1. 概率统计

§1 特殊分布的性质

1. χ^2: $X_1,\cdots,X_n \overset{i.i.d.}{\sim} N(0,1)$, $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$, $\mu=n$, $\sigma^2=2n$
2. t: $X\sim N(0,1)$, $Y\sim\chi^2(n)$ 独, $Z=\frac{X}{\sqrt{Y/n}}\sim t(n)$, $\mu=0$, $\sigma^2=\frac{n}{n-2}$
3. F: ① $X\sim\chi^2(n_1)$, $Y\sim\chi^2(n_2)$ 独, $Z=\frac{X/n_1}{Y/n_2}\sim F(n_1,n_2)$
   ② $F_{1-\alpha}(n_1,n_2)=\frac{1}{F_\alpha(n_2,n_1)}$

§2 重要定理

$A=\begin{bmatrix}a&b\\c&d\end{bmatrix}\Rightarrow A^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d&-b\\-c&a\end{bmatrix}$

1. Cauchy-Schwarz inequality: $[E(XY)]^2\le E(X^2)E(Y^2)$
2. Law of large number: $Z=\frac1n\sum_{i=1}^n X_i\to_p\mu$
3. Central limit theorem: $\frac{\bar X-\mu}{\sigma/\sqrt n}\to_d N(0,1)$ $(n>30)$
4. The Delta method: $Z=f(Y)\approx f(\mu)+(Y-\mu)f'(\mu)$
   $E(Z)\approx f(\mu)$, $var(Z)\approx\sigma_Y^2[f'(\mu)]^2$
   假设 $E(Y_i)=\mu_i$, $var(Y_i)\propto V(\mu_i)$ 求 $Z=f(Y)$
   $\Rightarrow f(\mu)\propto\int\frac{1}{\sqrt{V(\mu)}}d\mu$ $E[X]=E[E[X|Y]]$ $P(X)=E[P(X|Y)]$

§3 统计概念

$Var(X)=E[Var(X|Y)]+Var[E(X|Y)]$

1. p-value: $\hat p=\inf\{\alpha\in[0,1]:T\in R_\alpha\}\overset{H_0}{\sim}Unif(0,1)$
2. Fish Info.: $I(\theta)=E(\frac{\partial}{\partial\theta}\log f(x_1,\cdots,x_n,\theta))^2=-E\frac{\partial^2}{\partial\theta^2}\log f(\cdots)$
   $\sqrt{I(\theta)}(\hat\theta-\theta)\xrightarrow{d}N(0,1)$
3. Linear Regression: $X\in R^{n\times p}\Rightarrow\hat\beta=(X^TX)^{-1}X^TY\sim N(\beta,\sigma^2(X^TX)^{-1})$
4. Logistic Regression: $P(y_i=1)=\frac{1}{1+\exp(-X_i^T\beta)}$ (P(X)也是这个公式)
   $\Rightarrow l(\beta)=-\sum_{i=1}^n\log(1+\exp(-X_i^T\beta))-\sum_{i=1}^n(1-y_i)X_i^T\beta$

2. Potential Outcome Framework

1. risk difference: $RD=P(Y=1|Z=1)-P(Y=1|Z=0)$
   $=\frac{P_{11}}{P_{11}+P_{10}}-\frac{P_{01}}{P_{01}+P_{00}}=0$ 若 $Z\perp Y$

	Y=1	Y=0
Z=1	P_{11}	P_{10}
Z=0	P_{01}	P_{00}

2. Potential outcome framework: $Y_i=W_i\cdot Y_i(1)+(1-W_i)\cdot Y_i(0)=\begin{cases}Y_i(1) & 若\ W_i=1\\ Y_i(0) & 若\ W_i=0\end{cases}$
   ① no inference: unit i 的 outcomes 独立于其他 units 的 treatments W_j
   ② consistency: 没有 extra outcomes (即 treatment levels 有 well-defined)
   ③ Stable Unit Treatment Value Assumption (SUTVA): 满足上述两点 assumptions
3. 重要运算: $W_i^2=W_i$; $W_i(1-W_i)=0$; $(1-W_i)^2=1-W_i$
   ① $W_iY_i=W_i(W_iY_i(1)+(1-W_i)Y_i(0))=W_iY_i(1)$
   ② $(1-W_i)Y_i=(1-W_i)(W_iY_i(1)+(1-W_i)Y_i(0))=(1-W_i)Y_i(0)$
4. ATE: $\tau=\begin{cases}EY_i(1)-EY_i(0) & 若\ (Y_i(1),Y_i(0))\ 为\ i.i.d.\ random\ vector\\ \frac1n\sum_{i=1}^n Y_i(1)-\frac1n\sum_{i=1}^n Y_i(0) & 若\ Y_i(1),Y_i(0)\ 为\ fixed\ values\end{cases}$
5. Causal graph: ① 是 directed acyclic graph ② Nodes 表示 features
   ③ Edges 表示 causal relations e.g. $\begin{cases}X=U_X\\ Y=\frac{X}{3}+U_Y\\ Z=\frac{Y}{16}+U_Z\end{cases}$

$$\begin{tikzcd} U_X \arrow[r] & X \arrow[d] \\ U_Y \arrow[r] & Y \arrow[d] \\ U_Z \arrow[r] & Z \end{tikzcd}$$

3. CRE

§1 Completely randomized experiment (CRE)

1. CRE: ① 共 n 个 units ② fix n_0 和 n_1 满足 $n_0+n_1=n$
   则 CRE 有 $P(\mathbf W=\mathbf w)=\frac{1}{\binom{n}{n_1}}$ $\mathbf w=(w_1,\cdots,w_n)$ 满足 $w_i\in\{0,1\}$, $\sum w_i=n_1$
   注: $W_i\perp(Y_i(0),Y_i(1))$, $\forall i\in\{1,\cdots,n\}$, 但 $\mathbf W$ 间不独立
2. W_i 与 W_iW_j 的期望:
   $EW_i=P(W_i=1)=\frac{1}{\binom{n}{n_1}}\times\binom{n-1}{n_1-1}=\frac{n_1}{n}$
   $EW_iW_j=P(W_i=1,W_j=1)=\frac{1}{\binom{n}{n_1}}\times\binom{n-2}{n_1-2}=\frac{n_1(n_1-1)}{n(n-1)}$
3. ATE的canonical estimator: $\hat\tau=\frac{1}{n_1}\sum_{i=1}^n W_iY_i-\frac{1}{n_0}\sum_{i=1}^n(1-W_i)Y_i=\frac{1}{n_1}\sum_{W_i=1}Y_i-\frac{1}{n_0}\sum_{W_i=0}Y_i$
4. $\hat\tau$ 的 unbiasedness: $E\hat\tau=\tau$

证明: ① $Y_i(1),Y_i(0)$ 为 fixed
$E\hat\tau=E[\frac1{n_1}\sum_{i=1}^n W_iY_i(1)-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0)]=\frac{1}{n_1}\sum_{i=1}^n E[W_i]Y_i(1)-\cdots$
$=\frac1{n_1}\sum_{i=1}^n\frac{n_1}{n}Y_i(1)-\frac1{n_0}\sum_{i=1}^n(1-\frac{n_1}{n})Y_i(0)=\tau$
② $Y_i(1),Y_i(0)$ 为 random:
$E\hat\tau=E[\frac1{n_1}\sum_{i=1}^n W_iY_i(1)-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0)]=\frac1{n_1}\sum_{i=1}^n E[W_i]E[Y_i(1)]-\cdots$
$=\frac1{n_1}\sum_{i=1}^n\frac{n_1}{n}E[Y_i(1)]-\frac1{n_0}\sum_{i=1}^n(1-\frac{n_1}{n})E[Y_i(0)]=\tau$

