Matrix Algebra for OLS Estimator

Big Picture

- Matrix algebra can produce compact notation.
- Some packages such as Matlab are matrix-oriented.
- Excel spreadsheet is just a matrix.

Dependent Variable

- Suppose the sample consists of n observations.
- The dependent variable is denoted as an $n \times 1$ (column) vector

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- The subscript indexes the observation.
- We use boldface for vector and matrix.

Independent Variables

- Suppose there are k independent variables and a constant term. In the spreadsheet there are $k+1$ columns and n rows.
- Mathematically that spreadsheet corresponds to an $n \times (k+1)$ matrix, denoted by $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

where x_{ij} is the i-th observation of the j-th independent variable.

Linear Regression Model

- Define β as a $(k+1) \times 1$ vector of coefficients

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

and $\mathbf{U}$ as an $n \times 1$ vector of error terms. The linear multiple regression model in matrix form is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

- Read Appendix D of the textbook.
- The key to work with matrix is keeping track of the dimension.

First Order Conditions of Minimizing RSS

- The OLS estimators are obtained by minimizing residual sum squares (RSS). The first order conditions are

$$\frac{\partial \texttt{RSS}}{\partial \hat{\beta}_j} = 0 \Rightarrow \sum_{i=1}^{n} x_{ij}\hat{u}_i = 0, \qquad (j = 0, 1, \ldots, k)$$

  where $\hat{u}$ is the residual. We have a system of $k+1$ equations.

- This system of equations can be written in matrix form as

$$\mathbf{X}'\widehat{\mathbf{U}} = \mathbf{0}$$

  where $\mathbf{X}'$ is the transpose of $\mathbf{X}$. Notice boldface $\mathbf{0}$ denotes a $(k+1) \times 1$ vector of zeros.

OLS Estimators in Matrix Form

- Let $\hat{\beta}$ be a $(k+1) \times 1$ vector of OLS estimates. We have

$$\mathbf{X}'\widehat{\mathbf{U}} = \mathbf{0} \tag{1}$$

$$\Rightarrow \quad \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0} \tag{2}$$

$$\Rightarrow \quad \mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\hat{\beta} \tag{3}$$

$$\Rightarrow \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \tag{4}$$

  where $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse matrix of $\mathbf{X}'\mathbf{X}$. That inverse exists if $\mathbf{X}$ has column rank $k+1$, that is, there is no perfect multicollinearity. One example of perfect multicollinearity is dummy variable trap.

- Stata command `reg` uses the formula (4) to compute $\hat{\beta}$.

An Important Result

- We can show

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \tag{5}$$

$$\Rightarrow \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X}\beta + \mathbf{U})) \tag{6}$$

$$\Rightarrow \quad \hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{U}) \tag{7}$$

  where we use the property of inverse matrix and identity matrix: $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}$, and $\mathbf{I}\beta = \beta$.

- This shows $\hat{\beta}$ in general differs from β due to the error $\mathbf{U}$.

- β is an (unknown) constant, while $\hat{\beta}$ is a random variable because $\mathbf{U}$ is random. $\hat{\beta}$ varies across different samples. The distribution of $\hat{\beta}$ is called sampling distribution.

Statistical Properties of OLS Estimator I

Under the assumptions of (1) random sample (or iid sample), and (2) $E(u_i|x_1, \ldots, x_k) = 0$ we have

$$E(\hat{\beta}|\mathbf{X}) = E(\beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{U})|\mathbf{X}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{U}|\mathbf{X}) = \beta$$

Then the law of iterated expectation implies that

$$E(\hat{\beta}) = E(E(\hat{\beta}|\mathbf{X})) = E(\beta) = \beta$$

So under certain assumptions the OLS estimator is unbiased.

Statistical Properties of OLS Estimator II

Most likely $\hat{\beta}$ is biased for two reasons:

1. The sample is not iid. For example, time series data are most often dependent. So $\hat{\beta}$ in a time series regression usually is biased.
2. $E(\mathbf{U}|\mathbf{X}) \neq 0$, which can be attributed to omitted variable, simultaneity and measurement error.

Statistical Properties of OLS Estimator III

Under the additional assumptions of (3) $E(u_i^2|x_1, \ldots, x_k) = \sigma^2$ (homoskedasticity) we have

$$E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \sigma^2\mathbf{I}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix. Under these three assumptions the conditional variance-covariance matrix of OLS estimator is

$$E((\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{8}$$

By default command `reg` uses formula (8) to report standard error, t value, etc. Remember they are valid only if homoskedasticity holds.

Heteroskedasticity

If heteroskedasticity is present (still assuming independence), we have

$$E(u_i^2|x_1,\ldots,x_k) = \sigma_i^2 \neq \texttt{constant}$$

$$E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \Omega \equiv \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}$$

which is a diagonal matrix but the terms on the diagonal line are not constant. In this case the correct variance-covariance matrix is

$$E((\hat{\beta}-\beta)(\hat{\beta}-\beta)'|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \qquad (9)$$

White Sandwich Estimator

Halbert White (*Econometrica*, 1980) suggests that estimating the unknown Ω with

$$\widehat{\Omega} = \begin{pmatrix} \hat{u}_1^2 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{u}_n^2 \end{pmatrix}$$

We can show

$$\mathbf{X}'\widehat{\Omega}\mathbf{X} = \sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'$$

where $\mathbf{x}_i'$ is the i-th row of $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

The command `reg y x, r` uses the White Sandwich Estimator

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\widehat{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \tag{10}$$

to compute the heteroskedasticity-robust standard error, t value, etc.

