Ec. 2120 August 28, 2024
Harvard Economics

LECTURE NOTE 2

CONDITIONAL EXPECTATION

0. INTRODUCTION

This note introduces a -precise- probability framework that can be used to derive various approximations. The framework is based on finite state space for clarity (and to avoid unnecessary for now) measure theoretic issues. This note is important because the framework easily leads to Bayes rule, the law of iterated expectation, and the law of total probability. These are all useful. The note also applies the finite state space setup to regressions with discrete regressors, deriving short and long regression formulas that are useful and insightful.

1. FINITE STATE SPACE

The basic concepts of probability can be presented with simplicity and precision in the case of a finite state space. The measure theory details needed for a precise treatment of an infinite state space can be avoided, letting us concentrate on the concepts and building intuition. There is a finite set

$$S = \{s_1, \ldots, s_M\}$$

whose elements are the *states of nature*. The points in S are also called elementary outcomes or elementary events. A *probability measure* P assigns a real number $P(s)$ to each state $s \in S$. One and only one of the states will occur, and this is reflected in their probabilities being nonnegative and summing to one:

$$P(s_j) \geq 0, \quad \sum_{j=1}^{M} P(s_j) = 1.$$

An *event* A is a subset of S. It occurs if it contains the state that occurs. So the probability of A is the sum of the probabilities of the states in A:

$$P(A) = \sum_{s \in A} P(s).$$

Let $\mathcal{B}$ denote the set of events. Then we can regard P as a mapping from the set of events to the interval $[0, 1]$ of real numbers between 0 and 1:

$$P\colon \mathcal{B} \to [0, 1].$$

A *probability space* consists of the triple $(S, \mathcal{B}, P\}$.

A *random variable* Y is a mapping from S to the real numbers $\mathcal{R}$:

$$Y\colon S \to \mathcal{R}.$$

It's *expectation* $E(Y)$ is formed by evaluating Y at each state and forming a weighted average, weighting by the probabilities of the states:

$$E(Y) = \sum_{j=1}^{M} Y(s_j) P(s_j).$$

There is an equivalent way to write $E(Y)$, in terms of the distinct values that Y can take on: $\mathcal{Y} = \{y_1, \ldots, y_L\}$ (where L must be $\leq M$). The probability that $Y = y_l$ is the probability of the event

$$A_l = \{s \in S : Y(s) = y_l\}, \quad \text{so} \quad P(A_l) = \sum_{s \in A_l} P(s),$$

and

$$E(Y) = \sum_{j=1}^{M} Y(s_j) P(s_j) = \sum_{l=1}^{L} y_l \Bigg(\sum_{j : Y(s_j) = y_l} P(s_j) \Bigg) = \sum_{l=1}^{L} y_l P(Y = y_l).$$

We can think of the random variable Y inducing a new probability space, with state space equal to $\{y_1, \ldots, y_L\}$ (the image of S under the mapping Y). Now the probability measure is the distribution of Y, which is given by

$$F_Y(B) = \sum_{y \in B} f_Y(y) \quad \text{for} \quad B \subset \{y_1, \ldots, y_L\},$$

with

$$f_Y(y) = \sum_{s:Y(s)=y} P(s) \quad \text{for} \quad y \in \{y_1, \ldots, y_L\}.$$

I shall refer to f_Y as a *density* function. (In this discrete case, it is also known as a probability mass function. In the continuous case, the counterpart is known as a probability density function. It is convenient to use the term density in both cases. In the discrete case, f_Y can be regarded as the density of F_Y with respect to counting measure.) Using this induced probability space, the expectation of Y is a weighted average of the values of Y (the states), with weights given by the density of Y:

$$E(Y) = \sum_{l=1}^{L} y_l f_Y(y_l).$$

1.1 *Conditional Probability*

Suppose we know that an event B, with $P(B) > 0$, has occurred. If the state s_j is not in B, then we define its conditional probability given B to be 0. If $s_j \in B$, then define

$$P(s_j \,|\, B) = \frac{P(s_j)}{\sum_{s \in B} P(s)} = P(s_j)/P(B).$$

These conditional probabilities are nonnegative and sum to one:

$$P(s_j \,|\, B) \geq 0, \quad \sum_{j=1}^{M} P(s_j \,|\, B) = 1.$$

So $P(\,\cdot \,|\, B)$ is a probability measure that assigns an event A the probability

$$P(A \,|\, B) = \frac{\sum_{s \in A \cap B} P(s)}{P(B)} = P(A \cap B)/P(B).$$

For states s_j and s_k in B, we preserve the ratio of probabilities:

$$P(s_j \mid B)/P(s_k \mid B) = P(s_j)/P(s_k),$$

while ensuring that the conditional probability of B is one:

$$\sum_{s \in B} P(s \mid B) = 1.$$

(We could work with a smaller state space, dropping the states not in B because they have zero conditional probability.)

