Chapter 11

Multivariate Regression

11.1 Introduction

Multivariate regression is a system of regression equations. Multivariate regression is used as reduced form models for instrumental variable estimation (Chaper 12), vector autoregressions (Chapter 15), demand systems (demand for multiple goods), and other contexts.

Multivariate regression is also called by the name **systems of regression equations**. Closely related is the method of **Seemingly Unrelated Regressions** (SUR) introduced in Section 11.7.

Most of the tools of single equation regression generalize to multivariate regression. A major difference is a new set of notation to handle matrix estimators.

11.2 Regression Systems

A univariate linear regression equation equals $Y = X'\beta + e$ where Y is scalar and X is a vector. **Multivariate regression** is a system of m linear regressions, and equals

$$Y_j = X_j'\beta_j + e_j \tag{11.1}$$

for $j = 1, ..., m$. Here we use the subscript j to denote the j^{th} dependent variable, not the i^{th} individual. As an example, Y_j could be expenditures by a household on good category j (e.g., food, housing, transportation, clothing, recreation). The regressor vectors X_j are $k_j \times 1$ and e_j is an error. The coefficient vectors β_j are $k_j \times 1$. The total number of coefficients are $\overline{k} = \sum_{j=1}^{m} k_j$. The regressors can be common across j or can vary across j. In the household expenditure example the regressors X_j are typically common across j, and include variables such as household income, number and ages of family members, and demographic characteristics. The regression system specializes to univariate regression when $m = 1$.

Define the $m \times 1$ error vector $e = (e_1, ..., e_m)'$ and its $m \times m$ covariance matrix $\Sigma = \mathbb{E}\left[ee'\right]$. The diagonal elements are the variances of the errors e_j and the off-diagonals are the covariances across variables.

We can group the m equations (11.1) into a single equation as follows. Let $Y = (Y_1, ..., Y_m)'$ be the $m \times 1$ vector of dependent variables. Define the $m \times \overline{k}$ matrix of regressors

$$\overline{X} = \begin{pmatrix} X_1' & 0 & \cdots & 0 \\ \vdots & X_2' & & \vdots \\ 0 & 0 & \cdots & X_m' \end{pmatrix}$$

and the $\overline{k} \times 1$ stacked coefficient vector

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

The m regression equations can be jointly written as

$$Y = \overline{X}\beta + e. \tag{11.2}$$

This is a system of m equations.

For n observations the joint system can be written in matrix notation by stacking. Define

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \qquad \overline{\boldsymbol{X}} = \begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_n \end{pmatrix}$$

which are $mn \times 1$, $mn \times 1$, and $mn \times \overline{k}$, respectively. The system can be written as $\boldsymbol{Y} = \overline{\boldsymbol{X}}\beta + \boldsymbol{e}$.

In many applications the regressor vectors X_j are common across the variables j, so $X_j = X$ and $k_j = k$. By this we mean that the same variables enter each equation with no exclusion restrictions. Several important simplifications occur in this context. One is that we can write (11.2) using the notation

$$Y = \boldsymbol{B}'X + e \tag{11.3}$$

where $\boldsymbol{B} = \left(\beta_1, \beta_2, \cdots, \beta_m\right)$ is $k \times m$. Another is that we can write the joint system of observations in the $n \times m$ matrix notation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1' \\ \vdots \\ Y_n' \end{pmatrix}, \qquad \boldsymbol{E} = \begin{pmatrix} e_1' \\ \vdots \\ e_n' \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}.$$

Another convenient implication of common regressors is that we have the simplification

$$\overline{X} = \begin{pmatrix} X' & 0 & \cdots & 0 \\ 0 & X' & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & X' \end{pmatrix} = \boldsymbol{I}_m \otimes X'$$

where $\otimes$ is the Kronecker product (see Appendix A.21).

11.3 Least Squares Estimator

The equations (11.1) can be estimated by least squares. This takes the form

$$\widehat{\beta}_j = \left(\sum_{i=1}^{n} X_{ji}X_{ji}'\right)^{-1}\left(\sum_{i=1}^{n} X_{ji}Y_{ji}\right).$$

An estimator of β is the stacked vector

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_m \end{pmatrix}.$$

We can alternatively write this estimator using the systems notation

$$\widehat{\beta} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{Y}\right) = \left(\sum_{i=1}^{n}\overline{X}_i'\overline{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'Y_i\right). \tag{11.4}$$

To see this, observe that

$$\begin{aligned}
\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}} &= \begin{pmatrix} \overline{X}_1' & \cdots & \overline{X}_n' \end{pmatrix}\begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_n \end{pmatrix} \\
&= \sum_{i=1}^{n}\overline{X}_i'\overline{X}_i \\
&= \sum_{i=1}^{n}\begin{pmatrix} X_{1i} & 0 & \cdots & 0 \\ \vdots & X_{2i} & & \vdots \\ 0 & 0 & \cdots & X_{mi} \end{pmatrix}\begin{pmatrix} X_{1i}' & 0 & \cdots & 0 \\ \vdots & X_{2i}' & & \vdots \\ 0 & 0 & \cdots & X_{mi}' \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n}X_{1i}X_{1i}' & 0 & \cdots & 0 \\ \vdots & \sum_{i=1}^{n}X_{2i}X_{2i}' & & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^{n}X_{mi}X_{mi}' \end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\overline{\boldsymbol{X}}'\boldsymbol{Y} &= \begin{pmatrix} \overline{X}_1' & \cdots & \overline{X}_n' \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\
&= \sum_{i=1}^{n}\overline{X}_i'Y_i \\
&= \sum_{i=1}^{n}\begin{pmatrix} X_{1i} & 0 & \cdots & 0 \\ \vdots & X_{2i} & & \vdots \\ 0 & 0 & \cdots & X_{mi} \end{pmatrix}\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{mi} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n}X_{1i}Y_{1i} \\ \vdots \\ \sum_{i=1}^{n}X_{mi}Y_{mi} \end{pmatrix}.
\end{aligned}$$

Hence

$$\begin{aligned}
\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{Y}\right) &= \left(\sum_{i=1}^{n}\overline{X}_i\overline{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_iY_i\right) \\
&= \begin{pmatrix} \left(\sum_{i=1}^{n}X_{1i}X_{1i}'\right)^{-1}\left(\sum_{i=1}^{n}X_{1i}Y_{1i}\right) \\ \vdots \\ \left(\sum_{i=1}^{n}X_{mi}X_{mi}'\right)^{-1}\left(\sum_{i=1}^{n}X_{mi}Y_{mi}\right) \end{pmatrix} \\
&= \widehat{\beta}
\end{aligned}$$

as claimed.

The $m \times 1$ residual vector for the i^{th} observation is $\widehat{e}_i = Y_i - \overline{\boldsymbol{X}}_i'\widehat{\beta}$. The least squares estimator of the $m \times m$ error covariance matrix is

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i \widehat{e}_i'. \tag{11.5}$$

In the case of common regressors, the least squares coefficients can be written as

$$\widehat{\beta}_j = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_{ji}\right)$$

and

$$\widehat{\boldsymbol{B}} = \left(\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_m\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right). \tag{11.6}$$

In Stata, multivariate regression can be implemented using the `mvreg` command.

11.4 Expectation and Variance of Systems Least Squares

We can calculate the finite-sample expectation and variance of $\widehat{\beta}$ under the conditional expectation assumption

$$\mathbb{E}[e \mid X] = 0 \tag{11.7}$$

where X is the union of the regressors X_j. Equation (11.7) is equivalent to $\mathbb{E}\left[Y_j \mid X\right] = X_j'\beta_j$, which means that the regression model is correctly specified.

We can center the estimator as

$$\widehat{\beta} - \beta = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{e}\right) = \left(\sum_{i=1}^{n} \overline{X}_i'\overline{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \overline{X}_i' e_i\right).$$

Taking conditional expectations we find $\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta$. Consequently, systems least squares is unbiased under correct specification.

To compute the variance of the estimator, define the conditional covariance matrix of the errors of the i^{th} observation $\mathbb{E}\left[e_i e_i' \mid X_i\right] = \Sigma_i$ which in general is a function of X_i. If the observations are mutually independent then

$$\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right] = \mathbb{E}\left[\left.\begin{pmatrix} e_1 e_1' & e_1 e_2' & \cdots & e_1 e_n' \\ \vdots & \ddots & & \vdots \\ e_n e_1' & e_n e_2' & \cdots & e_n e_n' \end{pmatrix}\right| \boldsymbol{X}\right] = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \Sigma_n \end{pmatrix}.$$

Also, by independence across observations,

$$\operatorname{var}\left[\left.\sum_{i=1}^{n} \overline{X}_i' e_i \right| \boldsymbol{X}\right] = \sum_{i=1}^{n} \operatorname{var}\left[\left.\overline{X}_i' e_i \right| X_i\right] = \sum_{i=1}^{n} \overline{X}_i' \Sigma_i \overline{X}_i.$$

It follows that

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n} \overline{X}_i'\Sigma_i\overline{X}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

When the regressors are common so that $\overline{X}_i = \boldsymbol{I}_m \otimes X_i'$ then the covariance matrix can be written as

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{I}_m \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(\Sigma_i \otimes X_i X_i'\right)\right)\left(\boldsymbol{I}_m \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

If the errors are conditionally homoskedastic

$$\mathbb{E}\left[ee' \mid X\right] = \Sigma \tag{11.8}$$

then the covariance matrix simplifies to

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'\Sigma\overline{X}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

If both simplifications (common regressors and conditional homoskedasticity) hold then we have the considerable simplication

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \Sigma \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

11.5 Asymptotic Distribution

For an asymptotic distribution it is sufficient to consider the equation-by-equation projection model in which case

$$\mathbb{E}\left[X_j e_j\right] = 0. \tag{11.9}$$

First, consider consistency. Since $\widehat{\beta}_j$ are the standard least squares estimators, they are consistent for the projection coefficients β_j.

