one type of "intelligence," including a kind (such as "test-wiseness") which may lead to scholastic but not necessarily to material success (in addition to its effect via schooling). In any case, there is some statistical evidence for the existence of a second factor. The presence of such a factor is indicated by the squared canonical correlations. In the "no factor" model we would expect $\underset{\sim}{\overline{R}}$ to be proportional to $\underset{\sim}{R}$ but reduced by $\frac{1}{p}$ from averaging over families with p members. Then all the roots of $\underset{\sim}{\overline{R}}$ in the metric of $\underset{\sim}{R}$ would be $\frac{1}{p}$. With data on pairs we would expect all the squared canonical correlations to be .5. Actually we get $\rho = (.75, .63, .45)$. So clearly there is at least one factor in the data and in terms of the unexplained variance, i.e., 1.0-.50 = .50, the first factor (.75-.50) accounts for 50% of it.

To assess a second factor we construct a second index from the Y, O, and S reduced form residuals which is most highly correlated with a set of family indicator dummy variables, subject to the restriction of being uncorrelated with the first canonical index. Then ρ_2 gives the squared multiple correlation between the index and the family dummy variables. We get $\rho_2 = .63$ which is not very close to .5 and in terms of the unexplained variance (net of the first pair of canonical variables) the second factor accounts for 26% of it. An alternative interpretation of these variance ratios is that they are the principal components of $\underset{\sim}{\theta}$ in the metric of

$\underset{\sim}{R} = \underset{\sim}{\theta} + \underset{\sim}{\Sigma}$. For the components are $\psi_h/(1 + \psi_h)$ where the ψ_h are the roots of $\underset{\sim}{\theta}$ relative to $\underset{\sim}{\Sigma}$. It is shown in Appendix A, Chapter 3 that

$$\begin{aligned}\psi_h/(1 + \psi_h) &= (p\rho_h - 1)/(p-1) \\ &= (\rho_h - \frac{1}{p})/(1 - \frac{1}{p})\end{aligned}$$

i.e., the fraction of the unexplained variance accounted for by factor h. Now with one factor $\underset{\sim}{\theta} = \underset{\sim}{d}\underset{\sim}{d}'$ and $\psi_1 = \underset{\sim}{d}' \underset{\sim}{\Sigma}^{-1} \underset{\sim}{d}$ is a generalized reduced form signal noise ratio. Thus $\psi_1/(1 + \psi_1)$ gives the fraction of the residual variance accounted for by the systematic family factor. With 2 factors ψ_2 is the signal-noise ratio net of the first factor. The sum of the principal components

$$\frac{\psi_1}{1+\psi_1} + \frac{\psi_2}{1+\psi_2} = .76$$

is the total fraction of the residual variance accounted for by systematic factors and one third of it is due to the second factor.

A likelihood ratio (L.R.) test for two factors vs. one factor is derived in the Appendix. Conditional on the reduced form slope coefficients δ the test is

$$\begin{aligned}-2 \log (\text{L.R.}) = 2(L_2 - L_1) &= -pq \log p(1-\rho_2)/(p-1) \\ &\quad + q\log (\frac{1}{\rho_2} - 1)/(p-1) \to \chi^2(2).\end{aligned}$$

This test statistic is a measure of how far ρ_2 is from $\frac{1}{p}$ (or how far

$\psi_2/(1+\psi_2)$ is from zero). For $\rho_2 = \frac{1}{p}$ the likelihood ratio is one, implying no evidence for a second factor, and for $\rho_2 = 1.0$ it is zero. The unconditional test includes a comparison of the generalized variances of the GLS reduced form residuals for the one and two factor models and also evaluates the difference (if any) in the estimates of ρ_1. The unconditional test results in a quite unlikely value of 10.9. Also, we see in table 6 that the approximation of $\underset{\sim}{\theta} + \frac{1}{p}\underset{\sim}{\Sigma}$ to $\underset{\sim}{\overline{R}}$ is considerably improved in the 2 factor model.

So we turn to the question of what structural inferences can be made from a two factor reduced form. Now the structural form is:

$$y_k = \underset{\sim}{X}\,\underset{\sim}{\alpha}_k + y_{K+1}\beta_k + (f_1+g_1)\gamma_k + (f_2+g_2)\eta_k + u_k \qquad k = 1,\ldots K$$

$$y_{K+1} = \underset{\sim}{X}\,\underset{\sim}{\alpha}_{K+1} + (f_1+g_1)\gamma_{K+1} + (f_2+g_2)\eta_{K+1} + u_{K+1} .$$

In the reduced form we have:

$$\underset{\sim}{d}_1 = \begin{bmatrix} \gamma_1 + \beta_1\gamma_3 \\ \gamma_2 + \beta_2\gamma_3 \\ \gamma_3 \end{bmatrix} \qquad \underset{\sim}{d}_2 = \begin{bmatrix} \eta_1 + \beta_1\eta_3 \\ \eta_2 + \beta_2\eta_3 \\ \eta_3 \end{bmatrix}$$

$$\underset{\sim}{D} = (\underset{\sim}{d}_1\ \underset{\sim}{d}_2)$$

Table 6: Two Factor Model; Reduced Form Calculations

$$\underset{\sim}{R} = \begin{bmatrix} .437 & .232 & .930 \\ & .488 & 1.168 \\ & & 11.194 \end{bmatrix} , \quad \underset{\sim}{\overline{R}} = \begin{bmatrix} .314 & .157 & .554 \\ & .243 & .533 \\ & & 6.525 \end{bmatrix}$$

Canonical weights: $\underset{\sim}{c}_1' = (1.767 \;\; .297 \;\; -.190)$

$\underset{\sim}{c}_2' = (.506 \;\; -.885 \;\; .308)$

Squared canonical correlation coefficients: $\rho_1=.75$, $\rho_2=.63$, $\rho_3=.45$

$$\underset{\sim}{\Theta} = \begin{bmatrix} .191 & .087 & .192 \\ & .042 & .016 \\ & & 2.170 \end{bmatrix} , \quad \underset{\sim}{\Sigma} = \begin{bmatrix} .246 & .144 & .737 \\ & .446 & 1.152 \\ & & 9.025 \end{bmatrix}$$

$$\text{plim} \frac{p}{p-1} (\underset{\sim}{\overline{R}} - \frac{1}{p}\underset{\sim}{R}) = \underset{\sim}{\Theta}$$

$$\frac{p}{p-1} (\underset{\sim}{\overline{R}} - \frac{1}{p}\underset{\sim}{R}) = \begin{bmatrix} .190 & .082 & .178 \\ & -.001 & -.101 \\ & & 1.855 \end{bmatrix}$$

$$\text{plim } \underset{\sim}{\overline{R}} = \underset{\sim}{\Theta} + \frac{1}{p}\underset{\sim}{\Sigma} = \begin{bmatrix} .314 & .159 & .561 \\ & .265 & .592 \\ & & 6.682 \end{bmatrix}$$

and letting $\underset{\sim}{\Phi}$ be the covariance matrix of the family factor gives

$$\underset{\sim}{\theta} = \underset{\sim}{D}\underset{\sim}{\Phi}\underset{\sim}{D}' .$$

