Lecture 2: Regression Model

Zhentao Shi

September 13, 2018

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

1 Conditional Expectation

A regression model can be written as

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

where $m(x) = E[y|x]$ is called the *conditional mean function*, and $\epsilon = y - m(x)$ is called the *regression error*. Such an equation holds for (y, x) that follows any joint distribution, as long as $E[y|x]$ exists. The error term ϵ satisfies these properties:

- $E[\epsilon|x] = 0$,
- $E[\epsilon] = 0$,
- $E[h(x)\epsilon] = 0$, where h is a function of x.

The last property implies that ϵ is uncorrelated with any function of x.

If we are interested in predicting y given x, then the conditional mean function $E[y|x]$ is "optimal" in terms of the *mean squared error* (MSE).

As y is not a deterministic function of x, we cannot predict it with certainty. In order to evaluate different methods of prediction, we must have a criterion for comparison. For an arbitrary prediction method $g(x)$, we employ a *loss function* $L(y, g(x))$ to measure how wrong is the prediction, and the expected value of the loss function is called the *risk*, and is denoted as $R(y, g(x))$.

There are many choices of loss functions. A particularly convenient one is the *quadratic loss function*, defined as

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

The risk corresponding to the quadratic loss is

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

and it is called the MSE.

Due to its operational ease, MSE is one of the most widely used criterion. Under MSE, the conditional expectation function happens to be the best prediction method for y given x. In other words, the conditional mean function $m(x)$ minimizes the MSE.

The claimed optimality can be confirmed by "guess-and-verify." For an arbitrary $g(x)$, the risk is decomposed into three terms

$$E\left[(y-g(x))^2\right] = E\left[(y-m(x))^2\right] + 2E\left[(y-m(x))(m(x)-g(x))\right] + E\left[(m(x)-g(x))^2\right].$$

The first term is irrelevant to $g(x)$. The second term $2E\left[\epsilon(m(x)-g(x))\right] = 0$ is again irrelevant of $g(x)$. The third term, obviously, is minimized at $g(x) = m(x)$.

2 Linear Projection

As discussed in the previous section, we are interested in the conditional mean function $m(x)$. However, remind that

$$m(x) = E[y|x] = \int yf(y|x)\,\mathrm{d}y$$

is a complex function of x, as it depends on the joint distribution of (y, x).

A particular form of the conditional mean function is a linear function

$$m(x) = x'\beta.$$

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

$$y = x_1\beta_2 + x_2\beta_2 + x_1x_2\beta_3 + e,$$

then $\frac{\partial}{\partial x_1}m(x_1, x_2) = \beta_1 + x_2\beta_3$, which is nonlinear in x_1, while it is still linear in the parameter β if we define a set of new regressors as $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = (x_1, x_2, x_1x_2)$.

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

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

Even though in general $m(x) \neq x'\beta$, the linear form $x'\beta$ is still useful as an approximation, as will be clear soon. Therefore, we may write the linear regression model, or the *linear projection model*, as

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

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

If a constant is included in x as a regressor, we have $E[e] = 0$.

The coefficient β in the linear projection model has a straightforward closed-form. Multiplying x on both sides and taking expectation, we have $E[xy] = E[xx']\beta$. If $E[xx']$ is invertible, we can explicitly solve

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

Now we justify $x'\beta$ as an approximation to $m(x)$. Indeed, $x'\beta$ is the optimal *linear* predictor in terms of MSE; in other words,

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

This fact can be verified by taking the first-order condition of the above minimization problem

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

In the meantime, $x'\beta$ is also the best *linear* approximation to $m(x)$. If we replace y in the optimization problem by $m(x)$, we solve the minimizer as

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

Thus β is also the best linear approximation to $m(x)$ in terms of MSE.

2.1 Omitted Variable Bias

We write the *long regression* as

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

and the *short regression* as

$$y=x_1'\gamma_1+\gamma_2+u.$$

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

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

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

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

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 $\beta_2=0$.

Obviously we prefer to run the long regression to attain β_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 γ_1 is bigger or smaller than β_1.