因果推論 HW1

29236015 新庄紘己

October 21, 2023

1

デザインに基づいた統計的推測の利点

- 処置割り振りプロセスの情報を用いることで、実験の状況をそのまま反映し、分析する事ができる。
- 完全ランダム化実験など、データの抽出が独立にならない状況も、モデル化することができる。
- 母集団が国や州の場合、無作為抽出の仮定は満たされない。そのような有限標本においても、厳密な p 値の計算が可能な場合がある。

欠点

- 複雑な推定量になると、理論が難しい。
- 単純な帰無仮説出ない限り、必ずしも標本理論ベースよりも優れた統計手法を用いれるとは限らない。
- 研究が蓄積している段階であり、利用可能な状況が限られている。

2

2.1

まず、β の推定量を導出するために、Regression Anatomy を用いる。W_i に X_i を回帰した以下のモデル

$$W_i = \delta_0 + \delta_1 X_i + e_i \tag{1}$$

を考え、$(\hat{\delta_0}, \hat{\delta_1})^T$ を OLS 推定量とする。ここで、$\hat{W_i} = \hat{\delta_0} + \hat{\delta_1} X_i$ が成り立つ。$\tilde{W_i} = W_i - \hat{W_i}$ とおく。ここで、

$$\begin{aligned}\frac{Cov(Y_i, \tilde{W_i})}{Var(\tilde{W_i})} &= \frac{Cov(\alpha + \beta W_i + \gamma X_i + u_i, \tilde{W_i})}{Var(\tilde{W_i})} \\ &= \frac{\alpha Cov(1, \tilde{W_i})}{Var(\tilde{W_i})} + \frac{\beta Cov(W_i, \tilde{W_i})}{Var(\tilde{W_i})} + \frac{\gamma Cov(X_i, \tilde{W_i})}{Var(\tilde{W_i})} + \frac{Cov(u_i, \tilde{W_i})}{Var(\tilde{W_i})}\end{aligned}$$

が成り立つ。$\tilde{W_i}$ はモデル (1) の残差なので,$(1, X_i, u_i)^T$ と直交する。$\hat{W_i}$ は非確率変数なので、

$$\frac{Cov(Y_i, \tilde{W_i})}{Var(\tilde{W_i})} = \frac{\beta Cov(\tilde{W_i} + \hat{W_i}, \tilde{W_i})}{Var(\tilde{W_i})} = \frac{\beta Cov(\tilde{W_i}, \tilde{W_i})}{Var(\tilde{W_i})} = \beta$$

$\bar{\tilde{W}_i} = 0$ より、$\hat{\beta}$ の確率極限は

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{I} Y_i \tilde{W_i} - \bar{Y_i}\bar{\tilde{W}_i}}{\frac{1}{n}\sum_{i=1}^{I} {\tilde{W_i}}^2 - \left(\bar{\tilde{W}_i}\right)^2} \to_p \frac{\mathbb{E}\left[Y_i \tilde{W_i}\right]}{\mathbb{E}\left[{\tilde{W_i}}^2\right]} \tag{2}$$

である。

$\tau_i = Y_i(1) - Y_i(0)$ とする。ATE を $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$ と定義する。。$Y_i = W_iY_i(1) + (1 - W_i)Y_i(0) = Y_i(0) + W_i\tau_i$ という関係を用いると、(2) 式は次のように書き換えられる。

$$\frac{\mathbb{E}\left[Y_i\tilde{W}_i\right]}{\mathbb{E}\left[\tilde{W}_i^2\right]} = \frac{\mathbb{E}\left[Y_i\tilde{W}_i\right]}{\mathbb{E}\left[\tilde{W}_i^2\right]} = \frac{\mathbb{E}\left[Y_i(0)\tilde{W}_i\right]}{\mathbb{E}\left[\tilde{W}_i^2\right]} + \frac{\mathbb{E}\left[W_i\tilde{W}_i\tau_i\right]}{\mathbb{E}\left[\tilde{W}_i^2\right]}$$

