Econometrics

Instructor: Alexei Onatski

Part I: Basic regression.

Session 1.

References on Bruce Hansen's textbook (2019 version) will be abbreviated with H. Say, H 1.1 means chapter 1.1 from Bruce Hansen's textbook.

Basic probability concepts. Conditional Expectation Function and Best Linear Predictor. (H 2.1-2.12 and H 2.18-2.20)
Causal effects. (H 2.30)
Multivariate normal distribution. (H 5.2-5.3)
Analogy principle. OLS. Geometric interpretation of OLS. Unbiasedness of OLS. Quality of an estimator. (H 3.1-3.7 and H 4.5).

Random variables and their distributions.

A traditional goal of statistical analysis is to recover the joint probability distribution of random variables from observations of these variables. Random variables are best modelled as functions from a probability space to $\mathbb{R}$. A probability space is a triple $(\Omega, \mathcal{A}, \mathbb{P})$ of the space Ω of elementary outcomes of a random experiment, a sigma-algebra $\mathcal{A}$ of the sub-sets of elementary events, and the probability measure $\mathbb{P}$. In this course we will mostly ignore the measure-theoretic aspects of random variables. Interested students should consult Billingsley (1995) or Pollard (2002). However, thinking about random variables as functions will become particularly useful when we discuss convergence concepts (Session 4).

Recall that the distribution of an absolutely continuous random variable X is characterized by its probability density function $f_X(x)$ such that $\Pr(X \in A) = \int_A f_X(x)dx$. (For discrete random variables, $f_X(x)$ is called the probability mass function and the integral is taken with respect to a counting measure.) The cumulative distribution function of X is defined as $F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(x)dx$. The τ-th quantile of X is defined as $F_X^{-1}(\tau) = \inf(x : F(x) \geq \tau)$. For a function g, the expectation of $g(X)$ is defined as $Eg(X) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$. The row k-th moment of X is defined as EX^k. Note that EX^k may not exist (the integral may be infinite), then we say that X does not have k-th moment. EX is called the mean of X and $E(X - E(X))^2 = EX^2 - (EX)^2$ is called the variance of X.

Example 1 *Normal random variable with mean μ and variance σ^2, $X \sim N(\mu, \sigma^2)$ has density*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-(x-\mu)^2/2\sigma^2\right\}.$$

The 0.975-th quantile of the standard normal $N(0,1)$ is 1.96...

Example 2 *Random variable with Student's $t(k)$ distribution has density*

$$f_X(x) = c\left(1 + x^2/k\right)^{-(k+1)/2},$$

where c is a constant such that $\int f_X(x)\,dx = 1$. Clearly, $t(k)$ distribution has only $k-1$ finite moments. For this reason, stock returns are better modeled by $t(k)$ than by $N(\mu, \sigma^2)$.

Joint distribution of absolutely continuous random variables X and Y is characterized by the probability density function $f_{XY}(x,y)$ such that $\Pr(X \in A, Y \in B) = \int_{A\times B} f_{XY}(x,y)dxdy$. The cumulative distribution function is defined as $F_{XY}(x,y) = \Pr(X \leq x, Y \leq y)$. Marginal distributions integrate out all but one variable

$$f_X(x) = \int_y f_{XY}(x,y)dy \text{ and } f_Y(y) = \int_x f_{XY}(x,y)dx.$$

If X and Y are independent, then $f_{XY}(x,y) = f_X(x) f_Y(y)$ and $F_{XY}(x,y) = F_X(x) F_Y(y)$.

Example 3 *Bi-variate normal $Z \equiv (X,Y)' \sim N(\mu, \Sigma)$ with $\mu = (\mu_X, \mu_Y)'$ and $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$ has density*

$$\frac{1}{2\pi |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(Z-\mu)' \Sigma^{-1} (Z-\mu)\right\}.$$

This definition remains almost the same for a k-dimensional normal Z (the only difference is that 2π in the denominator becomes $(2\pi)^{k/2}$). Of course, in k-dimensional case, μ is a k-dimensional vector and Σ is a $k \times k$ matrix. The off-diagonal elements of Σ are the covariances between the elements of Z and the diagonal elements are the variances of the elements of Z.

Example 4 *Bi-variate Archimedean copula.*

$$F_{XY}(x,y) \equiv C_\phi(x,y) = \phi^{-1}(\phi(x) + \phi(y)),$$

where $\phi : [0,1] \to [0,\infty]$ is convex strictly decreasing continuous function with $\phi(0) = \infty$ and $\phi(1) = 0$. Marginal distributions are uniform $U[0,1]$. For arbitrary marginals, consider

$$F_{XY}(x,y) = C_\phi(F_X(x), F_Y(y)).$$

For more on copulas (which we will not cover in this course) see e.g. Nelsen (199) and Fan and Patton (2014).

The majority of econometric studies focus only on some aspects of multivariate distribution. In the bivariate case, often there is a distinction between explanatory variable and the dependent variable and one wants to understand how one "affects" the other "on average".

