A Paradoxical Result in Estimating Regression Coefficients

研究方法

112354010 陳品華

Contents

1 Introduction

- This article presents a **counterintuitive result** regarding the **estimation of a regression slope coefficient** (β_1).
- The **precision of the slope estimator** can **deteriorate** when **additional information is used** to estimate its value.

A. pooled estimate of the variance $Var(\tilde{\beta}_1)$ > not pooled estimate of the variance $Var(\hat{\beta}_1)$

B. actual variance is known $Var(\tilde{\beta}_1^*)$ > actual variance is unknown $Var(\hat{\beta}_1)$

2 Simple Linear Regression Model

$y = \beta_0 + \beta_1 x + \varepsilon\,, E(\varepsilon) = 0\,, \mathrm{Var}(\varepsilon) = \sigma^2$ with x independent of ε

where $E(y|x) = \beta_0 + \beta_1 x\,,\ \mathrm{Var}(y|x) = \sigma^2$

$$E\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\ ,\ \mathrm{cov}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}\ ,\ \beta_0 = \mu_y - \beta_1 \mu_x$$

2 Estimating the $Var(\hat{\beta}_1)$ (Least Squares Estimator)

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_1) = \frac{1}{n-3}\left(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2\right)$$

<proof>

$$Var(\hat{\beta}_1) = Var\left(E(\hat{\beta}_1|x)\right) + E(Var(\hat{\beta}_1|x))$$

$$= Var(\beta_1) + E\left(\frac{\sigma^2}{S_x^2(n-1)}\right)$$

$$= 0 + \frac{\sigma^2}{\sigma_x^2} E\left(\frac{(n-1)S_x^2}{\sigma_x^2}\right)^{-1}$$

$$= \frac{\sigma_y^2 - \beta_1^2\sigma_x^2}{(n-3)\sigma_x^2}$$

$$= \frac{1}{n-3}\left(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2\right)$$

2 Estimating the $Var(\tilde{\beta}_1^*)$ (σ_x^2 is known)

$$\tilde{\beta}_1^* = \frac{s_{xy}}{\sigma_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x^2}$$

$$Var(\tilde{\beta}_1^*) = \frac{1}{n-1}\left(\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2\right)$$

<proof>

$$Var(\tilde{\beta}_1^*) = E\left(Var(\tilde{\beta}_1^* | x)\right) + Var(E(\tilde{\beta}_1^* | x))$$

$$= E\left(\frac{\sigma^2 S_x^2}{S_x^4 (n-1)}\right) + Var\left(\frac{S_x^2}{\sigma_x^2}\beta_1\right)$$

$$= \frac{\sigma^2}{\sigma_x^2 (n-1)} + \frac{2\beta_1^2}{n-1}$$

$$= \frac{1}{n-1}\left(\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2\right)$$

2 Estimating the $Var(\tilde{\beta}_1)$ (2-Samples pooled estimator)

Treatment 1	Treatment 2
$\boldsymbol{x}_{11}$ $\boldsymbol{y}_{11}$	$\boldsymbol{x}_{12}$ $\boldsymbol{y}_{12}$
$\boldsymbol{x}_{21}$ $\boldsymbol{y}_{21}$	$\boldsymbol{x}_{22}$ $\boldsymbol{y}_{22}$
. . .	. . .
$\boldsymbol{x}_{n_1 1}$ $\boldsymbol{y}_{n_1 1}$	$\boldsymbol{x}_{n_2 2}$ $\boldsymbol{y}_{n_2 2}$

x_{c1} σ_x^2 x_{c2} σ_x^2

y_1 σ_{x1}^2 y_2 σ_{x2}^2

n_1 σ_{y1}^2 n_2 σ_{y2}^2

$$\tilde{\beta}_1 = \frac{n_1 + n_2 - 2}{n_1 - 1} \left(\sum_{j=1}^{2} x'_{cj}\, x_{cj}\right)^{-1} x'_{c1} y_1$$

$$Var(\tilde{\beta}_1) \geq \frac{\sigma_1^2}{\sigma_x^2 (n_1 - 1)} \left(\frac{n_1 + n_2 - 2}{n_1 + n_2}\right)^2 + \frac{2\beta_1^2}{n_1 - 1} \left(\frac{n_2 - 1}{n_1 + n_2}\right)$$

