LECTURE NOTE 2

CONDITIONAL EXPECTATION

1. FUNCTIONAL FORM

The linear predictor is very flexible because we are free to construct transformations of the original variables. For example, if EXP is a measure of years of job market experience, we can set $X_1 = \text{EXP}$ and $X_2 = \text{EXP}^2$. Then evaluating

$$E^*(Y \mid 1, X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

at $\text{EXP} = c$ gives

$$\beta_0 + \beta_1 c + \beta_2 c^2,$$

and we can do this evaluation for several interesting values for experience.

The same point applies with two or more original variables. Suppose that in addition to EXP we have EDUC, a measure of years of education. We can set $X_1 = \text{EDUC}$, $X_2 = \text{EXP}$, and $X_3 = \text{EDUC} \cdot \text{EXP}$. Then evaluating

$$E^*(Y \mid 1, X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

at $\text{EDUC} = c$ and $\text{EXP} = d$ gives

$$\beta_0 + \beta_1 c + \beta_2 d + \beta_3 c \cdot d,$$

and we can do this evaluation for several interesting values for education and experience.

2. CONDITIONAL EXPECTATION

Suppose that we start with a single original variable Z and develop linear predictors of Y based on Z that are increasingly flexible. To be specific, consider using a polynomial of order M:

$$E^*(Y \mid 1, Z, Z^2, \ldots, Z^M).$$

The expectation of the squared prediction error cannot increase as M increases, because the coefficients on the additional terms are allowed to be 0. So

$$E[Y - E^*(Y \mid 1, Z, Z^2, \ldots, Z^M)]^2$$

is decreasing as $M \to \infty$ and must approach a limit (since it is nonnegative). We shall assume that the linear predictor itself approaches a limit, and we shall identify this limit with the conditional expectation, $E(Y \mid Z)$:

$$E(Y \mid Z) = \lim_{M \to \infty} E^*(Y \mid 1, Z, Z^2, \ldots, Z^M).$$

This limit is in a mean square sense:

$$\lim_{M \to \infty} E[E(Y \mid Z) - E^*(Y \mid 1, Z, Z^2, \ldots, Z^M)]^2 = 0.$$

We can think of the conditional expectation as providing the best prediction of Y given Z, with (essentially) no constraint on the functional form of the predictor.

Let U be notation for the prediction error:

$$U \equiv Y - E(Y \mid Z).$$

Then U is orthogonal to any power of Z:

$$\langle U, Z^j \rangle = E(UZ^j) = 0 \qquad (j = 0, 1, 2, \ldots).$$

Because general functions of Z can be approximated (in mean square) by polynomials in Z, we have

$$\langle U, g(Z) \rangle = E[Ug(Z)] = 0$$

for (essentially) arbitrary functions $g(\cdot)$.

In the population, we shall generally prefer to work with the conditional expectation. The linear predictor remains useful, however, because it has a direct sample counterpart: the sample linear predictor or least-squares fit. We shall use a (population) linear predictor to approximate the conditional expectation, and then use use a least squares fit to estimate the linear predictor.

It is useful to have notation for evaluating the conditional expectation at a particular value for Z:

$$r(z) \equiv E(Y \mid Z = z).$$

The function $r(\cdot)$ is called the *regression function*. The regression function evaluated at the random variable Z is the conditional expectation: $r(Z) = E(Y \mid Z)$. Because the regression function may be complicated, we may want to approximate it by a simpler function that would be easier to estimate. For example, $E^*[r(Z) \mid 1, Z]$ is a minimum mean-square error approximation that uses a linear function of Z. This turns out to be the same as the linear predictor of Y given Z:

Claim 1. $E^*[r(Z) \mid 1, Z] = E^*(Y \mid 1, Z) = \beta_0 + \beta_1 Z$.

