Wooldridge, Introductory Econometrics, 2d ed.

Chapter 3: Multiple regression analysis: Estimation

In multiple regression analysis, we extend the simple (two-variable) regression model to consider the possibility that there are additional explanatory factors that have a systematic effect on the dependent variable. The simplest extension is the "three-variable" model, in which a second explanatory variable is added:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{1}$$

where each of the slope coefficients are now partial derivatives of y with respect to the x variable which they multiply: that is, holding x_2 fixed, $\beta_1 = \partial y / \partial x_1$. This extension also allows us to consider nonlinear relationships, such as a polynomial in z, where $x_1 = z$ and $x_2 = z^2$. Then, the regression is linear in x_1 and x_2, but nonlinear in $z : \partial y / \partial z = \beta_1 + 2\beta_2 z$. The key assumption for this model, analogous to that which we specified for the simple regression model, involves the independence of the error process u and both regressors, or explanatory variables:

$$E(u \mid x_1, x_2) = 0. \tag{2}$$

This assumption of a zero conditional mean for the error process implies that it does not systematically vary with the $x's$ nor with any linear combination of the $x's$; u is independent, in the statistical sense, from the distributions of the $x's$.

The model may now be generalized to the case of k regressors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{3}$$

where the β coefficients have the same interpretation: each is the partial derivative of y with respect to that x, holding all other $x's$ constant (*ceteris paribus*), and the u term is that nonsystematic part of y not linearly related to any of the $x's$. The dependent variable y is taken to be linearly related to the $x's$, which may bear any relation to each other (e.g. polynomials or other transformations) as long as there are no exact linear dependencies among the regressors. That is, no x variable can be an exact linear transformation of another, or the regression estimates cannot be calculated. The independence assumption now becomes:

$$E\left(u \mid x_1, x_2, ..., x_k\right) = 0. \tag{4}$$

Mechanics and interpretation of OLS

Consider first the "three-variable model" given above in (1). The estimated OLS equation contains the parameters of interest:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \tag{5}$$

and we may define the ordinary least squares criterion in terms of the OLS residuals, calculated from a sample of size n, from this expression:

$$\min S = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2}\right)^2 \tag{6}$$

where the minimization of this expression is performed with respect to each of the three parameters, $\{b_0, b_1, b_2\}$. In the case of k regressors, these expressions include terms in b_k, and the minimization is performed with respect to the $(k+1)$ parameters $\{b_0, b_1, b_2, ...b_k\}$. For this to be feasible, $n > (k+1)$: that is,

we must have a sample larger than the number of parameters to be estimated from that sample. The minimization is carried out by differentiating the scalar S with respect to each of the $b's$ in turn, and setting the resulting first order condition to zero. This gives rise to $(k+1)$ simultaneous equations in $(k+1)$ unknowns, the regression parameters, which are known as the **least squares normal equations**. The normal equations are expressions in the sums of squares and cross products of the y and the regressors, including a first "regressor" which is a column of $1's$, multiplying the constant term. For the "three-variable" regression model, we can write out the normal equations as:

$$\begin{aligned} \sum y &= nb_0 + b_1 \sum x_1 + b_2 \sum x_2 \\ \sum x_1 y &= b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \\ \sum x_2 y &= b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \end{aligned} \tag{7}$$

Just as in the "two-variable" case, the first normal equation can be interpreted as stating that the regression surface (in 3-space) passes through the multivariate point of means $\{\bar{x}_1, \bar{x}_2, \bar{y}\}$. These three equations may be uniquely solved, by normal algebraic techniques or linear algebra, for the estimated least squares parameters.

This extends to the case of k regressors and $(k+1)$ regression parameters. In each case, the regression coefficients are considered in the *ceteris paribus* sense: that each coefficient measures the partial effect of a unit change in its variable, or regressor, holding all other regressors fixed. If a variable is a component of more than one regressor–as in a polynomial

relationship, as discussed above–the total effect of a change in that variable is additive.

Fitted values, residuals, and their properties

Just as in simple regression, we may calculate fitted values, or predicted values, after estimating a multiple regression. For observation i, the fitted value is

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} \tag{8a}$$

and the residual is the difference between the actual value of y and the fitted value:

$$e_i = y_i - \hat{y}_i \tag{9}$$

As with simple regression, the sum of the residuals is zero; they have, by construction, zero covariance with each of the x variables, and thus zero covariance with $\hat{y};$ and since the average residual is zero, the regression surface passes through the multivariate point of means, $\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k, \bar{y}\}$.