Conditional expectation function (CEF)

Conditional distribution of Y given X is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}, \text{ if } f_X(x) \neq 0$$

Conditional expectation $E(Y|X=x)$ is defined as

$$E(Y|X=x) = \int_y y f_{Y|X}(y|x)dy$$

Often, we will skip $X=x$ having in mind that $E(Y|X)$ is a function of random variable X. Hence, it is itself a random variable.

Law of Iterated Expectations

The unconditional mean equals the mean of the conditional expectation.

$$E\left(Y\right) = E\left(E\left(Y|X\right)\right).$$

Indeed,

$$\begin{aligned} E\left(Y\right) &= \int_y y f_Y(y)dy = \int_y \int_x y f_{XY}(x,y)dxdy = \int_x \int_y y f_{XY}(x,y)dydx \\ &= \int_x \int_y y f_{Y|X}(y|x) f_X(x)dydx = \int_x E(Y|X=x) f_X(x)dx = E\left(E\left(Y|X\right)\right). \end{aligned}$$

A more general version of the law of iterated expectation is

$$E\left(Y|X\right) = E\left(E\left(Y|X,Z\right)|X\right).$$

Properties of CEF.

In principle, instead of using conditional means to describe economic relationship, we could use other measures of central tendency such as median for example. It turns out that

in some sense CEF predicts Y better than any other function X. Let us measure the accuracy of prediction by expected squared error criterion. That is to get the most accurate predictor we must find a function $g(X)$ such that

$$g(X) = \arg\min_{g(X)} E(Y - g(X))^2.$$

Theorem. CEF is the best predictor of Y given X.

Proof:

$$E(Y - g(X))^2 = E\left[Y - E(Y|X) + E(Y|X) - g(X)\right]^2$$

$$= E\left[Y - E(Y|X)\right]^2 + 2E\left(\left[Y - E(Y|X)\right]\left[E(Y|X) - g(X)\right]\right) + E\left[E(Y|X) - g(X)\right]^2.$$

Let us look at the middle term $2E\left(Z\right)$, where

$$Z = \left[Y - E(Y|X)\right]\left[E(Y|X) - g(X)\right].$$

By the law of iterated expectations, $2E\left(Z\right) = 2E\left(E\left(Z|X\right)\right)$. On the other hand,

$$\begin{aligned} E\left(Z|X\right) &= E\left\{\left[Y - E(Y|X)\right]\left[E(Y|X) - g(X)\right]|X\right\} = \left[E(Y|X) - g(X)\right] E\left\{\left[Y - E(Y|X)\right]|X\right\} \\ &= \left[E(Y|X) - g(X)\right]\left[E\left(Y|X\right) - E(Y|X)\right] = 0 \end{aligned}$$

Therefore, $2E\left(Z\right) = 0$ and

$$E(Y - g(X))^2 = E\left[Y - E(Y|X)\right]^2 + E\left[E(Y|X) - g(X)\right]^2 \geq E\left[Y - E(Y|X)\right]^2$$

and CEF is the best conditional predictor of Y. □

Good motivation to study CEF! CEF is also called *population regression*

Why "regression"?

This word was coined by Francis Galton who lived in 19th century and was a very interesting guy. At some point of his life Francis Galton studied seeds of sweet peas plant. He measured weights of many seeds and planted these seeds to get offsprings. He then made a scatterplot of weight of offspring seeds vs. weight of parent seeds. Galton divided all parent seeds into 7 different groups according to their weight. He then plot the line connecting median weight of offspring seeds for different parental groups. He noted that the slope of this line was substantially less than 1, and interpreted this as the tendency of offsprings to be more mediocre than their parents. Galton called this phenomenon *regression* to the mean.

Conditional Quantile Function

According to the expected squared error criterion, the best predictor of Y given X is $E(Y|X)$. If we change the criterion to the expected absolute error, then, the best predictor would become the conditional median. Note that the absolute error of predictor $g(X)$ can be written as

$$|Y - g(X)| = \mathbf{1}(Y > g(X))(Y - g(X)) + \mathbf{1}(Y < g(X))(g(X) - Y),$$

where $\mathbf{1}(\cdot)$ is an indicator function equal to one if the argument is true and zero if it is false. Hence,

$$E(|Y - g(X)| \,|X = x) = \int_{g(x)}^{\infty} (y - g(x)) f_{Y|X}(y|x)\, dy + \int_{-\infty}^{g(x)} (g(x) - y) f_{Y|X}(y|x)\, dy.$$

Differentiating w.r.t. $g(x)$, we get the f.o.c. for the minimum of the conditional absolute prediction error

$$-\int_{g(x)}^{\infty} f_{Y|X}(y|x)\, dy + \int_{-\infty}^{g(x)} f_{Y|X}(y|x)\, dy = 0.$$

This means that $g(x)$ must be the median of the conditional distribution of Y given $X = x$. Therefore, the conditional median $g(X) = med(Y|X)$ must be the best predictor of Y given X according to the expected absolute error criterion.

