Modelos para el propensity score que contemplan la positividad.

Datos Faltantes

- $Y_i \in \mathbb{R}$ respuesta de interés perdida al azar.

- $A_i \in \{0, 1\}$, $A_i = 1$ si Y_i es observada $A_i = 0$ en otro caso.

- $\mathbf{X}_i \in \mathbb{R}^m$ vector de variables adicionales.

- Datos observados: $O_1, \cdots, O_n$ i.i.d., donde

$$O_i = \begin{cases} \mathbf{X}_i, Y_i, A_i & \text{si } A_i = 1 \\ X_i, A_i & \text{si } A_i = 0. \end{cases}$$

Objetivo

Estimar el parámetro desconocido $\mu_0 \in \mathbb{R}$

$$\mu_0 =_E [Y] .$$

Identificabilidad.

Supuestos:

1. Missing at Random (MAR)

$$Y \text{ es independiente de } A, \text{ dado } \mathbf{X}.$$

2. Positividad

$$P(A = 1|\mathbf{X}) \geq a > 0.$$

Identificabilidad

:

- Los supuestos garantizan que

$$\mu_0 = E[Y] = E\left[\frac{AY}{P(A=1|\mathbf{X})}\right].$$

Estimación

$$\mu_0 = E[Y] = E\left[\frac{AY}{P(A = 1|\mathbf{X})}\right]. \tag{1}$$

Modelo para el Propensity: (MP)

$$P(A = 1|\mathbf{X}) = \pi(\mathbf{X}, \alpha_0).$$

$\widehat{\alpha}$ estimador de maxima verosimilitud para α_0 bajo MP.

$$\widehat{\mu}_{ipw} = \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(\mathbf{X}_i, \widehat{\alpha})}. \tag{2}$$

Hipótesis de positividad

$$\pi(\mathbf{X}, \alpha_0) = expit(\alpha_0^T \mathbf{X}) = \frac{e^{\alpha_0^T \mathbf{X}}}{1 + e^{\alpha_0^T \mathbf{X}}}.$$

1. $\pi(\mathbf{X}_i\, \alpha_0)$ muy chico,
2. $\pi(\mathbf{X}_i, \widehat{\alpha})$ muy chico,
3. $\widehat{\mu}_{ipw} = \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(\mathbf{X}_i, \widehat{\alpha})}$ inestable.

Modelo Propuesto

$$\pi(\mathbf{X}, \alpha) = (1 - \varepsilon)F(\beta^T \mathbf{X}) + \varepsilon,$$

1. F conocida.
2. $\alpha = (\epsilon, \beta)$ parámetro desconocido $\varepsilon \in \mathbb{R}_{>0}$, $\beta \in \mathbb{R}^p$.

Identificabilidad

$$\pi(\mathbf{X}, \alpha) = (1 - \varepsilon)F(\beta^T \mathbf{X}) + \varepsilon, \tag{3}$$

Supuestos

1. Para todo β
$$\inf_{\mathbf{X}}\{\beta^T \mathbf{X}\} = -\infty.$$
2. X no está concentrado en ningún hiperplano.
3. F inyectiva.

Simulación

- $n = 1000$
- $\mathbf{X} = (X_1, X_2)$ normales independientes con media 0 y desvío 1.3.
- $P(A = 1|\mathbf{X}) = (1 - \varepsilon)\text{expit}(\boldsymbol{\beta}^T\mathbf{X}) + \varepsilon$, $\boldsymbol{\beta} = (0, -1, -1)$, $\epsilon = 0, 0.1, 0.2, 0.3, 0.4$.
- $Y = \boldsymbol{\gamma}^T\mathbf{X} + u$ con $u \sim N(0, 1)$ y $\boldsymbol{\gamma} = (0, -3, 10)$.

Estimadores calculados

$$\pi_1(\mathbf{X}, \alpha) = (1 - \varepsilon)\text{expit}(\boldsymbol{\beta}^T \mathbf{X}) + \epsilon. \quad (m1)$$

$$\pi_2(\mathbf{X}, \alpha) = \text{expit}(\boldsymbol{\alpha}^T \mathbf{X}). \quad (m2)$$

1. $\widehat{\alpha}_{pos} = (\widehat{\varepsilon}, \widehat{\beta})$.
2. $\widehat{\alpha}_{expit}$.
3. $\widehat{\mu}_{ipw,1} = \sum_{i=1}^{n} \frac{A_i Y_i}{\pi_1(\mathbf{X}_i, \widehat{\alpha}_{pos})}$.
4. $\widehat{\mu}_{ipw,2} = \sum_{i=1}^{n} \frac{A_i Y_i}{\pi_2(\mathbf{X}_i, \widehat{\alpha}_{expit})}$.