Second, consider the asymptotic distribution. Our single equation theory implies that the $\widehat{\beta}_j$ are asymptotically normal. But this theory does not provide a joint distribution of the $\widehat{\beta}_j$ across j, which we now derive. Since the vector

$$\overline{X}_i' e_i = \begin{pmatrix} X_{1i}e_{1i} \\ \vdots \\ X_{mi}e_{mi} \end{pmatrix}$$

is i.i.d. across i and mean zero under (11.9), the central limit theorem implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overline{X}_i' e_i \underset{d}{\longrightarrow} \mathrm{N}(0,\Omega)$$

where

$$\Omega = \mathbb{E}\left[\overline{X}_i' e_i e_i' \overline{X}_i\right] = \mathbb{E}\left[\overline{X}_i'\Sigma_i\overline{X}_i\right].$$

The matrix Ω is the covariance matrix of the variables $X_{ji}e_{ji}$ across equations. Under conditional homoskedasticity (11.8) the matrix Ω simplifies to

$$\Omega = \mathbb{E}\left[\overline{X}_i'\Sigma\overline{X}_i\right] \tag{11.10}$$

(see Exercise 11.1). When the regressors are common it simplies to

$$\Omega = \mathbb{E}\left[ee' \otimes XX'\right] \tag{11.11}$$

(see Exercise 11.2). Under both conditions (homoskedasticity and common regressors) it simplifies to

$$\Omega = \Sigma \otimes \mathbb{E}\left[XX'\right] \tag{11.12}$$

(see Exercise 11.3).

Applied to the centered and normalized estimator we obtain the asymptotic distribution.

Theorem 11.1 Under Assumption 7.2, $\sqrt{n}\left(\widehat{\beta}-\beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$ where $\boldsymbol{V}_\beta = \boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1}$ and

$$\boldsymbol{Q} = \mathbb{E}\left[\overline{X}'\overline{X}\right] = \begin{pmatrix} \mathbb{E}\left[X_1X_1'\right] & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \mathbb{E}\left[X_mX_m'\right] \end{pmatrix}.$$

For a proof, see Exercise 11.4.

When the regressors are common the matrix $\boldsymbol{Q}$ simplifies as

$$\boldsymbol{Q} = \boldsymbol{I}_m \otimes \mathbb{E}\left[XX'\right] \tag{11.13}$$

(See Exercise 11.5).

If both the regressors are common and the errors are conditionally homoskedastic (11.8) then we have the simplification

$$\boldsymbol{V}_\beta = \Sigma \otimes \left(\mathbb{E}\left[XX'\right]\right)^{-1} \tag{11.14}$$

(see Exercise 11.6).

Sometimes we are interested in parameters $\theta = r(\beta_1, ..., \beta_m) = r(\beta)$ which are functions of the coefficients from multiple equations. In this case the least squares estimator of θ is $\widehat{\theta} = r(\widehat{\beta})$. The asymptotic distribution of $\widehat{\theta}$ can be obtained from Theorem 11.1 by the delta method.

Theorem 11.2 Under Assumptions 7.2 and 7.3, $\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\theta\right)$ where $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}$ and $\boldsymbol{R} = \frac{\partial}{\partial\beta} r\left(\beta\right)'$.

For a proof, see Exercise 11.7.

Theorem 11.2 is an example where multivariate regression is fundamentally distinct from univariate regression. Only by treating least squares as a joint estimator can we obtain a distributional theory for a function of multiple equations. We can thereby construct standard errors, confidence intervals, and hypothesis tests.

11.6 Covariance Matrix Estimation

From the finite sample and asymptotic theory we can construct appropriate estimators for the variance of $\widehat{\beta}$. In the general case we have

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'\widehat{e}_i\widehat{e}_i'\overline{X}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

Under conditional homoskedasticity (11.8) an appropriate estimator is

$$\widehat{\boldsymbol{V}}^0_{\widehat{\beta}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'\widehat{\Sigma}\overline{X}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

When the regressors are common then these estimators equal

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \left(\boldsymbol{I}_m \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(\widehat{e}_i\widehat{e}_i' \otimes X_i X_i'\right)\right)\left(\boldsymbol{I}_m \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

and $\widehat{\boldsymbol{V}}^0_{\widehat{\beta}} = \widehat{\Sigma} \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, respectively.

Covariance matrix estimators for $\widehat{\theta}$ are found as

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\theta}} &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\widehat{\boldsymbol{R}} \\ \widehat{\boldsymbol{V}}^0_{\widehat{\theta}} &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}^0_{\widehat{\beta}}\widehat{\boldsymbol{R}} \\ \widehat{\boldsymbol{R}} &= \frac{\partial}{\partial \beta} r\left(\widehat{\beta}\right)'.\end{aligned}$$

Theorem 11.3 Under Assumption 7.2, $n\widehat{\boldsymbol{V}}_{\widehat{\beta}} \xrightarrow[p]{} \boldsymbol{V}_\beta$ and $n\widehat{\boldsymbol{V}}^0_{\widehat{\beta}} \xrightarrow[p]{} \boldsymbol{V}^0_\beta$.

For a proof, see Exercise 11.8.

11.7 Seemingly Unrelated Regression

Consider the systems regression model under the conditional expectation and homoskedasticity assumptions

$$\begin{aligned}Y &= \overline{X}\beta + e \\ \mathbb{E}\left[e \mid X\right] &= 0 \\ \mathbb{E}\left[ee' \mid X\right] &= \Sigma.\end{aligned} \tag{11.15}$$

Since the errors are correlated across equations we consider estimation by Generalized Least Squares (GLS). To derive the estimator, premultiply (11.15) by $\Sigma^{-1/2}$ so that the transformed error vector is i.i.d. with covariance matrix $\boldsymbol{I}_m$. Then apply least squares and rearrange to find

$$\widehat{\beta}_{\text{gls}} = \left(\sum_{i=1}^{n}\overline{X}_i'\Sigma^{-1}\overline{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'\Sigma^{-1}Y_i\right). \tag{11.16}$$

(see Exercise 11.9). Another approach is to take the vector representation

$$\boldsymbol{Y} = \overline{\boldsymbol{X}}\beta + \boldsymbol{e}$$

and calculate that the equation error $\boldsymbol{e}$ has variance $\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}'\right] = \boldsymbol{I}_n \otimes \Sigma$. Premultiply the equation by $\boldsymbol{I}_n \otimes \Sigma^{-1/2}$ so that the transformed error has covariance matrix $\boldsymbol{I}_{nm}$ and then apply least squares to find

$$\widehat{\beta}_{\text{gls}} = \left(\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \Sigma^{-1}\right)\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \Sigma^{-1}\right)\boldsymbol{Y}\right) \tag{11.17}$$

(see Exercise 11.10).

Expressions (11.16) and (11.17) are algebraically equivalent. To see the equivalence, observe that

$$\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \Sigma^{-1}\right)\overline{\boldsymbol{X}} = \begin{pmatrix} \overline{X}_1' & \cdots & \overline{X}_n' \end{pmatrix}\begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ \vdots & \Sigma^{-1} & & \vdots \\ 0 & 0 & \cdots & \Sigma^{-1} \end{pmatrix}\begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_n \end{pmatrix}$$
$$= \sum_{i=1}^{n} \overline{X}_i' \Sigma^{-1} \overline{X}_i$$

and

$$\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \Sigma^{-1}\right)\boldsymbol{Y} = \begin{pmatrix} \overline{X}_1' & \cdots & \overline{X}_n' \end{pmatrix}\begin{pmatrix} \Sigma^{-1} & 0 & \cdots & 0 \\ \vdots & \Sigma^{-1} & & \vdots \\ 0 & 0 & \cdots & 0^{-1} \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$
$$= \sum_{i=1}^{n} \overline{X}_i' \Sigma^{-1} Y_i.$$

Since Σ is unknown it must be replaced by an estimator. Using $\widehat{\Sigma}$ from (11.5) we obtain a feasible GLS estimator.

$$\widehat{\beta}_{\text{sur}} = \left(\sum_{i=1}^{n} \overline{X}_i' \widehat{\Sigma}^{-1} \overline{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \overline{X}_i' \widehat{\Sigma}^{-1} Y_i\right)$$
$$= \left(\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \widehat{\Sigma}^{-1}\right)\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\left(\boldsymbol{I}_n \otimes \widehat{\Sigma}^{-1}\right)\boldsymbol{Y}\right). \tag{11.18}$$

This is the **Seemingly Unrelated Regression (SUR)** estimator as introduced by Zellner (1962).

The estimator $\widehat{\Sigma}$ can be updated by calculating the SUR residuals $\widehat{e}_i = Y_i - \overline{X}_i'\widehat{\beta}_{\text{sur}}$ and the covariance matrix estimator $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i \widehat{e}_i'$. Substituted into (11.18) we obtain an iterated SUR estimator. This can be iterated until convergence.

Under conditional homoskedasticity (11.8) we can derive its asymptotic distribution.

Theorem 11.4 Under Assumption 7.2 and (11.8)

$$\sqrt{n}\left(\widehat{\beta}_{\text{sur}} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta^*\right)$$

where $\boldsymbol{V}_\beta^* = \left(\mathbb{E}\left[\overline{X}'\Sigma^{-1}\overline{X}\right]\right)^{-1}$.

For a proof, see Exercise 11.11.

Under these assumptions, SUR is more efficient than least squares.

Theorem 11.5 Under Assumption 7.2 and (11.8)

$$\boldsymbol{V}_\beta^* = \left(\mathbb{E}\left[\overline{X}'\Sigma^{-1}\overline{X}\right]\right)^{-1} \leq \left(\mathbb{E}\left[\overline{X}'\overline{X}\right]\right)^{-1}\mathbb{E}\left[\overline{X}'\Sigma\overline{X}\right]\left(\mathbb{E}\left[\overline{X}'\overline{X}\right]\right)^{-1} = \boldsymbol{V}_\beta$$

and thus $\widehat{\beta}_{\text{sur}}$ is asymptotically more efficient than $\widehat{\beta}_{\text{ols}}$.