If underprediction is marginally less or more costly as overprediction, it makes sense to minimize the expectation of

$$\tau \mathbf{1}(Y > g(X))(Y - g(X)) + (1 - \tau)\mathbf{1}(Y < g(X))(g(X) - Y)$$

with $\tau \in (0, 1)$. For example, parameter $\tau < 1/2$ would correspond to situations where the underprediction is less costly than overprediction. Following the same logic as above, we can show that the corresponding best predictor would be τ-th quantile $\tau(X)$ of the conditional distribution of Y given X.

Causality. Econometricians do use regressions for prediction purposes. However, more often, they are interpreting them as measuring causal effects of explanatory variables on the dependent variable. When does such an interpretation make sense? Suppose that we would like to know what the causal effect of a change in characteristic x_1 of an individual (for example, x_1 may be years of education) on their characteristic y (for example, wage) is. Beside x_1, there will exist tons of other characteristics affecting y_1. Some of these, collected in vector x_2 may be observed, while others, collected in vector ε may be unobserved. Different

combinations of x_1, x_2, and ε will "result" in some particular values of y so we can write

$$y = h(x_1, x_2, \varepsilon)$$

for an unknown function h. In this framework, we may define the causal effect of x_1 on y as

$$C(x_1, x_2, \varepsilon) = \frac{\partial}{\partial x_1} h(x_1, x_2, \varepsilon),$$

that is, the change in y due to a change in x_1, holding x_2 AND ε constant.

Remark 5 *When x_1 is discrete, $\frac{\partial}{\partial x_1}$ can be replaced by a finite difference operator.*

To be concrete, I studied in a school, and a couple of universities for 21 years in total. What would be the effect on my wage had I studied 22 years instead? In other words, what is

$$C(21, x_2, \varepsilon) = h(22, x_2, \varepsilon) - h(21, x_2, \varepsilon)?$$

Note that $C(21, x_2, \varepsilon)$ would describe the causal effect for myself because I am characterized by particular values of x_2 and ε. This effect may be very different for my colleague who would have different x_2 and ε. In short,

<u>the causal effect is specific to an individual.</u>

Continuing with the example, although I know what $h(21, x_2, \varepsilon)$ is (my current wage), I have no idea what $h(22, x_2, \varepsilon)$ may be. In short,

<u>the causal effect is unobserved</u>.

The heterogeneity and unobservability of the causal effects creates huge problems for estimation.

Remark 6 *In the standard introductory graduate econometrics courses, $h(x_1, x_2, \varepsilon)$ is usually assumed to be equal $x_1'\beta_1 + x_2'\beta_2 + \varepsilon$. Under this model, the individual-specific nature of the causal effect disappears, which, of course, makes the model less realistic.*

A popular example arises in the analysis of treatment effects with a binary treatment x_1. Let $x_1 = 1$ indicate that the treatment, for example, taking a pill, has been received, and $x_1 = 0$ indicates that it has not been received. The outcome y may be, say, blood pressure. Then, we can write

$$\begin{aligned} y(1) &= h(1, x_2, \varepsilon), \\ y(0) &= h(0, x_2, \varepsilon). \end{aligned}$$

Variables $y(1)$ and $y(0)$ are called potential outcomes. $y(1)$ is the blood pressure in a hypothetical situation when the pill were taken, and $y(0)$ is the blood pressure in a hypothetical situation when the pill were not taken. In reality, the patient either takes the pill or not, so we either observe $y(1)$ or $y(0)$, but not both. The causal effect of taking the pill is defined as

$$C(0, x_2, \varepsilon) = y(1) - y(0).$$

Note that it depends on x_2 AND ε, so may be totally individual-specific. Since $y(1)$ and $y(0)$ are not simultaneously observed, and since they may be absolutely different for different individuals, there is no hope for estimating $C(0, x_2, \varepsilon)$ using data. Therefore, people focus on the estimation of what is called the average causal effect.

Average Causal Effects and regression.

Definition 7 *The average causal effect of x_1 on y conditional on x_2 is*

$$\begin{aligned} ACE(x_1, x_2) &= E(C(x_1, x_2, \varepsilon) | x_1, x_2) \\ &= \int \left[\frac{\partial}{\partial x_1} h(x_1, x_2, \varepsilon) \right] f(\varepsilon | x_1, x_2) \, d\varepsilon, \end{aligned}$$

where $f(\varepsilon | x_1, x_2)$ is the conditional density of ε given x_1 and x_2.

For example, suppose there are two types of professors. For the first type, change in education from 21 to 22 leads to no change in wage, whereas, for the second type, it leads to an increase in wage by £5,000. If 25% of professors are of the first type and 75% of them are of the second type, we would have

$$ACE(21, x_2 = professor) = 0 \times \frac{1}{4} + 5,000 \times \frac{3}{4} = 3,750$$

Here x_2 is an observed indicator variable for occupation. Had x_2 been, say, 'banker' the average causal effect would be different from $£3,750$.

We would like to emphasize the fact that ACE depends on what we are conditioning on. For example, the observed variables might have included not only occupation, but, say, post code. Then, the average causal effect of the extra year of education might have been different for professors living in areas with different post codes.