§2 Fisher randomization test

1. hypothesis: 令 $Y_i(1),Y_i(0)$ 为 fixed values.
   $H_0:Y_i(1)=Y_i(0)$ v.s. $H_1:Y_i(1)=Y_i(0)+C$ for all units $i=1,\cdots,n$
2. Fisher randomization test 的 algorithm
   Input: ① observations (W_i,Y_i), $i=1,\cdots,n$ ② number of repetitions B
   Step 1: 根据 §1/3 求出 estimator $\hat\tau$
   For $b=1,2,\cdots,B$, 重复:
   Step 2: 根据 §1/1 重新抽取 $\mathbf W^*=(W_1^*,\cdots,W_n^*)$
   Step 3: 求出 new estimator: $\hat\tau_b^*=\frac1{n_1}\sum W_i^*Y_i-\frac1{n_0}\sum(1-W_i^*)Y_i$
   ① 若 $C>0$, 则 $\hat p=\frac1B\sum_{b=1}^B 1_{\hat\tau_b^*>\hat\tau}$ (也可以取≥)
   ② 若 $C<0$, 则 $\hat p=\frac1B\sum_{b=1}^B 1_{\hat\tau_b^*<\hat\tau}$ (也可以取≥)
3. H0下 p-value 服从均匀分布: Under H_0: $Y_i(0)=Y_i(1)$ $(=Y_i)$,
   $\hat\tau=\frac1{n_1}\sum_{i=1}^n W_iY_i-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i=\frac1{n_1}\sum_{i=1}^n W_iY_i(1)-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(1)$
   $\hat\tau_b^*=\frac1{n_1}\sum_{i=1}^n W_i^*Y_i-\frac1{n_0}\sum_{i=1}^n(1-W_i^*)Y_i=\frac1{n_1}\sum_{i=1}^n W_i^*Y_i(1)-\frac1{n_0}\sum_{i=1}^n(1-W_i^*)Y_i(1)$
   $\Rightarrow\hat\tau_b^*$ 和 $\hat\tau$ i.i.d. distributed $\Rightarrow\hat p\to P(\hat\tau_b^*\ge\hat\tau)=1-F_{\hat\tau^*}(\hat\tau)=1-F_{\hat\tau}(\hat\tau)\sim Unif(0,1)$
   注: under H_1: $Y_i(1)=Y_i(0)+C$, 有 $\hat\tau_b^*$ 和 $\hat\tau$ 不满足 i.i.d. distributed:
   $\hat\tau=\frac1{n_1}\sum_{i=1}^n W_i[Y_i(0)+C]-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0)=\frac1{n_1}\sum_{i=1}^n W_iY_i(0)-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0)+\frac1{n_1}\sum_{i=1}^n W_iC$
   $\hat\tau_b^*=\frac1{n_1}\sum_{i=1}^n W_i^*(Y_i(0)+W_iC)-\frac1{n_0}\sum_{i=1}^n(1-W_i^*)(Y_i(0)+W_iC)=\cdots+\frac1{n_1}\sum_{i=1}^n W_i^*W_iC+\frac1{n_0}\sum_{i=1}^n(1-W_i^*)W_iC$

§3 Neyman's theorem ($Y_i(0),Y_i(1)$ 为 fixed value)

1. $\hat\tau$ 的 expecation, variance, 与 variance 的 estimator:
   $\hat\tau=\frac1{n_1}\sum_{i=1}^n W_iY_i-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i$, 则:
   ① $E\hat\tau=\tau$ ② $Var(\hat\tau)=\frac{n_0\sigma_1^2}{(n-1)\times n_1}+\frac{n_1\sigma_0^2}{(n-1)\times n_0}+\frac{2\rho}{n-1}$
   其中, $S_1=\frac1n\sum_{i=1}^n Y_i^2(1)$, $S_0=\frac1n\sum_{i=1}^n Y_i^2(0)$, $\mu_1=\frac1n\sum_{i=1}^n Y_i(1)$, $\mu_0=\frac1n\sum_{i=1}^n Y_i(0)$
   $\sigma_1^2=S_1-\mu_1^2$, $\sigma_0^2=S_0-\mu_0^2$, $\rho=\frac1n\sum_{i=1}^n Y_i(1)Y_i(0)-\mu_1\mu_0$
   ③ 定义 $\hat S_1=\frac1{n_1}\sum_{i=1}^n W_i(Y_i-\frac1{n_1}\sum_{i=1}^n W_iY_i)^2$
   $\hat S_0=\frac1{n_0}\sum_{i=1}^n(1-W_i)(Y_i-\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i)^2$
   则 $V=\frac{\hat S_1}{n_1}+\frac{\hat S_0}{n_0}$ 是 conservative estimator for $Var(\hat\tau)$

证明: ② $Var(\hat\tau)=Var(\frac1{n_1}\sum_{i=1}^n W_iY_i(1))+Var(\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))$
$-2\cdot Cov(\frac1{n_1}\sum_{i=1}^n W_iY_i(1),\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))$

(1) $Var(\frac1{n_1}\sum_{i=1}^n W_iY_i(1))=\frac1{n_1^2}\sum_{i=1}^n\sum_{j=1}^n E[W_iW_j]Y_i(1)Y_j(1)-[\frac1n\sum_{i=1}^n Y_i(1)]^2$
$=\frac1{n_1^2}\sum_{i=1}^n Y_i(1)^2\times\frac{n_1}{n}+\frac1{n_1^2}\sum_{i\ne j}Y_i(1)Y_j(1)\times\frac{n_1(n_1-1)}{n(n-1)}-[\frac1n\sum_{i=1}^n Y_i(1)]^2$
$=\frac1{nn_1}\sum_{i=1}^n Y_i(1)^2+\frac{n_1-1}{n_1n(n-1)}[[\sum_{i=1}^n Y_i(1)]^2-\sum_{i=1}^n Y_i(1)^2]-\frac1{n^2}[\sum_{i=1}^n Y_i(1)]^2$
$=\frac{n_0}{n_1n(n-1)}\sum_{i=1}^n Y_i(1)^2-\frac{n_0}{n_1n^2(n-1)}[\sum_{i=1}^n Y_i(1)]^2$
$=\frac{n_0}{n_1(n-1)}[\frac1n\sum_{i=1}^n Y_i(1)^2-(\frac1n\sum_{i=1}^n Y_i(1))^2]=\frac{n_0}{n_1(n-1)}\sigma_1^2$

(2) $Var(\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))=\frac{n_1}{n_0(n-1)}\sigma_0^2$

(3) $Cov(\frac1{n_1}\sum_{i=1}^n W_iY_i(1),\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))=\frac1{n_1n_0}\sum_{i=1}^n\sum_{j=1}^n Y_i(1)Y_j(0)\cdot(E[W_i(1-W_j)]-E[W_i]E[1-W_j])$
$=\frac1{n_1n_0}\sum_{i=1}^n\sum_{j=1}^n Y_i(1)Y_j(0)(\frac{n_1}{n}-E[W_iW_j]-\frac{n_1n_0}{n^2})$
$=\frac1{n_1n_0}\sum_{i\ne j}Y_i(1)Y_j(0)(\frac{n_1}{n}-\frac{n_1(n_1-1)}{n(n-1)}-\frac{n_1n_0}{n^2})+\frac1{n_1n_0}\sum_{i=1}^n Y_i(1)Y_i(0)(\frac{n_1}{n}-\frac{n_1}{n}-\frac{n_1n_0}{n^2})$
$=\frac{1}{n^2(n-1)}(\sum_{i=1}^n\sum_{j=1}^n Y_i(1)Y_j(0)-\sum_{i=1}^n Y_i(1)Y_i(0))-\frac1{n^2}\sum_{i=1}^n Y_i(1)Y_i(0)$
$=\frac1{n-1}(\frac1n\sum_{i=1}^n Y_i(1)\times\frac1n\sum_{j=1}^n Y_j(0)-\frac1n\sum_{i=1}^n Y_i(1)Y_i(0))=-\frac{\rho}{n-1}$

2. $\hat\tau$ 的 asymptotic distribution
   ① $\frac{\hat\tau-\tau}{Var(\hat\tau)}\to N(0,1)$ ② $\hat S_1\xrightarrow{P}\sigma_1^2$ ③ $\hat S_0\xrightarrow{P}\sigma_0^2$

§4 Neyman's theorem ($Y_i(0),Y_i(1)$ 为 random vector)

1. $\hat\tau$ 的 expecation, variance, asymptotic distribution, 与 σ_1^2 & σ_0^2 的 estimator
   令 ① $EY_i(0)=\mu_0$ ② $EY_i(1)=\mu_1$ ③ $Var(Y_i(0))=\sigma_0^2$
   ④ $Var(Y_i(1))=\sigma_1^2$ ⑤ $Cov(Y_i(0),Y_i(1))=\rho_{XY}$ ⑥ $0<C_1<\frac{n_1}{n_0}<C_2$
   则 ① $E\hat\tau=\tau$ ② $Var(\hat\tau)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0}$ ③ $\frac{\hat\tau-\tau}{Var(\hat\tau)}\xrightarrow{d}N(0,1)$
   ④ σ_1^2,σ_0^2 的 unbiased estimator 为:
   $\hat\sigma_1^2=\frac1{n_1-1}\sum_{i=1}^n(W_iY_i)^2-\frac{n_1}{n_1-1}\hat\mu_1^2=\frac{1}{n_1-1}\sum_{i=1}^n W_i(Y_i-\hat\mu_1)^2$
   $\hat\sigma_0^2=\frac1{n_0-1}\sum_{i=1}^n((1-W_i)Y_i)^2-\frac{n_0}{n_0-1}\hat\mu_0^2=\frac1{n_0-1}\sum_{i=1}^n(1-W_i)(Y_i-\hat\mu_0)^2$
   其中, $\hat\mu_1=\frac1{n_1}\sum_{i=1}^n W_iY_i$ $\hat\mu_0=\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i$

证明: ② $Var(\hat\tau)=Var(\frac1{n_1}\sum_{i=1}^n W_iY_i(1))+Var(\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))$
$-2\cdot Cov(\frac1{n_1}\sum_{i=1}^n W_iY_i(1),\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))$