So, if we scale f_1 and f_2 to have unit variance then:

$$\underset{\sim}{\Theta} = \underset{\sim}{d}_1\underset{\sim}{d}_1' + \underset{\sim}{d}_2\underset{\sim}{d}_2' + (\underset{\sim}{d}_1\underset{\sim}{d}_2' + \underset{\sim}{d}_2\underset{\sim}{d}_1')\ r_f$$

where r_f is the correlation between f_1 and f_2. Similarly

$$\underset{\sim}{\Sigma} = \underset{\sim}{d}_1\ \underset{\sim}{d}_1'\ \tau_1 + \underset{\sim}{d}_2\underset{\sim}{d}_2'\tau_2 + (\underset{\sim}{d}_1\underset{\sim}{d}_2' + \underset{\sim}{d}_2\underset{\sim}{d}_1')\ \sqrt{\tau_1\tau_2}\ \ r_g + \underset{\sim}{V}$$

with $\tau_1 = \sigma^2_{g_1} \big/ \sigma^2_{f_1}$, $\tau_2 = \sigma^2_{g_2} \big/ \sigma^2_{f_2}$, and r_g is the correlation between the individual components (g_1 and g_2) of the two unobservables.

Clearly, the model is highly underidentified. But a substantial simplification results from limiting our extension to a second factor which has only a family component ($\tau_2 = 0$). Examples would be family wealth or measures of family background such as father's occupational status or father's schooling. Then r_g equals zero and the structure of $\underset{\sim}{\Sigma}$ is identical to the one factor case. So if we can obtain $\underset{\sim}{d}_1$ up to a scale factor then the argument of Section 3 will identify the structural parameters. The problem is to retrieve $\underset{\sim}{d}_1$ from $\underset{\sim}{\theta}$. If we knew r_f we could factor $\underset{\sim}{\Phi} = \underset{\sim}{P}\ \underset{\sim}{P}'$, let $\underset{\sim}{\tilde{D}} = \underset{\sim}{D}\underset{\sim}{P}$, and obtain the factorization $\underset{\sim}{\theta} = \underset{\sim}{\tilde{D}}\ \underset{\sim}{\tilde{D}}'$. The general solution to this equation is $\underset{\sim}{\tilde{D}} = \underset{\sim}{\tilde{D}}_o\ \underset{\sim}{T}$ where $\underset{\sim}{\tilde{D}}_o$ is any solution and $\underset{\sim}{T}$ is a rotation, $\underset{\sim}{T}'\underset{\sim}{T} = \underset{\sim}{I}$. So we must condition on both r_f and (in our two factor case) a rotation angle ζ. Since ζ is difficult to interpret we instead specify r_f and λ_1. However, the relationship between λ_1 and ζ is neither one-to-one nor onto. We have to solve a cubic equation to obtain ζ from λ_1 and this can have multiple solutions or no (admissible) solution at all.

Over the range of correlations (r_f) considered we cannot obtain a value for $\lambda_1 = \sigma^2_{f_1} / \sigma^2_{a_1}$ as high as .75 for any ζ. We can obtain values as low as zero but they violate the restriction that $\sigma^2_{u_1} > 0$. In fact table 7 gives all values of λ_1 (at .05 intervals) that satisfy the inequality restrictions. When there is more than one rotation for a given λ_1 then the one with the lower value of β_1 is reported. It turns out that for the other rotation the γ's are not all positive as we would expect them to be for an ability variable with a genetic component.[9]

Putting together the restrictions that $0 \leq \lambda_1 \leq 1$ and $\sigma > 0$ produces a lower bound on β_1 of .046 corresponding to $r_f = .40$ and $\lambda_1 = .50$. (The bound also occurs for $r_f = .20$ and $\lambda_1 = .46$.) The upper bound is .26 (higher values imply $\sigma^2_w < 0$) and if we add the restriction $\gamma > 0$ it is .082 attained at $r_f = 0.0$, $\lambda_1 = .63$ (and at positive correlations for somewhat higher values of λ_1). The status of these bounds is identical to the bounds in the extended one factor model. They are all based on the same 2 factor M.L. reduced form estimates and are simply different equally likely ways of interpreting them. Obtaining estimates outside the bound would require imposing restrictions which would reduce the likelihood and be testable. In our case the bound is not vacuous but neither is it particularly sharp. The schooling coefficient in the income equation could be as much as 44% lower than its OLS value of .082.

Our separability restriction that $\gamma_3 = 0$ is not testable by itself in the two factor model. Complete separability requires $\eta_1 = \eta_2 = 0$ and $r_f = 0$ in addition to $\gamma_3 = 0$ and implies that

[9] We are free to change the signs of all the γ's and η's simultaneously; however, we can't change just the γ's alone without changing the sign of the correlation r_f which we assume is positive. There are only two rotations that satisfy the constraints $\sigma > 0$.

Table 7: Lower Bound on β Rotations; Restricted Two Factor Model

	$\frac{\sigma^2_{f_1}}{\sigma^2_{a_1}}$	β_1	β_2	γ_1	γ_2	γ_3	η_1	η_2	η_3
r_f=0.0	.50	.056	.120	.404	.127	.560	-.115	-.231	1.362
	.60	.076	.125	.416	.182	.188	-.035	-.198	1.461
	.65	.086	.130	.414	.225	-.170	.052	-.155	1.463
r_f=.20	.50	.056	.119	.412	.129	.577	-.198	-.257	1.244
	.60	.074	.124	.424	.177	.250	-.133	-.239	1.402
	.65	.084	.129	.425	.220	-.089	-.053	-.210	1.488
r_f=.40	.50	.046	.118	.437	.121	.719	-.306	-.286	1.030
	.60	.068	.122	.449	.170	.404	-.256	-.284	1.264
	.65	.077	.126	.455	.205	.156	-.206	-.275	1.404
	.70	.090	.133	.442	.267	-.401	-.075	-.234	1.587
r_f=.50	.55	.052	.119	.465	.138	.701	-.355	-.303	.991
	.60	.063	.121	.471	.164	.536	-.331	-.306	1.130
	.65	.073	.124	.479	.198	.306	-.292	-.305	1.296
	.70	.082	.129	.482	.239	-.019	-.227	-.295	1.483

