M1 INTERMEDIATE ECONOMETRICS

LINEAR REGRESSION MODEL

Koen Jochmans

August 10, 2025

1. Conditional expectation function

We are interested in the relationship between a scalar dependent variable (also referred to as outcome variable) Y and a dependent variable (also called regressor) X that, in this section, we also take to be a scalar. A useful starting point is the conditional expectation function (CEF) which, as a function of x, is

$$m(x) = \mathbb{E}(Y|X = x).$$

It is the mean of the conditional distribution of Y when X is held fixed at the value x.

1.1. Binary independent variable

Suppose, first, that $X \in \{0, 1\}$. As an example one can think about Y as the wage and X as a binary measure of education. Say that $X = 0$ corresponds to low education and $X = 1$ to high education. Then $\mathbb{E}(Y|X = 0)$ is the expected wage given low education $\mathbb{E}(Y|X = 1)$, is the expected wage given high education.

As X can take on only two values the CEF is linear. First, we can write

$$\mathbb{E}(Y|X = x) = \beta_1(1 - x) + \beta_2 x = \beta_1 + (\beta_2 - \beta_1)\, x.$$

Then $\beta_1 = \mathbb{E}(Y|X = 0)$ and $\beta_2 = \mathbb{E}(Y|X = 1)$ and so the coefficients β_1 and

β_2 correspond to the expected wage level for the different values that X can take. A second parametrization is

$$\mathbb{E}(Y|X=x) = \beta_1 + \beta_2 x,$$

in which case $\beta_1 = \mathbb{E}(Y|X=0)$ and

$$\beta_2 = \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0).$$

In this parametrization, the expected wage for the low-educated is chosen as baseline and β_2 gives the expected wage differential from being highly educated relative to low educated. That is, β_2 captures an average marginal effect.

Outside the simple binary case, $m(x)$ will not be linear in x, in general.

1.2. Discrete (countable) independent variable

If X can take on a countable number of values, say $X \in \{1, \ldots, m\}$, it is always possible to parametrize the CEF so that it is linear in transformed independent variables. Define the m dummy variables $T_1(X), \ldots, T_m(X)$ via

$$T_x(X) = \begin{cases} 1 & \text{if } X = x \\ 0 & \text{otherwise} \end{cases},$$

for $x = 1, 2, \ldots, m$. Then

$$\mathbb{E}(Y|X=x) = \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots + \beta_m T_m(x) = \beta_x$$

is linear in the m dummies. This is called a saturated specification of the CEF. It is a direct generalization of our first specification in the previous subsection. Note that the CEF is not linear in the original independent variable, however. The difference

$$\beta_{x_2} - \beta_{x_1} = \mathbb{E}(Y|X = x_2) - \mathbb{E}(Y|X = x_1)$$

for any two values x_1 and x_2 gives the expected change in Y when X changes from x_1 to x_2. In our wage illustration we could think about X as the years of education (or experience, or tenure, for example), so that $\beta_{x_2} - \beta_{x_1}$ would yield the expected returns to education from following x_2 years of education compared to x_1. This can different for any pair (x_1, x_2). In contrast, if the CEF would be of the form $m(x) = x\beta$ then an increase in X by one unit would amount to the same expected change in Y no matter the initial level of X. In our example, this would mean that the returns to an additional year of education are constant.

1.3. Continuous independent variable

When $X \in \mathbb{R}$ the average marginal effect at x is the derivative of the CEF at x, that is,

$$m'(x) = \frac{\partial m(x)}{\partial x} = \frac{\partial\, \mathbb{E}(Y|X = x)}{\partial x},$$

which changes with x unless m' is constant, in which case m must be a linear function. This will not be true in general. An exception is the case where (Y, X) are jointly normally distributed.

The CEF admits, under regularity conditions, a series representation as

$$m(x) = \sum_{i=1}^{+\infty} \beta_i\, T_i(x)$$

where the $T_1(X), T_2(X), \ldots$ are functions such as orthogonal polynomials or power series.