Suppose that the collection of disjoint subsets $\{B_i\}$ forms a partition of S: $B_i \cap B_j = \emptyset$ if $i \neq j$ and $\cup_j B_j = S$. Then $A = \cup_j (A \cap B_j)$ implies

$$P(A) = \sum_j P(A \cap B_j),$$

which gives the *partition formula*

$$P(A) = \sum_j P(A \mid B_j) P(B_j).$$

Combining the partition formula with the definition of conditional probability gives *Bayes' rule* (or *Bayes' theorem*):

$$\begin{aligned} P(B_i \mid A) &= \frac{P(A \mid B_i) P(B_i)}{P(A)} \\ &= \frac{P(A \mid B_i) P(B_i)}{\sum_j P(A \mid B_j) P(B_j)}. \end{aligned}$$

The definition of the random variable Y is unchanged, and its expectation with respect to the conditional distribution is

$$E(Y \mid B) = \sum_{j=1}^{M} Y(s_j) P(s_j \mid B).$$

Let X be another random variable, with values $\mathcal{X} = \{x_1, \ldots, x_K\}$. Then the subset of states for which $X = x_k$ is an event:

$$B = \{s \in S : X(s) = x_k\},$$

and there is a corresponding conditional probability measure:

$$P(s_j \,|\, X = x_k) = P(s_j)/P(X = x_k) \quad \text{if} \quad X(s_j) = x_k,$$

and $P(s_j \,|\, X = x_k) = 0$ otherwise. The expectation of Y with respect to this conditional distribution is

$$E(Y \,|\, X = x_k) = \sum_{i=1}^{M} Y(s_j) P(s_j \,|\, X = x_k).$$

We can assemble these conditional expectations for different values of X into a function:

$$r\colon \{x_1, \ldots, x_K\} \to \mathcal{R},$$

with

$$r(x) = E(Y \,|\, X = x).$$

This is known as the *regression function*, and it will play a very important role in our course. (Our definition only applies where $P(X = x) \neq 0$; r can be assigned an arbitrary value at other points in $\mathcal{X}$.)

Bayes' rule takes the form

$$P(X = x_k \,|\, Y = y_j) = \frac{P(Y = y_j \,|\, X = x_k) P(X = x_k)}{\sum_l P(Y = y_j \,|\, X = x_l) P(X = x_l)}.$$

1.2 *Joint Distribution*

We can think of the pair of random variables (X, Y) inducing a new probability space, where the states are the values (x, y) in the Cartesian product

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) : x \in \{x_1, \ldots, x_K\}, y \in \{y_1, \ldots, y_L\}\}.$$

The probability measure on this state space is the *joint distribution* of (X, Y), which is given by

$$F_{XY}(C) = \sum_{(x,y)\in C} f_{XY}(x,y) \quad \text{for} \quad C \subset \mathcal{X} \times \mathcal{Y},$$

with

$$f_{XY}(x,y) = P(X = x, Y = y) = \sum_{s\in S: X(s)=x, Y(s)=y} P(s) \quad \text{for} \quad (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

This *joint density* f_{XY} implies *marginal densities* for X and Y:

$$f_X \colon \mathcal{X} \to [0,1], \quad f_X(x) = \sum_{y\in\mathcal{Y}} f_{XY}(x,y),$$

$$f_Y \colon \mathcal{Y} \to [0,1], \quad f_Y(y) = \sum_{x\in\mathcal{X}} f_{XY}(x,y).$$

The joint and marginal densities can be used to construct a *conditional density*: if $P(X = x) \neq 0$,

$$\begin{aligned} f_{Y\,|\,X}(y\,|\,x) &= f_{XY}(x,y)/f_X(x) \\ &= P(X = x, Y = y)/P(X = x) \\ &= P(Y = y\,|\,X = x). \end{aligned}$$

For any $x \in \mathcal{X}$ with $f_X(x) \neq 0$, $f_{Y\,|\,X}(\,\cdot\,|\,x)$ is a (conditional) density on $\mathcal{Y}$. These conditional densities can be combined with the marginal density f_X to get back the joint density:

$$f_{XY}(x,y) = f_{Y\,|\,X}(y\,|\,x) f_X(x),$$

where the right-hand side is interpreted to be 0 when $f_X(x) = 0$ (and $f_{Y\,|\,X}(y\,|\,x)$ is not defined). This factorization of a joint density into a conditional density and a marginal density can be very useful.

