Section 4 Basics of Multiple Regression

Nearly all econometric applications require more than one explanatory variable. Thus, we need to extend the case of bivariate (simple) regression to multiple regression, involving multiple regressors.

Omitted variable bias

- What happens if we leave out a relevant regressor?
- Suppose that the true model is $y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e_i$, so that the true effect of x_2 on y is β_2. (Note that x_1 is the constant 1 multiplied by β_1, the intercept term.) Instead of fitting this model, we instead fit a simple regression model $y_i = \gamma_1 + \gamma_2 x_2 + v_i$. Will our estimate $\hat{\gamma}_2$ be a good (unbiased, at least) estimate of the true effect β_2? No, in general it will be biased:
    - The error term in the estimated simple regression model will include the effect of x_3 in addition to the true error e: $v_i = \beta_3 x_3 + e_i$
    - Applying the simple-regression expected value formula from earlier,

$$\hat{\gamma}_2 = \beta_2 + \left[\frac{\sum_{i=1}^{N} \left(x_{i,2} - \bar{x}_2 \right) v_i}{\sum_{i=1}^{N} \left(x_{i,2} - \bar{x}_2 \right)^2} \right]$$

$$= \beta_2 + \left[\frac{\sum_{i=1}^{N} \left(x_{i,2} - \bar{x}_2 \right) \left(\beta_3 x_{i,3} + e_i \right)}{\sum_{i=1}^{N} \left(x_{i,2} - \bar{x}_2 \right)^2} \right].$$

      Assuming the standard OLS assumptions are correct for the two-variable model, the e term in the numerator has expectation of zero. The cross-product term in the numerator has probability limit $\beta_3 \text{cov}(x_2, x_3)$. The denominator has plim of the variance of x_2. Thus $\text{plim}\left(\hat{\gamma}_2\right) = \beta_2 + \beta_3 \frac{\text{cov}\left(x_2, x_3\right)}{\text{var}\left(x_2\right)}$. Note that the ratio in this expression is (the expectation of) the regression slope coefficient we would get by regressing x_3 on x_2.
    - Thus, the omission of x_3 from the regression biases the OLS estimate of the coefficient on x_2 unless one of two conditions is true:
        - x_3 doesn't really belong in the regression ($\beta_3 = 0$), **or**
        - x_2 and x_3 are uncorrelated (cov = 0).
    - This is known as **omitted-variable bias**.
        - The bias has the sign of the product of β_3 and $\text{cov}(x_2, x_3)$.

        - Effect of education on wage, with ability omitted
            - Education and ability are probably positively correlated (cov > 0)
            - True effect of education on wage $\beta_2 > 0$
            - True effect of ability on wage is positive $\beta_3 > 0$
            - Bias in $\hat{\beta}_2$ will have sign of β_3 cov > 0, so we would expect coefficient of education to be biased upward.
            - Education is "proxying" for an omitted variable with which it is correlated, in addition to having its own effect.
    - Omitted-variable bias is a ubiquitous problem in econometrics because there are always potential explanatory variables that cannot be observed and included in the regression. It is extremely important to think about what variables are omitted, and how their effects are being picked up by the included variables.

Multiple regression

- $y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \ldots + \beta_K x_{i,K} + e_i$, for $i = 1, 2, \ldots, N$.
    - β_1 is the intercept term and can be thought of as the coefficient on $x_{i,1} \equiv 1$.
    - β_j for $j = 2, 3, \ldots, K$ is the partial effect of x_j on y.
- We can extend our least-squares (or method-of-moments or maximum-likelihood) analysis of the bivariate case to multiple regression easily. We now require that the expectation of the error term conditional on *each* of the *k* regressors be zero, which implies that the expected value of the product of each regressor with the error is zero and that the overall expected value of the error term is zero.
    - The population sum-of-squares function is

      $S = \sum_{i=1}^{N} \hat{e}_i^2 = \sum_{i=1}^{N} \left(y_i - b_1 - b_2 x_{i,2} - \ldots - b_K x_{i,K} \right)^2$, where the *b* coefficients are the OLS estimators that minimize *S*.
    - To minimize *S* we take the partial derivatives of *S* with respect to each *b* and set to zero:
        - $\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{N} \left(y_i - b_1 - b_2 x_{i,2} - \ldots - b_K x_{i,K} \right) = 0,$ which implies that $\sum_{i=1}^{N} \hat{e}_i = 0$, reflecting the population moment assumption that $E(e_i) = 0$ and
        - $\frac{\partial S}{\partial b_j} = -2 \sum_{i=1}^{N} x_{i,j} \left(y_i - b_1 - b_2 x_{i,2} - \ldots - b_K x_{i,K} \right) = 0, \quad j = 2, 3, \ldots, K,$ which reflects the population covariance assumption that $\text{cov}(e_i, x_{i,j}) = 0$
    - These are the OLS normal equations for multiple regressions. They are a set of *K* linear equations that can be solved for the *K* coefficient estimates.

- The solution, of course, is messy, but it can be described very compactly by the matrix notation that we developed for the bivariate case.

