This material comes from Hansen Chapter 2.

Linear Regression Notes 2: Conditional Expectation and Projection

Conditional expectation and projection

H: 2.5

We will spend much time this semester thinking about **conditional expectations**, that is, $\mathbb{E}[Y|X = x]$. This is the mean of Y conditional on X taking the particular value x. The book sometimes uses the short-hand notation $m(x) := \mathbb{E}[Y|X = x]$. You can think of $\mathbb{E}[Y|X = x]$ as a function — that is, when you plug in a new value of x, the value of the conditional expectation function can change. For example, if Y is a person's earnings and X is their years of education, the $\mathbb{E}[Y|X = 12]$ could differ (perhaps substantially) from, say, $\mathbb{E}[Y|X = 16]$. Sometimes it will be useful to view $\mathbb{E}[Y|X]$ as a function of the random variable X; in this case, $\mathbb{E}[Y|X]$ is itself random (because X is random). This is different from when you plug in a particular value of x, $\mathbb{E}[Y|X = x]$ is no longer random; it is equal to some number (though, in most cases, it is unlikely that we know the value of this sort of population quantity).

Law of iterated expectations

H: 2.7

One of the most useful tools (that you likely will be familiar with already) is the law of iterated expectations. We'll provide a simple version and a general version (both require the regularity condition that $\mathbb{E}|Y| < \infty$). The simple version is

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

In words, this says that the expectation of the conditional expectation is equal to the unconditional expectation. Here is an example. Continue to suppose that Y is a person's earnings and X is their years of education. $\mathbb{E}[Y|X = x]$ can vary arbitrarily for different values of x. But if you know $\mathbb{E}[Y|X = x]$ for all possible values of x, this will pin down the value of $\mathbb{E}[Y]$ (i.e., if you know mean earnings for all years of education, then you also should be able to recover the overall mean value of earnings). And, in particular, the law of iterated expectations says that the overall mean is equal to the mean of the conditional expectations (i.e., it puts more "weight'' on conditional expectations of relatively common values of X).

A more general version of the law of iterated expectations is the following: for any two random vectors X_1 and X_2,

$$\mathbb{E}[Y|X_1] = \mathbb{E}[\mathbb{E}[Y|X_1, X_2]|X_1]$$

The inside expectation is an expectation of Y conditional on X_1 and X_2, the outside expectation is over the distribution of X_2 conditional on X_1.

CEF Error

H: 2.8

We define the CEF error as the difference between Y and $\mathbb{E}[Y|X]$. That is,

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

Rearranging terms also implies the following expression that we will use often

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

That is, the actual outcome Y is equal to the CEF plus the CEF error. Notice that (besides regularity conditions that expectations exist), we are not making any assumptions here — we can essentially always write that Y is equal to its conditional expection plus CEF error.

A fundamental property of the CEF error is that $\mathbb{E}[e|X] = 0$. Let us show why this holds

$$\begin{aligned}
\mathbb{E}[e|X] &= \mathbb{E}[Y - m(X)|X] \\
&= \mathbb{E}[Y|X] - \mathbb{E}[m(X)|X] \\
&= m(X) - m(X) \\
&= 0
\end{aligned}$$

From the law of iterated expectations, this also implies that

$$\mathbb{E}[e] = \mathbb{E}\big[\underbrace{\mathbb{E}[e|X]}_{=0}\big] = 0$$

The condition that $\mathbb{E}[e|X] = 0$ is called **mean independence**. It is weaker than "full'' independence. To give an example, even if $\mathbb{E}[e|X] = 0$, it could be the case that $\mathbb{E}[e^2|X]$ varies with X which would imply that e and X are not independent.

We can also define the variance of the CEF error as

$$\sigma^2 = \text{var}(e) = \mathbb{E}[(e - \mathbb{E}[e])^2] = \mathbb{E}[e^2]$$

σ^2 measures the amount of variation in Y that is not accounted for by $\mathbb{E}[Y|X]$.

Best Predictor

H: 2.11

Next, we will show that the conditional expectation function is the "best' ' predictor of Y given X. We can write any predictor as a function $g(X)$; i.e., a function that takes values of X and makes predictions about what the outcome will be. In order to evaluate how well a predictor makes predictions, we need some criteria. The most common criteria is **mean squared prediction error**. This is given by

$$\mathbb{E}[(Y - g(X))^2]$$

The squared difference between Y and its prediction $g(X)$ is a measure of the distance between Y and $g(X)$ — in particular, it is always non-negative and gets larger as when Y and $g(X)$ are further away from each other. The outside expectation averages this distance over the distribution of Y and X.