For a proof, see Exercise 11.12.

An appropriate estimator of the variance of $\widehat{\beta}_{\text{sur}}$ is

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \left(\sum_{i=1}^{n} \overline{X}_i' \widehat{\Sigma}^{-1} \overline{X}_i\right)^{-1}.$$

Theorem 11.6 Under Assumption 7.2 and (11.8) $n\widehat{\boldsymbol{V}}_{\widehat{\beta}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$.

For a proof, see Exercise 11.13.

In Stata, the seemingly unrelated regressions estimator is implemented using the `sureg` command.

Arnold Zellner

Arnold Zellner (1927-2010) of the United States was a founding father of the econometrics field. He was a pioneer in Bayesian econometrics. One of his core contributions was the method of Seemingly Unrelated Regressions.

11.8 Equivalence of SUR and Least Squares

When the regressors are common across equations $X_j = X$ it turns out that the SUR estimator simplifies to least squares.

To see this, recall that when regressors are common this implies that $\overline{X} = \boldsymbol{I}_m \otimes X'$. Then

$$\begin{aligned}
\overline{X}_i' \widehat{\Sigma}^{-1} &= (\boldsymbol{I}_m \otimes X_i) \widehat{\Sigma}^{-1} \\
&= \widehat{\Sigma}^{-1} \otimes X_i \\
&= \left(\widehat{\Sigma}^{-1} \otimes \boldsymbol{I}_k\right) (\boldsymbol{I}_m \otimes X_i) \\
&= \left(\widehat{\Sigma}^{-1} \otimes \boldsymbol{I}_k\right) \overline{X}_i'.
\end{aligned}$$

Thus

$$\begin{aligned}
\widehat{\beta}_{\text{sur}} &= \left(\sum_{i=1}^{n} \overline{X}_i' \widehat{\Sigma}^{-1} \overline{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \overline{X}_i' \widehat{\Sigma}^{-1} Y_i\right) \\
&= \left(\left(\widehat{\Sigma}^{-1} \otimes \boldsymbol{I}_k\right) \sum_{i=1}^{n} \overline{X}_i' \overline{X}_i\right)^{-1} \left(\left(\widehat{\Sigma}^{-1} \otimes \boldsymbol{I}_k\right) \sum_{i=1}^{n} \overline{X}_i' Y_i\right) \\
&= \left(\sum_{i=1}^{n} \overline{X}_i' \overline{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \overline{X}_i' Y_i\right) = \widehat{\beta}_{\text{ols}}.
\end{aligned}$$

A model where regressors are not common across equations is nested within a model with the union of all regressors included in all equations. Thus the model with regressors common across equations

is a fully unrestricted model, and a model where the regressors differ across equations is a restricted model. Thus the above result shows that the SUR estimator reduces to least squares in the absence of restrictions, but SUR can differ from least squares otherwise.

Another context where SUR=OLS is when the variance matrix is diagonal, $\Sigma = \operatorname{diag}\{\sigma_1^2, ..., \sigma_m^2\}$. In this case $\Sigma^{-1/2}\overline{X}_i = \overline{X}_i \operatorname{diag}\{\boldsymbol{I}_{k_1}\sigma_1^{-1/2}, ..., \boldsymbol{I}_{k_m}\sigma_m^{-1/2}\}$ from which you can calculate that $\widehat{\beta}_{\text{sur}} = \widehat{\beta}_{\text{ols}}$. The intuition is that there is no difference in systems estimation when the equations are uncorrelated, which occurs when Σ is diagonal.

11.9 Maximum Likelihood Estimator

Take the linear model under the assumption that the error is independent of the regressors and multivariate normally distributed. Thus $Y = \overline{X}\beta + e$ with $e \sim \mathrm{N}(0, \Sigma)$. In this case we can consider the maximum likelihood estimator (MLE) of the coefficients.

It is convenient to reparameterize the covariance matrix in terms of its inverse $\boldsymbol{S} = \Sigma^{-1}$. With this reparameterization the conditional density of Y given $X = x$ equals

$$f\left(y \mid x\right) = \frac{\det\left(\boldsymbol{S}\right)^{1/2}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\left(y - x\beta\right)' \boldsymbol{S}\left(y - x\beta\right)\right).$$

The log-likelihood function for the sample is

$$\ell_n(\beta, \boldsymbol{S}) = -\frac{nm}{2}\log(2\pi) + \frac{n}{2}\log\left(\det\left(\boldsymbol{S}\right)\right) - \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \overline{X}_i\beta\right)' \boldsymbol{S}\left(Y_i - \overline{X}_i\beta\right).$$

The maximum likelihood estimator $\left(\widehat{\beta}_{\text{mle}}, \widehat{\boldsymbol{S}}_{\text{mle}}\right)$ maximizes the log-likelihood function. The first order conditions are

$$0 = \frac{\partial}{\partial\beta}\ell_n(\beta, \boldsymbol{S})\Big|_{\beta=\widehat{\beta}, \boldsymbol{S}=\widehat{\boldsymbol{S}}} = \sum_{i=1}^{n}\overline{X}_i\widehat{\boldsymbol{S}}\left(Y_i - \overline{X}_i\widehat{\beta}\right)$$

and

$$0 = \frac{\partial}{\partial\boldsymbol{S}}\ell_n(\beta, \Sigma)\Big|_{\beta=\widehat{\beta}, \boldsymbol{S}=\widehat{\boldsymbol{S}}} = \frac{n}{2}\widehat{\boldsymbol{S}}^{-1} - \frac{1}{2}\operatorname{tr}\left(\sum_{i=1}^{n}\left(Y_i - \overline{X}_i\widehat{\beta}\right)\left(Y_i - \overline{X}_i\widehat{\beta}\right)'\right).$$

The second equation uses the matrix results $\frac{\partial}{\partial\boldsymbol{S}}\log\left(\det\left(\boldsymbol{S}\right)\right) = \boldsymbol{S}^{-1}$ and $\frac{\partial}{\partial\boldsymbol{B}}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{B}\right) = \boldsymbol{A}'$ from Appendix A.20.

Solving and making the substitution $\widehat{\Sigma} = \widehat{\boldsymbol{S}}^{-1}$ we obtain

$$\widehat{\beta}_{\text{mle}} = \left(\sum_{i=1}^{n}\overline{X}_i'\widehat{\Sigma}^{-1}\overline{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\overline{X}_i'\widehat{\Sigma}^{-1}Y_i\right)$$

$$\widehat{\Sigma}_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{X}_i\widehat{\beta}\right)\left(Y_i - \overline{X}_i\widehat{\beta}\right)'.$$

Notice that each equation refers to the other. Hence these are not closed-form expressions but can be solved via iteration. The solution is identical to the iterated SUR estimator. Thus the iterated SUR estimator is identical to MLE under normality.

Recall that the SUR estimator simplifies to OLS when the regressors are common across equations. The same occurs for the MLE. Thus when $\overline{X}_i = \boldsymbol{I}_m \otimes X_i'$ we find that $\widehat{\beta}_{\text{mle}} = \widehat{\beta}_{\text{ols}}$ and $\widehat{\Sigma}_{\text{mle}} = \widehat{\Sigma}_{\text{ols}}$.

11.10 Restricted Estimation

In many multivariate regression applications it is desired to impose restrictions on the coefficients. In particular, cross-equation restrictions (for example, imposing Slutsky symmetry on a demand system) can be quite important and can only be imposed by a multivariate estimation method. Estimation subject to restrictions can be done by minimum distance, maximum likelihood, or the generalized method of moments.

Minimum distance is a straightforward application of the methods of Chapter 8 to the estimators presented in this chapter, as such methods apply to any asymptotically normal estimator.

Imposing restrictions on maximum likelihood is also straightforward. The likelihood is maximized subject to the imposed restrictions. One important example is explored in detail in the following section.

Generalized method of moments estimation of multivariate regression subject to restrictions will be explored in Section 13.18. This is a particularly simple and straightforward way to estimate restricted multivariate regression models and is our generally preferred approach.

11.11 Reduced Rank Regression

One context where systems estimation is important is when it is desired to impose or test restrictions across equations. Restricted systems are commonly estimated by maximum likelihood under normality. In this section we explore one important special case of restricted multivariate regression known as reduced rank regression. The model was originally proposed by Anderson (1951) and extended by Johansen (1995).

The unrestricted model is

$$\begin{aligned} Y &= \boldsymbol{B}'X + \boldsymbol{C}'Z + e \\ \mathbb{E}\left[ee' \mid X, Z\right] &= \Sigma \end{aligned} \tag{11.19}$$

where $\boldsymbol{B}$ is $k \times m$, $\boldsymbol{C}$ is $\ell \times m$, $Y \in \mathbb{R}^m$, $X \in \mathbb{R}^k$, and $Z \in \mathbb{R}^\ell$. We separate the regressors as X and Z because the coefficient matrix $\boldsymbol{B}$ will be restricted while $\boldsymbol{C}$ will be unrestricted.

The matrix $\boldsymbol{B}$ is full rank if

$$\operatorname{rank}(\boldsymbol{B}) = \min(k, m).$$

The reduced rank restriction is $\operatorname{rank}(\boldsymbol{B}) = r < \min(k, m)$for some known r.

The reduced rank restriction implies that we can write the coefficient matrix $\boldsymbol{B}$ in the factored form $\boldsymbol{B} = \boldsymbol{G}\boldsymbol{A}'$ where $\boldsymbol{A}$ is $m \times r$ and $\boldsymbol{G}$ is $k \times r$. This representation is not unique as we can replace $\boldsymbol{G}$ with $\boldsymbol{G}\boldsymbol{Q}$ and $\boldsymbol{A}$ with $\boldsymbol{A}\boldsymbol{Q}^{-1\prime}$ for any invertible $\boldsymbol{Q}$ and the same relation holds. Identification therefore requires a normalization of the coefficients. A conventional normalization is $\boldsymbol{G}'\boldsymbol{D}\boldsymbol{G} = \boldsymbol{I}_r$ for given $\boldsymbol{D}$.

