Seminar at Institute of Statistics, NCU, 2011, 11, 30 (Wed)

A goodness-of-fit test for parametric models based on dependently truncated data

by Takeshi Emura

Joint work with Dr. Yoshihiko Konno

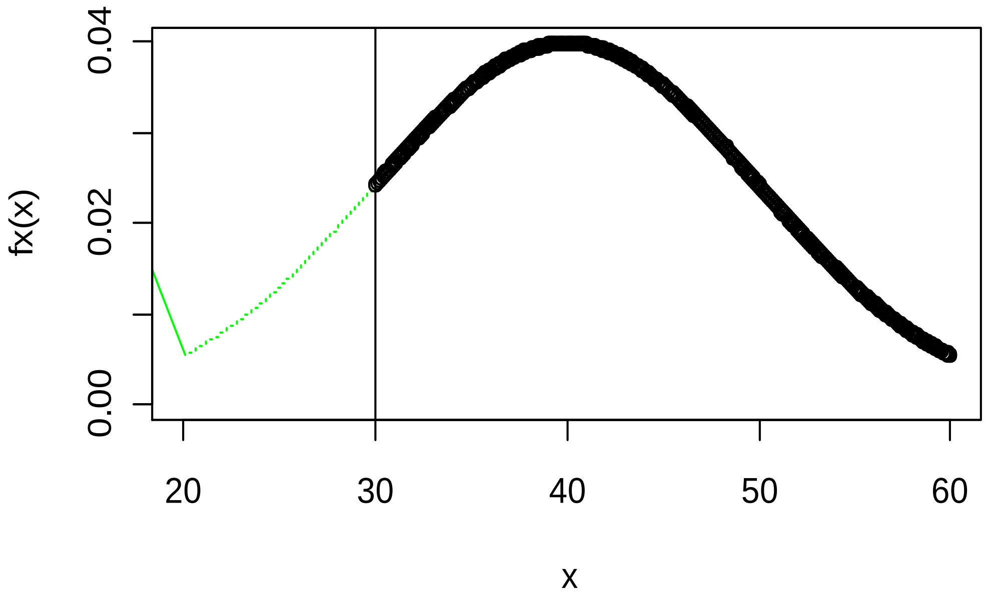
Outlines

- Truncated data
- Goodness-of-fit via parametric Bootstrap
- Goodness-of-fit via multiplier central limit theorem (proposed)
- Simulations, Data analysis
- Summary

Truncated data

- Data from truncated normal

$$\{X_j; j=1,\ldots,n\} \quad \text{subject to } l \le X_j$$



$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}$$

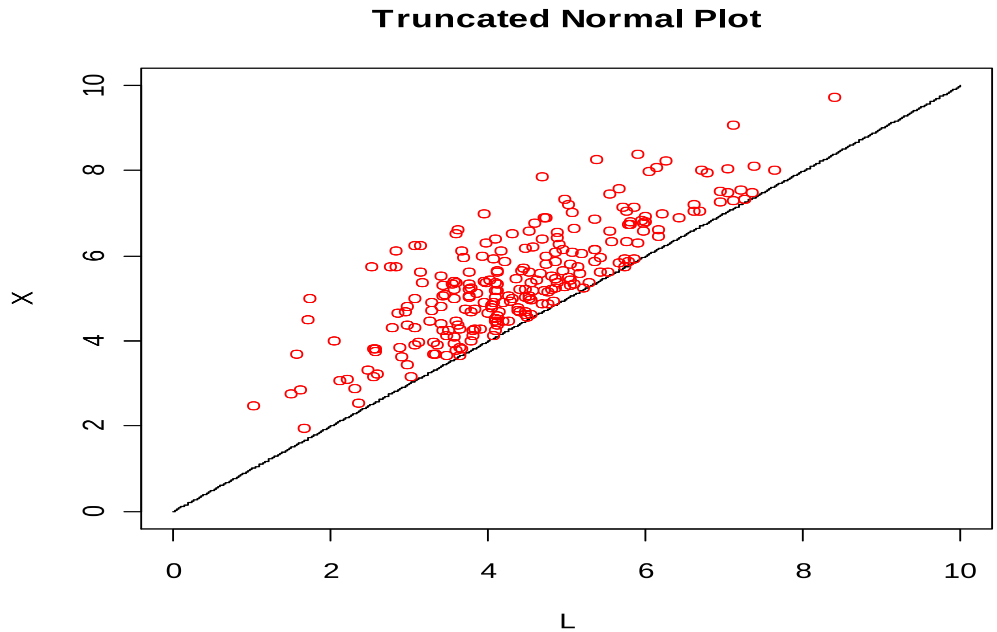
- Estimation via MLE; $(\hat{\mu}_X, \hat{\sigma}_X^2)$

  (numerical solution; Cohen, 1991);

Truncated data

- Left-truncated data
  (vast literature; e.g., Klein & Moeschberger, 2003)

  $\{(L_j, X_j)\,(j=1,\ldots,n)\}$ subject to $L_j \le X_j$



L~N(5,2),
X~N(5,2),
Cov(L,X)=0.5

- Estimation via MLE; $(\hat{\mu}_L, \hat{\mu}_X, \hat{\sigma}_L^2, \hat{\sigma}_X^2, \hat{\sigma}_{LX})$
  (Emura & Konno, 2010, Statistical Papers)

Truncated data

$$\{(L_j, X_j)(j=1,\ldots,n)\} \quad \text{subject to } L_j \le X_j$$

- Estimation of $\Pr(L \le l)$ & $\Pr(X \le x)$

1) Under $L \perp X$ with marginals unspecified

(Lynden-Bell, 1971; Woodroofe, 1985, etc.)

