ECON 640 Multivariate OLS (Matrix Notation)

Dr. Yang Liang

Introduction

This document outlines the key properties of the Multivariate Regression Model, including its interpretation, estimation methods, and statistical properties. These notes are based on Chapters 6-9 and 18 from Stock and Watson.

Some Important Matrix Rules

if A and B are vectors of the same dimension, then A'B will give you a scalar that

$$A'B = B'A$$
$$A'B + B'A = 2A'B = 2B'A$$

Matrix Differentiation

$$\frac{\partial a'X}{\partial b} = a, \quad \frac{\partial b'a}{\partial X} = a$$
$$\frac{\partial b'Ab}{\partial b} = (A + A')b = 2Ab \quad \text{if } A \text{ is symmetric}$$

1 Multivariate Regression in Scalar Notation

When dealing with more than one explanatory variable, the regression model is extended as follows:

$$y_i = \beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k + u_i$$

The goal is to minimize the sum of squared errors (you should take the FOC yourself):

$$\sum_{i=1}^{n} (y_i - (b_1 + x_{2i}b_2 + \cdots + x_{ki}b_k))^2$$

This leads to the first-order conditions for minimizing the sum of squared residuals:

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} e_i x_{2i} = 0, \quad \ldots, \quad \sum_{i=1}^{n} e_i x_{ki} = 0$$

where the residual e_i is defined as:

$$e_i = y_i - (\hat{\beta}_1 + x_{2i}\hat{\beta}_2 + \cdots + x_{ki}\hat{\beta}_k)$$

Solving these k equiations with k unknowns $\hat{\beta}_1, \ldots, \hat{\beta}_k$ is in principle no problem
But... much easier to switch to matrix notation

2 Matrix Notation

The multivariate regression model can be written in matrix notation as:

$$y = X\beta + u$$

where:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{21} & \dots & x_{k1} \\ 1 & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \dots & x_{kn} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

The least squares estimator of β, denoted by $\hat{\beta}$, is the solution to:

$$\hat{\beta} = (X'X)^{-1}X'y$$

The least squares estimator of β ($\hat{\beta}$) is the value of β that minimizes the sum of squared deviations of the y_i's from the fitted line, given by:

$$\hat{y}_i = x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \cdots + x_{ik}\hat{\beta}_k$$

We call these deviations the residuals.

2.1 Derive OLS - Define Residuals in Matrix

The residuals are given by:

$$e_i = y_i - \hat{y}_i$$

The least squares estimator of β minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = e'e = (y - \hat{y})'(y - \hat{y}) = (y - X\hat{\beta})'(y - X\hat{\beta})$$

where e is the vector of residuals and $\hat{y}$ is the vector of fitted values.

2.2 Derive OLS - First Order Conditions in Matrix

The first-order conditions for minimizing $(y - X\hat{\beta})'(y - X\hat{\beta})$ are:

$$\frac{\partial (y - Xb)'(y - Xb)}{\partial b} = \frac{\partial (y'y + (Xb)'Xb - y'Xb - b'X'y)}{\partial b} = \frac{\partial (y'y + b'(X'X)b - 2y'Xb)}{\partial b}$$

$$= \frac{\partial (y'y + b'X'Xb - 2b'(X'y))}{\partial b} = 2X'Xb - 2X'y = 0$$

Solving this gives:

$$X'y = X'X\hat{\beta} \quad \text{or} \quad \hat{\beta} = (X'X)^{-1}X'y$$

This is the ordinary least squares (OLS) estimator in general format.

2.3 Projection and Annihilation Matrix Properties (important!)

The OLS estimator is simply the linear projection of Y onto the X space. The projection matrix maps any vector in X space onto itself, while the annihilation matrix maps any vector orthogonal to X space onto zero.

$$P_X = X(X'X)^{-1}X', \quad P_X \cdot X = X$$
$$M_X = I - P_X, \quad M_X P_X = 0, \quad M_X X = 0$$

Geometry of OLS (please draw the triangle yourself, also important!)

