25/10/23

OLS tries to find the best linear combination

$x_i'\beta = x_1\beta_1 + x_2\beta_2$

OLS estimator. (coefficient)

$\beta \in \mathbb{R}^p$ is the solution to:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum (y_i - x_i'\beta)^2$$

We usually specify OLS by equations.

$y_i = x_i'\beta + \underset{\sim}{\epsilon}$

unobserved term.

Suppose this is your data:

Height	Male.	Taipei
170	1	1
180	1	0
165	0	1

We can consider:

① $Height = \beta_0 + \beta_1 Male + \varepsilon_i$

$\Rightarrow \min_{\beta_0\beta_1} \frac{1}{n}\sum (y_i - \beta_0 - \beta_1 Male)^2$

② $Height = \beta_0 + \beta_1 Male + \beta_2 Taipei + \varepsilon_i$

$\Rightarrow \min_{\beta_0 \beta_1 \beta_2} \frac{1}{n}\sum (y_i - \beta_0 - \beta_1 Male - \beta_2 Taipei)^2$

These are specifications of estimators, not the specification of the DGP (Data Generating Process)

Two main applications of OLS

① Predict on:
given new observation, x_{n+1}:

How to predict?

$(\hat{y}_{n+1} = x_{n+1}'\hat{\beta})$

③ estimate marginal effect:

if $x_j \uparrow$ by 1 unit
how much does y change?

(Ans: β_j)

$y_i = x_{i1}+\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p + \varepsilon_i$

Ex. CLTV prediction

Customer Life Time Value.

RFM framework.

Recency: when was the last time this customer make a transaction?

Frequency: how frequent a customer make its transaction.

Monetary: avg. sales per trans.

We can predict CLTV by

$CLTV_i = \beta_0 + \beta_1 R_i + \beta_2 F_i + \beta_3 M_i + \varepsilon_i$

Ex. (marketing mix)

$Sales_t$: sale at day t.

IG_t: advertisement spending (on IG) at day t

FB_t: " " (on FB) " " "

$Google_t$: " " (on Google) " "

Regress:

$$Sales_t = \beta_0 + \beta_1 IG_t + \beta_2 FB_t + \beta_3 Google_t + \varepsilon_t.$$

You should spend more on platform on higher β.

OLS estimator:

= some sorts of sample mean.

This can be generalized.

$$E[Y|X] = \int_{-\infty}^{\infty} y f(y|x)\, dy$$

where $f(y|x)$ is the conditional pdf.

$E[Height \mid Male = 1]$

$\Rightarrow$ Average height male.

$E[Height \mid Male = 0]$

$\Rightarrow$ Average height female.

Rmrk: It's useful to think of

$$E[Y|X=x] = f(x)$$

as a function of x

$$f(Male) = E[Y|Male] = \begin{cases} \text{avg. height women}: Male = 0 \\ \text{avg. height male}: Male = 1 \end{cases}$$

Note that $E[Y|X]$ $(=f(X))$ is a transformation of X, $E[Y|X]$ is also a R.V. (Random Variable)

P.S. ☆

$E[Y|X=x]$ is constant. $E[Y|X]$ is R.V.

<u>Proposition</u> (Law of Iterated expectation)

$$E[Y]=E[\underbrace{E[Y|X]}_{\text{is itself a R.V., its expectation is } Y}]$$ ⇒ it's an accounting equation (恒等式)

$$E[Y|X]=\begin{cases}\text{Male's average.}\\ \text{Female's average.}\end{cases}$$

Quiz 2:

$\min_{c\in\mathbb{R}} E[(Y-c)^2]$, what is $c^*=$?

Ans: $\bar{Y}$ or $E[Y]$

Recall Quiz 1:

$\min_{a\in\mathbb{R}} \frac{1}{n}\sum(y_i-a)^2$

$a=\bar{y}_i$

Proof:

$$\begin{aligned}
&E[(Y-c)^2]\\
&=E[((Y-EY)+(EY-c))^2]\\
&=E[(Y-EY)^2+2(Y-EY)(EY-c)+(EY-c)^2]\\
&=E[(Y-EY)^2]+\underbrace{2E[(Y-EY)(EY-c)]}_{(EY-c)2(\underbrace{E[Y-EY]}_{EY-EY=0})}+E[(EY-c)^2]\\
&\qquad\qquad\qquad\qquad =0\\
&=E[(Y-EY)^2]+\underbrace{(EY-c)^2}_{\geq 0}\\
&\geq E[(Y-EY)^2]=Var(Y)
\end{aligned}$$

$E[(Y-c)^2]\geq Var(Y)$ for any $c\in\mathbb{R}$

and equality holds when $c=EY \Rightarrow c^*=EY$

Prop. Conditional Mean: is the best predictor.