2) Under Copula model on (L, X)

(Chaieb et al., 2006; Emura, Wang & Hung, 2011; Emura and Wang, submit to *JMVA*, talk in TMS)

3) Under full parametric model

(Emura and Konno, 2010; Emura and Konno, submit to *CSDA*)

goodness-of-fit, focus of this talk

Truncated data

- Set up

  Population random variable; (L^O, X^O)

  If $L^O > X^O$, data is truncated (nothing observed)

  If $L^O \le X^O$, a pair $(L^O, X^O) = (L_j, X_j)$ is observed

Post-truncated data;

$$\{(L_j, X_j)\,(j = 1, \ldots, n)\} \quad \text{subject to } L_j \le X_j$$

$$i.i.d. \text{ from } \Pr(L^O \in dl, X^O \in dx \mid L^O \le X^O) = \frac{f_{L^O X^O}(l, x)\mathbf{I}(l \le x)}{\Pr(L^O \le X^O)}$$

Truncated data

- <u>Maximum likelihood</u>

$$\hat{\boldsymbol{\theta}} = \arg\max L(\boldsymbol{\theta})$$

where $L(\boldsymbol{\theta}) = c(\boldsymbol{\theta})^{-n} \prod_j f_{\boldsymbol{\theta}}(L_j, X_j)$

and $c(\boldsymbol{\theta}) = \Pr(L^O \leq X^O)$

*We need the form of $c(\boldsymbol{\theta}) = \Pr(L^O \leq X^O)$

$$= \iint_{l \leq x} f_{\boldsymbol{\theta}}(l, x) dl dx$$

Most commonly used distributions do not permit an explicit form of $c(\boldsymbol{\theta})$ (Emura and Konno, 2010)

Some useful bivariate models

Example 1: Bivariate *t* or bivariate normal

$$\boldsymbol{\theta}' = (\boldsymbol{\mu}' = (\mu_L, \mu_X), \sigma_L^2, \sigma_X^2, \sigma_{LX}, v)$$

$$f_{\boldsymbol{\theta}}(l, x) = \frac{(\sigma_L^2 \sigma_X^2 - \sigma_{LX}^2)^{-1/2} \Gamma\{(v+2)/2\}}{\pi\, \Gamma(v/2) v} \left\{ 1 + \frac{Q^2(l, x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{v} \right\}^{-(v+2)/2}$$

$$c(\boldsymbol{\theta}) = \Pr(L^O \le X^O) = \Psi\left(\frac{\mu_X - \mu_L}{\sqrt{\sigma_X^2 + \sigma_L^2 - 2\sigma_{LX}}}; v \right)$$

Example 2: Bivariate Poisson (Holgate, 1964) :

$$f_{\boldsymbol{\theta}}(l, x) = e^{-\lambda_L - \lambda_X - \alpha} \sum_{w=0}^{l} \frac{\lambda_L^{l-w} \lambda_X^{x-w} \alpha^w}{(l-w)!(x-w)!w!}$$

$$c(\boldsymbol{\theta}) = \Pr(L^O \le X^O) = e^{-\lambda_L - \lambda_X} \sum_{l=0}^{\infty} \frac{\lambda_L^l}{l!} S_V(l) \qquad S_V(l) = \sum_{v=l}^{\infty} \frac{\lambda_X^v e^{-\lambda_X}}{v!}$$

Example 3: Zero-modified Poisson (Dietz and Böhning, 2000)

$$f_{\boldsymbol{\theta}}(l, x) = p^l (1-p)^{1-l} \frac{(\lambda_X)^x e^{-\lambda_X}}{x!}, \qquad l = 0, 1; \quad x = 0, 1, 2, \ldots$$

$$c(\boldsymbol{\theta}) = \Pr(L^O \le X^O) = 1 - p e^{-\lambda_X}$$

Truncated data

Although parametric modeling easily incorporates the dependence structure between L^O and X^O, it involves strong distributional assumptions

If the goodness-of-fit tests are used, the robustness concern about the parametric analysis can be circumvented.

Let $\Im = \{ f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta \}$ be a given parametric family.

$H_0 : f \in \Im$ against $H_1 : f \notin \Im$.