Equivalently, the reduced rank restriction can be imposed by requiring that $\boldsymbol{B}$ satisfy the restriction $\boldsymbol{B}\boldsymbol{A}_\perp = \boldsymbol{G}\boldsymbol{A}'\boldsymbol{A}_\perp = 0$ for some $m \times (m - r)$ coefficient matrix $\boldsymbol{A}_\perp$. Since $\boldsymbol{G}$ is full rank this requires that $\boldsymbol{A}'\boldsymbol{A}_\perp = 0$, hence $\boldsymbol{A}_\perp$ is the orthogonal complement of $\boldsymbol{A}$. Note that $\boldsymbol{A}_\perp$ is not unique as it can be replaced by $\boldsymbol{A}_\perp\boldsymbol{Q}$ for any $(m - r) \times (m - r)$ invertible $\boldsymbol{Q}$. Thus if $\boldsymbol{A}_\perp$ is to be estimated it requires a normalization.

We discuss methods for estimation of $\boldsymbol{G}$, $\boldsymbol{A}$, Σ, $\boldsymbol{C}$, and $\boldsymbol{A}_\perp$. The standard approach is maximum likelihood under the assumption that $e \sim \mathrm{N}(0, \Sigma)$. The log-likelihood function for the sample is

$$\begin{aligned} \ell_n(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{C}, \Sigma) = &-\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log(\det(\Sigma)) \\ &- \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{A}\boldsymbol{G}'X_i - \boldsymbol{C}'Z_i\right)'\Sigma^{-1}\left(Y_i - \boldsymbol{A}\boldsymbol{G}'X_i - \boldsymbol{C}'Z_i\right). \end{aligned}$$

Anderson (1951) derived the MLE by imposing the constraint $\boldsymbol{B}\boldsymbol{A}_\perp = 0$ via the method of Lagrange multipliers. This turns out to be algebraically cumbersome.

Johansen (1995) instead proposed the following straightforward concentration method. Treating $\boldsymbol{G}$ as if it is known, maximize the log-likelihood with respect to the other parameters. Resubstituting these estimators we obtain the concentrated log-likelihood function with respect to $\boldsymbol{G}$. This can be maximized to find the MLE for $\boldsymbol{G}$. The other parameter estimators are then obtain by substitution. We now describe these steps in detail.

Given $\boldsymbol{G}$ the likelihood is a normal multivariate regression in the variables $\boldsymbol{G}'X$ and Z, so the MLE for $\boldsymbol{A}$, $\boldsymbol{C}$ and Σ are least squares. In particular, using the Frisch-Waugh-Lovell residual regression formula we can write the estimators for $\boldsymbol{A}$ and Σ as

$$\widehat{\boldsymbol{A}}(\boldsymbol{G}) = \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}$$

and

$$\widehat{\Sigma}(\boldsymbol{G}) = \frac{1}{n}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right)$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}$.

Substituting these estimators into the log-likelihood function we obtain the concentrated likelihood function, which is a function of $\boldsymbol{G}$ only.

$$\begin{aligned}
\widetilde{\ell}_n(\boldsymbol{G}) &= \ell_n\left(\boldsymbol{G}, \widehat{\boldsymbol{A}}(\boldsymbol{G}), \widehat{\boldsymbol{C}}(\boldsymbol{G}), \widehat{\Sigma}(\boldsymbol{G})\right) \\
&= \frac{m}{2}\left(n\log(2\pi) - 1\right) - \frac{n}{2}\log\left[\det\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right)\right] \\
&= \frac{m}{2}\left(n\log(2\pi) - 1\right) - \frac{n}{2}\log\left(\det\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)\right) - \frac{n}{2}\log\left[\frac{\det\left(\boldsymbol{G}'\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\right)\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)}\right].
\end{aligned}$$

The third equality uses Theorem A.1.8. The MLE $\widehat{\boldsymbol{G}}$ for $\boldsymbol{G}$ is the maximizer of $\widetilde{\ell}_n(\boldsymbol{G})$, or equivalently equals

$$\begin{aligned}
\widehat{\boldsymbol{G}} &= \underset{\boldsymbol{G}}{\operatorname{argmin}} \frac{\det\left(\boldsymbol{G}'\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\right)\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)} \\
&= \underset{\boldsymbol{G}}{\operatorname{argmax}} \frac{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)} \\
&= \{v_1, ..., v_r\}
\end{aligned} \tag{11.20}$$

which are the generalized eigenvectors of $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}$ with respect to $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ corresponding to the r largest generalized eigenvalues. (Generalized eigenvalues and eigenvectors are discussed in Section A.14.) The estimator satisfies the normalization $\widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}} = \boldsymbol{I}_r$. Letting v_j^* denote the eigenvectors of (11.20) we can also express $\widehat{\boldsymbol{G}} = \left\{v_m^*, ..., v_{m-r+1}^*\right\}$.

This is computationally straightforward. In MATLAB, for example, the generalized eigenvalues and eigenvectors of a matrix $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ are found using the command `eig(A,B)`.

Given $\widehat{\boldsymbol{G}}$, the MLE $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{C}}$, $\widehat{\Sigma}$ are found by least squares regression of Y on $\widehat{\boldsymbol{G}}'X$ and Z. In particular, $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}$ because $\widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}} = \boldsymbol{I}_r$.

We now discuss the estimator $\widehat{\boldsymbol{A}}_\perp$ of $\boldsymbol{A}_\perp$. It turns out that

$$\begin{aligned}\widehat{\boldsymbol{A}}_\perp &= \underset{\boldsymbol{A}}{\operatorname{argmax}} \frac{\det\left(\boldsymbol{A}'\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right)\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\boldsymbol{A}\right)} \\ &= \{w_1, ..., w_{m-r}\}\end{aligned} \tag{11.21}$$

the eigenvectors of $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ associated with the largest $m-r$ eigenvalues.

By the dual eigenvalue relation (Theorem A.5), equations (11.20) and (11.21) have the same non-zero eigenvalues λ_j and the associated eigenvectors v_j^* and w_j satisfy the relationship

$$w_j = \lambda_j^{-1/2}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}v_j^*.$$

Letting $\Lambda = \operatorname{diag}\{\lambda_m, ..., \lambda_{m-r+1}\}$ this implies

$$\{w_m, ..., w_{m-r+1}\} = \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left\{v_m^*, ..., v_{m-r+1}^*\right\}\Lambda = \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widehat{\boldsymbol{A}}\Lambda.$$

The second equality holds because $\widehat{\boldsymbol{G}} = \left\{v_m^*, ..., v_{m-r+1}^*\right\}$ and $\widehat{\boldsymbol{A}} = \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}}$. Since the eigenvectors w_j satisfy the orthogonality property $w_j'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}w_\ell = 0$ for $j \neq \ell$, it follows that

$$0 = \widehat{\boldsymbol{A}}_\perp'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\{w_m, ..., w_{m-r+1}\} = \widehat{\boldsymbol{A}}_\perp'\widehat{\boldsymbol{A}}\Lambda.$$

Since $\Lambda > 0$ we conclude that $\widehat{\boldsymbol{A}}_\perp'\widehat{\boldsymbol{A}} = 0$ as desired.

The solution $\widehat{\boldsymbol{A}}_\perp$ in (11.21) can be represented several ways. One which is computationally convenient is to observe that

$$\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}} = \boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y} = \widetilde{\boldsymbol{E}}'\widetilde{\boldsymbol{E}}$$

where $\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}} = \boldsymbol{I}_n - (\boldsymbol{X},\boldsymbol{Z})\left((\boldsymbol{X},\boldsymbol{Z})'(\boldsymbol{X},\boldsymbol{Z})\right)^{-1}(\boldsymbol{X},\boldsymbol{Z})'$ and $\widetilde{\boldsymbol{E}} = \boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}$ is the residual matrix from the unrestricted multivariate least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The first equality follows by the Frisch-Waugh-Lovell theorem. This shows that $\widehat{\boldsymbol{A}}_\perp$ are the generalized eigenvectors of $\widetilde{\boldsymbol{E}}'\widetilde{\boldsymbol{E}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ corresponding to the $m-r$ largest eigenvalues. In MATLAB, for example, these can be computed using the `eig(A,B)` command.

Another representation is to write $\boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ so that

$$\widehat{\boldsymbol{A}}_\perp = \underset{\boldsymbol{A}}{\operatorname{argmax}} \frac{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \frac{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}.$$

We summarize our findings.

Theorem 11.7 The MLE for the reduced rank model (11.19) under $e \sim \mathrm{N}(0, \Sigma)$ is given as follows. Let $\widetilde{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{X}}$ be the residual matrices from multivariate regression of $\boldsymbol{Y}$ and $\boldsymbol{X}$ on $\boldsymbol{Z}$, respectively. Then $\widehat{\boldsymbol{G}}_{\text{mle}} = \{v_1, ..., v_r\}$, the generalized eigenvectors of $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}$ with respect to $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ corresponding to the r largest eigenvalues $\widehat{\lambda}_j$. $\widehat{\boldsymbol{A}}_{\text{mle}}$, $\widehat{\boldsymbol{C}}_{\text{mle}}$ and $\widehat{\Sigma}_{\text{mle}}$ are obtained by the least squares regression

$$Y_i = \widehat{\boldsymbol{A}}_{\text{mle}}\widehat{\boldsymbol{G}}'_{\text{mle}}X_i + \widehat{\boldsymbol{C}}'_{\text{mle}}Z_i + \widehat{e}_i$$
$$\widehat{\Sigma}_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i\widehat{e}_i'.$$

Let $\widetilde{\boldsymbol{E}}$ be the residual matrix from a multivariate regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. Then $\widehat{\boldsymbol{A}}_{\perp}$ equals the generalized eigenvectors of $\widetilde{\boldsymbol{E}}'\widetilde{\boldsymbol{E}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ corresponding to the $m-r$ smallest eigenvalues. The maximized likelihood equals

$$\ell_n = \frac{m}{2}\left(n\log(2\pi) - 1\right) - \frac{n}{2}\log\left(\det\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)\right) - \frac{n}{2}\sum_{j=1}^{r}\log\left(1-\widehat{\lambda}_j\right).$$

An R package for reduced rank regression is "RRR". I am unaware of a Stata command.