Two factor structural model:

$$y = X\underset{\sim}{\alpha} + y_s\beta + (f_1 + g_1)\gamma + f_2\eta + \mu$$

$$\sigma^2_{f_2} / \sigma^2_{a_2} = 1.0 \quad r_f = \sigma_{f_1 f_2} / \sigma_{f_1}\sigma_{f_1}$$

$$\sigma_{f_1} = \sigma_{f_2} = 1.0$$

The reduced form is:

$$\underset{\sim}{\bar{Y}} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} , \quad \underset{\sim}{\Theta} = \begin{bmatrix} \underset{\sim}{\bar{Y}}\underset{\sim}{\bar{Y}}' & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \beta_1^2 & \beta_1\beta_2 & \beta_1 \\ & \beta_2^2 & \beta_2 \\ & & 1 \end{bmatrix} \eta_3^2$$

$$\underset{\sim}{\Sigma} = \tau_1 \begin{bmatrix} \underset{\sim}{\bar{Y}}\underset{\sim}{\bar{Y}}' & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sigma_{u_1}^2 + \beta_1^2\sigma_w^2 & \beta_1\beta_2\sigma_w^2 & \beta_1\sigma_w^2 \\ & \sigma_{u_2}^2 + \beta_2^2\sigma_w^2 & \beta_2\sigma_w^2 \\ & & \sigma_w^2 \end{bmatrix} .$$

So there are 2 restrictions across $\underset{\sim}{\Theta}$ and $\underset{\sim}{\Sigma}$:

$$\beta_1 = \theta_{13} / \theta_{33} = \sigma_{13} / \sigma_{33}$$

$$\beta_2 = \theta_{23} / \theta_{33} = \sigma_{23} / \sigma_{33} \quad .$$

In fact, we get $\theta_{13} / \theta_{33} = .088$, $\sigma_{13} / \sigma_{33} = .082$ which is not bad. But $\theta_{23} / \theta_{33} = .007$ and $\sigma_{23} / \sigma_{33} = .13$. So it appears that $\eta_2 \neq 0$ and a L.R. test for both restrictions gives $6.72 \sim \chi^2(2)$. Also in table 8 we can see that $r_f = 0$ or $.2$ and a λ_1 between .60 and .65 result in γ_3 and η_1 being essentially zero whereas η_2 is not negligible and in fact is negative. This possibly reflects anomalies in the construction of the status scale and we hope to return to this in the future. But even in the two factor model there is some indication of a partial recursiveness with $\gamma_3 = 0$, $r_f = 0$ and $\eta_1 = 0$. Then β_1 is estimable by either the constrained (smoothed) between family regression $\theta_{13} / \theta_{33}$ or by the constrained within family regression $\sigma_{13} / \sigma_{33}$. The later estimate ($\sigma_{13} / \sigma_{33} = .082$) is more robust since it only requires $\gamma_3 = 0$. For then the factor with an individual component causes no bias ($\gamma_3 = 0$) and the second factor, being purely family, is swept out by using the (constrained) within family moments.

7. Summary and Discussion

This paper dealt with two topics, the substantive problem of "ability bias" in estimates of returns to schooling and a somewhat novel econometric approach to estimation in the presence of unobservable variables. From a substantive point of view the new econometric methods did not produce results which differed greatly from those based on simpler methods. This is either satisfying or disappointing, depending on one's point of view. An elaborate procedure, designed to detect possible sources of bias, yielded little evidence of such bias. It is quite likely that important unobserved variables have been left-out from our schooling-achievement model but they are not of the type one usually associates with the notion of intellectual "ability". There is a significant positive relationship between disturbances in the income and occupation equations but it seems to have little to do with the disturbances in the schooling equation. There is some indication of a negative relationship between family components in the schooling and occupation equations, but little evidence of a strong relationship between unobservable family components in the schooling and income equations, implying little bias in the estimates of schooling coefficients which ignore such connections.

These conclusions are limited to the particular data set analyzed and the range of alternative hypotheses investigated. Since estimates of bias in the schooling coefficient depend crucially on the relationship between the left-out ability variable and the level of schooling in the sample studied, there is no reason to expect that they would generalize to different populations with a different ability-schooling nexus. It does appear, though, that there was little relationship, at the beginning of this century in Indiana, between the distribution of "ability" and the distribution of schooling, particularly if "ability" is assumed to have a significant family component. This may have changed over time, however, as the schooling system developed and became more selective. We do intend, therefore, to replicate our analysis on a more recent set of brothers taken from the 1966-1969 National Longitudinal Survey of Young Men.

Besides bringing us into amore recent period, the NLS data will allow us to overcome several other limitations of the Gorseline sample. It will have more background data on parental status and wealth allowing for a "cleaner" and clearer interpretation of the unobservable, making the various no-correlation assumptions more palatable. Moreover, the availability of some direct measures of "ability", such as IQ test scores will provide an explicit test of such interpretations. Also, given a larger number of indicators we may be able to dispense with the use of the rather ambiguous measure of "occupation." The whole notion of "occupation" deserves more study and the variable itself needs rescaling in any case.

From a statistical point of view, our work can be viewed as an extension of the error-components literature to the simultaneous equations systems context or alternatively as an extension and specialization of the resurgent path-analysis literature to the error-components case. The connections between our work and these fields are discussed at length in Griliches (1973) and will not be reproduced here. We should note, however, explicitly the similarity of some of our results to those of Hauser and Goldberger (1971) and the work of Jöreskog, especially his "Factoring the Multitest-multioccasion correlation matrix" (1970). Our results can also be related to the weighted regression technique of Frisch and Koopmans, with the weighting scheme derived from within families replication.

Appendix: Data and Variables

The data are taken from D.E. Gorseline, The Effect of Schooling Upon Income, Indiana University, 1932, and are based on interviews and mail surveys undertaken by Gorseline in 1928. The collection procedures and caveats are described rather clearly in his book and will not be reproduced here. He collected "usuable" data on 172 sets of brothers or a total of 68 individuals. Limiting ourselves, in this paper, solely to pairs of brothers, we have 156 pairs or a total of 312 individuals.

Schooling in this study is measured by the "probable grade of school attained" rather than by the reported years of school attained" rather than by the reported years of school attended, defined as "the grade in which the man who filled out the questionnaire was when he stopped going to school". It is taken from Tables LIV-LVII of the book. The "probably" enters into the definition because Gorseline often adjusted or estimated this number on the basis of other information in the sample.

"Income" is net earnings for the calendar year 1927 plus the imputed value of home consumed food for farmers and retail businessmen and the imputed value of housing when supplied with the job(e.g., parsonage for ministers).

"Occupation" is as of 1928(tables XCIII-XCVI). It was scaled according to Duncan's SES scale. Since the names given did not always correspond to a standard list of occupations, some of the attribution is arbitrary and may be subject to error. Moreover, it is not clear whether the SES scale is the best for our purposes or that it applies without further adjustment to the situation as it existed in 1928. We are currently reviewing our assignment procedures and are planning to experiment with alternative occupational scalings. This may lead to some changes in the results reported above.

The only other variable used in this paper is the age of brothers as of 1928, taken from Tables LXXXIX-XCII and the age at which they stopped going to school, from tables LXXX-LXXXV. More information was available on other characteristics of the sample, but in general it was not complete, not covering most of the brothers in the sample. Among other variables tried but not reported on in this paper was the rate at which schooling was completed as a measure of ability, test-scores for a subsample of brothers, and the birth-order of brothers.