Goodness-of-fit

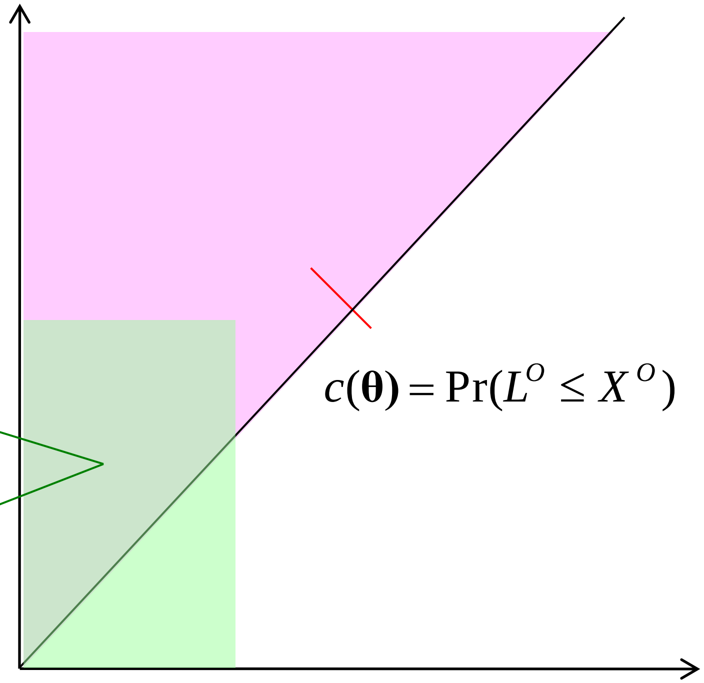
- Cramér-von Mises type statistic

$$C = \iint_{l \le x} \{ \hat{F}(l,x) - F_{\hat{\boldsymbol{\theta}}}(l,x)/c(\hat{\boldsymbol{\theta}}) \}^2 d\hat{F}(l,x)$$

$$= \sum_j \{ \hat{F}(L_j, X_j) - F_{\hat{\boldsymbol{\theta}}}(L_j, X_j)/c(\hat{\boldsymbol{\theta}}) \}^2,$$

where $\hat{F}(l,x) = \sum_j \mathbf{I}(L_j \le l, X_j \le x)/n$

$$F_{\boldsymbol{\theta}}(l,x) = \begin{cases} \iint_{u \le l, u \le v \le x} f_{\boldsymbol{\theta}}(u,v)\,du\,dv & \text{continuous;} \\ \sum_{u \le l} \sum_{u \le v \le x} f_{\boldsymbol{\theta}}(u,v) & \text{discrete.} \end{cases}$$



- Bootstrap has been suggested under complete data: (p.279 of *Asymptotic Statistics* by van der Vaart, 1998; ; Sec 4.1 of Genest & Remillard, 2008, Ann. Poincare 2009)

Some useful bivariate models

Example 1: Bivariate normal

$$f_{\boldsymbol{\theta}}(l, x) = \text{Bivariate normal density}$$

$$F_{\boldsymbol{\theta}}(l, x) = \int_{-\infty}^{(l-\mu_L)/\sigma_L} \left[\Phi\left\{ \frac{x - \mu_X - \sigma_{LX} s / \sigma_L}{\sqrt{\sigma_X^2 - \sigma_{LX}^2 / \sigma_L^2}} \right\} - \Phi\left\{ \frac{\mu_L + \sigma_L s - \mu_X - \sigma_{LX} s / \sigma_L}{\sqrt{\sigma_X^2 - \sigma_{LX}^2 / \sigma_L^2}} \right\} \right] \phi(s) ds$$

Example 2: Bivariate Poisson (Holgate, 1964) :

$$f_{\boldsymbol{\theta}}(l, x) = e^{-\lambda_L - \lambda_X - \alpha} \sum_{w=0}^{l} \frac{\lambda_L^{l-w} \lambda_X^{x-w} \alpha^w}{(l-w)!(x-w)!w!} \qquad F_{\boldsymbol{\theta}}(l, x) = e^{-\lambda_L - \lambda_X - \alpha} \sum_{u=0}^{l} \sum_{v=u}^{x} \sum_{w=0}^{u} \frac{\lambda_L^{u-w} \lambda_X^{v-w} \alpha^w}{(u-w)!(v-w)!w!}$$

Example 3: Zero-modified Poisson (Dietz and Böhning, 2000)

$$f_{\boldsymbol{\theta}}(l, x) = p^l (1-p)^{1-l} \frac{(\lambda_X)^x e^{-\lambda_X}}{x!}, \qquad l = 0, 1; \quad x = 0, 1, 2, \ldots$$

$$F_{\boldsymbol{\theta}}(l, x) = \begin{cases} (1-p) e^{-\lambda_X} \sum_{u=0}^{x} (\lambda_X^x / u!) & \text{if } l = 0 \;; \\ e^{-\lambda_X} \left(\sum_{u=0}^{x} \lambda_X^x / u! - p \right) & \text{if } l = 1, \end{cases}$$

Parametric Bootstrap

Step 0: Calculate the statistic $C = \sum_i \{ \hat{F}(L_i, X_i) - F_{\hat{\boldsymbol{\theta}}}(L_i, X_i) / c(\hat{\boldsymbol{\theta}}) \}^2$.