11.12 Principal Component Analysis

In Section 4.21 we described the Duflo, Dupas, and Kremer (2011) dataset which is a sample of Kenyan first grade test scores. Following the authors we focused on the variable *totalscore* which is each student's composite test score. If you examine the data file you will find other pieces of information about the students' performance, including each student's score on separate sections of the test, with the labels *wordscore* (word recognition), *sentscore* (sentence recognition), *letterscore* (letter recognition), *spellscore* (spelling), *additions_score* (addition), *substractions_score* (subtraction), *multiplications_score* (multiplication). The "total" score sums the scores from the individual sections. Perhaps there is more information in the section scores. How can we learn about this from the data?

Principal component analysis (PCA) addresses this issue by ordering linear combinations by their contribution to variance.

Definition 11.1 Let X be a $k \times 1$ random vector.
The **first principal component** is $U_1 = h_1' X$ where h_1 satisfies

$$h_1 = \underset{h'h=1}{\operatorname{argmax}} \operatorname{var}\left[h'X\right].$$

The **second principal component** is $U_2 = h_2' X$ where

$$h_2 = \underset{h'h=1, h'h_1=0}{\operatorname{argmax}} \operatorname{var}\left[h'X\right].$$

In general, the j^{th} **principal component** is $U_j = h_j' X$ where

$$h_j = \underset{h'h=1, h'h_1=0,\ldots,h'h_{j-1}=0}{\operatorname{argmax}} \operatorname{var}\left[h'X\right].$$

The principal components of X are linear combinations $h'X$ ranked by contribution to variance. By the properties of quadratic forms (Section A.15) the weight vectors h_j are the eigenvectors of $\Sigma = \operatorname{var}[X]$.

Theorem 11.8 The principal components of X are $U_j = h_j' X$, where h_j is the eigenvector of Σ associated with the j^{th} ordered eigenvalue λ_j of Σ.

Another way to see the PCA construction is as follows. Since Σ is symmetric the spectral decomposition (Theorem A.3) states that $\Sigma = \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}'$ where $\boldsymbol{H} = [h_1, ..., h_k]$ and $\boldsymbol{D} = \operatorname{diag}(d_1, ..., d_k)$ are the eigenvectors and eigenvalues of Σ. Since Σ is positive semi-definite the eigenvalues are real, non-negative, and ordered $d_1 \geq d_2 \geq \cdots \geq d_k$. Let $U = (U_1, ..., U_k)$ be the principal components of X. By Theorem 11.8, $U = \boldsymbol{H}'X$. The covariance matrix of U is

$$\operatorname{var}[U] = \operatorname{var}\left[\boldsymbol{H}'X\right] = \boldsymbol{H}'\Sigma\boldsymbol{H} = \boldsymbol{D}$$

which is diagonal. This shows that $\operatorname{var}\left[U_j\right] = d_j$ and the principal components are mutually uncorrelated. The relative variance contribution of the j^{th} principal component is $d_j / \operatorname{tr}(\Sigma)$.

Principal components are sensitive to the scaling of X. Consequently, it is recommended to first scale each element of X to have mean zero and unit variance. In this case Σ is a correlation matrix.

The sample principal components are obtained by replacing the unknowns by sample estimators. Let $\widehat{\Sigma}$ be the sample covariance or correlation matrix and $\widehat{h}_1, \widehat{h}_2, ..., \widehat{h}_k$ its ordered eigenvectors. The sample principal components are $\widehat{h}_j' X_i$.

To illustrate we use the Duflo, Dupas, and Kremer (2011) dataset. In Table 11.1 we display the seven eigenvalues of the sample correlation matrix for the seven test scores described above. The seven eigenvalues sum to seven because we have applied PCA to the correlation matrix. The first eigenvalue is 4.0, implying that the first principal component explains 57% of the variance of the seven test scores. The second eigenvalue is 1.0, implying that the second principal component explains 15% of the variance. Together the first two components explain 72% of the variance of the seven test scores.

In Table 11.2 we display the weight vectors (eigenvectors) for the first two principal components. The weights for the first component are all positive and similar in magnitude. This means that the first

Table 11.1: Eigenvalue Decomposition of Sample Correlation Matrix

	Eigenvalue	Proportion
1	4.02	0.57
2	1.04	0.15
3	0.57	0.08
4	0.52	0.08
5	0.37	0.05
6	0.29	0.04
7	0.19	0.03

Table 11.2: Principal Component Weight Vectors

	First	Second
words	0.41	−0.32
sentences	0.32	−0.49
letters	0.40	−0.13
spelling	0.43	−0.28
addition	0.38	0.41
subtraction	0.35	0.52
multiplication	0.33	0.36

principal component is similar to a simple average of the seven test scores. This is quite fascinating. This is consistent with our intuition that a simple average (e.g. the variable *totalscore*) captures most of the information contained in the seven test scores. The weights for the second component have a different pattern. The four literacy scores receive negative weight and the three math scores receive positive weight with similar magnitudes. This means that the second principal component is similar to the difference between a student's math and verbal test scores. Taken together, the information in the first two principal components is equivalent to "average verbal" and "average math" test scores. What this shows is that 57% of the variation in the seven section test scores can be explained by a simple average (e.g. *totalscore*), and 72% can be explained by averages for the verbal and math halves of the test.

In Stata, principal components analysis can be implemented with the `pca` command. In R use `prcomp` or `princomp`. All three can be applied to either covariance matrices (unscaled data) or correlation matrices (normalized data) but they have different default settings. The Stata `pca` command by default normalizes the observations. The R commands by default do not normalize the observations.

11.13 Factor Models

Closely related to principal components are factor models. These are statistical models which decompose random vectors into common factors and idiosyncratic errors. Factor models are popular throughout the social sciences. Consequently a variety of estimation methods have been developed. In the next few sections we focus on methods which are popular among economists.

Let $X = (X_1, ..., X_k)'$ be a $k \times 1$ random vector (for example the seven test scores described in the previous section). Assume that the elements of X are scaled to have mean zero and unit variance.

A **single factor model** for X is

$$X = \lambda F + u \tag{11.22}$$

where $\lambda \in \mathbb{R}^k$ are **factor loadings**, $F \in \mathbb{R}$ is a **common factor**, and $u \in \mathbb{R}^k$ is a random error. The factor F is individual-specific while the coefficient λ is common across individuals. The model (11.22) specifies that correlation between the elements of X is due to the common factor F. In the student test score example it is intuitive to think of F as a student's scholastic "aptitude"; in this case the vector λ describes how scholastic aptitude affects the seven subject scores.

A multiple factor model has $r < k$ factors. We write the model as

$$X = \Lambda F + u \tag{11.23}$$

where Λ is a $k \times r$ matrix of factor loadings and $F = (F_1, ..., F_r)'$ is an $r \times 1$ vector of factors. In the student test score example possible factors could be "math aptitude", "language skills", "social skills", "artistic ability", "creativity", etc. The factor loading matrix Λ indicates the effect of each factor on each test score. The number of factors r is taken as known. We discuss selection of r later.

The error vector u is assumed to be mean zero, uncorrelated with F, and (under correct specification) to have mutually uncorrelated elements. We write its covariance matrix as $\Psi = \mathbb{E}\left[uu'\right]$. The factor vector F can either be treated as random or as a regressor. In this section we treat F as random; in the next we treat F as regressors. The random factors F are assumed mean zero and are normalized so that $\mathbb{E}\left[FF'\right] = \boldsymbol{I}_r$.

The assumptions imply that the correlation matrix $\Sigma = \mathbb{E}\left[XX'\right]$ equals

$$\Sigma = \Lambda\Lambda' + \Psi. \tag{11.24}$$

The factor analysis literature often describes $\Lambda\Lambda'$ as the **communality** and the idiosyncratic error matrix Ψ as the **uniqueness**. The former is the portion of the variance which is explained by the factor model and the latter is the unexplained portion of the variance.

The model is often[1] estimated by **maximum likelihood**. Under joint normality of (F, u) the distribution of X is $\mathrm{N}\left(0, \Lambda\Lambda' + \Psi\right)$. The parameters are Λ and $\Psi = \operatorname{diag}\left(\psi_1, ..., \psi_k\right)$. The log-likelihood function of a random sample $(X_1, ..., X_n)$ is

$$\ell_n(\Lambda, \Psi) = -\frac{nk}{2}\log(2\pi) - \frac{n}{2}\log\det\left(\Lambda\Lambda' + \Psi\right) - \frac{n}{2}\operatorname{tr}\left(\left(\Lambda\Lambda' + \Psi\right)^{-1}\widehat{\Sigma}\right). \tag{11.25}$$

The MLE $\left(\widehat{\Lambda}, \widehat{\Psi}\right)$ maximizes $\ell_n(\Lambda, \Psi)$. There is not an algebraic solution so the estimator is found using numerical methods. Fortunately, computational algorithms are available in standard packages. A detailed description and analysis can be found in Anderson (2003, Chapter 14).

The form of the log-likelihood is intriguing. Notice that the log-likelihood is only a function of the observations through its correlation matrix $\widehat{\Sigma}$, and only a function of the parameters through the population correlation matrix $\Lambda\Lambda' + \Psi$. The final term in (11.25) is a measure of the match between $\widehat{\Sigma}$ and $\Lambda\Lambda' + \Psi$. Together, we see that the Gaussian log-likelihood is essentially a measure of the fit of the model and sample correlation matrices. It is therefore not reliant on the normality assumption.

