4 비추정과 회귀추정

4.1비추정(ratio estimation)

예. 총 쌀생산량(τ_y)

$\hat{\tau}_y = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\tau_x = b\tau_x$(평수:$x_i$, 생산량 : y_i, 총평수 : τ_x, 평당생산량 :b)

(y_i, x_i)를 동시에 측정

모합 : $\tau_y = \sum_{i=1}^{N} y_i, \tau_x = \sum_{i=1}^{N} x_i$ ⇨ 표본합 : $t_y = \sum_{i=1}^{n} y_i$, $t_x = \sum_{i=1}^{n} x_i$

모평균 : $\mu_y = \tau_y/N$, $\mu_x = \tau_x/N$ ⇨ 표본평균 : $\bar{y} = t_y/n$, $\bar{x} = t_x/n$

모비 : $\beta = \tau_y/\tau_x = \mu_y/\mu_x$ ⇨ 표본비 : $b = t_y/t_x = \bar{y}/\bar{x}$

모상관계수 : $\rho = \dfrac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)/N}{\sigma_x\sigma_y} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$

⇨ 표본상관계수 : $\gamma = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/(n-1)}{s_y s_x} = \dfrac{s_{xy}}{s_x s_y}$

➡ $\hat{\tau}_y = b\tau_x$, $\hat{\mu}_y = b\mu_x$

- b의 기대값과 분산(n이 클 때)

$E(b) \approx \beta$, $Var(b) \approx \dfrac{1}{\mu_x^2}\dfrac{\sigma_\epsilon^2}{n}\dfrac{N-n}{N}$, $\sigma_\epsilon^2 = \dfrac{\sum_{i=1}^{N}\epsilon_i^2}{N} = \dfrac{\sum_{i=1}^{N}(y_i-\beta x_i)^2}{N}$

$\widehat{Var}(b) \approx \dfrac{1}{\mu_x^2}\dfrac{s_e^2}{n}\dfrac{N-n}{N}(\approx \dfrac{1}{\bar{x}^2}\dfrac{s_e^2}{n}\dfrac{N-n}{N})$, $s_e^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-1} = \dfrac{\sum_{i=1}^{n}(y_i-bx_i)^2}{n-1}(= s_y^2 + b^2 s_x^2 - 2bs_{xy})$

95% 신뢰구간 : $b \pm 2\widehat{se}(b)$

- 모합 τ_y에 대한 추정

$\hat{\tau}_y = b\tau_x$, $\widehat{Var}(\hat{\tau}_y) = \tau_x^2\widehat{Var}(b) \approx \tau_x^2\dfrac{1}{\mu_x^2}\dfrac{s_e^2}{n}\dfrac{N-n}{N} = N^2\dfrac{s_e^2}{n}\dfrac{N-n}{N}(\tau_x = N\mu_x)$

95% 신뢰구간 : $\hat{\tau}_y \pm 2\widehat{se}(\hat{\tau}_y)$

- 모평균 μ_y에 대한 추정

$\hat{\mu}_y = b\mu_x$, $\widehat{Var}(\hat{\mu}_y) = \mu_x^2\widehat{Var}(b) \approx \dfrac{s_e^2}{n}\dfrac{N-n}{N}$

95% 신뢰구간 : $\hat{\mu}_y \pm 2\widehat{se}(\hat{\mu}_y)$

- 비율 p에 대한 추정

$\hat{p} = \bar{y}/\bar{x}$

예 4.1 전나무들의 평균 나이를 추정(나무의 나이는 나무의 지름에 비례)

수목원에 있는 나무 N=1132 전나무 모두에 대해서 지름을 측정하였고, 지름의 모평균 $\mu_x = 10.3$. n=20 그루의 나무를 임의표집.