Relationship between vectors:

$$X\hat{\beta} = P_X Y$$

Orthogonality condition:

$$\hat{u} = M_X Y \quad \text{(orthogonal)}$$

3 Statistical Properties of the OLS Estimator in Matrix Notation

Note that:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + u) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\ &= \beta + (X'X)^{-1}X'u\end{aligned}$$

Unbiasedness of $\hat{\beta}$

$$\begin{aligned}E(\hat{\beta}) &= E\left[(X'X)^{-1}X'Y\right] = E\left[(X'X)^{-1}X'(X\beta + u)\right] \\ &= \beta + E\left[(X'X)^{-1}X'u\right]\end{aligned}$$

Given that $E(u|X) = 0$:

$$E\left[(X'X)^{-1}X'u\right] = 0$$

Therefore:

$$E(\hat{\beta}) = \beta$$

Variance of $\hat{\beta}$

$$\hat{\beta} = \beta + (X'X)^{-1}X'u$$

Then, get rid of the constant β

$$\text{Var}(\hat{\beta}) = \text{Var}[(X'X)^{-1}X'u]$$

Expanding the variance and conditioning on X:

$$\text{Var}(\hat{\beta}|\mathbf{X}) = (X'X)^{-1}X' \cdot \text{Var}(u) \cdot X(X'X)^{-1}$$

Since $\text{Var}(u) = \sigma^2 I$:

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2 (X'X)^{-1} X'X (X'X)^{-1}$$

Simplifying the expression:

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2 (X'X)^{-1}$$

Which is also called the Variance-Covariance Matrix

4 Prove the unbiasedness again, this time using scalar notation while incorporating matrix concepts

We know:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})X_i} = \sum_{i=1}^n W_i Y_i$$

You should think of W_i as the equivalent of $(X'X)^{-1}X'$ in matrix form.

$$\text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

Proof: First, note that for any constant c, (**why?**)

$$\sum_{i=1}^n W_i c = 0.$$

then,

$$E\left[\hat{\beta}_1 | X_1, \ldots, X_n\right] = E\left[\sum_{i=1}^n W_i Y_i \middle| X_1, \ldots, X_n\right]$$

$$= \sum_{i=1}^n W_i E\left[Y_i | X_1, \ldots, X_n\right]$$

$$= \sum_{i=1}^n W_i \left(\beta_0 + \beta_1 X_i\right)$$

$$= \beta_1 \sum_{i=1}^n W_i X_i$$

$$= \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})X_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1$$

5 Using the matrix structure, you can now visualize the elements within the matrix notation and rewrite the above proofs in scalar notation.

This will help study the large sample properties of the OLS estimator as now it can be written as a function of n (sample size)

We can write the regression model (for each draw) as:

$$y_i = x_i \beta + u_i$$

Let $\hat{\beta}_n$ be the OLS estimator based on n observations:

$$\hat{\beta}_n = \left[\sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\sum_{i=1}^{n} x_i' y_i \right]$$

$$= \beta + \left[\sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\sum_{i=1}^{n} x_i' u_i \right]$$

$$= \beta + \left[\frac{1}{n} \sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} x_i' u_i \right]$$

5.1 Law of Large Numbers

Assuming that all relevant moments exist, all elements of

$$\frac{1}{n} \sum_{i=1}^{n} x_i' x_i \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} x_i' u_i$$

converge to the corresponding elements of $E[x_i x_i']$ and $E[x_i u_i]$ by the Law of Large Numbers.

5.2 Convergence to β

Now assume that $E[x_i u_i] = 0$. Since

$$\left[\frac{1}{n} \sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} x_i' u_i \right]$$

is a continuous function of the elements of $\frac{1}{n} \sum_{i=1}^{n} x_i x_i'$ and $\frac{1}{n} \sum_{i=1}^{n} x_i u_i$, this implies that:

$$\left[\frac{1}{n} \sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} x_i' u_i \right] \xrightarrow{p} E[x_i' x_i]^{-1} E[x_i' u_i] = 0$$

Hence, we have:

$$\hat{\beta}_n = \beta + \left[\frac{1}{n} \sum_{i=1}^{n} x_i' x_i \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} x_i' u_i \right] \xrightarrow{p} \beta$$

5.3 Rate of Convergence

Also,

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i' x_i\right]^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i' u_i\right]$$

