Limited Dependent Variables : Part II

Truncated and Censored Samples

$$y_i = x_i'\beta + \epsilon_i \qquad i = 1, \ldots, N$$

In usual case, we have all observations on y_i & x_i with no irregularities

- Censoring occurs when we observe x_i for the entire sample but for some observations we only have limited information about the dependent variable

Example (Tobin's study of household expenditures)

Consumer maximizes utility by purchasing durable goods under constraint that total expenditures do not exceed income

expenditure on durables $\geq$ cost of least expensive durable good

If available income $<$ least expensive durable good
no expenditure is observed

Don't know how much a household would have spent if a durable good could be purchased for less than the least expensive item.

y_i = observed consumption of durables

t_i = least expensive durable good

y_i^* = optimal consumption

$y_i = y_i^*$ if $y_i^* >$

Example (Lang's study of Prestige of 1st academic job of Biochemists)

y_i = prestige of 1st academic job

Prestige is measured on a continuous scale from 1 – 5 with ~~schools~~ schools from

1 – 1.99 = adequate

2 – 2.99 = good

3 – 3.99 = strong

3.99 + = distinguished

Prestige is not ~~measured for~~ coded as 1 for

(1) Graduate programs (2) rated below adequate

Truncation excludes all observations based on characteristics of the dependent variable

Example (Hausman & Wise's 1977 analysis of New Jersey negative income tax experiment)

Goal: Estimate earning functions for low income individuals

Individuals with ~~income~~ earnings less that 1.5 * poverty level were excluded from sample

Inference problems

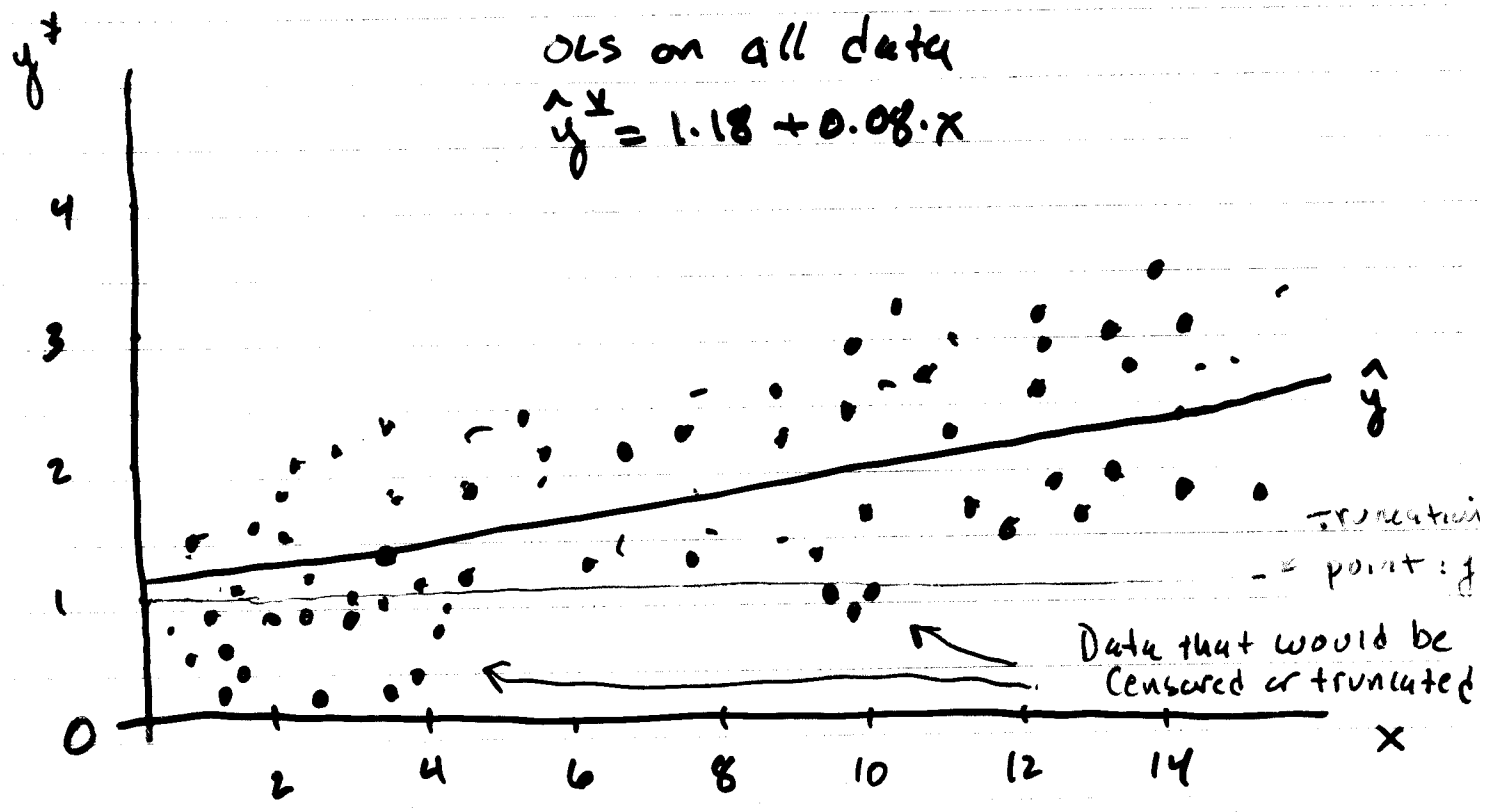
- Inference about entire population in presence of censoring & truncation
- Inference about subpopulation

