Model Uncertainty and Incomplete Data Analysis

John Copas

University of Warwick UK

jbc@stats.warwick.ac.uk

Ref:

Copas, J.B. and Eguchi, S (2005)

Local Model Uncertainty and Incomplete Data Bias (with discussion)

to appear in JRSSB

How to measure the health risk of passive smoking?

Case-control studies ....

Case = non-smoker with cancer

Control = non-smoker without cancer

Exposed = smoker in household

Unexposed = non-smoking household

e.g. German study:

	Case	control	
Exp	12	15	27
Non-exp	11	30	41
	23	45	68

Relative risk = 2.13 (0.75, 6.05)

Meta analysis of 37 studies gives

Relative risk = 1.24 (1.12, 1.35)

Risk highly significant ($P < 0.01\%$)

.... *BUT* ....

there are **some nasty problems ...**

Publication bias — not all studies are reviewed

Confounding — effect may be partly explained by differences on other variables

Measurement error — very crude measure of exposure

All these are problems of *incomplete data*:

we would like to measure z but can only measure y

e.g.

$z =$ data on all studies + selection indicators,

$y =$ data on selected studies only

z=(response, treatment, potential confounders)

y=(response, treatment)

z=(disease status, true exposure, measurement error)

y=(disease status, observed exposure = true +error)

In all cases we can write $z = h(y)$

The basic model §2

Model: $z \sim f_Z(z, \theta)$

$$\begin{aligned} \Rightarrow y &\sim f_Y(y, \theta) \\ &= \int_{z|y} f_Z(z, \theta) J dz \end{aligned}$$

Data on $z \rightarrow$ MLE $= \hat{\theta}_Z$

Data on $y \rightarrow$ MLE $= \hat{\theta}_Y$

If f_Z is correct, asymptotically

$$\mathrm{E}(\hat{\theta}_Z) = \mathrm{E}(\hat{\theta}_Y) = \theta$$

BUT

for θ to be estimable from y, f_Z must make *untestable (ignorability) assumptions*, **which may be wrong**

A Simple Example ...

Pre-referendum poll

non	oui	total
550	450	1000

$$P(non) = \theta$$

Naive model: $x \sim binomial(1000, \theta)$

$$\Rightarrow \hat{\theta} = 0.55(0.52, 0.58)$$

BUT what about non-response?

$$P(non) = \theta, \quad P(respond) = \psi$$

We are assuming the MAR (Missing At Random) model : —

	non	oui
respond	$\theta\psi$	$(1-\theta)\psi$
refuse	$\theta(1-\psi)$	$(1-\theta)(1-\psi)$

BUT the correct model is : —

	non	oui
respond	θp_1	$(1-\theta)p_0$
refuse	$\theta(1-p_1)$	$(1-\theta)(1-p_0)$

$$\Rightarrow x \sim binomial(1000, \theta^*)$$

where

$$\theta^* = \frac{\theta p_1}{\theta p_1 + (1-\theta)p_0} = \frac{\rho\theta}{\rho\theta + (1-\theta)}$$

and

$$\rho = \frac{p_1}{p_0} = \text{ relative risk}$$

Note: MAR $\Leftrightarrow \rho = 1 \Leftrightarrow \theta^* = \theta$

Points to note: —

- $\hat{\theta}$ is unbiased only if $\rho = 1$ (MAR)
- Inference is sensitive to the value of ρ
- It is impossible to estimate ρ from these data
- It is impossible to estimate θ unless we make unverifiable assumptions
- Bayesian inference about θ will be sensitive to the prior on ρ

For example, suppose $\rho \sim N(1, \tau^2)$ with a vague prior on θ ...

Local mis-specification of f_Z §4

If

$$g_Z(z) = f_Z(z,\theta)\exp\{\epsilon u_Z(z,\theta)\}$$

then

$$\int g_Z dz = \int f_Z(1+\epsilon u_Z)dz = 1 + \epsilon \mathrm{E}_f(u_Z)$$

Hence assume

$$\mathrm{E}_f(u_Z) = 0, \mathrm{E}_f(u_Z^2) = 1$$

$$\begin{aligned} \epsilon &= \text{“mis-specification distance”} \\ &\simeq \{2 \times KL(f_Z, g_Z)\}^{\frac{1}{2}} \end{aligned}$$

and

$$u_Z = \text{“mis-specification direction”}$$

Local mis-specification of f_Y

If $z \sim g_Z$ then

$$\begin{aligned} y \sim g_Y &= \int_{z|y} f_Z(z,\theta)\{1+\epsilon u_Z(z,\theta)\} J dz \\ &= f_Y(y,\theta)\exp\{\epsilon u_Y(y,\theta)\} \end{aligned}$$

where

$$u_Y(y,\theta) = \mathrm{E}_f\{u_Z(z,\theta)|y\}$$

Compare ...

