Estimation of Treatment Effects

Prof. Bernd Fitzenberger, Ph.D.

Humboldt-University Berlin – Winter Semester 2016/17

0. Introductory Material

Linear Regression Model, Conditional Expectation, and Causal Interpretation

- y explained/dependent/response variable
- $x = (x_1, ..., x_k)$ explanatory / independent variables, regressors, control variables, covariates (x is observed)

Structural conditional expectation (CE): $E(y|w, c)$

Based on random sample of (y, w, c) we can estimate the effect of w on y holding c constant.

Complications arise when there is no random sample of (y, w, c)

- $\rightarrow$ measurement error
- $\rightarrow$ simultaneous determination of y, w, c
- $\rightarrow$ some variables we would like to control for (elements of c) cannot be observed

$\Rightarrow$ CE of interest involves data for which the econometrician cannot collect data or requires an experiment that cannot be carried out.

Identification assumptions:

$\rightarrow$ Can recover structural CE of interest

Definition CE:

y (random variable) explained variable, $x \equiv (x_1, x_2, ..., x_k)$ $(1 \times k)$-vector of explanatory variables, $E(|y|) < \infty$

then function $\mu : \mathbb{R}^k \rightarrow \mathbb{R}$

(CE) $E(y|x_1, x_2, ..., x_k) = \mu(x_1, x_2, ..., x_k)$ or $E(y|x) = \mu(x)$

Distinguish

$E(y|x)$: random variable because x is a random variable

from

$E(y|x = x_0)$: conditional expectation when x takes specific value x_0

$\rightarrow$ Distinction most of the time not important
$\rightarrow$ Use $E(y|x)$ as short hand notation

Parametric model for $E(y|x)$ where $\mu(x)$ depends on a finite set of unknown parameters

Examples:

(i) $E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

(ii) $E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2$

(iii) $E(y|x_1, x_2) = \exp[\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2] \quad \text{with} \;\; y \geq 0, \; x_1 > 0$

(i) is linear in parameters and explanatory variables

(ii) is linear in parameters and nonlinear in explanatory variables

(iii) is nonlinear in both

Partial Effect:

- Continuous x_i, and differentiable μ

$$\Delta E(y|x) = \frac{\partial \mu}{\partial x_j} \Delta x_j \quad \text{holding } \; x_1, ..., x_{j-1}, x_{j+1}, ..., x_k \text{ fixed}$$

$\widehat{=}$ ceteris paribus effect for propertly specified population model

- Discrete $x_j : x_{j,0} \rightarrow x_{j,1}$

$$\Delta E(y|x) = E(y|x_1, ..., x_{j-1}, x_{j,1}, x_{j+1}, ..., x_k) - E(y|x_1, ..., x_{j-1}, x_{j,0}, x_{j+1}, ..., x_k)$$

Examples:

ad i) $\frac{\partial E(y|x)}{\partial x_1} = \beta_1 = \text{constant}$

ad ii) $\frac{\partial E(y|x)}{\partial x_1} = \beta_1 + \beta_4 x_2$, i.e. partial effect of x_1 varies with x_2

ad iii) $\frac{\partial E(y|x)}{\partial x_1} = exp[\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2] \frac{\beta_1}{x_1}$ $\rightarrow$ highly nonlinear

(Partial) Elasticity (only continuous case)

$$\frac{\partial E(y|x)}{\partial x_j} \cdot \frac{x_j}{E(y|x)} = \frac{\partial \log E(y|x)}{\partial \log x_j}$$

(Partial) Semielasticity:

$$\frac{\partial E(y|x)}{\partial x_j} \cdot \frac{1}{E(y|x)} = \frac{\partial \log E(y|x)}{\partial x_j}$$

Average Partial Effect (APE, 'integrate out distribution of x'):

$$E_x\left\{\Delta E(y|x)\right\} = E_x\left\{\frac{\partial \mu}{\partial x_i}\Delta x_j\right\}$$

Examples:

ad i) APE$=\beta_1$

ad ii) APE$=\beta_1 + \beta_4 E x_2$

ad iii) APE$= E\left\{exp[\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2]\frac{\beta_1}{x_1}\right\}$