Proof. Let U denote the prediction error:

$$U \equiv Y - E(Y \mid Z) = Y - r(Z). \tag{1}$$

Then U is orthogonal to any function of Z:

$$E[Ug(Z)] = 0,$$

and so is orthogonal to 1 and to Z:

$$E(U) = E(UZ) = 0.$$

This implies that the linear predictor of U given 1, Z is 0, and applying that to (1) gives

$$0 = E^*(U \mid 1, Z) = E^*(Y \mid 1, Z) - E^*[r(Z) \mid 1, Z]. \quad \diamond$$

The conditional expectation of Y given two (or more) variables Z_1 and Z_2 can also be viewed as a limit of increasingly flexible linear predictors:

$$E(Y \mid Z_1, Z_2) = \lim_{M\to\infty} E^*(Y \mid 1, Z_1, Z_2, Z_1^2, Z_1Z_2, Z_2^2, \ldots,$$
$$Z_1^M, Z_1^{M-1}Z_2, \ldots, Z_1Z_2^{M-1}, Z_2^M).$$

The regression function is defined as

$$r(z_1, z_2) \equiv E(Y \mid Z_1 = z_1, Z_2 = z_2).$$

As above, we can use the linear predictor to approximate the regression function. For example, the proof of claim 1 can be used to show that

$$E^*[r(Z_1, Z_2) \mid 1, Z_1, Z_2, Z_1^2, Z_1Z_2, Z_2^2] = E^*(Y \mid 1, Z_1, Z_2, Z_1^2, Z_1Z_2, Z_2^2).$$

We shall conclude this section by deriving the iterated expectations formula and then using it to obtain an omitted variables formula.

Claim 2 (Iterated Expectations). $E[E(Y \mid Z_1, Z_2) \mid Z_1] = E(Y \mid Z_1)$.
(Equivalently: $E[r(Z_1, Z_2) \mid Z_1] = r(Z_1)$.)
Proof. Let U denote the prediction error:

$$U \equiv Y - E(Y \mid Z_1, Z_2) = Y - r(Z_1, Z_2). \tag{2}$$

Then U is orthogonal to any function of (Z_1, Z_2):

$$E[Ug(Z_1, Z_2)] = 0,$$

and so is orthogonal to any function of Z_1:

$$E[Ug(Z_1)] = 0.$$

This implies that $E(U \mid Z_1) = 0$, and applying that to (2) gives

$$0 = E(U \mid Z_1) = E(Y \mid Z_1) - E[r(Z_1, Z_2) \mid Z_1]. \quad \diamond$$

Claim 3 (Omitted Variable Bias). If

$$E(Y \mid Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$$

and

$$E(Z_2 \mid Z_1) = \gamma_0 + \gamma_1 Z_1,$$

then

$$E(Y \mid Z_1) = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1) Z_1.$$

Proof.

$$\begin{aligned} E(Y \mid Z_1) &= E[E(Y \mid Z_1, Z_2) \mid Z_1] \\ &= E(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 \mid Z_1) \\ &= \beta_0 + \beta_1 Z_1 + \beta_2 E(Z_2 \mid Z_1) \\ &= \beta_0 + \beta_1 Z_1 + \beta_2(\gamma_0 + \gamma_1 Z_1) \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1) Z_1. \quad \diamond \end{aligned}$$

Note that here we assume that the regression function for Y on Z_1 and Z_2 is linear in Z_1 and Z_2, and that the regression function for Z_2 on Z_1 is linear in Z_1. It then follows that the regression function for Y on Z_1 is linear in Z_1, and the coefficients are related to the coefficients in the long regression function in the same way as in claim 1 in Note 1.

3. DISCRETE REGRESSORS

Suppose that Z_1 and Z_2 take on only a finite set of values:

$$Z_1 \in \{\lambda_1, \ldots, \lambda_J\}, \quad Z_2 \in \{\delta_1, \ldots, \delta_K\}.$$

Construct the following *dummy variables*:

$$\begin{aligned} X_{jk} &= \begin{cases} 1, & \text{if } Z_1 = \lambda_j,\ Z_2 = \delta_k; \\ 0, & \text{otherwise;} \end{cases} \\ &= 1(Z_1 = \lambda_j, Z_2 = \delta_k) \qquad (j = 1, \ldots, J;\ k = 1, \ldots, K). \end{aligned}$$

These are indicator variables that equal 1 if a particular value of (Z_1, Z_2) occurs, and equal 0 otherwise. We use the notation $1(B)$ for the indicator function that equals 1 if the event B occurs and equals 0 otherwise.

