Chapter 2

Regression, Projection and Causality

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

Machine learning is a big basket that contains the regression models. We motivate the conditional expectation model from the perspective of prediction. We view a regression as *supervised learning*. 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. 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 conditional mean function (CEF)* $m(x) = E[y|x] = \int yf(y|x)\,\mathrm{d}y$ *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

$$E[h(x)\epsilon] = E[E[h(x)\epsilon|x]] = E[h(x)E[\epsilon|x]] = 0.$$

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

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

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

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

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

$$\min_{b \in \mathbb{R}^K} E\left[\left(y - x'b\right)^2\right]. \tag{2.1}$$

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

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

if $E\left[xx'\right]$ 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 $\frac{\partial}{\partial x_1} m\left(x_1, x_2\right) = \beta_1 + 2x_1\beta_3$, 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

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

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

$$y = x'\beta + e$$
$$E[xe] = 0,$$

where $e = y - x'\beta$ is called the *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.

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

regression coefficient

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

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 *infeasible*. 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.

2.3 Causality

2.3.1 Structure and Identification

Unlike physical laws such as Einstein's mass–energy equivalence $E = mc^2$ and Newton's universal gravitation $F = Gm_1m_2/r^2$, economic phenomena can rarely be summarized in such a minimalistic style. When using experiments to verify physical laws, scientists often manage to come up with smart design in which signal-to-noise ratio is so high that small disturbances are kept at a negligible level. On the contrary, economic laws do not fit a laboratory for experiment. What is worse, the subjects in economic studies — human beings — are heterogeneous and with many features that are hard to control. People from distinctive cultural and family backgrounds respond to the same issue differently and researchers can do little to homogenize them. The signal-to-noise ratio in economic laws are often significantly lower than that of physical laws, mainly due to the lack of laboratory setting and the heterogeneous nature of the subjects.

Educational return and the demand-supply system are two classical topics in econometrics. A person's incomes is determined by too many random factors in the academic and career path that is impossible to exhaustively observe and control. The observable prices and quantities are outcomes of equilibrium so the demand and supply affect each other.

Generations of thinkers have been debating the definitions of causality. In economics, an accepted definition is *structural causality*. Structural

causality is a thought experiment. It assumes that there is a DGP that produces the observational data. If we can use data to recover the DGP or some features of the DGP, then we have learned causality or some implications of causality.

A key issue to resolve before looking at the realized sample is *identification*. We say a model or DGP is *identified* if the each possible parameter of the model under consideration generates distinctive features of the observable data. A model is *under-identified* if more than one parameter in the model can generate exact the same features of the observable data. In other words, a model is under-identified if from the observable data we cannot trace back to a unique parameter in the model. A correctly specified model is the prerequisite for any discussion of identification. In reality, all models are wrong. Thus when talking about identification, we are indulged in an imaginary world. If in such a thought experiment we still cannot unique distinguish the true parameter of the data generating process, then identification fails. In other words, we cannot determine what is the true model no matter how large the sample is.

2.3.2 Treatment Effect

We narrow down to the framework of the relationship between y and x. One question of particular interest is *treatment effect*. The treatment effect is how much y will change if we change a variable of interest, say d, by one unit while keeping all other variables (including the unobservable

variables) the same. The Latin phrase *ceteris paribus* means "keep all other things constant."

Example 2.1. During the 2020 covid-19 pandemic, Hong Kong's unemployment rate rose to a high-level and consumption collapsed. In order to boost the economy, some Hong Kong residents were qualified in receiving 10,000 HKD cash allowance from the government. We are interested to learn how much does the 10,000 HKD allowance increase people's consumption. For an individual, we imagine two parallel worlds: one with the cash allowance and one without. The difference of the consumption in the world with the allowance, denoted $Y(1)$, and that in the world without the allowance, denoted $Y(0)$, is the treatment effect of that particular person. This thought experiment is called the *potential outcome framework*.