1.4. Best predictor

Consider the problem of predicting the value of Y based on a realization of X. The choice of predictor amounts to the choice of a function $\tilde{m}$ that maps X to a prediction about Y. The error made by predictor $\tilde{m}$ is the difference

$$Y - \tilde{m}(X).$$

To decide what is a good predictor ex ante we need to decide how we penalize errors and then integrate over the distribution of errors. A popular approach is based on expected squared loss. This amounts to choosing $\tilde{m}$ as to minimize

$$\mathbb{E}((Y - \tilde{m}(X))^2).$$

The solution is to set $\tilde{m} = m$, the CEF. To see this let

$$e = Y - m(X)$$

and note that, by definition,

$$\mathbb{E}(e|X = x) = 0.$$

This implies, in particular, that e is orthogonal to any transformation of X, that is,

$$\mathbb{E}(g(X)e) = \mathbb{E}(g(X)\mathbb{E}(e|X)) = 0.$$

Indeed,

$$\mathbb{E}((Y-\tilde{m}(X))^2)=\mathbb{E}((e+m(X)-\tilde{m}(X))^2)=\mathbb{E}(e^2)+\mathbb{E}((m(X)-\tilde{m}(X))^2)$$

which clearly reaches its minimum of $\mathbb{E}(e^2)$ at the CEF.

2. Simple linear regression

2.1. Linear approximation to the CEF

When the CEF m is nonlinear we may still consider a linear approximation to it. A simple linear approximation is $\beta_1+\beta_2 x$ for chosen coefficients (β_1,β_2). We can again choose these coefficients by minimizing an expected squared loss objective,

$$\min_{b_1,b_2}\mathbb{E}((m(X)-(b_1+b_2X))^2).$$

Let $e(b_1,b_2)=Y-(b_1+b_2X)$. The first-order conditions to this problem are

$$\mathbb{E}(\quad e(b_1,b_2))=0,$$

$$\mathbb{E}(Xe(b_1,b_2))=0.$$

So the coefficients (β_1,β_2) have to be chosen so that the error has mean zero and is uncorrelated to the regressor. The solution is

$$\beta_2=\frac{\mathbb{E}(m(X)X)-\mathbb{E}(m(X))\mathbb{E}(X)}{\mathbb{E}(X^2)-\mathbb{E}(X)^2}=\frac{\text{cov}(m(X),X)}{\text{var}(X)}=\frac{\text{cov}(Y,X)}{\text{var}(X)}$$

and $\beta_1=\mathbb{E}(m(X))-\beta_2\mathbb{E}(X)=\mathbb{E}(Y)-\beta_2\mathbb{E}(X)$. Thus, we can always write

$$Y=\beta_1+\beta_2X+e$$

where $e=e(\beta_1,\beta_2)$ is such that $\mathbb{E}(e)=0$ and $\mathbb{E}(Xe)=0$. Note that this error is only orthogonal to X, and so, contrary to the deviation of Y from its CEF, not mean-independent of it. Of course, when the CEF is linear, its best linear approximation is equal to the function itself, and both notions of error coincide.

2.2. Regression to the mean

As is clear from its expression, the intercept term β_1 accounts for the fact that Y and or X may have non-zero mean. A regression of Y on X without a constant term amounts to choosing the single slope coefficient to satisfy $\mathbb{E}(Xe) = 0$, but not $\mathbb{E}(e) = 0$. The resulting coefficient is $\mathbb{E}(XY)/\mathbb{E}(X^2)$. This co-incides with β_2 above only when both $\mathbb{E}(Y) = 0$ and $\mathbb{E}(X) = 0$ hold.

2.3. Relation to correlation coefficient

The linear regression coefficient β_2 is related to the correlation coefficient, which is

$$\operatorname{corr}(Y, X) = \frac{\operatorname{cov}(Y, X)}{\operatorname{std}(Y) \operatorname{std}(X)}.$$

The latter is unit-less. It lies between -1 and 1 and measures the strength of linear association between the two variables. We have

$$\beta_2 = \frac{\operatorname{cov}(Y, X)}{\operatorname{var}(X)} = \operatorname{corr}(Y, X) \frac{\operatorname{std}(Y)}{\operatorname{std}(X)},$$

which is not unit-less. Moreover, the interpretation of β_2 is that, when X changes by Δx units of X, we predict a change in Y of $\Delta y = \beta_2 \Delta x$ units of Y. If X captures the number of years of schooling and Y is yearly wage in euro. Then we predict an increase of β_2 euro in yearly wage for every additional year of schooling.

2.4. A generalization

The simple regression can equally be performed on a transformation of the original independent variable. For example, if $m(x)$ is concave in x a sensible simple regression would be of the form

$$Y = \beta_1 + \beta_2 \log(X) + e.$$

Other transformation are clearly also possible. In the linear-log specification a Δx change in x leads to a change in our prediction of $\Delta y = \beta_2 \Delta x/x$. Here, $\Delta x/x$ is a relative change in the regressor value. So a 1% increase corresponds to a change of $\Delta y = \beta_2/100$ units of the outcome variable.