APE's in cases ii and iii can be estimated by sample averages of the expressions evaluated at the sample estimates of the coefficients $\hat{\beta}$

Error form of models of conditional expectations

We can always write

(1) $y = E(y|x) + u \quad \text{where } u = y - E(y|x)$

and it follows by definition:

(2) $E(u|x) = 0$

Implications:

(i) $E(u) = 0$

(ii) u is uncorrelated with any function of $x_1, ..., x_k$

Implication (i) and (ii) follows from the law of iterated expectations

$$\underline{\text{LIE}} : E(y|x) = E[E(y|w)|x] \quad \text{if } x = f(w)$$

i.e. $\{$Information set incorporated in $x\} \subseteq \{$Information set incorporated in $w\}$

i) $E(y|x) = E[E(y|w)|x]$
$\rightarrow$integrating out w wrt x: $\int yf(y|x)dy = \int[\int yf(y|w,x)dy]f(w|x)dw$

ii) $E(y|x) = E[E(y|x)|w]$
Knowing w implies knowing x
$\rightarrow$ Routinely used in the course

'The smaller information set always dominates'

Therefore

$$E(u) = E_x[E(u|x)] = E_x 0 = 0$$

which gives implication (i) and

$$E(u|f(x)) = E[E(u|x)|f(x)] = E[0|f(x)] = 0$$

which gives implication (ii).

Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

with

$$E(u|x_1, x_2) = 0$$

implies:

$E(u) = 0, Cov(x_1, u) = 0, Cov(x_2, u) = 0$ and u is also uncorrelated with $x_1^2, x_2^2, x_1 x_2, \exp(x_1)$ *etc.*

i.e. the functional form of $E(y|x)$ is properly specified.

We have $\beta_2 = \frac{\partial E(y|x_1,x_2)}{\partial x_2}$ because $E(u|x_1, x_2) = 0$, i.e. u is uncorrelated with any function of x_2. Thus β_2 describes the mean impact of x_2 on y.

$\boxed{E(u|x_1, x_2) = 0 \text{ sometimes called mean independence}}$

We have:

$$\text{Independence} \underset{\not\Leftarrow}{\Rightarrow} \text{Mean Independence} \underset{\not\Leftarrow}{\Rightarrow} \text{Uncorrelatedness}$$

Mean independence defines a Conditional Expectation

Uncorrelatedness defines a Linear Projection

Different nested sets of conditioning variables

Important special case: $w = (x, z)$

$$\begin{aligned}
\underbrace{E(y|x)}_{\mu_1(x)} &= E[\underbrace{E(y|x,z)}_{\mu_2(x,z)}|x] \\
\underbrace{\mu_1}_{\text{observed}} &= E[\mu_2(x, \underbrace{z}_{\text{unobserved}})|x]
\end{aligned}$$

Identification problem: Can we link the estimable $\mu_1(x)$ to the structural $\mu_2(x, z)$ which is the causal relationship of interest?

$$\underbrace{x, z}_{\text{more information}} \quad \text{versus} \quad \underbrace{x}_{\text{less information}}$$

$$\begin{aligned} \mu_1(x,z) &= E(y|x,z) \\ \mu_2(x) &= E(y|x) \end{aligned}$$

By LIE, we have ('integrating z out')

$$\mu_2(x) = E(y|x) = E[E(y|x,z)|x] = E[\mu_1(x,z)|x]$$

$\rightarrow$ allows to study effects of omitted regressors/unobserved components z on the relationship between y and x.

Example: Wage Equation

$$
\begin{aligned}
& E(wage|educ, exper) \\
&= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 educ \cdot exper \\
&= E(wage|educ, exper, exper^2, educ \cdot exper)
\end{aligned}
$$

by LIE, i.e. it is redundant to condition on $exper^2$ and $educ \cdot exper$.

Conditional Variance

The conditional variance of y given x is defined as

$$\begin{aligned} Var(y|x) = E(u^2|x) &\equiv \sigma^2(x) \equiv E[(y - E(y|x))^2|x] \\ &= E(y^2|x) - [E(y|x)]^2 \end{aligned}$$

Note: $\sigma^2(x)$ is a random variable when x is viewed as a random vector.