It is often of interest to estimate the factors F_i. Given Λ the equation $X_i = \Lambda F_i + u_i$ can be viewed as a regression with coefficient F_i. Its least squares estimator is $\widehat{F}_i = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i$. The GLS estimator (taking into account the covariance matrix of u_i) is $\widehat{F}_i = \left(\Lambda'\Psi^{-1}\Lambda\right)^{-1}\Lambda'\Psi^{-1}X_i$. This motivates the **Bartlett scoring** estimator

$$\widetilde{F}_i = \left(\widehat{\Lambda}'\widehat{\Psi}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\Psi}^{-1}X_i.$$

The idealized version satisfies

$$\widehat{F}_i = \left(\Lambda'\Psi^{-1}\Lambda\right)^{-1}\Lambda'\Psi^{-1}\left(\Lambda F_i + u_i\right) = F_i + \left(\Lambda'\Psi^{-1}\Lambda\right)^{-1}\Lambda'\Psi^{-1}u_i$$

[1] There are other estimators used in applied factor analysis. However there is little reason to consider estimators beyond the MLE of this section and the principal components estimator of the next section.

which is unbiased for F_i and has variance $\left(\Lambda'\Psi^{-1}\Lambda\right)^{-1}$. Thus the Barlett scoring estimator is typically described as "unbiased" though this is actually a property of its idealized version $\widehat{F}_i$.

A second estimator for the factors can be constructed from the multivariate linear projection of F on X. This is $F = \boldsymbol{A}X + \xi$ where the coefficient matrix $\boldsymbol{A}$ is $r \times k$. The coefficient matrix equals

$$\boldsymbol{A} = \mathbb{E}\left[FX'\right]\mathbb{E}\left[XX'\right]^{-1} = \Lambda'\Sigma^{-1},$$

the second equation using $\mathbb{E}\left[FX'\right] = \mathbb{E}\left[F(\Lambda F + u)'\right] = \mathbb{E}\left[FF'\right]\Lambda' + \mathbb{E}\left[Fu'\right] = \Lambda'$. The predicted value of F_i is $F_i^* = \boldsymbol{A}X_i = \Lambda'\Sigma^{-1}X_i$. This motivates the **regression scoring** estimator

$$\overline{F}_i = \widehat{\Lambda}'\widehat{\Sigma}^{-1}X_i.$$

The idealized version F_i^* has conditional expectation $\Lambda'\Sigma^{-1}\Lambda F_i$ and is thus biased for F_i. Hence the regression scoring estimator $\overline{F}_i$ is often described as "biased". Some algebraic manipulations reveal that F_i^* has MSE $\boldsymbol{I}_r - \Lambda'\left(\Lambda'\Lambda + \Psi\right)^{-1}\Lambda$ which is smaller (in a positive definite sense) than the MSE of the idealized Bartlett estimator $\widehat{F}_i$.

Which estimator is preferred, Bartlett or regression scoring? The differences diminish when k is large so the choice is most relevant for small to moderate k. The regression scoring estimator has lower approximate MSE, meaning that it is a more precise estimator. Thus based on estimation precision this is our recommended choice.

The factor loadings Λ and factors F are not separately identified. To see this, notice that if you replace (Λ, F) with $\Lambda^* = \Lambda\boldsymbol{G}$ and $F^* = \boldsymbol{G}'F$ where $\boldsymbol{G}$ is $r \times r$ and orthonormal then the regression model is identical. Such replacements are called "rotations" in the factor analysis literature. Any orthogonal rotation of the factor loadings is an equally valid representation. The default MLE outputs are one specific rotation; others can be obtained by a variety of algorithms (which we do not review here). Consequently it is unwise to attribute meaning to the individual factor loading estimates.

Another important and tricky issue is selection of the number of factors r. There is no clear guideline. One approach is to examine the principal component decomposition, look for a division between the "large" and "small eigenvalues, and set r to equal to the number of "large" eigenvalues. Another approach is based on testing. As a by-product of the MLE (and standard package implementations) we obtain the LR test for the null hypothesis of r factors against the alternative hypothesis of k factors. If the LR test rejects (has a small p-value) this is evidence that the given r may be too small.

In Stata, the MLE $\left(\widehat{\Lambda}, \widehat{\Psi}\right)$ can be calculated with the `factor, ml factors(r)` command. The factor estimates $\widetilde{F}_i$ and $\overline{F}_i$ can be calculated by the `predict` command with either the `barlett` or `regression` option, respectively. In R, the command `factanal(X,factors=r,rotation="none")` calculates the MLE $\left(\widehat{\Lambda}, \widehat{\Psi}\right)$ and also calculates the factor estimates $\widetilde{F}_i$ and/or $\overline{F}_i$ using the `scores` option.

11.14 Approximate Factor Models

The MLE of the previous section is a good choice for factor estimation when the number of variables k is small and the factor model is believed to be correctly specified. In many economic applications of factor analysis, however, the number of variables is k is large. In such contexts the MLE can be computationally costly and/or unstable. Furthermore it is typically not credible to believe that the model is correctly specified; rather it is more reasonable to view the factor model as a useful approximation. In this section we explore an approach known as the approximate factor model with estimation by principal components. The estimation method is justified by an asymptotic framework where the number of variables $k \to \infty$.

The **approximate factor model** was introduced by Chamberlain and Rothschild (1983). It is the same as (11.23) but relaxes the assumption on the idiosyncratic error u so that the covariance matrix $\Psi = \mathbb{E}\left[uu'\right]$ is left unrestricted. In this context the Gaussian MLE of the previous section is misspecified.

Chamberlain and Rothschild (and the literature which followed) proposed estimation by least squares. The idea is to treat the factors as unknown regressors and simultaneously estimate the factors F_i and factor loadings Λ. We first describe the estimation method.

Let $(X_1, ..., X_n)$ be a sample centered at sample means. The least squares criterion is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \Lambda F_i)'(X_i - \Lambda F_i).$$

Let $\left(\widehat{\Lambda}, \widehat{F}_1, ..., \widehat{F}_n\right)$ be the joint minimizers. As Λ and F_i are not separately identified a normalization is needed. For compatibility with the notation of the previous section we use $n^{-1}\sum_{i=1}^{n}\widehat{F}_i\widehat{F}_i' = \boldsymbol{I}_r$.

We use a concentration argument to find the solution. As described in the previous section, each observation satisfies the multivariate equation $X_i = \Lambda F_i + u_i$. For fixed Λ this is a set of k equations with r unknowns F_i. The least squares solution is $\widehat{F}_i(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i$. Substituting this expression into the least squares criterion the concentrated least squares criterion for Λ is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \Lambda\widehat{F}_i(\Lambda)\right)'\left(X_i - \Lambda\widehat{F}_i(\Lambda)\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \Lambda\left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i\right)'\left(X_i - \Lambda\left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i'X_i - X_i'\Lambda\left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i\right)\\ &= \operatorname{tr}\left[\widehat{\Sigma}\right] - \operatorname{tr}\left[\left(\Lambda'\Lambda\right)^{-1}\Lambda'\widehat{\Sigma}\Lambda\right]\end{aligned}$$

where $\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}X_i X_i'$ is the sample covariance matrix. The least squares estimator $\widehat{\Lambda}$ minimizes this criterion. Let $\widehat{\boldsymbol{D}}$ and $\widehat{\boldsymbol{H}}$ be first r eigenvalues and eigenvectors of $\widehat{\Sigma}$. Using the normalization $\Lambda'\Lambda = \boldsymbol{I}_r$, from the extrema results of Section A.15 the minimizer of the least squares criterion is $\widehat{\Lambda} = \widehat{\boldsymbol{H}}$. More broadly any rotation of $\widehat{\boldsymbol{H}}$ is valid. Consider $\widehat{\Lambda} = \widehat{\boldsymbol{H}}\widehat{\boldsymbol{D}}^{1/2}$. Recall the expression for the factors $\widehat{F}_i(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_i$. We find that the estimated factors are

$$\widehat{F}_i = \left(\widehat{\boldsymbol{D}}^{1/2}\widehat{\boldsymbol{H}}'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{D}}^{1/2}\right)^{-1}\widehat{\boldsymbol{D}}^{1/2}\widehat{\boldsymbol{H}}' X_i = \widehat{\boldsymbol{D}}^{-1/2}\widehat{\boldsymbol{H}}' X_i.$$

We calculate that

$$n^{-1}\sum_{i=1}^{n}\widehat{F}_i\widehat{F}_i' = \widehat{\boldsymbol{D}}^{-1/2}\widehat{\boldsymbol{H}}'\widehat{\Sigma}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{D}}^{-1/2\prime} = \widehat{\boldsymbol{D}}^{-1/2}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{D}}^{-1/2\prime} = \boldsymbol{I}_r$$

which is the desired normalization. This shows that the rotation $\widehat{\Lambda} = \widehat{\boldsymbol{H}}\widehat{\boldsymbol{D}}^{1/2}$ produces factor estimates satisfying this normalization.

We have proven the following result.

Theorem 11.9 The least squares estimator of the factor model (11.23) under the normalization $n^{-1}\sum_{i=1}^{n}\widehat{F}_i\widehat{F}_i' = \boldsymbol{I}_r$ has the following solution:

1. Let $\widehat{\boldsymbol{D}} = \operatorname{diag}\left[\widehat{d}_1, ..., \widehat{d}_r\right]$ and $\widehat{\boldsymbol{H}} = \left[\widehat{h}_1, ..., \widehat{h}_r\right]$ be the first r eigenvalues and eigenvectors of the sample covariance matrix $\widehat{\Sigma}$.
2. $\widehat{\Lambda} = \widehat{\boldsymbol{H}}\widehat{\boldsymbol{D}}^{1/2}$.
3. $\widehat{F}_i = \widehat{\boldsymbol{D}}^{-1/2}\widehat{\boldsymbol{H}}' X_i$.

Theorem 11.9 shows that the least squares estimator is based on an eigenvalue decomposition of the covariance matrix. This is computationally stable even in high dimensions.