3

Comparison

Least Squares Estimator	σ_x^2 is known	2-samples pooled estimator
$Var(\hat{\beta}_1) = \frac{1}{n-3}(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2)$	$Var(\tilde{\beta}_1^*) = \frac{1}{n-1}(\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2)$	$Var(\tilde{\beta}_1) \gtrsim \frac{1}{n_1-1}(\frac{\sigma_{y1}^2}{\sigma_x^2} + \frac{n_2-n_1}{n_1+n_2}\beta_1^2)$

3 Comparison between $\hat{\beta}_1$(LSE) and $\tilde{\beta}_1^*$ (σ_x^2 is known)

In an extreme special case : $y = \beta_0 + \beta_1 x$

$\hat{\beta}_1 = \beta_1$	$\tilde{\beta}_1^* = \beta_1 \frac{S_x^2}{\sigma_x^2}$
perfect estimation	not a perfect estimator whenever $S_x^2 \neq \sigma_x^2$

3 Comparison between $Var(\hat{\beta}_1)$ and $Var(\tilde{\beta}_1)$

$$Var(\hat{\beta}_1) < Var(\tilde{\beta}_1^*)$$

Another insight : Cauchy - Schwartz inequality : $\sigma_{xy}^2 \leq \sigma_x^2 \sigma_y^2$

$$\beta_1^2 = \frac{\sigma_{xy}^2}{\sigma_x^4} \leq \frac{\sigma_y^2}{\sigma_x^2}$$

$Var(\hat{\beta}_1)$ approaches to 0	$Var(\tilde{\beta}_1^*)$ approaches to $\frac{2\sigma_y^2}{\sigma_x^2(n-1)}$

3 Comparison between $Var(\hat{\beta}_1)$ and $Var(\tilde{\beta}_1)$

$$Var(\hat{\beta}_1) < Var(\tilde{\beta}_1^*)$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{\sigma_x^2}\left(\frac{\sigma_x^2}{S_x^2}\right) = \tilde{\beta}_1^* w \quad , where\ w = \frac{\sigma_x^2}{S_x^2}$$

when $\tilde{\beta}_1^*$ overestimates β_1, the ratio w will tend to pull $\hat{\beta}_1$ down toward the true slope β_1

3 Comparison between $Var(\tilde{\beta}_1^*)$ and $Var(\tilde{\beta}_1)$

$Var(\tilde{\beta}_1^*) = \frac{1}{n-1}(\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2)$	$Var(\tilde{\beta}_1) \gtrsim \frac{1}{n_1-1}(\frac{\sigma_{y1}^2}{\sigma_x^2} + \frac{n_2-n_1}{n_1+n_2}\beta_1^2)$

pooled estimator $\tilde{\beta}_1$	Relationship between $Var(\tilde{\beta}_1^*)$ and $Var(\tilde{\beta}_1)$
n_1 is fixed $n_2 \longrightarrow \infty$	$Var(\tilde{\beta}_1)$ would converge to $Var(\tilde{\beta}_1^*)$
finite n_1 and n_2	$Var(\tilde{\beta}_1) < Var(\tilde{\beta}_1^*)$

3 Comparison between $Var(\hat{\beta}_1)$ and $Var(\tilde{\beta}_1)$

$Var(\hat{\beta}_1) = \frac{1}{n-3}(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2)$	$Var(\tilde{\beta}_1) \gtrsim \frac{1}{n_1 - 1}(\frac{\sigma_{y1}^2}{\sigma_x^2} + \frac{n_2 - n_1}{n_1 + n_2}\beta_1^2)$

pooled estimator $\tilde{\beta}_1$	Relationship between $Var(\hat{\beta}_1)$ and $Var(\tilde{\beta}_1)$
$n_1 \gg n_2$	$Var(\hat{\beta}_1) \approx Var(\tilde{\beta}_1)$

4 Simulation

population :

