Day 13

Regression

$\hat{y} = b_0 + b_1 x_i$

$\frac{\Delta \hat{y}}{\Delta x_i} = b_1 = \text{constant}$

$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2$

$= b_0 + b_1 \ln x_i$

$b_1 = \frac{\Delta \hat{y}}{\Delta x_i} = f(x_i)$

1. $\frac{\Delta \hat{y}}{\Delta x_i}$ depends on x_i 이유?

2. $\frac{\Delta \hat{y}}{\Delta x_i}$ depends on x_i

$y_i = b_0 + b_1 \text{edu}_i + u_i$

$\Delta y \neq 1$ 교육

$y_i = b_0 + b_1 \text{edu}_i + b_2 \text{sex} + b_3 \text{edu}_i * \text{sex}_i + u_i$

$\Delta y \neq 15$ 교육

$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$

$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$

$E(y|x) = \hat{y} = b_0 + b_1 x$

$E(y|x=1) = \hat{y}_1 = b_0 + b_1 x_1$

$E(y|x=0) = \hat{y}_0 = b_0 + b_1 x_0$

$E(y|x=1) - E(y|x=0) = b_1$

$E(y|x=1) - E(y|x=0) = a\hat{y}$ [illegible]

$E(b_0 + b_1 x_1) - (b_0 + b_1 x) = b_1$

- $E(y|x_i) = \hat{y} = b_0 + b_1 x_i$: OLS regression model

dependent var — regressor — ordinary least squares — OLS estimator — MLE estimator

$\hat{y}_i = b_0 + b_1 x_i$

$\min[u_1^2 + u_2^2 + u_3^2 + \dots + u_n^2]$

$\sum u_i^2$

Least squares principle: $\min \sum u_i^2 \Rightarrow b_0, b_1$

$\min[y_1 - (b_0 + b_1 x_1)]^2 + [y_2 - (b_0 + b_1 x_2)]^2 + \dots + (y_n - (b_0 + b_1 x_n))^2$

i	x_i	y_i	$b_0 + b_1 x_i = \hat{y}_i$	u_i
1	x_1	y_1	$\hat{y}_1 = b_0 + b_1 x_1$	$y_1 - \hat{y}_1$
2	x_2	y_2	$b_0 + b_1 x_2 = \hat{y}_2$	$y_2 - \hat{y}_2$
3	x_3	y_3		
⋮	⋮	⋮		⋮

$b_0 = 20$

$b_1 = 0.73$

OLS estimator

N

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
⋮	⋮	⋮
N	x_N	y_N

n

i	x_i	y_i
1	x_1	y_1
2	⋮	⋮
⋮	⋮	⋮
n	x_n	y_n

population regression line

OLS: $\hat{y}_i = \beta_0 + \beta_1 x_i$

unique population parameter — unique population parameter — point estimate

sample regression

OLS $\hat{y}_i = b_0 + b_1 x_i$

variable sample statistic — variable sample statistic

$\hat{y}_i = b_0 + b_1 x_i$

$\bar{x} = 0.73$ u

$b_1 = 0.73$ β_1

$[b_1 \pm 1.96 \cdot \sigma_{b_1}]$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

σ_{b_1} 2.5% b_1 $\beta_1 = 0$ 0.73 0.73

OLS regression:

$y_i = b_0 + b_1 x_i + u_i$

- how much variation in y is explained by x_i
- fraction of total variation in y that is explained by x_i

$\parallel$

R squared $= R^2$ = coefficient of determination $0 \le R^2 \le 1$

$R^2 = \frac{\text{variation in } y \text{, explained by } x}{\text{Total variation in } y} = 1 - \frac{\text{variation in } y \text{, explained by } u_i}{\text{Total variation in } y}$

$= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum u_i^2}{\sum(y_i - \bar{y})^2}$

$(y_i - \hat{y}_i)^2$

$\parallel$ 0

predictive power

$y_i = b_0 + b_1 x_i + u_i$

trt — all other factors

constant — $E(u_i|x_i)$ = constant

$E(u_i|x_i) = 0$:

$y_i = b_1 x_i + u_i$

$E(y_i) = b_1 E(x_i) + E(u_i)$

b_0

$E(u_i|x_i) = a_0 + a_1 x_i + \varepsilon_i$

$b_0 + b_1 x$, $x \in [\ ]$

$E(y_i|x+1) - E(y_i|x) = b_0 + b_1(x+1) + E(u_i|x+1)$

$- [b_0 + b_1 x + E(u_i|x)]$

$= b_1 + [E(u_i|x+1) - E(u_i|x)]$

β_1 — $=0$ — $\neq 0$

$\parallel$

$b_1 : \frac{\Delta y}{\Delta x}$

$y_i = b_0 + b_1 x_i + u_i$

$E(u_i|x_i) = 0$

$E(y_i|x_i) = b_0 + b_1 x_i + [E(u_i|x_i)]$

$a_0 + a_1 x_i$

$b_1 \neq \beta_1 + [E(u_i|x+1) - E(u_i|x)]$

$\parallel$ 0

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

$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + u_i$

Causal relationship

cause

y_i ← x_i

variation

y_i	b_0
100	b_0
80	b_0
50	b_0
⋮	b_0
	b_0
	b_0
	b_0

$y_i = b_0 + b_1 \text{CS}_i + u_i$

$E(u_i|x_i) = 0$

$y_i = b_0 + b_1 \text{CS} + b_2 \text{Inc}_i + u_i$

$y \leftarrow$ model $= b_0 + b_1 x_1 + b_2 x_2$

$y_i = b_0 + b_1 \text{CS}_i + u_i$

$E(y|X=1) = b_0 + b_1 1 = b_0 + b_1 = 100 + E(u_i|X=1)$

$E(y|X=0) = b_0 + b_1 \times 0 + E(u_i|X=0)$

$= b_0 + E(u_i|X=0)$

$b_0 + E(u_i|X=0)$

$b_0 = E(y|$ 소위 $)$

$E(u_i|x_i) \neq 0$

Question

$R^2 = \frac{SS\text{Model}}{SS\text{Total}} = 1 - \frac{SSEm}{SSTotal}$

$R^2 = 0.0001$

x , u

Prediction — Causal effect

$\hat{y}_i \approx y_i$ — $E(u_i|x_i) = 0$

$R^2 \uparrow$

$y_i = b_1 x_i + u_i$

$E(y_i) = E(b_1 x_i + u_i)$

$E(y_i) = b_1 E(x_i) + E(u_i)$

residual analysis

Spurious correlation

2020 2021

Price = f(cost, expectation, ms, ……)

$y_i \leftarrow x_i$

$R^2 = 0.99$