The factor estimates are the principal components scaled by the eigenvalues of $\widehat{\Sigma}$. Specifically, the j^{th} factor estimate is $\widehat{F}_{ji} = \widehat{d}_j^{-1/2}\widehat{h}_j' X$. Consequently many authors call this estimator the "principal-component method".

Unfortunately, $\widehat{\Lambda}$ is inconsistent for Λ if k is fixed, as we now show. By the WLLN and CMT, $\widehat{\Sigma} \underset{p}{\longrightarrow} \Sigma$ and $\widehat{\boldsymbol{H}} \underset{p}{\longrightarrow} \boldsymbol{H}$, the first r eigenvectors of Σ. When Ψ is diagonal, the eigenvectors of $\Sigma = \Lambda\Lambda' + \Psi$ do not lie in the range space of Λ except in the special case $\Psi = \sigma^2 \boldsymbol{I}_k$. Consequently the estimator $\widehat{\Lambda}$ is inconsistent.

This inconsistency should not be viewed as surprising. The sample has a total of nk observations and the model has a total of $nr + kr - r(r+1)/2$ parameters. Since the number of estimated pararameters is proportional to sample size we should not expect estimator consistency.

As first recognized by Chamberlain and Rothschild, this deficiency diminishes as k increases. Specifically, assume that $k \to \infty$ as $n \to \infty$. One implication is that the number of observations nk increase at a rate faster than n, while the number of parameters increase at a rate proportional to n. Another implication is that as k increases there is increasing information about the factors.

To make this precise we add the following assumption. Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues of a positive semi-definite matrix $\boldsymbol{A}$.

Assumption 11.1 As $k \to \infty$

1. $\lambda_{\max}(\Psi) \leq B < \infty$.
2. $\lambda_{\min}\left(\Lambda'\Lambda\right) \to \infty$ as $k \to \infty$.

Assumption 11.1.1 bounds the covariance matrix of the idiosyncratic errors. When $\Psi = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_k^2\right)$ this is the same as bounding the individual variances. Effectively Assumption 11.1.1 means that while the elements of u can be correlated they cannot have a correlation structure similar to that of a factor model. Assumption 11.1.2 requires the factor loading matrix to increase in magnitude as the number of variables increases. This is a fairly mild requirement. When the factor loadings are of similar magnitude across variables, $\lambda_{\min}\left(\Lambda'\Lambda\right) \sim k \to \infty$. Conceptually, Assumption 11.1.2 requires additional variables to add information about the unobserved factors.

Assumption 11.1 implies that in the covariance matrix factorization $\Sigma = \Lambda\Lambda' + \Psi$ the component $\Lambda\Lambda'$ dominates as k increases. This means that for large k the first r eigenvectors of Σ are equivalent to those of $\Lambda\Lambda'$, which are in the range space of Λ. This observation led Chamberlain and Rothschild (1983) to deduce that the principal components estimator is an asymptotic (large k) analog estimator for the factor loadings and factors. Bai (2003) demonstrated that the estimator is consistent as $n, k \to \infty$ jointly. The conditions and proofs are technical so are not reviewed here.

Now consider the estimated factors

$$\widehat{F}_i = \boldsymbol{D}^{-1/2}\boldsymbol{H}' X_i = \boldsymbol{D}^{-1}\Lambda' X_i$$

where for simplicity we ignore estimation error. Since $X_i = \Lambda F_i + u_i$ and $\Lambda'\Lambda = \boldsymbol{D}$ we can write this as

$$\widehat{F}_i = F_i + \boldsymbol{D}^{-1}\Lambda' u_i.$$

This shows that $\widehat{F}_i$ is an unbiased estimator for F_i and has variance $\operatorname{var}\left[\widehat{F}_i\right] = \boldsymbol{D}^{-1}\Lambda'\Psi\Lambda\boldsymbol{D}^{-1}$. Under Assumption 11.1, $\left\|\operatorname{var}\left[\widehat{F}_i\right]\right\| \leq B/\lambda_{\min}\left(\Lambda'\Lambda\right) \to 0$. Thus $\widehat{F}_i$ is consistent for F_i as $k \to \infty$. Bai (2003) shows that this extends to the feasible estimator as $n, k \to \infty$.

In Stata, the least squares estimator $\widehat{\Lambda}$ and factors $\widehat{F}_i$ can be calculated with the `factor, pcf factors(r)` command followed by `predict`. In R a feasible estimation approach is to calculate the factors by eigenvalue decomposition.

11.15 Factor Models with Additional Regressors

Consider the model

$$X = \Lambda F + \boldsymbol{B}Z + e$$

where X and e are $k \times 1$, Λ is $k \times r$, F is $r \times 1$, $\boldsymbol{B}$ is $k \times \ell$, and Z is $\ell \times 1$.

The coefficients Λ and $\boldsymbol{B}$ can be estimated by a combination of factor regression (either MLE or principal components) and least squares. The key is the following two observations:

1. Given $\boldsymbol{B}$, the coefficient Λ can be estimated by factor regression applied to $X - \boldsymbol{B}Z$.
2. Given the factors F, the coefficients Λ and $\boldsymbol{B}$ can be estimated by multivariate least squares of X on F and Z.

Estimation iterates between these two steps. Start with a preliminary estimator of $\boldsymbol{B}$ obtained by multivariate least squares of X on Z. Then apply the above two steps and iterate under convergence.

11.16 Factor-Augmented Regression

In the previous sections we considered factor models which decompose a set of variables into common factors and idiosyncratic errors. In this section we consider factor-augmented regression, which uses such common factors as regressors for dimension reduction.

Suppose we have the variables (Y, Z, X) where $Y \in \mathbb{R}$, $Z \in \mathbb{R}^\ell$, and $X \in \mathbb{R}^k$. In practice, k may be large and the elements of X may be highly correlated. The **factor-augmented regression** model is

$$\begin{aligned} Y &= F'\beta + Z'\gamma + e \\ X &= \Lambda F + u \\ \mathbb{E}[Fe] &= 0 \\ \mathbb{E}[Ze] &= 0 \\ \mathbb{E}\left[Fu'\right] &= 0 \\ \mathbb{E}[ue] &= 0, \end{aligned}$$

The random variables are $e \in \mathbb{R}$, $F \in \mathbb{R}^r$, and $u \in \mathbb{R}^k$. The regression coefficients are $\beta \in \mathbb{R}^k$ and $\gamma \in \mathbb{R}^\ell$. The matrix Λ are the factor loadings.

This model specifies that the influence of X on Y is through the common factors F. The idea is that the variation in the regressors is mostly captured by the variation in the factors, so the influence of the regressors can be captured through these factors. This can be viewed as a dimension-reduction technique as we have reduced the k-dimensional X to the r-dimensional F. Interest typically focuses on the regressors Z and its coefficients γ. The factors F are included in the regression as "controls" and its coefficient β is less typically of interest. Since it is difficult to interpret the factors F only their range space is identified it is generally prudent to avoid intrepreting the coefficients β.

The model is typically estimated in multiple steps. First, the factor loadings Λ and factors F_i are estimated by factor regression. In the case of principal-components estimation the factor estimates are the scaled[2] principal components $\widehat{F}_i = \widehat{\boldsymbol{D}}^{-1}\widehat{\Lambda}' X_i$. Second, Y is regressed on the estimated factors and the other regressors to obtain the estimator of β and γ. This second-step estimator equals (for simplicity assume there is no Z)

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{i=1}^{n} \widehat{F}_i \widehat{F}_i'\right)^{-1}\left(\sum_{i=1}^{n} \widehat{F}_i Y_i\right) \\ &= \left(\widehat{\boldsymbol{D}}^{-1}\widehat{\Lambda}' \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{\Lambda}\widehat{\boldsymbol{D}}^{-1}\right)^{-1}\left(\widehat{\boldsymbol{D}}^{-1}\widehat{\Lambda}' \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).\end{aligned}$$

Now let's investigate its asymptotic behavior. As $n \to \infty$, $\widehat{\Lambda} \xrightarrow[p]{} \Lambda$ and $\widehat{\boldsymbol{D}} \xrightarrow[p]{} \boldsymbol{D}$ so

$$\widehat{\beta} \xrightarrow[p]{} \beta^* = \left(\boldsymbol{D}^{-1}\Lambda' \mathbb{E}\left[XX'\right]\Lambda \boldsymbol{D}^{-1}\right)^{-1}\left(\boldsymbol{D}^{-1}\Lambda' \mathbb{E}\left[XY\right]\right). \tag{11.26}$$

Recall $\mathbb{E}\left[XX'\right] = \Lambda\Lambda' + \Psi$ and $\Lambda'\Lambda = \boldsymbol{D}$. We calculate that

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[(\Lambda F + u)\left(F'\beta + e\right)\right] = \Lambda\beta.$$

We find that the right-hand-side of (11.26) equals

$$\beta^* = \left(\boldsymbol{D}^{-1}\Lambda'\left(\Lambda\Lambda' + \Psi\right)\Lambda \boldsymbol{D}^{-1}\right)^{-1}\left(\boldsymbol{D}^{-1}\Lambda'\Lambda\beta\right) = \left(\boldsymbol{I}_r + \boldsymbol{D}^{-1}\Lambda'\Psi\Lambda\boldsymbol{D}^{-1}\right)^{-1}\beta$$

which does not equal β. Thus $\widehat{\beta}$ has a probability limit but is inconsistent for β as $n \to \infty$.

This deficiency diminishes as $k \to \infty$. Indeed,

$$\left\|\boldsymbol{D}^{-1}\Lambda'\Psi\Lambda\boldsymbol{D}^{-1}\right\| \le B\left\|\boldsymbol{D}^{-1}\right\| \to 0$$

as $k \to \infty$. This implies $\beta^* \to \beta$. Hence, if we take the sequential asymptotic limit $n \to \infty$ followed by $k \to \infty$, we find $\widehat{\beta} \xrightarrow[p]{} \beta$. This implies that the estimator is consistent. Bai (2003) demonstrated consistency under the more rigorous but technically challenging setting where $n, k \to \infty$ jointly. The implication of this result is that factor augmented regression is consistent if both the sample size and dimension of X are large.

