The Best Linear Prediction model

(only)

Assumption, \( y_i, x_i \) iid with plenty 6+ moments.

That is it.

No other assumptions.

Should this be even called a "model"?

Recall our

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

by LLN, CMT

\[ \beta_0 \]

\[ \beta = \arg \min_{\beta} E (y - x' \beta)^2 \]

\[ \hat{\beta} \]

\[ E_{xy} = E_{x' x'} \hat{\beta} \iff E (x' (y - x' \beta)) = 0 \]

\[ \beta_0 = \left( E_{xx'} \right)^{-1} E_{xy} \]

\[ \beta_0 \] is called BLR
Consider a "model":

\[ y = x^\prime \beta + \epsilon \]

\[ \text{i.i.d. data} \]

\[ E \epsilon | x = 0 \]

Is this a model? \\
for \( \beta_0 = (E x x')^{-1} E x y \) BLP. \\
let \( \epsilon = y - x^\prime \beta_0 \)

then \( E x \epsilon = E x (y - x^\prime \beta_0) \) \\
\[ = E x y - E x x^\prime \beta_0 \]
\[ = 0 \text{ if } \beta_0 = (E x x')^{-1} E x y. \]

Some textbooks said \\
"OLS is consistent if \( E \epsilon x = 0 \)." \\

What do you think of these textbooks? \\
consistent for what? for \( \beta_0 \)? \\
But what is \( \beta_0 \)? \\
Other than saying that we are running OLS using i.i.d. data, what else is in \( \beta_0 \)?
Suppose you want to conduct inference for $p_o = (E x' x)^{-1} E x y$ the BLP.

not for making causal statements but perhaps for constructively prediction intervals.

\[
\sum \left( \mathbf{b}_{oas} - p_o \right) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i \left( y_i - \mathbf{b}_{oas} \right)
\]

Recall that $E x_i (y_i - \mathbf{b}_{oas}) = 0$ by definition of $p_o$ as BLP.

then by CLT

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i (y_i - \mathbf{b}_{oas}) \overset{d}{\rightarrow} N(0, E (y_i - \mathbf{b}_{oas})^2 x_i x_i')
\]

So

\[
\sum \left( \mathbf{b}_{oas} - p_o, BLP \right) \overset{d}{\rightarrow} N(0, \mathbf{H}^{-1} \mathbf{S} \mathbf{H}^{-1})
\]

\[
\mathbf{H} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \Rightarrow \mathbf{H}
\]

\[
\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{b}_{oas}) x_i x_i' \Rightarrow \mathbf{S}
\]
This is the same as Huber-White heteroscedasticity robust std error.

In other words, the Huber-White std errors:

\[
\left( \frac{1}{n} \sum_{i=1}^{n} x_i x'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} e_i^2 x_i x'_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} x_i x'_i \right)^{-1}
\]

provides consistent inference of

\[
\hat{B}_0 = (B x x')^{-1} E x y
\]

the BLP based on \( B x x' \) without making any assumptions except that the data \( (y_i, x_i) \) is iid with enough number of moments.
"Conditional" Best Linear Predictor.

(Abadie, Imbens & Zheng, 2012)

AIZ defines C-BLP as

\[ \beta_{no} = \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i E(y_i | X_i) \]

Recall that the (unconditional) BLP is

\[ \beta_0 = \left( EX_i X_i' \right)^{-1} EX_i E(y_i | X_i) \]

Why do you care about \( \beta_{no} \)?

maybe you want make best (linear) prediction based on the same sample of \( X_i \), rather than a different set of draws of \( X_i \) from the population

\[ \beta_{no} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left( E(y_i | X_i) - X_i \beta \right)^2 \]
\[ \beta_{nl} = \beta_0 \] if \( E(y_i | x_i) = x_i' \beta_0 \) a linear function

otherwise \( \beta_{nl} \neq \beta_0 \) in general.

\[
\Delta_n (\beta_{nl} - \beta_{n0}) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i (y_i - E(y_i | x_i)) \\
\Delta_n (\beta_{n0} - \beta_0) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i (E(y_i | x_i) - x_i' \beta_0) \\
\text{for } \beta_0 = (E(x x')^{-1} E x y) \\
A = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \rightarrow H = E x x' \\
\text{Now} \\
\begin{pmatrix} \pm \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i (y_i - E(y_i | x_i)) \\ \pm \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i (E(y_i | x_i) - x_i' \beta_0) \end{pmatrix} \rightarrow N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right) \\
\text{where } \Sigma_1 = E (y_i | x_i) x_i x_i' \\
\Sigma_2 = E \left( E(y_i | x_i) - x_i' \beta_0 \right)^2 x_i x_i' \]
\[ \sum_2 \neq 0 \text{ in general if } \mathbb{E}(y_i | x_i) = x_i \beta_0 \]

\[ \sum_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^2 (y_i (x_i) x_i x_i) \]

where \( \hat{\sigma}^2 (y_i | x_i) \) is any consistent (nonparametric) estimate of \( \sigma^2 (y_i | x_i) \)

\[ \sum_1 = \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}(y_i | x_i) - x_i \beta_{ols} \right)^2 x_i x_i \]

where \( \mathbb{E}(y_i | x_i) \) is any consistent (nonparametric) estimate of \( \mathbb{E}(y_i | x_i) \).

In general, Huber-White consistency estimates \( \sum = \sum_1 + \sum_2 \), but not \( \sum_1 \), unless \( \sum_2 = 0 \).

If \( x_i \)'s are random, \( B_n \)'s are random too.