Chapter 6

Extensions

I. Multi-Factor Models

Much of the methodology in Chapters 2 and 3 was confined to the one factor model. But before trying to develope general results for N factors we will want some guidance on what sorts of restrictions are reasonable to impose. Our empirical work in Chapters 4 and 5 made some beginnings in this direction. Chapter 5 considered a second factor with a purely family structure and developed some fairly useful bounds. Chapter 4 considered an extension to two distinct but correlated kinds of ability, scholastic (f_1) and economic (f_2). The substantive constraints are that f_1 is excluded from the income equation and f_2 is excluded from the schooling equation. The test is assumed to measure a combination of both kinds of ability. This model is not identified without proportionality constraints on the background variables and Section III of Chapter 2 shows that the problem cannot be solved by simply adding more indicators if they all depend on S.

I would like to sketch a possible attack, most of which is or soon will be operational using data from the National Longitudinal Survey (see Griliches (1974) for an overview of this data). The key is the availability of two tests which measure different combinations of the two unobservables:

$$
\begin{aligned}
T_1 &= & & \lambda_1 f_1 + \delta_1 f_2 + \nu_1 \\
S_1 &= & & \lambda_2 f_1 + \nu_2 \\
T_2 &= \gamma_{23} S_1 & & + \lambda_3 f_1 + \delta_3 f_2 + \nu_3 \\
S_2 &= \gamma_{24} S_1 & & + \lambda_4 f_1 + \nu_4 \\
Y_1 &= \quad \gamma_{45} S_2 & & + \delta_5 f_2 + \nu_5 \\
E &= \quad \gamma_{46} S_2 & & + \delta_6 f_2 + \nu_6 \\
Y_2 &= \quad \gamma_{47} S_2 + \gamma_{67} E & & + \delta_7 f_2 + \nu_7 \quad .
\end{aligned}
\tag{I.1}
$$

T_1 is the score on an IQ test given prior to the years of schooling represented by S_1 (or at least prior to the part of S_1 that shows significant variation in the sample). T_2 is the score on a test which differs from a standard IQ test and is given after S_1. The knowledge of the World of Work (KWW) test in the NLS data would seem to fulfill these requirements, with S_1 equal to years of schooling completed in 1966, the year the KWW test was given. Y_1 and Y_2 are the log of earnings in two different years, and E is a measure of job experience accumulated between the two years. There is also a variety of exogenous variables but we have surpressed them in order to see what sort of estimates can be obtained without the proportionality restrictions.

If $\lambda_1/\lambda_3 \neq \delta_1/\delta_3$, so that the two kinds of ability have differential effects on the two tests, then we can solve for f_2 in terms of T_1, T_2, and S_1. Then substituting this proxy into the Y_2 equation gives

$$(I.2) \qquad Y_2 = \gamma_{47}S_2 + \gamma_{67}E + \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 S_1 + \nu_7 - \alpha_1\nu_1 - \alpha_2\nu_3 \ .$$

This leaves an errors-in-variables problem in T_1 and T_2 and so we look for instruments. Y_1 is a candidate since the ν's are assumed to be uncorrelated. But Y_1 is the only excluded variable in (I.2) and two instruments are needed. More promising is a similar substitution in the Y_1 equation, since E and Y_2 can be used as instruments to identify γ_{45}. Then following the Corollary to Theorem 3, Chapter 2, we can use $\tilde{Y}_1 = Y_1 - \gamma_{45}S_2$ as a proxy for f_2 in the Y_2 equation with T_1, S_1, and T_2 as possible instruments. As in Theorem 3 a rank condition is needed to tell us which parameters are estimable from the IV equations.

II. A Parsimonious Model of Cross Equation Serial Correlation

A natural way to think about cross equation serial correlation is in terms of a common left out variable:

$$\text{(II.1)} \qquad y_{tk} = \underset{\sim}{x}_t{}'\underset{\sim}{\pi}_k + \varepsilon_{tk}$$

with $\varepsilon_{tk} = f_t d_k + \nu_{tk}$, $t = 1,\ldots, T$; $k = 1,\ldots, m$

where there are T observations, m equations, and the ν_{tk} are serially uncorrelated. An appropriate prior for the f_t could be based on a low order autoregressive-moving average process. Considerable analytic simplification of the likelihood function would be possible, along the lines of the Chapter 3 Appendix.

A more standard approach to this problem would be based on a matrix generalization of an autoregressive-moving average process: $\underset{\sim}{A}(L)\underset{\sim}{\varepsilon}_t = \underset{\sim}{C}(L)\ \underset{\sim}{w}_t$ where $\underset{\sim}{A}$ and $\underset{\sim}{C}$ are m x m low order matrix polynomials and $\underset{\sim}{w}_t$ is serially uncorrelated. The advantage of our approach is that it is much less parameter expensive.

As always, the troublesome question of how many factors must be faced. The answer will depend on the way in which the unobservable is being used. If it is "just" serial correlation which is to be swept out but not explained, e.g., to avoid biasing the coefficient of a lagged dependent variable, then we could try to estimate the number of factors by overfitting. But it may be that f is a substantive unobservable that we want to measure. An example could be using data on the term structure of interest rates along with price data in order to measure expected inflation.

III. ML Regions

In the errors-in-variables model of Chapter 2 (Theorem 4) and in model 4 in Chapter 3 we have a simple description of the ML region in the unidentified case. Our proof of Theorem 4, Chapter 2 shows that $\underset{\sim}{\Sigma}_2 - \tau\underset{\sim}{\sigma}_{21}\underset{\sim}{\sigma}_{21}'$ is positive definite. So $\tau \leq \min_{\underset{\sim}{\ell}} \underset{\sim}{\ell}'\underset{\sim}{\Sigma}_2\underset{\sim}{\ell} \,/(\underset{\sim}{\ell}'\underset{\sim}{\sigma}_{21})^2 = 1/(\underset{\sim}{\sigma}_{21}'\underset{\sim}{\Sigma}_2^{-1}\underset{\sim}{\sigma}_{21})$. Therefore the net reliability $\rho_N = 1/(\tau\sigma_{11}) \geqq \underset{\sim}{\sigma}_{21}'\underset{\sim}{\Sigma}_2^{-1}\underset{\sim}{\sigma}_{21}/\sigma_{11}$. Using $1-\rho = (1-\rho_N)(1-R^2_{y_1.\underset{\sim}{x}})$ together with $1-R^2_{y_1.\underset{\sim}{x},y_2\ldots y_m} = (1-R^2_{y_1.\underset{\sim}{x}})(1-\underset{\sim}{\sigma}_{21}'\underset{\sim}{\Sigma}_2^{-1}\underset{\sim}{\sigma}_{21}/\sigma_{11})$ gives the following bound on the reliability $\rho = 1 - \sigma_1^{\;2}/\sigma_{y_1}^{\;2}$: $0 \leqq \rho \geqq R^2_{y_1.\underset{\sim}{x},\, y_2,\ldots,y_m}$.
This interval for ρ generates the ML regions for the other structural parameters since the proof of Theorem 4 shows that given ρ (or τ) the reduced form can be uniquely solved for the structural parameters.