나무20	나이	지름	전차$e_i = y_i - bx_i$
합	$t_y = 2148$	$t_x = 188.1$	
평균	$\bar{y} = 107.4$	$\bar{x} = 9.405$	
표준편차	$s_y = 28.67$	$s_x = 1.829$	$s_e = 17.94$
비	$b = 11.42$		

$\hat{\mu}_y = b\mu_x = 11.42(10.3) = 117.6$

$\widehat{Var}(\hat{\mu}_y) \approx \frac{s_e^2}{n}\frac{N-n}{N} = 15.81$

예 4.2 1998년도 대비 2000년의 소나무 생존률을 95% 신뢰구간 추정

지역 10	00모목수	98묘목수	전차$e_i = y_i - bx_i$
합	$t_y = 10753$	$t_x = 15720$	
평균	$\bar{y} = 1075.3$	$\bar{x} = 1572$	
표준편차	$s_y = 987.07$	$s_x = 1452.3$	$s_e = 45.038$
비	$b = 0.684$		

$\widehat{Var}(b) \approx \frac{1}{\bar{x}^2}\frac{s_e^2}{n}\frac{N-n}{N} = 0.0041$(유한모집단수정을 무시)

4.2 표본크기

$z_{\alpha/2}\sqrt{Var(\hat{\theta})} = B$:오차한계

$$z_{\alpha/2}\sqrt{\frac{1}{\mu_x^2}\frac{\sigma_\epsilon^2}{n}\frac{N-n}{N}} = B \Rightarrow n = \frac{N\sigma_\epsilon^2}{NB^2\mu_x^2/z_{\alpha/2}^2+\sigma_\epsilon^2} = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2}, D = \left(\frac{B\mu_x}{z_{\alpha/2}}\right)^2$$

- μ_x와 σ_ϵ^2의 추정량

$$\hat{\mu}_x = \frac{\sum_{i=1}^{n'} x_i}{n'},\ \hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n'}(y_i - bx_i)^2}{n'-1}$$

- 모합 τ_y에 대한 추정

$$n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2}, D = \left(\frac{B}{z_{\alpha/2}N}\right)^2$$

- 모평균 μ_y에 대한 추정

$$n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2}, D = \left(\frac{B}{z_{\alpha/2}}\right)^2$$

예 4.3 1,000 채의 아파트를 보유한 아파트 단지 관리인이 현재 아파트 1채당 평균 거주인이 몇 명인지 알고 싶어 표본조사를 계획.(오차한계 $B=\pm 0.5$)

1) 특이 값 존재

아파트10	올해	작년	전차$e_i = y_i - bx_i$
합	$t_y = 43$	$t_x = 42$	
평균	$\overline{y} = 4.3$	$\overline{x} = 4.2$	
표준편차	$s_y = 1.64$	$s_x = 1.55$	$s_e = 1.612$
비	$b = 1.02 (r = 0.50)$		

$$n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2} = 39.9, D = \left(\frac{B}{z_{\alpha/2}}\right)^2 = 1/16$$

2) 특이 값 제거

아파트9	올해	작년	전차$e_i = y_i - bx_i$
합	$t_y = 36$	$t_x = 39$	
평균	$\overline{y} = 4.0$	$\overline{x} = 4.33$	
표준편차	$s_y = 1.41$	$s_x = 1.58$	$s_e = 0.818$
비	$b = 0.923 (r = 0.84)$		

$$n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2} = 10.6, D = \left(\frac{B}{z_{\alpha/2}}\right)^2 = 1/16$$

4.3 층화표집에서 비추정

분리비추정(separate ratio estimation) : 각 층별로 독립적으로 비추정을 하여 층합을 추정한 후 그 층별결과를 합하는 방법

병합비추정(combined ratio estimation, 결합비추정)) : 모든 층을 통틀어 단 하나의 비를 추정해서 모합을 추정하는 방법

‣어느 방법이든 층마다 표본크기가 커야 쓸모가 있다.