- Because we invested in matrix notation for the bivariate model, there is very little that needs to be changed to extend the model from two variables to many. In matrix form:
  - $\mathbf{y}$ is an $N \times 1$ column vector as before:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

  - $\mathbf{X}$ is now an $N \times K$ matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} & \cdots & x_{1,K} \\ 1 & x_{2,2} & x_{2,3} & \cdots & x_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,2} & x_{N,3} & \cdots & X_{N,K} \end{pmatrix}.$$

  - $\boldsymbol{\beta}$ is now a $K \times 1$ column vector of coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}.$$

  - And $\mathbf{e}$ is as before an $N \times 1$ vector of the error terms:

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}.$$

  - As in the bivariate case, we can write this system of N equations as $\mathbf{y = X\beta + e}$.
  - As in the bivariate case, the OLS coefficient estimator is $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$.
  - As in the bivariate case, the predicted values of $\mathbf{y}$ are $\hat{\mathbf{y}} = \mathbf{Xb}$ and the residuals are $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Xb}$.
  - Note that

$$\mathbf{X'X} = \begin{pmatrix} N & \sum_{i=1}^{N} x_{i,2} & \cdots & \sum_{i=1}^{N} x_{i,K} \\ \sum_{i=1}^{N} x_{i,2} & \sum_{i=1}^{N} x_{i,2}^2 & \cdots & \sum_{i=1}^{N} x_{i,2}x_{i,K} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{i,k} & \sum_{i=1}^{N} x_{i,2}x_{i,K} & \cdots & \sum_{i=1}^{N} x_{i,K}^2 \end{pmatrix},$$

$$\mathbf{X'y} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i,2} y_i \\ \vdots \\ \sum_{i=1}^{N} x_{i,K} y_i \end{pmatrix}.$$

So the **X′X** and **X′y** matrices include all the first and second moment information for the sample: *N*, the sum of each variable, the sum of the squares of each variable, and the sums of the cross-products of each pair of variables. In particular, **X′X** is often called the "moment matrix" of the regressors.

- **Estimating the error variance**
  - As before, we calculate the standard error of the regression (estimate) (SEE, SER, or RMSE)
  - $s^2 = \hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N} \hat{e}_i^2$. (Note that HGL uses degrees of freedom correction here but not in simple regression.)
  - The degrees of freedom are now $N - K$ because K degrees of freedom have been "used up" in the calculation of **b**.

OLS assumptions in multiple regression

- We need to add one assumption: that the regressors are not perfectly collinear
  - **MR1**: $y_i = \beta_1 + \beta_2 x_{i,2} + \ldots + \beta_K x_{i,K} + e_i, \quad i = 1, 2, \ldots, N.$
  - **MR2**: $E(e_i) = 0$.
  - **MR3**: $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$ (homoskedasticity)
  - **MR4**: $\text{cov}(e_i, e_j) = 0, \quad \forall i \neq j.$ (no autocorrelation)
  - **MR5**: The values of *x* are non-random and not perfectly collinear: No perfect multicollinearity
    - Intuitively, it means that within the sample, no variable can be expressed as an exact linear function of the other variables (including the constant 1). Note that nonlinear functions are OK: we can include both age and age-squared, for example. But if we define a work experience variable as age – education – 6 (as is often done), then we can't include age, education, and experience in the regression because experience is a linear function of age and education.
    - The most common violation of this is the "dummy variable trap" in which we include a dummy for male, a dummy for female, and a constant. If all observations in the sample are either male or female, then

the two dummies add up to 1, which equals the constant term. Thus, we have perfect multicollinearity and cannot perform the regression.

- Perfect multicollinearity also results when one variable is equal to a constant (zero or one, if a dummy is turned on or off for every observation)
- Mathematically, the assumption of no perfect multicollinearity means that the **X** matrix has full column rank (rank *K*), so that the **X′X** matrix is non-singular and has an inverse.
- What happens if regressors are *nearly* collinear? Then it becomes impossible for OLS to distinguish between the effects of nearly collinear regressors.
  - The **X′X** matrix is nearly singular, which means that the diagonal elements of its inverse are very large (kind of like dividing by zero, note the simple-regression formula for the slope estimator requires variation in *x*).
  - The large diagonal elements of **X′X** lead to large estimated standard errors of the coefficients, accurately reflecting the problem that OLS has in estimating the effects of individual variables.

- **MR6**: (optional) $e_i \sim N(0,\sigma^2)$

Distribution of OLS multiple-regression estimators

- If the error term is classical (including homoskedasticity), then we showed before that the covariance matrix of the coefficient estimator is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- The Gauss-Markov Theorem also tells us that OLS is BLUE in the multiple-regression case under the following classical conditions:
  - $E(\mathbf{e}\,|\,\mathbf{X}) = \mathbf{0}_N$,

    $E(\mathbf{e}\mathbf{e}'\,|\,\mathbf{X}) = \sigma^2\mathbf{I}_N$,

    X has full column rank.
  - In the case of multiple regression, "best" means that the covariance matrix of any other estimator can be shown to be "larger" than the OLS estimator's by a positive definite matrix.
  - If the error term is conditionally normally distributed, then the OLS estimator is also normally distributed (and the *t* statistics follow the *t* distribution with $N-K$ degrees of freedom).