When we run regression of y on x_1 and x_2, we might hope that the slope of the CEF with respect to x_1 equals the average causal effect conditional on x_2. Unfortunately, in general this is not so. Consider a general model for CEF

$$E(y | x_1, x_2) = m(x_1, x_2).$$

In linear regression, we assume that $m(x_1, x_2)$ is a linear function, but nonparametric regression methods (which we do not cover in this class) allow estimation of $m(x_1, x_2)$ that belong to rather general classes of functions. The regression slope with respect to x_1 equals

$$\begin{aligned}
\frac{\partial}{\partial x_1} m(x_1, x_2) &= \frac{\partial}{\partial x_1} \int h(x_1, x_2, \varepsilon) f(\varepsilon | x_1, x_2) \, d\varepsilon \\
&= \int \frac{\partial}{\partial x_1} h(x_1, x_2, \varepsilon) f(\varepsilon | x_1, x_2) \, d\varepsilon \\
&\quad + \int h(x_1, x_2, \varepsilon) \frac{\partial}{\partial x_1} f(\varepsilon | x_1, x_2) \, d\varepsilon \\
&= ACE(x_1, x_2) + \int h(x_1, x_2, \varepsilon) \frac{\partial}{\partial x_1} f(\varepsilon | x_1, x_2) \, d\varepsilon \\
&\neq ACE(x_1, x_2).
\end{aligned}$$

It is because the conditional density of $f(\varepsilon | x_1, x_2)$ may depend on x_1 so that $\frac{\partial}{\partial x_1} f(\varepsilon | x_1, x_2) \neq 0$.

For example, consider the two types of professors mentioned above. Suppose further that the potential outcome (wages) given $x_1 = 21$ or 22 can only take on two values, £65,000 and £70,000. Suppose the following joint distribution of $(x_1, y(21), y(22))$:

$$\begin{aligned}
(x_1, y(21), y(22)) &= (21, 65000, 65000) \quad 10\% \\
(x_1, y(21), y(22)) &= (21, 65000, 70000) \quad 30\% \\
(x_1, y(21), y(22)) &= (22, 65000, 70000) \quad 60\%
\end{aligned}$$

There are no observables, except x_1 (and $y(x_1)$). This situation is compatible with structural function $y = h(x_1, \varepsilon)$ defined by, for example,

$\varepsilon \backslash x_1$	21	22
1	65000	65000
2	65000	70000

In the assumed population, there are no people with $(x_1 = 22, \varepsilon = 1)$, 10% of people have $(x_1 = 21, \varepsilon = 1)$, 30% of people have $(x_1 = 21, \varepsilon = 2)$, and 60% of people have $(x_1 = 22, \varepsilon = 2)$.

The conditional expectation of y given x_1 is as follows

$$\begin{aligned}
E[y | x_1 = 21] &= 65000 \times 1 = 65,000 \\
E[y | x_1 = 22] &= 70,000 \times 1 = 70,000
\end{aligned}$$

So, the regression slope is

$$\frac{70,000-65,000}{22-21}=5,000$$

On the other hand, the average causal effect is

$$ACE\left(21\right)=0\times\frac{1/10}{4/10}+5,000\times\frac{3/10}{4/10}=3,750.$$

The reason for the discrepancy is that the conditional distribution of ε given x_1 changes when x_1 increases from 21 to 22. For $x_1=21$, it is 25% mass on $\varepsilon=1$ and 75% mass on $\varepsilon=2$, whereas for $x_1=22$, it is 100% on $\varepsilon=2$. Regression compares wage outcomes of subpopulations with dramatically different distributions of the individual-specific characteristics (in the sense that $\varepsilon|x_1=21$ and $\varepsilon|x_1=22$ are very different distributions), and thus, fails to identify the ACE.

The regression analysis has a causal interpretation if $\frac{\partial}{\partial x_1}f\left(\varepsilon|x_1,x_2\right)=0$, that is when conditional density of ε does not depend on x_1. This would be the case if the following assumption holds.

Definition 8 *(Conditional Independence Assumption) Conditional on x_2, the random variables x_1 and ε are statistically independent.*

Theorem 9 *In the model $y=h\left(x_1,x_2,\varepsilon\right)$, the CIA assumption implies that the regression slope with respect to x_1 equals the average causal effect of x_1 on y conditional on x_2.*

Best linear predictor.

CEF in the bi-variate normal case.

Let us compute $E\left(Y|X\right)$ when $\left(X,Y\right)'\sim N(\mu,\Sigma)$ with $\mu=\left(\mu_X,\mu_Y\right)'$ and $\Sigma=\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}$.

