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 $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$, so that the true effect of X_1 on Y is β_1. Instead of fitting this model, we instead fit a simple regression model $Y_i = \gamma_0 + \gamma_1 X_1 + v_i$. Will our estimate $\hat{\gamma}_1$ be a good (unbiased, at least) estimate of the effect β_1? No, in general it will be biased:
    - The error term in the estimated simple regression model will include the effect of X_2 in addition to the true error u: $v_i = \beta_2 X_2 + u_i$
    - Applying the simple-regression expected value formula from earlier,

$$\hat{\gamma}_1 = \beta_1 + \left[\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_{1,i} - \bar{X}_1\right)v_i}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{1,i} - \bar{X}_1\right)^2} \right]$$

$$= \beta_1 + \left[\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_{1,i} - \bar{X}_1\right)\left(\beta_2 X_{2,i} + u_i\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{1,i} - \bar{X}_1\right)^2} \right].$$

      Assuming the standard OLS assumptions are correct for the two-variable model, the u term in the numerator has expectation of zero. The cross-product term in the numerator has probability limit $\beta_2 \text{cov}(X_1, X_2)$. The denominator has plim of the variance of X_1. Thus $\text{plim}\left(\hat{\gamma}_1\right) = \beta_1 + \beta_2 \frac{\text{cov}\left(X_1, X_2\right)}{\text{var}\left(X_1\right)}$. Note that the ratio in this expression is (expectation of) the regression slope coefficient we would get by regressing X_2 on X_1.
    - Thus, the omission of X_2 from the regression biases the OLS estimate of the coefficient on X_1 unless one of two conditions is true:
        - X_2 doesn't really belong in the regression ($\beta_2 = 0$), **or**
        - X_1 and X_2 are uncorrelated (cov = 0).
    - This is known as **omitted-variable bias**.
        - The bias has the sign of the product of β_2 and $\text{cov}(X_1, X_2)$.
        - Temperature/rainfall example on daily umbrella sales
            - Rainfall and temp are negatively correlated (cov < 0)

        - True effect of temp on umbrella sales is zero ($\beta_1 = 0$)
        - True effect of rainfall on umbrella sales is positive ($\beta_2 > 0$)
        - Bias in $\hat{\beta}_1$ will have sign of β_2 cov < 0, so we would expect temperature to have a negative estimated coefficient, even though the true effect is zero.
        - Temperature is "proxying" for an omitted variable with which it is correlated.
    - 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

- $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i$, for $i = 1, 2, \ldots, n$.
    - β_0 is the intercept term and can be thought of as the coefficient on $X_{0,i} \equiv 1$.
    - β_j for $j = 1, 2, \ldots, k$ is the partial effect of X_j on Y.
- We can extend our method-of-moments 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 moment conditions are
        - $Eu_i = 0$
        - $E\left[\left(X_{j,i} - E\left(X_j\right)\right)u_i\right] = 0, \quad j = 1, 2, \ldots k.$
    - The corresponding sample conditions are
        - $\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \ldots - \hat{\beta}_k X_{k,i} = 0,$
        - $\sum_{i=1}^{n} \left(X_{j,i} - \bar{X}_j\right)\hat{u}_i = \sum_{i=1}^{n} \left(X_{j,i} - \bar{X}_j\right)\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \ldots - \hat{\beta}_k X_{k,i}\right) = 0, \quad j = 1, 2, \ldots k.$
    - These are the OLS normal equations for multiple regressions. They are a set of $k + 1$ linear equations that can be solved for the $k + 1$ 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:

- Y is an $n \times 1$ column vector as before:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

- X is now an $n \times (k + 1)$ matrix:

$$X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{k,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{k,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{k,n} \end{pmatrix}.$$

- β is now a $(k + 1) \times 1$ column vector of coefficients

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

- And u is as before an $n \times 1$ vector of the error terms:

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

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

$$X'X = \begin{pmatrix} n & \sum_{i=1}^{n} X_{1,i} & \cdots & \sum_{i=1}^{n} X_{k,i} \\ \sum_{i=1}^{n} X_{1,i} & \sum_{i=1}^{n} X_{1,i}^2 & \cdots & \sum_{i=1}^{n} X_{1,i} X_{k,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{k,i} & \sum_{i=1}^{n} X_{1,i} X_{k,i} & \cdots & \sum_{i=1}^{n} X_{k,i}^2 \end{pmatrix},$$

$$X'Y = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{1,i} Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{k,i} Y_i \end{pmatrix}.$$

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."

