Sample selection models for non-Gaussian response
a general proposal

Adelchi Azzalini
University of Padua, Italy
adelchi.azzalini@unipd.it
Guest visitor: Department of Statistics, University of Pretoria
joint work with Hyoung-Moon Kim and Hea-Jung Kim

DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) — 23rd March 2018

Note: opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the CoE-MaSS

Sample selection, general

- Denote by Y the variable of interest (target) and by Y_{obs} the sampling variable (actual observations)
- Ideally

$$Y \equiv Y_{\text{obs}}$$

- In some cases, the two variables do not coincide
- Usual source of problem is some censoring mechanism
- typically this occurs in observational studies
- The term ‘sample selection’ commonly related to Heckman work (1976, 1979), although earlier work exist (Gronau, 1974)

Key example

- $Y \sim \mathrm{N}(\mu, \sigma^2)$ is of interest
- consider case where Y is associated to U, assume specifically

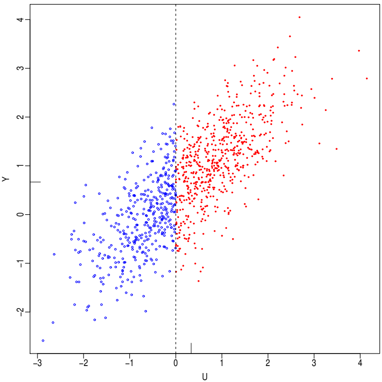
$$\begin{pmatrix} U \\ Y \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} \tau \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$$

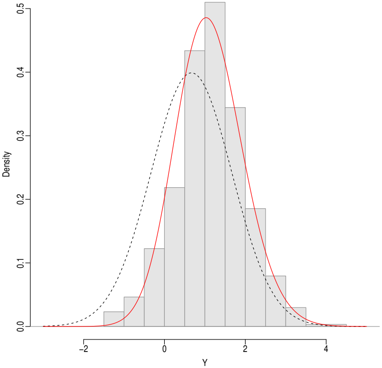
- suppose we observe Y conditionally on $U \geq 0$
- distribution of observed Y_{obs} values is

$$f_{\text{obs}}(x) = \underbrace{\frac{1}{\sigma}\varphi(z)}_{\mathrm{N}(\mu,\sigma^2)} \underbrace{\left[\Phi\left(\frac{\tau+\rho z}{\sqrt{1-\rho^2}}\right)/\Phi(\tau)\right]}_{\text{perturbation factor}}, \qquad z = \left(\frac{x-\mu}{\sigma}\right)$$

Key example, visually

$$\begin{pmatrix} U \\ Y \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}, \begin{pmatrix} 1 & 3/4 \\ 3/4 & 1 \end{pmatrix} \right), \qquad n = 1000 \text{ sample values}$$





Classical real case (Heckman, 1979)

- Y represents women wage: $Y_1, \ldots, Y_n$
- $Y_i = \underbrace{x_i^\top \beta}_{\mu_i} + \varepsilon_i$, where x_i are covariates, interest in $\beta \in \mathbb{R}^p$
- $U_i = \underbrace{w_i^\top \gamma}_{\tau_i} + \zeta_i$, where w_i are covariates, $\gamma \in \mathbb{R}^p$
- (U_i, Y_i) jointly normal, individuals behave independently
- if $U_i \leq 0$ the woman decides not to work
- we do not observe the latent variable U_i, but only

$$D_i = \begin{cases} 1 & \text{if } U_i > 0 \text{ (i.e. the woman works)} \\ 0 & \text{otherwise} \end{cases}$$

- available information is of the form:

work/no work (d) :	1 0 1 1 1 0 1 1 0 . . .
salary (y_{obs}) :	y ? y y y ? y y ? . . .

Likelihood function of Heckman's model

- Notation: d_i is realized value of D_i, y_i is realized valued of Y_i
- available data:

  $d_1, \ldots, d_n$: work (yes/no), y_i: wage, only when $d_i = 1$
- $\mathbb{P}\{D_i = 1\} = \Phi(\tau_i)$
- PDF of $(Y_i|D_i = 1) =$ (Normal PDF)$(y_i) \times$ (perturbation factor)

$$\begin{aligned}\log L &= \sum_{d_i=1} \log\left[\mathbb{P}\{D_i = 1\} \times f(y_i|D_i = 1)\right] + \sum_{d_i=0} \log\, \mathbb{P}\{D_i = 0\} \\ &= \sum_{d_i=1} \log\Big[\,\underbrace{f(y_i)}_{\mathrm{N}(\mu_i,\sigma^2)} \times \mathbb{P}\{D_i = 1|y_i\}\,\Big] + \sum_{d_i=0} \log\,[1 - \Phi(\tau_i)]\end{aligned}$$

where

$$\mathbb{P}\{D_i = 1|y_i\} = \Phi\left(\frac{\tau + \rho z_i}{\sqrt{1-\rho^2}}\right), \qquad z_i = \left(\frac{y_i - \mu_i}{\sigma}\right)$$

