EC203: Applied Econometrics
Classic Linear Regression Model: assumptions, failures and consequences

Dr. Tom Martin

University of Warwick

The CLRM

The population model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

The CLRM assumptions:

1. $E[\epsilon] = 0$
2. No perfect multicollinearity and all $X's$ must exhibit some variation.
3. $E[\epsilon | X_1, ..., X_k] = 0$
4. $Cov(\epsilon_i, \epsilon_j | X_1, ..., X_k) = 0$ for all $i \neq j$
5. $V(\epsilon | X_1, ..., X_k) = \sigma^2$
6. $\epsilon | X_1, ..., X_k \sim N(0, \sigma^2)$

The CLRM

The sample model:

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} + e_i$$

- Where $a, b_1, ...b_k$ are OLS estimators of $\alpha, \beta_1, ..., \beta_k$.
- If all CLRM assumptions hold then it is the case OLS estimators are BLUE:
  - Best (smallest variance among all linear unbiased estimators)
  - Linear
  - Unbiased ($E[b_j] = \beta_j$)
  - Estimator

Recall: the CLRM

Three important results:

1. If assumptions 1, 2 and 3 hold then OLS estimators are unbiased: $E[b_j] = \beta_j$, for all $j = 1, ..., k$. However, the only strong assumption one needs to think carefully about is assumption 3, the conditional independence assumption: $E[\epsilon|\mathbf{X}] = 0$.

Recall, the conditional independence assumption is equivalent to the selection effect we discussed in the introductory lecture. That is, if we are confident that $E[\epsilon|X] = 0$, this is the same as being confident that there is no selection effect, such that the observed effect equates to the causal effect. See the regression motivation lecture.

Recall: the CLRM

2. If assumptions 1, 2, 3, 4 and 5 hold then the OLS estimators are unbiased and the most efficient among all linear unbiased estimators: $Var(b_j) \leq Var(b_j^*)$, for all $j = 1, ..., k$.

3. If assumptions 1, 2, 3, 4, 5 and 6 hold then the OLS estimators are unbiased, most efficient, and have a normal distribution: $b_j \sim N(\beta_j, Var(b_j))$ for all $j = 1, ..., k$. (BLUE)

The CLRM

Three important questions:

1. What are the consequences for the OLS estimators if one or more of the CLRM assumptions fail?

2. How do we test if the CLRM assumptions hold?

3. What are the potential solutions if CLRM assumptions fail?

We spend the rest of the course answering these three questions.

The CLRM

Outline of problems and consequences for OLS estimators:

1. Non-normality:
   - The assumption $\epsilon | X_1, ..., X_k \sim N(0, \sigma^2)$ does not hold.
   - The error terms are not normally distributed.
   - Main consequence: t and F-statistics are invalidated.

2. Heteroskedasticity:
   - The assumption $V(\epsilon | X_1, ..., X_k) = \sigma^2$ does not hold.
   - The error terms are not constant. The errors are not homoskedastic, they are heteroskedastic.
   - Main consequence: t and F-statistics are invalidated.

The CLRM

3. Serial correlation in the errors:
   - The assumption $Cov(\epsilon_i, \epsilon_j | X_1, ..., X_k) = 0$ for all $i \neq j$ does not hold.
   - Non-zero correlation between the error terms.
   - Main consequence: t and F-statistics are invalidated.

4. Perfect multicollinearity:
   - A perfect linear relationship between X's in the model.
   - The the more practical problem is very strong multicollinearity.
   - Main consequence of strong multicollinearity: the variance of OLS estimators increases and, therefore, effects are more difficult to pick up.

The CLRM

5. Endogenous variables: Failure of the conditional independence assumption, $E[\epsilon|X_1, ...X_k] \neq 0$. Due to:

- Misspecified functional form (covered)
- Measurement error (covered)
- Omitted variables (covered)
- Reverse causality or simultaneity (covered)
- Sample selection issues (not covered)
- Main consequences: OLS estimators are biased (and t and F statistics are invalidated).

The CLRM

After considering problems 1-5 in more detail we focus on possible solutions for endogeneity:

- Experimental: randomised control trials with regression analysis.
- Quasi-experimental: instrumental variables.
- Non-experimental: panel data methods.
- Regression discontinuity designs
- Combinations of the above.