(1) $Var(\frac1{n_1}\sum_{i=1}^n W_iY_i(1))=\frac1{n_1^2}\sum_{i=1}^n\sum_{j=1}^n E[W_iW_j]E[Y_i(1)Y_j(1)]-[\frac1n\sum_{i=1}^n\mu_1]^2$
$=\frac1{n_1^2}\sum_{i=1}^n\frac{n_1}{n}\times(\mu_1^2+\sigma_1^2)+\frac1{n_1^2}\sum_{i\ne j}\frac{n_1(n_1-1)}{n(n-1)}\times\mu_1^2-\mu_1^2$
$=\frac1{n_1}(\mu_1^2+\sigma_1^2)+\frac{n_1-1}{n_1}\mu_1^2-\mu_1^2=\frac1{n_1}\sigma_1^2$

(2) $Var(\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))=\frac1{n_0}\sigma_0^2$

(3) $Cov(\frac1{n_1}\sum_{i=1}^n W_iY_i(1),\frac1{n_0}\sum_{i=1}^n(1-W_i)Y_i(0))=\frac1{n_1n_0}\sum_{i=1}^n\sum_{j=1}^n\{E[W_i(1-W_j)Y_i(1)Y_j(0)]-E[W_iY_i(1)]E[(1-W_j)Y_j(0)]\}$
$=\frac1{n_1n_0}\sum_{i=1}^n\mu_1\mu_0(E[W_i(1-W_j)]-E[W_i]E[1-W_j])-\frac1{n_1n_0}\cdot n\cdot\frac{n_1}{n}\mu_1\cdot\frac{n_0}{n}\mu_0$
$=\frac1{n_1n_0}\sum\mu_1\mu_0(\frac{n_1}{n}-\frac{n_1(n_1-1)}{n(n-1)}-\frac{n_1n_0}{n^2})-\frac1{n_1n_0}\cdot n\cdot\frac{n_1}{n}\mu_1\frac{n_0}{n}\mu_0$
$=\frac1n\mu_1\mu_0-\frac1n\mu_1\mu_0=0$

证明: ④ $E\sum_{i=1}^n(W_iY_i)^2=\sum_{i=1}^n E[W_i]\cdot E[Y_i(1)^2]=n_1(\sigma_1^2+\mu_1^2)$
$E\hat\mu_1^2=\frac1{n_1^2}\sum_{i=1}^n\sum_{j=1}^n E[W_iW_jY_i(1)Y_j(1)]=\frac1{n_1^2}\sum_{i=1}^n E[W_iY_i^2(1)]+\frac1{n_1^2}\sum_{i\ne j}E[W_iW_j]E[Y_i(1)]E[Y_j(1)]$
$=\frac1{n_1^2}\cdot n\times\frac{n_1}{n}(\mu_1^2+\sigma_1^2)+\frac{n(n-1)}{n_1^2}\times\frac{n_1(n_1-1)}{n(n-1)}\cdot\mu_1^2=\mu_1^2+\frac{\sigma_1^2}{n_1}$

2. Test $H_0:\tau=0$ v.s. $H_1:\tau\ne0$
   Input: ① observations (W_i,Y_i), $i=1,\cdots,n$ ② nominal type-I error α
   Step 1: 求出 estimator $\hat\tau,\hat\sigma_1^2,\hat\sigma_0^2$
   Step 2: 求出 $\hat T=|\hat\tau|/\sqrt{\frac{\hat\sigma_1^2}{n_1}+\frac{\hat\sigma_0^2}{n_0}}$ Step 3: Rej if $\hat T>z_{1-\alpha/2}$; $\hat p=2-2\Phi(\hat T)$

§5 Stratified randomized experiments (SRE)

1. Stratified randomized experiments (SRE)
   ① 共 n 个 units K 组, 每组有 $(n_1,\cdots,n_K)$ 个, $\sum n_i=n$
   ② fix $(n_1^{(0)},\cdots,n_K^{(0)})$ 与 $(n_1^{(1)},\cdots,n_K^{(1)})$
   则 ① SRE 在每组内进行 CRE, 即 group j 内, $P(\mathbf W_j=\mathbf w_j)=\frac{1}{\binom{n_j}{n_j^{(1)}}}$
   其中, $\mathbf W_j=(W_{j,1},W_{j,2},\cdots,W_{j,n_j})$, $W_{j,i}\in\{0,1\}$, $\sum W_{j,i}=n_j^{(1)}$
   ② 组间独立, 满足 $P(\mathbf W_1=\mathbf w_1,\cdots,\mathbf W_K=\mathbf w_K)=\prod_{j=1}^K\frac{1}{\binom{n_j}{n_j^{(1)}}}$
   ③ Propensity score 为: $P_j=\frac{n_j^{(1)}}{n_j}$
2. ATE 与 estimator
   ① group j 内部, $\tau_j=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}(1)-\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}(0)$
   $\hat\tau_j=\frac1{n_j^{(1)}}\sum_{i=1}^{n_j}W_{j,i}Y_{j,i}-\frac1{n_j^{(0)}}\sum_{i=1}^{n_j}W_{j,i}Y_{j,i}$
   ② 总体, $\tau=\frac1n\sum_{j=1}^K\sum_{i=1}^{n_j}Y_{j,i}(1)-\frac1n\sum_{j=1}^K\sum_{i=1}^{n_j}Y_{j,i}(0)=\sum_{j=1}^K\lambda_j\tau_j$
   $\hat\tau=\sum_{j=1}^K\lambda_j\hat\tau_j$ 其中 $\lambda_k=\frac{n_k}{n}$
3. Fisher setting下 $\hat\tau$ 的 expectation 和 variance
   令 $Y_{j,i}(0),Y_{j,i}(1)$ 为 fixed values.
   则 ① $E\hat\tau=\tau$ ② $Var(\hat\tau)=\sum_{j=1}^K\lambda_j^2(\frac{n_j^{(0)}\sigma_{j,1}^2}{(n_j-1)\times n_j^{(1)}}+\frac{n_j^{(1)}\sigma_{j,0}^2}{(n_j-1)\times n_j^{(0)}}+\frac{2\rho_j}{n_j-1})$
   其中, $S_{j,1}=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}^2(1)$, $S_{j,0}=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}^2(0)$, $\mu_{j,1}=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}(1)$, $\mu_{j,0}=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}(0)$
   $\sigma_{j,1}^2=S_{j,1}-\mu_{j,1}^2$, $\sigma_{j,0}^2=S_{j,0}-\mu_{j,0}^2$, $\rho_j=\frac1{n_j}\sum_{i=1}^{n_j}Y_{j,i}(1)Y_{j,i}(0)-\mu_{j,1}\mu_{j,0}$
4. Fisher setting下的 test
   Input: ① observations $(W_{j,i},Y_{j,i})$, $j=1,\cdots,K$, $i=1,\cdots,n_j$ ② number of repetitions B
   Step 1: 求出 estimator $\hat\tau$
   For $b=1,2,\cdots,B$, 重复: $\hat\tau_b^*=\sum_{j=1}^K\lambda_j\hat\tau_{b,j}^*$
   Step 2: 根据 §1/1 重新抽取 $\mathbf W^*=(\mathbf W_1^*,\cdots,\mathbf W_K^*)$
   Step 3: 求出 new estimator: $\hat\tau_{b,j}^*=\frac1{n_j^{(1)}}\sum_{i=1}^{n_j}W_{j,i}^*Y_{j,i}-\frac1{n_j^{(0)}}\sum_{i=1}^{n_j}W_{j,i}^*Y_{j,i}$
   ① 若 $C>0$, 则 $\hat p=\frac1B\sum_{b=1}^B 1_{\hat\tau_b^*>\hat\tau}$ (也可以取≥)
   ② 若 $C<0$, 则 $\hat p=\frac1B\sum_{b=1}^B 1_{\hat\tau_b^*<\hat\tau}$ (也可以取≥)
5. Random setting下的 $\hat\tau$ 的 expecation, variance, asymptotic distribution
   令 ① $EY_{j,i}(0)=\mu_j^{(0)}$ ② $EY_{j,i}(1)=\mu_j^{(1)}$
   ③ $Var(Y_{j,i}(0))=\sigma_{j,0}^2$ ④ $Var(Y_{j,i}(1))=\sigma_{j,1}^2$
   则 ① $E\hat\tau=\tau$ ② $Var(\hat\tau)=\sum_{j=1}^K\lambda_j^2(\frac{\sigma_{j,1}^2}{n_j^{(1)}}+\frac{\sigma_{j,0}^2}{n_j^{(0)}})$
   ③ $\sigma_{j,1}^2,\sigma_{j,0}^2$ 的 unbiased estimator 为:
   $\hat\sigma_{j,1}^2=\frac{1}{n_j^{(1)}-1}\sum_{i=1}^{n_j}(W_{j,i}Y_{j,i})^2-\frac{n_j^{(1)}}{n_j^{(1)}-1}\hat\mu_j^{(1)2}=\frac{1}{n_j^{(1)}-1}\sum_{i=1}^{n_j}W_{j,i}(Y_{j,i}-\hat\mu_j^{(1)})^2$
   $\hat\sigma_{j,0}^2=\frac{1}{n_j^{(0)}-1}\sum_{i=1}^{n_j}((1-W_{j,i})Y_{j,i})^2-\frac{n_j^{(0)}}{n_j^{(0)}-1}\hat\mu_j^{(0)2}=\frac{1}{n_j^{(0)}-1}\sum_{i=1}^{n_j}(1-W_{j,i})(Y_{j,i}-\hat\mu_j^{(0)})^2$
   其中, $\hat\mu_j^{(1)}=\frac{1}{n_j^{(1)}}\sum_{i=1}^{n_j}W_{j,i}Y_{j,i}$, $\hat\mu_j^{(0)}=\frac{1}{n_j^{(0)}}\sum_{i=1}^{n_j}(1-W_{j,i})Y_{j,i}$
   ④ $Var(\hat\tau)$ 的 unbiased estimator 为 $\hat S^2=\sum_{j=1}^K\lambda_j^2(\frac{\hat\sigma_{j,1}^2}{n_j^{(1)}}+\frac{\hat\sigma_{j,0}^2}{n_j^{(0)}})$
   ⑤ $\frac{\hat\tau-\tau}{\sqrt{\hat S^2}}\xrightarrow{d}N(0,1)$
6. Random setting下的 test: $\tau=0$ v.s. $H_1:\tau\ne0$
   Input: ① observations $(W_{j,i},Y_{j,i})$ ② nominal type-I error α
   Step 1: 求出 estimator $\hat\tau,\hat S^2$
   Step 2: 求出 $\hat T=|\hat\tau|/\hat S$ Step 3: Rej if $\hat T>z_{1-\alpha/2}$; $\hat p=2-2\Phi(\hat T)$

