Classical Linear Regression Model

In this handout we will revisit the Classical Linear Regression Model (CLRM) (see Wooldridge 2010, chaps. 1–2). The goal of this week's lecture is to:

1. understand the model specification;
2. it's underlying assumptions;
3. and the appropriate interpretation (correlation vs causation).

Model Specification

The linear population regression model is given by,

$$\begin{aligned} Y_i =& X_i'\beta + \varepsilon_i \\ =& \beta_1 \mathbf{1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + \varepsilon_i \end{aligned}$$

for $i = 1, 2, ..., n$. Where,

- i: unit of observation; e.g. individual, firm, union, political party, etc.
- $Y_i \in \mathbb{R}$: scalar random variable.
- $X_i \in \mathbb{R}^k$: k-dimensional (column[1]) vector of regressors.[2]
- β: k-dimensional, non-random vector of unknown population parameters.
- ε_i: *unobserved*, random error term.[3]

[1] My notation assumes that X_i is a column vector, which makes $X_i'\beta$ a scalar. Wooldridge (2010) uses the notation $X_i\beta$, implying that X_i is a row vector. This is a matter of preference.

[2] You might also refer to the vector of regressors as covariates or explanatory variables. Some texts will use the term independent variables, but this name implies a specific relationship between Y and X that need not hold.

[3] This is **NOT** the residual.

The linear population regression equation is **linear in parameters**. This is an important assumption that does NOT restrict the model from being non-linear in regressors. For example, the equation

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i2}^2 + \varepsilon_i$$

non-linear in X_{i2}, but still linear in parameters. In contrast, the equation

$$Y_i = \beta_1 + \beta_2 X_{i2} + (\beta_2 \beta_3) X_{i3} + \varepsilon_i$$

is non-linear in parameters.

Intercept

The constant (intercept) in the equation serves an important purpose. While there is no *a priori* reason for the model to have a constant term, it does ensure that the error term is mean zero.

Proof. Suppose $E[\varepsilon_i] = \gamma$.

We can then define a new error term, $\upsilon_i = \varepsilon_i - \gamma$, such $E[\upsilon_i] = \gamma$. The population regression model can be rewritten as,

$$\begin{aligned} Y_i =& X_i'\beta + \upsilon_i + \gamma \\ =& \underbrace{(\beta_1 + \gamma)}_{\tilde{\beta}_1} \mathbf{1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \upsilon_i \end{aligned}$$

The model has a new intercept $\tilde{\beta}_1 = \beta_1 + \gamma$, but the other parameters remain unchanged. □

Matrix notation

For a sample of n observations, we can stack the unit-level linear regression equation into a vector,

$$Y = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_n'\beta \end{bmatrix}}_{n\times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & & \\ \vdots & & \ddots & \\ X_{n1} & & & X_{nk} \end{bmatrix}}_{n\times k} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{k\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = X\beta + \varepsilon$$

Notice, in matrix notation, you lose the transpose from $X_i'\beta$. Apart from the absence of the i subscript, this is a useful way of knowing the dimension of the equation (in my notes). You MUST always write $X\beta$ and not βX. For the scalar case, $X_i'\beta = \beta' X_i$, but for the vector case βX is not defined since β is $k \times 1$ and X is $n \times k$.

CLRM Assumptions

Assumption CLRM 1. Population regression equation is linear in parameters:

$$Y = X\beta + \varepsilon$$

Assumption CLRM 2. Conditional mean independence of the error term:

$$E[\varepsilon|X] = 0$$

Together, CLRM 1. and CLRM 2. imply that

$$E[Y|X] = X\beta$$

This means that the Conditional Expectation Function is known and linear in parameters.

Conditional mean independence implies - by the Law of Iterated Expectations - mean independence of the error term,

$$E[\varepsilon|X] = 0 \Rightarrow E[E[\varepsilon|X]] = E[\varepsilon] = 0$$

and uncorrelatedness,

$$E[\varepsilon|X] = 0 \Rightarrow E[\varepsilon X] = 0$$

Note, neither of the above statements hold the other way around. Mean independence does not imply conditional mean independence and uncorrelatedness (zero correlation/covariance) does not imply conditional mean independence.

Uncorrelatedness rules out linear relationships between the regressors and error term while conditional mean independence rules out non-linear relationships too.

In general, distributional independence implies mean independence which then implies uncorrelatedness.

In the case joint-normally distributed random variables, uncorrelatedness implies independence. That is, if

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

Then $\sigma_{12} = \sigma_{21} = 0 \iff f_1 * f_2 = f_{12}$.

We will later show that uncorrelatedness is sufficient for consistency of the Ordinary Least Squares estimater, while conditional mean independence is required for unbiasedness of OLS.

