

The Classical Linear Regression Model (without Geometry)

Byron Tsang

01/08/2008

N. ABOUT ME

I am a 5th year graduate student at the Department of Economics, and my main fields of research are macroeconomics and labor economics. I have written on the US nominal term structure, consumption and real interest rates, adolescence behavior, and household consumption and fertility. Being a research assistant for 3 years, I have never taught any undergraduate class.

In my spare time I indiscriminately read books about philosophy and history.

I. OLS ASSUMPTIONS

Let's begin with some "fundamental" questions: What is a model? What does a model do? What is a good model? What can we learn from a model? If you do not have a good answer to each of these questions, you can never appreciate empirical works in economics and use econometrics in your own research appropriately.

An econometric model, like any model, is based on assumptions. The following four assumptions give us the classical linear regression (CLR) model:

1) **Linearity**: The model specifies a linear relationship between the LHS variable y_i (a random variable) and the RHS variables $x_{i1}, x_{i2}, \dots, x_{iK}$ (a random vector), for $i = 1, 2, \dots, N$:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{iK}\beta_K + \varepsilon_i,$$

or in matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{X}_{N \times K} = [\mathbf{x}_1' \quad \mathbf{x}_2' \quad \dots \quad \mathbf{x}_N'].$$

Think for a second for the dimensions of the vectors/matrices. The β 's are unknown parameters that we would like to learn more about, and ε_i is the unobserved error/disturbance term. The first RHS variable is usually the constant one and β_1 is the intercept. The subscript i may stand for the i th individual or the i th period of time. To avoid confusion, I do not call the LHS variable the “dependent variable” and the RHS variable the “independent/explanatory variable”. The linearity assumption only says that the parameters enter the regression linearly, and the RHS variables $x_{i1}, x_{i2}, \dots, x_{iK}$ can take any form (e.g. you can have age and the square of age on the RHS).

There is nothing sacred about linearity: as it is simple and easy to manipulate, it is a good assumption to start with. Notice that, just like any assumption in a model, linearity gives you a simplified/approximated version of the real world.

2) **Full Rank/ No Multicollinearity**: The rank of the $N \times K$ matrix \mathbf{X} is K (i.e. it has full rank) with probability 1 (Remember that \mathbf{X} is a matrix of random variables) .

3) **Strict Exogeneity**: The expectation of the error term of any observation i conditional of the RHS variables for all observations does not depend on the RHS variables and is a constant value of zero:

$$E(\varepsilon_i | \mathbf{X}) = 0 \quad \text{for } i = 1, 2, \dots, N.$$

This is an important assumption and brings up a lot of difficulties for empirical work. What does it say? It says at least 1) using the law of iterated expectation, the unconditional mean for the error term is also zero:

$$E(\varepsilon_i) = E(E(\varepsilon_i | \mathbf{X})) = 0;$$

and 2) any of the RHS variable is **orthogonal** to the error term:

$$E(\mathbf{x}_{jk} \varepsilon_i) = E(E(\mathbf{x}_{jk} \varepsilon_i | \mathbf{x}_{jk})) = E(\mathbf{x}_{jk} E(\varepsilon_i | \mathbf{x}_{jk})) = 0.$$

Think for a second of how likely these criteria are met: The error term this period is orthogonal to the RHS variables for all periods and, equivalently, the RHS variables this period are orthogonal to the error terms for all periods. For most applications in

time-series macroeconomics strict exogeneity just cannot hold. Consider the simple Phillips curve:

$$\pi_t = \beta(u_t - \bar{u}) + \varepsilon_t$$

In words, inflation in period t depends on 1) the difference between unemployment rate in period t and the natural rate of unemployment and 2) an error term which includes all other factors that affect inflation in period t . Strict exogeneity requires that unemployment in period t be orthogonal to the error terms in all periods, and it clearly cannot be true. Think about why in a cross-section context the strict exogeneity assumption is more likely to hold.

Does it mean that we should abandon empirical studies in time-series macroeconomics? No, and I will tell you why in a week or two.

4) **Spherical Error Variance:** This assumption can be broken into two parts:

a) **Homoskedasticity:** $E(\varepsilon_i^2 | \mathbf{X}) = \sigma^2 > 0$ for $i = 1, 2, \dots, N$ (i.e. the conditional variance of the error term is not a function of the RHS variables);

b) **No Serial Correlation:** $E(\varepsilon_i \varepsilon_j | \mathbf{X}) = 0$ for $i, j = 1, 2, \dots, N$ and $i \neq j$ (i.e. the error terms are not correlated across observations).

