제 4장: 비추정량과 회귀추정량: 보조변수를 이용한 추정량

4.1 개요와 정의

1. 모집단의 보조정보가 알려져 있는 경우. 연구변수와 상관이 높은 경우 보조변수(X)와 연구의 대상인 변수(Y)사이에 어떤 관계를 파악하여 이 관계를 Y의 모수 추정에 이용하는 방법.

$$Y \approx bX \qquad \text{or} \qquad Y \approx a + bX$$

2. 비추정법: 보조변수(X)와 연구변수(Y) 사이에 비례관계가 성립할 때 이 비 값을 이용하여 모집단의 평균과 총계를 추정하는 방법.

3. 회귀추정법: 보조변수(X)와 연구변수(Y) 사이에 선형관계가 성립할 때 이 선형관계를 이용하여 모집단의 평균과 총계를 추정하는 방법.

4. 비슷한 개념으로 사후 층화를 생각할 수 있다.

5. 예제:

   - 교재의 나무의 평균나이
   - 아파트단지의 평균 세대원 수
   - 소비자물가지수: 두 시점사이의 물건의 구매값들의 비. $X =$기준 시점의 물건의 값, $Y =$현재 물건의 값.
   - 병원의 의약품 소비: $X =$병원의 병상수, $Y =$병원에서 소비하는 의약품의 양.

4.2 비추정법

1. 모합: τ_x, τ_y;

모평균: μ_x, μ_y;

모비:

$$\beta = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}.$$

2. 단순임의비복원 추출을 가정한다.

   각 모집단위 마다 (x_i, y_i)를 동시에 관측

   크기 N인 모집단으로 부터 n개의 표본 $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$을 단순임의비복원 추출로 관측한다.

   표본합: t_x, t_y;

   표본평균: $\overline{Y}$, $\overline{X}$;

   표본비:

$$b = \frac{t_y}{t_x} = \frac{\overline{Y}}{\overline{X}}.$$

   따라서 τ_y와 μ_y에 대한 비 추정치는

$$\underline{\widehat{\tau}_y = b\tau_x}, \qquad \underline{\widehat{\mu}_y = b\mu_x}$$

3. <u>모집단 비 β에 대한 추정</u>

   - 추정량:

$$b = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \frac{\overline{Y}}{\overline{X}}.$$

   - 추정량의 분산:

$$\text{var}(b) = \frac{N-n}{n(N-1)} \left(\frac{1}{\mu_x^2}\right) \sigma_r^2 \approx \frac{N-n}{nN} \left(\frac{1}{\mu_x^2}\right) \sigma_r^2$$

     이고 여기서

$$\sigma_r^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \beta x_i)^2. \quad (\sum_{i=1}^N y_i = \beta \sum_{i=1}^N x_i).$$

- 분산 추정량:

$$\widehat{\mathrm{var}}(b) = \frac{N-n}{nN}\left(\frac{1}{\mu_x^2}\right)s_r^2$$

이고 여기서

$$s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - bX_i)^2.$$

4. <u>모총계 τ_y에 대한 추정</u>

- 추정량:

$$\widehat{\tau}_y = b\tau_x.$$

- 추정량의 분산:

$$\mathrm{var}\left(\widehat{\tau}_y\right) = \tau_x^2\ \mathrm{var}(b) \approx N^2\mu_x^2\frac{N-n}{nN}\left(\frac{1}{\mu_x^2}\right)\sigma_r^2 = \frac{N(N-n)}{n}\sigma_r^2.$$

- 분산 추정량:

$$\widehat{\mathrm{var}}(\widehat{\tau}_y) = \tau_x^2\ \widehat{\mathrm{var}}(b) = \frac{N(N-n)}{n}s_r^2$$

이다.