Predicted Values

The vector of predicted (fitted) values is

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} && (11) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} && (12) \\ &= \mathbf{P}\mathbf{Y} && (13) \end{aligned}$$

where

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is called projection matrix. It is a symmetric idempotent matrix satisfying

$$\mathbf{P} = \mathbf{P}', \quad \mathbf{P}\mathbf{P} = \mathbf{P}, \quad \mathbf{P}\mathbf{X} = \mathbf{X}.$$

Residuals

We can show the vector of residuals is

$$\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{M}\mathbf{Y}$$

where

$$\mathbf{M} \equiv \mathbf{I} - \mathbf{P}$$

is another symmetric idempotent matrix satisfying

$$\mathbf{M} = \mathbf{M}', \quad \mathbf{M}\mathbf{M} = \mathbf{M}, \quad \mathbf{P}\mathbf{M} = \mathbf{0}$$

Exercise: prove that

$$\mathbf{M}\mathbf{X} = \mathbf{0} \tag{14}$$

$$\widehat{\mathbf{U}} = \mathbf{M}\mathbf{U} \tag{15}$$

Frisch Waugh Theorem I

Using the partitioned (block) matrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$

we can write

$$\mathbf{Y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \widehat{\mathbf{U}}$$

Consider

$$\mathbf{M}_2 \equiv \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$$

The homework will ask you to prove:

$$\mathbf{M}_2\mathbf{M} = \mathbf{M}$$

That means

$$\mathbf{M}_2\mathbf{Y} = \mathbf{M}_2\mathbf{X}_1\hat{\beta}_1 + \widehat{\mathbf{U}}$$

Frisch Waugh Theorem II

After pre-multiplying both sides by $\mathbf{X}_1'$ we have

$$\begin{aligned} \hat{\beta}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}) && (16) \\ &= (\widehat{\mathbf{r}}'\widehat{\mathbf{r}})^{-1}(\widehat{\mathbf{r}}'\mathbf{Y}) \quad (\texttt{FW Theorem}) && (17) \end{aligned}$$

where $\widehat{\mathbf{r}} \equiv \mathbf{M}_2\mathbf{X}_1$ is the residual of regressing $\mathbf{X}_1$ onto $\mathbf{X}_2$.

Frisch Waugh Theorem III

The FW theorem states that

- $\hat{\beta}_1$ can be obtained in a two-step procedure. In step I, regress $\mathbf{X}_1$ onto $\mathbf{X}_2$ and save the residual $\widehat{\mathbf{r}}$. In step two, regress $\mathbf{Y}$ onto $\widehat{\mathbf{r}}$.
- $\hat{\beta}_1$ measures the effect of $\mathbf{X}_1$ on $\mathbf{Y}$, after the effect of $\mathbf{X}_2$ has been netted out.

Two Important Results

The homework will ask you to prove (based on the FW theorem)

$$\hat{\beta}_1 = \beta_1 + (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{M}_2\mathbf{U}) \tag{18}$$

$$E((\hat{\beta}_1 - \beta_1)'(\hat{\beta}_1 - \beta_1)|\mathbf{X}) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \tag{19}$$

The second result implies that if $\mathbf{X}_1$ and $\mathbf{X}_2$ are highly correlated (called multicollinearity), then the variance for $\hat{\beta}_1$ will be big. How about t value and p value?

Testing Linear Restrictions I

Consider a $q \times (k+1)$ matrix $\mathbf{R}$ and the null hypothesis

$$H_0 : \mathbf{R}\beta = \mathbf{c}. \tag{20}$$

This hypothesis involves multiple restrictions if $q > 1$, and can be tested by using Wald or F test.

Testing Linear Restrictions II

It is straightforward to show under H_0 (20) and the assumption of homoskedasticity,

$$\mathbf{R}\hat{\beta} \sim N(\mathbf{R}\beta, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') = N(\mathbf{c}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$$

Therefore

$$\begin{aligned} \texttt{Wald Test} &= (\mathbf{R}\hat{\beta} - \mathbf{c})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{c}) \quad (21) \\ &\sim \chi^2(q) \quad (\texttt{as } n \to \infty) \quad (22) \end{aligned}$$

where $\chi^2(q)$ denotes the chi-squared distribution with degree of freedom q. We get the F test after dividing the Wald test by q.

Generalized Least Squares (GLS)

The GLS estimator is more efficient (having smaller variance) than OLS in the presence of heteroskedasticity. Consider a three-step procedure:

1. Regress $\log(\hat{u}_i^2)$ onto x, keep the fitted value $\hat{g}_i$, and compute $\hat{h}_i = e^{\hat{g}_i}$

2. Construct

$$\mathbf{X}'\tilde{\Omega}^{-1}\mathbf{X} = \sum_{i=1}^{n} \hat{h}_i^{-1}\mathbf{x}_i\mathbf{x}_i', \quad \mathbf{X}'\tilde{\Omega}^{-1}\mathbf{Y} = \sum_{i=1}^{n} \hat{h}_i^{-1}\mathbf{x}_i\mathbf{y}_i \tag{23}$$

3. The feasible GLS estimator is

$$\hat{\beta}^{\texttt{fgls}} = (\mathbf{X}'\tilde{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\tilde{\Omega}^{-1}\mathbf{Y})$$

Delta Method

Consider a nonlinear function of OLS estimator $g(\hat{\beta})$. The delta method can be used to compute the variance-covariance matrix of $g(\hat{\beta})$. The key is the first-order Taylor expansion:

$$g(\hat{\beta}) \approx g(\beta) + \frac{dg}{dx}(\hat{\beta} - \beta)$$

where $\frac{dg}{dx}$ is the first order derivative of $g(\,)$ evaluated at β. As a result

$$\texttt{var}(g(\hat{\beta})) = \left(\frac{dg}{dx}\right) \texttt{var}(\hat{\beta}) \left(\frac{dg}{dx}\right)' \quad (\texttt{Delta Method})$$