Lecture 8

Violation of the Classical Assumptions:

Multicollinearity

Ananya Iyengar

February 2026

St. Stephen's College

This corresponds to Chapter 8 of Studenmund (2017).

Starting this lecture, we will look at violations of the CLRM assumptions and their implications for interpreting model coefficients. We will begin with violations of multicollinearity.

1. **Perfect Multicollinearity** is a situation where an independent variable is a linear combination of others. The variation in one independent variable can be *completely* explained by the variation in another. For example, $X_1 = \gamma_1 + \gamma_2 X_2$.

2. With perfect multicollinearity, (1) $\hat{\beta}$ is indeterminate and (2) $se(\hat{\beta}) = \infty$. We have seen before in the simple regression case, perfect multicollinearity corresponds to $X = c$, a constant.

3. Perfect multicollinearity is relatively easy to identify and avoid. However, imperfect multicollinearity is a larger practical problem. Consider a non-exact but hihgly correlated relationship of the form $X_1 = \gamma_1 + \gamma_2 X_2 + u$, where u is a stochastic error term. This implies that there is some part of the variation in X_1 that X_2 does not explain.

4. Severe and imperfect multicollinearity has the following consequences:

   - $\hat{\beta}$**s remain unbiased**. We do not need the "no perfect multicollinearity" assumption to prove unbiasedness. The sampling distribution of $\hat{\beta}$ will be centered around the true population value β.

- **Variance of $\hat{\beta}$ goes up.** Intuitively, in the simple regression case, $\sum x_i^2$ is small, which is the denominator of the standard error term. OLS continues to remain BLUE.
- **The magnitude of the observed t-statistic falls.** Low t-scores imply that the individual coefficients of a regression with high multicollinearity may tend to not be statistically significant. This also means that the size of confidence intervals ($\hat{\beta} \pm t_{\alpha/2} se(\hat{\beta})$) will widen.
- **The regression is sensitive to changes in specifications**. Dropping a statistically insignificant variable may change the estimates and significance of other explanatory variables.
- **Overall fit of the regression is unaffected**. The overall explanatory power of the regression is not affected by high imperfect multicollinearity. We would expect the R^2 to remain approximately the same. Moreover, it is possible that the F-test rejects the null while individual t-statistics imply no statistical significance.

5. How does one identify multicollinearity?

   (a) Compute the simple correlation $\rho \in [-1, 1]$ between two variables if we suspect pairwise linear relationships between variables.

   (b) In a situation where there may be more than 2 linearly correlated variables, we compute the **Variance Inflation Factor (VIF)**.

   The procedure for computing the VIF is as follows. Suppose there are k regressors $X_1, ...X_i..., X_k$. The coefficients associated with each of these X_is are given by β_i. Then, to examine if a given β_i is affected due to high imperfect multicollinearity:

   i. Auxiliary Regression of X_i on all other X_{-i}s. For $i = 1$, this is $X_1 = \gamma_1 + \gamma_2 X_2 + ... + \gamma_k X_k + v$, where v is the stochastic error term.

   ii. Find the coefficient of determination R_i^2 for the i^{th} auxiliary regression.

   iii. $VIF = \frac{1}{1-R_i^2}$. Higher the R_i^2, higher the degree of multicollinearity.

   iv. As a rule of thumb, $VIF(\beta_i) > 5$ is used as a heuristic to diagnose high degrees of multicollinearity.

   The VIF is a sufficient test of multicollinearity but not necessary. Specifically, (1) the rule of thumb is arbitrary and (2) two variables may have high pairwise correlation but the VIF may be lower than arbitrary thresholds. Here, theory must dictate remedies.

6. Ways to remedy multicollinearity:

   (a) Do nothing! Specificity of the case matters; it is possible that coefficients are significant, one risks specification errors (omitted variable bias) and the specification may have predictive capacity.

   (b) When multicollinearity is such that individual coefficients are insignificant and the F-test suggests goodness of fit, *and* we are certain that the multicollinear variables are not imperative to the regression, we can drop the redundant variable. In fact, this can often mean remedying a specification error.

   (c) Increasing the sample size reduces scope for error and also counters the small t-value. However, there are practical limitations to this remedy, especially with time series data.

7. *Multicollinearity is often left unadjusted.* Omitted variables can lead to *bias* (expected sign $\times$ correlation) which can lead to t-values misbehaving.