Next, we show that $m(X)$ minimizes mean squared prediction error. To this end, notice that

$$\begin{aligned}
\mathbb{E}[(Y - g(X))^2] &= \mathbb{E}\Big[\big((Y - m(X)) + (m(X) - g(X))\big)^2\Big] \\
&= \mathbb{E}\Big[\big(e + (m(X) - g(X))\big)^2\Big] \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2]
\end{aligned}$$

where the first equality holds by adding and subtracting $m(X)$, the second equality holds by the definition of e, and the last equality holds by squaring the main term and pushing the expectation through the sum. Now, let's consider each term. First, $\mathbb{E}[e^2] = \sigma^2$ which does not depend on our prediction function $g(X)$. Next, consider the middle term

$$\begin{aligned}
\mathbb{E}[e(m(X) - g(X))] &= \mathbb{E}[(m(X) - g(X))\mathbb{E}[e|X]] \\
&= 0
\end{aligned}$$

where the first equality holds by the law of iterated expectations and the second equality holds because $\mathbb{E}[e|X] = 0$. Finally, the third term is minimized at 0 by setting $g(X) = m(X)$. This implies that setting $g(X) = m(X)$ minimizes mean squared prediction error.

Marginal Contrasts

H 2.14

Following the textbook, we'll use the shorthand notation $m(x) := \mathbb{E}[Y|X = x]$. We will often be interested in comparing average outcomes across observations whose value of one regressor, say X_1, differs but that have the same values of the other regressors. This is what we'll call the **marginal**

contrast for X_1 (the textbook calls this the **regression derivative**). When X_1 is continuously distributed, it is given by

$$MC_1(x) := \frac{\partial\, \mathbb{E}[Y|X = x]}{\partial\, x_1}$$

When X_1 is discrete, the marginal contrast is given by

$$MC_1(x) := \mathbb{E}[Y|X_1 = x_1 + 1, X_2 = x_2, \dots, X_k = x_k] - \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \dots, X_k = x_k]$$

and, in the leading case where X_1 is binary (so it only takes the values 0 or 1), this is $\mathbb{E}[Y|X_1 = 1, X_2 = x_2, \dots, X_k = x_k] - \mathbb{E}[Y|X_1 = 0, X_2 = x_2, \dots, X_k = x_k]$. There is nothing special about X_1, and we can likewise define marginal contrasts for $X_2, \dots, X_k$; for example, $MC_2(x)$ is the marginal contrast for X_2.

The marginal contrast should be interpreted as: comparing observations that have the same values of $X_2, \dots, X_k$, those whose X_1 is one unit higher have, on average, an outcome that is $MC_1(x)$ higher. Notice that this interpretation compares *different* observations; it does not say that increasing X_1 for a particular observation would change its outcome by $MC_1(x)$.

Side-Comment: Marginal contrasts are often called "partial effects" or "marginal effects." I use "contrast" because the word "effect" makes it sound like X_1 *causes* Y to change, which we will not be able to claim without additional assumptions. We will come back to when marginal contrasts have a causal interpretation later in the course.

Example: Let Y be a person's earnings, X_1 be a binary variable indicating whether or not they graduated from college, and X_2 be their age, and suppose that

$$m(x_1, x_2) = 20{,}000 + 15{,}000x_1 + 1{,}000x_2 + 200x_1x_2$$

Then, the marginal contrast for college is

$$MC_1(x_2) = m(1, x_2) - m(0, x_2) = 15{,}000 + 200x_2$$

For example, $MC_1(30) = 21{,}000$. This means that, comparing people who are 30 years old, college graduates earn, on average, \$21,000 more than non-college graduates. Among 50 year olds, the difference is \$25,000. Similarly, the marginal contrast for age is $MC_2(x_1) = 1{,}000 + 200x_1$: among college graduates, people who are one year older earn, on average, \$1,200 more, while, among non-college graduates, the difference is \$1,000.
Notice that $MC_1(30) = 21{,}000$ is *not* the same as saying that going to college would increase a particular 30 year old's earnings by \$21,000. College graduates and non-college graduates may differ in other ways (e.g., ability or family background) that are also related to earnings.

Some Comments

- First, marginal contrasts hold the other regressors constant. But they do not hold other variables that are not in the model constant (e.g., ability in the example above).
- Second, you should notice that $MC_1(x)$ is a function of x. If you plug in different values of x, then the value of this function could change (as in the example above, where the marginal contrast for college depends on age).
- Third, marginal contrasts are about averages. Even among people with the same age, the difference in earnings between college graduates and non-college graduates varies (perhaps tremendously) across people; the marginal contrast compares the *average* earnings of the two groups.