Corresponding to the formal equivalence between this errors-in-variables model and the replication model, we can apply a similar argument to the proof of Theorem 1, Chapter 3. There the bound on $\tau = \sigma_g^{\;2}/\sigma_f^{\;2}$ is that $\underset{\sim}{\Sigma} - \tau\underset{\sim}{d}\underset{\sim}{d}'$ is positive definite and so $\tau \leqq 1/\underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$. Since $\lambda = \sigma_f^{\;2}/(\sigma_f^{\;2} + \sigma_g^{\;2}) = 1/(1 + \tau)$, we have $0 \leqq \lambda \geqq \psi/(1 + \psi)$ where $\psi = \underset{\sim}{d}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{d}$. It is shown in the Chapter 3 Appendix that the ML estimate of $\psi/(1 + \psi)$ is $(\rho - \frac{1}{p})/(1 - \frac{1}{p})$ where ρ is the largest squared canonical correlation between $\underset{\sim}{y}' - \underset{\sim}{x}'\underset{\sim}{\Pi}$ and a set of group indicator dummy variables. In Section 6 of Chapter 5 $(\rho - \frac{1}{p})/(1 - \frac{1}{p})$ is interpreted as a generalized R^2. The bound on λ generates the bounds for the other structural parameters.

IV. A Production Function Example

Consider the following Cobb-Douglas production model:

$$\text{(IV.1)} \qquad y_{it} = \sum_n \beta_n x_{nit} + f_i + u_{it}, \qquad \begin{array}{l} i = 1,\ldots,q \\ t = 1,\ldots,T \end{array}$$

where y and the x's are the logs of output and the observable inputs, and f is intended to capture the effects of omitted inputs which do not vary over the sample period. In an agricultural context f could include measures of soil quality or average differences in climate. Another possibility is the quality of management or entrepreneurial capacity. The variable factors are determined by the following factor demand relationships:

$$\text{(IV.2)} \qquad p_{nit} + x_{nit} - y_{it} = v_{nit}, \qquad n = 1,\ldots,N$$

where the p's are the logs of the deflated factor prices. Note that we are suppressing the intercepts and will not be exploiting any information they may contain, as in Klein's (1953) factor share method. Thus we can allow for (or test) imperfections in the product or factor markets in the form of constant demand or supply elasticities. Also we will not have to make arithmetic vs. geometric mean distinctions in specifying that firm's maximize "on average."

Clearly there is no identification problem if we can observe exogenous factor price variation. We can simply use the prices as instruments with the f_i picked up by a set of firm dummy variables. But to take advantage of an informative prior for the

f_i, we have to face a simultaneity problem. For under decreasing returns ($\sum_n \beta_n < 1$) the firms with more of the fixed factors will use more of the variable factors. So we set up a reduced form and try applying our prior there:

$$\text{(IV.3)} \qquad y_{it} = -\eta\sum_n \beta_n p_{nit} + \eta f_i + \eta u_{it} + \eta \sum_n \beta_n v_{nit}$$

$$x_{kit} = -\eta\sum_n \beta_n p_{nit} - p_{kit} + \eta f_i + \eta u_{it} + \eta\sum_n \beta_n v_{nit} + v_{kit} \; ,$$

$$k = 1,\ldots,N$$

with $\eta = (1/(1-\sum_n \beta_n)$.

Note that the output elasticities can be identified from a covariance analysis of any one of the reduced form equations. We can simultaneously exploit all of the equations together with some of the between firm variation by applying the GLS estimator in (III.7), Chapter 3.

First reparameterize in terms of y and the logarithmic factor shares $s_n = p_n + x_n - y$:

$$\text{(IV.4)} \qquad y_{it} = -\eta\sum_n \beta_n p_{nit} + \eta f_i + \eta u_{it} + \eta \sum_n \beta_n v_{nit}$$

$$s_{nit} = v_{nit}, \qquad n = 1,\ldots,N \; .$$

This fits our model 1 framework with $\underset{\sim}{\Sigma}$ unrestricted and $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$. There is also the restriction that $d_k = 0$ for $k \geq 2$.

$\underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix}$ can be obtained from the GLS estimate of the

reduced form slope coefficients:

$$\text{(IV.5)}\qquad \underset{\sim}{\delta}^{*} = (\underset{\sim}{H}_{W} + \underset{\sim}{H}_{B})^{-1}(\underset{\sim}{H}_{W}\underset{\sim}{\delta}_{W}^{GLS} + \underset{\sim}{H}_{B}\underset{\sim}{\delta}_{B}^{GLS})$$

$$\underset{\sim}{H}_{W} = \underset{\sim}{\Sigma}^{-1} * \underset{\sim}{\overline{W}}$$

$$\underset{\sim}{H}_{B} = \frac{1}{T}(\underset{\sim}{\Theta} + \frac{1}{T}\underset{\sim}{\Sigma})^{-1} * \underset{\sim}{\overline{B}}.$$

The GLS estimator $\underset{\sim}{\delta}^{*}$ pools two other GLS estimators, $\underset{\sim}{\delta}_{W}^{GLS}$ and $\underset{\sim}{\delta}_{B}^{GLS}$. The within firm $\underset{\sim}{\delta}_{W}^{GLS}$ corresponds to an efficient use of Mundlak's (1963) analysis of covariance approach, recognizing that each of the reduced form equations is informative about $\underset{\sim}{\beta}$. The other term, $\underset{\sim}{\delta}_{B}^{GLS}$, is new; it reflects the exchangeable prior bringing in some of the between firm variation. This may be quite valuable if most of the sample variation is between firms, reflecting location differences, etc. With firm effects h_{ki} in the factor demand equations we would have the multi-factor version of model 1. The GLS estimates would still be given by (IV.5) but $\underset{\sim}{\Theta}$ would be less constrained.

I next want to take up the Marachak-Andrews (1944) case in which there is no observable price variation. The estimation techniques generated by this extreme case are quite relevant to panel data since much of the price variation may reflect permanent location differences which are confounded with the firm effects f_i. Also the observed price variation may be mostly quality variation. For example, let X_1 be hours when in fact the relevant quantity variable is $\tilde{X}_{1it} = Q_i X_{it}$, where Q_i is a labor augmenting qualtiy index reflecting average labor quality in the i^{th} firm. Then the relevant price variable is obtained by dividing the total wage bill by the number of "efficiency units" of labor

$$\tilde{P}_1 = P_1 X_1 / \tilde{X}_1 \ .$$

So in logs we have

$$\text{(IV.6)} \qquad \tilde{x}_{1it} = x_{1it} + q_i$$

$$\tilde{p}_{1it} = p_{1it} - q_i$$

We have to add $\beta_1 q_i$ to the structural form of the production function in (IV.1). But this can be absorbed in f_i. The factor demand relationships don't have to be altered because $\tilde{p}_1 + \tilde{x}_1 = p_1 + x_1$; i.e., the total factor compensation is correctly measured. Problems arise only in trying to disaggregate the wage bill into a price and a quantity. So we need methods which do not depend on such a division.