By the Law of Large Numbers:

$$\frac{1}{n}\sum_{i=1}^{n} x_i' x_i \xrightarrow{p} E[x_i' x_i]$$

and hence:

$$\left[\frac{1}{n}\sum_{i=1}^{n} x_i' x_i\right]^{-1} \xrightarrow{p} E[x_i' x_i]^{-1}$$

The OLS estimator has several important properties. In large samples, under the assumption that the error term u_i is homoskedastic and normally distributed, the estimator $\hat{\beta}$ is unbiased and has the following distribution:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma)$$

where Σ is the variance-covariance matrix of the OLS estimator.
Proof: By the Central Limit Theorem:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \xrightarrow{d} N(0, E[x_i u_i u_i' x_i'])$$

Therefore:

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right]^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i\right]$$

$$\xrightarrow{d} E[x_i x_i']^{-1} \times N(0, E[x_i u_i u_i' x_i'])$$

Or equivalently:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N\left(0, E[x_i x_i']^{-1} E[x_i u_i u_i' x_i'] E[x_i x_i']^{-1}\right)$$

Estimating the Variance of $\hat{\beta}_n$ and obtaining standard errors

This implies that the variance of $\hat{\beta}_n$ can be approximated by:

$$\widehat{V}[\hat{\beta}_n] = \left[\sum_{i=1}^{n} x_i' x_i\right]^{-1} \left[\sum_{i=1}^{n} x_i' e_i^2 x_i\right] \left[\sum_{i=1}^{n} x_i' x_i\right]^{-1}$$

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} x_i' x_i\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} x_i' e_i^2 x_i\right] \left[\frac{1}{n}\sum_{i=1}^{n} x_i' x_i\right]^{-1}$$

where $e_i = y_i - x_i\hat{\beta}$.

We sometimes write:

$$\hat{\beta}_n \overset{approx}{\sim} N\left(\beta, \widehat{V}[\hat{\beta}_n]\right)$$

The approximate 95% confidence interval for $\hat{\beta}_{n,j}$ is given by:

$$\hat{\beta}_{n,j} \pm 1.96\sqrt{\widehat{V}[\hat{\beta}_n]_{jj}}$$

To test the null hypothesis:

$$H_0 : \beta_j = \beta_j^0$$

The test statistic is:

$$t = \frac{\hat{\beta}_{n,j} - \beta_j^0}{\sqrt{\widehat{V}[\hat{\beta}_n]_{jj}}}$$

6 Method of Moments

A completely new estimator but is equivalent to the OLS estimator

Let

$$u_i = Y_i - \beta_0 - \beta_1 X_i \quad \text{(key is always in the error term)}$$

We start by taking the average of the residuals:

$$\frac{1}{n}\sum u_i = \frac{1}{n}\sum(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{1}{n}\sum(Y_i - \beta_0 - \beta_1 X_i) \xrightarrow{p} 0 \quad \text{as } n \to \infty \tag{1}$$

Using the expectations:

$$E(u_i|X) = 0 \quad \text{and} \quad E(u_i X_i|X) = 0$$

Taking the expectation of the residuals with respect to X_i:

$$\frac{1}{n}\sum u_i X_i = \frac{1}{n}\sum X_i(Y_i - \beta_0 - \beta_1 X_i) = \frac{1}{n}\sum(X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) \xrightarrow{p} 0 \quad \text{as } n \to \infty \tag{2}$$

Hence, we can estimate these two moments using our available data by setting:

$$\frac{1}{n}\sum(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{1}{n}\sum(X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) = 0$$

Expanding and rewriting the equations we get:

$$\beta_0 + \left(\frac{1}{n}\sum X_i\right)\beta_1 = \frac{1}{n}\left(\sum Y_i\right)$$

$$\left(\frac{1}{n}\sum X_i\right)\beta_0 + \left(\frac{1}{n}\sum X_i^2\right)\beta_1 = \frac{1}{n}\left(\sum X_i Y_i\right)$$

times n on both sides for both equations, we can write down the following:

$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Hence:

$$\begin{bmatrix} \hat{\beta_0} \\ \hat{\beta_1} \end{bmatrix}_{\textbf{Method of Moments}} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

Is this the same as the OLS estimator?

