Econometrics

Fundamental concepts with matrix notation

2019DMB02

CLRM assumptions and properties

NULL SPACE

0.1 Classical Linear Regression Model

In this model, typically the variable in question (dependent variable) is related in some manner to other variables called regressors or **explanatory variables**. Now consider that we observe a sample of n values for all the variables. Let y_i be the i^{th} observation of the dependent variable and the corresponding i^{th} observations of the K regressors be denoted by the vector:

$$(x_{i1}, x_{i2}, \cdots, x_{iK}) \tag{1}$$

This collection of n observations is what essentially constitutes our **sample** or **data**. Now we will go through the assumptions of the CLRM step by step:

0.1.1 Linearity Assumption

Relationship between the dependent variable and regressors is assumed to be linear:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \epsilon_i \tag{2}$$

Note that here β values are the unknown parameters that are to be estimated and ϵ represents the unobserved error term. This equation is what is known as the **regression function** and the β values are known as the **regression coefficients**. A key point to note is the interpretation : β_2 is the change in the dependent variable when the second regressor increases by one unit while all other regressors are held constant. This is essentially the marginal effect of one regressor over the dependent variable and can be expressed as the partial derivative:

$$\frac{\partial y_i}{\partial x_{i2}} = \beta_2 \tag{3}$$

The error term is the part of the dependent variable that is left **unexplained** by the regressors.

0.1.2 Matrix notation

We are mostly dealing with many observations and hence many equations, so it is useful to introduce matrix notation. The K dimensional column vectors $\boldsymbol{x_i}$ and $\boldsymbol{\beta}$ are defined as:

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \tag{4}$$

Recall that vector inner products or dot products over those two vectors can be defined as:

$$\boldsymbol{x_i'\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \tag{5}$$

With this, we can write our original regression function as:

$$y_i = \boldsymbol{x_i'\beta} + \epsilon_i \tag{6}$$

Now putting all n observations together, we can define the following vectors and matrices, while noting that dimensions of these are written in parantheses:

$$\underset{(n\times1)}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \underset{(n\times1)}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \underset{(n\times K)}{\boldsymbol{X}} \begin{bmatrix} \boldsymbol{x_1'} \\ \boldsymbol{x_2'} \\ \vdots \\ \boldsymbol{x_n'} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix} \tag{7}$$

Traditionally the number of rows in these matrices correspond to the number of observations. The $\boldsymbol{X}$ matrix is called the **data matrix**. Again we can write the regression equation along with associated matrix and vector dimensions below:

$$\underset{n\times1}{\boldsymbol{y}} = \underset{n\times K}{\boldsymbol{X}} \underset{K\times1}{\boldsymbol{\beta}} + \underset{n\times1}{\epsilon} \tag{8}$$

0.1.3 Strict exogeneity assumption

This assumption states the following:

$$E(\epsilon_i|\boldsymbol{X}) = 0 \tag{9}$$

Note that the expectation is conditional on the regressors for all observations. This can also be written as follows :

$$E(\epsilon_i|\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}) = 0 \tag{10}$$

We can go a bit indepth with this argument, for a given observation i we can write the joint distribution of the $nK + 1$ random variables as $f(\epsilon_i, \boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ and the conditional distribution as $f(\epsilon_i|\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$. Therefore we say that the conditional mean of this distribution is 0 as is given by equation 10 above. Further, an important implication of this is that the **unconditional mean** of the error term is 0. From the **law of total probability** we can say that:

$$E[E(\epsilon_i|\boldsymbol{X})] = E(\epsilon_i) = 0 \tag{11}$$

We can state another important implication of this property that if the combined moment of two random variables is 0 $(E(x\epsilon) = 0)$ under strict exogeneity the regressors are **orthogonal** to the error term vector for all observations:

$$E(x_{jk}\epsilon_i) = 0,\ (i, j = 1, \cdots, n|k = 1, \cdots, K) \tag{12}$$

$$E(\boldsymbol{x_j} \cdot \epsilon_i) = \begin{bmatrix} E(x_{j1}\epsilon_i) \\ E(x_{j2}\epsilon_i) \\ \vdots \\ E(x_{jK}\epsilon_i) \end{bmatrix} = \underset{K\times 1}{\mathbf{0}} \tag{13}$$

These orthogonality conditions further tell us that the **covariance** between the error terms and regressors is 0:

$$Cov(\epsilon_i, x_j) = E(x_{jk}\epsilon) - E(x_{jk})E(\epsilon_i) = 0 \tag{14}$$

0.1.4 No Multicollinearity assumption

This basically means that the rank of the $(n \times K)$ data matrix $\boldsymbol{X}$ is K. We can say that the regressors are not correlated with each other. In terms of rank this would imply that the matrix has **full column rank** which inturn means that the matrix has K linearly independent columns, which essentially means that the regressors are not correlated.

