PUAD 7130: Multiple Regression Estimation

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

September 30, 2021

Overview

Motivation

- Previously, we built the simple, bivariate linear regression model to help us estimate the effect some explanatory variable, x, had one some dependent variable, y.
- But our assumption that x is uncorrelated with u is probably unrealistic.
- We'd like to say that x affects y, *ceteris paribus*. So how do we get there using regression?
- Simple. We "control" for other factors simultaneously via "multiple regression" as we attempt to isolate x's affect on y.

Introduction to multiple regression

- Consider an econometric model with two explanatory variables, "education" and "experience," with a dependent variable of "wage":

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + u_i.$$

- Had we omitted experience as a right-hand-side covariate, its effect would be in the error term.
- Thus, removing it from the error term allows us to estimate the effect of education on wage while *holding experience constant*.
- Similarly, we needn't (and shouldn't) be limited to just two explanatory variables.

Generalized multiple regression

- Multiple regression analysis allows us to examine the effect of x on y while controlling for potentially many other variables.
- To generalize to the k^{th} variable, we get the following econometric model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} + \hat{u}_i, \qquad (1)$$

- Each x_k represents some independent variable, $\hat{\beta}_0$ is the intercept, the other $\hat{\beta}_k$s now represent the "partial slope coefficients" for each variable, and $\hat{u}_i$ represents the unaccounted variance (or error).
- As before, we continue to assume that what is contained in $\hat{u}$ is uncorrelated with the independent variables.

Estimating the multiple regression coefficients

- Previously, we used calculus and the scalar form of the linear model to derive via closed-form solution the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ that minimize the sum of squared errors.
- We want to use the same logic in multiple regression, but the equations we derived for the bivariate model are no longer appropriate since they don't account for the linear effect of other independent variables.
- We could continue in the scalar form, but the addition of more variables makes the work rather tedious. Instead, we'll derive a solution via matrix form.

Data in the matrix form

- We begin with a population regression function. We'll use subscript, t to denote observation 1 through n and k to denote parameters:

$$y_t \quad = \quad \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t. \qquad (2)$$

- We can take equation (2) and rewrite this as a series of equations:

$$\begin{array}{c} y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1 \\ y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2 \\ \hline \\ y_n = \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \ldots + \beta_k x_{kn} + u_n \end{array} \qquad (3)$$

Data in the matrix form (cont'd.)

- Looking at equation (3), we can see that really all we have here is a matrix.
- For each observation, t, define a $1 \times (k+1)$ vector, $\boldsymbol{x}_t = (1, x_{t1}, \ldots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ be a $(k+1) \times 1$ vector of all parameters. (Note the use of bold-face.)
- Finally, let $\boldsymbol{u}$ be the $n \times 1$ vector of unobservable errors.

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \ldots \\ u_n \end{bmatrix} \tag{4}$$

Data in the matrix form (cont'd.)

- Next, let $\boldsymbol{X}$ be the $n \times (k+1)$ vector of observations.
- Thus, with no alternative in meaning, we can rewrite equation (2) using matrix notation:

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{u} \tag{5}$$

- To multiply two matrixes together, $\boldsymbol{A} \times \boldsymbol{B}$, the column dimensions of $\boldsymbol{A}$ must be equal to the row dimensions of $\boldsymbol{B}$. Thus a matrix with dimensions 1×3 (1 row and 3 columns) can be multiplied by a matrix with dimensions 3×1.
- Because the product of two matrixes takes the shape of its constituent outside elements,[1] $\boldsymbol{X}\boldsymbol{\beta}$ is $n \times 1$.

[1] For example, multiplying a matrix with dimensions 2×3 by another matrix with dimensions 3×4 yields a matrix with dimensions 2×4.