Using the joint density, we can express the regression function as

$$r(x) = E(Y\,|\,X = x) = \sum_{y\in\mathcal{Y}} y f_{Y\,|\,X}(y\,|\,x).$$

1.3 *Optimal Prediction*

In Note 1 we developed an optimal linear predictor:

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

The optimality of the predictor is in minimizing mean-square error:

$$(\beta_0, \beta_1) = \arg\min_{c_0, c_1} E(Y - c_0 - c_1 X)^2.$$

If we are not allowed to use X, the prediction problem is

$$\min_c E(Y - c)^2,$$

with solution

$$\beta = \arg\min_c E(Y - c)^2$$

that satisfies the orthogonality condition

$$E[(Y - \beta) \cdot 1] = 0,$$

so $\beta = E(Y)$ and

$$E^*(Y \mid 1) = E(Y).$$

Given that $E(Y)$ is an optimal predictor when we do not condition on anything, it is plausible that the regression function $r(x) = E(Y \mid X = x)$ would be an optimal predictor when we condition on X. So consider the following prediction problem:

$$\min_g E[Y - g(X)]^2,$$

where g can be any function defined on $\mathcal{X} = \{x_1, \ldots, x_K\}$. We can evaluate the expectation using the joint density of (X, Y) and its factorization into a conditional density for Y given X and the marginal density of X:

$$\begin{aligned} E[Y - g(X)]^2 &= \sum_{x,y} [y - g(x)]^2 f_{XY}(x, y) \\ &= \sum_x \left(\sum_y [y - g(x)]^2 f_{Y \mid X}(y \mid x) \right) f_X(x). \end{aligned}$$

Because $f_X(x) \geq 0$, we cannot do better than minimizing the inner sum separately at each value of x for which $f_X(x) > 0$, which gives the expectation of Y with respect to the conditional distribution given $X = x$:

$$\arg\min_c \sum_y (y - c)^2 f(y \mid x) = E(Y \mid X = x).$$

So the optimal choice of the function g is the regression function r.

Note: This is an important observation mainly that the function that does "best" in terms of minimizing expected square loss is the conditional expectation function. In economics, the conditional expectation is of primary interest. And we will see below the sense in which linear functions can come close in approximating this general function.

The linear predictor $E^*(Y \mid 1, X)$ is also minimizing mean-square error, but subject to the restriction that the function g be linear in X (i.e., a linear combination of 1 and X). So in general it will not do as well as the regression function (unless the regression function is linear):

$$E(Y - \beta_0 - \beta_1 X)^2 \geq E[Y - r(X)]^2,$$

with strict inequality ($>$) unless $r(X) = \beta_0 + \beta_1 X$. However, we can include additional variables in the linear predictor, and these variables can be given, nonlinear functions of X, such as $X^2, X^3, \ldots$. In this way a linear predictor can approximate the regression function. This is a fruitful idea that carries over to a general (possibly infinite) state space.

1.4 *General State Space*

I mainly want to introduce some notation for the general case and give a rough indication of how it relates to the case of a finite state space. A good source for a careful treatment is Billingsley (1995).

As before, an event A is a subset of the state space S, but now some regularity may be needed on the subsets that are assigned probabilities. $\mathcal{B}$ denotes this set of measurable

subsets of S. For example, S could be the interval $[a, b]$ on the real line $\mathcal{R}$, and the uniform distribution assigns

$$P([c, d]) = \frac{d - c}{b - a} \qquad (a \leq c < d \leq b).$$

Let λ denote Lebesgue measure on $\mathcal{R}$, so that $\lambda([c, d]) = d - c$. If A is a measurable subset of $[a, b]$, then

$$P(A) = \frac{1}{b - a}\lambda(A).$$

As before, a random variable Y is a mapping from S to $\mathcal{R}$,

$$Y \colon S \to \mathcal{R},$$

but now some regularity may be needed, and we work with measurable functions. If Y only takes on a finite set of values $\{y_1, \ldots, y_L\}$, then

$$E(Y) = \sum_l y_l P(Y = y_l).$$

For a general measurable function, we can think of approximating it by a simple function that takes on only a finite set of values, evaluate the expectation for the simple function, and take a limit. The general notation is

$$E(Y) = \int_S Y(s)\, dP(s),$$

representing integration with respect to the (probability) measure P. In the above example with a uniform distribution on the interval $[a, b]$, we have

$$E(Y) = \frac{1}{b - a} \int_{[a,b]} Y(s) d\lambda(s),$$

representing integration with respect to Lebesgue measure. This has the alternative notation $\int_{[a,b]} Y(s)\, ds$. The symbol for the dummy variable of integration is arbitrary, so could also write $\int_{[a,b]} Y(x)\, d\lambda(x)$ or $\int_{[a,b]} Y(x)\, dx$.

