Binary Response variable Estimation

· 로지스틱 함수.

$$f(x) = \frac{e^x}{1+e^x} \qquad -\infty < x < \infty, \quad 0 \le y \le 1.$$

실수값을 확률값으로 mapping 해줌.

· glm approach

· glm component : random component / linear predictor / link function.

- random component : $\underline{Y} = (y_1, y_2, \cdots y_n)'$. which is r.v.

  $Y \underset{\text{from}}{\sim}$ exponential family.

  이유 : Model likelihood Equation, Asymptotic Dist, fit Algorithms (ridge, LASSO, linear reg, logistic ...)

  에 대한 generalized representation. $S(x) \xrightarrow{D}$ certain dist. $N(\mu, \sigma^2)$

- linear predictor : LC of $\underbrace{\text{Design Matrix}}_{G(=X)}$ and $\underbrace{\text{Parameter}}_{\beta}$ $\doteq G \cdot \beta$. $\mathbb{R}_{n \times p}$, $\mathbb{R}_{p \times 1}$

- link function : link Linear Predictor with $E(y)_{(=\mu)}$

$$g(\widehat{E(y)}) = X\hat{\beta}. \qquad g(\mu) = X\beta.$$

연결함수의 특징 : 단조성(monotonicity), 미분가능성(differentiable)

이유 !?

모형 fitting algorithm을 구하기 위해 : Iterative WLS를 적용할 때 사용하는데 이 때 도함수가 사용됨 $g'(\cdot)$

로지스틱 선형회귀의 경우, $g(\mu) = \underbrace{\text{logit}(\mu)}_{\text{로짓함수}}$.

$$\text{logit}(\mu) = \ln \frac{\mu}{1-\mu}.$$

Applied to logistic regression

$$g(\mu_i) = X \cdot \beta. \iff \ln \frac{\mu}{1-\mu} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_i$$

$$\iff \underset{E(Y_i|X_i)}{\mu_i} = \frac{1}{1+e^{-\beta \cdot X_i}} = p_i$$

$$E(Y|X_1 \cdots X_n) = P(Y=1|X_1 \cdots X_n) \cdot Y=1 + \cancel{P(Y=0|X_1 \cdots X_n) \cdot Y=0}.$$

$$= P(Y=1|X_1 \cdots X_n)$$

$$= \frac{1}{1+e^{-\beta \cdot X_i}}$$

Logit & Probit.

$$\text{logit}(u) = \ln \frac{u}{1-u}.$$

$$\text{probit}(u) = \Phi(u) = P(Z \le u).\ (\text{cdf of standard Normal}).$$

link function이 logit/probit 일때,

$E(Y|X)$는 각각 logistic regression / probit regression 가 된다.

$$\frac{1}{1+e^{-\beta \cdot X}} \qquad \Phi(\beta X)$$

별거 없다!

T/C group 이 대하여 각각

Logit/Probit 모형을 통해 propensity score π_x 를 추정하고,
(regression)

propensity score 값이 차이가 크게 나지 않는 observation 들을.

matching 하여, T/C group이 비슷한 covariate 값들을 갖게끔

matching 시켜준다.

이를 통해 DiD. Causal Inference 를 할 수 있다.

(trt 를 받았을 확률은 어떤거니 !).

trt 를 받았을 확률 $\gamma(z)$ 을 안다면,

$D=\{1, 0\}$ 이었는데 soft clustering 을 통해 (기여된 x가 어떤

cluster에 속할지가 soft하게 나뉘게 됨. (trt O trt X).

경제력 공변량들. 방법

월별변수

ex) 증가소득, 오래된 지원금

표기되면

assignment prob에 대한 힌트를 알면 → 공변량의 parameter들을

어떻게 변화시켜야

Maximize assignment prob을

할수 있을지 나옴.