Deriving the $\hat{\beta}$ coefficients via matrix form

- As in the bivariate case, we want to minimize the sum of squared errors. For any $(k+1) \times 1$ vector of coefficients, $\boldsymbol{b}$, we get:

$$SSR(\boldsymbol{b}) \equiv \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t \boldsymbol{b})^2. \quad (6)$$

- We want to minimize the expression in equation (6) such that:

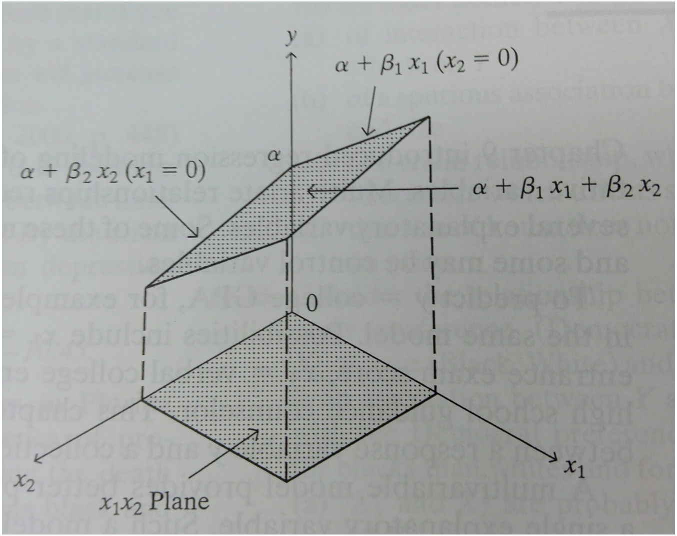
$$\frac{\partial SSR(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{b}} = -2 \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t \hat{\boldsymbol{\beta}}) \boldsymbol{x}_t. \quad (7)$$

Deriving the $\hat{\beta}$ coefficients via matrix form (cont'd.)

- Next, we want to set equation (7) equal to zero and solve for our vector of coefficients:

$$\begin{aligned}
\sum_{t=1}^{n} \boldsymbol{x}_t'(y_t - \boldsymbol{x}_t\hat{\boldsymbol{\beta}}) &= \mathbf{0}, \\
\sum_{t=1}^{n} \boldsymbol{x}_t'(y_t - \boldsymbol{x}_t\hat{\boldsymbol{\beta}}) &= \mathbf{0}, \\
\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \mathbf{0}, \\
\boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}, \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \qquad (8)
\end{aligned}$$

A 2-d multiple regression plane



How to interpret multiple regression OLS output

- We refer to each $\hat{\beta}_k$ as the "partial slope coefficient."
- For each independent variable, we say, "A one-unit change in X_k has a corresponding $\hat{\beta}_k$ effect on Y, *ceteris paribus*.
- By holding other variables constant, we can isolate the effect of a given X on Y.
- But remember, any leftover or unaccounted for variance is going into $\hat{u}$.

Why Multiple Regression?

- Wouldn't it be easier to estimate bivariate regressions piecemeal?
- Well, yes. But that doesn't mean we should. We use multiple regression to combat endogeneity in the error term.
- Remember, what explains the error in u_i? It's anything not included as an independent variable. By including more IVs, we are sucking error variance out of u_i.

Assumption 1: Linearity

- Just as with the bivariate OLS model, we require linearity in our parameter.
- Again, we can allow non-linearity in our variables, just not in our partial slope coefficients.

Assumption 2: Random sampling

- As with the bivariate model, we assume that our data derive from a random sampling method.
- This doesn't guarantee us a representative sample, but it maximizes the likelihood we got one.
- As we mentioned above, though, absent time-series features in the data, we can assume random sampling for most cross-sectional research designs.

Assumption 3: No perfect collinearity

- This one's a little different from the bivariate context.
- If some variable, x is *perfectly collinear* with some other variable, y, then OLS can only estimate parameters for one of these variables.
- We're unlikely to run across many truly perfectly collinear relationships in the wild.
- Odds are when you see it is due to dummy variables. With categorical controls, we need to omit a "reference category."

Assumption 4: Zero conditional mean

- As with the bivariate model, we assume that the expected value of the error term is equal to zero, conditional upon the covariates, $E(u \mid \boldsymbol{X}) = 0$.
- The way violations of this assumption usually emerge is via omitted variable bias.
- If Corr$(\boldsymbol{X}, \boldsymbol{u}) > 0$, the simplest fix is to identify what variables are missing in the model and to put them there. Then they're out of the error, and your $\boldsymbol{X}$ can't covary with it.