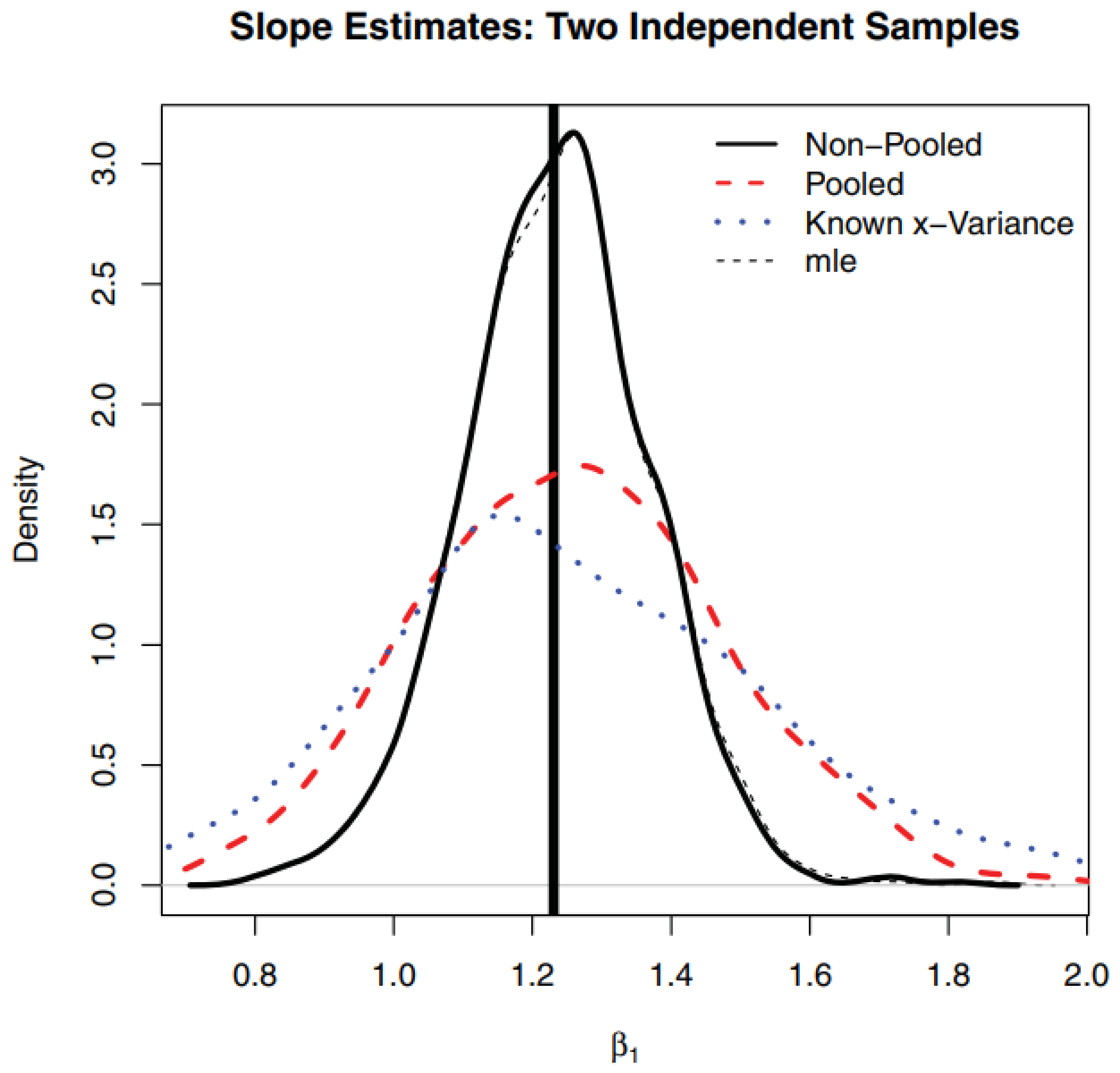
group 1 : $(x, y) \sim BN(1, 2, 1.3^2, 2^2, 0.8)$

group 2 : $(x, y) \sim BN(1, 2.5, 1.3^2, 2.5^2, 0.7)$

sample :

