Multiple Regression Analysis: Estimation

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Outline

Motivation.

Mechanics and Interpretation.

Expected Values and Variances of the Estimators.

Motivation for multiple regression

Consider the following results of a regression of the *number of crimes* reported in Milwaukee on the *search volume (on Google) for the term "ice cream"*

- which I'm using as a proxy for ice cream sales.
- A couple caveats about these data.

Dep. Var.: Crimes	**Coefficient ($\widehat{\beta}$)**	**Std. Err.**	**t**
Ice Cream	1.5437	.4306	3.58
N=82; R^2=0.1384			

Motivation for multiple regression (continued)

Ice cream is somehow an input into or a reason for committing crimes?

- Evidenced by its positive association with crime.

Common sense and past experiences with ice cream, however, argue that there is probably no causal connection between the two.

- A spurious correlation that
- disappears when we control for a confounding variable that is related to both x and y.

In this case summarized as “good weather”,

- i.e., people eat more ice cream when it is warm and also go outside more when it is warm (up to a point, at least).
- The increase in social interaction occasioned by warm weather, then, creates more opportunities for conflicts and crimes, according to this informal theory, while *coincidentally* leading to more ice cream sales, as well.

Motivation for multiple regression (concluded)

To use regression analysis to disconfirm the theory that ice cream causes more crime, perform a regression that controls for the effect of weather in some way.

Either,

- Examine sub-samples of days in which the weather is (roughly) the same but ice cream consumption varies, or
- Explicitly control for the weather by including it in a multiple regression model.

Multiple regression defined

Multiple regression expands the regression model using more than 1 regressor / explanatory variable / “independent variable”.

For 2 regressors, we would model the following relationship.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

and estimate separate effects (β_1 and β_2)—for each explanatory variable (x_1 and x_2).

Assuming enough data (and reasons for adding additional variables), the model with $k > 1$ regressors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

would not be anything fundamentally different from a simple regression.

Multiple OLS and simple OLS

Multiple regression relaxes the assumption that all other factors are held fixed.

Instead of, $E(error|x_1) = 0$, one needs only assume that $E(error|x_1, x_2 \ldots x_k) = 0$.

- A more realistic assumption when dealing with non-experimental data.

One other way of thinking about multiple regression:

- simple regression as a special case—in which $\{x_2, \ldots x_k\}$ have been relegated to the error term and treated as mean independent of x_1.
- I.e., simple regression has you estimating:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon; \varepsilon \equiv \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u.$$

Multiple OLS and simple OLS (continued)

To demonstrate why this could be a bad strategy, consider the ice cream–crime problem again, assuming that the true relationship follows:

$$crime = \beta_0 + \beta_1 icecream + \beta_2 temperature + u;$$
$$E(u|icecream, temperature) = 0.$$

My informal theory states that $\beta_1 = 0$ and $\beta_2 > 0$.

Multiple OLS and simple OLS (continued)

Estimating the simple regression between ice cream and crime, was as if the model was transformed:

$$crime = \beta_0 + \beta_1 icecream + \varepsilon;\ \varepsilon = \beta_2 temperature + u,$$

which I estimated under the assumption that: $E(\varepsilon|icecream) = 0$.

However this likely a bad assumption, since it is probable that:

$$E(temperature|icecream) \neq 0.$$

Multiple OLS and simple OLS (concluded)

High (low) values of *icecream* correspond with high (low) values of *temperature*.

The assumptions underlying simple regression analysis state that when the conditional mean independence is violated, bias is introduced in OLS.

- This bias would explain the positive estimate for β_1 shown above.

We will examine the source of the bias more closely and how to estimate its direction later in this chapter.

First we turn our attention back to the technical aspects of estimating the OLS parameters with multiple regressors.

Results, controlling for temperature

Now the coefficient on *ice cream* is much smaller (closer to 0) and the effect of temperature is large and significant.

- It's not precisely zero, but we can no longer reject the null that it has zero effect on crime.

Dep. Var.: Crimes	**Coefficient ($\widehat{\beta}$)**	**Std. Err.**	**t**
Ice Cream	0.4760	.4477	1.06
Temperature	1.9492	.4198	4.64
N=82; R^2=0.3231			

Mechanics and interpretation of OLS

Even with $k > 1$ regressors, OLS minimizes the sum of the squares of the residuals.