There are two instances where the simple regression of y on x_1 will yield the same coefficient as the multiple regression of y on x_1 and x_2, with respect to x_1. In general, the simple regression coefficient will not equal the multiple regression coefficient, since the simple regression ignores the effect of x_2 (and considers that it can be viewed as nonsystematic, captured in the error $u)$. When will the two coefficients be equal? First, when the coefficient of x_2 is truly zero–that is, when x_2 really does not belong in the model. Second, when x_1 and x_2 are uncorrelated in the sample. This is likely to be quite rare in actual data. However, these two cases suggest when the two coefficients will be similar; when x_2 is relatively unimportant in explaining $y,$ or when it is

very loosely related to x_1.

We can define the same three sums of squares–SST, SSE, $SSR-$as in simple regression, and R^2 is still the ratio of the explained sum of squares (SSE) to the total sum of squares (SST). It is no longer a simple correlation (e.g. r_{yx}) squared, but it still has the interpretation of a squared simple correlation coefficient: the correlation between y and $\hat{y}$, $r_{\hat{y}y}$. A very important principle is that R^2 never decreases when an explanatory variable is added to a regression–no matter how irrelevant that variable may be, the R^2 of the expanded regression will be no less than that of the original regression. Thus, the regression R^2 may be arbitrarily increased by adding variables (even unimportant variables), and we should not be impressed by a high value of R^2 in a model with a long list of explanatory variables.

Just as with simple regression, it is possible to fit a model through the origin, suppressing the constant term. It is important to note that many of the properties we have discussed no longer hold in that case: for instance, the least squares residuals (e_is) no longer have a zero sample average, and the R^2 from such an equation can actually be negative–that is, the equation does worse than the "model" which specifies that $\hat{y} = \bar{y}$ for all i. If the population intercept β_0 differs from zero, the slope coefficients computed in a regression through the origin will be biased. Therefore, we often will include an intercept, and let the data determine whether it should be zero.

Expected value of the OLS estimators

We now discuss the statistical properties of the OLS estimators of the parameters in the population regression function. The population model is taken to be (3). We assume that we have a random sample of size n on the variables of the model. The multivariate analogue to our assumption about the error process is now:

$$E\left(u \mid x_1, x_2, ..., x_k\right) = 0 \tag{10}$$

so that we consider the error process to be independent of each of the explanatory variables' distributions. This assumption would not hold if we misspecified the model: for instance, if we ran a simple regression with inc as the explanatory variable, but the population model also contained inc^2. Since inc and inc^2 will have a positive correlation, the simple regression's parameter estimates will be biased. This bias will also appear if there is a separate, important factor that should be included in the model; if that factor is correlated with the included regressors, their coefficients will be biased.

In the context of multiple regression, with several independent variables, we must make an additional assumption about their measured values:

Proposition 1 *In the sample, none of the independent variables x may be expressed as an exact linear relation of the others (including a vector of 1s).*

Every multiple regression that includes a constant term can be considered as having a variable $x_{0i} = 1 \; \forall i$. This proposition

states that each of the other explanatory variables must have nonzero sample variance: that is, it may not be a constant in the sample. Second, the proposition states that there is no **perfect collinearity**, or **multicollinearity**, in the sample. If we could express one x as a linear combination of the other x variables, this assumption would be violated. If we have perfect collinearity in the regressor matrix, the OLS estimates cannot be computed; mathematically, they do not exist. A trivial example of perfect collinearity would be the inclusion of the same variable twice, measured in different units (or via a linear transformation, such as temperature in degrees F versus C). The key concept: each regressor we add to a multiple regression must contain information at the margin. It must tell us something about y that we do not already know. For instance, if we consider x_1 : proportion of football games won, x_2 : proportion of games lost, and x_3: proportion of games tied, and we try to use all three as explanatory variables to model alumni donations to the athletics program, we find that there is perfect collinearity: since for every college in the sample, the three variables sum to one by construction. There is no information in, e.g., x_3 once we know the other two, so including it in a regression with the other two makes no sense (and renders that regression uncomputable). We can leave any one of the three variables out of the regression; it does not matter which one. Note that this proposition is not an assumption about the population model: it is an implication of the sample data we have to work with. Note also that this only applies to linear relations among the explanatory variables:

a variable and its square, for instance, are not linearly related, so we may include both in a regression to capture a nonlinear relation between y and x.