Step 1: Generate $(L_j^{(b)}, X_j^{(b)})$ which follows the truncated distribution of $F_{\hat{\boldsymbol{\theta}}}(l, x)$,

subject to $L_j^{(b)} \leq X_j^{(b)}$, for $b = 1, 2, \ldots, B,\ j = 1, 2, \ldots, n$.

Step 2: Calculate $\{ C^{*(b)}; b = 1, 2, \ldots, B \}$,

where $C^{*(b)} = \sum_i \{ \hat{F}^{(b)}(L_i^{(b)}, X_i^{(b)}) - F_{\hat{\boldsymbol{\theta}}^{(b)}}(L_i^{(b)}, X_i^{(b)}) / c(\hat{\boldsymbol{\theta}}^{(b)}) \}^2$ and where

$\hat{F}^{(b)}(l, x)$ and $\hat{\boldsymbol{\theta}}^{(b)}$ are the empirical CDF and MLE based on

$\{ (L_j^{(b)}, X_j^{(b)}); j = 1, 2, \ldots, n \}$.

Step 3: Reject H_0 at the 100α % significance level if $\sum_{b=1}^{B} \mathbf{I}(C^{*(b)} \geq C) / B < \alpha$.

Parametric Bootstrap

- Some simulation results under bivariate normal model (Emura and Konno, 2010)

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_L \\ \mu_X \end{bmatrix} = \begin{bmatrix} 120-62.63 \\ 60.82 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_L^2 & \sigma_{LX} \\ \sigma_{LX} & \sigma_X^2 \end{bmatrix} = \begin{bmatrix} 19.64^2 & 19.64 \times 16.81 \rho_{LX} \\ 19.64 \times 16.81 \rho_{LX} & 16.81^2 \end{bmatrix}$$

Table 1: Type I error rates of the the Cramér-von-Mises type goodness-of-fit test at level α based on 300 replications. The cut-off value is obtained by the parametric bootstrap-based procedure based on 1,000 resamplings.

	$\rho_{X^O Y^O}$				
	-0.70	-0.35	0.00	0.35	0.70
$\alpha = 0.10$	0.113	0.090	0.107	0.123	0.083
$\alpha = 0.05$	0.056	0.040	0.047	0.060	0.050
$\alpha = 0.01$	0.013	0.007	0.007	0.023	0.013

Multiplier method

- Parametric Bootstrap is computationally very intensive, especially in Step 2:

Step 2: Calculate $\{ C^{*(b)}; b = 1, 2, \ldots, B \}$,

where $C^{*(b)} = \sum_i \{ \hat{F}^{(b)}(L_i^{(b)}, X_i^{(b)}) - F_{\hat{\boldsymbol{\theta}}^{(b)}}(L_i^{(b)}, X_i^{(b)}) / c(\hat{\boldsymbol{\theta}}^{(b)}) \}^2$ and where

$\hat{F}^{(b)}(l, x)$ and $\hat{\boldsymbol{\theta}}^{(b)}$ are the empirical CDF and MLE based on

$\{ (L_j^{(b)}, X_j^{(b)}); j = 1, 2, \ldots, n \}$.

- Maximizing likelihood functions *B* times often encounter the problem of local maxima.

Proposed method

- Asymptotic linear expression

$$\sqrt{n}\{\hat{F}(l,x)-F_{\hat{\boldsymbol\theta}}(l,x)/c(\hat{\boldsymbol\theta})\}=\frac{1}{\sqrt{n}}\sum_j \hat{V}_j(l,x;\hat{\boldsymbol\theta})+o_P(1),$$

where $\hat{V}_j(l,x;\boldsymbol\theta)=\mathbf{I}(L_j\le l,\ X_j\le x)-F_{\boldsymbol\theta}(l,x)/c(\boldsymbol\theta)-\mathbf{g}'_{\boldsymbol\theta}(l,x)\{i_n(\boldsymbol\theta)/n\}^{-1}\dot{l}_j(\boldsymbol\theta),$

$\mathbf{g}_{\boldsymbol\theta}(l,x)=c(\boldsymbol\theta)^{-2}\{\dot{F}_{\boldsymbol\theta}(l,x)c(\boldsymbol\theta)-F_{\boldsymbol\theta}(l,x)\dot{c}(\boldsymbol\theta)\},$

$\dot{F}_{\boldsymbol\theta}(l,x)=\partial F_{\boldsymbol\theta}(l,x)/\partial\boldsymbol\theta,\quad \dot{c}(\boldsymbol\theta)=\partial c(\boldsymbol\theta)/\partial\boldsymbol\theta$

- Let Z_j be random variable with $\mathrm{E}(Z_j)=0,\ Var(Z_j)=1$

$$\frac{1}{\sqrt{n}}\sum_j Z_j\hat{V}_j(l,x;\hat{\boldsymbol\theta})$$

Conditional on data, only random quantities are

$$\{Z_j;\ j=1,2,\ldots,n\}$$

By the multiplier central limit theorem, the dsitribution of

$C = \sum_j \{ \hat{F}(L_j, X_j) - F_{\hat{\boldsymbol{\theta}}}(L_j, X_j) / c(\hat{\boldsymbol{\theta}}) \}^2$ is approximated by

$$\frac{1}{n^2} \sum_i \left[\sum_j Z_j \hat{V}_j (L_i, X_i; \hat{\boldsymbol{\theta}}) \right]^2 = \left\| \frac{\mathbf{Z}' \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})}{n} \right\|^2$$

Step 0: Calculate the statistic $C = \sum_i \{ \hat{F}(L_i, X_i) - F_{\hat{\boldsymbol{\theta}}}(L_i, X_i) / c(\hat{\boldsymbol{\theta}}) \}^2$ and matrix

$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$.