Assumption CLRM 3. Homoskedasticity: $Var(\varepsilon|X) = E[\varepsilon\varepsilon'|X] = \sigma^2 I_n$

CLRM 3. states that the variance of the error term is independent of X and constant across units. The diagonal nature of the covariance matrix also implies that the error terms are uncorrelated across units in the data. Note, this does not imply independence of the error terms across units.

Models with heteroskedasticity relax the assumption of constant variance, allowing for a richer variance-covariance matrix that typically depends on X.

This assumption is unlikely to hold in time-series models where units represent repeated observations across time. Such violations are referred to as serial correlation or autocorrelation.

Even in cross-sectional data settings, you can have non-zero correlations across units in the data. A common instance of this is the case of clustering. Clustering can occur when units experience common/correlated 'shocks'; for example, the data contains groups of students from the same classroom who have a the same teacher. This can also be the result of clustered sampling, a common practice in multi-stage survey design.

Assumption CLRM 4. Full rank: $rank(X) = k$[4]

CLRM 4. is some time referred to as the absence of perfect (or exact) collinearity. Do not confuse this with multicollinearity. Multicollinearity occurs when regressors are highly (linearly) correlated with one another, yielding imprecise estimates.

Assumption CLRM 5. Normality of the error term: $U|X \sim N(0, \sigma^2 I_n)$

Assumption CLRM 6. Observations $\{(Y_i, X_i) : i = 1, ..., n\}$ are independently and identically distributed (iid).

CLRM 5 & 6 are not part of the Classical assumptions, but do simplify the problem of inference. Note, CLRM 5 implies independence across error terms, not implied by CLRM 4.

Non-random X

There is an alternative version of the CLRM in which X is a non-random, matrix of regressors/predictors. With X fixed, the error term is the only only random variable in the model. CLRM assumptions 1 and 4 remain the same, while CLRM 2, 3, 5, and 6 become:

Assumption CLRM 2$^\mathbf{a}$. Mean independence of the error term:

$$E[\varepsilon] = 0$$

Assumption CLRM 3$^\mathbf{a}$. Homoskedasticity: $Var(\varepsilon) = \sigma^2 I_n$

Assumption CLRM 5$^\mathbf{a}$. Normality of the error term: $U \sim N(0, \sigma^2 I_n)$

[4] See extra material on Linear Algebra. Since X is a random variable we should add to the assumption: $rank(X) = k$ *almost surely* (abbreviated a.s.). This means that the set of events in which X is not full rank occur with probability 0. The reason for this addition is that such a set of events may not be empty.

Assumption CLRM 6$^{\mathbf{a}}$. Observations $\{\varepsilon_i : i = 1, ..., n\}$ are independently and identically distributed (iid).

Identification

CLRM 1,2 and 4. are the *identifying* assumptions of the model. These assumptions allow us to write the parameter of interest as a set of 'observable' moments in the data. We can demonstrate this as follows.

Proof. Start with CLRM 2.

$$E[\varepsilon_i|X_i] = 0$$

Pre-multiply by the vector X_i,

$$X_i E[\varepsilon_i|X_i] = 0$$

Since the expectation is conditional on X_i, we can bring X_i inside the expectation function,

$$E[X_i\varepsilon_i|X_i] = 0$$

This conditional expectation is a random-function of X_i. If we take the expectation of this function w.r.t. X, we achieve the aforementioned result that conditional mean independence implies zero covariance,

$$E\left[E[X_i\varepsilon_i|X_i]\right] = E[X_i\varepsilon_i] = 0$$

Now substitute in for ε_i using the linear regression model from CLRM 1 and separate the resulting two terms,

$$\begin{aligned} E[X_i(Y_i - X_i'\beta)] &= 0 \\ \Rightarrow E[X_iX_i']\beta &= E[X_iY_i] \end{aligned}$$

Since β is a non-random vector, we can remove it from the expectation function.

Now we have a system of linear equations (of the form $Av = b$) with a unique solution if and only if the matrix $E[X_iX_i']$ is invertible. For the inverse of $E[X_iX_i']$ to exist, we require CLRM 3.

$$\beta = E[X_iX_i']^{-1}E[X_iY_i]$$

□

We cannot compute β because we do not know the joint distribution of (Y_i, X_i) needed to solve for the variance-covariance matrices. However, β is (point) identified because both Y and X are observed in the data and the parameters are "pinned down" by a unique set of 'observable' moments in the data.