As before, the random variable Y induces a new probability space. We can take the state space to be the real line $\mathcal{R}$, with probability measure given by

$$F_Y(B) = P\{s : Y(s) \in B\}$$

(with B a suitable measurable subset of $\mathcal{R}$). Using this induced probability space, the expectation of Y is a weighted average of the values of Y (the states), with weights given by the distribution of Y:

$$E(Y) = \int y \, dF_Y(y).$$

If the induced distribution for Y is uniform on $[a, b]$, then

$$E(Y) = \frac{1}{b-a} \int_a^b y \, dy.$$

If X has a uniform distribution on the interval $[a, b]$, then

$$P(X = c) = \lim_{d \downarrow c} P(c \leq X \leq d) = \lim_{d \downarrow c} \frac{d-c}{b-a} = 0.$$

So some care is needed in defining conditional probability and conditional expectation given an event B like $X = c$ that has zero probability. In Section 3 I shall use optimal prediction to motivate a general version of conditional expectation (along the lines of Section 1.3 above), and then the conditional probability of a set can be obtained from the conditional expectation of the indicator function for that set. The general notation for the conditional probability measure is $P(\,\cdot \,|\, B)$ and

$$P(A \,|\, B) = \int_A dP(s \,|\, B), \quad E(Y \,|\, B) = \int Y(s) \, dP(s \,|\, B).$$

As before, a pair of random variables (X, Y) induces a new probability space, with state space $\mathcal{R}^2 = \mathcal{R} \times \mathcal{R}$ and with probability measure given by the joint distribution of (X, Y):

$$F_{XY}(B \times A) = P\{s : X(s) \in B, Y(s) \in A\}.$$

This joint distribution implies marginal distributions for X and Y:

$$F_X(B) = F_{XY}(B \times \mathcal{R}), \quad F_Y(A) = F_{XY}(\mathcal{R} \times A).$$

The notation for conditional distribution is

$$F_{Y \mid X}(A \mid x) = P(Y \in A \mid X = x).$$

The joint distribution can be recovered from the conditional distribution and the marginal distribution:

$$F_{XY}(B \times A) = \int_B F_{Y \mid X}(A \mid x)\, dF_X(x).$$

The notation for the regression function is

$$r(x) = E(Y \mid X = x) = \int_S Y(s)\, dP(s \mid X = x) = \int y\, dF_{Y \mid X}(y \mid x).$$

Bayes' rule takes the form

$$P(X \in B \mid Y \in A) = \frac{\int_B P(Y \in A \mid X = x)\, dF_X(x)}{\int P(Y \in A \mid X = x)\, dF_X(x)}.$$

If F_{XY} has a *joint density* f_{XY} (with respect to Lebesgue measure on $\mathcal{R}^2$), then

$$\begin{aligned} F_{XY}(B \times A) &= \int_{B \times A} f_{XY}(x, y)\, dx\, dy \\ &= \int_A \left(\int_B f_{XY}(x, y)\, dx \right) dy = \int_B \left(\int_A f_{XY}(x, y)\, dy \right) dx. \end{aligned}$$

The *marginal densities* are given by

$$f_X(x) = \int f_{XY}(x, y)\, dy, \quad f_Y(y) = \int f_{XY}(x, y)\, dx.$$

At a point x where $f_X(x) \neq 0$, we can define the *conditional density*

$$f_{Y \mid X}(y \mid x) = f_{XY}(x, y)/f_X(x),$$

and use it to obtain the conditional distribution:

$$F_{Y\,|\,X}(A\,|\,x) = \int_A f_{Y\,|\,X}(y\,|\,x)\,dy.$$

The joint density factors into the product of the conditional density and the marginal density:

$$f_{XY}(x,y) = f_{Y\,|\,X}(y\,|\,x) f_X(x),$$

so that

$$\int_B \left(\int_A f_{Y\,|\,X}(y\,|\,x)\,dy \right) f_X(x)\,dx = \int_B \left(\int_A f_{XY}(x,y)\,dy \right) dx = F_{XY}(B \times A).$$

Bayes' rule takes the form

$$f_{X\,|\,Y}(x\,|\,y) = \frac{f_{Y\,|\,X}(y\,|\,x) f_X(x)}{\int f_{Y\,|\,X}(y\,|\,x) f_X(x)\,dx}.$$

2. FUNCTIONAL FORM

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

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

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

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

at EDUC $= c$ and 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.