Claim 4. $E(Y \mid Z_1, Z_2) = E^*(Y \mid X_{11}, \ldots, X_{J1}, \ldots, X_{1K}, \ldots, X_{JK})$

Proof. Any function $g(Z_1, Z_2)$ can be written as

$$g(Z_1, Z_2) = \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{jk} X_{jk}$$

with $\gamma_{jk} = g(\lambda_j, \delta_k)$. So searching over functions g to find the best predictor is equivalent to searching over the coefficients γ_{jk} to find the best linear predictor. $\diamond$

So the conditional expectation function can be expressed as a linear combination of the dummy variables:

$$E(Y \mid Z_1, Z_2) = \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{jk} X_{jk}$$

with

$$\beta_{jk} = E(Y \mid Z_1 = \lambda_j, Z_2 = \delta_k).$$

Note this requires that we use a complete set of dummy variables, with one for each value of (Z_1, Z_2). In this discrete regressor case, there is a concrete form for the notion that conditional expectation is a limit of increasingly flexible linear predictors. Here the limit is achieved by using a complete set of dummy variables in the linear predictor.

There is a sample analog to this result, using least-squares fits. The basic data consist of (y_i, z_{i1}, z_{i2}) for each of $i = 1, \ldots, n$ members of the sample. Construct the dummy variables

$$x_{i,jk} = 1(z_{i1} = \lambda_j, z_{i2} = \delta_k)$$

and the matrices

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x_{jk} = \begin{pmatrix} x_{1,jk} \\ \vdots \\ x_{n,jk} \end{pmatrix} \qquad (j = 1, \ldots, J;\ k = 1, \ldots, K).$$

The coefficients in the least-squares fit are obtained from

$$\min ||y - \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}x_{jk}||^2,$$

where the minimization is over $\{b_{jk}\}$ and the inner product is

$$\langle y, x_{jk}\rangle = \frac{1}{n}\sum_{i=1}^{n} y_i x_{i,jk}.$$

Claim 5.

$$b_{lm} = \frac{\sum_{i=1}^{n} y_i x_{i,lm}}{\sum_{i=1}^{n} x_{i,lm}} \qquad (l = 1, \ldots, J;\ m = 1, \ldots, K).$$

Proof. The residual from the least-squares fit must be orthogonal to each of the dummy variables:

$$\langle y - \sum_{j,k} b_{jk}x_{jk}, x_{lm}\rangle = 0.$$

The dummy variables are orthogonal to each other:

$$\langle x_{jk}, x_{lm}\rangle = 0$$

unless $j = l$ and $k = m$. So we have

$$\begin{aligned}\langle y - \sum_{j,k} b_{jk}x_{jk}, x_{lm}\rangle &= \langle y, x_{lm}\rangle - \sum_{j,k} b_{j,k}\langle x_{jk}, x_{lm}\rangle \\ &= \langle y, x_{lm}\rangle - b_{lm}\langle x_{lm}, x_{lm}\rangle \\ &= 0.\end{aligned}$$

So

$$b_{lm} = \frac{\langle y, x_{lm}\rangle}{\langle x_{lm}, x_{lm}\rangle} = \frac{\sum_{i=1}^{n} y_i x_{i,lm}}{\sum_{i=1}^{n} x_{i,lm}}.$$

(Note that $x_{i,lm}^2 = x_{i,lm}$ because $x_{i,lm}$ equals 0 or 1.) $\diamond$

Note that $\sum_i y_i x_{i,lm}$ is summing the y values for the observations with $(z_{i1}, z_{i2}) = (\lambda_l, \delta_m)$, and $\sum_i x_{i,lm}$ is the number of observations with this value for (z_{i1}, z_{i2}). So the coefficient b_{lm} is a subsample mean, for the subsample with $(z_{i1}, z_{i2}) = (\lambda_l, \delta_m)$. In order to stress this interpretation as a subsample mean, we shall use the notation

$$\bar{y} \,|\, \lambda_l, \delta_m \equiv \frac{\sum_{i=1}^{n} y_i x_{i,lm}}{\sum_{i=1}^{n} x_{i,lm}}.$$