§6 CRE 的 regression methods

1. CRE 的 regression form 与 least square estimator:
   令 $\begin{cases}Y_i(0)=\mu_0+\varepsilon_i(0), & \mu_0:=EY_i(0)\\ Y_i(1)=\mu_1+\varepsilon_i(1)=\mu_0+\tau+\varepsilon_i(1), & \mu_1:=EY_i(1)\end{cases}$
   $\Rightarrow Y_i=\mu_0+W_i\tau+\eta_i$ 其中 $\eta_i:=W_i\varepsilon_i(1)+(1-W_i)\varepsilon_i(0)$
   $\Rightarrow\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}=\begin{bmatrix}1&W_1\\1&W_2\\\vdots&\vdots\\1&W_n\end{bmatrix}\begin{bmatrix}\mu_0\\\tau\end{bmatrix}+\begin{bmatrix}\eta_1\\\eta_2\\\vdots\\\eta_n\end{bmatrix}:=X\beta+\eta\Rightarrow\begin{bmatrix}\hat\mu_0\\\hat\tau\end{bmatrix}=(X^TX)^{-1}X^TY$
2. regression method 的 error $e_2=(0,1)^T$
   $\begin{bmatrix}\hat\mu_0\\\hat\tau\end{bmatrix}-\begin{bmatrix}\mu_0\\\tau\end{bmatrix}=(X^TX)^{-1}X^T\eta\Rightarrow\sqrt n(\hat\tau-\tau)=\frac1{\sqrt n}e_2^T(\frac{X^TX}{n})^{-1}\begin{bmatrix}\sum_{i=1}^n W_i\varepsilon_i(1)+(1-W_i)\varepsilon_i(0)\\\sum_{i=1}^n W_i\varepsilon_i(1)\end{bmatrix}$
   其中, $\frac{X^TX}{n}=\begin{bmatrix}1&\frac1n\sum_{i=1}^n W_i\\\frac1n\sum_{i=1}^n W_i&\frac1n\sum_{i=1}^n W_i^2\end{bmatrix}=\begin{bmatrix}1&n_1/n\\n_1/n&n_1/n\end{bmatrix}\Rightarrow(\frac{X^TX}{n})^{-1}=\frac{1}{p-p^2}\begin{bmatrix}p&-p\\-p&1\end{bmatrix}$, $p:=\frac{n_1}{n}$
   $\Rightarrow\sqrt n(\hat\tau-\tau)=\frac{1}{\sqrt n p(1-p)}\times((1-p)\sum_{i=1}^n W_i\varepsilon_i(1)-p\sum_{i=1}^n(1-W_i)\varepsilon_i(0))$
3. 若存在 covariates $X_i=(X_{i1},\cdots,X_{ip})$, $\sum_{i=1}^n X_{ij}=0$
   令 $\begin{cases}Y_i(0)=\mu_0+X_i^T\beta+\varepsilon_i(0),\\ Y_i(1)=\mu_0+X_i^T\beta+\tau+\varepsilon_i(1)\end{cases}$
   $\Rightarrow Y_i=\mu_0+X_i^T\beta+W_i\tau+\eta_i$ 其中 $\eta_i:=W_i\varepsilon_i(1)+(1-W_i)\varepsilon_i(0)$
   $\Rightarrow[\hat\mu_0\ \hat\tau\ \hat\beta^T]^T=(X^TX)^{-1}X^TY$

4. Observational study

§1 Identification 与 IPW estimator

1. Settings: ① 共 n 个 units $(X_i,Y_i(0),Y_i(1),W_i)$ $(Y_i(0),Y_i(1))\perp W_i$ 不一定
   ② Observation 为 (X_i,Y_i,W_i), 其中 $Y_i=W_iY_i(1)+(1-W_i)Y_i(0)$
2. 困难之处: $E[Y_i(1)]=P(W_i=1)\cdot E[Y_i(1)|W_i=1]+P(W_i=0)\cdot E[Y_i(1)|W_i=0]$
   $E[Y_i(0)]$ 同理 $=P(W_i=1)\cdot E[Y_i|W_i=1]+P(W_i=0)\cdot\underline{E[Y_i(1)|W_i=0]}$ 无法估
3. Identification: 即可被写作 function of the distribution of the observed data
   e.g. condition: $\begin{cases}E[Y_i(1)|W_i=1,X_i]=E[Y_i(1)|W_i=0,X_i]\\ E[Y_i(0)|W_i=1,X_i]=E[Y_i(0)|W_i=0,X_i]\end{cases}$
   $\Rightarrow\begin{cases}E[Y_i(1)|X_i]=\sum_{j=0}^1E[Y_i(1)|W_i=j,X_i]\cdot P(W_i=j)=E[Y_i|W_i=1,X_i]\\ E[Y_i(0)|X_i]=\sum_{j=0}^1E[Y_i(0)|W_i=j,X_i]\cdot P(W_i=j)=E[Y_i|W_i=0,X_i]\end{cases}$
   $\Rightarrow$ identification formula:
   $\tau=E[Y_i(1)]-E[Y_i(0)]=E[E[Y_i(1)|X_i]-E[Y_i(0)|X_i]]$
   $=\int\{E[Y_i|t,W_i=1]-E[Y_i|t,W_i=0]\}\cdot f_X(t)\,dt$
4. Ignorability / Unconfoundedness assumption: $\{Y_i(1),Y_i(0)\}\perp W_i|X_i$
5. propensity score: $p(X_i):=E[W_i|X_i]$
6. Unconfoundedness assumption 下的 τ 为 identifiable: identification formula:
   $\tau=E[\frac{W_iY_i}{p(X_i)}]-E[\frac{(1-W_i)Y_i}{1-p(X_i)}]$

证明: $E[\frac{W_iY_i}{p(X_i)}]=E[\frac{W_iY_i(1)}{p(X_i)}]=E\{E[\frac{W_iY_i(1)}{p(X_i)}|X_i]\}$
$=E\{\frac{1}{p(X_i)}E[W_iY_i(1)|X_i]\}=E\{\frac{1}{p(X_i)}E[W_i|X_i]E[Y_i(1)|X_i]\}=E[Y_i(1)]$

7. Inverse propensity score weighting (IPW) estimator:
   $\hat\tau=\frac1n\sum_{i=1}^n\frac{W_iY_i}{p(X_i)}-\frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i}{1-p(X_i)}$
8. IPW estimator 的 expecation, variance, 与 asymptotic distribution
   令 ① $\{Y_i(1),Y_i(0)\}\perp W_i|X_i$ ② $\eta<p(X)<1-\eta$ ③ $p(X_i)$ 已知
   则 ① $E\hat\tau=\tau$ ② $Var(\sqrt n(\hat\tau-\tau))=E[\frac{Y_i^2(1)}{p(X_i)}]+E[\frac{Y_i^2(0)}{1-p(X_i)}]-\tau^2$

③ $E[\frac{Y_i^2(1)}{p(X_i)}]$ 和 $E[\frac{Y_i^2(0)}{1-p(X_i)}]$ 的 unbiased estimator 为：

$$\hat S_1^2 = \frac1n\sum_{i=1}^n \frac{W_iY_i^2}{p(X_i)^2} \quad 和 \quad \hat S_0^2 = \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i^2}{[1-p(X_i)]^2}$$