With $k = 2$ regressors:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \rightarrow \hat{u}_i^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right)^2.$$

$$Sum\ of\ Squared\ Residuals\ (SSR) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right)^2$$

Note is the change in subscripting: each observation is indexed by "i" as before, but the two "x" variables are now distinguished from one another by a "1" and a "2" in the subscript.

Minimizing SSR

For $k > 1$ regressors:

$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right)^2 .$$

Once again minimization is accomplished by differentiating the above line with respect to each of the $k + 1$ statistics,

- $\hat{\beta}_0$ and
- all the (*k*) slope parameters.

The first order conditions for the minimum are:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = 0 \text{ and } \frac{\partial SSR}{\partial \hat{\beta}_j} = 0 \; \forall \; j \in \{1,2,\ldots k\}.$$

Minimizing SSR (continued)

"Simultaneously choose all the 'betas' to make the regression model fit as closely as possible to all the data points."

The partial derivatives for the problem look like this:

$$(1)\frac{\partial SSR}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right)$$

$$(2)\frac{\partial SSR}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right)$$

$$\ldots(k+1)\frac{\partial SSR}{\partial \hat{\beta}_k} = -2\sum_{i=1}^{n} x_{ik}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right).$$

Example with $k = 2$

As a representative (and comparatively easy to solve) example, consider the solution when $k = 2$.

Setting (1) equal to zero (FOC) and solving for $\hat{\beta}_0$:

$$2n\hat{\beta}_0 - 2\sum_{i=1}^{n} y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + 2\hat{\beta}_1 \sum_{i=1}^{n} x_{i2} = 0$$

$$(4)\ \hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} - \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}\right] = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

Example with $k = 2$ (continued)

Doing the same thing with (2) and (3) gives you:

$$(2)\ \frac{\partial SSR}{\partial \hat{\beta}_1} = 0 \rightarrow \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 = \left[\sum_{i=1}^{n} x_{i1} y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} - \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2}\right]$$

$$(3)\ \frac{\partial SSR}{\partial \hat{\beta}_2} = 0 \rightarrow \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \left[\sum_{i=1}^{n} x_{i2} y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_{i2} - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2}\right]$$

Example with $k = 2$ (continued)

Substituting (4):

$$(2)\ \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 = \left[\sum_{i=1}^{n} x_{i1} y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2\right) n \bar{x}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2}\right]$$

$$(3)\ \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \left[\sum_{i=1}^{n} x_{i2} y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2\right) n \bar{x}_2 - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2}\right]$$

Example with $k = 2$ (continued)

Solving (3) for $\hat{\beta}_2$:

$$(3) \rightarrow \hat{\beta}_2 \left(\sum_{i=1}^{n} x_{i2}^2 - n\bar{x}_2^2 \right) = \left[\sum_{i=1}^{n} x_{i2} y_i - n\bar{x}_2 \bar{y} + n\bar{x}_2 \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2} \right]$$

$$\Leftrightarrow \hat{\beta}_2 = \left(\sum_{i=1}^{n} x_{i2}^2 - n\bar{x}_2^2 \right)^{-1} \left[\sum_{i=1}^{n} x_{i2} y_i - n\bar{x}_2 \bar{y} + n\bar{x}_2 \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2} \right]$$

$$\Leftrightarrow \hat{\beta}_2 = \big(Var(x_2)\big)^{-1} \big[Cov(x_2, y) - \hat{\beta}_1 \big(Cov(x_1, x_2) \big) \big]$$

Example with $k = 2$ (continued)

Solve it simultaneously with (2) by substituting it in:

$$(2) \rightarrow \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 = \left[Cov(x_1, y) + \hat{\beta}_1 n \bar{x}_1^2 - \hat{\beta}_2 \big(Cov(x_1, x_2)\big)\right]; \text{now}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 = Cov(x_1, y) + \hat{\beta}_1 n \bar{x}_1^2 - \frac{Cov(x_2, y)Cov(x_1, x_2) - \hat{\beta}_1 \big(Cov(x_1, x_2)\big)^2}{Var(x_2)}$$

Example with $k = 2$ (continued)