Step 1: Generate $Z_j^{(b)} \sim N(0,1);\ b = 1, 2, \ldots, B,\ j = 1, 2, \ldots, n$.

Step 2: Calculate $\{ C^{(b)}; b = 1, 2, \ldots, B \}$,

where $C^{(b)} = \| (\mathbf{Z}^{(b)})' \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) / n \|$ and $\mathbf{Z}^{(b)} = (Z_1^{(b)}, \ldots, Z_n^{(b)})'$.

Step 3: Reject H_0 at the 100α % significance level if $\sum_{b=1}^{B} \mathbf{I}(C^{(b)} \geq C) / B < \alpha$.

Let $G_n(l, x) = \sqrt{n}\{\hat{F}(l, x) - F_{\hat{\boldsymbol{\theta}}}(l, x)/c(\hat{\boldsymbol{\theta}})\}$ and $G_n^{(b)}(l, x) = n^{-1/2}\sum_j Z_j^{(b)}\hat{V}_j(l, x; \hat{\boldsymbol{\theta}})$

for $b = 1, 2, \ldots, B$.

Theorem 2: *. Under* H_0,

$$(G_n, G_n^{(1)}, ..., G_n^{(B)}) \to (G_{\boldsymbol{\theta}}, G_{\boldsymbol{\theta}}^{(1)}, ..., G_{\boldsymbol{\theta}}^{(B)})$$

in $D\{(-\infty, \infty)^2\}^{\otimes(B+1)}$, *where* $G_{\boldsymbol{\theta}}$ *is the mean zero Gaussian process whose covariance for* $(l, x), (l^*, x^*) \in (-\infty, \infty)^2$ *is given as*

$$\begin{aligned} Cov\{G_{\boldsymbol{\theta}}(l, x), G_{\boldsymbol{\theta}}(l^*, x^*)\} &= E\{V_j(l, x; \boldsymbol{\theta})V_j(l^*, x^*; \boldsymbol{\theta})\} \\ &= F_{\boldsymbol{\theta}}(l \wedge l^*, x \wedge x^*)/c(\boldsymbol{\theta}) - F_{\boldsymbol{\theta}}(l, x)F_{\boldsymbol{\theta}}(l^*, x^*)/c(\boldsymbol{\theta})^2 - \mathbf{g}_{\boldsymbol{\theta}}(l, x)' I^{-1}(\boldsymbol{\theta})\mathbf{g}_{\boldsymbol{\theta}}(l^*, x^*), \end{aligned}$$

where $a \wedge b \equiv \min(a, b)$, *and* $G_{\boldsymbol{\theta}}^{(1)}, ..., G_{\boldsymbol{\theta}}^{(B)}$ *are independent copies of* $G_{\boldsymbol{\theta}}$.

Simulation under the null

(L^O, X^O) follows the bivariate normal distribution with mean vector and covariance matrix given respectively by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_L \\ \mu_X \end{bmatrix} = \begin{bmatrix} 120\text{-}62.63 \\ 60.82 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_L^2 & \sigma_{LX} \\ \sigma_{LX} & \sigma_X^2 \end{bmatrix} = \begin{bmatrix} 19.64^2 & 19.64 \times 16.81 \rho_{LX} \\ 19.64 \times 16.81 \rho_{LX} & 16.81^2 \end{bmatrix},$$

where $\rho_{LX} = 0.70$, 0.35, 0, -0.35, or -0.70. In this design, the population parameters of Japanese test scores (mean=60.82, SD=16.81) and English test scores (mean=62.63 SD=19.64) are determined by the record for National Center Test for University for 2008 in Japan.

Compare the sample mean and standard deviation (SD) of $\{C^{(b)}; b = 1, 2, \ldots, 1000\}$ based on multiplier and parametric Bootstrap