An alternative way to capture some of the nonlinearity in the CEF when fully saturating is not possible would be to consider a predictor of the form

$$\beta_1 + \beta_2 X + \beta_3 X^2 + \cdots + \beta_k X^{k-1}$$

for some integer k. This is a truncated power series. For example, when $k = 3$ we obtain $\beta_1 + \beta_2 X + \beta_3 X^2$ which can capture linear, convex, and concave functions. The approximation becomes increasingly flexible as k grows. The power series is nonlinear in the original independent variable but linear in the coefficient vector. It generalizes the simple regression, which corresponds to the case $k = 1$, and is a special case of multiple regression, which we turn to next.

3. Multiple regression

Now consider the case where we have multiple independent variables. Say that $X = (X_1, \ldots, X_k)'$ is now a $k \times 1$ (column) vector. One of the regressors, usually the first one, can be taken to be a constant term, i.e., $\mathbb{P}(X_1 = 1) = 1$.

3.1. CEF and best linear predictor

The concept of CEF and (best) linear predictor are unchanged. The CEF is now a vector-valued function. In particular, m is still the best predictor under expected squared loss. While it may still be possible to write down saturated specifications when all independent variables can take on a countable number of values, the number of parameters needed to do so becomes large very quickly. This will cause complications when turning to estimation later on. We may again consider a linear approximation to the CEF, now of the form

$$X'\beta = X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k$$

or, indeed, in transformations of the original vector X, the important part is that the approximation is linear in the coefficient vector β. This vector solves

$$\min_b \mathbb{E}((m(X) - X'b)^2) = \min_b \mathbb{E}((Y - X'b)^2).$$

The first-order condition to this problem now is the set of k normal equations

$$\mathbb{E}(X(Y - X'b)) = 0.$$

Provided that the $k \times k$ matrix $\mathbb{E}(XX')$ is full rank, the unique solution is equal to

$$\beta = \mathbb{E}(XX')^{-1}\mathbb{E}(XY).$$

When the rank condition fails there exist multiple solutions to the normal equations. This is usually referred to as a the multicolinearity problem. For prediction purposes this multitude of solutions is inconsequential. The minimal-norm solution is $\mathbb{E}(XX')^*\mathbb{E}(XY)$. where * denote the Moore-Penrose pseudo inverse.

3.2. Short and long regressions, and partitioned regression

Suppose that X is bivariate, that is, $X = (X_1, X_2)'$. The long regression is the regression of Y on X,

$$Y = X_1\beta_1 + x_2\beta_2 + e$$

where e is defined through the two orthogonality conditions $\mathbb{E}(X_1e) = 0$ and $\mathbb{E}(X_2e) = 0$. The short regression is the regression of Y on X_1 alone. Its coefficient is equal to

$$\frac{\mathbb{E}(X_1Y)}{\mathbb{E}(X_1^2)} = \frac{\mathbb{E}(X_1(X_1\beta_1 + X_2\beta_2 + e))}{\mathbb{E}(X_1^2)} = \beta_1 + \frac{\mathbb{E}(X_1X_2)}{\mathbb{E}(X_1^2)}\beta_2.$$

The short and long regressions give the same coefficient only if it holds that $\mathbb{E}(X_1X_2) = 0$; when $\beta_2 = 0$ the short and long regression are the same specification. The short regression can lead to an over- or undervaluation of the impact of X_1, depending on the sign and magnitude of $\mathbb{E}(X_1X_2)/\mathbb{E}(X_1^2)$ and of β_2. In a regression of wages on education we would, for example, want to control for other factors that affect the wage and correlate with the education level, such as tenure, experience, and location, among other things.

Reversely, the long regression coefficient is

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1X_1) & \mathbb{E}(X_1X_2) \\ \mathbb{E}(X_1X_2) & \mathbb{E}(X_2X_2) \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}(X_1Y) \\ \mathbb{E}(X_2Y) \end{pmatrix}.$$

The matrix inverse on the right-hand side of this equation takes the simple form

$$\frac{1}{\mathbb{E}(X_1^2)\mathbb{E}(X_2^2)-\mathbb{E}(X_1X_2)^2}\begin{pmatrix} \mathbb{E}(X_2^2) & -\mathbb{E}(X_1X_2) \\ -\mathbb{E}(X_1X_2) & \mathbb{E}(X_1^2) \end{pmatrix}.$$

Hence,

$$\beta_1 = \frac{\mathbb{E}(X_2^2)\mathbb{E}(X_1Y)-\mathbb{E}(X_1X_2)\mathbb{E}(X_2Y)}{\mathbb{E}(X_1^2)\mathbb{E}(X_2^2)-\mathbb{E}(X_1X_2)^2}.$$

It is useful to inspect this formula in some detail.