We have

$$\Sigma^{-1}=\frac{1}{\sigma_X^2\sigma_Y^2-\sigma_{XY}^2}\begin{pmatrix}\sigma_Y^2 & -\sigma_{XY}\\ -\sigma_{XY} & \sigma_X^2\end{pmatrix}$$

and therefore,

$$\begin{aligned}f_{XY}\left(x,y\right) &= \frac{\exp\left\{-\frac{1}{2\left(\sigma_X^2\sigma_Y^2-\sigma_{XY}^2\right)}\left[\left(x-\mu_X\right)^2\sigma_Y^2-2\left(x-\mu_X\right)\left(y-\mu_Y\right)\sigma_{XY}+\left(y-\mu_Y\right)^2\sigma_X^2\right]\right\}}{2\pi\sqrt{\sigma_X^2\sigma_Y^2-\sigma_{XY}^2}}\\ &= \frac{\exp\left\{-\frac{1}{2\sigma_{Y|X}^2}\left[\left(x-\mu_X\right)^2\frac{\sigma_Y^2}{\sigma_X^2}-2\left(x-\mu_X\right)\left(y-\mu_Y\right)\frac{\sigma_{XY}}{\sigma_X^2}+\left(y-\mu_Y\right)^2\right]\right\}}{2\pi\sqrt{\sigma_X^2\sigma_Y^2-\sigma_{XY}^2}},\end{aligned}$$

where $\sigma^2_{Y|X} = \sigma^2_Y - \sigma^2_{XY}/\sigma^2_X$. Note that

$$\begin{aligned} & -2\left(x-\mu_X\right)\left(y-\mu_Y\right)\frac{\sigma_{XY}}{\sigma^2_X} + \left(y-\mu_Y\right)^2 \\ = & \left(\left(y-\mu_Y\right) - \left(x-\mu_X\right)\frac{\sigma_{XY}}{\sigma^2_X}\right)^2 - \left(x-\mu_X\right)^2\left(\frac{\sigma_{XY}}{\sigma^2_X}\right)^2 \end{aligned}$$

Therefore,

$$f_{XY}\left(x,y\right) = \exp\left\{-\frac{\left(\left(y-\mu_Y\right) - \left(x-\mu_X\right)\frac{\sigma_{XY}}{\sigma^2_X}\right)^2}{2\sigma^2_{Y|X}}\right\} g(x)$$

and

$$f_{Y|X}\left(y|x\right) = \exp\left\{-\frac{\left(\left(y-\mu_Y\right) - \left(x-\mu_X\right)\frac{\sigma_{XY}}{\sigma^2_X}\right)^2}{2\sigma^2_{Y|X}}\right\} \frac{g(x)}{f_X\left(x\right)}.$$

But such $f_{Y|X}\left(y|x\right)$ has the form of a univariate normal variable with mean $\mu_Y + \left(x-\mu_X\right)\frac{\sigma_{XY}}{\sigma^2_X}$ and variance $\sigma^2_{Y|X}$. Therefore,

$$E\left(Y|X\right) = \mu_Y + \left(X-\mu_X\right)\frac{\sigma_{XY}}{\sigma^2_X} = \beta_1 + \beta_2 X$$

with

$$\beta_1 = \mu_Y - \mu_X\beta_2 \text{ and } \beta_2 = \frac{\sigma_{XY}}{\sigma^2_X} \equiv \frac{Cov\left(X,Y\right)}{Var\left(X\right)}.$$

Hence, in the case of bi-variate normal distribution, the population regression is **linear**. This generalizes to the general multivariate normal distribution.

Best Linear Predictor (BLP)

Whenever the conditional expectation $E\left(Y|X\right)$ is linear, it has form $\beta_1 + \beta_2 X$ with

$$\beta_1 = EY - \left(EX\right)\beta_2 \text{ and } \beta_2 = \frac{Cov\left(X,Y\right)}{Var\left(X\right)},$$

just as in the case of the bi-variate normal distribution. Indeed, we know that CEF minimizes the expected squared error, therefore β_1 and β_2 must solve

$$\min_{\beta_1,\beta_2} E(Y-\beta_1-\beta_2 X)^2 = \min_{\alpha,\beta}\left\{EY^2 - 2E\left[Y(\beta_1+\beta_2 X)\right] + E\left(\beta_1+\beta_2 X\right)^2\right\}$$

The first order conditions are:

$$\begin{aligned} -2EY + 2E(\beta_1 + \beta_2 X) &= 0 \\ -2E(YX) + 2E(X(\beta_1 + \beta_2 X)) &= 0 \end{aligned}$$

Solving these two equations we obtain the above formulae for β_1 and β_2. Note that we can compute β_1 and β_2 even if CEF is not linear. We can just use above formulas. In such a case $\beta_1 + \beta_2 X$ is called the best linear predictor of Y. Sometimes, it is denoted as $E^*(Y|X)$ to emphasize the difference with the CEF.

Errors of prediction of CEF and BLP

What are errors of prediction of CEF and BLP? Let us denote errors of prediction as e. For CEF we have

$$e = Y - E(Y|X),$$

so that

$$\begin{aligned} Ee &= 0 \\ E(eX) &= 0 \\ E(e|X) &= 0 \end{aligned}$$

Exercise: ξ and η are two r.v. $E\xi = 0$ implies that $Cov(\xi, \eta) = E(\xi\eta)$

For BLP we have

$$e = (Y - EY) - \frac{Cov(X,Y)}{Var(X)}(X - EX),$$

so that

$$\begin{aligned} Ee &= 0 \\ E(eX) &= 0. \end{aligned}$$

That is for BLP the errors of prediction are simply uncorrelated with X. However, they may depend on X in a non-linear way. Technically, it is much easier to estimate BLP than CEF. It is a good idea to look at the plot of residuals after fitting a straight line to the data. If residuals show some pattern then this indicates undermodeling of CEF by BLP. More on CEF and BLP can be found in Goldberger (1991, ch.4-5)

Estimating parameters of BLP.