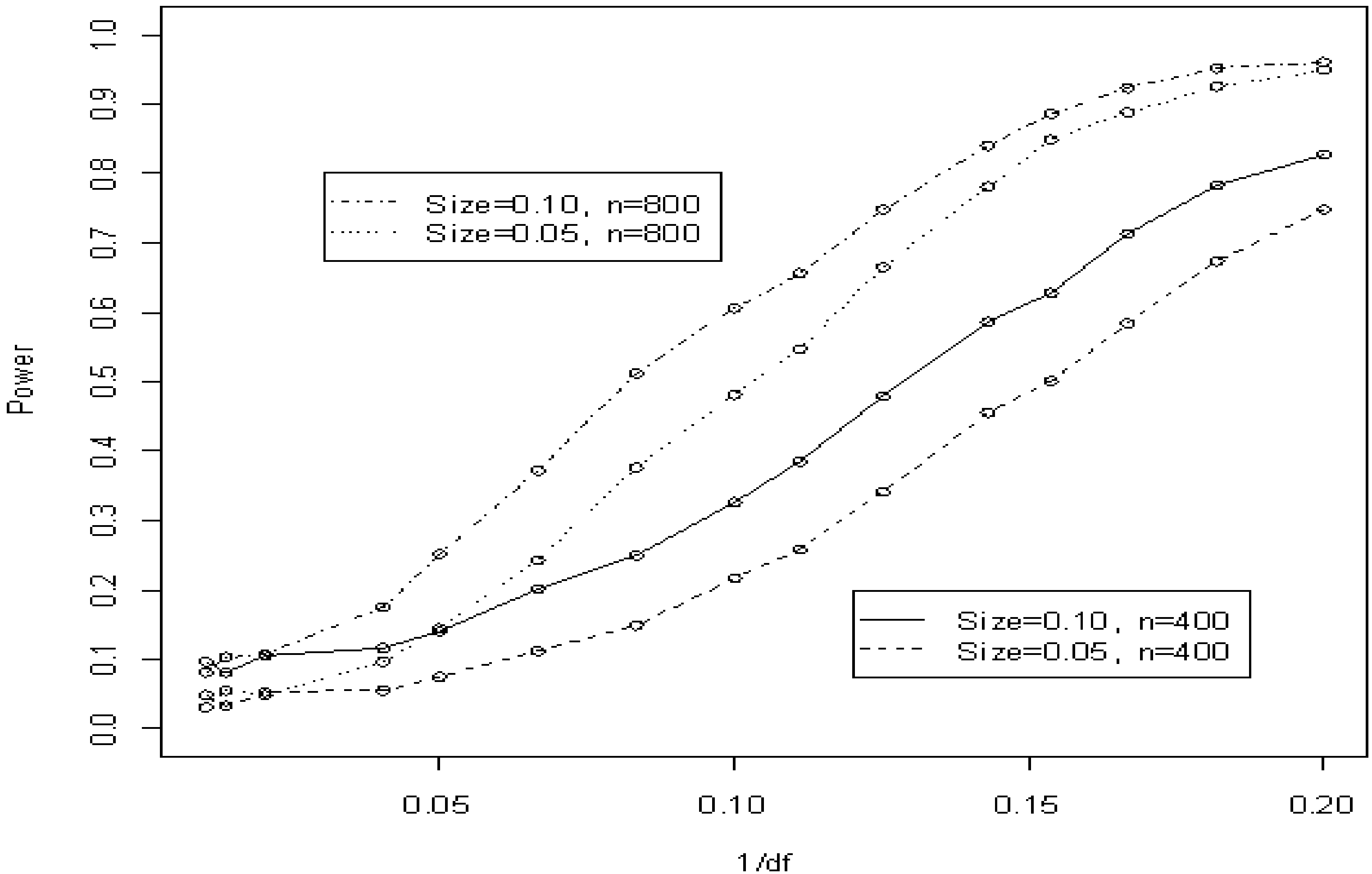
Table 3: A comparison of the multiplier method and parametric Bootstrap: $n = 400$

		Elapsed time (in second)		Resampling mean and SD (in parenthesis)	
		Multiplier	Parametric Bootstrap	Multiplier	Parametric Bootstrap
Bivariate Normal					
	$\rho_{X^oY^o}$ =0.70	23.17	1402.31	0.0621 (0.0314)	0.0603 (0.0304)
	$\rho_{X^oY^o}$ =0.35	22.84	1463.01	0.0606 (0.0269)	0.0581 (0.0252)
	$\rho_{X^oY^o}$ =0.00	23.04	1539.29	0.0597 (0.0219)	0.0586 (0.0225)
	$\rho_{X^oY^o}$ =-0.35	23.37	1380.09	0.0598 (0.0234)	0.0577 (0.0209)
	$\rho_{X^oY^o}$ =-0.70	26.77	1296.32	0.0591 (0.0225)	0.0569 (0.0217)
Bivariate Poisson					
$l_L = 1$,	$\rho_{X^oY^o}$ =0.70	12.48	10563.95	0.0608 (0.0413)	0.0614 (0.0426)
$l_X = 1$	$\rho_{X^oY^o}$ =0.35	12.40	12265.27	0.0505 (0.0353)	0.0543 (0.0404)
$l_L = 1$,	$\rho_{X^oY^o}$ =0.70	12.41	10632.34	0.0703 (0.0381)	0.0703 (0.0380)
$_X = 2$	$\rho_{X^oY^o}$ =0.35	12.62	13138.48	0.0676 (0.0395)	0.0692 (0.0413)

NOTE: Elapsed time is calculated by proc.time( ) in R. The number of resamplings is 1,000 for both multiplier and parametric Bootstrap methods.

Power study

we generated data from the bivariate *t*-distribution (Lang et al., 1989) while we performed the goodness-of-fit test under the null hypothesis of the bivariate normal distribution.



