Econ 2120: Section 1

Part II - The Conditional Expectation

Ashesh Rambachan

Fall 2018

Outline

The Conditional Expectation

- Optimal prediction
- Projection interpretation
- Best linear approximation
- Limit of flexible linear predictors

Discrete Regressors

Outline

The Conditional Expectation
- Optimal prediction
- Projection interpretation
- Best linear approximation
- Limit of flexible linear predictors

Discrete Regressors

Optimal prediction

Recall: The solution to

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

is given by

$$g(X) = E[Y|X].$$

We interpreted this as: $E[Y|X]$ is the optimal mean-square-error predictor of Y.

Projection interpretation

We will now interpret the conditional expectation as an orthogonal projection.

Vector space: $\mathcal{H}$ is the space of all linear combinations of Y and $h(X)$, where $h(\cdot)$ is any function with $E[h(X)^2] < \infty$.

Subspace: $\mathcal{H}_X$ is the space of all random variables $h(X)$.

Inner product: $\langle W_1, W_2 \rangle = E[W_1 W_2]$.

Projection: Find $g \in H_X$ to solve

$$\min_{g \in H_X} \| Y - \hat{Y} \|^2$$

where $\hat{Y} = g(X)$. Let $U = Y - \hat{Y}$. Solution will satisfy

$$\langle U, h(X) \rangle = E[Uh(X)] = 0$$

for all $h \in H_x$.

Projection interpretation

We'll show that $E[Y|X]$ solves this orthogonality condition. We have

$$\begin{aligned} E[(Y - E[Y|X])h(X)] &= E_X[E_{Y|X}[(Y - E[Y|X])h(X)]] \\ &= E_X[E_{Y|X}[(Y - E[Y|X])]h(X)] = 0 \end{aligned}$$

So, the conditional expectation solves the minimum-norm problem when we project Y onto the space of all functions $h(X)$.

The best linear predictor solved this problem when we projected onto the space of linear functions of X.

Projection interpretation

We can also show that if

$$E[Ug(X)] = 0$$

for all g, then $E[U|X] = 0$.

By the projection theorem, the solution g^* to

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

satisfies $E[(U - g^*(X))g(X)] = 0$ and we know that $g^*(X) = E[U|X]$.

Why is this useful? You may be used to seeing someone write

$$Y = g(X) + U$$

and then make some assumptions about U. What they specify about U tells you what projection problem $g(X)$ solves i.e. what subspace we're projecting Y onto.

Projection interpretation

Interpreted $E[Y|X]$ as the orthogonal projection of Y onto the space of all functions of X.

Can we connect $E[Y|X]$ to the best linear predictor $E^*[Y|1, X]$? Yes - in two ways.

- $E^*[Y|1, X]$ as the best linear approximation to $E[Y|X]$.
- $E[Y|X]$ as the limit of increasingly flexible linear predictors.

$E^*[Y|1, X]$ as best linear approximation to $E[Y|X]$

For simplicity, let $r(X) = E[Y|X]$. This could be a very complex function and so, we may wish to approximate it.

Claim: $E^*[r(X)|1, X] = E^*[Y|1, X]$.

Let $U = Y - r(X)$. We know $U \perp g(X)$ for all g. In particular, we have $U \perp 1, X$. So,

$$E^*[U|1, X] = E^*[Y|1, X] - E^*[r(X)|1, X] = 0.$$

$\Rightarrow$ "The best linear predictor is the best linear approximation of the conditional expectation."

$\Rightarrow$ "The best linear predictor is the orthogonal projection of $E[Y|X]$ onto the space of linear functions of X."

$E[Y|X]$ as increasingly flexible linear predictors

Start with single variable X. Consider the best linear predictor of Y using a polynomial of order M:

$$E^*[Y|1, X, X^2, \ldots, X^M]$$

As M increases, the squared prediction error cannot increase since we are projecting onto larger and larger subspaces.

So

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

is decreasing in M and bounded below at 0. It must have a limit.