A major use of regression analysis is to measure the effect of one variable holding constant other variables. Consider, for example, the effect on Y of a change from $Z_1 = c$ to $Z_1 = d$, holding Z_2 constant at $Z_2 = e$. Let θ denote this effect:

$$\begin{aligned} \theta &= E(Y \,|\, Z_1 = d, Z_2 = e) - E(Y \,|\, Z_1 = c, Z_2 = e) \\ &= r(d, e) - r(c, e). \end{aligned}$$

This is a predictive effect. It measures how the prediction of Y changes as we change the value for one of the predictor variables, holding constant the value of the other predictor variable. In the case of discrete regressors with a complete set of dummy variables, this predictive effect has a sample analog:

$$\hat{\theta} = (\bar{y} \,|\, d, e) - (\bar{y} \,|\, c, e).$$

We estimate θ by comparing two subsample means. The individuals in the first subsample have $z_{i1} = c$, and the individuals in the second subsample have $z_{i1} = d$. In both subsamples, all individuals have the same value for z_2: $z_{i2} = e$. So the sense in which z_2 is being held constant is clear: all individuals in the comparison of means have the same value for z_2.

In general there is a different effect θ for each value of Z_2, and we may want to have a way to summarize these effects. This is discussed in the next section.

4. AVERAGE PARTIAL EFFECT

Recall our definition of the regression function:

$$r(s, t) = E(Y \,|\, Z_1 = s, Z_2 = t).$$

Consider the predictive effect based on comparing $Z_1 = c$ with $Z_1 = d$, with $Z_2 = t$:

$$r(d, t) - r(c, t).$$

Instead of reporting a different effect for each value of Z_2, we can evaluate the effect at the random variable Z_2:

$$r(d, Z_2) - r(c, Z_2).$$

This gives a random variable, and we can take its expectation:

$$\theta = E[r(d, Z_2) - r(c, Z_2)].$$

We shall refer to this as an *average partial effect*. It is "partial" in the sense of holding Z_2 constant.

Note that

$$\theta = E[r(d, Z_2)] - E[r(c, Z_2)].$$

This is not the same, in general, as

$$\begin{aligned}\tilde{\theta} &= E[r(d, Z_2) \,|\, Z_1 = d] - E[r(c, Z_2) \,|\, Z_1 = c] \\ &= E(Y \,|\, Z_1 = d) - E(Y \,|\, Z_1 = c).\end{aligned}$$

An integral notation may be helpful. Define

$$\begin{aligned}&F(B \,|\, s, t) = \text{Prob}(Y \in B \,|\, Z_1 = s, Z_2 = t), \\ &F_{Z_2}(B) = \text{Prob}(Z_2 \in B), \\ &F_{Z_2}(B \,|\, s) = \text{Prob}(Z_2 \in B \,|\, Z_1 = s).\end{aligned}$$

Then

$$r(s, t) = \int y dF(y \,|\, s, t),$$

and

$$\theta = \int [r(d, t) - r(c, t)] dF_{Z_2}(t).$$

We have $\theta = \tilde{\theta}$ if

$$F_{Z_2}(B \,|\, c) = F_{Z_2}(B \,|\, d) = F_{Z_2}(B)$$

for all B; $\theta = \tilde{\theta}$ for all c and d if Z_1 and Z_2 are independent.