Let $g(\cdot)$ be a function on

$\mathcal{X} \to \mathcal{Y}$, that $x \mapsto g(x)$

(you can think of $g(\cdot)$ as a predictor of y based on x)

Consider

$$\min \underbrace{E[(Y - \overset{\text{a prediction}}{g(X)})^2]}_{\text{prediction error.}}$$

Then, $g(x) = E[Y|X=x]$ is the solution.

So, conditional mean is the best predictor (under squared loss)

The proof is similar to $\min_{c \in \mathbb{R}} E[(Y-c)^2]$

Objective function on $E[(Y-g(X))^2]$

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

$$\Rightarrow E[(Y-E(Y|X))^2] + 2E[(Y-E(Y|X)) \cdot (E(Y|X)-g(X))]$$
$$+ E[(E(Y|X) - g(X))^2]$$

$\Rightarrow$ The mid. term.

$$E[2(Y-g^*(X))(g^*(X)-g(X))]$$

Let $g^*(X) = E(Y|X)$

$$\to 2E[E[(Y-g^*(X))\underbrace{(g^*(X)-g(X))}_{\text{Constant (given } X)}|X]]$$

$$= E[(g^*(X)-g(X))(E[Y-g^*(X)|X])]$$

$0 \because E[Y-g^*(X)|X] = E[Y|X] - g^*(X)$

$g^*(X) = E[Y|X]$.

$= 0$

After showing mid. term $= 0$

$$E[(Y-g(X))^2]$$
$$= E[(Y-g^*(X))^2] + E[\underbrace{(g^*(X)-g(X))^2}_{\geq 0}]$$
$$\geq E[(Y-g^*(X))^2]$$
$$= E[(Y-E(Y|X))^2]$$

Note: We've shown the for any $g(\cdot)$. $E[(Y-g(X))^2]$

$$\geq \underbrace{E[(Y-E(Y|X))^2]}_{\text{lower bound.}}$$

and lower bound is attained when $g(X) = E[Y|X]$

So conditional E is the best predictor.

In the proof : for any $g(\cdot)$

$$E[(Y-g(X))^2]$$

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

and. $E[(Y-g(X))^2] \geq E[(Y-E(Y|X))^2]$

$$g^*(X) = E[Y|X]$$

Remark : $E[Y|X]$ is best for L_2-loss

$E[(Y-g(X))^2]$, but not necessarilly for other loss function.

e.g. $\min E[|Y-g(X)|]$.

Remark 2:

While we know $E[Y|X]$, we still need to estimate $E[Y|X]$.

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

OLS:

$$\min_{\beta} \frac{1}{n}\sum(y_i - x_i'\beta)^2$$

while $n \to \infty$

$\frac{1}{n}\sum(y_i - x_i'\beta)^2 \to E[(y_i - x_i'\beta)^2]$ by L.L.N.

Replacing $g(x)$ to $x'\beta$

$$\min_{\beta} E[(y_i - x_i'\beta)^2] \Leftrightarrow \min E[(E(Y|X)-X'\beta)^2]$$

Since $E[(y-g(x))^2] = \underbrace{E[(y-E(Y|X))^2]}_{\text{does not depend on } g} + E[(E(Y|X)-X\beta)]$

So OLS is equivalently $\min_{\beta} E[(E(Y|X)-X'\beta)^2]$ i.e. OLS is the best linear approximation of $E[Y|X]$.

Remark: even if the relationship is not causal, the regression on is still useful for prediction.

(we only need correlation for prediction)

We only need causality if we want to predict effect of intervention.

Failure to identification

eg.1 $n=4$ Can't identify (Situation 1)

Height	Female	Male
170	0	0
180	0	0
163	1	1
157	1	1

The regression:

$y_i = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Taipei} + \varepsilon$

if all female are born in Taipei
male not Taipei.

(Situation 2)

eg.2 $n=3$ $p=4$

y_0	x_1	x_2	x_3	x_4
1	1	0	0	0
3	0	1	0	0
4	0	0	1	0

all 0

reg.: $y_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

$\frac{1}{n}\sum(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4})^2$

$\Rightarrow (\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$ is a solution

$\Rightarrow (\beta_1^*, \beta_2^*, \beta_3^*, \tilde{\beta}_4)$ $\tilde{\beta}_4$ can be any value.

"Observational equivalence"

State 1: Host 1 is popular. but Host 2 is not.

State 2: Host 2 is popular. but Host 1 is not.

But if host 1 & 2 always partner up in TV show
you can't distinguish whether State 1 or 2 is true since they are
"observationally equivalent".

other example:

S1: $Height = \beta_0 + 5 \cdot Male + 3\, Taipei + \varepsilon_i$

S2: $Height = \beta_0 + 3\, Male + 5\, Taipei + \varepsilon_i$

If All males are from Taipei,
you can't distinguish state 1 & 2.

End of the Lecture!!!