However, in reality one and only one scenario happens, which echos the saying of ancient Greek philosopher Heraclitus (553 BC--475 BC): "You cannot step into the same river twice." While the individual treatment effect is infeasible, we evaluate the treatment effect at the population level. The *average treatment effect* (ATE) is defined as

$$ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)].$$

Notice that $E[Y(1)]$ and $E[Y(0)]$ are still infeasible. Now, let

$$D = 1\{\text{treatment received}\}$$

be a binary variable signifying the treatment, and $E[Y(1)|D=1]$ and $E[Y(0)|D=0]$ are feasible from the data. If the two potential outcomes $(Y(1), Y(0))$ are independent of the assignment D, then $E[Y(1)] = E[Y(1)|D=1]$ and $E[Y(0)] = E[Y(0)|D=0]$ so that ATE can be estimated from the data in a feasible way as

$$ATE = E[Y(1)|D=1] - E[Y(0)|D=0].$$

Therefore, to evaluate ATE ideally we would like use a lottery to randomly decide that some people receive the treatment (treatment group, with $D = 1$) and the others do not (control group, with $D = 0$).

When we have other control variables, we can also define a finer treatment effect conditional on x:

$$ATE(x) = E[Y(1)|x] - E[Y(0)|x].$$

ATE is the average effect in the population of individuals when we hypothetical give them the treatment, keeping all other factors x constant. If conditioning on x, the treatment D is independent of $(Y(1), Y(0))$, then ATE becomes feasible:

$$ATE(x) = E[Y(1)|D=1,x] - E[Y(0)|D=0,x]$$

The important condition $((Y(1), Y(0)) \perp D)|x$ is called the *conditional in-*

dependence assumption (CIA).

Example 2.2. CIA is more plausible than full independence. Consider the example $Y(1) = x + u(1)$, $Y(0) = x + u(0)$ and $D = 1\{x + u_d \geq 0\}$. If $((u(0), u(1)) \perp u_d) | x$, then CIA is satisfied. Obviously $(Y(1), Y(0))$ and D are dependent however, since x is involved in both random variables.

2.3.3 ATE and CEF

In the previous section the treatment D is binary. Now we consider a continuous treatment D. Suppose the DGP is $Y = h(D, x, u)$ where D and x are observable and u is unobservable. It is natural to define ATE with the continuous treatment (Hansen's book Chapter 2.30 calls it *average causal effect*) as

$$ATE(d, x) = E\left[\lim_{\Delta \to 0} \frac{h(d + \Delta, x, u) - h(d, x, u)}{\Delta}\right] = E\left[\frac{\partial}{\partial d} h(d, x, u)\right],$$

where the continuous differentiability of $h(d, x, u)$ at d is implicitly assumed. Unlike the binary treatment case, here d explicitly shows up in $ATE(d, x)$ because the effect can vary at different values of d. ATE here is the average effect in the population of individuals if we hypothetical move d a little bit, keeping all other factors x constant.

In the previous sections, we focused on the CEF $m(d, x)$, where d is added to x as an additional variable of interest. We did not intend to model the underlying economic mechanism $h(D, x, u)$, which may be very

complex. Can we learn the $ATE\left(d,x\right)$ which bears the structural causal interpretation, from the mechanical $m\left(d,x\right)$ which merely cares about best prediction? The answer is positive under CIA: $\left(u\perp D\right)|x$.

$$\begin{aligned}\frac{\partial}{\partial d}m\left(d,x\right) &= \frac{\partial}{\partial d}E\left[y|d,x\right] = \frac{\partial}{\partial d}E\left[h\left(d,x,u\right)|d,x\right] = \frac{\partial}{\partial d}\int h\left(d,x,u\right)f\left(u|d,x\right)du \\ &= \int \frac{\partial}{\partial d}\left[h\left(d,x,u\right)f\left(u|d,x\right)\right]du \\ &= \int \left[\frac{\partial}{\partial d}h\left(d,x,u\right)\right]f\left(u|d,x\right)du + \int h\left(d,x,u\right)\left[\frac{\partial}{\partial d}f\left(u|d,x\right)\right]du,\end{aligned}$$

where the second line implicitly assumes interchangeability between the integral and partial derivative. Under CIA, $\frac{\partial}{\partial d}f\left(u|d,x\right)=0$ and the second term drops out. Thus

$$\frac{\partial}{\partial d}m\left(d,x\right) = \int \left[\frac{\partial}{\partial d}h\left(d,x,u\right)\right]f\left(u|d,x\right)du = E\left[\frac{\partial}{\partial d}h\left(d,x,u\right)\right] = ATE\left(d,x\right).$$

This is an important result. It says that if CIA holds, we can learn the causal effect of d on y by the partial derivative of CEF conditional on x. In particular, if we further assume a linear CEF $m\left(d,x\right)=\beta_1 d+\beta_2'x$, then the causal effect is the coefficient β_1.