Write down the linear model, and calculate the OLS estimator, then compare!

$$Y_i = \begin{pmatrix} Y_1 \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_n \end{pmatrix}$$

Calculating $X'X$:

$$X'X = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}$$

Calculating $X'Y$

$$X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta_0} \\ \hat{\beta_1} \end{bmatrix}_{\textbf{OLS}} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

7 The Frisch–Waugh–Lovell (FWL) theorem

The Frisch–Waugh–Lovell (FWL) theorem provides an elegant insight into how ordinary least squares (OLS) operates when there are multiple regressors. It shows that the coefficient on one regressor (or group of regressors) can be obtained by first removing the influence of other variables from both the dependent variable and the variable(s) of interest, and then regressing the residuals on each other (a dimension reduction process).

Model Setup

Consider the linear model:

$$y = X_1\beta_1 + X_2\beta_2 + u$$

where

- y is an $n \times 1$ vector of observations on the dependent variable.
- X_1 is an $n \times k_1$ matrix of regressors of interest.
- X_2 is an $n \times k_2$ matrix of control regressors.
- u is the $n \times 1$ vector of errors.

The OLS estimator for the full model is:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

Theorem Statement

Frisch–Waugh–Lovell Theorem:

Let $M_2 = I - P_2$ denote the residual-maker matrix, where $P_2 = X_2(X_2'X_2)^{-1}X_2'$ is the projection matrix onto the column space of X_2. Then:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2y$$

That is, $\hat{\beta}_1$ can be obtained in three steps:

1. Regress y on X_2 and obtain the residuals M_2y.
2. Regress each column of X_1 on X_2 and obtain the residuals M_2X_1.
3. Regress M_2y on M_2X_1 (without a constant, why?).

8 Intuition

- M_2y represents the part of y unexplained by X_2.
- M_2X_1 represents the part of X_1 unexplained by X_2.

Thus, the FWL theorem tells us that the effect of X_1 on y after controlling for X_2 is simply the relationship between their residualized versions.

Stata Example

```
// =============================================
// Example:  FrischWaughLovell  Theorem in Stata
// =============================================

// Load example data
sysuse auto, clear

// Full model
reg price weight mpg, vce(robust)
est store full

// Step 1: Residualize y on control
reg price mpg
predict double ytilde, resid

// Step 2: Residualize X1 on control
reg weight mpg
predict double wtilde, resid

// Step 3: Regression of residuals on residuals (no constant)
```

```
reg ytilde wtilde, nocons vce(robust)
est store fwl

// Compare coefficients
esttab full fwl, se stats(N r2)

// ---------------------------------------------
// Visualization: Added-variable (partial) plot
// ---------------------------------------------
// Manual FWL visualization (custom twoway)
reg price mpg
predict double ytilde2, resid
reg weight mpg
predict double wtilde2, resid

twoway ///
 (scatter ytilde2 wtilde2, msize(small) mlabel(make)) ///
 (lfit ytilde2 wtilde2), ///
 ytitle("price   mpg") xtitle("weight   mpg") ///
 title("FWL Visualization: residualized price vs residualized weight") ///
 legend(off)
```

Measures of Fit

The goodness of fit of a regression model is measured by the R^2 statistic:

$$R^2 = 1 - \frac{SSR}{TSS}$$

where SSR is the sum of squared residuals and TSS is the total sum of squares. The adjusted R^2 corrects for the number of regressors.
Let

$$\hat{u}_i = Y_i - \hat{Y}_i \quad \text{(residuals)}$$

where

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad \text{(fitted values)}$$

then

$$(1) \quad \Rightarrow \quad \sum_{i=1}^{n} \hat{u}_i = 0$$

$$(2) \quad \Rightarrow \quad \sum_{i=1}^{n} X_i \hat{u}_i = 0$$

Easy to show that

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \\
&= RSS + ESS
\end{aligned}
$$

Define

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \\
&= \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \\
&0 \leq R^2 \leq 1
\end{aligned}
$$

F-tests for Joint Hypotheses

(optional)

The F-statistic is used to test joint hypotheses about the regression coefficients. For example, to test whether $\beta_2 = 0$ and $\beta_3 = 0$, we use:

$$
F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n-k-1)}
$$

where q is the number of restrictions.

Stata practice

install the package to use `bcuse` by baum to load SW datasets:
`http://fmwww.bc.edu/ec-p/data/stockwatson/datasets.list.html` or use `sysuse nlsw88.dta`