First we will look at Mundlak's (1963) modification of Hoch's (1958) direct least squares approach. Hoch's idea was that if the disturbance in the production function is random not only to the econometrician but also to the firm, then it will not be "transmitted" to the factor demand decisions. In that case we can rewrite (IV.2) as

$$p_n + x_n - (y - u) = v_n$$

or

$$\text{(IV.7)} \qquad s_n = v_n - u \ , \qquad\qquad n = 1,\ldots,N.$$

As Mundlak pointed out, this assumption becomes more tenable when we have replication on the firms, thereby allowing us to distinguish the part of the residual that is of a more permanent nature. For the firm effects f_i, although random to me are probably known to the firm and hence transmitted to the factor demand system. Under these assumptions $\underset{\sim}{\beta}$ can be estimated from a covariance analysis of the structural form of the production function. But using just the within firm deviations may throw out most of the sample information. The cure can be worse than the disease.

In order to see what improvements are possible we use the following version of the reduced form:

$$\text{(IV.8)} \qquad y_{it} = \eta f_i + u_{it} + \eta \sum_n \beta_n v_{nit}$$

$$x_{kit} = \eta f_i + \eta \sum_n \beta_n v_{nit} + v_{kit}, \qquad k = 1,\ldots,N.$$

This fits our modell framework with $\underset{\sim}{\Theta} = \underset{\sim}{d}\underset{\sim}{d}'$ and all of the elements of $\underset{\sim}{d}$ restricted to be equal. Assuming that $\underset{\sim}{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$ is independent of $\underset{\sim}{u}$ with $E(\underset{\sim}{v}\underset{\sim}{v}') = \underset{\sim}{V}$, we have

$$\text{(IV.9)} \qquad \Sigma_{11} = \eta^2(\underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta}) + \sigma_u^2$$

$$\underset{\sim}{\Sigma}_{12} = \underset{\sim}{\Sigma}'_{21} = \eta^2(\underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta})\underset{\sim}{\ell}'_N + \eta\underset{\sim}{\beta}'\underset{\sim}{V}$$

$$\underset{\sim}{\Sigma}_{22} = \underset{\sim}{V} + \eta(\underset{\sim}{\ell}_N\underset{\sim}{\beta}'\underset{\sim}{V} + \underset{\sim}{V}\underset{\sim}{\beta}\underset{\sim}{\ell}'_N) + \eta^2(\underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta})\underset{\sim}{\ell}_N\underset{\sim}{\ell}'_N .$$

Given $\underset{\sim}{\Sigma}$ we can uniquely solve for $\underset{\sim}{\beta}$, $\underset{\sim}{V}$, and σ_u^2. For example $\underset{\sim}{\Sigma}_{12} = \underset{\sim}{\beta}'\underset{\sim}{\Sigma}_{22}$ and so

$$\text{(IV.10)} \qquad \underset{\sim}{\beta} = \underset{\sim}{\Sigma}_{22}^{-1}\underset{\sim}{\Sigma}_{21} .$$

This is just OLS using the constrained within firm moments ($\underset{\sim}{\Sigma} = \underset{\sim}{R} - \underset{\sim}{d}\underset{\sim}{d}'$). It differs from Mundlak's estimator in that instead of $\underset{\sim}{R} - \overline{\underset{\sim}{R}}$ we only subtract off $\underset{\sim}{d}\underset{\sim}{d}'$, a matrix of rank one, thereby using more of the between firm variation.

Next we will consider the case in which all of the production function disturbance u_{it} is transmitted to the factor demand equations. We assume as before that v is independent of u with an arbitrary covariance structure $E(\underset{\sim}{v}\underset{\sim}{v}') = \underset{\sim}{V}$. Mundlak argued that the independence of u and v is more plausible after removing firm effects. Then the logarithmic factor shares can be used as instruments for the x's in an equation with firm dummies. So Mundlak's suggestion is to apply the Hoch-Thiel instrumental variable estimator to the within firm variation.

Our extension is based on assigning the f_i an exchangeable prior in the following reduced form:

$$\text{(IV.11)} \qquad y_{it} = \eta f_i + \eta u_{it} + \eta \sum_n \beta_n v_{nit}$$

$$x_{kit} = \eta f_i + \eta u_{it} + \eta \sum_n \beta_n v_{nit} + v_{kit} \qquad k = 1,\dots,N.$$

We have the same restrictions on d as before but now

$$\text{(IV.12)} \qquad \Sigma_{11} = \eta^2(\sigma_u^2 + \underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta})$$

$$\underset{\sim}{\Sigma}_{12} = \underset{\sim}{\Sigma}_{21}' = \eta^2(\sigma_u^2 + \underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta})\underset{\sim}{\ell}_N' + \eta\underset{\sim}{\beta}'\underset{\sim}{V}$$

$$\underset{\sim}{\Sigma}_{22} = \underset{\sim}{V} + \eta(\underset{\sim}{\ell}_N\underset{\sim}{\beta}'\underset{\sim}{V} + \underset{\sim}{V}\underset{\sim}{\beta}\underset{\sim}{\ell}_N') + \eta^2(\sigma_u^2 + \underset{\sim}{\beta}'\underset{\sim}{V}\underset{\sim}{\beta})\underset{\sim}{\ell}_N\underset{\sim}{\ell}_N' \ .$$

Note that

$$\underset{\sim}{\beta}'(\underset{\sim}{\Sigma}_{22} - \underset{\sim}{\Sigma}_{12}\underset{\sim}{\ell}_N') = \underset{\sim}{\Sigma}_{12} - \Sigma_{11}\underset{\sim}{\ell}_N'$$

and so we can solve for $\underset{\sim}{\beta}$:

$$(\text{IV.13}) \qquad \underset{\sim}{\beta} = (\underset{\sim}{\Sigma}_{22} - \underset{\sim}{\ell}_N \underset{\sim}{\Sigma}_{21})^{-1}(\underset{\sim}{\Sigma}_{21} - \underset{\sim}{\ell}_N \Sigma_{11}).$$

This is the Hoch-Thiel estimator based on the constrained within firm moments. As in (IV.10) we are using $\underset{\sim}{R} - \underset{\sim}{d}\underset{\sim}{d}'$ instead of the unconstrained $\underset{\sim}{R} - \underset{\sim}{\overline{R}}$.

It is disturbing that the appropriate estimation technique depends so critically on whether or not u is transmitted. The technique which is consistent for one case is not for the other. So we want to develop a more robust approach. Also we have neglected the possibility of a firm structure in the factor demand residuals. One could argue that there "shouldn't" be persistent errors in choosing factor ratios; but in fact firm effects have been observed in factor demand relationships (e.g., Ringstad (1971)). Also they could reflect demand and supply elasticities differing across firms. In any event these firm effects provide another potential source of identifying information.

