

**Notation:** in this note,  $y$  is a scale random variable, and  $x$  is a  $K \times 1$  random vector.

## 1 Conditional Expectation

We view the regression as a problem of supervised learning. Supervised learning uses a function of  $x$ , say,  $g(x)$ , to predict  $y$ . We do not  $x$  can perfectly predict  $y$ ; otherwise their relationship is deterministic. The prediction error

$$e = y - g(x)$$

depends on the choice of  $g$ . There are numerous possible choices of  $g$ . Which one is the best? To answer this question, we need to decide a criterion to compare different  $g$ , which is called the *loss function*  $L(y, g(x))$ . A particularly convenient one is the *quadratic loss*, defined as

$$L(y, g(x)) = (y - g(x))^2.$$

Since the data is random, we average the loss function across the joint distribution of  $(y, x)$  as  $R(y, g(x)) = E[L(y, g(x))]$ , which is called the *risk*. For the quadratic loss function, the corresponding risk

$$R(y, g(x)) = E[(y - g(x))^2],$$

is called the *mean squared error* (MSE). MSE is a deterministic quantity since the randomness is integrated out.

What is the optimal choice of  $g$  if we aim to minimize the MSE?

**Proposition 1.** *The conditional mean function  $m(x) = E[y|x]$  minimizes MSE.*

Before we prove the above proposition, we first discuss some properties of the conditional mean function. Obviously

$$y = m(x) + (y - m(x)) = m(x) + \epsilon,$$

where  $\epsilon \equiv y - m(x)$  is called the *regression error*. This equation holds for  $(y, x)$  following any joint distribution, as long as  $E[y|x]$  exists. The error term  $\epsilon$  satisfies these properties:

- $E[\epsilon|x] = E[y - m(x)|x] = E(y|x) - m(x) = 0$ ,
- $E[\epsilon] = E[E[\epsilon|x]] = E(0) = 0$ ,
- For any function  $h(x)$ , we have  $E[h(x)\epsilon] = E[E[h(x)\epsilon|x]] = E[h(x) \cdot E[\epsilon|x]] = 0$ .

The last property implies that  $\epsilon$  is uncorrelated with any function of  $x$ . In particular, when  $h$  is the identity function  $h(x) = x$ , we have  $E[x\epsilon] = \text{cov}(x, \epsilon) = 0$ .

*Proof of Proposition 1.* The optimality of the conditional mean can be confirmed by “guess-and-verify.” For an arbitrary  $g(x)$ , the MSE can be decomposed into three terms

$$\begin{aligned} E[(y - g(x))^2] &= E[(y - m(x) + m(x) - g(x))^2] \\ &= E[(y - m(x))^2] + 2E[(y - m(x))(m(x) - g(x))] + E[(m(x) - g(x))^2]. \end{aligned}$$

.The first term is irrelevant to  $g(x)$ . The second term

$$\begin{aligned} 2E[(y - m(x))(m(x) - g(x))] &= 2E[\epsilon(m(x) - g(x))] \\ &= 2E[E[\epsilon(m(x) - g(x)) | x]] \\ &= 2E[(m(x) - g(x)) E[\epsilon | x]] = 0. \end{aligned}$$

is again irrelevant of  $g(x)$ . The third term, obviously, is minimized at  $g(x) = m(x)$ .  $\square$

Our perspective so far deviates from mainstream econometric textbooks, most of which start the regression model by assuming that the dependent variable  $y$  is generated by, or is modeled as,  $g(x) + \epsilon$  for some unknown function  $g(\cdot)$  and some error term  $\epsilon$  such that  $E[\epsilon | x] = 0$ . Instead, we take a predictive framework regardless the data generating process. What we observe are  $y$  and  $x$  and we are solely interested in seeking a function  $g(x)$  to predict  $y$  as accurately as possible under the MSE criterion.

## 2 Linear Projection

As discussed in the previous section, the conditional mean function  $m(x)$  is the function that minimizes the MSE. However,

$$m(x) = E[y | x] = \int y f(y | x) dy$$

is a complex function of  $x$ , as it depends on the joint distribution of  $(y, x)$ , which is almost always unknown in practice. Now let us make the prediction task even simpler. How about we minimize the MSE within all linear functions in the form of  $g(x) = x'b$  for  $\beta \in \mathbb{R}^K$ ? The minimization problem is

$$\min_{b \in \mathbb{R}^K} E[(y - x'b)^2]. \quad (1)$$

Take the first-order condition of the MSE

$$\frac{\partial}{\partial b} E[(y - x'b)^2] = -2E[x(y - x'b)] = 0.$$

Rearrange the above equation and we solve the optimal  $b$  as

$$\beta = (E[xx'])^{-1} E[xy]$$

if  $E[xx']$  is invertible. The function  $x'\beta$  is called the *best linear projection* of  $y$  on  $x$ , and the vector  $\beta$  is called the *linear projection coefficient*.

