Section 2*

Econ 240A - Second Half

Ingrid Haegele†
University of California, Berkeley

October 19, 2018

*These section notes rely on the notes prepared and revised by Markus Pelger, Raffaele Saggio and Seongjoo Min.

†E-mail: inha@berkeley.edu

1 Conditional Expectation

The classical linear regression theory is basically a side product of two concepts: (i) conditional expectation and (ii) projection theory. In this section, I am going to start by properly defining the concept of conditional expectation focusing on its probabilistic foundation.

Notation: I will follow the lecture notes in the sense that capital letters, such as X, denote random variables. If we write $X = x$ it means that the random variable X is taking one particular value in its support, namely x.

Definition 1: Let (X,Y) denote two continuous random variables. The joint CDF of (X,Y) is

$$F(x, y) = \Pr(X \leq x, Y \leq y) \tag{1}$$

The joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) \tag{2}$$

Definition 2: For any measurable function $g(x, y)$

$$\mathrm{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \tag{3}$$

Definition 3: The marginal distribution of X is

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dx dy \tag{4}$$

Definition 4: X and Y are independent if and only if

$$f(x, y) = f_X(x) f_Y(y) \tag{5}$$

Corollary: If X and Y are independent (which we usually write as $X \perp Y$) then $\mathrm{E}(g(X)h(Y)) = \mathrm{E}(g(X))\,\mathrm{E}(h(Y))$.

Definition 5: Let $X = x$ with $f_X(x) > 0$. The conditional density of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} \tag{6}$$

Definition 6: The conditional mean or **conditional expectation function (CEF)** of Y given $X = x$ is defined as

$$\mathrm{E}(Y|X = x) = \mu_{Y|X}(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \tag{7}$$

The CEF is a generalization of the mean, providing a measure of the central tendency of one variable given another. The notation $\mu_{Y|X}(x)$ stresses the fact that this is a function, in the sense that when evaluated at $X = x$ it outputs a number. For instance, if Y is earnings and X is education, then $\mu_{Y|X}(6)$ is the expected value of earnings in the subpopulation that has six years of schooling. Notice that if we write $\mathrm{E}(Y|X) = \mu_{Y|X}(X)$ then this object now represents a random variable.

Another important concept we need to introduce is the Law of Iterated Expectations.
Proposition 1: *Simple Law of Iterated Expectations*

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y] \tag{8}$$

Proposition 2: *Law of Iterated Expectations*

$$\mathrm{E}[\mathrm{E}[Y|X, Z]|X] = \mathrm{E}[Y|X]. \tag{9}$$

Proposition 3: *Conditioning Theorem*
For any function g(x):

$$\mathrm{E}[g(X)Y|X] = g(X)\,\mathrm{E}[Y|X] \tag{10}$$

1.1 The Conditional Expectation Function Error

Suppose I want to predict the wage of an individual in the United States last year. Let's call this individual i. A natural starting point is the (unconditional) mean of wages. But we can do more. Suppose you have information on the individual's gender. In most countries, it is probably the case that $\mathrm{E}(wage|gender = male) > \mathrm{E}(wage|gender = female)$.

Hence we can improve our initial guessing point, the unconditional mean, by using $\mu_{Y|X}(male)$ if i is a man or $\mu_{Y|X}(female)$ if i is a female. Clearly we can keep iterating

on our prediction by using more and more information in our conditioning set for i, i.e. X_i, by letting X_i include information on $\{gender, education, race, city, AFQTscore, \ldots\}$.

At the end of the day, we can use the information included in X to predict as best as we can the random variable Y_i, the wage of individual i. However, it is probably the case that we cannot have *all* the possible information on i, hence we are always ending up having an error when predicting the random variable Y_i using $\mathrm{E}(Y_i|X_i = x)$.

Definition 7: The CEF error e is defined as follows

$$e = Y - \mu_{Y|X}(X). \tag{11}$$

Notice that therefore any random variable, Y can be written as follows

$$Y = \mu_{Y|X}(X) + e \tag{12}$$

We can easily derived the properties of e using the linearity and iteration properties of the conditional expectation.