④ $Var(\sqrt n(\hat\tau-\tau))$ 的 unbiased estimator 为

$$\hat S^2 = \hat S_1^2+\hat S_0^2-\hat\tau^2 = \frac1n\sum_{i=1}^n\frac{W_iY_i^2}{p(X_i)^2} + \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i^2}{[1-p(X_i)]^2} - \hat\tau^2$$

⑤ $\hat\tau \xrightarrow{P}\tau$, $\hat S^2 \xrightarrow{P} Var(\sqrt n(\hat\tau-\tau))$

⑥ $\frac{\sqrt n(\hat\tau-\tau)}{\sqrt{\hat S^2}} \xrightarrow{d} N(0,1)$

证明: ④ $Var(\hat\tau) = \frac1n Var(\frac{W_iY_i(1)}{p(X_i)}) + \frac1n Var(\frac{(1-W_i)Y_i(0)}{1-p(X_i)})$
$- 2\cdot Cov(\frac1n\sum_{i=1}^n\frac{W_iY_i(1)}{p(X_i)}, \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i(0)}{1-p(X_i)})$

(1) $Var(\frac{W_iY_i(1)}{p(X_i)}) = E\{E[\frac{W_iY_i(1)^2}{p(X_i)^2}|X_i]\} - (E[Y_i(1)])^2$
$= E\{\frac{1}{p(X_i)^2}\cdot E[W_i|X_i]\cdot E[Y_i(1)^2|X_i]\} - (E[Y_i(1)])^2 = E[\frac{Y_i(1)^2}{p(X_i)}] - (E[Y_i(1)])^2$

(2) $Var(\frac{(1-W_i)Y_i(0)}{1-p(X_i)}) = E[\frac{Y_i(0)^2}{1-p(X_i)}] - (E[Y_i(0)])^2$

(3) $Cov(\frac1n\sum_{i=1}^n\frac{W_iY_i(1)}{p(X_i)}, \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i(0)}{1-p(X_i)}) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n E[\frac{W_iY_i(1)}{p(X_i)}\cdot\frac{(1-W_j)Y_j(0)}{1-p(X_j)}] - E[Y_i(1)]E[Y_i(0)]$
$= \frac{1}{n^2}\sum_{i\neq j} E[Y_i(1)]E[Y_j(0)] - E[Y_i(1)]E[Y_i(0)] = -\frac1n E[Y_i(1)]E[Y_i(0)]$

证明: ③ $E[\hat S_1^2] = E[\frac{1}{p(X_i)^2}\cdot E[W_iY_i(1)^2|X_i]] = E[\frac{Y_i(1)^2}{p(X_i)}]$

证明: ⑤ 由于 $\frac{W_iY_i(1)}{p(X_i)}$ 为 i.i.d.，且 $E[\frac{W_iY_i^2}{p(X_i)^2}] \le \frac{1}{\eta^2}E[Y_i(1)^2] < \infty$

LLN ⇒ $\begin{cases}\frac1n\sum_{i=1}^n\frac{W_iY_i}{p(X_i)} \xrightarrow{P} E[Y_i(1)] \\ \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i}{1-p(X_i)}\xrightarrow{P} E[Y_i(0)]\end{cases}$ $\begin{cases}\frac1n\sum_{i=1}^n\frac{W_iY_i^2}{p(X_i)^2}\xrightarrow{P}E[\frac{Y_i^2(1)}{p(X_i)}] \\ \frac1n\sum_{i=1}^n\frac{(1-W_i)Y_i^2}{[1-p(X_i)]^2}\xrightarrow{P} E[\frac{Y_i^2(0)}{1-p(X_i)}]\end{cases}$

⇒ $\hat\tau\xrightarrow{P}\tau$, $\hat S^2 \xrightarrow{P} Var(\sqrt n(\hat\tau-\tau))$

9. Test $H_0:\tau=0$ vs. $H_1:\tau\neq0$

Input: ① observations (X_i, W_i, Y_i), $i=1,\dots,n$ ② propensity score $p(\cdot)$

Step 1: 求出 estimator $\hat\tau$, $\hat S^2$

Step 2: 求出 $\hat T = \sqrt n\,\hat\tau/\hat S$ Step 3: Rej if $\hat T > z_{1-\alpha/2}$; $\hat p = 2-2\Phi(\hat T)$

§2 Propensity score 的估计和相关定理

1. Propensity score 的估计 ($X_i\in\{x_1,x_2,\dots,x_K\}$)

$$p(x_k) = \frac{P(W_i=1,X_i=x_k)}{P(X_i=x_k)} \Rightarrow \hat p_k = \hat p(x_k) = \frac{\sum_{i=1}^n 1_{W_i=1,X_i=x_k}}{\sum_{i=1}^n 1_{X_i=x_k}}$$

2. Propensity score 的估计 (continuous covariates)

① Assume $W_i|X_i \overset{iid}{\sim} Bernoulli(p(X_i))$, $L=\prod_{i=1}^n p(X_i)^{W_i}(1-p(X_i))^{(1-W_i)}$

⇒ $l = \sum_{i=1}^n W_i\log(p(X_i)) + \sum_{i=1}^n(1-W_i)\log(1-p(X_i))$

② Assume $p(X_i) = \frac{\exp(-X_i^T\beta)}{1+\exp(-X_i^T\beta)} \iff \log(p(X_i)) = -X_i^T\beta - \log(1+\exp(-X_i^T\beta))$

⇒ $l(\beta) = -\sum_{i=1}^n W_iX_i^T\beta - \sum_{i=1}^n\log(1+\exp(-X_i^T\beta))$

⇒ $\frac{\partial l(\beta)}{\partial\beta_k} = \sum_{i=1}^n\frac{X_{ik}\exp(-X_i^T\beta)}{1+\exp(-X_i^T\beta)} - \sum_{i=1}^n X_{ik}W_i$ ③ $\hat p(X_i) = \frac{\exp(-X_i^T\hat\beta)}{1+\exp(-X_i^T\hat\beta)}$

3. $p(X_i)$ summarizes X_i: $W\perp(Y(1),Y(0))|X \Rightarrow W\perp(Y(1),Y(0))|p(X)$

证明: $P(W=1,(Y(1),Y(0))\in A\,|\,p(X))$
$=E_X[P(W=1,(Y(1),Y(0))\in A\,|\,p(X),X)\,|\,p(X)]$
$=E_X[P(W=1,(Y(1),Y(0))\in A\,|\,X)\,|\,p(X)]$
$=E_X[P(W=1|X)\,|\,p(X)]\cdot E_X[P((Y(1),Y(0))\in A\,|\,X)\,|\,p(X)]$
$=p(X)\cdot P((Y(1),Y(0))\in A\,|\,p(X))$

4. propensity score 的 balancing property ① $W\perp X\,|\,p(X)$

② 对 ∀ function $h(\cdot)$，有 $E[\frac{Wh(X)}{p(X)}] = E[\frac{(1-W)h(X)}{1-p(X)}]$

证明: ① $P(W=1,X\in A\,|\,p(X)) = E[P(W=1,X\in A\,|\,X)\,|\,p(X)] = p(X)\cdot P(X\in A\,|\,p(X))$

证明: ② $\Rightarrow E[\frac{Wh(X)}{p(X)}] = E[\frac{(1-W)h(X)}{1-p(X)}]$

$\begin{cases}E[\frac{Wh(X)}{p(X)}] = E[E[\frac{Wh(X)}{p(X)}|X]] = E[\frac{h(X)}{p(X)}\cdot E[W|X]] = E[h(X)] \\ E[\frac{(1-W)h(X)}{1-p(X)}] = E[E[\frac{(1-W)h(X)}{1-p(X)}|X]] = E[\frac{h(X)}{1-p(X)}\cdot E[1-W|X]] = E[h(X)]\end{cases}$

§3 Observational study 下计算 ATE 的 regression method

1. Observational study 下计算 ATE 的 model

$\begin{cases}Y_i(0)=X_i^T\beta_0+\varepsilon_i(0) \iff W_iY_i = (W_iX_i^T)\beta_0 + W_i\varepsilon_i(0) \\ Y_i(1)=X_i^T\beta_1+\varepsilon_i(1) \iff (1-W_i)Y_i = ((1-W_i)X_i^T)\beta_1 + (1-W_i)\varepsilon_i(1)\end{cases}$

⇒ $\tau = E[Y_i(1)]-E[Y_i(0)] = E[X_i]^T(\beta_1-\beta_0) = \mu_X^T(\beta_1-\beta_0)$

2. model-based estimation ① $\hat\mu_X = \frac1n\sum_{i=1}^n X_i$ ② $\hat\tau = \hat\mu_X^T(\hat\beta_1-\hat\beta_0)$

③ $\hat\beta_1 = (\sum_{i=1}^n W_iX_iX_i^T)^{-1}\sum_{i=1}^n W_iX_iY_i = \beta_1 + (\sum_{i=1}^n W_iX_iX_i^T)^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1)$

