USC, Fall 2016, Economics 513

Lecture 4: The Classical Linear Regression Model II

The CLR model

The Classical Linear Regression model is a set of assumptions under which the OLS estimator $\hat{\beta}$ is optimal.

Re-label the variables $x_1, \ldots, x_{K-1}$ as $x_2, \ldots, x_K$ and introduce the variable

$$x_1 \equiv 1$$

i.e. $x_{i1} = 1, i = 1, \ldots, n$. Also re-label $\beta_0, \ldots, \beta_{K-1}$ to $\beta_1, \ldots, \beta_K$.

In the new notation for $i = 1, \ldots, n$

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \qquad (2)$$

This is the multiple linear regression model. The relation (2) is linear in $x_1, \ldots, x_K$ and also linear in $\beta_1, \ldots, \beta_K$. The latter is essential.

Assumption 1 in the new notation:

$$\mathrm{E}(\varepsilon_i | x_{i1}, \ldots, x_{iK}) = 0, i = 1, \ldots, n \qquad (3)$$

Note that this implies that $\mathrm{E}(\varepsilon_i) = 0$ by the Law of Iterated Expectations.

By Assumption 1 also for $k = 1, \ldots, K$ by the Law of Iterated Expectations

$$\mathrm{E}(\varepsilon_i | x_{ik}) = \mathrm{E}\left[\mathrm{E}(\varepsilon_i | x_{i1}, \ldots, x_{iK}) | x_{i1}, \ldots, x_{i,k-1}, x_{i,k+1}, \ldots, x_{iK}\right] = 0$$

so that

$$\mathrm{E}(x_{ik}\varepsilon_i) = \mathrm{E}[\mathrm{E}(x_{ik}\varepsilon_i | x_{ik})] = \mathrm{E}[x_{ik}\mathrm{E}(\varepsilon_i | x_{ik})] = 0$$

i.e. ε_i is uncorrelated with the independent variables.

In matrix notation (2) and (3) are

$$y = X\beta + \varepsilon$$

and

$$\mathrm{E}(\varepsilon_i | x_i) = 0, i = 1, \ldots, n$$

with

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \qquad (4)$$

Note that x_i is a $K \times 1$ vector that has the observations on the independent variables for i, i.e. the i-th row of X written as a column vector.

The CLR model makes assumptions mainly on the conditional distribution of ε given the independent variables $x_1, \ldots, x_K$.

Some assumptions are essential, some are convenient initial assumptions and can/will be relaxed.

The CLR model is appropriate in many practical situations and is the starting point for the use of mathematical statistical inference in the linear regression model.

The CLR model assumptions

The observations satisfy

$$y = X\beta + \varepsilon$$

Assumption 1: Fundamental assumption

$$\mathrm{E}(\varepsilon|X) = 0$$

Assumption 2: Spherical disturbances

$$\mathrm{E}(\varepsilon\varepsilon'|X) = \sigma^2 I$$

Assumption 3: Full rank

$$\mathrm{rank}(X) = K$$

Discussion of the assumptions

Assumption 1 is shorthand for

$$\mathrm{E}(\varepsilon_i|X) = 0, i = 1, \ldots, n$$

Hence this is equivalent to

$$\mathrm{E}(\varepsilon_i|x_1, \ldots, x_n) = 0, i = 1, \ldots, n$$

Compare this with

$$\mathrm{E}(\varepsilon_i|x_i) = 0, i = 1, \ldots, n$$

By the law of iterated expectations, the current assumption implies the latter. The current assumption states that not only x_i but also $x_j, j \neq i$ is not related to ε_i. This is not stronger than the previous assumption if ε_i, x_i are independent for $i = 1, \ldots, n$ as in a random sample from a population. If these are not independent, as e.g. in time-series data, then this additional assumption may be too strong. As we will see for independent observations OLS will estimate partial effects even if $\mathrm{E}(\varepsilon_i|x_i) = 0, i = 1, \ldots, n$ or if $\mathrm{E}(x_{ik}\varepsilon_i) = 0$ for all k.

Assumption 1 is satisfied if the distribution of ε_i does not depend on X. In that case we can treat X as a matrix of known constants. Therefore instead of Assumption 1 one sometimes sees

Assumption 1': X is a matrix of known constants determined independently of ε.

Note: If we can choose X we should do so independently of ε.

Next, we consider assumption 2

Note

$$\varepsilon\varepsilon' = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \varepsilon_n\varepsilon_1 & \cdots & \cdots & \varepsilon_n^2 \end{bmatrix}$$

so that

$$\mathrm{E}(\varepsilon\varepsilon'|X) = \begin{bmatrix} E(\varepsilon_1^2|X) & E(\varepsilon_1\varepsilon_2|X) & \cdots & E(\varepsilon_1\varepsilon_n|X) \\ E(\varepsilon_2\varepsilon_1|X) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1|X) & \cdots & \cdots & E(\varepsilon_n^2|X) \end{bmatrix}$$

In other words, $\mathrm{E}(\varepsilon\varepsilon'|X)$ is the (conditional on X) variance matrix of the vector ε.

Hence Assumption 2 implies that for the conditional variance

$$\mathrm{E}(\varepsilon_i^2|X) = \sigma^2, i = 1, \ldots, n$$

This is called homoskedasticity.

Assumption 2 also implies that for the conditional covariance

$$\mathrm{E}(\varepsilon_i\varepsilon_j|X) = 0$$

Hence, given X the random errors are uncorrelated.

Example of failure of homoskedasticity: random coefficients

Random coefficient model

$$y = \beta_0 + (\beta_1 + u)x + \varepsilon$$

$$= \beta_0 + \beta_1 x + \varepsilon + ux$$

where the (random) variable u captures population variation in the coefficient on x.

