The EIV regression model                                    Stat 523

5-16-19

"errors in variables" also called

"measurement errors" model

(1)

$i=1,\dots,n \qquad X_i \sim N(\mu_i, \sigma_x^2)$

$$Y_i \sim N(\beta_0+\beta_1\mu_i, \sigma_y^2)$$

Assume $X_1,\dots,X_n, Y_1,\dots,Y_n$ are independent

$$L = \left(\frac{1}{\sigma_x\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma_x^2}\sum(x_i-\mu_i)^2}\left(\frac{1}{\sigma_y\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma_y^2}\sum(y_i-\beta_0-\beta_1\mu_i)^2}$$

Let $\lambda = \frac{\sigma_x^2}{\sigma_y^2} \qquad \sigma_y = \frac{\sigma_x}{\sqrt{\lambda}}$

(2)

$$L = (2\pi)^{-n}(\sigma_x^2)^{-\frac{n}{2}}\left(\frac{\sigma_x^2}{\lambda}\right)^{-\frac{n}{2}} \cdot$$

$$e^{-\frac{1}{2\sigma_x^2}\underbrace{\left[\sum_{i=1}^n(x_i-\mu_i)^2+\lambda\sum_{i=1}^n(y_i-\beta_0-\beta_1\mu_i)^2\right]}_{\star}}$$

Maximizing L with respect to $\mu_i$ is equivalent

to minimizing $\star$ w.r.t. $\mu_i$

$$\frac{\partial\star}{\partial\mu_i} = \cancel{2}(x_i-\mu_i)(\cancel{-}1) + \cancel{2}\lambda(y_i-\beta_0-\beta_1\mu_i)(\cancel{-}\beta_1) \overset{set}{=} 0$$

$$x_i + \lambda\beta_1 y_i - \lambda\beta_0\beta_1 = \mu_i + \lambda\beta_1^2\mu_i$$

③

$$\hat{\mu}_i = \frac{x_i + \lambda\beta_1(y_i - \beta_0)}{1 + \lambda\beta_1^2}$$

Substitute this into ☆:

$$☆ = \frac{\lambda}{1+\lambda\beta_1^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Let $\beta_0^* = \sqrt{\lambda}\,\beta_0$, $\beta_1^* = \sqrt{\lambda}\,\beta_1$, $y_i^* = \sqrt{\lambda}\,y_i$

then $$☆ = \frac{\sum(y_i^* - \beta_0^* - \beta_1^* x_i)^2}{1+\beta_1^{*2}}$$

④

So $\beta_0^*$ and $\beta_1^*$ are the solutions to the orthogonal least squares problem (solved last time)

Then $\hat{\beta}_0 = \frac{1}{\sqrt{\lambda}}\beta_0^*$, $\hat{\beta}_1 = \frac{1}{\sqrt{\lambda}}\beta_1^*$

$$\ln L = -n\ln(2\pi) - \frac{n}{2}\ln\sigma_x^2 - \frac{n}{2}\ln\sigma_y^2 - \frac{1}{2\sigma_x^2}\sum(x_i - \hat{\mu}_i)^2 - \frac{1}{2\sigma_y^2}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\mu}_i)^2$$

$$\frac{\partial \ln L}{\partial \sigma_x^2} = -\frac{n}{2}\frac{1}{\sigma_x^2} - \frac{1}{2}\frac{-1}{(\sigma_x^2)^2}\sum(x_i - \hat{\mu}_i)^2 \overset{\text{set}}{=} 0$$

$$\hat{\sigma}_x^2 = \frac{\sum(X_i - \hat{\mu}_i)^2}{n}$$ (5)

Similarly, $$\hat{\sigma}_Y^2 = \frac{\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{\mu}_i)^2}{n}$$

and $$\hat{\lambda} = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$$

(6)

A return to Bayesian hypothesis testing

Our previous approach ignored the differences between Type I & Type II errors

Let $p_0 = P(H_0)$ (prior probability of $H_0$)

$p_1 = P(H_1)$ $\quad (p_0 + p_1 = 1)$

Define a loss function $$\mathcal{L} = \begin{cases} 0 & \text{if correct decision} \\ l_{01} & \text{if commit Type I} \\ l_{10} & \text{if commit Type II} \end{cases}$$