$\hat\beta_0 = (\sum_{i=1}^n(1-W_i)X_iX_i^T)^{-1}\sum_{i=1}^n(1-W_i)X_iY_i = \beta_0 + (\sum_{i=1}^n(1-W_i)X_iX_i^T)^{-1}\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0)$

证明: $\sum_{i=1}^n W_iX_iY_i = \sum_{i=1}^n W_iX_iY_i(1) = \sum_{i=1}^n W_iX_i(X_i^T\beta_1+\varepsilon_i(1)) = (\sum_{i=1}^n W_iX_iX_i^T)\beta_1 + \sum_{i=1}^n W_iX_i\varepsilon_i(1)$

⇒ $\hat\beta_1 = (\sum_{i=1}^n W_iX_iX_i^T)^{-1}\sum_{i=1}^n W_iX_iY_i = \beta_1 + (\sum_{i=1}^n W_iX_iX_i^T)^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1)$

3. $\hat\beta_1$ 和 $\hat\beta_0$ 的 unbiasedness

$E[(\sum_{i=1}^n W_iX_iX_i^T)^{-1}\sum_{i=1}^n W_iX_i\varepsilon_i(1)\,|\,X] = \sum_{i=1}^n E[(\sum_{i=1}^n W_iX_iX_i^T)^{-1}W_iX_i\,|\,X_i]\cdot E[\varepsilon_i(1)] = 0$

4. $\hat\beta_1,\hat\beta_0$ 和 $\hat\tau$ 的 consistency ① $\hat\beta_1\xrightarrow{P}\beta_1$ ② $\hat\beta_0\xrightarrow{P}\beta_0$ ③ $\hat\tau\xrightarrow{P}\tau$

§4 Doubly robust estimation

1. Doubly robust estimation

$\begin{cases}\hat\mu_1^\dagger = \hat\mu_X\hat\beta_1 + \frac1n\sum_{i=1}^n\frac{W_i}{\hat p(X_i)}(Y_i - X_i^T\hat\beta_1) \\ \hat\mu_0^\dagger = \hat\mu_X\hat\beta_0 + \frac1n\sum_{i=1}^n\frac{1-W_i}{1-\hat p(X_i)}(Y_i-X_i^T\hat\beta_0)\end{cases} \Rightarrow \hat\tau^\dagger = \hat\mu_1^\dagger - \hat\mu_0^\dagger$

注: ① 若 $\hat p\approx p$, $\frac1n\sum_{i=1}^n\frac{W_i}{\hat p(X_i)}(Y_i-X_i^T\hat\beta_1) \approx \frac1n\sum_{i=1}^n\frac{W_i}{p(X_i)}(Y_i-X_i^T\beta_1)$

$\approx E[E[\frac{W_i}{p(X_i)}(Y_i(1)-X_i^T\beta_1)|X]] = E[Y_i(1)] - \mu_X^T\beta_1$

⇒ $\hat\mu_1^\dagger \approx \mu_X\beta_1 + E[Y_i(1)] - \mu_X^T\beta_1 \approx E[Y_i(1)]$

② 若 model assumption 成立, $\frac1n\sum_{i=1}^n\frac{W_i}{\hat p(X_i)}(Y_i - X_i^T\hat\beta_1) \approx \frac1n\sum_{i=1}^n\frac{W_i}{\hat p(X_i)}(Y_i(1)-X_i^T\beta_1)$

$\approx \frac1n\sum_{i=1}^n\frac{W_i}{\hat p(X_i)}\varepsilon_i(1)\approx 0 \Rightarrow \hat\mu_1^\dagger \approx \mu_X\beta_1 + 0 \approx Y_i(1)$

2. oracle double-debiased estimator: $\hat\mu_0^* = \hat\mu_X\beta_0 + \frac1n\sum_{i=1}^n\frac{1-W_i}{1-p(X_i)}(Y_i - X_i^T\beta_0)$

$\hat\mu_1^* = \hat\mu_X\beta_1 + \frac1n\sum_{i=1}^n\frac{W_i}{p(X_i)}(Y_i-X_i^T\beta_1) \Rightarrow \hat\tau^* = \hat\mu_1^*-\hat\mu_0^*$

$$=\frac1n\sum_{i=1}^n\underbrace{\left\{(1-\frac{W_i}{p(X_i)})X_i^T\beta_1 - (1-\frac{1-W_i}{1-p(X_i)})X_i^T\beta_0 + \frac{W_iY_i(1)}{p(X_i)} - \frac{(1-W_i)Y_i(0)}{1-p(X_i)}\right\}}_{:=Z}$$

3. oracle double-debiased estimator 的 variance 的 estimator

$n\cdot Var(\hat\tau^*) = Var(Z) = E[Z^2] - E[Z]^2$

$$\Rightarrow \hat T^{*2} = \frac1n\sum_{i=1}^n\left\{(1-\frac{W_i}{p(X_i)})X_i^T\beta_1 - (1-\frac{1-W_i}{1-p(X_i)})X_i^T\beta_0 + \frac{W_iY_i(1)}{p(X_i)} - \frac{(1-W_i)Y_i(0)}{1-p(X_i)}\right\}^2 - \hat\tau^{*2}$$

4. 使用 doubly robust estimator 进行 test $H_0:\tau=0$ v.s. $H_1:\tau>0$

① 若 oracle double-debiased estimator 可以求出：

$$\hat L^* = \frac{\sqrt n\,\hat\tau^*}{\hat T^*} \quad 和 \quad \hat p^* = 1-\Phi(\hat L^*)$$

② 若 oracle double-debiased estimator 不可求出：

$$\hat T^{\dagger2} = \frac1n\sum_{i=1}^n\left\{(1-\frac{W_i}{\hat p(X_i)})X_i^T\hat\beta_1 - (1-\frac{1-W_i}{1-\hat p(X_i)})X_i^T\hat\beta_0 + \frac{W_iY_i}{\hat p(X_i)} - \frac{(1-W_i)Y_i}{1-\hat p(X_i)}\right\}^2 - \hat\tau^{\dagger2}$$

$$\hat L^\dagger = \frac{\sqrt n\,\hat\tau^\dagger}{\hat T^\dagger} \quad 和 \quad \hat p^\dagger = 1-\Phi(\hat L^\dagger)$$

§5. CATE 和 Meta learner

1. Conditional ATE

$\tau(X) := \mu_1(X)-\mu_0(X) := E[Y_i(1)\,|\,X_i=x] - E[Y_i(0)\,|\,X_i=x]$

2. T-learner 的 model

$\begin{cases}Y_i(0) = X_i^T\beta_0 + \varepsilon_i(0) \\ Y_i(1) = X_i^T\beta_1 + \varepsilon_i(1)\end{cases}$ 其中 $\varepsilon_i(0)\overset{i.i.d.}{\sim}N(0,\sigma_0^2)$, $\varepsilon_i(1)\overset{i.i.d.}{\sim}N(0,\sigma_1^2)$,

⇒ $\tau(x) = E[Y_i(1)|X_i=x] - E[Y_i(0)|X_i=x] = x^T(\beta_1-\beta_0)$

3. T-learner estimation

$\hat\beta_1 = (\sum_{i=1}^n W_iX_iX_i^T)^{-1}\cdot\sum_{i=1}^n W_iX_iY_i$, $\hat\beta_0 = (\sum_{i=1}^n(1-W_i)X_iX_i^T)^{-1}\cdot\sum_{i=1}^n(1-W_i)X_iY_i$

⇒ $\hat\tau(x) = x^T(\hat\beta_1-\hat\beta_0)$

4. $\hat\beta_1$ 和 $\hat\beta_0$ 的近似 令 ① $\Sigma_1 = E[W_iX_iX_i^T] = E[p(X_i)X_iX_i^T]$

② $\Sigma_0 = E[(1-W_i)X_iX_i^T] = E[(1-p(X_i))X_iX_i^T]$

则 ① $\hat\beta_1\approx\beta_1 + \frac1n\Sigma_1^{-1}\sum_{i=1}^n W_iX_i\varepsilon_i(1)$ ② $\hat\beta_0\approx\beta_0 + \frac1n\Sigma_0^{-1}\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0)$

5. $\hat\tau(x)$ 的 variance: $Var(\sqrt n(\hat\tau(x)-\tau(x))) = \sigma_1^2x^T\Sigma_1^{-1}x^T + \sigma_0^2x^T\Sigma_0^{-1}x$

证明: $\hat\tau(x)-\tau(x) = \frac1n x^T\Sigma_1^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1) - \frac1n x^T\Sigma_0^{-1}\cdot\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0)$

① $E[\frac1n x^T\Sigma_1^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1)] = E[\frac1n x^T\Sigma_1^{-1}\cdot\sum_{i=1}^n W_iX_i]\cdot\underbrace{E[\varepsilon_i(1)]}_{=0} = 0$