Data analysis

- Data : the average scores of LSAT (the national law test) and average GPA (graduate point average) for N=82 American law schools

$$(\mathrm{LSAT}_j, \mathrm{GPA}_j) \quad \text{for} \quad j = 1, 2, ..., N$$

Use only $(\mathrm{LSAT}_j, \mathrm{GPA}_j)$ whose sum of LSAT and $100 \times \mathrm{GPA}$ are

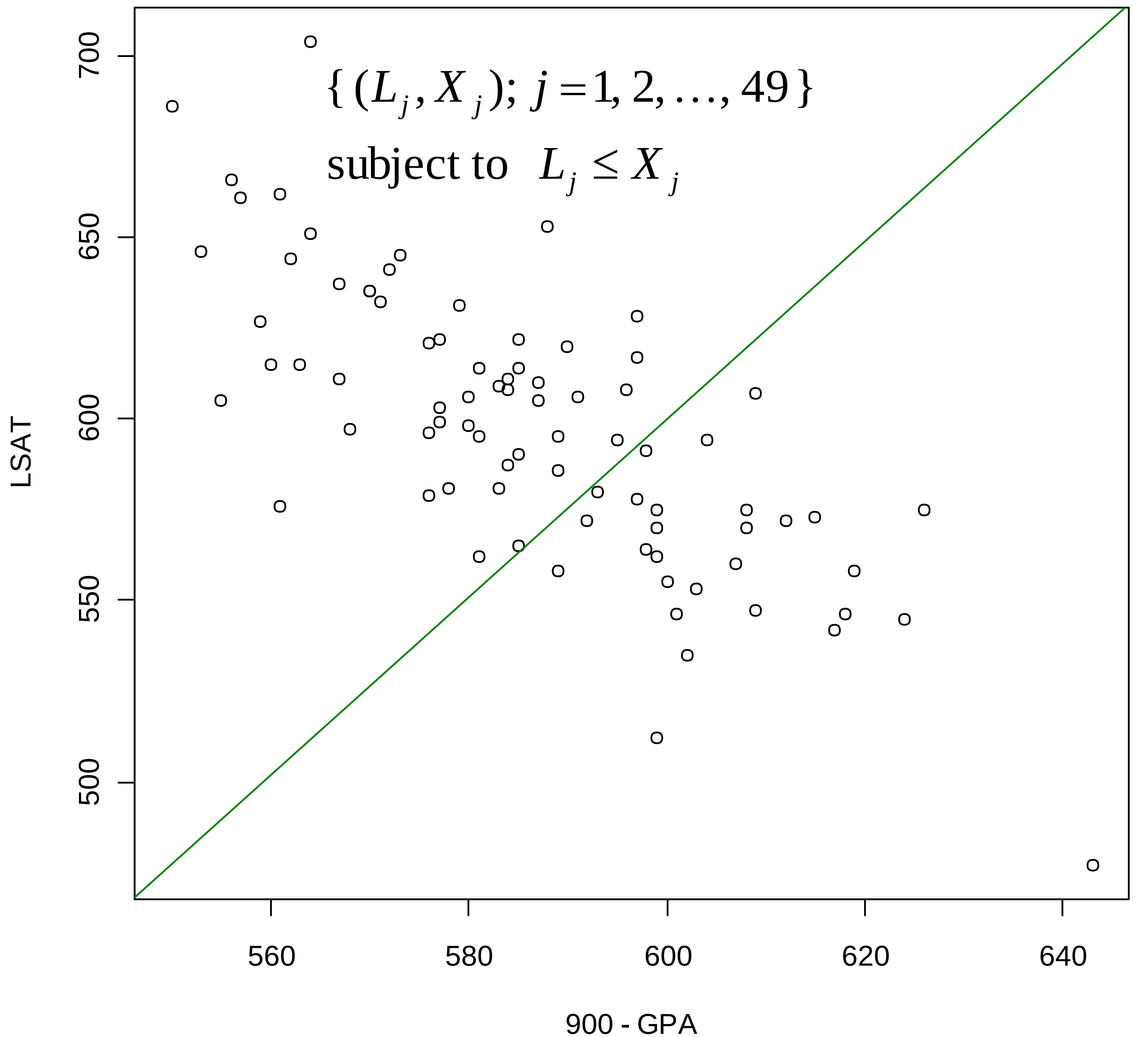
above a threshold $600 + 100 \times 3.0 = 900$.