Goodness of fit

- Standard error of the regression is similar to bivariate case, but with $n - k - 1$ degrees of freedom.
    - There are n pieces of information in the dataset. We use $k + 1$ of them to minimally define the regression function (estimate the $k + 1$ coefficients). There are $n - (k + 1) = n - k - 1$ degrees of freedom left.
    - $SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n} \hat{u}_i^2} = \sqrt{\frac{SSR}{n-k-1}}.$
- R^2 is defined the same way: the share of variance in Y that is explained by the set of explanatory variables:
    - $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}.$
    - However, adding a new regressor to the equation *always* improves R^2 (unless it is totally uncorrelated with the previous residuals), so we would expect an equation with 10 regressors to have a higher R^2 than one with only 2. To correct for this, we often use an **adjusted R^2** that corrects for the number of degrees of freedom:

      $$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}.$$

    - Three properties of $\bar{R}^2$:
        - $\bar{R}^2 < R^2$ whenever $k > 0$.

- Adding a regressor generally decreases *SSR*, but also increases *k*, so the effect on $\bar{R}^2$ is ambiguous. Choosing a regression to maximize $\bar{R}^2$ is not recommended, but it's better than maximizing R^2.
- $\bar{R}^2$ can be negative if *SSR* is close to *TSS*, because $\frac{n-1}{n-k-1} > 1$.

OLS assumptions in multiple regression

- We need to add one assumption: that the regressors are not perfectly collinear
  - **Assumption #1**: $E\left(u_i \mid X_{1,i}, X_{2,i}, \ldots, X_{k,i}\right) = 0.$
  - **Assumption #2**: $\left(Y_i, X_{1,i}, X_{2,i}, \ldots, X_{k,i}\right), i = 1, 2, \ldots, n$ are IID.
  - **Assumption #3**: All variables have non-zero, finite fourth moments, so large outliers are unlikely.
  - **Assumption #4**: 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* + 1), 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.

Distribution of OLS multiple-regression estimators

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

    $E(uu' \mid X) = \sigma_u^2 I_n$,

    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 $n - k - 1$ degrees of freedom).
- Under the more general assumptions above, OLS can be shown to be asymptotically normal.
  - The covariance matrix of the estimated coefficient vector is below

    $\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0_{k+1}, \Sigma_{\sqrt{n}\left(\hat{\beta}-\beta\right)}\right)$,

    $\Sigma_{\sqrt{n}\left(\hat{\beta}-\beta\right)} \equiv Q_X^{-1} \Sigma_V Q_X^{-1}$,

    $Q_X \equiv E\left(X_i X_i'\right)$,

    $\Sigma_V \equiv E\left(V_i V_i'\right)$,

    $V_i \equiv X_i u_i$.

    (Note that X_i is a column vector of the observations of the ith variable, which is a row vector in the X matrix.)
  - Compare this formula to the bivariate slope variance:

    $$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{1}{n} \frac{\operatorname{var}\left[\left(X_i - EX\right) u_i\right]}{\left[\operatorname{var}\left(X_i\right)\right]^2}.$$

    - The variance of the cross-product of X and u lives on in the Σ_V matrix, which is the covariance matrix of all the X variables with u.

        - The squared variance of X in the denominator becomes the inverse Q_X matrices that are pre- and post-multiplied by Σ_V.
- In practice, we don't know the Q and Σ matrices, so they must be estimated in order to get standard errors (and covariances among coefficients if we want them).
    - Approximate $Q_X \equiv E\left(X_i X_i'\right)$ by the sample moment $\frac{1}{n}X'X$, recalling that the inner product of the X matrix is the sum of squares and cross-products. (Note that X_i is a $(k + 1)$-element column vector corresponding to the ith observation. It is the ith *row* of the X matrix. So $X_i X_i'$ is a $(k + 1) \times (k + 1)$ matrix corresponding to the covariances of *a single observation*. By the IID assumption, these covariances are the same for all observations. The $X'X$ matrix divided by n calculates the sample covariances averaging across all observations.)
    - To approximate Σ_V we want sample moments for $\Sigma_V = E\left(V_i V_i'\right) = E\left(X_i u_i u_i' X_i'\right)$. But since u_i is a scalar, $E\left(X_i u_i u_i' X_i'\right) = E\left(u_i^2 X_i X_i'\right)$, for which the corresponding sample moments are $\hat{\Sigma}_{\hat{V}} \equiv \frac{1}{n-k-1}\sum_{i=1}^{n} X_i X_i' \hat{u}_i^2$.
    - Once we have calculated $\hat{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)} \equiv \left(\frac{X'X}{n}\right)^{-1} \hat{\Sigma}_{\hat{V}} \left(\frac{X'X}{n}\right)^{-1}$, the estimate of the heteroskedasticity-robust covariance matrix, the square roots of the diagonal elements are the robust standard errors of the coefficient estimators.