CEF and BLP are ideal objects. They depend on the parameters of the joint distribution

of the dependent and explanatory variables. We may try to estimate them given some data. Often the estimation problem will be specified in the following form. Suppose the dependent variable is y and the explanatory variables are collected in k-dimensional vector x (x usually includes a constant to accommodate the intercept). Estimate β in the linear model:

$$y = x'\beta + \varepsilon$$

along with one of the following assumptions on ε,

1) $E(\varepsilon) = 0$ and $E(\varepsilon|x) = 0$ which assumes that CEF is the linear function $x'\beta$.

2) $E(\varepsilon) = 0$ and $E(\varepsilon x) = 0$ which states that ε and x are uncorrelated and therefore β are coefficients in BLP.

Note that condition 1) implies condition 2) but not vice versa:

$$E(\varepsilon x) = E(E(\varepsilon x|x)) = E(xE(\varepsilon|x)) = E(x \times 0) = 0$$

Obviously, we must base our estimation on the available data. Suppose that we have n observations of the dependent variable y, say mortality rate of children in a country, and independent or explanatory variables x, say constant, the country's log(GDP), climate conditions, number of hospital beds per capita etc. Let us organize these data into a table:

Country i	Dependent variable mortality rate	Independent variables constant,log(GDP), climate, hospital beds,...
1	y_1	$x_{11} = 1, x_{12}, ..., x_{1k}$
2	y_2	$x_{21} = 1, x_{22}, ..., x_{2k}$
$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n1} = 1, x_{n2}, ..., x_{nk}$

.

We define vector Y and matrix X

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ and } X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}.$$

The transpose of the i-th row of the matrix X is denoted as:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}.$$

This can be interpreted as the $i-$th observation of the vector of independent variables x. We assume that the observations of the data (y_i, x_i) are independent and come from the same joint distribution.

We want to estimate BLP (or in the case when CEF is linear, CEF) of the form:

$$\begin{aligned} E^*(y_i|x_i) &= x_i'\beta = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}' \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \\ &= \beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k \end{aligned}$$

where β is known as the coefficient on x_i that yields the best prediction of y_i.

Analogy Principle

We know that

$$\beta = \arg\min_b E(y - x'b)^2$$

We however, do not know the population expectation of $(y - x'b)^2$. We may estimate β by minimizing a sample analog to the population moment, that is

$$\hat{\beta}_{OLS} = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'b)^2. \tag{1}$$

This method of estimation is due to Gauss and is called minimum sum of squared errors method (msse).

The first order conditions are

$$0 = \frac{2}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{OLS})(-x_i)$$

Solving for $\hat{\beta}_{OLS}$, we get

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i$$

or, in matrix notation,

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y.$$

For $X'X$ to be invertible, $X'X$ needs to have full rank (so X has full column rank, a condition known as no perfect multicollinearity assumption). If it does not, then there are infinite number of solutions to the first order conditions. This may happen, for example, if some components of x are linear combinations of the others as in the dummy variable trap). Then, we say that the parameters of the BLP are not identified. That is, cannot be determined even if the infinite amount of data is available.

The analogy principle can also be used to derive parameters of **linear** quantile functions. Recall that these parameters must be equal to

$$\beta = \arg\min_b E\left\{\tau \mathbf{1}\left(y > x'b\right)\left(y - x'b\right) + \left(1-\tau\right)\mathbf{1}\left(y < x'b\right)\left(x'b - y\right)\right\}$$

By the analogy principle, we have

$$\beta = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}\left\{\tau \mathbf{1}\left(y_i > x_i'b\right)\left(y_i - x_i'b\right) + \left(1-\tau\right)\mathbf{1}\left(y_i < x_i'b\right)\left(x_i'b - y_i\right)\right\}.$$

Details can be found in an authoritative book on quantile regression Koenker (2005).

Flexibility

Now when we estimated the coefficients in BLP you may ask so what? The world is non-linear and using BLP will usually tell us little about CEF. Note, however, that BLP must be linear with respect to parameters, not necessarily with respect to variables. A general way to make BLP a good description of CEF is to add squares or higher orders of the explanatory variables or/and transform dependent or explanatory variables so that the relationship between transformed variables appear to be linear. One such famous transformation is called the Box-Cox transformation

$$y \mapsto \frac{y^\lambda - 1}{\lambda}.$$

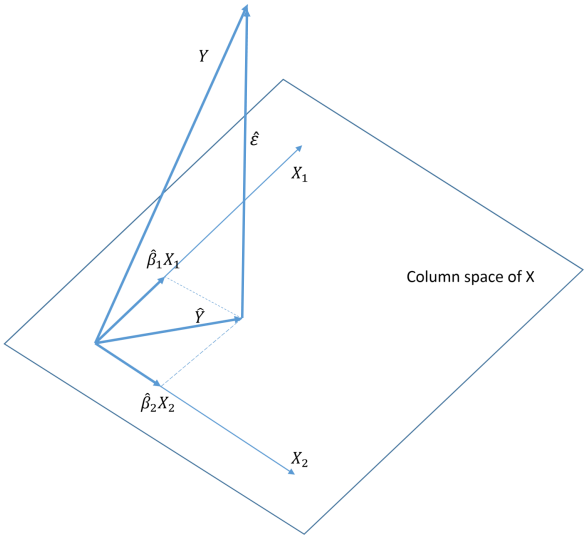
This converge to $\ln y$ when $\lambda \to 0$. Parameter λ is not known, but may be estimated, say, by nonlinear least squares.