0.1.5 Homoskedasticity assumption

This is the assumption of constant variance of error terms and given by:

$$E(\epsilon_i^2|\boldsymbol{X}) = \sigma^2 \tag{15}$$

To see this more clearly, we define the variance as follows :

$$Var(\epsilon_i|\boldsymbol{X}) = E(\epsilon_i^2|\boldsymbol{X}) - E(\epsilon_i|\boldsymbol{X})^2 = E(\epsilon_i^2|\boldsymbol{X}) = constant \tag{16}$$

This argument can further extend to the assumption of **no serial correlation** between the error terms which can be stated as:

$$Cov(\epsilon_i, \epsilon_j|\boldsymbol{X}) = 0 \tag{17}$$

Noting that the (i, j) element of the $(n \times n)$ matrix $\boldsymbol{\epsilon\epsilon'}$ is basically $\epsilon_i\epsilon_j$, we can write the no serial correlation assumption as follows:

$$E(\boldsymbol{\epsilon\epsilon'}|\boldsymbol{X}) = Var(\boldsymbol{\epsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n \tag{18}$$

0.1.6 CLRM for Random Variables

Note that the sample $(\boldsymbol{y}, \boldsymbol{X})$ would be a **random sample** if $(y_i, \boldsymbol{x_j})$ were independently and identically distributed across all observations. Now since ϵ is essentially a function of $(y_i, \boldsymbol{x_i})$ and we know that $(y_i, \boldsymbol{x_i})$ is independent of $(y_j, \boldsymbol{x_j})$ then we can also say that $(\epsilon_i, \boldsymbol{x_i})$ is independent of $\boldsymbol{x_j}$. This condition of independence allows us to state the following:

$$E(\epsilon_i|\boldsymbol{X}) = E(\epsilon_i|\boldsymbol{x_i}) \tag{19}$$

$$E(\epsilon_i\epsilon_j|\boldsymbol{X}) = E(\epsilon_i\boldsymbol{x_i})E(\epsilon_j\boldsymbol{x_j}) \tag{20}$$

An important implication of this independence is that the joint distribution of $(\epsilon_i, \boldsymbol{x_i})$ actually does not depend on i at all. Therefore the **unconditional variance** of the error term is constant across all i. Therefore **unconditional homoskedasticity** holds.

0.1.7 Fixed regressors

In the previous equations we have been conditioning various variables on $\boldsymbol{X}$ as if it represents a set of random variables. However, our assumption in the CLRM states that the regressors are **fixed** or deterministic. This further gives us clarification that the conditional distribution of the error term $f(\epsilon|\boldsymbol{X})$ is the same of the unconditional distribution of the error term $f(\epsilon_i)$.

0.2 The Algebra of Least Squares

We will now describe the computational procedure for obtaining the OLS estimate $\boldsymbol{b}$ of the unknown regression coefficients $\boldsymbol{\beta}$. Now OLS primarily minimizes the **sum of squared residuals**. The residuals are typically given by:

$$y_i - \boldsymbol{x_i'b} \tag{21}$$

This is the **residual** for observation i. From this we can get the sum of squared residuals (SSR) as:

$$SSR(\boldsymbol{b}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'b})^2 = (\boldsymbol{y} - \boldsymbol{Xb})'(\boldsymbol{y} - \boldsymbol{Xb}) \tag{22}$$

This above expression is also known as the **residual sum of squares** (RSS). We can see that since the residual essentially depends on $\boldsymbol{b}$, it is a function of $\boldsymbol{b}$. The procedure for us then is to basically select $\boldsymbol{b}$ such that the SSR function is minimized.

$$\boldsymbol{b} = argmin_{\boldsymbol{b}}SSR(\boldsymbol{b}) \tag{23}$$

0.2.1 Normal Equations

The traditional way to solve the above stated minimization problem is to actually derive the first order conditions by setting the partial derivatives equal to zero. We start by writing the SSR function as follows:

$$SSR(\boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{24}$$

Opening up the transpose operator in the RHS we get:

$$= (\boldsymbol{y}' - \boldsymbol{b}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{25}$$

Note that we can write the above result since $(\boldsymbol{X}\boldsymbol{b})' = \boldsymbol{b}'\boldsymbol{X}'$. Further expanding the brackets we get:

$$SSR = \boldsymbol{y'y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \tag{26}$$

