Lecture 2

Where are we going?

a) descriptive modeling (weeks 1-5)

b) causal modeling (weeks 6-10)

c) prediction (weeks 11-15)

descriptive modeling

Often we are interested in trying to *summarize the relationship* between some "outcome" y and some other variables $x = (x_1, x_2 ... x_J)$.

- we **aren't** necessarily trying to measure the causal effect of x_j on y

- we are trying to take account of the fact that y may be strongly related to some $x_j's$ and only weakly related to others.

- e.g.: what is the relationship between earnings (y), gender (x_1), and other characteristics, like education (x_2)?

- our benchmark: $E[y|x]$ the "conditional expectations function"

- benchmark: $E[y|x]$ or CEF

- we are going to approximate this with a "linear regression function"

- we'll consider 2 regression functions:

$-$the "population" regression: the function we could estimate with ∞ sample

$-$the "sample" regression: the one we can actually estimate with a given sample

Two items of review: vector notation and conditional probability.

a. vectors

Suppose we are interested in the OLS regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{1}$$

where $i = 1...N$ indexes elements of a sample. Here (x_{1i}, x_{2i}, y_i) are observed values of two *covariates a*nd our *outcome of interest* (y) for unit i. We can define the 3-row vectors x_i and β :

$$x_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Using these vectors we can write the model in vector notation:

$$y_i = x_i'\beta + u_i \tag{2}$$

What happens when we differentiate the dot product $x_i'\beta$ w.r.t. β?

$$\frac{\partial(x_i'\beta)}{\partial\beta} = \begin{bmatrix} \frac{\partial(x_i'\beta)}{\partial\beta_1} \\ \frac{\partial(x_i'\beta)}{\partial\beta_2} \\ \dots \\ \frac{\partial(x_i'\beta)}{\partial\beta_K} \end{bmatrix} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{iK} \end{bmatrix} = x_i$$

Suppose we have k equations in k unknowns of the form:

$$\begin{aligned} a_{11}b_1 + a_{12}b_2 + ... + a_{1k}b_k &= c_1 \\ a_{21}b_1 + a_{22}b_2 + ... + a_{2k}b_k &= c_2 \\ ... \\ a_{k1}b_1 + a_{k2}b_2 + ... + a_{kk}b_k &= c_k \end{aligned}$$

This system can be represented as the matrix equation $Ab = c$, where

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ ... & & & ... \\ a_{lk1} & a_{k2} & ... & a_{kk} \end{pmatrix}, \ c = \begin{pmatrix} c_1 \\ c_2 \\ ... \\ c_k \end{pmatrix}$$

A unique solution for b will exist if A has "full rank": then A is "invertible" and $b = A^{-1}c$.

b. Review of probability

x, y are two r.v.'s, joint p.d.f $f(x, y)$

marginal densities $f(x) = \int_y f(x, y)dy$, $f(y) = \int_x f(x, y)dx$

conditional density

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

Note that this means $f(x, y) = f(y|x)f(x)$.

$E[y] = \int_y y f(y)dy$

$E[y|x] = \int_y y f(y|x)dy$

Law of interated expectations (LIE):

$$E[E[y|x]] = E[y]$$

Proof:

$$\begin{aligned}
E[E[y|x]] &= \int_x E[y|x] f(x) dx \\
&= \int_x \int_y y f(y|x) dy f(x) dx \\
&= \int_y \int_x y f(y|x) f(x) dx dy \\
&= \int_y y \left(\int_x f(x,y) dx \right) dy \\
&= \int_y y f(y) dy
\end{aligned}$$

Two excellent properties of CEF $E[y_i|x_i]$

1. We can always write $y_i = E[y_i|x_i] + \varepsilon_i$ where $E[\varepsilon_i|x_i] = 0$

and $E[\varepsilon_i h(x_i)] = 0$ for *any* function of x.

Proof: we first show $E[\varepsilon_i|x_i] = 0$:

$$
\begin{aligned}
E[\varepsilon_i|x_i] &= E[(y_i - E[y_i|x_i])|x_i] \\
&= E[y_i|x_i] - E[E[y_i|x_i]|x_i] = 0
\end{aligned}
$$

Next, using LIE:

$$
\begin{aligned}
E[\varepsilon_i h(x_i)] &= E[E[\varepsilon_i h(x_i)|x_i]] \\
&= E[h(x_i)E[\varepsilon_i|x_i]] = 0
\end{aligned}
$$

2. the function $m(x_i) = E[y_i|x_i]$ minimizes $E[(y_i - m(x_i))^2]$

Proof:

$$
\begin{aligned}
y_i - m(x_i) &= y_i - E[y_i|x_i] + E[y_i|x_i] - m(x_i) \\
\Rightarrow (y_i - m(x_i))^2 &= \varepsilon_i^2 + (E[y_i|x_i] - m(x_i))^2 \\
&\quad +2\varepsilon_i(E[y_i|x_i] - m(x_i)) \\
\Rightarrow E[(y_i - m(x_i))^2] &= E[\varepsilon_i^2] + E[(E[y_i|x_i] - m(x_i))^2] \\
&\quad +2E[\varepsilon_i(E[y_i|x_i] - m(x_i))]
\end{aligned}
$$

But the last term is 0, so the minimizing choice is $m(x_i) = E[y_i|x_i]$!