Theorem 1: Unbiasedness of OLS

- Under assumptions 1 through 4, we may conclude that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- Again, please note that there is no guarantee that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, merely that, *in expectation,* due primarily to random sampling and zero mean variance, we have no a priori reason to suspect a biased set of coefficients.

Inclusion of irrelevant variables

- We've touched on omitted variables, but what if we include some variable—call it a—that *doesn't* belong in the regression model? Will this bias our results?
- Suppose we have variables x_1 and x_2, which are theoretically linked to outcomes in y. Then we have:

$$y_i \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 c_i + \hat{u}_i.$$

- If c is truly irrelevant to outcomes in y, then its partial slope coefficient is $\hat{\beta}_3 = 0$.
- Obviously, then, the effect of c will be canceled out, and we'll just be left with the relevant independent variables and the error term.

Omitted variable bias

- But now it's time to prove that omitting a relevant variable is not without penalty to $\hat{\boldsymbol{\beta}}$.
- Suppose we specify an *under-specified* regression model: $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{u}_i$.
- And now suppose we should have included another variable, x_2. In the true model, we would observe: $y = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i$.
- Then we see that $\tilde{u}_i = \hat{\beta}_2 x_{i2} + \hat{u}_i.$, and so long as $\hat{\beta}_2 \neq 0$, $\tilde{u}_i \neq \hat{u}_i$, which ensures that $\tilde{\beta}_1 \neq \hat{\beta}_1$.
- Hence, omitted variables can bias coefficients for the variables that are included in an OLS model.

Assumption 5: Homoskedasticity

- As in the case of the bivariate model, we assume that the error term has the same variance given any values in the independent variables.
- Put differently, OLS assumes that $\text{Var}(u \mid \boldsymbol{X}) = \sigma^2$, where σ^2 denotes the error variance of the model.

Theorem 2: Sampling variance estimates

- Under assumptions 1 through 5, we can estimate the variance of our slope coefficients conditional on the sample values of the independent variable:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}, \quad (9)$$

  where R_j^2 represents the R^2 from regressiong x_j on all other independent varialbes (and including an intercept).
- We then that as variance in x increases, variance in $\hat{\beta}$ increases.
- And as components in $\boldsymbol{X}$ become more linearly related, variance in the coefficients increases—a problem known as multicollinearity. (Remember, to *perfect* multicollinearity allowed.)
- Importantly, if assumption 5 fails and we have heteroskedasticity, we will not have efficient standard errors for the slope coefficients.

When is multicollinearity a potential problem?

- We have a method for measuring the amount of multicollinearity in our variables, known as variance inflation factor (VIF).
- Formally, we calculate VIF for each coefficient as $VIF_j = \frac{1}{1-R_j^2}$.
- By tradition, VIF measures greater than or equal to 10 tend to raise an eyebrow and may justify considering dropping the offending variables from the regression.

Theorem 3: Standard errors of the esimates

- Under assumptions 1 through 5, $E(\hat{\sigma}) = \sigma^2$. Taking the root, we denote $\hat{\sigma}$ the standard error of the regression.
- With $\hat{\sigma}$ in hand, we can now estimate the standard error of $\hat{\beta}_j$, which is what allows us to hypothesis-test in OLS:

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum(x_{ij} - \bar{x})(1 - R_j^2)}} \tag{10}$$

Theorem 4: The Gauss-Markov Theorem

- Under assumptions 1 through 5, OLS produces $\hat{\boldsymbol{\beta}}$, are the best, linear, unbiased, estimates of $\boldsymbol{\beta}$. (BLUE)
- By "best," we refer to minimal variance—i.e., efficiency (follows from Assumptions 3 and 5).
- By "linear" we refer to the parameters (follows from Assumption 1).
- By "unbiased," we refer to the $\hat{\beta}$ estimates themselves (follows from Assumptions 2 and 4).
- By "estimates," we refer to the method of minimizing the sum of squared error variance (OLS itself).

Discussion

- When our dependent variable is measured continuously, OLS is, given the Gauss-Markov assumptions are satisfied, the most appropriate statistical regression technique.
- Multiple regression allows us to estimate the effect of some x on y even while controlling for other variables.
- As we proceed, we'll learn how to make inferences using OLS, how to test OLS assumptions, and how to revise our regression tehcniques when Gauss-Markov assumptions are not met.