Chapter 2

Regression and Projection

Notation: In this note, y is a scale random variable, and $x = (x_1, \ldots, x_K)'$ is a $K \times 1$ random vector. Throughout this course, a vector is a *column* vector, i.e. a one-column matrix.

2.1 Conditional Expectation

Up to now, we know the conditional expectation is an important object for causality interpretation under CIA. Next, we motivate it from the perspective of prediction. Supervised learning uses a function of x, say, $g(x)$, to predict y. x cannot perfectly predict y; otherwise their relationship is deterministic. The prediction error $y - g(x)$ depends on the choice of g. There are numerous possible choices of g. Which one is the best? Notice that this question is not concerned about the underlying data generating process (DGP) of the joint distribution of (y, x). We want to find a general rule to achieve accurate prediction of y given x, no matter how this pair of variables is generated.

To answer this question, we need to decide a criterion to compare different g. Such a criterion 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 are random, $L(y, g(x))$ is also random. "Random" means uncertainty: sometimes *this* happens, and sometimes *that* happens. To get rid of the uncertainty, we average the loss function with respect to the joint distribution of (y, x) as $R(y, g(x)) = E[L(y, g(x))]$, which is called *risk*. Risk is a deterministic quality. For the quadratic loss function, the corresponding risk is

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

is called the *mean squared error* (MSE). MSE is the most widely used risk measure, although there exist many alternative measures, for example the *mean absolute error* (MAE) $E[|y - g(x)|]$. The popularity of MSE comes from its convenience for analysis in closed-form, which MAE does not enjoy due to its nondifferentiability. This is similar to the choice of utility functions in economics. There are only a few functional forms for the utility, for example CRRA, CARA, and so on. They are popular because they lead to close-form solutions that are easy to handle. Now our quest is narrowed to: What is the optimal choice of g if we minimize the MSE?

Proposition 2.1. *The CEF $m(x)$ 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 $\epsilon := y - m(x)$ 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 ϵ 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

$$E[h(x)\epsilon] = E[E[h(x)\epsilon|x]] = E[h(x)E[\epsilon|x]] = 0. \tag{2.1}$$

The last property implies that ϵ 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 2.1. The optimality of the CEF can be confirmed by "guess-and-verify." For an arbitrary $g(x)$, the MSE can be decomposed into three terms

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

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

$$2E[(y - m(x))(m(x) - g(x))] = 2E[\epsilon(m(x) - g(x))] = 0$$

by invoking (2.1) with $h(x) = m(x) - g(x)$. The second term is again irrelevant of $g(x)$. The third term, obviously, is minimized at $g(x) = m(x)$. □

Our perspective so far deviates from many econometric textbooks that assume that the dependent variable y is generated as $g(x) + \epsilon$ for some unknown function $g(\cdot)$ and error term ϵ such that $E[\epsilon|x] = 0$. Instead, we take a predictive approach regardless the DGP. 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.

2.2 Linear Projection

The CEF $m(x)$ is the function that minimizes the MSE. However, $m(x) = E[y|x]$ is a complex function of x, for it depends on the joint distribution of (y, x), which is mostly 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 $h(x) = h(x; b) = x'b$ for $b \in \mathbb{R}^K$? The minimization problem is

$$\min_{b \in \mathbb{R}^K} E\left[\left(y - x'b\right)^2\right]. \tag{2.2}$$

Take the first-order condition of the MSE

$$\frac{\partial}{\partial b} E\left[\left(y - x'b\right)^2\right] = E\left[\frac{\partial}{\partial b}\left(y - x'b\right)^2\right] = -2E\left[x\left(y - x'b\right)\right],$$

where the first equality holds if $E\left[\left(y - x'b\right)^2\right] < \infty$ so that the expectation and partial differentiation is interchangeable, and the second equality hods by the chain rule and the linearity of expectation. Set the first order condition to 0 and we solve

$$\beta = \arg\min_{b \in \mathbb{R}^K} E\left[\left(y - x'b\right)^2\right]$$

in the closed-form

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

if $E[xx']$ is invertible. Notice here that b is an arbitrary K-vector, while β is the optimizer. The function $x'\beta$ is called the *best linear projection* (BLP) of y on x, and the vector β is called the *linear projection coefficient*.

Remark 2.1. 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

$$y = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + e,$$

then $\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + 2x_1\beta_3$, which is nonlinear in x_1, while it is still linear in the parameter $\beta = (\beta_1, \beta_2, \beta_3)$ if we define a set of new regressors as $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = \left(x_1, x_2, x_1^2\right)$.

Remark 2.2. If (y, x) is jointly normal in the form

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}\right)$$

where ρ 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 liner function of x. In this example, the CEF is linear.

Remark 2.3. Even though in general $m(x) \neq x'\beta$, the linear form $x'\beta$ is still useful in approximating $m(x)$. That is, $\beta = \arg\min_{b \in \mathbb{R}^K} E\left[(m(x) - x'b)^2\right]$.

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

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

Thus β is also the best linear approximation to $m(x)$ under MSE. □

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

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

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

Exercise 2.1. Show (a) $E[xe] = 0$. (b) If x contains a constant, then $E[e] = 0$.

2.2.1 Omitted Variable Bias

We write the *long regression* as

$$y = x_1'\beta_1 + x_2'\beta_2 + \beta_3 + e_\beta,$$

and the *short regression* as

$$y = x_1'\gamma_1 + \gamma_2 + e_\gamma,$$

where e_β and e_γ are the projection errors, respectively. If β_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

$$\tilde{y} = \tilde{x}_1'\beta_1 + \tilde{x}_2'\beta_2 + \tilde{e}_\beta,$$

and the short regression becomes

$$\tilde{y} = \tilde{x}_1'\gamma_1 + \tilde{e}_\gamma,$$

where *tilde* denotes the demeaned variable.

Exercise. Show $\tilde{e}_\beta = e_\beta$ and $\tilde{e}_\gamma = e_\gamma$.

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

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

where the third line holds as $E\left[\tilde{x}_1\tilde{e}_\beta\right]=0$. Therefore, $\gamma_1=\beta_1$ if and only if $E\left[\tilde{x}_1\tilde{x}_2'\right]\beta_2=0$, which demands either $E\left[\tilde{x}_1\tilde{x}_2'\right]=0$ or $\beta_2=0$.

Exercise 2.2. Show that $E\left[\left(y-x_1'\beta_1-x_2'\beta_2-\beta_3\right)^2\right]\leq E\left[\left(y-x_1'\gamma_1-\gamma_2\right)^2\right]$.

Obviously we prefer to run the long regression to attain β_1 if possible, for it is a more general model than the short regression and achieves no larger variance in the projection error. However, sometimes x_2 is unobservable so the long regression is unavailable. This example of omitted variable bias is ubiquitous in applied econometrics. Ideally we would like to directly observe some regressors but in reality we do not have them at hand. We should be aware of the potential consequence when the data are not as ideal as we have wished. When only the short regression is available, in some cases we are able to sign the bias, meaning that we can argue whether γ_1 is bigger or smaller than β_1 based on our knowledge.

`Zhentao Shi. January 10, 2026`