Collect all the $\hat{\beta}_1$ terms.

$$\hat{\beta}_1 \left[\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2 - \frac{\left(Cov(x_1, x_2)\right)^2}{Var(x_2)} \right] = Cov(x_1, y) - \frac{Cov(x_2, y)Cov(x_1, x_2)}{Var(x_2)}$$

$$\Leftrightarrow \hat{\beta}_1 \left[Var(x_1) - \frac{\left(Cov(x_1, x_2)\right)^2}{Var(x_2)} \right] = Cov(x_1, y) - \frac{Cov(x_2, y)Cov(x_1, x_2)}{Var(x_2)}$$

$$\Leftrightarrow \hat{\beta}_1 \left[\frac{Var(x_1)Var(x_2) - \left(Cov(x_1, x_2)\right)^2}{Var(x_2)} \right] = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_2)}$$

Example with $k = 2$ (concluded)

Simplifying gives you:

$$\hat{\beta}_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \big(Cov(x_1, x_2)\big)^2}$$

Then you can solve for $\hat{\beta}_2$ which, after a lot of simplification, is:

$$\hat{\beta}_2 = \frac{Cov(x_2, y)Var(x_1) - Cov(x_1, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \big(Cov(x_1, x_2)\big)^2}.$$

Lastly, substitute the above expressions into the solution for the intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2.$$

Multiple OLS and variance/covariance

Examine the solutions closely.

They depend, as with simple regression, only on the variances and covariances of the regressors and their covariances with y.

- This is true generally of OLS, even when there are many explanatory variables in the regression.

As this number grows, even to 2, the solution becomes difficult to work out with algebra, but software (like STATA) is very good at performing the calculations and solving for the $k + 1$ estimates, even when k is quite large.

- How software works out the solution: matrix algebra.

Multiple regression and "partialling out"

The "Partialling Out" Interpretation of Multiple Regression is revealed by the matrix and non-matrix estimate of $\hat{\beta}_1$.

- What goes into $\hat{\beta}_1$ in a multiple regression is the variation in x_1 that cannot be "explained" by its relation to the other x variables.
- The covariance between this residual variation in x_1, not explained by other regressors, and y is what matters.

It also provides a good way to think of $\hat{\beta}_1$ as the partial effect of x_1, holding other factors fixed.

It is estimated using variation in x_1 that is independent of variation in other regressors.

More.

Expected value of the OLS estimators

Now we resume the discussion of a misspecified OLS regression, e.g., *crime* on *ice cream,* by considering the assumptions under which OLS yields unbiased estimates and how misspecifying a regression can produce biased estimates.

- Also what is the direction of the bias?

There is a population model that is linear in parameters.

Assumption MLR.1 states this model which represents the true relationship between y and all x.

I.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u.$$

OLS under assumptions MLR 1-4

The sample of n observations is assumed random and indexed by i (MLR.2).

From simple regression, we know that there must be variation in x for an estimate to exist.

With multiple regression, each regressor must have (at least some) variation that is not explained by the other regressors.

- The other regressors cannot "partial out" all of the variation in, say, x_1 while still estimating β_1.
- Perfect multicollinearity is ruled out (by Assumption MLR.3).

When you have a set of regressors that violates this assumption, one of them must be removed from the regression.

- Software will usually do this automatically for you.

OLS under assumptions MLR 1-4 (continued)

Finally multiple regression assumes a less restrictive version of mean independence between regressors and error term.

Assumption MLR.4 states that the error term has zero conditional mean, i.e., conditional on all regressors.

$$E(u|x_1, x_2, \ldots, x_k) = 0.$$

This can fail if the estimated regression is misspecified, in terms of the functional form of one of the variables or the omission of a relevant regressor.

The model contains exogenous regressors, with the alternative being an endogenous regressor (explanatory variable).

- This is when some x_j is correlated with an omitted variable in the error term.

Unbiasedness

Under the four assumptions above, the OLS estimates are unbiased, i.e.,

$$E(\hat{\beta}_j) = \beta_j \forall\ j.$$

This includes cases in which variables have no effect on y and in which $\beta_j = 0$.

- This is the "overspecified" case in the text, in which an irrelevant variable is included in the regression.
- Though this does not affect the unbiased-ness of OLS, it does impact the variance of the estimates—sometimes in undesirable ways we will study later.