Geometric interpretation

Y can be thought of as a vector in n-dimensional space with coordinates $y_1, y_2, ..., y_n$. Similarly, all columns of X are vectors in n-dimensional space. For example, the first column is the vector with coordinates $1, 1, ..., 1$. Then, Xb can be thought of as a linear combination of the columns of X with coefficients $b_1, b_2, .., b_k$. OLS is trying to find b such that Xb is as

close to Y as possible. Geometrically, this means that the length of the vector $Y - Xb$ is as small as possible. This is achieved by choosing b so that Xb is the orthogonal projection of Y on the space spanned by X (so to speak, shadow of Y when "sun is shining down" on the space X).

Let us illustrate such an interpretation by considering $n \times 2$ matrix X. Let X_1 and X_2 be the first and the second columns of X. Y, X_1, and X_2 are interpreted as vectors in n-dimensional space. OLS projects Y on the space spanned by X_1 and X_2.



Unbiasedness of OLS.

We have

$$E\hat{\beta} = E\left[(X'X)^{-1}X'Y\right] = E\left[E((X'X)^{-1}X'Y|X)\right] = E\left[(X'X)^{-1}X'E(Y|X)\right]$$

Assuming that CEF is linear, $E(Y|X) = X\beta$ and we can continue

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

Hence, if CEF is linear, OLS is an unbiased estimator of β. Moreover, the above arguments reveal that

$$E\left(\hat{\beta}|X\right) = \beta.$$

That is, whatever the value of X is, $\hat{\beta}$ is unbiased for β conditionally on this value. This is stronger than the unconditional unbiasedness.

Quality of an estimator.

Is OLS a good estimator of β? The answer depends on the criteria for the quality of estimators. Intuitively, an estimator is good if it is close to the true value of β and it is bad if it is far away from β. In 18th century, Laplace noted that the estimation problem is similar to games of chance. The researcher is playing game with God who gives him data on which then estimation is based. The researcher almost always looses because his estimator will in general be different from the true value. At best, the researcher can break even so that his estimator is equal to the estimand.

In analogy to the gambling situation, Laplace proposed to consider loss of the researcher which is proportional to the deviation of the true parameter from his guess.

$$Loss = \left|\hat{\beta} - \beta\right|$$

Then the researcher's risk when he place the estimation game with God is equal to his expected loss

$$Risk = E\left|\hat{\beta} - \beta\right|.$$

Gauss proposed to consider loss proportional not to an absolute deviation but to square of the deviation. In such a case the risk is equal to

$$Risk = E(\hat{\beta} - \beta)^2$$

Good estimator will minimize risk (or variance) of the estimator

$$E(\hat{\beta} - \beta)^2 \to \min.$$

Gauss also noted that if researcher wants to avoid systematic mistakes, he must consider unbiased estimators

$$E(\hat{\beta} - \beta) = 0.$$

In this case the risk is equal to Variance of $\hat{\beta}$.

In a multivariate case (when β is a vector), the mean squared error criterion

$$E\left(\left(\hat{\beta} - \beta\right)'\left(\hat{\beta} - \beta\right)\right) = E\left(\left\|\hat{\beta} - \beta\right\|^2\right) \to \min$$

is used. The mean squared error can be decomposed in the bias and variance part as follows.

$$\begin{aligned} E\left(\left\|\hat{\beta}-\beta\right\|^2\right) &= E\left(\left\|\hat{\beta}-E\hat{\beta}+E\hat{\beta}-\beta\right\|^2\right) \\ &= \underbrace{E\left(\left\|\hat{\beta}-E\hat{\beta}\right\|^2\right)}_{variance}+\underbrace{\left\|E\hat{\beta}-\beta\right\|^2}_{bias^2}. \end{aligned}$$

Interestingly, estimators minimizing the mean squared error are not in general unbiased. It may pay off to introduce some bias by shrinking $\hat{\beta}$ towards zero so that the variance is reduced.

Shrinkage.

