Econometrics – Matrices for Linear Regressions

William M Volckmann II

January 6, 2017

1 Introduction

Suppose we have a bunch of data. In particular, we've made t observations of k variables. In other words, for the tth observation, we've recorded data $y_t, x_{t1}, x_{t2}, \ldots, x_{tk}$. We want to take all of this data and come up with an equation that best describes the value of y given some values $x_1, \ldots, x_k$. And we want to do it with matrices.

Assuming that y is linearly related to $x_1, \ldots, x_k$, we want to find the values of the parameters $\beta_0, \ldots, \beta_k$ that come "closest" to satisfying the system of equations

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk}. \end{aligned}$$

Letting $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$, write the system as

$$\begin{aligned} y_1 &= \mathbf{x}_1\boldsymbol{\beta} + u_1 \\ &\vdots \\ y_n &= \mathbf{x}_n\boldsymbol{\beta} + u_n, \end{aligned}$$

where the u_t terms represent **error terms**. Our model will almost certainly not be a perfect fit. But we're trying to get it as close as we can, and the error terms represent any error the model might introduce. This can be expressed entirely in terms of matrices as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \qquad \Longrightarrow \qquad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

2 Estimating the Parameters

It is exceedingly unlikely that we will find a $\boldsymbol{\beta}$ vector that will exactly satisfy the system of equations – that is why we include the error terms u_t. Thus, the **fitted values** that our model predicts will usually be at least a little bit off the true value. For instance, even though $y_t = 444$, our model might predict,

$$\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 x_{t1} + \ldots + \widehat{\beta}_k x_{tk} = 444.44,$$

where the hats represent numbers derived in the model. The difference between the real value and the fitted value is called the **residual**, and it is given by

$$y_t - \widehat{y} - t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_{t1} - \ldots - \widehat{\beta}_k x_{tk}.$$

We can write the matrix of residuals as $\mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{x}\widehat{\boldsymbol{\beta}}$.

In order to come up with the best possible estimate, we want to minimize these residuals in some overall sense. One way to measure the "total" residual is by squaring each residual and summing them, known as the **sum of squared residuals (SSR)**. So to estimate the $\boldsymbol{\beta}$ vector, we aim to minimize the sum of square residuals. We find the minimizing values for $\boldsymbol{\beta}$ by solving

$$\arg\min_{b} \mathrm{SSR}(\mathbf{b}) = \arg\min_{b} \sum_{t=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2. \tag{1}$$

One of the reasons we like to use squared residuals is because each term is convex, and the sum of a finite number of convex functions is itself convex. Thus, first order conditions are both necessary and sufficient for minimization, making this a rather ordinary multivariable calculus problem. That is, the solution must solve

$$\frac{\partial \mathrm{SSR}(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{x}_t'(y_t - \mathbf{x}_t\mathbf{b}) = 0.$$

We can, of course, drop the -2.

A good question to ask is why the $\mathbf{x}_t'$ is transposed. That is a matter of matrix multiplication compatibility – we cannot multiply two $1 \times k$ matrices together. Thus, the matrix equivalent of "pre-squaring" $\mathbf{x}_t$ is

$$\mathbf{x}_t'\mathbf{x}_t = \begin{bmatrix} 1 \\ x_{t1} \\ x_{t2} \\ \vdots \\ x_{tk} \end{bmatrix} \begin{bmatrix} 1 & x_{t1} & x_{t2} & \dots & x_{tk} \end{bmatrix} = \begin{bmatrix} 1 & x_{t1} & x_{t2} & \dots & x_{tk} \\ x_{t1} & x_{t1}^2 & x_{t1}x_{t2} & \dots & x_{t1}x_{tk} \\ x_{t2} & x_{t2}x_{t1} & x_{t2}^2 & \dots & x_{t1}x_{tk} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{tk} & x_{tk}x_{t1} & x_{tk}x_{t2} & \dots & x_{tk}^2 \end{bmatrix}.$$

What this amounts to is the first order condition

$$\begin{aligned}\sum_{t=1}^{n}(y_t - b_0 - b_1 x_{t1} - \ldots - b_k x_{tk}) &= 0,\\ \sum_{t=1}^{n} x_{t1}(y_t - b_0 - b_1 x_{t1} - \ldots - b_k x_{tk}) &= 0,\\ &\vdots\\ \sum_{t=1}^{n} x_{tk}(y_t - b_0 - b_1 x_{t1} - \ldots - b_k x_{tk}) &= 0.\end{aligned}$$