Expected value and variance of OLS estimators: under MLR assumptions 1-4

Unbiasedness does not apply when you *exclude* a variable from the population model when estimating it.

- This is called underspecifying the model.
- It occurs when, because of an empiricist's mistake or failure to observe a variable in data, a relevant regressor is excluded from the estimation.
- For the sake of concreteness, consider an example of each.

1. An empiricist regresses *crime* on *ice cream* because he has an axe to grind about ice cream or something like that, and he wants to show that it causes crime.
2. An empiricist regresses *earnings* on *education* because he cannot observe other determinants of workers' productivity ("ability") in data.

Bias from under-specification

So in both cases, there is a population ("true") model that includes two variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

but the empiricist estimates a version that excludes x_2, generating estimates different from unbiased OLS—which are denoted with the "tilde" on the next line.

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1;\ x_2 \text{ is omitted from the estimation. So,}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}.$$

Omitted variable bias

If the true relationship includes x_2, however, the bias in this estimate is revealed by substituting the population model for y_i.

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

Simplify the numerator:

$$\beta_0 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1) + \beta_1 \sum_{i=1}^{n} x_{i1}(x_{i1} - \bar{x}_1) + \beta_2 \sum_{i=1}^{n} x_{i2}(x_{i1} - \bar{x}_1) + \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i \,.$$

Then take expectations:

$$\Leftrightarrow E(\tilde{\beta}_1) = \frac{0 + \beta_1 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 + \beta_2 Cov(x_1, x_2) + 0}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2},$$

since the regressors are uncorrelated with the error from the correctly specified model.

Omitted variable bias (concluded)

A little more simplifying gives:

$$\Leftrightarrow E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}.$$

The estimator using the misspecified model is equal to the true parameter (β_1) plus a second term that captures the bias.

This bias is non-zero unless one of two things is true:

1. β_2 is zero. There is no bias from excluding an irrelevant variable.
2. $Cov(x_1, x_2)$ is zero. This means that x_1 is independent of the error term, varying it does not induce variation in an omitted variable.

Direction of omitted variable bias

The textbook calls the fraction in the bias term, $\hat{\delta}_1$,

- the regression coefficient you would get if you regressed x_2 on x_1.

$$\hat{\delta}_1 \equiv \frac{Cov(x_1, x_2)}{Var(x_1)}$$

This term, along with β_2, helps predict the direction of the bias.

Specifically, if $\hat{\delta}_1$ and β_2 are the same sign (+/-), the bias is positive, and if they are opposite signs, the bias is negative.

1. $\hat{\delta}_1, \beta_2$ same sign $\rightarrow E(\tilde{\beta}_1) > \beta_1$ "upward bias"
2. $\hat{\delta}_1, \beta_2$ opposite sign $\rightarrow E(\tilde{\beta}_1) < \beta_1$ "downward bias"

Omitted variable bias (concluded)

One more way to think about bias: the ceteris paribus effect of x_1 on y.

$\tilde{\beta}_1$ estimates this effect, but it does so by lumping the true effect in with an unknown quantity of misleading effects, e.g., you take the population model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \text{and estimate } y = \beta_0 + \beta_1 x_1 + v;$$
$$v \equiv \beta_2 x_2 + u. \text{ So,}$$

$$\frac{\partial E(y|x)}{\partial x_1} = \beta_1 + \frac{\partial v}{\partial x_1} = \beta_1 + \beta_2 \frac{\partial E(x_2|x_1)}{\partial x_1}.$$

The partial derivative of x_2 with respect to x_1 is the output you would get if you estimated a regression of x_2 on x_1, i.e., $\hat{\delta}_1$.

Standard error of the OLS estimator

Why is variance of an estimator ("standard error") so important?

- Inference: next chapter.

Generating and reporting only point estimates is irresponsible.

- It leads your (especially an uninitiated) audience to an overly specific conclusion about your results.
- This is bad for you (empiricist) and them (policy makers, customers, other scholars) because they may base decisions on your findings, e.g., choosing a marginal income tax rate depends crucially on labor supply elasticity.