Sesgo

ε	0	0.1	0.2
$\widehat{\epsilon}$	$8.4e-06$	0.062	0.039
$\widehat{\beta}$	(-0.164, -0.171)	(-0.173, -0.1708)	(-0.118, -0.123)
$\widehat{\mu}_{ipw,1}$	-0.61	-0.091	-0.106
$\widehat{\mu}_{ipw,2}$	-0.1	3.832	2.91

ε	0.3	0.4
$\widehat{\epsilon}$	0.033	0.024
$\widehat{\beta}$	(-0.145,-0.157)	(-0.144,-0.153)
$\widehat{\mu}_{ipw1}$	-0.254	-0.1544572
$\widehat{\mu}_{ipw2}$	2.637	2.166

ECM

ε	0	0.1	0.2
$\widehat{\epsilon}$	3.35e-08	0.0058	0.0033
$\widehat{\beta}$	(0.042, 0.051)	(0.053, 0.051)	(0.0489, 0.185)
$\widehat{\mu}_{ipw1}$	0.421	0.124	0.158
$\widehat{\mu}_{ipw2}$	0.13	36.223	14.3

ε	0.3	0.4
$\widehat{\epsilon}$	0.0054	0.0046
$\widehat{\beta}$	(0.032,0.029)	(0.042,0.042)
$\widehat{\mu}_{ipw1}$	0.191	0.175
$\widehat{\mu}_{ipw2}$	8.112288	5.156

Simulación con covariables uniformes

ε	0	0.1	0.2
$\widehat{\epsilon}$	$3.56 * 10^{-8}$	$7.2 * 10^{-9}$	$4.2 * 10^{-6}$

ε	0.3	0.4
$\widehat{\epsilon}$	$3.5 * 10^{-6}$	$2.2 * 10^{-5}$

Causalidad

- Se quiere decidir si un tratamiento tiene un cierto efecto sobre una variable respuesta.

- Enfoque causal:

  - Mundo donde TODOS siguen un tratamiento$\longrightarrow Z_T$.

  - Mundo donde TODOS siguen un control$\longrightarrow Z_C$.

- Objetivo: identificar y estimar

$$ATE = E[Z_T] - E[Z_C].$$

Identificabilidad y estimación de $E[Z_T]$ y $E[Z_C]$.

- En este mundo ALGUNOS siguen control y OTROS tratamiento midiéndose siempre la respuesta Z.

- Z_T y Z_C no son observables.

- Consistencia: Si en este mundo se sigue tratamiento $Z_T = Z$.

- Luego, si en cada individuo medimos A, Z, $\mathbf{X}$ y asumimos MAR y positividad, podemos pensar que Z_T está perdida en una submuestra ($A = 0$).

Identificabilidad y Estimación de $E[Z_T]$ y $E[Z_C]$.

$$ATE = E\left(\frac{ZA}{p(A=1|\mathbf{X})}\right) - E\left(\frac{Z(1-A)}{(1-p(A=1|\mathbf{X}))}\right).$$

$$\widehat{ATE} = \sum_{i=1}^{n} \frac{Z_i A_i}{\pi(\mathbf{X}_i, \widehat{\alpha})} - \sum_{i=1}^{n} \frac{Z_i\,(1-A_i)}{1-\pi(\mathbf{X}_i, \widehat{\alpha})}.$$

PROPUESTA

$$\pi(\mathbf{X}, \alpha) = (1 - \varepsilon - \delta)F(\beta^T \mathbf{X}) + \varepsilon. \tag{4}$$

Supuestos para identificar:

1. Para todo β

$$\inf_{\mathbf{X}} \{\beta^T \mathbf{X}\} = -\infty \quad \text{y} \quad \sup_{\mathbf{X}} \{\beta^T \mathbf{X}\} = +\infty.$$

2. $\mathbf{X}$ no está concentrado en ningún hiperplano.
3. F inyectiva.

$$\pi(\mathbf{X}, \alpha) = (1 - \varepsilon - \delta) F(\beta^T \mathbf{X}) + \varepsilon.$$

Estudiamos varias simulaciones con resultados parecidos al caso anterior (Datos Faltantes.)

Cosas por ver....

1. Datos reales
2. Distribución asintótica del ATE.