*Remark 2.* The linear function is not as restrictive as one might thought. It can be used to produce some nonlinear (in random variables) effect if we re-define  $x$ . For example, if

$$y = x_1\beta_2 + x_2\beta_2 + x_1^2\beta_3 + e,$$

then  $\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + x_1\beta_3$ , which is nonlinear in  $x_1$ , while it is still linear in the parameter  $\beta$  if we define a set of new regressors as  $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = (x_1, x_2, x_1^2)$ .

*Remark 3.* If  $(y, x)$  is jointly normal in the form  $\begin{pmatrix} y \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}\right)$  where  $\rho$  is the correlation coefficient, then

$$E[y|x] = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) = \left( \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x \right) + \rho \frac{\sigma_y}{\sigma_x} x,$$

is a linear function of  $x$ . In this example, the conditional mean coincides with a linear function.

*Remark 4.* Even though in general  $m(x) \neq x'\beta$ , the linear form  $x'\beta$  is still useful in approximating  $m(x)$ . That is,  $\beta = \arg \min_{b \in \mathbb{R}^k} E[(m(x) - x'b)^2]$ .

*Proof.* The first-order condition gives  $\frac{\partial}{\partial b} E[(m(x) - x'b)^2] = -2xE[(m(x) - x'b)] = 0$ . Rearrange the terms and obtain  $E[x \cdot m(x)] = E[xx']\beta$ . When  $E[xx']$  is invertible, we solve

$$\beta = (E[xx'])^{-1} E[x \cdot m(x)] = (E[xx'])^{-1} E[E(xy|x)] = (E[xx'])^{-1} E(xy).$$

Thus  $\beta$  is also the best linear approximation to  $m(x)$  under MSE.  $\square$

We may rewrite the linear regression model, or the *linear projection model*, as

$$\begin{aligned} y &= x'\beta + e \\ E[xe] &= 0, \end{aligned}$$

where  $e = y - x'\beta$  is called the *projection error*, to be distinguished from  $\epsilon = y - m(x)$ .

**Exercise:** show (a)  $E[xe] = 0$ . (b) If  $x$  contains a constant, then  $E[e] = 0$ .

## 2.1 Omitted Variable Bias

We write the *long regression* as

$$y = x'_1\beta_1 + x'_2\beta_2 + \beta_3 + \epsilon,$$

and the *short regression* as

$$y = x'_1\gamma_1 + \gamma_2 + u.$$

If  $\beta_1$  in the long regression is the parameter of interest, omitting  $x_2$  as in the short regression will render *omitted variable bias* (meaning  $\gamma_1 \neq \beta_1$ ) unless  $x_1$  and  $x_2$  are uncorrelated.

We first demean all the variables in the two regressions, which is equivalent as if we project out the effect of the constant. The long regression becomes

$$\tilde{y} = \tilde{x}'_1\beta_1 + \tilde{x}'_2\beta_2 + \tilde{\epsilon},$$

and the short regression becomes

$$\tilde{y} = \tilde{x}'_1\gamma_1 + \tilde{u},$$

where *tilde* denotes the demeaned variable.

After demeaning, the cross-moment equals to the covariance. The short regression coefficient

$$\begin{aligned}
\gamma_1 &= (E [\tilde{x}_1 \tilde{x}'_1])^{-1} E [\tilde{x}_1 \tilde{y}] \\
&= (E [\tilde{x}_1 \tilde{x}'_1])^{-1} E [\tilde{x}_1 (\tilde{x}'_1 \beta_1 + \tilde{x}'_2 \beta_2 + e)] \\
&= (E [\tilde{x}_1 \tilde{x}'_1])^{-1} E [\tilde{x}_1 \tilde{x}'_1] \beta + (E [\tilde{x}_1 \tilde{x}'_1])^{-1} E [\tilde{x}_1 \tilde{x}'_2] \beta_2 \\
&= \beta_1 + (E [\tilde{x}_1 \tilde{x}'_1])^{-1} E [\tilde{x}_1 \tilde{x}'_2] \beta_2.
\end{aligned}$$

Therefore,  $\gamma_1 = \beta_1$  if and only if  $E [\tilde{x}_1 \tilde{x}'_2] \beta_2 = 0$ , which demands either  $E [\tilde{x}_1 \tilde{x}'_2] = 0$  or  $\beta_2 = 0$ .

Obviously we prefer to run the long regression to attain  $\beta_1$  if possible, as it is a model general model than the short regression. However, sometimes  $x_2$  is simply unobservable so the long regression is infeasible. When only the short regression is available, in some cases we are able to sign the bias, meaning that we know whether  $\gamma_1$  is bigger or smaller than  $\beta_1$ .