② $E[\frac1n x^T\Sigma_0^{-1}\cdot\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0)] = E[\frac1n x^T\Sigma_0^{-1}\cdot\sum_{i=1}^n(1-W_i)X_i]\cdot\underbrace{E[\varepsilon_i(0)]}_{=0} = 0$

③ $E\{[\frac1n x^T\Sigma_1^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1)]^2\} = \frac1n E\{(x^T\Sigma_1^{-1}X_i\cdot W_i)^2\cdot\varepsilon_i(1)^2\} = \frac{\sigma_1^2}{n}\cdot x^T\Sigma_1^{-1}x$

④ $E\{[\frac1n x^T\Sigma_0^{-1}\cdot\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0)]^2\} = \frac1n E\{(x^T\Sigma_0^{-1}X_i\cdot(1-W_i))^2\varepsilon_i(0)^2\} = \frac{\sigma_0^2}{n}\cdot x^T\Sigma_0^{-1}x$

⑤ $Cov(\frac1n x^T\Sigma_1^{-1}\cdot\sum_{i=1}^n W_iX_i\varepsilon_i(1), \frac1n x^T\Sigma_0^{-1}\cdot\sum_{i=1}^n(1-W_i)X_i\varepsilon_i(0))$
$= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n E[x^T\Sigma_1^{-1}W_iX_i\varepsilon_i(1)\cdot x^T\Sigma_0^{-1}(1-W_j)X_j\varepsilon_j(0)] = 0$

6. S-learner 的 model (assumption 和 T-learner 一样)

$Y_i = (W_iX_i^T)\beta_1 + W_i\varepsilon_i(1) + ((1-W_i)X_i^T)\beta_0 + (1-W_i)\varepsilon_i(0)$

$= \underbrace{X_i^T\beta_0 + (W_iX_i^T)(\beta_1-\beta_0)}_{:=H(X_i,W_i)=E[Y_i|X_i,W_i]} + \underbrace{W_i\varepsilon_i(1) + (1-W_i)\varepsilon_i(0)}_{:=\eta_i}$

⇒ $\tau(x) = x^T(\beta_1-\beta_0) = H(x,1) - H(x,0)$

7. S-learner estimation

$$\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix} = \begin{bmatrix}X_1^T & W_1X_1^T\\X_2^T & W_2X_2^T\\\vdots&\vdots\\X_n^T & W_nX_n^T\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1-\beta_0\end{bmatrix} + \begin{bmatrix}\eta_1\\\eta_2\\\vdots\\\eta_n\end{bmatrix} := Z\begin{bmatrix}\beta_0\\\beta_1-\beta_0\end{bmatrix} + \eta$$

$$\Rightarrow\begin{bmatrix}\hat\beta_0\\\hat\beta_1-\hat\beta_0\end{bmatrix} = (Z^TZ)^{-1}Z^TY \Rightarrow \hat\tau(x) = \hat H(x,1) - \hat H(x,0)$$

§6 Regression discontinuity design

1. Before-after (BA) design: $\hat\tau = \frac1N\sum_{i=1}^N(Y_{ic} - Y_{ic-1})$

2. RDD: W_i 取决于 X_i 是否 crossing c. 即 $W_i = 1(X_i\ge c)$ 或 $W_i = 1(X_i\le c)$

3. RD setting 下, unconfoundness condition (关于 X) 永远成立：

$P(W_i=1,(Y_i(0),Y_i(1))\in A\,|\,X_i=x) = 1_{x\ge c}\cdot P(Y_i(0),Y_i(1))\in A)\,|\,X_i=x)$

4. $\tau(x_0)$ 的 identification: 令 ① $EY(1)|X=x$ 在 $x\ge x_0$ 上连续

② $EY(0)|X=x$ 在 $x\le x_0$ 上连续, 若 ① $\gamma(x_0+\varepsilon) = EY(1)|X=x_0+\varepsilon$

② $\gamma(x_0-\varepsilon) = EY(0)|X=x_0-\varepsilon$

则 identification formula 为: $\gamma(x_0) = \lim_{\varepsilon\to0,\varepsilon>0}\gamma(x_0+\varepsilon) + \lim_{\varepsilon\to0,\varepsilon<0}\gamma(x_0-\varepsilon)$

5. T-learner model: $\hat\beta_0(x_0)$ $\hat\beta_1(x_0)$

$\begin{cases}Y_i(1) = f_1(X_i)+\varepsilon_i(1) = f_1(x_0) + f_1'(x_0)(X_i-x_0) + \varepsilon_i(1) + Ch^2,\ X_i\in[x_0,x_0+h) \\ Y_i(0) = f_0(X_i)+\varepsilon_i(0) = f_0(x_0) - f_0'(x_0)(X_i-x_0) + \varepsilon_i(0) + Ch^2,\ X_i\in(x_0-h,x_0]\end{cases}$

6. Local linear regression

① $K(\cdot)\ge0$ 满足 $\int_{[-1,1]}K(x)dx=1$, $\int_{[-1,1]}xK(x)dx=0$

② 令 $W_h = diag(K(\frac{|X_1-x_0|}{h}), K(\frac{|X_2-x_0|}{h}),\dots,K(\frac{|X_n-x_0|}{h}))$

则 $\begin{bmatrix}\hat\beta_0(x_0)\\\hat\beta_1(x_0)\end{bmatrix} = \arg\min\sum_{i=1}^n K(\frac{|X_i-x_0|}{h})\cdot(Y_i-\beta_0-\beta_1(X_i-x_0))^2$

$= \arg\min|W_h^{1/2}y - W_h^{1/2}X\beta|_2^2$

$= (X^TW_hX)^{-1}X^TW_hy$

7. Local linear regression estimator 的 bias 和 variance

令 $y = \begin{bmatrix}1 & X_1-x_0\\1 & X_2-x_0\\\vdots&\vdots\\1 & X_n-x_0\end{bmatrix}\begin{bmatrix}f(x_0)\\f'(x_0)\end{bmatrix} + b+\epsilon$, where $|b_i|\le Ch^2$.

则 $\hat\beta(x_0) = (X^TW_hX)^{-1}X^TW_h(X\begin{bmatrix}f(x_0)\\f'(x_0)\end{bmatrix} + b+\epsilon) = \begin{bmatrix}f(x_0)\\f'(x_0)\end{bmatrix}$

$+(X^TW_hX)^{-1}X^TW_hb + (X^TW_hX)^{-1}X^TW_h\epsilon.$

$|e_i^T(X^TW_hX)^{-1}X^TW_hb| \le |e_i^T(X^TW_hX)^{-1}X^TW_h^{1/2}|_2\times|W_h^{1/2}b|_2$

$\le\sqrt{\frac1n\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)b_i^2}\times\sqrt{e_i^T(\frac{X^TW_hX}{n})^{-1}e_i}.$

$\frac1n\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)\approx EK(\frac{X_1-x_0}{h}) = \int_{(x_0-h,x_0+h)}K(\frac{y-x_0}{h})f(y)dy\approx hf(x_0),$

$E(X_1-x_0)K\left(\frac{X_1-x_0}{h}\right) = \int_{x_0-h}^{x_0+h}(y-x_0)K\left(\frac{y-x_0}{h}\right)f(y)dy\approx h^2f(x_0)\int_{[-1,1]}zK(z)dz,$

$\frac1nX^TW_hX = \frac1n\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)\begin{bmatrix}1 & X_i-x_0\\X_i-x_0 & (X_i-x_0)^2\end{bmatrix}\approx\begin{bmatrix}EK(\frac{X_1-x_0}{h}) & E(X_1-x_0)\times K(\frac{X_1-x_0}{h})\\E(X_1-x_0)\times K(\frac{X_1-x_0}{h}) & E(X_1-x_0)^2K(\frac{X_1-x_0}{h})\end{bmatrix}$

$\approx hf(x_0)\begin{bmatrix}1 & 0\\0 & h^2\int_{[-1,1]}z^2K(z)dz\end{bmatrix}.$

$\Rightarrow e_1^T(\frac{X^TW_hX}{n})^{-1}e_1\approx\frac{1}{hf(x_0)}$

⇒ bias 为 $O(h^2)$

注: 若研究 derivative: $\frac{\sqrt{hf(x_0)}\times h^2}{\sqrt{h^3f(x_0)\int_{[-1,1]}z^2K(z)dz}} = O(h),$

$E(e_i^T(X^TW_hX)^{-1}X^TW_h\epsilon)^2|X = \frac{\sigma^2}{n}e_i^T(\frac{X^TW_hX}{n})^{-1}\frac{X^TW_h^2X}{n}(\frac{X^TW_hX}{n})^{-1}e_i.$

有 $\frac1nX^TW_h^2X = \frac1n\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)^2\begin{bmatrix}1 & X_i-x_0\\X_i-x_0 & (X_i-x_0)^2\end{bmatrix}$

$\approx\begin{bmatrix}EK(\frac{X_1-x_0}{h})^2 & E(X_1-x_0)K(\frac{X_1-x_0}{h})^2\\E(X_1-x_0)K(\frac{X_1-x_0}{h})^2 & E(X_1-x_0)^2K(\frac{X_1-x_0}{h})^2\end{bmatrix}$

$\approx hf(x_0)\begin{bmatrix}\int_{[-1,1]}K^2(z)dz & 0\\0 & h^2\int_{[-1,1]}z^2K^2(z)dz\end{bmatrix}$

covariance matrix

$\approx\frac{\sigma^2}{n}\begin{bmatrix}\frac{\int_{[-1,1]}K^2(z)dz}{hf(x_0)} & 0\\0 & \frac{\int_{[-1,1]}z^2K^2(z)dz}{h^3f(x_0)(\int_{[-1,1]}z^2K^2(z)dz)^2}\end{bmatrix}$

⇒ variance 为 $O(\frac{1}{nh})$

⇒ h 取 $O(n^{-1/5})$

8. 对 T-learner 使用 LLR

Input: Observations (X_i,y_i), the break point x_0, the bandwidth h.