The problem can be formulized by considering a model with partial transmission (cf. Mundlak and Hoch (1965)). Mundlak (1963) has shown that this case can arise from aggregation over different stages of the production process. We decompose u_{it} into $u_{1it} + u_{2it}$ and assume that only u_{1it} is transmitted. This model cannot be identified from the within firm variation $\underset{\sim}{\Sigma}$. For before there was a one-to-one relationsip between $\underset{\sim}{\Sigma}$ and the structural parameters, leaving no degrees of freedom to determine how σ_u^2 splits into $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$. But there is some hope if we can modify the factor share equations to include firm effects:

$$\text{(IV.14)} \quad s_{nit} = p_{nit} + x_{nit} - y_{nit} = h_{ni} + v_{nit}. \qquad n = 1,\ldots,N.$$

Then the crucial question is what distributional assumptions to make for f and $\underset{\sim}{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_N \end{pmatrix}$. I want to argue that it's reasonable to assume that $\underset{\sim}{h}$ is independent of f. This may seem implausible since the h's are to some extent profit maximization errors and may reflect the same underlying managerial ability that is contained in f. But it's the absolute value of h_n that reflects how well a first order condition is being satisfied, with allocative ability inversely related to $|h_n|$. So we need not expect a simple linear relationship between f and h_n; e.g., $|h_n|$ could be an exact function of f but so long as the sign of h_n is independent of f there will be no correlation between f and h_n. Furthermore, as Welch (1970) and Nelson (1970) have emphasized, it may be incorrect to think of f as primarily reflecting entrepreneurial skill. For f is the addition to output holding other inputs constant when, in fact, the true contribution of entrepreneurial skill may be in choosing the proper levels for the other inputs.

Although we cannot recover the output elasticities from $\underset{\sim}{\Sigma}$, we now have a much more interesting between firm $\underset{\sim}{\Theta}$. For the vector of reduced form firm effects is

$$\text{(IV.15)} \quad \begin{bmatrix} \eta f_i + \eta \sum_n \beta_n h_{ni} \\ \eta f_i + \eta \sum_n \beta_n h_{ni} + h_{1i} \\ \vdots \\ \eta f_i + \eta \sum_n \beta_n h_{ni} + h_{Ni} \end{bmatrix}$$

which is formally identical to the within firm effects in the complete transmission case. So $\underset{\sim}{\Theta}$ has the same structure as $\underset{\sim}{\Sigma}$ in (IV.12) and analagous to (IV.13) we have

$$\text{(IV.16)} \quad \underset{\sim}{\beta} = (\underset{\sim}{\Theta}_{22} - \underset{\sim}{\ell}_N \underset{\sim}{\Theta}_{21})^{-1} (\underset{\sim}{\Theta}_{21} - \ell_N \Theta_{11}).$$

This is the Hoch-Thiel method applied to the between firm variation. It has some intuitive appeal relative to applying it to the within firm $\underset{\sim}{\Sigma}$. First, most of the relevant variation may be in $\underset{\sim}{\Theta}$, reflecting permanent location differences that are swept away in $\underset{\sim}{\Sigma}$. Second, it is easier to specify how much of the residual is transmitted. For since the firm effects are relatively constant, it's reasonable to assume they are not random to the firm and hence are fully transmitted.

Bibliography

Anderson, T.W. and H. Rubin, 1956. "Statistical Inference in Factor Analysis" in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, J. Neyman ed., Berkeley: University of California Press, 11-150.

Balestra, P. and M. Nerlove, 1966. "Pooling Cross Section and Time Series Data in the Stimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, 34, 585-612.

Brundy, J.M and D.W. Jorgenson, 1974. "Consistent and Efficient Estimation of Systems of Simultaneous Equations by Means of Instrumental Variables," in P. Zarembka ed., *Frontiers of Econometrics*, New York: Academic Press.

Cagan, P., 1965. "Measuring Quality Changes and the Purchasing Power of Money: An Exploratory Study of Automobiles," *National Banking Review*, 3, 217-236. (Reprinted in Griliches, 1971).

Chamberlain, G. and Z. Griliches, 1974. "Unobservables with a Variance Components Structure: Ability, Schooling and the Economic Success of Brothers," *International Economic Review*, forthcoming.

DeFinetti, B., 1937. "La Prévision: Ses Lois Logiques, ses Sources Subjectives," *Ann. Inst. H. Poincaré* Vol. 7, 1. English Translation in Kyburg and Smokler (1964).

Dhondt, A., 1960. "Sur une Généralisation d'une Théorème de R. Frisch en Analyse de Conluence, "*Chahiers du Centre d'Etudes de Recherche Opérationnelle,* 2(1), Brussels.

Duesenberry, J.S., 1952. *Income, Saving, and the Theory of Consumer Behavior.* Cambridge, Mass.: Harvard University Press.

Duncan, O.D., 1968. "Ability and Achievement," *Eugenics Quarterly,* 15, 1-11.

Eisner, R., 1958. "The Permanent Income Hypothesis: A Comment," *American Economic Review*, 48, 972-990.

Friedman, M., 1957. *A Theory of the Consumption Function*. New York: National Bureau of Economic Research (distr. Princeton University Press).

Frisch, R., 1934. *Statistical Concluence Analysis by Means of Complete Regression Systems.* Oslo: University Economics Institute, Publication No. 5.

Gantmacher, F. R., 1959. *Matrix Theory*, Vol. 1. New York: Chelsea.

Geraci, V., 1974. "Simultaneous Equation Models with Measurement Error," SSRI Workshop Series 7407, Madison: University of Wisconsin. unpublished.

Geraci, V. and A.S. Goldberger, 1971. "Simultaneity and Measurement Error," SSRI Workshop Series 7125, Madison: University of Wisconsin, unpublished.

Goldberger, A.S., 1971. "Econometrics and Psychometrics: A Survey of Communalities," Psychometrika, 36, 83-107.

__________, 1972. "Maximum LIkelihood Estimation of Regression Models Containing Unobservable Variables," International Economic Review, 13(1), 1-15.

__________, 1972a. "Structural Equation Methods in the Social Sciences," Econometrica, 40(6), 976-1002.

__________, 1974. "Unobservable Variables in Econometrics" in P. Zarembka ed., Frontiers of Econometrics, New York: Academic Press.

Good, I.J., 1965. The Estimation of Probabilities: An Essay on Modern Bayesian Methods, Research Monograph No. 30. Cambridge, Mass.: MIT Press.

Goreseline, D.E., 1932. The Effect of Schooling Upon Income. Bloomington: Indiana University Press.

Griliches, Z., 1970. "Notes on the Role of Education in Production Functions and Growth Accounting," in W.L. Hansen ed., Education, Income and Human Capital, NBER Studies in Income and Wealth, Vol. 35, 71-127.

__________ , ed., 1971. Price Indices and Quality Change, Cambridge, Mass.; Harvard University Press.

__________ , 1973. "Errors in Variables and Other Unobservables," Econometrica, 42(6), 971-998.

Griliches, Z. and W.M. Mason, 1972. "Education, Income, and Ability," Journal of Political Economy, 80(3), Part II, S47-103.