Property 1: *The CEF error is conditionally mean zero*

$$\mathrm{E}[e|X] = 0 \tag{13}$$

Proof:

$$\mathrm{E}[e|X] = \mathrm{E}(Y - \mu_{Y|X}(X)|X) = \mu_{Y|X}(X) - \mu_{Y|X}(X) = 0 \tag{14}$$

Property 2: *The CEF error is unconditionally mean zero*

$$\mathrm{E}(e) = 0 \tag{15}$$

Proof:

$$\mathrm{E}[e] = \mathrm{E}[\mathrm{E}(e|X)] = \mathrm{E}(0) = 0 \tag{16}$$

Property 3: *The CEF error is uncorrelated with any function of X*
Let h(X) be such that $\mathrm{E}\,|h(X)e| < \infty$. Then e is uncorrelated with h(X), i.e.

$$\mathrm{E}[h(X)e] = \mathrm{E}[h(X)\,\mathrm{E}(e|X)] = 0 \tag{17}$$

2 Linear Predictor

2.1 Motivation

We have learned that the conditional mean $E[Y|X]$ is the best predictor of Y among all functions of X in the sense that it has the lowest mean squared error. However, in practice the particular functional form of the CEF is unknown. Unless we have few discrete covariates and fully saturate the model (including interaction terms), the true CEF is unlikely to be linear. Nevertheless, we can define an approximation to the CEF by the linear function that minimizes the mean squared prediction error among all linear predictors. We call this the best linear predictor (BLP).

2.2 Regularity Conditions

In order to derive the best linear predictor, we need to introduce auxiliary assumptions, the so-called regularity conditions.

1. $\mathbb{E}[Y^2] < \infty$
2. $\mathbb{E}\|X\|^2 = \mathbb{E}[X'X] < \infty$
3. $\mathbb{E}[(a'X)^2] > 0, \ \forall \ a \in \mathbb{R}^K, \ a \neq 0$

Assumptions 1 and 2 imply that the variables Y and X have finite means, variances and covariances, which is a required assumption in order to guarantee that $E[XY]$ and $E[XX]$ exist.

Assumptions 1 and 2 imply that $\mathbb{E}[XY]$ exists:

$$\begin{aligned} \|\mathbb{E}[XY]\| &\leq \mathbb{E}[\|XY\|] && \text{(by Jensen's Ineqaulity)} && (18)\\ &\leq \mathbb{E}[\|X\|^2]^{1/2}\mathbb{E}[Y^2]^{1/2} && \text{(by Cauchy-Schwartz Inequality)} && (19)\\ &< \infty && && (20) \end{aligned}$$

Assumption 2 implies that $\mathbb{E}[XX']$ exists:

$$\begin{aligned}
\|\mathbb{E}[XX']\|_F &\leq \mathbb{E}[\|XX'\|_F] && \text{(by Jensen's Ineqaulity)} \\
&= \mathbb{E}[\sqrt{tr(XX'XX')}] \\
&= \mathbb{E}[\sqrt{tr\big(X(X'X)X'\big)}] \\
&= \mathbb{E}\Big[\sqrt{X'X\ tr(XX')}\Big] \\
&= \mathbb{E}\big[\sqrt{(X'X)(X'X)}\big] \\
&= \mathbb{E}[X'X] \\
&< \infty
\end{aligned}$$

where $\|\cdot\|_F$ is called the Frobenius norm. For an $N \times K$ matrix A, it is defined as

$$\|A\|_F = \sqrt{\sum_{n=1}^{N}\sum_{k=1}^{K}[A]_{nk}^2} = \sqrt{tr(A'A)}$$

Assumption 3 imposes that the matrix $E[XX']$ is of full rank and thus invertible.

2.3 Derivation of BLP

Let Y denote a scalar random variable and X a K x 1 random vector. The population linear prediction of Y on X is defined as $E^*[Y|X] = X'\beta$ where β is chosen to minimize the risk under squared error loss:

$$\beta = \underset{b\in\mathbb{R}^K}{\operatorname{argmin}}\, E\left[(Y - X'b)^2\right]$$

The FOC of this minimization problem $E[X(Y - X'\beta)]$ implies that the prediction error U (defined as $Y - X'\beta$) is uncorrelated with X.
By rearranging we obtain the expression for β :

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

.

This leads to the defintion of the **best linear predictor (BLP)** of Y given X (under squared error loss) as

$$\mathbb{E}^*[Y|X] = X'\beta$$

so that

$$Y = \mathbb{E}^*[Y|X] + U_i$$

where the projection error is given by:

$$U = Y - \mathbb{E}^*[Y|X]$$

By the properties of linear projections and assuming X contains a constant, two important properties of the projection error are $E[XU] = 0$ and $E[U] = 0$.