1. Concatenate $Z_i = (1\ X_i)^T$.
2. Calculate

$\hat\beta_1(x_0) = (\sum_{i=1}^n Z_iZ_i^T\times K\left(\frac{X_i-x_0}{h}\right)\times1_{X_i\ge x_0})^{-1}\sum_{i=1}^n Z_iY_i\times K\left(\frac{X_i-x_0}{h}\right)\times1_{X_i\ge x_0}$

$\hat\beta_0(x_0) = (\sum_{i=1}^n Z_iZ_i^T\times K\left(\frac{X_i-x_0}{h}\right)\times1_{X_i<x_0})^{-1}\sum_{i=1}^n Z_iY_i\times K\left(\frac{X_i-x_0}{h}\right)\times1_{X_i<x_0}$

3. $\hat\tau(x_0) = e_1^T\hat\beta_{1,1}(x_0) - e_1^T\hat\beta_{0,1}(x_0)$.

9. S-learner model

令 $U_i = (X_i-x_0)_+ = \begin{cases}X_i-x_0 \text{ if } X_i\ge x_0,\\0 \text{ otherwise}\end{cases}$

$V_i = (X_i-x_0)_- = \begin{cases}X_i-x_0 \text{ if } X_i<x_0,\\0 \text{ otherwise}\end{cases}$

$Y_i = \begin{cases}f_1(X_i)+\epsilon_i(1) \text{ if } X_i\ge x_0,\\f_0(X_i)+\epsilon_i(0) \text{ if } X_i<x_0\end{cases}$

$= f_0(x_0) + W_i\tau(x_0) + f_1'(x_0)U_i + f_0'(x_0)V_i + b_i + W_i\epsilon_i(1) + (1-W_i)\epsilon_i(0).$

$\begin{bmatrix}\hat f_0(x_0)\\\hat\tau(x_0)\\\hat f_1'(x_0)\\\hat f_0'(x_0)\end{bmatrix} = \arg\min\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)\times(Y_i-Z_i^T\beta)^2$, where $Z_i = (1,W_i,U_i,V_i)^T$

and $\hat\tau(x_0) = e_2^T[\hat f_0(x_0)\ \hat\tau(x_0)\ \hat f_1'(x_0)\ \hat f_0'(x_0)]^T$

10. T-learner estimator 的 variance

Notice that

$Y_i\times1_{X_i\ge x_0} = Y_i(1)1_{X_i\ge x_0} = Z_i^T[f_1(x_0)\ f_1'(x_0)]^T\times1_{X_i\ge x_0} + b_i\times1_{X_i\ge x_0} + \epsilon_i(1)\times1_{X_i\ge x_0},$

error becomes

$\hat\beta_1(x_0) - [f_1(x_0)\ f_1'(x_0)]^T\approx(\frac1n\sum_{i=1}^n Z_iZ_i^T\times K\left(\frac{X_i-x_0}{h}\right)\times1_{X_i\ge x_0})^{-1}$
$\frac1n\sum_{i=1}^n Z_i\epsilon_i(1)K\left(\frac{X_i-x_0}{h}\right)1_{X_i\ge x_0}.$

Define the new "kernel"

$K_+(x) = K(x)1_{x\ge0}$ and $K_-(x) = K(x)1_{x<0},$

then the variance of the second term becomes

$\frac{\sigma_1^2}{n^2}\sum_{i=1}^n K_+\left(\frac{X_i-x_0}{h}\right)^2EZ_iZ_i^T\approx\frac{\sigma_1^2\times hg(x_0)}{n}\begin{bmatrix}\int_{[-1,1]}K_+(z)^2dz & h\int_{[-1,1]}zK_+(z)^2dz\\h\int_{[-1,1]}zK_+(z)^2dz & h^2h\int_{[-1,1]}z^2K_+(z)^2dz\end{bmatrix},$

where $g(\cdot)$ is the density of the random variable X_i. The first term satisfies

$\frac1n\sum_{i=1}^n Z_iZ_i^TK_+\left(\frac{X_i-x_0}{h}\right)\approx EZ_1Z_1^TK_+\left(\frac{X_i-x_0}{h}\right)\approx hg(x)\begin{bmatrix}\int_{[-1,1]}K_+(z)dz & h\int_{[-1,1]}zK_+(z)dz\\h\int_{[-1,1]}zK_+(z)dz & h^2\int_{[-1,1]}z^2K_+(z)dz\end{bmatrix}$

So the asymptotic variance of the estimator $\hat f_1(x_0)$ is

$$\frac{\sigma_1^2}{nh\times g(x_0)}e_1^TA^{-1}BA^{-1}e_1,$$

$A = \begin{bmatrix}\int_{[-1,1]}K_+(z)dz & h\int_{[-1,1]}zK_+(z)dz\\h\int_{[-1,1]}zK_+(z)dz & h^2\int_{[-1,1]}z^2K_+(z)dz\end{bmatrix}$

$B = \begin{bmatrix}\int_{[-1,1]}K_+(z)^2dz & h\int_{[-1,1]}zK_+(z)^2dz\\h\int_{[-1,1]}zK_+(z)^2dz & h^2h\int_{[-1,1]}z^2K_+(z)^2dz\end{bmatrix}$

and the asymptotic variance of the estimator $\hat f_0(x_0)$ is

$$\frac{\sigma_0^2}{nhg(x_0)}e_1^TC^{-1}DC^{-1}e_1,$$

$C = \begin{bmatrix}\int_{[-1,1]}K_-(z)dz & h\int_{[-1,1]}zK_-(z)dz\\h\int_{[-1,1]}zK_-(z)dz & h^2\int_{[-1,1]}z^2K_-(z)dz\end{bmatrix}$

$D = \begin{bmatrix}\int_{[-1,1]}K_-(z)^2dz & h\int_{[-1,1]}zK_-(z)^2dz\\h\int_{[-1,1]}zK_-(z)^2dz & h^2\int_{[-1,1]}z^2K_-(z)^2dz\end{bmatrix}$

$Cov(\hat\beta_1(x_0),\hat\beta_0(x_0))\approx Ee_1^T(\frac1n\sum_{i=1}^n Z_iZ_i^TK_-\left(\frac{X_i-x_0}{h}\right))^{-1}\frac1n\sum_{i=1}^n\sum_{j=1}^n Z_i\epsilon_i(0)K_-\left(\frac{X_i-x_0}{h}\right)$
$K_+(\frac{X_j-x_0}{h})\epsilon_j(1)Z_j^T(\frac1n\sum_{i=1}^n Z_iZ_i^TK_-\left(\frac{X_i-x_0}{h}\right))^{-1}e_1 = 0.$

Therefore, the asymptotic variance of the estimator $\hat\tau(x_0)$ is given by

$$\frac{\sigma_1^2}{nh\times g(x_0)}e_1^TA^{-1}BA^{-1}e_1 + \frac{\sigma_0^2}{nhg(x_0)}e_1^TC^{-1}DC^{-1}e_1.$$

11. T-learner 的 inference

$\hat g(x_0) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)$

Input: the bandwidth h for local linear regression
Observations (X_i,y_i), the break point x_0,
and the bandwidth h_1 for density estimation.

1. Calculate $\hat\beta_1(x_0)$ and $\hat\beta_0(x_0)$, and $\hat\tau(x_0)$ as in algorithm.
2. Estimate the variances $\hat\sigma_1^2$ and $\hat\sigma_0^2$.
3. Calculate the test statistics $\hat\sigma_1^2 = \sum_{i=1}^n 1_{X_i\ge x_0}(Y_i - \hat f_1(x_0) - \hat f_1'(x_0)X_i)^2/\sum_{i=1}^n 1_{X_i\ge0}$

$$\hat T = \frac{\sqrt{nh\times\hat g(x_0)}\times\hat\tau(x_0)}{\sqrt{\hat\sigma_1^2e_1^TA^{-1}BA^{-1}e_1 + \hat\sigma^2e_1^TC^{-1}DC^{-1}e_1}}.$$

4. Calculate the corresponding p-value based on the test.