1) 분리비추정

■ 모합 τ_y에 대한 추정

$$b_h = \frac{t_{yh}}{t_{xh}} = \frac{\bar{y}_h}{\bar{x}_h},\ \hat{\tau}_{yh} = b_h \tau_{xh},\ \hat{\tau}_{yRs} = \sum_{h=1}^{H} \hat{\tau}_{yh} = \sum_{h=1}^{H} b_h \tau_{xh}$$

$$\widehat{Var}(\hat{\tau}_{yRs}) = \widehat{Var}\left(\sum_{h=1}^{H} b_h \tau_{xh}\right) \qquad s_{eh}^2 = \frac{\sum_{i=1}^{n_h} e_{hi}^2}{n_h - 1} = \frac{\sum_{i=1}^{n_h} (y_{hi} - b_h x_{hi})^2}{n_h - 1}$$

$$= \sum_{h=1}^{H} \tau_{xh}^2 \widehat{Var}(b_h)$$

$$\approx \sum_{h=1}^{H} \tau_{xh}^2 \frac{1}{\mu_{xh}^2} \frac{s_{eh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

■ 모평균 μ_y에 대한 추정

$$\hat{\mu}_{yRs} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) b_h \mu_{xh},\ \widehat{Var}(\hat{\mu}_{yRs}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_{eh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

2) 병합비추정

$$\hat{\tau}_y = \sum_{h=1}^{H} N_h \bar{y}_h,\ \hat{\tau}_x = \sum_{h=1}^{H} N_h \bar{x}_h$$

$$\hat{\mu}_y = \bar{y}_{st} = \hat{\tau}_y / N,\ \hat{\mu}_x = \bar{x}_{st} = \hat{\tau}_x / N$$

■ 모합 τ_y에 대한 추정

$$b_c = \frac{\hat{\tau}_y}{\hat{\tau}_x} = \frac{\bar{y}_{st}}{\bar{x}_{st}},\quad \hat{\tau}_{yRc} = b_c \tau_x = \sum_{h=1}^{H} b_c \tau_{xh}$$

$$\widehat{Var}(\widehat{\tau_{yRc}}) = \widehat{Var}(b_c \tau_x) = \sum_{h=1}^{H} \tau_{xh}^2 \widehat{Var}(b_c),\ s_{eh}^2 = \frac{\sum_{i=1}^{n_h} e_{hi}^2}{n_h - 1} = \frac{\sum_{i=1}^{n_h} (y_{hi} - b_c x_{hi})^2}{n_h - 1}$$

$$\approx \sum_{h=1}^{H} \tau_{xh}^2 \frac{1}{\mu_{xh}^2} \frac{s_{eh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

$$= \sum_{h=1}^{H} N_h^2 \frac{s_{eh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

■ 모평균 μ_y에 대한 추정

$$\hat{\mu}_{yRc} = b_c \mu_x,\ \widehat{Var}(\hat{\mu}_{yRc}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_{eh}^2}{n_h} \frac{N_h - n_h}{N_h}$$

※ 층별 표본크기가 큰 경우에는 분리비추정이 유리하고, 작은 경우는 병합비추정이 유리하다.

예4.4 농가별 벼를 재배한 평균면적 추정

y : 농가별 벼를 재배한 경지의 면적

x : 농가별 총 경작 면적

층 경지면적	N_h	$\overline{y}_h$	$\overline{x}_h$	σ_{yh}^2	σ_{xyh}	σ_{xh}^2	β_h	n_h
1 ≦ 160	1580	19.40	82.56	312	494	2055	0.2350	70
2 > 160	430	51.63	244.85	922	858	7357	0.2109	30
모집단	2010	26.30	117.28	620	1453	7619	0.2242	100

1) 단순임의표집

$$\nu_1 = \frac{\sigma_y^2}{n} = 6.20$$

2) 단순임의표집에서 비추정

$$\nu_2 = \frac{\sigma_\epsilon^2}{n} = 3.51$$

3) 층화임의표집

$$\nu_3 = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{yh}^2}{n_h} = 4.16$$

4) 층화임의표집에서 분리비 추정

$$\nu_4 = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{\epsilon h}^2}{n_h} = 3.06$$

5) 층화임의표집에서 병합비 추정

$$\nu_5 = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_{\epsilon c}^2}{n_h} = 3.10$$