A point estimate outside the context of its standard error may prompt your audience to rash decisions that they would not make if they knew your estimates were not "pinpoint" accurate.

- This is bad for you because if they do the "wrong" thing with your results, they will blame you for the policy's failure.
- And empiricists, as a group, will have their credibility diminished a little, as well.

Variance of OLS estimators

Here we show that the standard error of the j^{th} estimate, $\hat{\beta}_j$, is:

$$se(\hat{\beta}_j) \equiv Var(\hat{\beta}_j)^{\frac{1}{2}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{(n-k-1)(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}, \text{ where}$$

$\hat{u}_i$ are the residuals from the regression,

k is the number of regressors, and

R_j^2 is the proportion of $x_j's$ variance explained by the other x variables.

Once again the estimate of the variance assumes homoskedasticity, i.e.,

$$Var(u|all\ x_j) = Var(u) = \sigma^2.$$

Variance of OLS estimators (continued)

Most of the components of the variance are familiar from simple regression.

- Subscripts (j) and that the degrees of freedom ($n - k - 1$ instead of $n - 1 - 1$) have been generalized for k regressors.

The term, $(1 - R_j^2)$, is new, reflecting the consequence of (imperfect) multicollinearity.

Perfect multicollinearity has already been ruled out by Assumption MLR.3, but the standard error of a multicollinear variable ($R_j^2 \to 1$) is very large because the denominator of the expression above will be very small ($(1 - R_j^2) \to 0$).

Variance of OLS estimators (continued)

Once again the variance of the errors is estimated without bias, i.e., ($E(\hat{\sigma}^2) = \sigma^2$), by the (degrees of freedom-adjusted) mean squared error:

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2 \; ; \hat{\sigma}^2 = \frac{SSR}{n-k-1}, \text{so } Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2},$$

Where the residuals are: $\hat{u}_i \equiv y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$.

Variance of OLS estimators (continued)

Observations about the standard error.

- When adding a (relevant) regressor to a model, the standard error of an existing regressor can increase or decrease.
- Generally the sum of squares of residuals decreases, since more information is being used to fit the model. But the degrees of freedom decreases as well.
- The standard error depends on Assumption MLR.5 (homoskedasticity).
- If it is violated, $\hat{\beta}_j$ is still unbiased, however, the above expression becomes a biased estimate of the variance of $\hat{\beta}_j$.
- The 8th chapter in the text is devoted to dealing with the problem of heteroskedasticity in regression.

Gauss-Markov Theorem and BLUE

If, however, Assumptions MLR 1-5 hold, the OLS estimators ($\hat{\beta}_0$ through $\hat{\beta}_k$) have minimal variance among the class of linear unbiased estimators,

- i.e., it is the "best" in the sense of minimal variance.

OLS is "BLUE", which stands for Best Linear Unbiased Estimator.

The 5 Assumptions that lead to this conclusion (MLR 1-5) are collectively known as the Gauss-Markov assumptions.

The BLUE-ness of the OLS estimators under those assumptions is known as the Gauss-Markov Theorem.

Conclusion

Multiple regression solves the problem of omitted variables that are correlated with the dependent variable and regressor.

It is estimated in a manner analogous to simple OLS and has similar properties:

- Unbiasedness under MLR 1-4,
- BLUE under MLR 1-5.

Its estimates are interpreted as partial effects of changing one regressor and holding the others constant.

The estimates have standard errors that can be used to generate confidence intervals and test hypotheses about the parameters' values.

Data on crime and ice cream

The unit of observation is daily, i.e., these are 82 days on which ice cream search volume and crime reports have been counted in the City of Milwaukee.

All the variables have had the “day of the week”-specific mean subtracted from them, because all of them fluctuate across the calendar week, and this avoids creating another spurious correlation through the days of the week.

Back.

OLS estimates using matrices

Understanding what software is doing when it puts out regression estimates is one reason for the following exercise.

Another reason is to demonstrate the treatment of OLS you will experience in graduate-level econometrics classes—one in which everything is espoused using matrix algebra.

Imagine the variables in your regression as a spreadsheet, i.e., with y in column 1, x_1 in column 2, and each row is a different observation.

With this image in mind, divide the spreadsheet into two matrices,* on which operations may be performed as they are on numbers ("scalars" in matrix jargon).