By linearity of the expectations operator, the numerator in this expression can be written as

$$\mathbb{E}(X_2^2)\,\mathbb{E}\Bigg(X_1\Bigg(Y-\frac{\mathbb{E}(X_2Y)}{\mathbb{E}(X_2^2)}X_2\Bigg)\Bigg) = \mathbb{E}(X_2^2)\,\mathbb{E}(X_1Y^{\perp}),$$

where

$$Y^{\perp} = Y - \frac{\mathbb{E}(X_2Y)}{\mathbb{E}(X_2^2)}X_2$$

is the prediction error of a regression of Y on X_2 alone. Moreover, because $\mathbb{E}(X_2Y^{\perp})=0$ by construction, $Y^{\perp}$ is the part of Y that is orthogonal to X_2. It is what remains of Y once its linear dependence on X_2 has been filtered out. We can write this decomposition as

$$Y = Y^* + Y^{\perp}, \qquad Y^* = \frac{\mathbb{E}(X_2Y)}{\mathbb{E}(X_2^2)}X_2,$$

where $\mathbb{E}(Y^*Y^{\perp})=0$.

In the same way, the denominator equals,

$$\mathbb{E}(X_2^2)\,\mathbb{E}\left(X_1\left(X_1-\frac{\mathbb{E}(X_2X_1)}{\mathbb{E}(X_2^2)}X_2\right)\right) = \mathbb{E}(X_2^2)\,\mathbb{E}(X_1X_1^{\perp})$$

where, now,

$$X_1 = X_1^* + X_1^{\perp}, \qquad X_1^* = \frac{\mathbb{E}(X_2X_1)}{\mathbb{E}(X_2^2)}X_2$$

and $\mathbb{E}(X_1^* X_1^\perp) = 0$, so that $X_1^\perp$ is the part of X_1 that is orthogonal to X_2.

Plugging these expressions into the formula for the coefficient on X_1 in the long regression gives

$$\beta_1 = \frac{\mathbb{E}(X_1 Y^\perp)}{\mathbb{E}(X_1 X_1^\perp)} = \frac{\mathbb{E}(X_1^\perp Y\ \ )}{\mathbb{E}(X_1^\perp X_1^\perp)},$$

which is the slope in a simple regression of Y on $X_1^\perp$. Here, the last transition follows from noting that

$$\mathbb{E}(X_1 X_1^\perp) = \mathbb{E}(X_1^* X_1^\perp) + \mathbb{E}(X_1^\perp X_1^\perp) = \mathbb{E}(X_1^\perp X_1^\perp),$$

because $\mathbb{E}(X_1^* X_1^\perp) = 0$, and

$$\begin{aligned} \mathbb{E}(X_1 Y^\perp) &= \mathbb{E}(X_1^* Y^\perp) + \mathbb{E}(X_1^\perp Y^\perp) \\ &= \mathbb{E}(X_1^\perp Y^\perp) \\ &= \mathbb{E}(X_1^\perp Y) - \mathbb{E}(X_1^\perp Y^*) \\ &= \mathbb{E}(X_1^\perp Y), \end{aligned}$$

again using that $\mathbb{E}(X_1^* Y^\perp) = 0$ and that $\mathbb{E}(X_1^\perp Y^*) = 0$.

The above goes through for general partitions of the vector X into two parts. This observation is known as the Frisch-Waugh-Lovell theorem. The coefficient on a single regressor in a multiple regression problem behaves like a partial correlation coefficient. To see this, for $X = (X_1, \dots, X_k)'$, let $X_{-1} = (X_2, \dots, X_k)'$. Then, writing

$$X_1 = X_1^* + X_1^\perp = X_{-1}'\, \mathbb{E}(X_{-1} X_{-1}')^{-1} \mathbb{E}(X_{-1} X_1) + X_1^\perp$$

where $X_1^\perp$ is defined through the condition $\mathbb{E}(X_{-1} X_1^\perp) = 0$, it continues to hold that

$$\beta_1 = \frac{\mathbb{E}(X_1^{\perp} Y \ \ )}{\mathbb{E}(X_1^{\perp} X_1^{\perp})}.$$

Compared to the simple linear regression the interpretation of a coefficient in the multiple regression model is in a ceteris paribus sense. For example, we predict a change of $\Delta y = \Delta x_1 \beta_1$ units in Y due to a change of Δx_1 units in X_1 while holding the other included regressors fixed at their observed values.