$E[Y|X]$ as increasingly flexible linear predictors

We assume that $E^*[Y|1, X, X^2, \ldots, X^M]$ has a limit and we call it

$$E[Y|X] = \lim_{M \to \infty} E^*[Y|1, X, X^2, \ldots, X^M].$$

Intuition:

As M increases, I am using increasingly flexible functions to predict Y.

As M increases, I am using increasingly flexible functions to approximate $E[Y|X]$.

$E[Y|X]$ as increasingly flexible linear predictors

With this definition of $E[Y|X]$, we know that $U = Y - E[Y|X]$ satisfies

$$U \perp X^j$$

for all $j \geq 0$. Because we can approximate general functions $g(X)$ by polynomials, we will have that

$$U \perp g(X)$$

.

Exercise

Let $E[Y|Z]$ be the conditional expectation of Y given Z. Let $U = Y - E[Y|Z]$ and let

$$V(Y|Z) = E[(Y - E[Y|Z])^2|Z] = E[U^2|Z]$$

be the conditional variance of Y given Z.

(1) Show that

$$V(U) = E[U^2] = E[V(Y|Z)].$$

(2) Show that

$$V(Y) = V(E[Y|Z]) + E[V(Y|Z)].$$

Outline

The Conditional Expectation
- Optimal prediction
- Projection interpretation
- Best linear approximation
- Limit of flexible linear predictors

Discrete Regressors

Discrete regressors

We interpreted $E^*[Y|1, X]$ as the best linear approximation of $E[Y|X]$.

When will $E^*[Y|1, X] = E[Y|X]$? Whenever $E[Y|X]$ is linear!

One case: whenever the regressors are **discrete**.

> The conditional expectation is a linear function of appropriately defined transformations of the discrete regressors.

Discrete regressors

To match the notation from lecture, consider regressors Z_1, Z_2. Assume that they take only a finite set of values

$$Z_1 \in \{\lambda_1, \ldots, \lambda_J\}$$

$$Z_2 \in \{\delta_1, \ldots, \delta_K\}.$$

Define

$$X_{jk} = \begin{cases} 1, & Z_1 = \lambda_j, \quad Z_2 = \delta_k \\ 0, & \text{otherwise} \end{cases}$$

or $X_{jk} = 1(Z_1 = \lambda_j, Z_2 = \delta_k)$.

Discrete regressors

Claim:

$$E[Y|Z_1, Z_2] = E^*[Y|X_{11}, \ldots, X_{J1}, \ldots, X_{1K}, \ldots, X_{JK}].$$

Why? 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 $\gamma_{jk} = g(\lambda_j, \delta_k)$.

Discrete regressors: sample analog

The data are $\{(y_i, z_{i,1}, z_{i,2})\}_{i=1}^n$. Construct the dummies

$$x_{i,jk} = 1(z_{i,1} = \lambda_j, z_{i,2} = \delta_k)$$

and let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x_{jk} = \begin{pmatrix} x_{1,jk} \\ \vdots \\ x_{n,jk} \end{pmatrix}$$

for $j = 1, \ldots, J$, $k = 1, \ldots, K$.

Discrete regressors: sample analog

The coefficients of the least-squares fit are

$$\min \|y - \sum_j \sum_k b_{jk} x_{jk}\|^2.$$

Claim:

$$b_{jk} = \frac{\sum_i y_i x_{i,jk}}{\sum_i x_{i,jk}} = \bar{y}_{jk}$$

where $\bar{y}_{jk}$ is the sample average of y_i over values with $z_{i1} = \lambda_j$ and $z_{ik} = \delta_k$.

Discrete Regressors: sample analog

Why? We know

$$\langle y - \sum_j \sum_k b_{jk} x_{jk}, x_{lm} \rangle = 0$$

for each l, m. Moreover, the dummies are orthogonal to each other

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

unless $j = l$, $k = m$. So, we have that

$$\langle y - \sum_j \sum_k b_{jk} x_{jk}, x_{lm} \rangle = \langle y, x_{lm} \rangle - b_{lm} \langle x_{lm}, x_{lm} \rangle = 0$$

.