Properties:

$$Var(a(x)y + b(x)|x) = [a(x)]^2 Var(y|x)$$

Decomposition of variance (corresponds to LIE)

$$\begin{aligned} Var(y) &= E[Var(y|x)] + Var(E(y|x)) \\ &= \underbrace{E[\sigma^2(x)]}_{\text{average conditional variance}} + \underbrace{Var(\mu(x))}_{\text{variance of condtional expectation}} \end{aligned}$$

$$\text{where } \mu(x) = E(y|x).$$

Extension (further conditioning variable z)

$$Var(y|x) = E[Var(y|x,z)|x] + Var[E(y|x,z)|x]$$

Consequently:

$$E[Var(y|x)] \geq E[Var(y|x,z)]$$

$\rightarrow$ further conditioning variables z reduce the average conditional variances.

Linear Projections

Even though a structural CE (conditional expectation) $E(y|x)$ is typically not a linear function of x, it is possible to use the linear projection of y on the random variables $(x_1, ..., x_k) =: x$

$$\underbrace{L(y|1, x_1, ..., x_k)}_{\text{(including an intercept)}} = L(y|1, x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

$$= \beta_0 + x\beta$$

$$\text{where } \beta := [Var(x)]^{-1} Cov(x, y)$$

$$\beta_0 = E(y) - E(x)\beta = E(y) - \beta_1 E(x_1) - ... - \beta_k E(x_k)$$

Variance–Covariance matrix is the $(k \times k)$-matrix

$$Var(x) = \begin{pmatrix} Var(x_1) & \dots & Cov(x_k, x_1) \\ Cov(x_2, x_1) & \ddots & \\ \vdots & & \\ Cov(x_k, x_1) & \dots & Var(x_k) \end{pmatrix} = E[(x - E(x))(x - E(x))']$$

Note:

$$x - E(x) = \begin{pmatrix} x_1 - E(x_1) \\ \vdots \\ x_k - E(x_k) \end{pmatrix}$$

and

$$(x - E(x))' = (x_1 - E(x_1), ..., x_k - E(x_k))$$

$$Cov(x, y) = \begin{pmatrix} Cov(x_1, y) \\ \vdots \\ Cov(x_k, y) \end{pmatrix} \qquad (k \times 1)\text{-vector}$$

Linear projection with a zero intercept

$$\begin{aligned} L(y|x) &= L(y|x_1,...,x_k) = \gamma_1 x_1 + ... + \gamma_k x_k = x\gamma \\ \text{where} \quad & \gamma := [E(x'x)]^{-1}E(x'y) \end{aligned}$$

The linear projection can be derived as the linear predictor minimizing the mean square prediction error ($\equiv$ Best linear predictor or least squares linear predictor), i.e.

$$\min_{b_0, b \in \mathbb{R}^k} E[(y - b_0 - xb)^2]$$

yields β and β_0 as defined.

Using the linear projection

$$L(y|x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

define the error term u by

$$u := y - L(y|x)$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

By definition of linear projections:

$$Eu = 0 \text{ and } Cov(x_j, u) = 0 \quad (j = 1, \ldots, k)$$

Note: This does not imply independence between x and u or mean independence $E(u|x) = 0$

Primary use of linear projections: Obtaining estimable equations involving the parameters of an underlying conditional expectation of interest. Appendix of WO Chapter 2 contains more results on conditional expectations etc. which will be useful later.

Derivation of the OLS Estimator and its Asymptotic Properties

Population equation of interest:

$$y = x\beta + u$$

where: x is a $1 \times K$ vector
$\beta = (\beta_1, \ldots, \beta_K)$
$x_1 \equiv 1$: with intercept

Sample of size N: $\{(x_i, y_i) : i = 1, \ldots, N\}$
i.i.d. random variables where x_i is $1 \times K$ and y_i is a scalar.