$n_1 = 50$

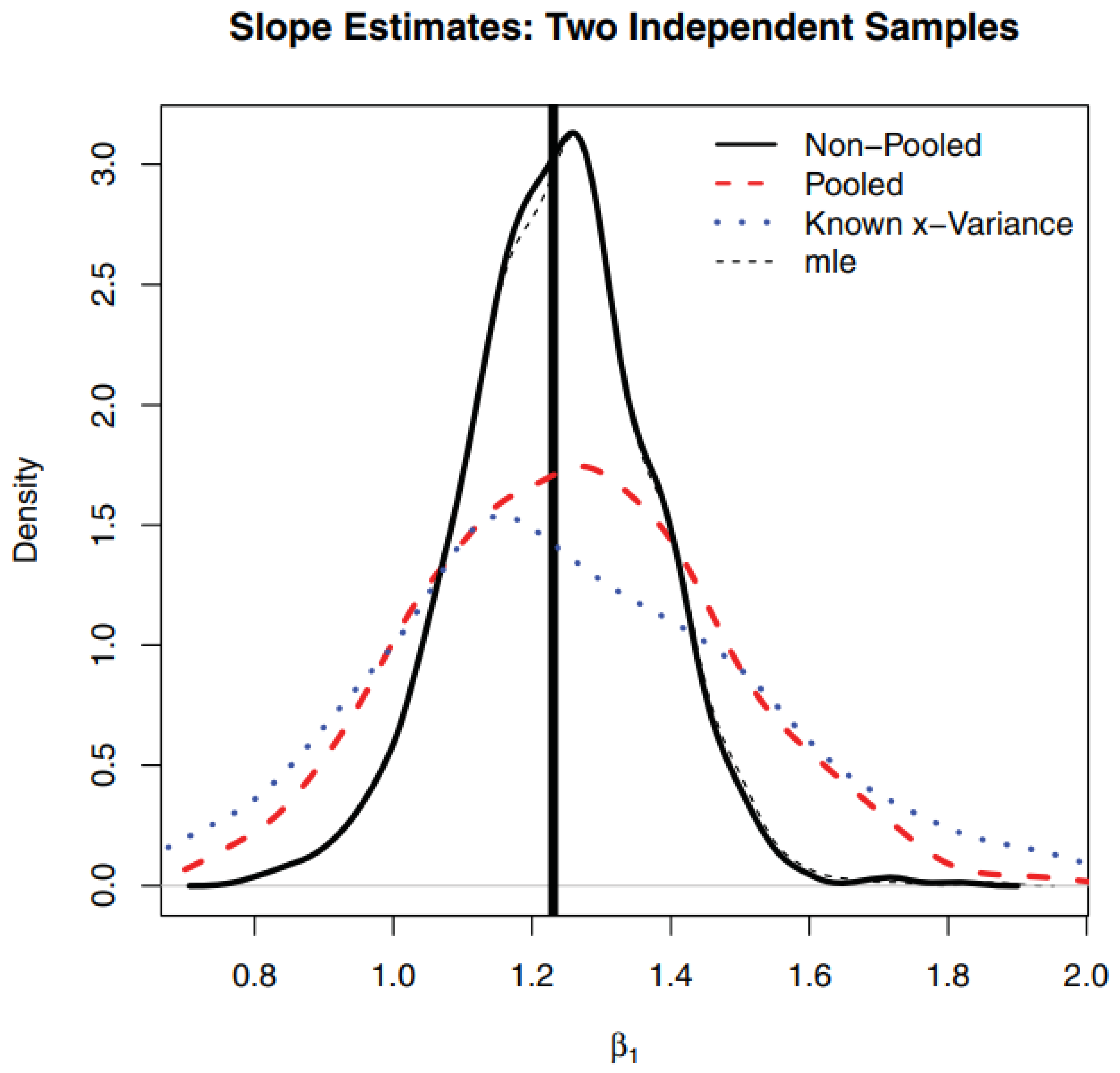
$n_2 = 50$



4 Simulation

<u>Group 1</u> ($\beta_1 = 1.23$)

1. $\sigma_{LSE} = 0.1372$
2. $\sigma_{MLE} = 0.1395$
3. $\sigma_{pooled} = 0.2197$
4. $\sigma_{\sigma_x^2 \ is \ known} = 0.2851$



5 Conclusion

- Using a **known variance** to estimate β_1 may lead to **an increase in the variance** of the estimator, which is **contrary to conventional statistical wisdom.**
- The authors point out that this paradoxical result has important implications for interpreting the results of randomized experiments and propose a new method for better estimating the relationship between predictor variables and outcomes.

Thanks!