So: we've established that if we want to find the function of x_i, $m(x_i)$ that gives the "best guess" for y_i in the sense of minimizing $E[(y_i - m(x_i))^2]$, then the best choice is $m(x_i) = E[y_i|x_i]$.

Problem: we don't know $f(y_i|x_i)$.

Solution: we'll use the "linear regression function": combination $x_i\beta$.

Recall: given a sample of size N the OLS regression coefficients β solve:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

Consider WLLN for the r.v. $(y_i - x_i'\beta)^2$:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i'\beta)^2 \to E[(y_i - x_i'\beta)^2]$$

The "infeasible" (or population) OLS estimator solves:

$$\min_{\beta} E[(y_i - x_i'\beta)^2]$$

What are the FOC? Consider the derivative w.r.t. jth element of β:

$$\begin{aligned} \frac{\partial x_i'\beta}{\partial \beta_j} &= x_{ji} \\ \Rightarrow \frac{\partial (y_i - x_i'\beta)^2}{\partial \beta_j} &= -2(y_i - x_i'\beta)x_{ji} \end{aligned}$$

So: the foc for the optimal choice β^* that solves:

$$\min_{\beta} E[(y_i - x_i'\beta)^2]$$

are:

$$\begin{aligned} E[-2x_i(y_i - x_i'\beta^*)] &= 0. \\ \Rightarrow E[x_i(y_i - x_i'\beta^*)] &= 0 \end{aligned}$$

How does $x_i'\beta^*$ relate to $E[y_i|x_i]$?

Property #1: If $E[y_i|x_i] = x_i'\beta^e$ then $\beta^* = \beta^e$.

Why? Recall that if we define the CEF error $\varepsilon_i = y_i - E[y_i|x_i]$,

$$E[x_i\varepsilon_i] = 0 \Rightarrow E[x_i(y_i - x_i'\beta^e)] = 0$$

Which implies that β^e satisfies the FOC for infeasible OLS.

This means that *if the true CEF is linear*, then the infeasible OLS represents the CEF.

This happens when $x's$ are dummies since $E[y_i|x_i]$ is $E[y_i|i$ in group k]

Property #2: $x_i'\beta^*$ is the "best" linear approx. to $E[y_i|x_i]$ (best as in *minimum-MSE*)

Proof:

$$\begin{aligned} y_i - x_i'\beta &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i'\beta \\ \Rightarrow (y_i - x_i'\beta)^2 &= \varepsilon_i^2 + (E[y_i|x_i] - x_i'\beta))^2 \\ &\quad +2\varepsilon_i(E[y_i|x_i] - x_i'\beta) \\ \Rightarrow E[(y_i - x_i'\beta)^2] &= E[\varepsilon_i^2] + E[(E[y_i|x_i] - x_i'\beta))^2] \\ &\quad +2E[\varepsilon_i(E[y_i|x_i] - x_i'\beta)] \end{aligned}$$

And as before, $E[\varepsilon_i(E[y_i|x_i] - x_i'\beta)] = 0$. So the infeasible OLS minimand is

$$E[(y_i - x_i'\beta)^2] = E[\varepsilon_i^2] + E[(E[y_i|x_i] - x_i'\beta))^2]$$

So what is β^*? Recall that the objective

$$min_\beta E[(y_i - x_i'\beta)^2]$$

has foc that imply:

$$\begin{aligned} E[x_i(y_i - x_i'\beta^*)] &= 0 \\ \Rightarrow E[x_i x_i']\beta^* &= E[x_i y_i] \\ \Rightarrow \beta^* &= [E[x_i x_i']]^{-1} E[x_i y_i] \end{aligned}$$

We can think of the "population regression" as:

$$y_i = x_i'\beta^* + u_i$$

Notice that $u_i = \varepsilon_i + \{E[y_i|x_i] - x_i'\beta^*\}$, and $E[x_i u_i] = 0$. (why?)

The feasible regression (OLS) minimizes

$$SSR = \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

The foc (in vector form) are:

$$\sum_{i=1}^{N} -2x_i(y_i - x_i'\beta) = 0$$

which implies that

$$\begin{aligned} \sum_{i=1}^{N} x_i x_i' \beta &= \sum_{i=1}^{N} x_i y_i \\ \Rightarrow \widehat{\beta} &= [\sum_{i=1}^{N} x_i x_i']^{-1} \sum_{i=1}^{N} x_i y_i \end{aligned}$$

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

Now using: $y_i = x_i'\beta^* + u_i$

$$\begin{aligned}
\widehat{\beta} &= [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1}\frac{1}{N}\left(\sum_{i=1}^{N} x_i (x_i'\beta^* + u_i)\right) \\
&= [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\beta^* + [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i u_i \\
&= \beta^* + [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i u_i
\end{aligned}$$

The deviation depends on a term that should be small