⑦

Conditional risk, given $H_0$ is

$$R_0 = E[\mathcal{L} \mid H_0]$$
$$= 0(1-\alpha) + l_{01}\alpha + l_{10}\cdot 0$$
$$= l_{01}\alpha$$

Conditional risk, given $H_1$ is

$$R_1 = E[\mathcal{L} \mid H_1]$$
$$= 0(1-\beta) + l_{01}\cdot 0 + l_{10}\beta = l_{10}\beta$$

⑧

Define $ECM = p_0 R_0 + p_1 R_1$

"expected cost of misclassification"

We would like to minimize this total risk

$$ECM = p_0 l_{01}\alpha + p_1 l_{10}\beta$$
$$= p_0 l_{01}\int_R L(\theta_0)\,d\vec{x} + p_1 l_{10}\int_A L(\theta_1)\,d\vec{x}$$

⑨

$$ECM = p_0 l_{01}\left[1 - \int_A L(\theta_0)\,d\vec{x}\right] + p_1 l_{10}\int_A L(\theta_1)\,d\vec{x}$$

$$= p_0 l_{01} + \int_A p_0 l_{01}(-L(\theta_0)) + p_1 l_{10} L(\theta_1)\ d\vec{x}$$

$$= p_0 l_{01} + \int_A \left(p_1 l_{10} L(\theta_1) - p_0 l_{01} L(\theta_0)\right) d\vec{x}$$

We can minimize ECM by choosing A so that the integral is minimized

⑩

That is, define

$$A = \{\vec{x} : p_1 l_{10} L(\theta_1) - p_0 l_{01} L(\theta_0) < 0\}$$

$$= \{\vec{x} : p_1 l_{10} L(\theta_1) < p_0 l_{01} L(\theta_0)\}$$

$$= \left\{\vec{x} : \frac{L(\theta_0)}{L(\theta_1)} > \frac{p_1 l_{10}}{p_0 l_{01}}\right\}$$

$$= \left\{\vec{x} : \Lambda > \frac{p_1 l_{10}}{p_0 l_{01}}\right\}$$

HW #5 due 5/23

p.602 #12.2,

12.4

**12.2** Show that the extrema of

$$f(b) = \frac{1}{1+b^2}\left[S_{yy} - 2bS_{xy} + b^2 S_{xx}\right]$$

are given by

$$b = \frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

Show that the "+" solution gives the minimum of $f(b)$.

**12.4** Consider the MLE of the slope in the EIV model

$$\hat{\beta}(\lambda) = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}},$$

where $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$ is assumed known.

(a) Show that $\lim_{\lambda\to 0}\hat{\beta}(\lambda) = S_{xy}/S_{xx}$, the slope of the ordinary regression of $y$ on $x$.

(b) Show that $\lim_{\lambda\to\infty}\hat{\beta}(\lambda) = S_{yy}/S_{xy}$, the reciprocal of the slope of the ordinary regression of $x$ on $y$.

(c) Show that $\hat{\beta}(\lambda)$ is, in fact, monotone in $\lambda$ and is increasing if $S_{xy} > 0$ and decreasing if $S_{xy} < 0$.

(d) Show that the orthogonal least squares line ($\lambda = 1$) is always between the lines given by the ordinary regressions of $y$ on $x$ and of $x$ on $y$.

(e) The following data were collected in a study to examine the relationship between brain weight and body weight in a number of animal species.

| Species | Body weight (kg) ($x$) | Brain weight (g) ($y$) |
|---|---|---|
| Arctic fox | 3.385 | 44.50 |
| Owl monkey | .480 | 15.50 |
| Mountain beaver | 1.350 | 8.10 |
| Guinea pig | 1.040 | 5.50 |
| Chinchilla | .425 | 6.40 |
| Ground squirrel | .101 | 4.00 |
| Tree hyrax | 2.000 | 12.30 |
| Big brown bat | .023 | .30 |

Calculate the MLE of the slope assuming the EIV model. Also, calculate the least squares slopes of the regressions of $y$ on $x$ and of $x$ on $y$, and show how these quantities bound the MLE.