Or, in as a system of matrix equations,

$$\begin{aligned}\mathbf{x}_1'(y_1 - \mathbf{x}_1\mathbf{b}) &= 0,\\ \mathbf{x}_2'(y_2 - \mathbf{x}_2\mathbf{b}) &= 0,\\ &\vdots\\ \mathbf{x}_n'(y_n - \mathbf{x}_n\mathbf{b}) &= 0.\end{aligned}$$

We can condense this into a single matrix as

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \qquad \Longrightarrow \qquad \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}. \tag{2}$$

Solve for $\mathbf{b}$ by pre-multipling both sides by $(\mathbf{X}'\mathbf{X})^{-1}$ and we find the **ordinary least squares (OLS)** estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3}$$

Equation (3) is the critical formula for the matrix analysis of multiple linear regressions. Of course, we have to assume that the columns of $\mathbf{X}$ are linearly independent in order for $\mathbf{X}'\mathbf{X}$ to be invertible. It might be tempting to use the fact that $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ to cancel out the two $\mathbf{X}'$ terms, but this operation is only valid if $\mathbf{X}$ itself is a square matrix (i.e. when

$n = k + 1$) so that it can actually be inverted. In general, it will not be.

Because $\mathbf{y} = \mathbf{X}\mathbf{b} - \mathbf{u}$, we can write the fitted values as $\widehat{\mathbf{u}} = \mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{y}}$. From equation (2), it follows that $\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0}$. Explicitly,

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \widehat{u_1} \\ \widehat{u_2} \\ \vdots \\ \widehat{u_n} \end{bmatrix} = \begin{bmatrix} \widehat{u_1} + \widehat{u_2} + \ldots + \widehat{u_n} \\ \vdots \end{bmatrix} = \mathbf{0}.$$

Point being, the sum of the fitted OLS residuals is zero when the intercept is included.

3 OLS Assumptions

In order for the OLS estimator given in equation (3) to be valid, we need to make a number of assumptions.

(a) It must possible to write the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. In other words, the model is **linear in parameters**.

(b) The matrix $\mathbf{X}$ must have rank $k+1$. This is so $\mathbf{X}'\mathbf{X}$ is nonsingular, and thus invertible, and thus a unique $\widehat{\boldsymbol{\beta}}$ exists. This is called **no perfect collinearity**.

(c) Each conditional expected error $E[u_t|\mathbf{X}] = 0$. This can be written as $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$. This is essentially saying that we are assuming random samples from the population and is referred to as **zero conditional mean**.

(d) $\text{Var}(u_t|\mathbf{X}) = \sigma^2$ and $\text{Cov}(u_t, u_s|\mathbf{X}) = 0$ for all $t \neq s$. In matrix form,

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

The first says that each error term has the same variance given $\mathbf{X}$, and is called **homoskedasticity**. The second says that error terms are uncorrelated, referred to as **no serial correlation**. Under random sampling, there is no serial correlation.

Theorem 1. *Under assumptions (a)-(c), the OLS estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for β.*

Proof. Recall that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, we can write

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

This is allowed because $\mathbf{X}'\mathbf{X}$ is both nonsingular and a square matrix and thus is invertible. Now take the conditional expectation on $\mathbf{X}$ to get

$$\begin{aligned}E\left[\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}\right] &= E\left[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\middle|\mathbf{X}\right]\\ &= E\left[\boldsymbol{\beta}\right] + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left[\mathbf{u}\middle|\mathbf{X}\right]\\ &= E\left[\boldsymbol{\beta}\right] + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{0}\\ &= E\left[\boldsymbol{\beta}\right]\\ &= \boldsymbol{\beta}.\end{aligned}$$

The law of iterated expectations then gives

$$E[\widehat{\boldsymbol{\beta}}] = E\left[E[\widehat{\boldsymbol{\beta}}|\mathbf{X}]\right] = E[\boldsymbol{\beta}] = \boldsymbol{\beta}.$$

Or you could simply note that $E[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$ has nothing to do with $\mathbf{X}$ so

$$E[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}. \qquad \square$$