Given the four assumptions: that of the population model, the random sample, the zero conditional mean of the u process, and the absence of perfect collinearity, we can demonstrate that the OLS estimators of the population parameters are unbiased:

$$Eb_j = \beta_j,\ j = 0, ..., k \tag{11}$$

What happens if we misspecify the model by including **irrelevant explanatory variables**: x variables that, unbeknowst to us, are not in the population model? Fortunately, this does not damage the estimates. The regression will still yield unbiased estimates of all of the coefficients, including unbiased estimates of these variables' coefficients, which are zero in the population. It may be improved by removing such variables, since including them in the regression consumes degrees of freedom (and reduces the precision of the estimates); but the effect of **overspecifying** the model is rather benign. The same applies to overspecifying a polynomial order; including quadratic and cubic terms when only the quadratic term is needed will be harmless, and you will find that the cubic term's coefficient is far from significant.

However, the opposite case–where we **underspecify** the model by mistakenly excluding a relevant explanatory variable– is much more serious. Let us formally consider the direction and size of bias in this case. Assume that the population model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{12}$$

but we do not recognize the importance of x_2, and mistakenly consider the relationship

$$y = \beta_0 + \beta_1 x_1 + u \tag{13}$$

to be fully specified. What are the consequences of estimating the latter relationship? We can show that in this case:

$$Eb_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} \tag{14}$$

so that the OLS coefficient b_1 will be biased–not equal to its population value of β_1, even in an expected sense–in the presence of the second term. That term will be nonzero when β_2 is nonzero (which it is, by assumption) and when the fraction is nonzero. But the fraction is merely a simple regression coefficient in the auxiliary regression of x_1 on x_2. If the regressors are correlated with one another, that regression coefficient will be nonzero, and its magnitude will be related to the strength of the correlation (and the units of the variables). Say that the auxiliary regression is:

$$x_1 = d_0 + d_1 x_2 + u \tag{15}$$

with $d_1 > 0$, so that x_1 and x_2 are positively correlated (e.g. as income and wealth would be in a sample of household data). Then we can write the bias as:

$$Eb_1 - \beta_1 = \beta_2 d_1 \tag{16}$$

and its sign and magnitude will depend on both the relation between y and x_2 and the interrelation among the explanatory variables. If there is no such relationship–if x_1 and x_2 are uncorrelated in the sample–then b_1 is unbiased (since in that

special case multiple regression reverts to simple regression). In all other cases, though, there will be bias in the estimation of the underspecified model. If the left side of (16) is positive, we say that b_1 has an upward bias: the OLS value will be too large. If it were negative, we would speak of a downward bias. If the OLS coefficient is closer to zero than the population coefficient, we would say that it is "biased toward zero" or attenuated.

It is more difficult to evaluate the potential bias in a multiple regression, where the population relationship involves k variables and we include, for instance, $k-1$ of them. All of the OLS coefficients in the underspecified model will generally be biased in this circumstance unless the omitted variable is uncorrelated with each included regressor (a very unlikely outcome). What we can take away as a general rule is the asymmetric nature of specification error: it is far more damaging to exclude a relevant variable than to include an irrelevant variable. When in doubt (and we almost always are in doubt as to the nature of the true relationship) we will always be better off erring on the side of caution, and including variables that we are not certain should be part of the explanation of y.

Variance of the OLS estimators

We first reiterate the assumption of homoskedasticity, in the context of the $k-$variable regression model:

$$Var\left(u \mid x_1, x_2, ..., x_k\right) = \sigma^2 \tag{17}$$

If this assumption is satisfied, then the error variance is identical for all combinations of the explanatory variables. If it is violated, we say that the errors are **heteroskedastic**, and

must be concerned about our computation of the OLS estimates' variances. The OLS estimates are still unbiased in this case, but our estimates of their variances are not. Given this assumption, plus the four made earlier, we can derive the sampling variances, or precision, of the OLS slope estimators:

$$Var\left(b_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)},\ j = 1, ..., k \tag{18}$$