$$\begin{aligned} s_Y(y,\theta) &= \frac{\partial \log f_Y(y,\theta)}{\partial\theta} \\ &= \frac{\partial}{\partial\theta}\log\int_{z|y} f_Z(z,\theta) J dz \\ &= \mathrm{E}_f\left\{\frac{\partial f_Z(z,\theta)}{\partial\theta}\Big|y\right\} \\ &= \mathrm{E}_f\{s_Z(z,\theta)\big|y\} \end{aligned}$$

Maximum Likelihood

MLE $\hat{\theta}_Z$ is given by

$$\frac{1}{n}\sum s_Z(z_i, \hat{\theta}_Z) = 0$$

If $z_i \sim g_Z$ and $n \to \infty$

$$\begin{aligned} 0 &= \int s_Z(z, \theta_Z) f_Z(z, \theta) \exp\{\epsilon u_Z(z, \theta)\} dz \\ &\sim \int \{s_Z(z, \theta) - I_Z(\theta_Z - \theta)\} f_Z\{1 + \epsilon u_Z\} dz \\ &= -I_Z(\theta_Z - \theta) + \epsilon \mathrm{E}(s_Z u_Z) \\ &\Rightarrow \theta_Z = \theta + \epsilon I_Z^{-1} \mathrm{E}(s_Z u_Z) \end{aligned}$$

Similarly

$$\theta_Y = \theta + \epsilon I_Y^{-1} \mathrm{E}(s_Y u_Y)$$

Hence

$$b_\theta = \theta_Y - \theta_Z = \epsilon \mathrm{E} u_Z \{I_Y^{-1} s_Y - I_Z^{-1} s_Z\}$$

For given “distance” ϵ, the *incomplete data bias* b_θ depends on the “direction” u_Z. If $||b_\theta|| = b_\theta^T I_Y b_\theta$,

$$\max_{u_Z|\epsilon} ||b_\theta|| = \epsilon^2(1 - \lambda_{\text{min}})$$

where λ_{min} is the smallest eigen value of the “relative efficiency matrix”

$$\Lambda = I_Y^{\frac{1}{2}} I_Z^{-1} I_Y^{\frac{1}{2}}$$

The worst case is when $u_Y(y, \theta) \in \langle s_Y(y, \theta) \rangle$

Example §5.1: *univariate missing data*

$$z = (t, r) \qquad y = (t^{(r)}, r)$$

where

$$\begin{aligned} t^{(r)} &= t \qquad\qquad\qquad \text{if } r = 1 \\ t^{(r)} &= (-\infty, +\infty) \text{ if } r = 0 \end{aligned}$$

Model f_Z assumes MAR:

$$f_Z = f_T(t, \theta)\psi^r(1-\psi)^{1-r}$$

Model g_Z allows for non-ignorable missing data:

$$g_Z = \{f_T(t, \theta)\}\{\psi^r(1-\psi)^{1-r}\}\{\exp[\epsilon u_Z(t, r)]\}$$

$\Rightarrow$

$$\log \frac{P(r=1|t)}{P(r=0|t)} = \log \frac{\psi}{1-\psi} + \epsilon\{u_Z(t, 1) - u_Z(t, 0)\}$$

Then

$$
\begin{aligned}
b_\theta^2 &\leq \epsilon^2 I_Y^{-1}(1-\lambda_{\mathrm{min}}) \\
&= \epsilon^2 I_Y^{-1}(1-\psi) \\
&= I_Y^{-1}\psi(1-\psi)^2 \mathrm{Var}\left\{\log \frac{P(r=1|t)}{P(r=0|t)}\right\}
\end{aligned}
$$

E.g. for binary data

$$f_T = \theta^t(1-\theta)^{(1-t)}$$

$$|b_\theta| \leq \theta(1-\theta)|(\rho-1)| + O(\rho-1)^2$$

where

$$\rho = \frac{P(r=1|t=1)}{P(r=1|t=0)}$$

Example §5.2: *potential confounder*

$$z = (t, x, c) \qquad y = (t, x)$$

where

$t =$ response, $x =$ treatment, $c =$ confounder

Randomized experiment $\Rightarrow$ x and c are independent

Observational data $\Rightarrow$ x and c may be correlated $\Rightarrow$ dependence of t on x is confounded with their dependence via c

$$f_Z(t, x, c, \theta) = f_{T|X,C}(t|x, c, \theta) f_X(x) f_C(c)$$

$$f_Y(t, x, \theta) = f_{T|X}(t|x, \theta) f_X(x)$$

$$g_Z = f_{T|X,C}(t|x,c,\theta) f_X(x) f_C(c) \exp\{\epsilon u(x,c)\}$$

$$\Rightarrow \log \frac{P\{t|x\}}{P\{t|\mathrm{do}(x)\}} = \epsilon \mathrm{E}\{u(x,c)|t,x\}$$

where

$$P\{t|x\} = \int P(t|x,c)P(c|x)dc$$

and

$$P\{t|\mathrm{do}(x)\} = \int P(t|x,c)P(c)dc$$

(these are the same if the experiment is randomized)