Once we have an estimate $\hat{r}$ of the regression function, we can form an estimate of θ by taking an average over the sample:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}[\hat{r}(d, z_{i2}) - \hat{r}(c, z_{i2})]. \tag{3}$$

We can obtain an estimate of r by first approximating the conditional expectation by a linear predictor, using a polynomial in Z_1 and Z_2:

$$\begin{aligned} E(Y \,|\, Z_1, Z_2) &\cong E^*(Y \,|\, \{Z_1^j \cdot Z_2^k\}_{j+k=0}^M) \\ &= \sum_{j,k:j+k=0}^{M} \beta_{jk} Z_1^j \cdot Z_2^k. \end{aligned}$$

We can use a least-squares fit to obtain estimates b_{jk} of the coefficients β_{jk}. Then we can use

$$\hat{r}(c, z_{i2}) = \sum_{j,k:j+k=0}^{M} b_{jk} c^j \cdot z_{i2}^k \quad \text{and} \quad \hat{r}(d, z_{i2}) = \sum_{j,k:j+k=0}^{M} b_{jk} d^j \cdot z_{i2}^k$$

in (3).

Now consider the special case of discrete regressors, with

$$Z_1 \in \{\lambda_1, \ldots, \lambda_J\}, \quad Z_2 \in \{\delta_1, \ldots, \delta_K\}.$$

In this case,

$$\theta = \sum_{k=1}^{K}[r(d, \delta_k) - r(c, \delta_k)]\mathrm{Prob}(Z_2 = \delta_k).$$

We can estimate θ using the sample analog

$$\hat{\theta} = \sum_{k=1}^{K}[(\bar{y} \,|\, d, \delta_k) - (\bar{y} \,|\, c, \delta_k)](n_k/n),$$

where n_k is the number of observations with $z_{i2} = \delta_k$:

$$n_k = \sum_{i=1}^{n} 1(z_{i2} = \delta_k),$$

and the mean of y for a subsample is

$$(\bar{y} \mid s, t) \equiv \frac{\sum_{i=1}^{n} y_i 1(z_{i1} = s, z_{i2} = t)}{\sum_{i=1}^{n} 1(z_{i1} = s, z_{i2} = t)}.$$

5. LOGS

Section 1 stressed that the linear predictor is flexible because we are free to construct transformations of the original variables. A transformation that is often used is the logarithm:

$$E^*(Y \mid 1, \log Z) = \beta_0 + \beta_1 \log Z.$$

(This is the log to the base e or natural logarithm ln.) In order to compare $Z = c$ and $Z = d$, we simply substitute:

$$\beta_1 \log d - \beta_1 \log c = \beta_1 \log(d/c).$$

A useful approximation here is

$$(\beta_1/100)[100 \log(d/c)] \cong (\beta_1/100)[100(\frac{d}{c} - 1)].$$

With this approximation, we can interpret $(\beta_1/100)$ as the (predictive) effect of a one per cent change in Z.

Now consider a log transformation of Y:

$$E^*(\log Y \mid 1, Z) = \beta_0 + \beta_1 Z.$$

We can certainly say that the predicted change in $\log Y$ is $\beta_1(d - c)$, and it is often useful to think of $100\beta_1(d - c)$ as a predicted percentage change in Y. We should note, however, that even if the conditional expectation of $\log Y$ is linear, so that

$$E(\log Y \mid Z) = \beta_0 + \beta_1 Z,$$

we cannot relate this to the conditional expectation of Y without additional assumptions.

To see this, define

$$U \equiv \log Y - E(\log Y \mid Z),$$

so that

$$E(U \mid Z) = 0.$$

Since $\log Y = \beta_0 + \beta_1 Z + U$, we have

$$\begin{aligned} Y &= \exp(\beta_0 + \beta_1 Z + U) \\ &= \exp(\beta_0 + \beta_1 Z) \cdot \exp(U). \end{aligned}$$

So

$$E(Y \mid Z) = \exp(\beta_0 + \beta_1 Z) \cdot E[\exp(U) \mid Z].$$

In general, $E(U \mid Z) = 0$ does not imply that $E[\exp(U) \mid Z]$ is a constant. If we make an additional assumption that U and Z are independent, then

$$E[\exp(U) \mid Z] = E[\exp(U)].$$

In that case,

$$\frac{E(Y \mid Z = d)}{E(Y \mid Z = c)} = \exp[\beta_1(d - c)] \cong \beta_1(d - c) + 1,$$

and

$$100\left[\frac{E(Y \mid Z = d)}{E(Y \mid Z = c)} - 1\right] \cong 100\beta_1(d - c).$$