For asymptotic normality of $\widehat{\beta}$ it turns out that we need to strengthen Assumption 11.1.2. The relevant condition is $n^{-1/2}\lambda_{\min}\left(\Lambda'\Lambda\right) \to \infty$. This is similar to the condition that $k^2/n \to \infty$. This is technical but can be interpreted as meaning that k is large relative to $\sqrt{n}$. Intuitively, this requires that the dimension of X is larger than sample size n.

In Stata, estimation takes the following steps. First, the `factor` command is used to estimate the factor model. Either MLE or principal components estimation can be used. Second, the `predict` command is used to estimate the factors, either by Barlett or regression scoring. Third, the factors are treated as regressors in an estimated regression.

11.17 Multivariate Normal*

Some interesting sampling results hold for matrix-valued normal variates. Let $\boldsymbol{Y}$ be an $n \times m$ matrix whose rows are independent and distributed $\mathrm{N}\left(\mu, \Sigma\right)$. We say that $\boldsymbol{Y}$ is **multivariate matrix normal**, and

[2]The unscaled principal components can equivalently be used if the coefficients $\widehat{\beta}$ are not reported. The coefficient estimates $\widehat{\gamma}$ are unaffected by the choice of factor scaling.

write $\boldsymbol{Y} \sim \mathrm{N}\left(\overline{\mu}, \boldsymbol{I}_n \otimes \Sigma\right)$, where $\overline{\mu}$ is $n \times m$ with each row equal to μ'. The notation is due to the fact that $\mathrm{vec}\left(\left(\boldsymbol{Y} - \mu\right)'\right) \sim \mathrm{N}\left(0, \boldsymbol{I}_n \otimes \Sigma\right)$.

Definition 11.2 If $n \times m$ $\boldsymbol{Y} \sim \mathrm{N}\left(\overline{\mu}, \boldsymbol{I}_n \otimes \Sigma\right)$ then $W = \boldsymbol{Y}'\boldsymbol{Y}$ is distributed **Wishart** with n degress of freedom and covariance matrix $\boldsymbol{\Sigma}$, and is written as $W \sim W_m\left(n, \Sigma\right)$.

The Wishart is a multivariate generalization of the chi-square. If $W \sim W_1\left(n, \sigma^2\right)$ then $W \sim \sigma^2 \chi^2_n$.

The Wishart arises as the exact distribution of a sample covariance matrix in the normal sampling model. The bias-corrected estimator of Σ is

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)\left(Y_i - \overline{Y}\right)'.$$

Theorem 11.10 If $Y_i \sim \mathrm{N}\left(\mu, \Sigma\right)$ are independent then $\widehat{\Sigma} \sim W_m\left(n-1, \frac{1}{n-1}\Sigma\right)$.

The following manipulation is useful.

Theorem 11.11 If $W \sim W_m\left(n, \Sigma\right)$ then for $m \times 1$ α, $\left(\alpha' W^{-1} \alpha\right)^{-1} \sim \dfrac{\chi^2_{n-m+1}}{\alpha' \Sigma^{-1} \alpha}$.

To prove this, note that without loss of generality we can take $\Sigma = \boldsymbol{I}_m$ and $\alpha'\alpha = 1$. Let $\boldsymbol{H}$ be $m \times m$ orthonormal with first row equal to α. so that $\boldsymbol{H}\alpha = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Since the distribution of $\boldsymbol{Y}$ and $\boldsymbol{Y}\boldsymbol{H}$ are identical we can without loss of generality set $\alpha = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Partition $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2]$ where $\boldsymbol{Y}_1$ is $n \times 1$, $\boldsymbol{Y}_2$ is $n \times (m-1)$, and they are independent. Then

$$\begin{aligned}
\left(\alpha' W^{-1} \alpha\right)^{-1} &= \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_1'\boldsymbol{Y}_1 & \boldsymbol{Y}_1'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\boldsymbol{Y}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right)^{-1} \\
&= \boldsymbol{Y}_1'\boldsymbol{Y}_1 - \boldsymbol{Y}_1'\boldsymbol{Y}_2\left(\boldsymbol{Y}_2'\boldsymbol{Y}_2\right)^{-1}\boldsymbol{Y}_2'\boldsymbol{Y}_1 \\
&= \boldsymbol{Y}_1'\boldsymbol{M}_2\boldsymbol{Y}_1 \sim \chi^2_{n-(m-1)}
\end{aligned}$$

where $\boldsymbol{M}_2 = \boldsymbol{I}_{m-1} - \boldsymbol{Y}_2\left(\boldsymbol{Y}_2'\boldsymbol{Y}_2\right)^{-1}\boldsymbol{Y}_2'$. The final distributional equality holds conditional on $\boldsymbol{Y}_2$ by the same argument in the proof of Theorem 5.7. Since this does not depend on $\boldsymbol{Y}_2$ it is the unconditional distribution as well. This establishes the stated result.

To test hypotheses about μ a classical statistic is known as **Hotelling's** T^2:

$$T^2 = n\left(\overline{Y} - \mu\right)'\widehat{\Sigma}^{-1}\left(\overline{Y} - \mu\right).$$

Theorem 11.12 If $Y \sim \mathrm{N}(\mu, \Sigma)$ then

$$T^2 \sim \frac{m}{(n-m)(n-1)} F(m, n-m)$$

a scaled F distribution.

To prove this recall that $\overline{Y}$ is independent of $\widehat{\Sigma}$. Apply Theorem 11.11 with $\alpha = \overline{Y} - \mu$. Conditional on $\overline{Y}$ and using the fact that $\widehat{\Sigma} \sim W_m\left(n-1, \frac{1}{n-1}\Sigma\right)$,

$$\begin{aligned}
\frac{n}{T^2} &= \left(\left(\overline{Y} - \Sigma\right)' \widehat{\Sigma}^{-1} \left(\overline{Y} - \Sigma\right)\right)^{-1} \\
&\sim \frac{\chi^2_{n-1-m+1}}{\left(\overline{Y} - \mu\right)' \left(\frac{1}{n-1}\Sigma\right)^{-1} \left(\overline{Y} - \mu\right)} \\
&\sim n(n-1)\frac{\chi^2_{n-m}}{\chi^2_m}.
\end{aligned}$$

Since the two chi-square variables are independent, this is the stated result.

A very interesting property of this result is that the T^2 statistic is a multivariate quadratric form in normal random variables, yet it has the exact F distribution.

11.18 Exercises

Exercise 11.1 Show (11.10) when the errors are conditionally homoskedastic (11.8).

Exercise 11.2 Show (11.11) when the regressors are common across equations $X_j = X$.

Exercise 11.3 Show (11.12) when the regressors are common across equations $X_j = X$ and the errors are conditionally homoskedastic (11.8).

Exercise 11.4 Prove Theorem 11.1.

Exercise 11.5 Show (11.13) when the regressors are common across equations $X_j = X$.

Exercise 11.6 Show (11.14) when the regressors are common across equations $X_j = X$ and the errors are conditionally homoskedastic (11.8).

Exercise 11.7 Prove Theorem 11.2.

Exercise 11.8 Prove Theorem 11.3.

Exercise 11.9 Show that (11.16) follows from the steps described.

Exercise 11.10 Show that (11.17) follows from the steps described.

Exercise 11.11 Prove Theorem 11.4.

Exercise 11.12 Prove Theorem 11.5.

Hint: First, show that it is sufficient to show that

$$\mathbb{E}\left[\overline{X}'\overline{X}\right]\left(\mathbb{E}\left[\overline{X}'\Sigma^{-1}\overline{X}\right]\right)^{-1}\mathbb{E}\left[\overline{X}'\overline{X}\right] \leq \mathbb{E}\left[\overline{X}'\Sigma\overline{X}\right].$$

Second, rewrite this equation using the transformations $U = \Sigma^{1/2}\overline{X}$ and $V = \Sigma^{1/2}\overline{X}$, and then apply the matrix Cauchy-Schwarz inequality (B.33).

Exercise 11.13 Prove Theorem 11.6.

Exercise 11.14 Take the model

$$\begin{aligned} Y &= \pi'\beta + e \\ \pi &= \mathbb{E}[X \mid Z] = \Gamma' Z \\ \mathbb{E}[e \mid Z] &= 0 \end{aligned}$$

where Y is scalar, X is a k vector and Z is an ℓ vector. β and π are $k \times 1$ and Γ is $\ell \times k$. The sample is $(Y_i, X_i, Z_i : i = 1, ..., n)$ with π_i unobserved.

Consider the estimator $\widehat{\beta}$ for β by OLS of Y on $\widehat{\pi} = \widehat{\Gamma}' Z$ where $\widehat{\Gamma}$ is the OLS coefficient from the multivariate regression of X on Z.

(a) Show that $\widehat{\beta}$ is consistent for β.

(b) Find the asymptotic distribution $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$ assuming that $\beta = 0$.

(c) Why is the assumption $\beta = 0$ an important simplifying condition in part (b)?

(d) Using the result in (c) construct an appropriate asymptotic test for the hypothesis $\mathbb{H}_0 : \beta = 0$.

Exercise 11.15 The observations are i.i.d., $(Y_{1i}, Y_{2i}, X_i : i = 1, ..., n)$. The dependent variables Y_1 and Y_2 are real-valued. The regressor X is a k-vector. The model is the two-equation system

$$\begin{aligned} Y_1 &= X'\beta_1 + e_1 \\ \mathbb{E}[Xe_1] &= 0 \\ Y_2 &= X'\beta_2 + e_2 \\ \mathbb{E}[Xe_2] &= 0. \end{aligned}$$

(a) What are the appropriate estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ for β_1 and β_2?

(b) Find the joint asymptotic distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(c) Describe a test for $\mathbb{H}_0 : \beta_1 = \beta_2$.