*"A rectangular array of numbers enclosed in parentheses. It is conventionally denoted by a capital letter."

OLS estimates using matrices (continued)

The dimensions of a matrix are expressed: $rows \times columns$.

- I.e., n observations (rows) of a variable y, would be an $n \times 1$ matrix ("Y"); n observations (rows) of k explanatory variables x, would be an $n \times k$ matrix ("X").

One operation that can be performed on a matrix is called transposition, denoted by an apostrophe behind the matrix's name.

- Transposing a matrix just means exchanging its rows and columns. E.g.,

$$X \equiv \begin{bmatrix} 5 & 2 & 1 \\ 3 & 0 & 1 \end{bmatrix} \rightarrow X' = \begin{bmatrix} 5 & 3 \\ 2 & 0 \\ 1 & 1 \end{bmatrix}.$$

Matrix multiplication

Another operation that can be performed on some pairs of matrices is multiplication, but only if they are conformable.

"Two matrices A and B of dimensions $m \times n$ and $n \times q$ respectively are conformable to form the product matrix AB, since the number of columns in A is equal to the number of rows in B. The product matrix AB is of dimension $m \times q$, and its ij^{th} element, c_{ij}, is obtained by multiplying the elements of the i^{th} row of A by the corresponding elements of the j^{th} column of B and adding the resulting products."

OLS estimates using matrices (continued)

Transposition and multiplication are useful in econometrics for obtaining a matrix of covariances between y and all the x, as well as among the x variables.

Specifically when you multiply Y by the transpose of X, you get a matrix ($k \times 1$) with elements,

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i2} y_i \\ \vdots \\ \sum_{i=1}^{n} x_{k2} y_i \end{bmatrix} = \begin{bmatrix} S_{y1} \\ S_{y2} \\ \vdots \\ S_{yk} \end{bmatrix};$$

x and *y* are expressed as deviations from their means.

OLS estimates using matrices (continued)

Similarly multiplying X by its own transpose gives you a matrix ($k \times k$) with elements,

$$X'X = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{i1}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1k} \\ \vdots & \ddots & \vdots \\ S_{k1} & \cdots & S_{kk} \end{bmatrix}.$$

OLS estimates using matrices (continued)

X'X and X'Y are matrices of the variances of the explanatory variables and covariances among all variables. They can be combined to obtain the matrix $(k \times 1)$ of coefficient estimates.

- This is the part for which computers are really indispensable.

First of all, you invert the matrix $X'X$. This means solving for the values that make the following hold:

$$(X'X)(X'X)^{-1} = I \Leftrightarrow \begin{bmatrix} s_{11} & \cdots & s_{1k} \\ \vdots & \ddots & \vdots \\ s_{k1} & \cdots & s_{kk} \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

OLS estimates using matrices (continued)

To illustrate using the case of $k = 2$, you have to solve for all the "a" terms using simultaneous equations.

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{so}$$

$$a_{11}s_{11} + a_{21}s_{12} = 1,$$
$$a_{11}s_{21} + a_{21}s_{22} = 0,$$
$$a_{12}s_{11} + a_{22}s_{12} = 0, \text{and}$$
$$a_{12}s_{21} + a_{22}s_{22} = 1.$$

OLS estimates using matrices (continued)

Solving the first two simultaneously gives you:

$$a_{11} = \frac{s_{22}}{s_{11}s_{22} - s_{12}s_{21}} \text{ and } a_{21} = \frac{s_{21}}{s_{12}s_{21} - s_{11}s_{22}}.$$

Similarly the last to give you:

$$a_{12} = \frac{s_{12}}{s_{12}s_{21} - s_{11}s_{22}} \text{ and } a_{22} = \frac{s_{11}}{s_{11}s_{22} - s_{12}s_{21}}.$$

To get the coefficient estimates, you do the matrix equivalent of dividing covariance by variance, i.e., you solve for the vector B:

$$\mathrm{B} = (X'X)^{-1}(X'Y).$$

OLS estimates using matrices (continued)