Some remarks and related work

- the resulting estimate is corrected for selection bias
- widely applied construction in socio-economic literature
- criticism: results strongly dependent on normality assumption
- Non-parametric and semi-parametric formulations exist, but not much used in practice; large datasets are required
- robust versions for continuous response (Marchenko & Genton, 2012; Zhelonkin *et alii*, 2016)
- less development for discrete response variables (probit adjusted 'á la Heckman': Van de Ven & Van Praag, 1981)
- recent work using copulae to regulate dependence (Marra & Wyszynski, 2016, 2017)

Our plan of work

- highlight connection with literature on ‘modulated symmetry’
- develop a general construction for selection distributions
- work in a (flexible) parametric context
- focus especially on discrete distributions

Symmetry-modulated distributions

- ‘Extendend skew-normal distribution’:

$$f_{\text{obs}}(x) = \underbrace{\frac{1}{\sigma}\varphi(z)}_{\mathrm{N}(\mu,\sigma^2)} \underbrace{\left[\Phi\left(\frac{\tau+\rho z}{\sqrt{1-\rho^2}}\right)/\Phi(\tau)\right]}_{\text{perturbation factor}}, \qquad z = \left(\frac{x-\mu}{\sigma}\right)$$

- this is an instance of a general construction of continuous type

$$f_{\text{obs}}(x) = f(x)\left[G(x)/\pi\right]$$

where

$$\begin{aligned} G(x) &= \mathbb{P}\{x \text{ is observed} \mid Y = x \text{ is sampled from } f\}, \\ \pi &= \mathbb{P}\{\text{actually observe the sampled value}\} = \mathbb{E}_f\{G(Y)\} \end{aligned}$$

- under appropriate symmetry conditions, $\pi = 1/2$ holds
- multivariate extensions are simple to obtain
- see Azzalini & Capitanio (2014) for an overview

Selection as modulation of a general distribution

$$f_{\text{obs}}(x) = f(x)G(x)/\pi \qquad\qquad (x \in \mathbb{R},\ \text{or a subset})$$

- adopt this construction with non-symmetric f, possibly discrete
- in general, main technical issue is computation of

$$\pi = \mathbb{P}\{\text{do observe a sampled valued}\} = \mathbb{E}_f\{G(Y)\}$$

- in the discrete case integration reduces to a summation
- in continuous case use numerical integration
- log-likelihood:

$$\begin{aligned}\log L &= \sum_{d_i=1} \log\left[f(y_i) \times \mathbb{P}\{D_i = 1|y_i\}\right] + \sum_{d_i=0} \log\ \mathbb{P}\{D_i = 0\}\\ &= \sum_{d_i=1} \log\{f(y_i)\ G(y_i)\} + \sum_{d_i=0} \log\left(1 - \pi_i\right)\end{aligned}$$

Selection model for binary case, response component

- The simplest case occurs with binary response:

$$\mathbb{P}\{Y = 1\} = \mu, \qquad \mathbb{P}\{Y = 0\} = 1 - \mu$$

- then

$$\pi = \mathbb{E}_f\{G(Y)\} = (1 - \mu)\, G(0) + \mu\, G(1)$$

- if $\mathbb{E}\{Y\}$ depends on covariates, then

$$\pi_i = (1 - \mu_i)\, G(0) + \mu_i\, G(1), \qquad \mu_i = \text{function}(x_i^\top \beta)$$

- most common choices are the logit and probit models:

$$\mu_i = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)}, \qquad \mu_i = \Phi(x_i^\top \beta)$$

- still need to introduce model for $G(\cdot)$ component...

Selection model for binary case, selection component

- conceptually convenient to introduce a latent variable

$$T \sim G_0$$

and some appropriate function $h(\cdot)$, to write

$$G(y) = G_0\{h(y)\} = \mathbb{P}\{T \leq h(y) | Y = y\}$$

- covariates are incorporated in $h(\cdot)$ through $\tau_i = w_i^\top \gamma$
- Instance A: $T \sim \mathrm{N}(0,1), \quad h(y) = \tau_i + \alpha\mu_i^{-1}y$

$$G(y) = \Phi(\tau_i + \alpha\mu_i^{-1}y)$$

- Instance B: $T \sim \mathrm{Expn}(1), \quad h(y) = \exp(\tau_i + \alpha\mu_i^{-1}y)$

$$G(y) = 1 - \exp\{-exp(\tau_i + \alpha\mu_i^{-1}y)\}$$

- Instance C, ...
  (ideally motivated by subject matter considerations)
- parameter α plays a similar role of ρ in Heckman's model

Other discrete distributions

- $Y_i \sim \text{Poisson}(\mu_i), \quad \mu_i = \exp(x_i^\top \beta)$
- approximate π by truncated sum

$$\pi_i \approx \sum_{k=0}^{K} \frac{e^{-\mu_i} \mu_i^k}{k!} G(k),$$

- options for $G(\cdot)$ as before
- Negative Binomial and other discrete distributions handled similarly