Observe $\{(L_j, X_j);\ j = 1, 2, \ldots, 49\}$,

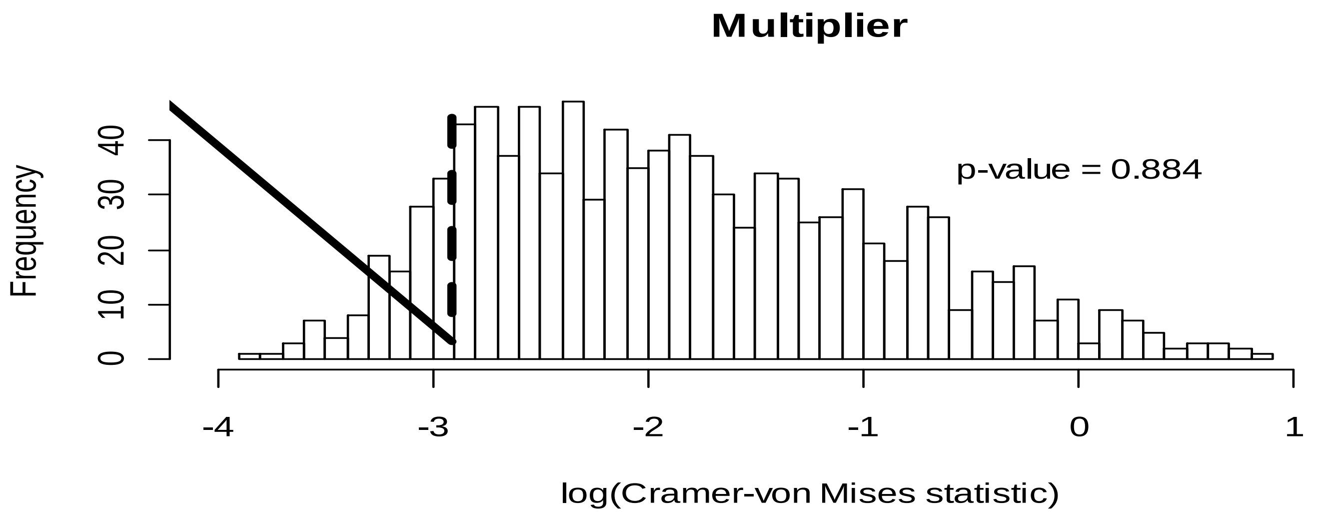
subject to $L_j \le X_j$, where $L_j = 900 - 100 \times \mathrm{GPA}_j$ and $X_j = \mathrm{LSAT}_j$.

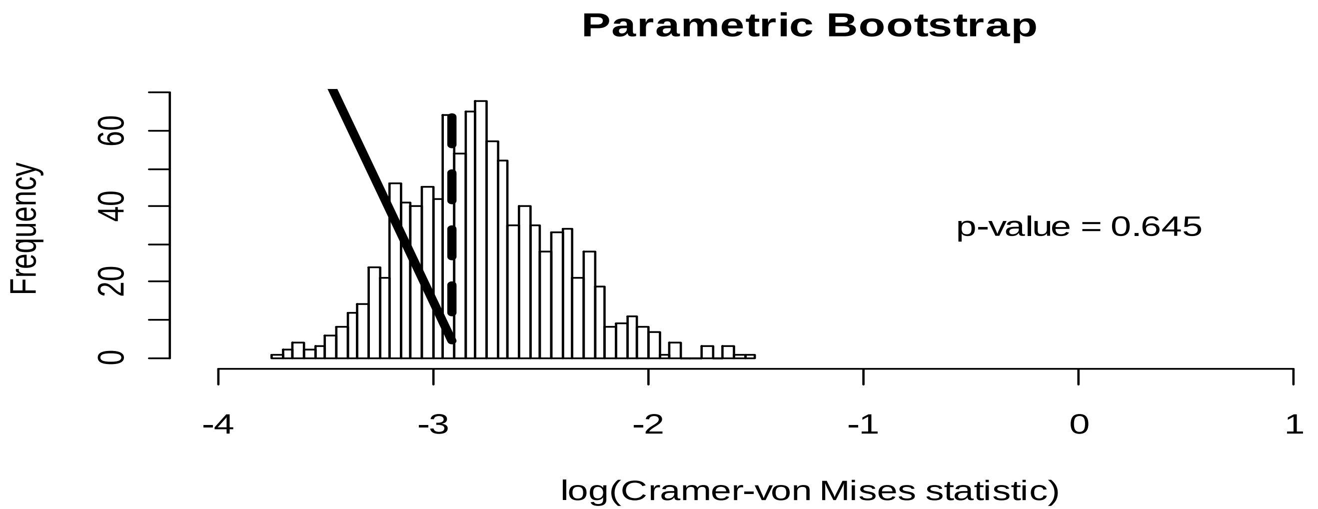
- Based only on observed data, we estimate the average LSAT score of N=82 schools

$$\{(900 - GPA_j, LSAT_j);\ j = 1, \ldots, 82\}$$



Computational time for the multiplier method = 1.25 second
Computational time for the parametric Bootstrap = 222.46 second





Summary

Why simulations based on the multiplier reduce computational cost ?

- The multiplier method involves only arithmetic operations (multiply, sum). On the other hand, the parametric Bootstrap involves nonlinear maximization to get $\hat{\boldsymbol{\theta}}^{(b)}$
- Re-sampling from a truncated parametric model is involved:

Accept-reject algorithm

(i) data (L, X) from the distribution function $F_{\hat{\boldsymbol{\theta}}}(l, x)$ is generated;

(ii) if $L \le X$, we accept the sample and set $(L_j^{(b)}, X_j^{(b)}) = (L, X)$;

otherwise we reject (L, X) and return to (i).

On the other hand, the multiplier only requires generating i.i.d. sequences

Thank you for your kind attention