Now since the expression $\boldsymbol{b'X'y}$ is a scalar it is equal to its transpose which is $\boldsymbol{y'Xb}$ and hence we can simply add the two to get:

$$SSR = \boldsymbol{y'y} - 2\boldsymbol{y'Xb} + \boldsymbol{b'X'Xb} \tag{27}$$

Now taking $\boldsymbol{a} = \boldsymbol{X'y}$ and taking $\boldsymbol{A} = \boldsymbol{X'X}$. Then we will get:

$$SSR = \boldsymbol{y'y} - 2\boldsymbol{ab} + \boldsymbol{bAb} \tag{28}$$

Now this is the equation that we will differentiate to get the optimal coefficient estimate vector. Note that since the $\boldsymbol{y'y}$ term does not depend on $\boldsymbol{b}$ we can essentially ignore it in differentiation. Taking partial derivatives with resepct to b we would get:

$$\frac{\partial \boldsymbol{a'b}}{\partial b} = \boldsymbol{a},\ and\ ,\ \frac{\partial(\boldsymbol{b'Ab})}{\partial \boldsymbol{b}} = 2\boldsymbol{Ab} \tag{29}$$

Where A is a square symmetric matrix. Now combining these derivative components with the total derivative for the SSR we would get:

$$\frac{\partial SSR}{\partial \boldsymbol{b}} = -2\boldsymbol{a} + 2\boldsymbol{Ab} \tag{30}$$

Now we will basically set this to zero, following the **first order conditions** and after substituting the requisite values for $\boldsymbol{a}$ and $\boldsymbol{A}$ we would obtain K **Normal equations** represented in matrix form.

$$\underset{(K\times K)}{\boldsymbol{X'X}}\underset{(K\times 1)}{\boldsymbol{b}} = \boldsymbol{X'y} \tag{31}$$

Now we can get the vector of residuals calculated using these $\boldsymbol{b}$ estimates as follows:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb},\ i^{th}\ element,\ e_i = y_i - \boldsymbol{x_i b} \tag{32}$$

In equation 32 if we premultiply $\boldsymbol{X}$ we would get:

$$\boldsymbol{X'e} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{Xb}) = 0 \tag{33}$$

In terms of an inner product expansion this can also be written as :

$$\boldsymbol{X'e} = \sum_{i=1}^{n} \boldsymbol{x_i} e_i = 0 \tag{34}$$

0.2.2 The estimator

Therefore after going through that process, we ultimately obtain K simultaneous equations in K unknowns where the unkowns are $\boldsymbol{b}$. Further the symmetric data matrix $\boldsymbol{X'X}$ is actually **positive definite** thereby ensuring that it has a minima and is also nonsingular (which basically means that its inverse exists). Therefore we can compute the coefficient estimate vector by:

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{35}$$

0.3 Additional concepts

We note that the **fitted value** for an observation i can be given by $\hat{y}_i = \boldsymbol{x_i'b}$ and the vector pf residuals is given by $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{\hat{y}}$. As some additional concepts we can write the **projection matrix P** and **annihilator matrix M** as follows:

$$\underset{(n\times n)}{\boldsymbol{P}} = \boldsymbol{X(X'X)^{-1}X'} \tag{36}$$

$$\underset{(n\times n)}{\boldsymbol{M}} = \boldsymbol{I}_n - \boldsymbol{P} \tag{37}$$

Following are some handy properties regarding them:

$$\boldsymbol{PX} = \boldsymbol{X},\ \boldsymbol{MX} = 0 \tag{38}$$

We note that the OLS estimate of error term variance is given by:

$$s^2 = \frac{SSR}{n-K} = \frac{\boldsymbol{e'e}}{n-K} \tag{39}$$

The overall variability in the dependent variable which is given by its sum of squares measure and can be decomposed into:

$$\boldsymbol{y'y} = (\boldsymbol{\hat{y}} + \boldsymbol{e})'(\boldsymbol{\hat{y}} + \boldsymbol{e}) \tag{40}$$

$$= \boldsymbol{\hat{y}'\hat{y}} + 2\boldsymbol{\hat{y}'e} + \boldsymbol{e'e} \tag{41}$$

$$= \boldsymbol{\hat{y}'\hat{y}} + 2\boldsymbol{b'X'e} + \boldsymbol{e'e} \tag{42}$$

$$= \boldsymbol{\hat{y}'\hat{y}} + \boldsymbol{e'e} \tag{43}$$

We got this result using the property that $\boldsymbol{X'e} = 0$. Finally, the coefficient of determination of R^2 can be given by:

$$R^2 = 1 - \frac{\boldsymbol{e'e}}{\boldsymbol{y'y}} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{44}$$

0.3.1 Desirable properties of OLS estimates

Here are some of the desirable properties of the OLS estimates:

- Unbiasesness. $E(\boldsymbol{b}) = \boldsymbol{\beta}$.
- Variance. $Var(\boldsymbol{b}) = \sigma^2(\boldsymbol{X'X})^{-1}$.
- Gauss Markov Theorem. $Var(\boldsymbol{b}) \leq Var(\boldsymbol{c})$.

References

[1] Fumio Hayashi - Econometrics