For the $k = 2$ case,

$$\mathrm{B} = \begin{bmatrix} \dfrac{s_{22}}{s_{11}s_{22} - s_{12}s_{21}} & \dfrac{s_{12}}{s_{12}s_{21} - s_{11}s_{22}} \\ \dfrac{s_{21}}{s_{12}s_{21} - s_{11}s_{22}} & \dfrac{s_{11}}{s_{11}s_{22} - s_{12}s_{21}} \end{bmatrix} \begin{bmatrix} s_{y1} \\ s_{y2} \end{bmatrix}.$$

Multiplying this out gives the same solution as minimizing the Sum of Squared Residuals.

$$\mathrm{B} = \begin{bmatrix} \dfrac{s_{22}s_{y1} - s_{12}s_{y2}}{s_{11}s_{22} - s_{12}s_{21}} \\ \dfrac{s_{11}s_{y2} - s_{21}s_{y1}}{s_{11}s_{22} - s_{12}s_{21}} \end{bmatrix} = \begin{bmatrix} \dfrac{Var(x_2)Cov(x_1, y) - Cov(x_1, x_2)Cov(x_2, y)}{Var(x_1)Var(x_2) - [Cov(x_1, x_2)]^2} \\ \dfrac{Var(x_1)Cov(x_2, y) - Cov(x_1, x_2)Cov(x_1, y)}{Var(x_1)Var(x_2) - [Cov(x_1, x_2)]^2} \end{bmatrix}$$

OLS estimates using matrices (concluded)

Solving OLS for $k > 2$ is not fundamentally different, and the solution will always be a function of the variances and covariances of the variables in the model.

- It just gets more difficult (at an accelerating rate) to write down solutions for the individual coefficients.
- It is left as an exercise to show that the "partialling out" interpretation of OLS holds.

$$\mathrm{B} = (X'X)^{-1}(X'Y) = \begin{bmatrix} \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2} \\ \vdots \\ \frac{\sum_{i=1}^{n} \hat{r}_{ik} y_i}{\sum_{i=1}^{n} \hat{r}_{k1}^2} \end{bmatrix},$$

in which $\hat{r}_{ij}$ is the residual for observation i from a regression of x_j on all the other regressors.

- $\hat{r}_{i1} \equiv x_{i1} - \hat{x}_{i1}$; $\hat{x}_{i1}$ is the fitted value.

Back.

Partialling out interpretation of OLS

The expression for $\hat{\beta}_1$ can be derived from the first order condition for x_1 from the least squares minimization.

This states that:

$$\frac{\partial SSR}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right) = 0.$$

Substituting in the definition of the residuals from the previous line, you have:

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right] = 0$$

Partialling out interpretation of OLS (continued)

$$\Leftrightarrow \sum_{i=1}^{n} \hat{x}_{i1}\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right] +$$

$$\sum_{i=1}^{n} \hat{r}_{i1}\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right] = 0.$$

The term in brackets is the OLS residual from regressing y on $\{x_1 \ldots x_k\}$, "$\hat{u}_i$", and $\hat{x}_1$ is a linear function of the other x variables, i.e.,

$$\hat{x}_{i1} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i2} + \ldots + \hat{\gamma}_k x_{ik},$$

. . .

Partialling out interpretation of OLS (continued)

. . . so one may write:

$$\sum_{i=1}^{n} \hat{x}_{i1}\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right] = \sum_{i=1}^{n} (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i2} + \ldots + \hat{\gamma}_k x_{ik})\hat{u}_i \,.$$

Since the OLS residuals are uncorrelated with each x_j, and the sum of the residuals is zero, this entire line equals zero and drops out.

Partialling out interpretation of OLS (continued)

Then you have:

$$\sum_{i=1}^{n} \hat{r}_{i1}\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}\right] = 0, \text{and}$$

the residuals ($\hat{r}_1$) are uncorrelated with the other regressors (2 through k) and sum to zero, as well. So,

$$\sum_{i=1}^{n} \hat{r}_{i1} y_i = \sum_{i=1}^{n} \hat{\beta}_1 (\hat{r}_{i1}^2 + \hat{r}_{i1}\hat{x}_{i1}) \Leftrightarrow \sum_{i=1}^{n} \hat{r}_{i1} y_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{r}_{i1}{}^2 .$$

Partialling out interpretation of OLS (concluded)

Then you get the "partialling out" expression:

$$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}{}^2}.$$

Back.