The key condition that links the CEF and the causal effect is CIA. CIA is not an innocuous assumption. In applications, our causal results are credible only when we can convincing defend CIA.

Exercise 2.3. Let factories' output be a Dobb-Douglas function $Y=AK^{\alpha}L^{\beta}$, where the capital level K and labor L as well as the output Y is observable,

while the "technology" A is unobservable. Take logarithm on both sides of the equation:

$$y = u + \alpha k + \beta l \tag{2.2}$$

where $y = \log Y$, $u = \log A$, $k = \log K$ and $l = \log L$. Suppose $\begin{pmatrix} u \\ k \\ l \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$ and $\alpha = \beta = 1/2$ make the true DGP. Here w and k are correlated, because factories of larger scale can afford robots to facilitate automation.

1. What is the partial derivative of CEF when we use k as a treatment variable for a fixed labor level l? (Hint: the CEF is a linear function thanks to the joint normality.)

2. Does it coincide with $\alpha = 1/2$, the coefficient in the causal model (2.2)? (Hint: No, because CIA is violated.)

Sometimes applied researchers assume by brute force that $y = m(d, x) + u$ is the DGP and $E[u|d, x] = 0$, where d is the variable of interest and x is the vector of other control variables. Under these assumptions,

$$ATE(d, x) = E\left[\frac{\partial}{\partial d}(m(d, x) + u)\,|d, x\right] = \frac{\partial m(d, x)}{\partial d} + \frac{\partial}{\partial d}E[u|d, x] = \frac{\partial m(d, x)}{\partial d},$$

where the second equality holds if $\frac{\partial}{\partial d} E[u|d,x] = E\left[\frac{\partial}{\partial d} u|d,x\right]$. At a first glance, it seems that the mean independence assumption $E[u|d,x] = 0$, which is weaker than CIA, implies the equivalence between $ATE(d,x)$ and $\partial m(d,x)/\partial d$ here. However, such slight weakening is achieved by a very strong assumption that the DGP $h(d,x,u)$ follows the additive separable form $m(d,x)+u$. Without economic theory to defend the choice of the assumed DGP $y = m(d,x)+u$, this is at best the *reduced-form* approach.

The *structural approach* here models the economic mechanism, guided by economic theory. The *reduced-form approach* is convenient and can document stylized facts when suitable economic theory is not immediately available. There are constant debates about the pros and cons of the two approaches; see *Journal of Economic Perspectives* Vol. 24, No. 2 Spring 2010. In macroeconomics, the so-called Phillips curve, attributed to A.W. Phillips about the negative correlation between inflation and unemployment, is a stylized fact learned from the reduced-form approach. The Lucas critique (Lucas, 1976) exposed its lack of microfoundation and advocated modeling deep parameters that are invariant to policy changes. The latter is a structural approach. Ironically, more 40 years has passed since the Lucas critique, equations with little microfoundation still dominate the analytical apparatus of central bankers. At present (as of 2020), applied econometric research in China is dominated by the reduced-form approach.

2.4 Summary

In this lecture, we cover the conditional mean function and causality. When we are faced with a pair of random variable (y, x) drawn from some joint distribution, the CEF is the best predictor. When we go further into the structural causality about some treatment d to the dependent variable y, under CIA we can find equivalence between ATE and the partial derivative of CEF. All analyses are conducted in population. We have not touched sample yet.

Historical notes: Regressions and conditional expectations are concepts from statistics and they are imported to econometrics in early time. Researchers at the Cowles Commission (now Cowles Foundation for Research in Economics) — Jacob Marschak (1898–1977), Tjalling Koopmans (1910–1985, Nobel Prize 1975), Trygve Haavelmo (1911–1999, Nobel Prize 1989) and their colleagues — were trailblazers of the econometric structural approach.

The potential outcome framework is not peculiar to economics. It is widely used in other fields such as biostatistics and medical studies. It was initiated by Jerzy Neyman (1894–1981) and extended by Donald B. Rubin (1943–), Professor of Statistics at Tsinghua University.

Extensive reading: Lewbel (2019) offers a comprehensive summary of identification in econometrics.

Zhentao Shi. September 11, 2020

Bibliography

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature 57*(4), 835–903. 2.4

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, Volume 1, pp. 19–46. 2.3