3. 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. In the case of a finite state space, this interpretation of conditional expectation as a limit of optimal linear predictors agrees with the definition given in Section 1. We shall see this connection explicitly in Section 4, which considers the case where Z takes on only a finite set of values (discrete regressor).

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:

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

This is 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:

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

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

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

Note: The takeaway is we have given a motivation for studying linear projections or least squares if indeed one cares about the conditional expectation function. In addition, when one studies *linear* models, it is not necessarily a strong restriction in the sense: we can make our linear regression rich enough to be able approximate arbitrary (smooth enough) functions.

We shall conclude this section by deriving the iterated expectations formula and then use 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):

$$E[U g(Z_1, Z_2)] = 0,$$

and so is orthogonal to any function of Z_1:

$$E[U g(Z_1)] = 0.$$

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.

4. 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

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

Predictive Effect: 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 $Z_2 = t$. Let γ denote this effect:

$$\begin{aligned}\gamma &= E(Y \mid Z_1 = d, Z_2 = t) - E(Y \mid Z_1 = c, Z_2 = t) \\ &= r(d,t) - r(c,t).\end{aligned}$$

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.

Note: Even though it is a predictive effect, it is not causal in the sense that it is not a causal predictive effect. It simply answers the question of how much would your prediction change if one variable's value changed.

In the case of discrete regressors with a complete set of dummy variables, this predictive effect has a sample analog:

$$\hat{\gamma} = (\bar{y} \mid d, t) - (\bar{y} \mid c, t).$$

We estimate γ 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: $z_{i2} = t$. 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 γ 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.

5. AVERAGE PARTIAL EFFECT

Recall our definition of the regression function:

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

Consider the predictive effect based on comparing $Z_1 = c$ and $Z_1 = d$, 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:

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

We shall refer to this as an *average partial effect*. It is "partial" in the sense of holding Z_2 constant (for example, if Z_2 is binary and takes two values $0, 1$, then the above will be a weighted sum of the regression function for the subpopulation with $Z_2 = 1$ and the subpopulation for which $Z_2 = 0$ where the weights are the proportions $P(Z_2 = 1)$)

Note: Since $r(d, Z_2) - r(c, Z_2)$ is a random variables, one can report statistics other than its mean. For example, one can take the median or any quantile of this random variable. Plotting its density or quantile function would also be helpful.

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

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

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:

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

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,

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

We can estimate γ using the sample analog

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

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

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

Note: Averaging over the *marginal* distribution of Z_2 may be give a misleading answer for the partial effect when Z_1 and Z_2 are correlated. When we are fixing the value of Z_1 at d for example, the conditional distribution of $Z_2|Z_1 = d$ may be different then unconditional distribution so then $\hat{\gamma}$ may confound the interpretation of this effect due to the correlation between Z_1 and Z_2.

6. EXAMPLE: POLYNOMIAL REGRESSORS

Consider the following quadratic polynomial approximation to a regression function:

$$E(Y|Z_1 = s, Z_2 = t) \cong \beta_0 + \beta_1 s + \beta_2 s^2 + \beta_3 t \cdot s + \beta_4 t + \beta_5 t^2.$$

Table 5.1 in Mincer (1974) provides the least-squares fit:

$$\hat{y} = 4.87 + .255s - .0029s^2 - .0043t \cdot s + .148t - .0018t^2,$$

with $y = \log(\text{earnings})$, $s =$ years of schooling, and $t =$ years of work experience. The data are from the 1 in 1000 sample, 1960 census with 1959 annual earnings, and sample size $n = 31093$.

The partial predictive effect of four years of college, holding work experience constant at t, is

$$E(Y|Z_1 = 16, Z_2 = t) - E(Y|Z_1 = 12, Z_2 = t)$$
$$\cong \beta_1 \cdot 4 + \beta_2(16^2 - 12^2) + \beta_3 \cdot 4 \cdot t.$$

The partial predictive effect of four years of high school, holding work experience constant at t is

$$E(Y|Z_1 = 12, Z_2 = t) - E(Y|Z_1 = 8, Z_2 = t)$$
$$\cong \beta_1 \cdot 4 + \beta_2(12^2 - 8^2) + \beta_3 \cdot 4 \cdot t.$$

Evaluating these expressions with $t = 0, 10, 20$ gives the following table:

Experience	Returns to college	Returns to high school
0	.70	.79
10	.52	.62
20	.35	.44

The term "returns to college" is related to the use of log(earnings) and is discussed below.

7. 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

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