ここで、右辺第 1 項の分子は、次のように書き換えられる。

$$\begin{aligned}
\mathbb{E}\left[Y_i(0)\tilde{W}_i\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_i(0)\tilde{W}_i \mid X_i\right]\right] && \because \text{繰り返し期待値の法則} \\
&= \mathbb{E}\left[\mathbb{E}[Y_i(0) \mid X_i]\mathbb{E}\left[\tilde{W}_i \mid X_i\right]\right] && \because \text{X を所与とすると、ポテンシャルアウトカムと処置割り当ては独立} \\
&= 0 && \because \text{モデル (1) の説明変数と残差は直行するので、} E\left[\tilde{W}_i \mid X_i\right] = 0
\end{aligned}$$

同様に右辺の第 2 項の分子は、

$$\begin{aligned}
\mathbb{E}\left[W_i\tilde{W}_i\tau_i\right] &= \mathbb{E}\left[\mathbb{E}[W_i\tilde{W}_i\tau_i \mid X_i]\right] && \because \text{繰り返し期待値の法則} \\
&= \mathbb{E}\left[\mathbb{E}[W_i\tilde{W}_i \mid X_i]\mathbb{E}[\tau_i \mid X_i]\right] && \because \text{X を所与とすると、ポテンシャルアウトカムと処置割り当ては独立} \\
&= \mathbb{E}[Var(W_i \mid X_i)\tau(X_i)]
\end{aligned}$$

最後の等号は、$\mathbb{E}[W_i\tilde{W}_i \mid X_i] = \mathbb{E}[W_i(W_i - \hat{W}_i) \mid X_i] = \mathbb{E}[W_i^2 - W_i\hat{W}_i) \mid X_i]$

$$\begin{aligned}
\mathbb{E}[Var(W_i \mid X_i)] &= \mathbb{E}\left[Var(\tilde{W}_i \mid X_i)\right] \\
&= \mathbb{E}\left[\left(\tilde{W}_i - \mathbb{E}[\tilde{W}_i \mid X_i]\right)^2\right] \because E\left[\tilde{W}_i \mid X_i\right] = 0 \\
&= \mathbb{E}\left[\tilde{W}_i^2\right]
\end{aligned}$$

3.

推定に用いたコードは、pdf の最後に示した。

(a)

$ATE = -0.09704$ である。t 値は -7.987 であり、燃費の悪い車への税金は、1% の有意水準で有意に、大気汚染を減少させるといえる。

(b)

$ATE = -0.09674$ である。また、t 値は -8.39 である。単回帰の場合と比較して、ATE の推定値は、変化しない。95% 信頼区間は、$[-0.11939, -0.07409]$ である。

(c)

$ATE = -0.09601$ である。

(d)

図より、重なりの条件は満たされている事がわかる。

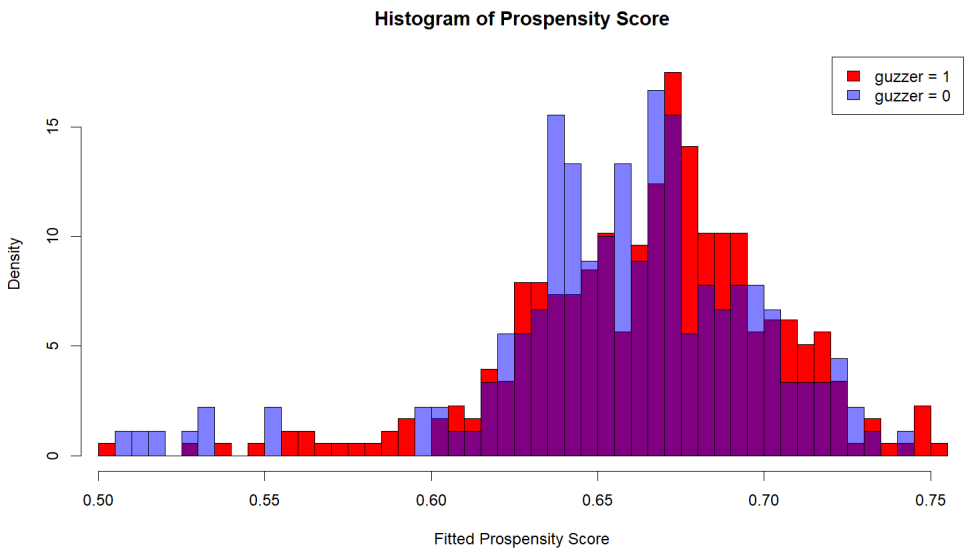

Figure1　Caption

(e)