5. <u>모평균 μ_y에 대한 추정</u>

- 추정량:

$$\widehat{\mu}_y = b\mu_x.$$

- 추정량의 분산:

$$\mathrm{var}\left(\widehat{\mu}_y\right) = \mu_x^2\ \mathrm{var}(b) = \mu_x^2\ \frac{N-n}{nN}\left(\frac{1}{\mu_x^2}\right)\sigma_r^2 = \frac{N-n}{nN}\sigma_r^2.$$

- 분산 추정량:

$$\widehat{\mathrm{var}}(\widehat{\mu}_y) = \mu_x^2\ \widehat{\mathrm{var}}(b) = \frac{N-n}{nN}s_r^2$$

이다.

6. 가중선형회귀를 이용한 비추정량의 계산:

7. 예제 4.1

4.3 표본의 크기

표본의 크기는 오차한계(B)를 이용하여 결정한다. 즉,

$$Z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta})} = B$$

를 만족시키는 n을 계산한다.

- 모비 β에 대한 표본크기:

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} \qquad D = \left(\frac{B\mu_x}{Z_{\alpha/2}}\right)^2.$$

- 모합 τ_y에 대한 표본크기:

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} \qquad D = \left(\frac{B}{Z_{\alpha/2}N}\right)^2.$$

- 모합 μ_y에 대한 표본크기:

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} \qquad D = \left(\frac{B}{Z_{\alpha/2}}\right)^2.$$

- 예제 4.3

4.4 SRS와 비추정량의 효율성 비교

1. $r_i \equiv y_i - \beta x_i$를 정의 한다. 즉, $y_i = \beta x_i + r_i$ 이고 $\sum_{i=1}^{N} r_i = 0$.

2. 각 추정량의 분산을 살펴보면

$$\mathrm{var}(\widehat{\mu}_y^{\mathrm{srs}}) = \frac{\sigma_y^2}{n}\frac{N-n}{N-1} \approx \frac{\sigma_y^2}{n}\frac{N-n}{N}$$

   과

$$\mathrm{var}(\widehat{\mu}_y^{\mathrm{ratio}}) = \mathrm{var}\big(\mu_x b\big) = \frac{\sigma_r^2}{n}\frac{N-n}{N}$$

   이다.

3. 여기서

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \beta x_i + \beta x_i - \mu_y\big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{r_i^2 - 2r_i(x_i - \mu_x)\beta + \beta^2(x_i - \mu_x)^2\right\} \\
&= \sigma_r^2 + \beta^2\sigma_x^2 - 2\beta\sigma(r, x)
\end{aligned}$$

   이고

$$\sigma(r, x) = \frac{1}{N}\sum_{i=1}^{N} r_i(x_i - \mu_x)$$

   이다.

4.

$$\mathrm{RE} = \frac{\mathrm{var}(\widehat{\mu}_y^{\mathrm{srs}})}{\mathrm{var}(\widehat{\mu}_y^{\mathrm{ratio}})} = \frac{\sigma_y^2}{\sigma_r^2} = 1 + \beta^2\frac{\sigma_x^2}{\sigma_r^2} - 2\beta\frac{\sigma_x}{\sigma_r}\mathrm{corr}(r, x)$$

   이고

$$\mathrm{corr}(r, x) = \frac{\sigma(r, x)}{\sigma_r\sigma_x}.$$

5. 따라서 위의 결과로 부터

- r와 x의 correlation이 약할 수록/ 독립에 가까울 수록 (대체적으로 x와 y의 correlation이 강할 수록) ratio estimator의 SRS estimator에 대한 상대 효율성이 좋아진다.
- r_i와 x_i의 비상관성/독립성 은 회귀분석에서와 같이 $R_i = Y_i - bX_i$ 와 X_i의 잔차도를 통하여 확인할 수 있다.

층화표집에서의 비추정량으로는 다음의 두 가지 형태가 있다.