Linear CEF

H: 2.15

An important special case is when $\mathbb{E}[Y|X] = X'\beta$; that is, when the CEF is linear. Importantly, unlike the previous generic discussion of CEFs, in many cases, it may be a strong assumption to impose a linear CEF.

Notation:

I'll follow the convention in the book by writing

$$\mathbb{E}[Y|X = x] = x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1} + \beta_k$$

so that the "intercept" is in the last position. More specifically,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

so that X and β are both $k \times 1$ vectors. And Y, the outcome, is a scalar.

When referring to particular observations, I'll use the notation

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix} \qquad \text{and} \qquad Y_i$$

where X_i is $k \times 1$ and Y_i is a scalar.

In this case, the marginal contrast is given by $MC_1(x) = \beta_1$, so that β_1 should be interpreted as: comparing observations with the same values of the other regressors, those whose X_1 is one unit higher have, on average, an outcome that is β_1 higher. This is a major simplification from the general version of $MC_1(x)$ that we discussed above. It has pros and cons. If the CEF really is linear, then (i) it will be much easier to estimate $m(x) = x'\beta$ (and, hence, $MC_1(x) = \beta_1$) than in the more general case. Additionally, given an estimate of β_1, it is easy to report/fully summarize the marginal contrast for X_1. The main disadvantage is that, often, we may not have a good reason to think/impose that the CEF is actually linear.

Linear CEF with Nonlinear Effects

H: 2.16-2.17

One way to add some additional "flexibility" to the linear CEF model that we have been discussing is to add higher order terms and interactions into the model. For example,

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$

This would still be considered a linear CEF as it is still *linear in the parameters* though it allows for nonlinear effects of the regressors. In particular, notice that

$$MC_1(x_1, x_2) = \beta_1 + 2x_1\beta_3 + x_2\beta_5$$

which can depend on the values of x_1 and x_2. To be clear, this is still less flexible than the general version of $MC_1(x)$ that we discussed earlier, but it does allow for marginal contrasts to depend on the values of different regressors.

One important special case of the nonlinear effects CEF mentioned above is when, say, x_2 is a binary variable. For simplicity, let's consider a slight modification of the previous model such that

$$m(x_1, x_2, x_3) = x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + x_3\beta_4$$

Suppose also that Y is a person's income, X_1 is a person's years of education, X_2 is whether or not a person is married, and X_3 is how old a person is. In this case,

$$MC_1(x_1, x_2, x_3) = \beta_1 + x_2\beta_3$$

which depends on whether $x_2 = 1$ or 0; that is $MC_1(x_1, 1, x_3) = \beta_1 + \beta_3$ while $MC_1(x_1, 0, x_3) = \beta_1$, so that (in the example), the marginal contrast for education can differ between married and unmarried people. Further, rearranging, we have that

$$\beta_3 = MC_1(x_1, 1, x_3) - MC_1(x_1, 0, x_3)$$

so that β_3 is equal to the difference between the marginal contrasts for education for married people

relative to unmarried people.

Please read H: 2.16-2.17 for additional discussions on nonlinear effects in linear CEF models. I think much of this will be review from your undergrad econometrics, but these are things that you should make sure that you fully understand. Relatedly, please read the discussion in H: 2.4 about taking the logarithm of positive random variables (which is relatively common in economics).

Best Linear Predictor

H: 2.18

In many cases, we may be hesitant (or not have a good reason to believe) that the CEF is actually linear. Even in this case, we can still "run a regression" of Y on X.

In this section, we'll consider the best *linear* predictor of Y given X. A linear predictor for Y is a function $X'b$ for some $b \in \mathbb{R}^k$. We will choose the best possible value of b. For this section, we will make the following regularity assumptions (Assumption 2.1 in the textbook)

1. $\mathbb{E}[Y^2] < \infty$
2. $\mathbb{E}||X||^2 < \infty$
3. $\mathbb{E}[XX']$ is positive definite

where $||x|| = (x'x)^{1/2}$ is the Euclidean length of the vector x. The first two assumptions imply that Y and X have finite second moments (and, therefore, finite means, variances, and covariances). As earlier, we will try to minimize mean squared prediction error; that is,

$$\beta = \underset{b}{\operatorname{argmin}}\ S(b)$$

where we define

$$S(b) = \mathbb{E}[(Y - X'b)^2]$$

We can use tools from calculus to solve this. As a first step, it is helpful to notice that

$$\mathbb{E}[(Y - X'b)^2] = \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$$