An alternative form of selection mechanism

- An interesting alternative for G is to take $T \sim \text{Expn}(1)$ and
$$h(y) = \exp(\tau) + \alpha\mu^{-1}\,y = \lambda + \eta\,y$$
leading to
$$G(y) = 1 - \exp\{-(\lambda + \eta\,y)\}$$
- Then for a positive response Y (discrete or continuous) get exactly
$$\pi = \int_0^\infty f(y)\left(1 - e^{-\lambda-\eta y}\right)\,\mathrm{d}y = 1 - e^{-\lambda}\,M(-\eta)$$
provided moment generating function $M(\cdot)$ of f is known
- restriction: requires $\alpha \geq 0$

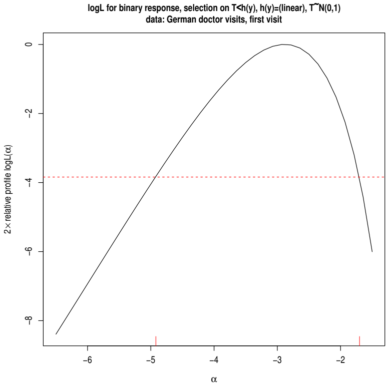
Computational aspects

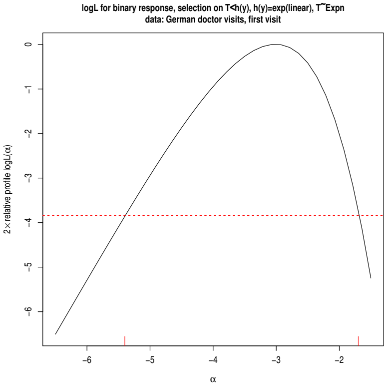
- parameters: α and $\theta = (\beta^\top, \gamma,^\top, \psi)$
  where ψ may be an additional parameter of f, e.g. dispersion
- to maximize $\log L$, consider profile log-likelihood
  $$\log L_p(\alpha) = \log L(\alpha, \hat\theta(\alpha))$$
  and evaluate over a grid of α values
- initial values of θ: take $\alpha = 0$ and fit two separate generalized linear models for Y and D
- first- and second-order derivatives of $\log L$ are available, for a given α, hence numerical maximization is speeded-up
- at the end of the process, retain $\hat\alpha$ which maximizes $\log L_p$ and the corresponding $\hat\theta(\hat\alpha)$
- standard errors from Hessian matrix of $\log L(\alpha, \hat\theta(\alpha))$

Numerical illustration with binary data

- Consider data of Riphahn et al. (2003) about usage preferences of German health insurance system
- Y_i: 'subject i makes at least one visit to the doctor in the year'
- D_i: 'subject i has subscribed for publich health insurance'
- the data have been fitted by Greene (2012, p. 921–2) using the bivariate probit method of Van de Ven & Van Praag (1981)
- we fit also our model described above
- general indication is broadly similar to earlier findings
- two different choices of $G(\cdot)$ produce almost identical anwsers (hence typical problem of classical Heckman model does not emerge)

Numerical illustration with binary data, $\log L_p$





Short summary of simulation work

- Various simulation experiments, whose basic structure was:
  - response: binary or Poisson variable,
  - selection: either earlier Instance A (normal T, linear h)
    or Instance B (exponential T, exponential h)
  - $\mu_i = x_i^\top \beta = 0.5 + 1.5\, x_i, \quad \tau_i = w_i^\top \gamma = 1 + x_i + 1.5\, w_i$
- Variants:
  - with or without ‘exclusion restriction’ (= without term $1.5\, w_i$)
  - increasing number of components in x_i and w_i to 6 and 7
  - Some experiments sampled data from a different dependence model (copula)
- Key finding: estimates of β remain nearly unbiased
  - even without exclusion restriction,
  - even sampling data from the ‘wrong’ dependence model

Summary remarks

- The proposed formulation is quite flexible,
  it allows many specifications
- Particularly suited for discrete response variables
- The response and the selection equations are chosen separately
- Estimation of the response equation appears robust
  to misspecification of the selection mechanism

Some references

- Azzalini, A. with the collaboration of Capitanio, A. (2014). *The Skew-Normal and Related Families*. Cambridge U. Press.
- Greene, W. H. (2012). *Econometric Analysis*, 7th edition. Pearson Education Ltd, Harlow.
- Heckman, J. J. (1976). *Ann. Econ. Socl. Measmnt.*, **5**, 475–492.
- Heckman, J. J. (1979). *Econometrica*, **47**, pp. 153-61.
- Marchenko and Genton (2012). *J. Amer. Statist. Assoc.*, **107**, 304–317.
- Marra, G. and Wyszynski, K. (2016). *CSDA* **104**, 110–129.
- Riphahn, R. R., Wambach, A. and Million, A. (2003). *J. Applied Econometrics*, **18**, 387–405.
- Van de Ven, W.P.M.M. and Van Praag, B.M.S. (1981). *J. Econometrics*, **17**, p.229–252. Corrigendum in Vol. **22** (1983), p. 395.
- Wyszynski, K. and Marra, G. (2017). *Comput. Stat.*, to appear.
- Zhelonkin, Genton & Ronchetti (2016). *JRSS-B* 78, 805–827
- this paper: prelim. `arXiv` (2016), to appear in *Stat. Methods & Appl.*