Is it a strong assumption? Consider the wage equation:

$$\ln wage_i = \alpha + \beta S_i + \varepsilon_i$$

In words, for individual i , log wage is a function of the years of school that individual has. If the variance of the error term depends on schooling, then the error term is heteroskedastic and not homoskedastic. For example, for very educated people the error term can have a larger variance than people with less education (a Ph.D. in economics vs. a Ph.D. in ... philosophy), and $E(\varepsilon_i^2 | \mathbf{X}) \neq \sigma^2$ but some function of schooling.

You may have heard before that homoskedasticity means the variance of the error term does not vary among observations. That is wrong. The correct definition is simply that the variance of the error term does not depend on the RHS variables.

Can you think of an example in which the error term is serially correlated?

Extra Fun: Spherical Disturbance – What Is So Spherical About It?

Consider a multivariate normal distribution:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

If we have spherical covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$, we can get a nicer form:

$$f(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \sigma^{-2} \mathbf{I} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Let's say we evaluate the function at some point:

$$f(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \sigma^{-2} \mathbf{I} (\mathbf{x} - \boldsymbol{\mu})\right) = c$$

Solving for a while:

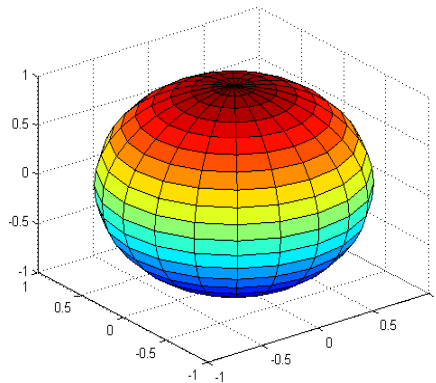
$$-2 \ln\left(c(2\pi\sigma^2)^{n/2}\right) \sigma^2 = (\mathbf{x} - \boldsymbol{\mu})' (\mathbf{x} - \boldsymbol{\mu})$$

Expanding for RHS, and assume it is 3-dimensional:

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + (x_3 - \mu_3)^2 = -2 \ln\left(c(2\pi\sigma^2)^{3/2}\right) \sigma^2$$

Gee, isn't it a sphere???

Below is the case for $-2 \ln\left(c(2\pi\sigma^2)^{3/2}\right) \sigma^2 = 1$ and $\mu_1 = \mu_2 = \mu_3 = 0$.



II. THE OLS ESTIMATOR

We do not observe the parameter β and we have to estimate it based on the LHS \mathbf{y} and RHS \mathbf{X} that we observe. One way to derive an estimator for β is through minimizing the **sum of squared residuals (SSR)** (hence the name “least squares”):

$$SSR(\tilde{\beta}) \equiv \sum_{i=1}^N (y_i - \mathbf{x}_i' \tilde{\beta})^2 = (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})$$

The vector $\tilde{\beta}$ is a candidate for the estimator for β . Why do we choose $\tilde{\beta}$ by minimizing the SSR? Intuitively, we choose $\tilde{\beta}$ such that the residual (i.e. the mistake) is as small as possible. Of course, this is not the only criterion we have, and we will consider other objective function later in this class.

Terminology: **Predicted value:** $\tilde{y}_i = \mathbf{x}_i' \tilde{\beta}$ and **residual:** $\tilde{\varepsilon}_i = \tilde{y}_i - \mathbf{x}_i' \tilde{\beta}$. Residual $\tilde{\varepsilon}_i$ is observable (numbers that come from your sample) and error ε_i is not (something exists in your imagined world). Also notice two important properties: $\sum_{i=1}^N \mathbf{x}_i \tilde{\varepsilon}_i = \mathbf{0}$, and, if

a constant is included, $\sum_{i=1}^N \tilde{\varepsilon}_i = 0$.

To derive the OLS estimator, we first rewrite the SSR:

$$\begin{aligned} SSR(\tilde{\beta}) &= (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\ &= (\mathbf{y}' - \tilde{\beta}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \tilde{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\beta} + \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\beta} + \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta} \end{aligned}$$

Now we differentiate the SSR with respect to $\tilde{\beta}$:

$$\frac{\partial SSR(\tilde{\beta})}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

Set it to zero and denote the solution from the first order condition as $\hat{\beta}$, we have the OLS estimator:

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

Question: Have we used any of the four assumptions to derive the OLS estimator?

Notice that an **estimator** is random, and it is based on the idea that the data generating process (DGP, the stochastic process that generated our sample (\mathbf{y}, \mathbf{X})) keeps giving us samples; we can talk about the estimator's statistical properties. An **estimate** is not random, and it is based on a particular sample we have; it is some number(s), and there are no statistical properties to speak of. Or, an estimator is a rule that takes a sample as input, and give an estimate as output.

Next time we will talk about the geometry of OLS. After that we will study the finite-sample (or small-sample) properties of the OLS estimator, and that is when we will make use of the four assumptions.