- 분리 비추정량 (separate ratio estimator)
- 병합 비추정량 (combined ratio estimator)

4.5 층화표집 - 분리비추정량

1. 모합 τ_y 에 대한 분리 비추정량:

   - 추정량:

$$b_h = \frac{t_{yh}}{t_{xh}} = \frac{\overline{Y}_h}{\overline{X}_h},$$

$$\widehat{\tau}_{yh} = b_h \tau_{xh},$$

$$\widehat{\tau}_{yRs} = \sum_{h=1}^{H} \widehat{\tau}_{yh} = \sum_{h=1}^{H} b_h \tau_{xh}.$$

   - 추정량 $\widehat{\tau}_{yRs}$ 의 분산:

$$\begin{aligned} \text{var}(\widehat{\tau}_{yRs}) &= \sum_{h=1}^{H} \tau_{xh}^2 \text{var}(b_h) \\ &\approx \sum_{h=1}^{H} \tau_{xh}^2 \frac{1}{\mu_{xh}^2} \frac{\sigma_{rh}^2}{n_h} \frac{N_h - n_h}{N_h - 1}. \end{aligned} \tag{1}$$

   - 분산 $\text{var}(\widehat{\tau}_{yRs})$ 의 추정량:

$$\widehat{\text{var}}(\widehat{\tau}_{yRs}) = \sum_{h=1}^{H} \tau_{xh}^2 \frac{1}{\mu_{xh}^2} \frac{s_{rh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

   이고

$$s_{rh}^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} \left(Y_{hj} - b_h X_{hj}\right)^2.$$

2. 모평균 μ_y 에 대한 분리 비추정량:

- 추정량:

$$\widehat{\mu}_{yRs} = \frac{1}{N}\widehat{\tau}_{yRs} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\frac{1}{N_h}\right) b_h\tau_{xh} = \sum_{h=1}^{H} W_h b_h \mu_{xh}$$

- 추정량 $\widehat{\mu}_{yRs}$ 의 분산:

$$\begin{aligned}
\mathrm{var}(\widehat{\mu}_{yRs}) &= \sum_{h=1}^{H} W_h^2 \mu_{xh}^2 \mathrm{var}(b_h) \\
&\approx \sum_{h=1}^{H} W_h^2 \mu_{xh}^2 \frac{1}{\mu_{xh}^2}\frac{\sigma_{rh}^2}{n_h}\frac{N_h - n_h}{N_h - 1} \\
&= \sum_{h=1}^{H} W_h^2 \frac{\sigma_{rh}^2}{n_h}\frac{N_h - n_h}{N_h - 1}
\end{aligned}$$

- 분산 $\mathrm{var}(\widehat{\mu}_{yRs})$ 의 추정량:

$$\widehat{\mathrm{var}}(\widehat{\mu}_{yRs}) = \sum_{h=1}^{H} W_h^2 \frac{s_{rh}^2}{n_h}\frac{N_h - n_h}{N_h}$$

  이고

$$s_{rh}^2 = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(Y_{hj} - b_h X_{hj}\right)^2.$$

4.6 층화표집 - 병합비추정량

$$\beta = \frac{\tau_y}{\tau_x}$$

를 이용

$$b_c = \frac{\widehat{\tau}_y}{\widehat{\tau}_x} = \frac{\overline{Y}_{\mathrm{st}}}{\overline{X}_{\mathrm{st}}}$$

의 추정량을 사용한다.

1. 모합 τ_y 에 대한 병합 비추정량:

   - 병합추정량:

$$\widehat{\tau}_{yRc} = b_c \tau_x.$$

- 추정량 $\widehat{\mu}_{yRc}$의 분산:

$$\operatorname{var}\left(\widehat{\tau}_{yRc}\right) = \tau_x^2 \widehat{\operatorname{var}}(b_c).$$