$ATE = -0.09574$ である。

(f)

$ATE = -0.09590$ である。

```
#library
library(estimatr)
library(fastDummies)

options(scipen = 1)

#data
df <- read.csv("pollution.csv")
df <- dummy_cols(df,select_columns = c("guzzler","traffic"))
head(df)

#a
lm_a <- lm_robust(polution <- pollution ~ 1 + guzzler_tax,data = df)
summary(lm_a)

#b
lm_b <- lm_robust(pollution ~ 1 + guzzler_tax + industrial + rainfall + train + traffic_high +
      traffic_low + traffic_medium,data = df)
summary(lm_b)

#c
#data separation
```

```r
df_w1 <- df[df$guzzler_tax == 1, ]
df_w0 <- df[df$guzzler_tax == 0, ]

mu1 <- lm_robust(pollution ~ 1 + industrial + rainfall + train + traffic_high + traffic_low +
    traffic_medium,data = df_w1)
mu0 <- lm_robust(pollution ~ 1 + industrial + rainfall + train + traffic_high + traffic_low +
    traffic_medium,data = df_w0)

means <- colMeans(df[, c("industrial", "rainfall", "train", "traffic_high","traffic_low","traffic
    _medium")], na.rm = TRUE)

ATE_c <- mu1$coef[1]-mu0$coef[1] + (mu1$coef[2]-mu0$coef[2])*means[1] + (mu1$coef[3]-mu0$coef
    [3])*means[2] + (mu1$coef[4]-mu0$coef[4])*means[3] + (mu1$coef[5]-mu0$coef[5])*means[4] +
    (mu1$coef[6]-mu0$coef[6])*means[5] + (mu1$coef[7]-mu0$coef[7])*means[6]
ATE_c

#d
#probit estimation
probit_d <- glm(guzzler_tax ~ industrial + rainfall + train + traffic_high + traffic_low +
    traffic_medium,family = binomial(link = "probit"),data = df)
summary(probit_d)

#model fit with dummy = 1 and 0 respectively
df_w1 <- df[df$guzzler_tax == 1, ]
df_w0 <- df[df$guzzler_tax == 0, ]

#model fit
fit_w1 <- predict(probit_d,newdata = df_w1,type = "response")
fit_w0 <- predict(probit_d,newdata = df_w0,type = "response")

#Plot set up
par(mfrow=c(1,1))
hist(predict_w1, col="red", main="Histogram␣of␣Prospensity␣Score", xlab="Fitted␣Prospensity␣Score"
    , xlim=c(min(fit_w1, fit_w0), max(fit_w1, fit_w0)),freq = FALSE,breaks = 70)
hist(predict_w0, col=rgb(0,0,1,0.5), add=TRUE,freq = FALSE,breaks = 70)
legend("topright", legend=c("guzzer␣=␣1", "guzzer␣=␣0"), fill=c("red", rgb(0,0,1,0.5)))
par(mfrow=c(1,1))

#e
df$e_hat <- predict(probit_d,type = "response")
df$e_hat
df$one_minus_e_hat <- 1-predict(probit_d,type = "response")

df$numerator <- (df$guzzler_tax-df$e_hat)*df$pollution
df$denominator <- df$e_hat*df$one_minus_e_hat
df$ATE <- df$numerator/df$denominator

ATE_e <- sum(df$ATE)/nrow(df)
ATE_e

```

```
#f
df$m1 <- predict(probit_d,type = "response")
df$m0 <- predict(probit_d,type = "response")

df$left_side <- (df$guzzler_tax*df$pollution)/df$e_hat - (df$guzzler_tax-df$e_hat)*df$m1/df$e_hat
df$right_side <- ((1-df$guzzler_tax)*df$pollution)/df$one_minus_e_hat - (-df$guzzler_tax+df$e_hat
    )*df$m0/df$one_minus_e_hat

ATE_f <- (sum(df$left_side)-sum(df$right_side))/nrow(df)
ATE_f
```