As a side-comment, notice that the function that we are minimizing is a scalar (we will likely see other parameters/estimators this semester that minimize or maximize some objective function — these objective functions always return a scalar). Second, the expansion in the previous equation may not be obvious; in general, I think you can do quite well at linear algebra by keeping track of the dimensions of the terms that you are working with. In particular, you would know that you were making a mistake if the dimension of any term above were not scalar. Finally, we combined

the two terms $\mathbb{E}[YX]'\beta$ and $\beta'\mathbb{E}[XY]$ for the middle term above — these are equal to the transpose of each other and since they are both scalars, they are exactly equal to each other.

Next, let's take the derivative of the previous equation, set it equal to 0 and solve for β. Before we do this, let's be clear about exactly what we are doing. We are taking the derivative of a function $S(b) : \mathbb{R}^k \to \mathbb{R}$ (that is a function that takes in a k dimensional vector and returns a scalar). And, in particular, this vector derivative is given by

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix}$$

which is a $k \times 1$ vector. As a side-comment, I follow the convention (which is also used in the textbook) of taking first derivatives of scalar-valued functions with respect to a vector "down" (so that the first derivative is $k \times 1$ vector rather than a $1 \times k$ vector) and (if needed) second derivatives "across'' (so that the second derivative would result in a $k \times k$ matrix). For more details about taking derivatives with respect to a vector, see the discussion on p.38 of the textbook and in Appendix A.20 (I consider this to be review material, but it is definitely material you should know).

Returning to our present problem, notice that

$$0 = \frac{\partial S(b)}{\partial b}\Big|_{b=\beta} = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta$$

In terms of mechanics, this derivative is just like the scalar case (the first term is a linear term and the second term is a quadratic term) except you just need to make sure that you get a $k \times 1$ vector (rather than, especially, a $1 \times k$ vector).

We can immediately solve the previous equation for β. The regularity conditions above imply that all of the moments here exist and that the matrix $\mathbb{E}[XX']$ is invertible.

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

In this context, we'll refer to β as the **linear projection coefficient** and $X'\beta$ as the **best linear predictor**

We can also define the **projection error**

$$e = Y - X'\beta$$

which is the difference between the actual value of Y and the best linear predictor of Y given X.

An important property of the projection error is that $\mathbb{E}[Xe] = 0$. To see this, notice that

$$\begin{aligned}\mathbb{E}[Xe] &= \mathbb{E}[X(Y - X'\beta)] \\ &= \mathbb{E}[XY] - \mathbb{E}[XX']\beta \\ &= \mathbb{E}[XY] - \mathbb{E}[XX']\mathbb{E}[XX']^{-1}\mathbb{E}[XY] \\ &= 0\end{aligned}$$

Notice that $\mathbb{E}[Xe]$ is a $k \times 1$ vector. When X includes an intercept (so that $X_k = 1$), this implies that $\mathbb{E}[e] = 0$.

Best linear approximation

H 2.25

Next, we show another interesting property/interpretation for the linear projection coefficient β. Suppose that we are interested in learning about the conditional expectation function $m(x) = \mathbb{E}[Y|X = x]$, but we have no reason to suppose that it is linear. Therefore, we might be interested in trying to construct the best linear approximation to $m(x)$. That is, let's consider choosing β in the following way

$$\beta = \underset{b}{\operatorname{argmin}}\ \mathbb{E}[(m(X) - X'b)^2]$$

Following roughly the same strategy as earlier, notice that

$$\mathbb{E}[(m(X) - X'b)^2] = \mathbb{E}[m(X)^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$$

Taking the derivative and setting equal to 0 implies that

$$0 = -2\mathbb{E}[Xm(X)] + 2\mathbb{E}[XX']\beta$$

which further implies that

$$\begin{aligned}\beta &= \mathbb{E}[XX']^{-1}\mathbb{E}[Xm(X)] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbb{E}[Y|X]] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY]\end{aligned}$$

where the second equality holds by the definition of $m(X)$ and the last equality holds by the law of iterated expectations.

This is exactly the same expression for β as we derived earlier under the motivation of best linear predictor. This implies that $X'\beta$ can additionally be interpreted as the best linear approximation to the underlying CEF — even if the CEF is nonlinear. This is a nice property for the linear

projection model to have. That being said, even the best linear approximation to a nonlinear CEF can sometimes be quite poor. See Section 2.28 for an example and some discussion.