For the composite disturbance, if $\mathrm{E}(u|x) = 0$

$$\mathrm{E}(\varepsilon + ux|x) = 0$$

but

$$\mathrm{Var}(\varepsilon + ux|x) = \sigma^2 + \sigma_{\varepsilon u} x + \sigma_u^2 x^2$$

with $\sigma_{\varepsilon u} = \mathrm{E}(\varepsilon u)$, $\sigma_u^2 = \mathrm{E}(u^2)$. Hence the composite error is heteroskedastic.

Failure of uncorrelated disturbances: serial correlation

Serial correlation of order 1 in disturbances

$$\varepsilon_i = \rho \varepsilon_{i-1} + u_i$$

Applies in time-series data.

In cross-sectional data we can consider a sample of students who are in different schools. The dependent variable is the test score and the independent variables are student background variables, e.g. parents education and income, and school input variables, e.g. teacher qualification. It is likely that there are omitted school inputs and this makes the ε_i and ε_j correlated if i and j are at the same school, but not if they are at different schools.

Failure of Assumption 3

If $\mathrm{rank}(X) = K$, then $Xa = 0$ if and only if $a = 0$ with a a $K \times 1$ vector, i.e. there is no (linear) relation between the K variables.

We conclude that $X'X$ is a positive-definite matrix and hence its inverse exists.

Example of failure of Assumption 3: wage equation

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, i = 1, \ldots, n$$

with

- x_2 = schooling (years)
- x_3 = age
- x_4 = potential experience (age-years in school-6)

Hence

$$x_4 = x_3 - x_2 - 6$$

Note that for the $n \times 4$ matrix X we have $Xa = 0$ with a the 4×1 vector

$$a = c \begin{pmatrix} 6 \\ 1 \\ -1 \\ 1 \end{pmatrix}$$

for any scalar constant c, because $c(6x_1 + x_2 - x_3 + x_4) = 0$. Therefore

$$X\beta = X(\beta + a)$$

Thus β and $\beta + a$ give the same right-hand side so that if we define

$$\tilde{\beta}_1 = \beta_1 + 6c, \tilde{\beta}_2 = \beta_2 + c, \tilde{\beta}_3 = \beta_3 - c, \tilde{\beta}_4 = \beta_4 + c$$

then also

$$y_i = \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \tilde{\beta}_3 x_{i3} + \tilde{\beta}_4 x_{i4} + \varepsilon_i, i = 1, \ldots, n$$

We cannot distinguish between β and $\tilde{\beta}$. For all c these parameters are observationally equivalent, because they both satisfy the CLR model.

If there are parameter vectors $\beta \neq \tilde{\beta}$ that observationally equivalent, then we say that β is not identified. Without identification we cannot do the usual statistical inference on β.

Problem is also clear if we substitute for x_4

$$y_i = (\beta_1 - 6\beta_4)x_{i1} + (\beta_2 - \beta_4)x_{i2} + (\beta_3 + \beta_4)x_{i3} + \varepsilon_i$$

If Assumption 3 almost fails, i.e. the independent variables satisfy almost a linear relation, then we say that there is multicollinearity among the columns of X. I will not discuss this issue since it is mainly a data issue.

By assumptions 1 and 2, $\mathrm{E}(\varepsilon|X) = 0$, $\mathrm{Var}(\varepsilon|X) = \sigma^2 I$. Sometimes it is assumed that

Assumption 4

$$\varepsilon|X \sim N(0, \sigma^2 I)$$

Why is the normal distribution a natural choice?

This assumption is useful if we want exact results on the distribution of the OLS estimator $\hat{\beta}$.

Appendix: Law of Iterated Expectations

Let y be earnings and x_1 be gender (1 is female) x_2 an indicator of post high school education (1 if post high school education). Then

$$E(y|x_1, x_2)$$

is the average earnings of individuals with gender x_1 and education level x_2.

The table gives an example. The numbers between parentheses are the cell means $E(y|x_1, x_2)$

	0 (M)	1 (F)
0 (L)	.4 (5)	.3 (4)
1 (H)	.2 (7)	.1 (6)
	.6	.4

The average earnings in this population is 5.2 and this can also be computed as the average over x_1, x_2 of the cell means

$$5.2 = E(y) = E(y|0,0)\Pr(x_1 = 0, x_2 = 0) + E(y|1,0)\Pr(x_1 = 1, x_2 = 0) +$$

$$E(y|0,1)\Pr(x_1 = 0, x_2 = 1) + E(y|1,1)\Pr(x_1 = 1, x_2 = 1) = E\left[E(y|x_1, x_2)\right]$$

In general

$$E(y) = E\left[E(y|x_1, x_2)\right]$$

Also the average earnings for males is computed by averaging the cell means for the males over the distribution of education level for males, i.e. 2/3 lower and 1/3 higher educated. The fraction lower educated among the males is computed as

$$\Pr(x_2 = 0|x_1 = 0) = \frac{\Pr(x_1 = 0, x_2 = 0)}{\Pr(x_1 = 0)} = \frac{.4}{.6} = 2/3$$

Hence

$$E(y|x_1 = 0) = E(y|0,0)\Pr(x_2 = 0|x_1 = 0) + E(y|0,1)\Pr(x_2 = 1|x_1 = 0) =$$

$$5 * 2/3 + 7 * 1/3 = 5.66$$

In general

$$E(y|x_1) = E\left[E(y|x_1, x_2)|x_1\right]$$

The general formulas are instances of the Law of Iterated Expectations, one of the most useful formulas in statistics.