For each observation

$$y_i = x_i\beta + u_i$$

Consistency

Assumption OLS.1: $E(x'u) = 0$

Assumption OLS.2: $rank(Ex'x) = K$
$\rightarrow$ expected outer product matrix has full rank, i.e.

$$Ex'x = \begin{pmatrix} 1 & Ex_2 & \dots & Ex_K \\ Ex_2 & Ex_2^2 & \dots & Ex_2 x_K \\ \vdots & \vdots & \ddots & \vdots \\ Ex_K & Ex_K x_2 & \dots & Ex_K^2 \end{pmatrix} \quad \text{is invertible}$$

Under OLS.1 and OLS.2, the parameter vector β is identified, which is equivalent to saying that β can be written in terms of population moments (and of course be solved for!)

To see this:

$$
\begin{aligned}
y &= x\beta + u \\
x'y &= x'x\beta + x'u \\
Ex'y &= Ex'x\beta + \underbrace{Ex'u}_{=0} \quad \text{by OLS.1} \\
\beta &= (Ex'x)^{-1}Ex'y \quad \text{by OLS.2}
\end{aligned}
$$

Because (x, y) is observed $\rightarrow$ β is identified.

Analogy principle:

Choose an estimator by turning the population relationship (based on the probability distribution for the data generating process) into its sample counterpart (based on the empirical distribution for the sample).

Here, the analogy principle implies the method-of-moments: Replace the population moments $E(x'y)$ and $E(x'x)$ (expected values) by their corresponding sample moments (averages).

$$E(x'y) \rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i' y_i$$

$$E(x'x) \rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i' x_i$$

$$\widehat{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' y_i \right) \qquad \text{with } y_i = x_i \beta + u_i$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' (x_i \beta + u_i) \right)$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right) \beta + \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' u_i \right)$$

$$= \beta + \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' u_i \right)$$

WO Theorem 4.1:

Under assumptions OLS.1 and OLS.2, the OLS estimator $\widehat{\beta}$ obtained from a random sample following the population model (5) is consistent for β.

$\rightarrow$ Simplicity should not undermine usefulness.
$\rightarrow$ Whenever estimable equation is of the form then consistency follows.

Under the assumption of theorem 4.1, $x\beta$ is the linear projection of y on x.

$\rightarrow$ OLS estimates linear projection consistently (also in cases such as y being a binary variable) ... and conditional expectations that are linear in parameters.

If either OLS.1 or OLS.2 fail, β is not identified
$\rightarrow$ typically because x and u are correlated.

OLS estimator not necessarily unbiased under OLS.1 and OLS.2 (Jensen's Inequality)

$$E\left[\left(\frac{1}{N}\sum_{i=1}^{N}x_i'x_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}x_i'u_i\right)\right]$$
$$\neq E\left[\frac{1}{N}\sum_{i=1}^{N}x_i'x_i\right]^{-1}\underbrace{E\left[\frac{1}{N}\sum_{i=1}^{N}x_i'u_i\right]}_{=0}$$

$\rightarrow E(u|x)=0 \quad$ implies $\quad E\hat{\beta}=\beta$ (unbiasedness) because of LIE.

We do not need to assume independence
$\rightarrow Var(u|x)$ unrestricted.

Asymptotic distribution of the OLS estimator

Rewrite

$$\widehat{\beta} = \beta + \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i' u_i \right)$$

as

$$\sqrt{N}(\widehat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i' u_i \right)$$

We know $\left[\left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} - A^{-1} \right] = O_p(1)$

Also $\{(x_i' u_i) : i = 1, 2 \ldots\}$ is i.i.d. sequence with $E x_i' u_i = 0$ and we assume each element has a finite variance. Then the central limit theorem implies:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i' u_i \xrightarrow{d} N(0, B)$$

where B is a $K \times K$ matrix: $B \equiv E(u^2 x' x)$

Recall: $x'x$ is the outer product of the $K \times 1$ row vector x

This implies

$$\sqrt{N}(\widehat{\beta} - \beta) = A^{-1} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i' u_i \right) + o_p(1)$$

Under Heteroskedasticity: $\sqrt{N}(\widehat{\beta} - \beta) \overset{a}{\sim} N(0, A^{-1} B A^{-1})$

Under Heteroskedasticity: $\sqrt{N}(\widehat{\beta}-\beta) \overset{a}{\sim} N(0, A^{-1}BA^{-1})$

Under Homoskedasticity:

Assumption OLS.3: $E(u^2x'x) = \sigma^2 Ex'x$

where $\sigma^2 = Eu^2 = Var(u)$

WO Theorem 4.2 (Asymptotic Normality of OLS):

Under Assumptions OLS.1 - OLS.3: $\sqrt{N}(\widehat{\beta} - \beta) \overset{a}{\sim} N(0, \sigma^2 A^{-1})$

Proof: Use $B = \sigma^2 A$ q.e.d.