For example, Charles Stein (1956) showed that if you are given an n-dimensional vector Y with independent normal coordinates having arbitrary means, $\mu = EY$, then the "natural" unbiased estimator $\hat{\mu} = Y$ does not minimize the mean squared error if $n \geq 3$. For large n, his logic is simple enough to illustrate here. We have

$$\|Y\|^2 = \|Y-\mu\|^2 + \|\mu\|^2 + 2\|\mu\| Z,$$

where

$$Z = \mu'(Y-\mu)/\|\mu\|$$

is a standard normal random variable. For large n, by the Central Limit Theorem,

$$\|Y-\mu\|^2 = n + O_p\left(\sqrt{n}\right)$$

so that

$$\|Y\|^2 = n + \|\mu\|^2 + O_p\left(\sqrt{n+\|\mu\|^2}\right).$$

Therefore, when we observe $\|Y\|^2$, we know that $\|\mu\|^2$ must be close to $\|Y\|^2 - n$. On the other hand, the estimator $\hat{\mu} = Y$ obviously satisfies

$$\|\hat{\mu}\|^2 = \|Y\|^2.$$

Hence, it does make sense to shrink the estimator at least by a factor $\sqrt{\left(\|Y\|^2 - n\right)/\|Y\|^2}$.

Remark 10 *A sequence x_n is said to be $O_p(f(n))$ (x_n is of probability order $O(f(n))$) if for each $\delta > 0$ there exists an $A_\delta > 0$ and $N_\delta > 0$ such that $\Pr(|x_n| \leq A_\delta f(n)) \geq 1-\delta$ for all $n > N_\delta$.*

It turns out that the following estimator

$$\hat{\mu}_{JS} = \left(1 - \frac{n-2}{\|Y\|^2}\right) Y$$

which is called the James-Stein estimator outperforms $\hat{\mu} = Y$ in terms of the mean squared errors for all (!) μ (we say that $\hat{\mu}$ is inadmissible in such a case)

Theorem 11 *Let $Y_i, i = 1, ..., n\,(n > 2)$ be independent $N(\mu_i, 1)$. Then*

$$\frac{1}{n} E \sum\nolimits_{i=1}^{n} \left(\hat{\mu}_{JS,i} - \mu_i\right)^2 = 1 - \frac{1}{n} E \frac{(n-2)^2}{\|Y\|^2} < 1$$

Proof: First, note that

$$\frac{1}{n} E \sum\nolimits_{i=1}^{n} \left(\hat{\mu}_{JS,i} - \mu_i\right)^2 = \frac{1}{n} E \sum\nolimits_{i=1}^{n} \left(Y_i - \mu_i\right)^2 + \frac{1}{n} E \frac{(n-2)^2}{\|Y\|^2} - \frac{2(n-2)}{n} \sum\nolimits_{i=1}^{n} E \frac{(Y_i - \mu_i) Y_i}{\|Y\|^2}$$

Focus on

$$E \frac{(Y_i - \mu_i) Y_i}{\|Y\|^2} = \int \frac{(y_i - \mu_i) y_i}{\|y\|^2} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum\nolimits_{i=1}^{n} \left(y_i - \mu_i\right)^2\right) dy_1 ... dy_n.$$

Performing integration by parts, we obtain

$$\begin{aligned} & \int \frac{(y_i - \mu_i) y_i}{\|y\|^2} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum\nolimits_{i=1}^{n} \left(y_i - \mu_i\right)^2\right) dy_1 ... dy_n \\ = & \int \left(\frac{\partial}{\partial y_i} \frac{y_i}{\|y\|^2}\right) \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum\nolimits_{i=1}^{n} \left(y_i - \mu_i\right)^2\right) dy_1 ... dy_n \end{aligned}$$

On the other hand,

$$\frac{\partial}{\partial y_i} \frac{y_i}{\|y\|^2} = \frac{1}{\|y\|^2} - \frac{2y_i^2}{\|y\|^4}$$

Therefore,

$$E \frac{(Y_i - \mu_i) Y_i}{\|Y\|^2} = E \frac{1}{\|Y\|^2} - E \frac{2Y_i^2}{\|Y\|^4}$$

and thus,

$$\begin{aligned}
\frac{1}{n}E\sum\nolimits_{i=1}^{n}\left(\hat{\mu}_{JS,i}-\mu_i\right)^2 &= \frac{1}{n}E\sum\nolimits_{i=1}^{n}\left(Y_i-\mu_i\right)^2+\frac{1}{n}E\frac{(n-2)^2}{\|Y\|^2} \\
&\quad -\frac{2(n-2)}{n}\sum\nolimits_{i=1}^{n}\left(E\frac{1}{\|Y\|^2}-E\frac{2Y_i^2}{\|Y\|^4}\right) \\
&= 1+\frac{1}{n}E\frac{\left[(n-2)^2-2n(n-2)+4(n-2)\right]}{\|Y\|^2} \\
&= 1-\frac{1}{n}E\frac{(n-2)^2}{\|Y\|^2}.\square
\end{aligned}$$

References

[1] Billingsley, P. (1995) *Probability and Measure*, John Wiley & Sons.

[2] Fan, Y., and Patton, A. J. (2014) "Copulas in Econometrics," *Annual Review of Economics* 6, 179–200.

[3] Goldberger, A. (1991) *A course in Econometrics*, Harvard University Press.

[4] Koenker, R. (2005) *Quantile Regression*. Cambridge University Press.

[5] Nelsen, R.B. (1999) *An introduction to copulas*. Springer-Verlag.

[6] Pollard, D. (2002) *A User's Guide to Measure Theoretic Probability*, Cambridge University Press.

[7] Stein, C. (1956) "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, Vol. 1, pp. 197-206.