β is not identified if the above system of linear equations does not have a unique solution. This will occur if two or more of the regressors are perfectly colinear.[5] β is also not be identified is the resulting expression for β includes 'objects' (moments, distribution/scale parameters) that are not 'observed' in the data. For example, if the error is not mean independent, the above expression will include a bias term that depends on $E[X_i'\varepsilon_i]$.

In this instance, the identification of β is scale dependent. That is, if we multiply Y_i by a scalar, β is multiplied by the same scalar. Consider cases where a researcher is modelling standardized test-scores.

Interpretation

In this linear regression model each slope coefficient has a partial derivative interpretation,

$$\beta_j = \frac{\partial E[Y_i|X_i]}{\partial X_{ij}}$$

or, as a vector,

$$\beta = \frac{\partial E[Y_i|X_i]}{\partial X_i} = \begin{bmatrix} \frac{\partial E[Y_i|X_i]}{\partial X_{i1}} \\ \vdots \\ \frac{\partial E[Y_i|X_i]}{\partial X_{ik}} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Note, the derivative is expressed in terms of changes in the *expected* value of Y_i (conditional on X_i), not Y_i itself. This is because Y_i is a random variable, but under CLRM 1 & 2

$$E[Y_i|X_i] = X_i'\beta$$

For a given value of X_i, the above expression is non-random.

As β_j is a partial derivative, its interpretation is one that "holds fixed" the value of other regressors (i.e. *ceteris paribus*). Because of this, many researchers apply the experimental language of control variables when interpretting regression coefficients.

Consider this simplified version of the linear regression model popularized by Mincer.

[5] There are many such failures of (parametric) identification in models that include dummy variables (or fixed effects). Earlier we saw that the intercept is not separately identified from the mean of the error term. Mean independence of the error term, $E[\varepsilon_i] = 0$, is required for us to separately 'identify' β_1.

$$WAGE_i = \gamma_1 + \gamma_2 EDU_i + \gamma_3 EXP_i + \upsilon_i$$

where,

- $WAGE_i$: individual wage (£'s)
- EDU_i: years of schooling/education
- EXP_i: years of experience

Assuming $E[\upsilon_i|S_i, EXP_i] = 0$, β_3 is the expected change in wages from an additional year of experience, holding fixed years of schooling. The model implies that if we were to consider two individuals with the same years of schooling, but a 1-year difference in work experience, then we would expect the more experienced worker to earn β_3 £'s more.

We would get the same interpretation from an experiment where we *control* individual schooling, but (randomly) vary years of experience by one year across units in the population.

Remember, for the above linear regression model, this interpretation is based on the assumption that the conditional expectation function is correctly specified. If no, then this interpretation is incorrect. Moreover, there are other ways to think about "controlling" for covariates that we will address towards the end of this module.

Semi-elasticities and elasticities

The original Mincer equation has the outcome as the log of wages,

$$\ln(WAGE_i) = \gamma_1 + \gamma_2 EDU_i + \gamma_3 EXP_i + \gamma_4 EXP_i^2 + \upsilon_i$$

The interpretation of β_3 is now in terms of expected log-points of wages.

$$\beta_3 = \frac{\partial E[ln(WAGE_i)|EDU_i, EXP_i]}{\partial EXP_i}$$

This can be converted into a percentage change in (expected) wages,

$$\%\Delta E[WAGE_i|EDU_i, EXP_i] = (\exp^{\beta_3} - 1) \times 100$$

For values of $\beta_3 \in [-0.1, 0.1]$ this value is closely approximated by $\beta_3 \times 100$.

Next, consider a model where the regressor is in logs, while the outcome remains in levels. For example, a model of commuting cost as a function of distance to work,

$$COST_i = \gamma_1 + \gamma_2 \ln(DIST_i) + +\nu_i$$

Here the interpretation of β_2 is,

$$\beta_2 = \frac{\partial E[COST_i | \ln(DIST_i)]}{\partial \ln(DIST_i)}$$

We can convert this to $\%\Delta$ in $DIST_i$, using the fact that a 1% change in distance implies a change in log points of $\ln(1.01) \approxeq 0.01$. Thus, we can approximate the expected change in cost by $\beta_2/100$.

Finally, when both the outcome and regressor are logged, the coefficient as an elasticity interpretation. For example, in the taxation literature, it is common to see taxable income modeled as a function of the (marginal) tax rate,

$$\ln(INC_i) = \beta_1 + \beta_2 \ln(RATE_i) + \xi_i$$

Here, β_2 has an tax elasticity interpretation,

$$\beta_2 = \frac{\partial E[\ln(INC_i) | \ln(RATE_i)]}{\partial \ln(RATE_i)} = \frac{\%\Delta E[INC_i | RATE_i]}{\%\Delta RATE_i}$$

Bibliography

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT press.