이제 $\widehat{\operatorname{var}}(b_c)$를 계산하여 보면

$$\operatorname{var}(b_c) = \operatorname{var}(b_c - \beta)$$

이고

$$\begin{aligned} b_c - \beta &= \frac{\overline{Y}_{\mathrm{st}}}{\overline{X}_{\mathrm{st}}} - \beta \\ &= \frac{1}{\overline{X}_{\mathrm{st}}}\left(\overline{Y}_{\mathrm{st}} - \beta \overline{X}_{\mathrm{st}}\right) \\ &\approx \frac{1}{\mu_x}\left(\overline{Y}_{\mathrm{st}} - \beta \overline{X}_{\mathrm{st}}\right) \\ &= \frac{1}{\mu_x}\left\{\sum_{h=1}^{H} \frac{N_h}{N}\left(\overline{Y}_h - \beta \overline{X}_h\right)\right\}. \end{aligned} \tag{2}$$

따라서

$$\begin{aligned} \operatorname{var}(b_c) &= \operatorname{var}(b_c - \beta) \\ &\approx \frac{1}{\mu_x^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{rh}^2}{n_h} \frac{N_h - n_h}{N_h - 1} \\ &\approx \frac{1}{\mu_x^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{rh}^2}{n_h} \frac{N_h - n_h}{N_h}. \end{aligned}$$

이고

$$\begin{aligned} \operatorname{var}\left(\widehat{\tau}_{yRc}\right) &= \tau_x^2 \operatorname{var}(b_c) \\ &\approx \tau_x^2 \cdot \frac{1}{\mu_x^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{rh}^2}{n_h} \frac{N_h - n_h}{N_h - 1}. \end{aligned} \tag{3}$$

- 분산 $\operatorname{var}(\widehat{\mu}_{yRc})$의 추정량:

$$\begin{aligned} \widehat{\operatorname{var}}\left(\widehat{\tau}_{yRc}\right) &= \tau_x^2 \widehat{\operatorname{var}}(b) \\ &= \tau_x^2 \cdot \frac{1}{\mu_x^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_{rh}^2}{n_h} \frac{N_h - n_h}{N_h}. \end{aligned} \tag{4}$$

2. 모평균 μ_y에 대한 병합 비추정량:

- 추정량:

$$\widehat{\mu}_{yRc} = b_c \mu_x.$$

- 추정량의 분산

$$\mathrm{var}(\widehat{\mu}_{yRc}) \approx \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{rh}^2}{\mu_h} \frac{N_h - n_h}{N_h - 1}.$$

- 분산의 추정량:

$$\widehat{\mathrm{var}}(\widehat{\mu}_{yRc}) \approx \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_{rh}^2}{\mu_h} \frac{N_h - n_h}{N_h}.$$

4.7 분리 비추정량과 병합 비 추정량의 비교

	분리비 추정법	병합비 추정법
층별모비	$--$	일정
x의 층별모합	알고있음	$--$
x의 전체모합	알고있음	알고있음
표본크기	커야함	$--$

4.7 회귀추정량

1. 모집단 $\{(x_i, y_i), i = 1, 2, \ldots, N\}$에서 n개의 표본 $\{(X_j, Y_j), j = 1, 2, \ldots, n\}$을 얻는다.

   변수 x_i와 y_i가 선형의 관계 $y = \alpha + \beta x$를 가지고 있음을 가정한다.

2. 회귀계수 α와 β에 대한 추정량은

$$\begin{aligned} \widehat{\beta} = b &= \frac{\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y})}{\sum_{j=1}^{n}(X_j - \overline{X})^2} = \frac{s_{XY}}{s_X^2} = \frac{R s_Y}{s_X}, \\ \widehat{\alpha} = a &= \overline{Y} - b\overline{X}. \end{aligned}$$

3. 따라서 y의 평균 μ_y의 회귀 추정량은

$$\widehat{\mu}_y^{\mathrm{reg}} = a + b \cdot \mu_x = \overline{Y} + b \cdot (\mu_x - \overline{X}).$$