where SST_j is the total variation in x_j about its mean, and R_j^2 is the R^2 from an auxiliary regression from regressing x_j on all other x variables, including the constant term. We see immediately that this formula applies to simple regression, since the formula we derived for the slope estimator in that instance is identical, given that $R_j^2 = 0$ in that instance (there are no other x variables). Given the population error variance σ^2, what will make a particular OLS slope estimate more precise? Its precision will be increased (i.e. its sampling variance will be smaller) the larger is the variation in the associated x variable. Its precision will be decreased, the larger the amount of variable x_j that can be "explained" by other variables in the regression. In the case of perfect collinearity, $R_j^2 = 1$, and the sampling variance goes to infinity. If R_j^2 is very small, then this variable makes a large marginal contribution to the equation, and we may calculate a relatively more precise estimate of its coefficient. If R_j^2 is quite large, the precision of the coefficient will be low, since it will be difficult to "partial out" the effect of variable j on y from the effects of the other explanatory variables (with which it is highly

correlated). However, we must hasten to add that the assumption that there is no perfect collinearity does not preclude R_j^2 from being close to unity–it only states that it is less than unity. The principle stated above when we discussed collinearity–that at the margin, each explanatory variable must add information that we do not already have, in whole or in large part–if that variable is to have a meaningful role in a regression model of y. This formula for the sampling variance of an OLS coefficient also explains why we might not want to overspecify the model: if we include an irrelevant explanatory variable, the point estimates are unbiased, but their sampling variances will be larger than they would be in the absence of that variable (unless the irrelevant variable is uncorrelated with the relevant explanatory variables).

How do we make (18) operational? As written, it cannot be computed, since it depends on the unknown population parameter σ^2. Just as in the case of simple regression, we must replace σ^2 with a consistent estimate:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-(k+1))} = \frac{\sum_{i=1}^{n} e_i^2}{(n-k-1)} \tag{19}$$

where the numerator is just SSR, and the denominator is the sample size, less the number of estimated parameters: the constant and k slopes. In simple regression, we computed s^2 using a denominator of 2: intercept plus slope. Now, we must account for the additional slope parameters. This also suggests that we cannot estimate a $k-$variable regression model without having a sample of size at least $(k+1)$. Indeed, just as two points define a straight line, the degrees of freedom in simple regression

will be positive iff $n > 2$. For multiple regression, with k slopes and an intercept, $n > (k + 1)$. Of course, in practice, we would like to use a much larger sample than this in order to make inferences about the population.

The positive square root of s^2 is known as the **standard error of regression**, or SER. (Stata reports s on the regression output labelled "Root MSE", or root mean squared error). It is in the same units as the dependent variable, and is the numerator of our estimated standard errors of the OLS coefficients. The magnitude of the SER is often compared to the mean of the dependent variable to gauge the regression's ability to "explain" the data.

In the presence of heteroskedasticity–where the variance of the error process is not constant over the sample–the estimate of s^2 presented above will be biased. Likewise, the estimates of coefficients' standard errors will be biased, since they depend on s^2. If there is reason to worry about heteroskedasticity in a particular sample, we must work with a different approach to compute these measures.

Efficiency of OLS estimators

An important result, which underlays the widespread use of OLS regression, is the **Gauss-Markov Theorem**, describing the relative efficiency of the OLS estimators. Under the assumptions that we have made above for multiple regression–and making no further distributional assumptions about the error process–we may show that:

Proposition 2 ***(Gauss-Markov)*** *Among the class of linear, unbiased estimators of the population regression function, OLS provides the best estimators, in terms of minimum sampling variance: OLS estimators are best linear unbiased estimators (**BLUE**).*

This theorem only considers estimators that have these two properties of linearity and unbiasedness. Linearity means that the estimator–the rule for computing the estimates–can be written as a linear function of the data y (essentially, as a weighted average of the y values). OLS clearly meets this requirement. Under the assumptions above, OLS estimators are also unbiased. Given those properties, the proof of the Gauss-Markov theorem demonstrates that the OLS estimators have the minimum sampling variance of any possible estimator: that is, they are the "best" (most precise) that could possibly be calculated. This theorem is not based on the assumption that, for instance, the u process is Normally distributed; only that it is independent of the x variables and homoskedastic (that is, that it is $i.i.d.$).