$$b_\theta^2 \leq I_{T|X}^{-1}(1-\lambda_{\min})\mathrm{Var}\left\{\log \frac{P(c|x)}{P(c)}\right\}$$

e.g. f_Z = normal linear model with

$$\mathrm{E}(t|x,c) = \alpha + \theta x + \gamma c$$

$$\mathrm{E}(t|x) = \alpha^* + \theta x$$

then

$$b_\theta^2 \leq I_{T|X}^{-1} \mathrm{cor}^2(t,c|x) \mathrm{cor}^2(c,x)$$

Worst case is when g_Z gives c a linear regression on x

Example: Meta Analysis of Case-Control Studies

$t =$ presence or absence of cancer ($t = 1, 0$)

$x =$ presence or absence of exposure ($x = 1, 0$)

Standard model is

$$\log P(cancer|x) = \psi + \theta x$$

$\Rightarrow \theta =$ log relative risk

2×2 table from jth case-control study gives

$$\hat{\theta}_j \sim N(\theta, \sigma_j^2)$$

Meta analysis weights $w_j = 1/(\sigma_j^2 + \tau^2)$ give

$$\begin{aligned}\tilde{\theta} &= \sum w_j \hat{\theta}_j / \sum w_j \\ &\sim N(\theta, 1/\sum w_j)\end{aligned}$$

$$\tilde{\theta} = 0.22(0.12, 0.32)$$

There will be many confounders c (e.g. quality of diet)

$$\log P(cancer|x,c) = \psi + \theta x + \alpha c$$

with $\mathrm{Var}(c) = 1$

$$\Rightarrow \lambda = 1 - \frac{\alpha^2}{\sigma_x^2}$$

$$\Rightarrow \mathrm{E}(\hat{\theta}_j) = \theta + \alpha\{\mathrm{E}(c|x=1) - \mathrm{E}(c|x=0)\}$$

$$= \theta \text{ if } x \text{ and } c \text{ are independent}$$

Suppose

$$c|x \sim N(\psi^* + \epsilon x, 1 - \epsilon^2\sigma_x^2)$$

$$\Rightarrow \mathrm{corr}(x,c) = \rho = \frac{\epsilon\sigma_x}{\sqrt{1+\epsilon^2\sigma_x^2}}$$

$$\mathrm{E}(\hat{\theta}_j) = \theta + \alpha\epsilon$$

Simplification: assume all studies are similar

Strength of confounder:

$$\lambda = 1 - \frac{\alpha^2}{\sigma_x^2}$$

Degree of non-ignorability:

$$\rho = \frac{\epsilon \sigma_x}{\sqrt{1 + \epsilon^2 \sigma_x^2}}$$

Resulting bias

$$\text{bias} = \alpha \epsilon$$

Sensitivity Analysis:

estimate σ_x, fix λ, plot bias against ρ

find smallest ρ such that

$$|\epsilon \alpha| \geq |\tilde{\theta}| - \frac{1.96}{\sqrt{\sum w_j}}$$

Publication Bias in Meta Analysis

$$z = (t, x, r), y = (t^{(r)}, x^{(r)}, r)$$

$t =$ study outcome

$x^2 = \mathrm{Var}(t)$

$r = 1$ published, $r = 0$ unpublished

$$t|x \sim N(\theta, x^2)$$

$$x \sim f_X(x)$$

$f_Z : r \perp t|x$

$g_Z : P(publish|t, x) = p(t, x)$

$$\tilde{\theta} = \frac{\sum x^{-2} t}{\sum x^{-2}}$$

$$\Rightarrow \text{bias} = \frac{\mathrm{E}\{x^{-2}(t-x)p(t,x)\}}{\mathrm{E}\{x^{-2}p(t,x)\}}$$

$$P(unpublished) = p = 1 - \mathrm{E}\{p(t,x)\}$$

THEOREM

If $\mathrm{E}\{p(t,x)|x\} \searrow x$ then

$$|\text{bias}| \leq \frac{\phi\{\Phi^{-1}(p)\}}{1-p} \frac{\mathrm{E}(x^{-1}|r=1)}{\mathrm{E}(x^{-2}|r=1)}$$

Ref: Copas and Jackson (2004) *A bound for publication bias based on the fraction of unpublished studies*, Biometrics, **60**, 146-153

Observed studies (t_i, x_i), $i = 1, 2, \cdots n$

Sensitivity Analysis:

Plot

$$B(m) = \frac{m+n}{n} \phi \left\{ \Phi^{-1} \left(\frac{n}{m+n} \right) \right\} \frac{\sum x_i^{-1}}{\sum x_i^{-2}}$$

against m for $m = 1, 2, \cdots$

Find smallest m such that

$$B(m) \geq \left| \frac{\sum t_i x_i^{-2}}{\sum x_i^{-2}} \right| - \frac{1.96}{\sqrt{\sum x_i^{-2}}}$$