4. 편의(bias):

$$\begin{aligned}
\widehat{\mu}_y^{\mathrm{reg}} - \mu_y &= \overline{Y} + b\big(\mu_x - \overline{X}\big) - \mu_y \\
&= \overline{Y} - \mu_y + (b - \beta)\big(\mu_x - \overline{X}\big) + \beta\big(\overline{X} - \mu_x\big).
\end{aligned}$$

따라서

$$\begin{aligned}
\mathrm{E}\big(\widehat{\mu}_y - \mu_y\big) &= \mathrm{E}\big[(b - \beta)\big(\mu_x - \overline{X}\big)\big] = -\mathrm{cov}\big(b, \overline{X}\big) \\
&= -\mathrm{E}\left[\frac{\sum_{i=1}^n R_i (X_i - \overline{X})(\overline{X} - \mu_x)}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \\
&\approx -\frac{\mathrm{E}\left[\sum_{i=1}^n R_i (X_i - \overline{X})(\overline{X} - \mu_x)\right]}{n\sigma_x^2}
\end{aligned}$$

가 된다. 단, $R_i = Y_i - \alpha - \beta X_i$.

5. 추정량의 분산:

$$\begin{aligned}
\begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \\
&= \Big(\sum_{i \in A} \tilde{x}_i \tilde{x}_i^\top\Big)^{-1} \Big(\sum_{i \in A} \tilde{x}_i y_i\Big).
\end{aligned}$$

$\tilde{X}_i = (1, X_i)^\top$로 정의되고 A는 표본의 index set.

이제 μ_y 의 회귀추정량은

$$
\begin{aligned}
\widehat{\mu}_y^{\text{reg}} &= (1, \mu_x)\Big(\sum_{i\in A}\tilde{x}_i\tilde{x}_i^\top\Big)^{-1}\Big(\sum_{i\in A}\tilde{x}_i y_i\Big) \\
&= \mu_y + (1, \mu_x)\left\{\Big(\sum_{i\in A}\tilde{x}_i\tilde{x}_i^\top\Big)^{-1}\sum_{i\in A}\tilde{x}_i y_i - \begin{pmatrix}\alpha \\ \beta\end{pmatrix}\right\} \\
&= \mu_y + (1, \mu_x)\left\{\Big(\sum_{i\in A}\tilde{x}_i\tilde{x}_i^\top\Big)^{-1}\sum_{i\in A}\tilde{x}_i\big(y_i - \alpha - \beta x_i\big)\right\} \\
&= \mu_y + (1, \mu_x)\Big(\sum_{i\in A}\tilde{x}_i\tilde{x}_i^\top\Big)^{-1}\sum_{i\in A}\tilde{x}_i\big(y_i - \alpha - \beta x_i\big) \\
&\approx \mu_y + (1, \mu_x)\begin{pmatrix} n & n\mu_x \\ n\mu_x & n(\mu_x^2 + \sigma_x^2)\end{pmatrix}^{-1}\sum_{i\in A}\tilde{x}_i\big(y_i - \alpha - \beta x_i\big) \\
&= \mu_y + (1, \mu_x)\begin{pmatrix} n & n\mu_x \\ n\mu_x & n(\mu_x^2 + \sigma_x^2)\end{pmatrix}^{-1}\begin{pmatrix}\sum_{i\in A}(y_i - \alpha - \beta x_i) \\ \sum_{i\in A} x_i(y_i - \alpha - \beta x_i)\end{pmatrix} \\
&= \mu_y + \frac{1}{n}\sum_{i\in A}(y_i - \alpha - \beta x_i). \qquad (5)
\end{aligned}
$$

가 된다. 따라서

$$
\text{var}\big(\widehat{\mu}^{\text{reg}}\big) \approx \frac{\sigma_e^2}{n}\frac{N-n}{N-1}, \quad \sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \alpha - \beta x_i\big)^2
$$

을 얻을 수 있다.

4.8 차이검정