4.4 회귀추정

$$y=\alpha+\beta x+\epsilon \Rightarrow \hat{y}=a+bx,\ \hat{\beta}=b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{s_{xy}}{s_x^2}=\frac{\gamma s_y}{s_x},\ \hat{\alpha}=a=\bar{y}-b\bar{x}$$

- 모평균 μ_y에 대한 추정

$$\hat{\mu}_y=a+b\mu_x=\bar{y}+b(\mu_x-\bar{x})\ (\triangledown\ \text{항상}\ (\bar{x},\bar{y})\text{를 통과})$$

$$E(\hat{\mu}_y-\mu_y)=-cov(\bar{x},b),\ \widehat{Var}(\hat{\mu}_y)\approx\frac{s_e^2}{n}\frac{N-n}{N}\approx\frac{MSE}{n}\frac{N-n}{N}$$

$$s_e^2\approx\frac{\sum_{i=1}^{n}e_i^2}{n-1}=\frac{\sum_{i=1}^{n}[(y_i-\bar{y})-b(x_i-\bar{x})]^2}{n-1}$$

- 모합 τ_y에 대한 추정

$$\hat{\tau}_y=N\ \hat{\mu}_y,\ \ \widehat{Var}(\hat{\tau}_y)\approx N^2\frac{s_e^2}{n}\frac{N-n}{N}$$

예 4.5 산림지역에서 죽은 나무의 숫자를 추정

x : 항공사진만으로 죽은 나무의 수

y : 실사를 통한 죽은 나무의 수

$\hat{y}=4.6581+0.6629x, r=0.874,\ s_e^2=129.6904/(24-1)=5.6387$

$\hat{\mu}_y=a+b\mu_x=4.6581+0.6629(12.2)=12.75$

1) 비추정

$$\widehat{se}(\hat{\mu}_y)=\sqrt{\frac{100-24}{100}\frac{5.6387}{24}}=0.423$$

95%신뢰구간 $\hat{\mu}_y\pm t_{(n-2):\alpha/2}\widehat{se}(\hat{\mu}_y)=(11.87,13.63)$

2) 단순임의표집

$$\widehat{se}(\bar{y})=\sqrt{\frac{100-24}{100}\frac{4.8812^2}{24}}=0.869$$

3) 죽은 나무 총수

$\hat{\tau}_y=N\ \hat{\mu}_y=100(12.75)=1275$

$\widehat{se}(\hat{\tau}_y)=N\widehat{se}(\hat{\mu}_y)=100(0.423)=42.3$

4.5 차이추정

$\beta = 1$임을 아는 경우

$y = \alpha + 1x + \epsilon \quad \Rightarrow \quad \hat{y} = a + x, \ \hat{\alpha} = a = \bar{y} - \bar{x}$

- 모평균 μ_y에 대한 추정

$$\hat{\mu}_y = (\bar{y} - \bar{x}) + \mu_x, \ \widehat{Var}(\hat{\mu}_y) = \widehat{Var}(\bar{e} + \mu_x) = \widehat{Var}(\bar{e}) \approx \left(\frac{N-n}{N}\right)\frac{s_e^2}{n}, \ s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}$$

- 모합 τ_y에 대한 추정

$$\hat{\tau}_y = N\,\hat{\mu}_y = \tau_x + N(\bar{y} - \bar{x}) = \tau_x + N\bar{e}$$

$$\widehat{Var}(\hat{\tau}_y) = N^2\,\widehat{Var}(\bar{e}) \approx N^2\left(\frac{N-n}{N}\right)\frac{s_e^2}{n}$$

예 4.6

$$\hat{\mu}_y = (\bar{y} - \bar{x}) + \mu_x = 1.1 + 100 = 101.1$$

$$\widehat{Var}(\hat{\mu}_y) \approx \left(\frac{N-n}{N}\right)\frac{s_e^2}{n} = \frac{300-10}{300}\frac{2.77}{10} = 0.268$$

4.6 층과 비-블록(block)과 공변수(covariate)