Hall, R.E., 1969, 1971. "The Measurement of Quality Change from Vintage Price Data, "Working Paper 144, C.R.M.S. and I.B.E.R., University of California (Berkeley). Reprinted in Griliches (1971).

Hannan, E.J., 1967. "Canonical Correlation and Multiple Equation Systems in Economics," Econometrica, 35(1), 123-138.

Hauser, R.M. and A.S. Goldberger, 1971. "The Treatment of Unobservable Variables in Path Analysis," in H.L. Costner ed., Sociological Methodology 1971. San Francisco: Jassey-Bass, 81-117.

Harberger, A.C., 1953. "On the Estimation of Economic Parameters," Cowles Commission Discussion Paper No. 2088, Chicago. unpublished.

Hewitt, E. and L.J. Savage, 1955. "Symmetric Measures on Cartesian Products," Transactions of the American Mathematical Society, 80, 470-501.

Holbrook, R. and F. Stafford, 1971. "The Propensity to Consume Separate Types of Income, A Generalized Permanent Income Hypothesis," Econometrica, 39(1), 1-22.

Howe, W.G., 1955. "Some Contributions to Factor Analysis," Report No. ORNL-1919, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Hurwicz, L. and T.W. Anderson, 1946. "Statistical Models with Disturbances in Equations and/or Disturbances in Variables," unpublished Cowles Commission memoranda in four parts: I Introduction by L Hurwicz, II Contemporaneous Systems by T.W. Anderson, III Lagged Systems by L. Hurwicz, and IV Some Notes on Tintner's Statistical Methods by T.W. Anderson.

Jencks, C. et al., 1972. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basis Books.

Jennrich, R.I. and S.M. Robinson, 1969. "A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis," Psychometrika, 34, 11-123.

Jöreskog, K.G., 1967. "Some Contributions to Maximum Likelihood Factor Analysis," Psychometrika, 32, 443-482.

__________, 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis," Psychometrika, 34, 183-202.

__________, 1970. "A General Method for Analysis of Covariance Structures," Biometrika, 57, 239-251.

__________, 1970a. "Factoring the Multitest-Multioccassion Correlation Matrix," in C.E. Lunneborg ed., Current Problems and Techniques in Multivariate Psychology. Seattle: University of Washington Press, 68-100.

__________, 1973. "Analysis of Covariance Structures," in P.E. Krishnaiah ed., Multivariate Analysis - III. New York: Academic Press.

Jöreskog, K. and A.S. Goldberger, 1973. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," SSRI Workshop Series 7328, Madison: University of Wisconsin. unpublished.

Keller, W.J., 1973. "A New Class of Limited Information Estimators for Simultaneous Equation Systems," Econometric Institute, Erasmas University, Rotterdam. unpublished.

Klein, L.R., 1953. A Textbook of Econometrics. Evaston: Row, Peterson and Company.

Kuh, E., 1959. "The Validity of Cross-Sectionally Estimated Behavior Equations in Time Series Applications," Econometrica, 27, 197-214.

Kyburg, H.E. and H.G. Smokler, eds., 1964. Studies in Subjective Probability. New York: Wiley.

Lawley, D.N., 1940. "The Estimation of Factor Loadings by the Method of Maximum Likelihood," Proceedings of the Royal Society of Edinburgh, Vol. 60.

Leamer, E.E., 1972. "A Class of Informative Priors and Distributed Lag Analysis," Econometrica, 40(6), 1059-1082.

Liviatan, N., 1961. "Erros in Variables and Engel Curve Analysis," Econometrica, 29, 336-362.

__________, 1963. "Tests of the Permanent-Income Hypothesis Based on a Reinterview Savings Survey," in C.F. Christ et al., Measurement in Economics, Studies in Memory of Yehuda Grunfeld. Stanford: Stanford University Press, 29-66.

Maddala, G.S., 1971. "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," Econometrica, 39(2), 341-358.

Marschak, J. and W.H. Andrews, 1944. "Random Simultaneous Equations and the Theory of Production," Econometrica, 12(3-4), 143-206.

Mazodier, P.A., 1971. The Econometrics of Error Components Models. unpublished Harvard University PhD dissertation.

Mundlak, Y., 1963. "Estimation of Production and Behavioral Relations from a Combination of Cross-Section and Time Series Data," in C.F. Christ et al., Measurement in Economics, Studies in Memory of Yehuda Grunfeld. Stanford: Stanford University Press, 138-166.

Mundlak, Y. and I. Hoch, 1965. "Consequences of Alternative Specifications in Estimation of Cobb-Douglas Production Functions," Econometrica, 33(4), 814-828.

Nelson, R.R., 1970. "Comments," in W.L. Hansen ed., Education, Income, and Human Capital, NBER Studies in Income and Wealth, Vol. 35, 124-127.

Nerlove, M., 1971. "A Note on Error Components Models," Econometrika, 39(2), 383-396.

Ohta, M. and Z. Griliches, 1973. "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis," Harvard Institute of Economic Research Discussion Paper No. 325, Cambridge, Mass., unpublished.

Rao, C.R. and S.K. Mitra, 1971. Generalized Inverse of Matrices and its Applications. New York: Wiley.

Rao, C.R., 1973. (second edition). Linear Statistical Inference and its Applications. New York: Wiley.

Riersol, O., 1945. "Confluence Analysis by Means of Instrumental Sets of Variables," *Arkiv for Mathematic, Astronomi och Fysik,* Vol. 32.

Ringstad, V., 1971. *Estimating Production Functions and Technical Change from Micro Data. An Exploratory Study of Individual Establishment Time Series from Norwegian Mining and Manufacturing 1959-1967.* Oslo: Central Bureau of Statistics.

Robinson, P.M., 1974. "Identification, Estimation and Large-Sample Theory for Regressions Containing Unobservable Variables," Harvard University, unpublished.

Schiller, R.J., 1973. "A Distributed Lag Estimator Derived from Smoothness Priors," *Econometrika,* 41(4), 775-788.

Uppsala Symposium on Psychological Factor Analysis, 1953. *Nordisk Psykologi's Monograph Series,* No. 3.

Wald, A., 1940. "The Fitting of Straight Lines if Both Variables are Subject to Errors," *Annals of Mathematical Statistics,* 11, 284-300.

Wallace, T.D. and A. Hussein, 1969. "The Use of Error Components Models in Combining Cross Section with Time Series Data," *Econometrika,* 37(1), 55-72.

Welch, R., 1970. "Education in Production," *Journal of Political Economy,* 78, 35-59.

Whittle, P., 1953. "A Principal Components and Least Squares Method of Factor Analysis," Skandinavisk Aktuarietidskrift, 35, 223-239.

Wilkinson, J.H., 1965. *The Algebraic Eigenvalue Problem.* Oxford: Oxford University Press.

Zellner, A., 1970. "Estimation of Regression Relationships Containing Unobservable Independent Variables," *International Economic Review,* 11, 441-454.