Practical usage:

Treat $\widehat{\beta}$ as approximately jointly normal with expected value β and Variance-Covariance-Matrix (VCOV) $V = \frac{\sigma^2}{N}[Ex'x]^{-1}$.

V is estimated by

$$\widehat{Avar(\widehat{\beta})} = \frac{\widehat{\sigma}^2}{N}\left[\frac{1}{N}\sum_{i=1}^{N} x_i' x_i\right]^{-1} = \widehat{\sigma}^2 \left(X'X\right)^{-1}$$

and

$$\widehat{\sigma}^2 \equiv s^2 = \frac{1}{N-K}\sum_{i=1}^{N}\widehat{u}_i^2$$

Heteroskedasticity

Failure of assumption OLS.3: $E(u^2x'x) = \sigma^2 E(x'x)$ has nothing to do with consistency of OLS estimator $\widehat{\beta}$ (WO theorem 4.1) and the proof of asymptotic normality is still valid but the final asymptotic variance is different.

Heteroskedasticity robust inference

Often we want to stick to the consistent estimator $\widehat{\beta}$
$\rightarrow$ because no correct specification of $Var(y_i|x_i)$ available
$\rightarrow$ WLS generally inconsistent for linear projections (e.g. when OLS.1 holds but $E(u|x) \neq 0$)

Appropriate asymptotic variance

Without OLS.3 the asymptotic variance of $\widehat{\beta}$ is $Avar(\widehat{\beta}) = \frac{1}{N} A^{-1} B A^{-1}$

A^{-1} is consistently estimated by $\left(\frac{1}{N} \sum_{i=1}^{N} x_i' x_i \right)^{-1} = \widehat{A}^{-1}$

B is consistently estimated by $\left(\frac{1}{N} \sum_{i=1}^{N} u_i^2 x_i' x_i \right)$

We replace the unobserved error terms u_i by the estimated residuals $\widehat{u}_i = y_i - x_i \widehat{\beta}$

$\widehat{B} = \frac{1}{N} \sum_{i=1}^{N} \widehat{u}_i^2 x_i' x_i \xrightarrow{p} B$

Heteroskedasticity-robust variance estimator

$$\widehat{Avar}(\widehat{\beta}) = \frac{1}{N}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} = (X'X)^{-1}\left(\sum_{i=1}^{N} \widehat{u}_i^2 x_i' x_i\right)(X'X)^{-1}$$

often called White standard errors, White-Eicker standard error, or Huber standard errors.

Typically with degrees-of-freedom adjustment to improve finite sample properties.

$$\widehat{Avar}(\widehat{\beta}) = \frac{1}{N-K}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} = (X'X)^{-1}\left(\frac{N}{N-K}\sum_{i=1}^{N} \widehat{u}_i^2 x_i' x_i\right)(X'X)^{-1}$$

t-statistics, χ^2-statistics (but not F-statistics based on comparison of sums of squared residuals in restricted and unrestricted model!) can be used in the usual way.

$H_0 : R\beta = r$

$$W = (R\widehat{\beta} - r)'(R\widehat{V}R')^{-1}(R\widehat{\beta} - r)$$

where $\widehat{V}$ is the heteroskedasticity consistent estimate of $Avar\widehat{\beta}$

Applied work often uses artificial (asymptotic) F-statistics $= \dfrac{W}{Q}$ and adjusts the degrees-of-freedom of $\hat{V}$ as above. This asymptotic F-statistic is assumed to be $F_{Q,N-K}$-distributed under H_0.