2.4 Properties of BLP

The BLP is the best fitting linear approximation to the conditional expectation function for Y. To see this rewrite the minimization problem as follows:

$$\begin{aligned}
\beta &= \underset{b}{\operatorname{argmin}}\, E\left[(Y - X'b)^2\right] \\
&= \underset{b}{\operatorname{argmin}}\, E\left[(Y - E[Y|X] + E[Y|X] - X'b)^2\right] \\
&= \underset{b}{\operatorname{argmin}}\, E\left[(Y - E[Y|X])^2 + (E[Y|X] - X'b)^2 + 2(Y - E[Y|X])(E[Y|X] - X'b)\right] \\
&= \underset{b}{\operatorname{argmin}}\, E\left[(E[Y|X] - X'b)^2\right]
\end{aligned}$$

where the last line follows from the fact that $E\left[(Y - E[Y|X])^2\right]$ is a constant and $E[2(Y - E[Y|X])(E[Y|X] - X'b)] = 0$ by iterated expectations. This way of writing the minimization problem reveals that the BLP chooses the coefficient vector b that minimizes its average squared deviation from the conditional expectation function.

Moreover, the BLP ($E^*[Y|X] = X'\beta$) coincides with the conditional expectation if the CEF is linear (if $E[Y|X] = X'b$):

$$\begin{aligned} E^*[Y|X] &= X'\beta \\ &= X'E[XX']^{-1}E[XY] \\ &= X'E[XX']^{-1}E[XE[Y|X]] \\ &= X'E[XX']^{-1}E[XX]b \\ &= X'b \\ &= E[Y|X] \end{aligned}$$

Another property of the BLP is the **law of iterated projections**:

$$E^*[Y|X] = E^*[E^*[Y|X, Z]|X]$$

This property is an important tool, for example if one is interested in comparing a long to a short regression.

2.5 Short and Long Regression

Understanding how long and short regressions compare is fundamental to understanding central topics in econometrics such as omitted variable bias. To develop more intution, we want to derive and compare the regression coefficients from both the long and the short regression.

The **long regression function** predicts Y only based on X and W

$$\mathbb{E}^*[Y|X, W] = X'\beta + W'\gamma$$

The **short regression function** predicts Y only based on X

$$\mathbb{E}^*[Y|X] = X'\delta$$

Which of the regressions (and thus which of the coefficients β or δ) is more interesting to the researcher depends on the specific economic question we are after.

Also, note that we can define the an **auxiliary regression function** that predicts W given X as

$$\mathbb{E}^*[W|X] = \Pi' X$$

As before, let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^K$. Also suppose thta $W \in \mathbb{R}^J$. Then Π is a $K \times J$ matrix defined as

$$\Pi = \mathbb{E}[XX']^{-1}\mathbb{E}[XW']$$

Define the error of the long regressionl as $U = Y - X'\beta - W'\gamma$. By construction, this error is uncorrelated with X and W, which implies $\mathbb{E}^*[U|X] = 0$. Note that we can write Y = $\mathrm{X}'\beta + W'\gamma + U$.

In order to examine the short regression coefficients, we rewrite the expression for the short regression $\mathbb{E}^*[Y|X] = X'\delta$ as follows:

$$\begin{aligned}
\mathbb{E}^*[Y|X] &= \mathbb{E}^*[X'\beta + W'\gamma + U|X] && (21)\\
&= \mathbb{E}^*\Big[X'\beta + W'\gamma|X\Big] + \underbrace{\mathbb{E}^*[U|X]}_{=0} && (22)\\
&= X'\beta + \mathbb{E}^*[W|X]'\gamma && (23)\\
&= X'\beta + X'\Pi\gamma && (24)\\
&= X'(\beta + \Pi\gamma) && (25)
\end{aligned}$$

This derivation reveals that the short coefficient δ equals the long coeffcient β plus the coeffcient on the omitted variable γ times the regression of the omitted variable on the included variable. This is the classical omitted variable bias formula. Based on this formula, we can predict how regression coefficients change when we add more regressors.

2.6 Firsch-Waugh-Lovell Theorem

The Frisch-Waugh-Lovell theorem allows us to fully clarify the anatomy of long regressions and is an important tool to know. It is also often referred to as residual linear regression. Consider $Y = X'\beta + W'\gamma + U$ as before. Start with an auxiliary regression of Y on W, which can be denoted as $\tilde{Y} = Y - \mathbb{E}^*[Y|W]$. Smilarily, consider the regression of X on W

as $\tilde{X} = X - \mathbb{E}^*[X|W]$. The Frisch-Waugh-Lovell theorem states that

$$\mathbb{E}^*[\tilde{Y}|\tilde{X}] = \tilde{X}'\beta$$

Why is the Frisch-Waugh-Lovell theorem useful? It tells us that we can produce the long regression coefficientβ by first partialling out W from X and Y and then regressing $\tilde{Y}$ on $\tilde{X}$. This shows that the long regression coefficient uses only the variation in X that cannot be predicted by W.