Results

OLS has severe problems

MLE ~~gain~~ is appropriate alternative.

True Regression: $y^* = 1.2 + 0.08x + \epsilon$

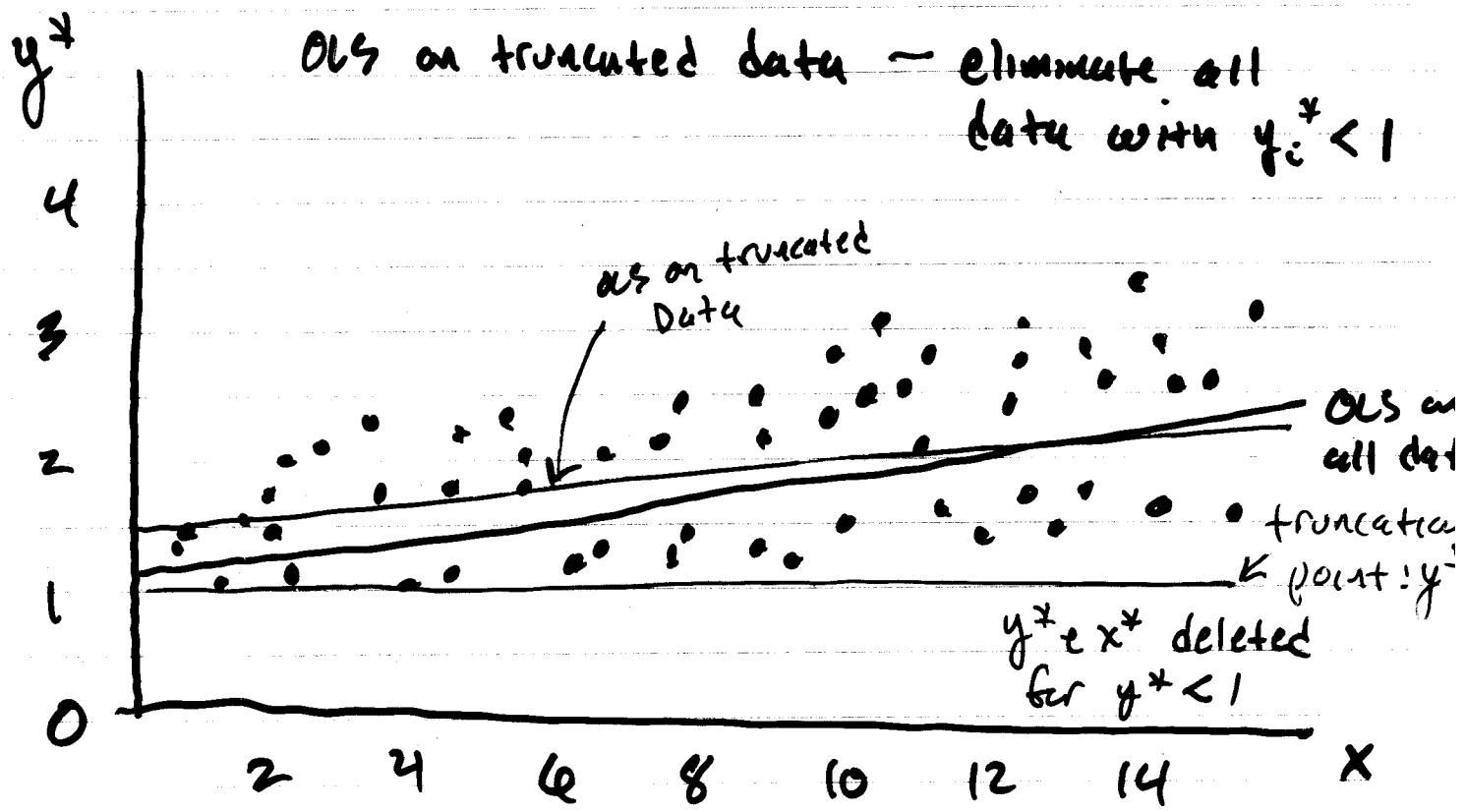


Censored Data: $y_i = y_i^* \quad \text{if } y_i^* > 1$

$= 0 \quad \text{if } y_i^* \leq 1$

Truncated Data: Only observe $y_i = y_i^*$ and x_i if $y_i^* > 1$

Truncated Regression: $\hat{y}^* = 1.41 + 0.06kx$



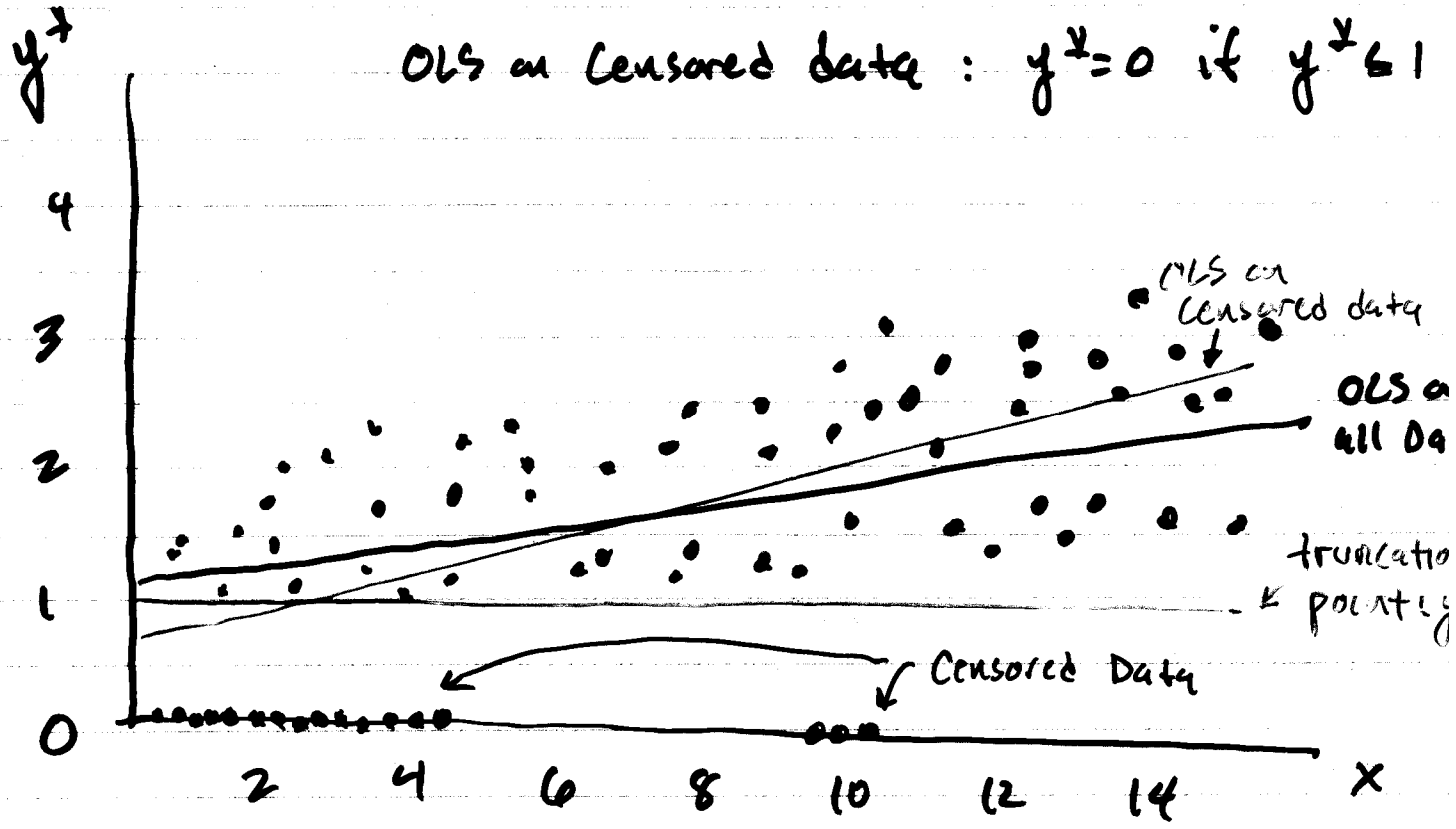
- OLS on truncated Data is biased and inconsistent
- Truncation causes correlation between X and ϵ in general

Here, observations with large negative errors have been deleted

intercept is overestimated

slope is underestimated

Censored Regression : $\hat{y}^* = 0.95 + 0.11x$



- OLS on censored Data creates biased and inconsistent estimates of intercept & slope

Here

Intercept is underestimated

Slope is over estimated

Truncated Regression Model

{ truncation from below

$$y_i = \beta' x_i + \epsilon_i$$

where we observe x_i and y_i only if $y_i > a$

$$\epsilon_i \sim \text{iid } N(0, \sigma^2)$$

$$\Rightarrow y_i \mid x_i \sim N(\beta' x_i, \sigma^2)$$

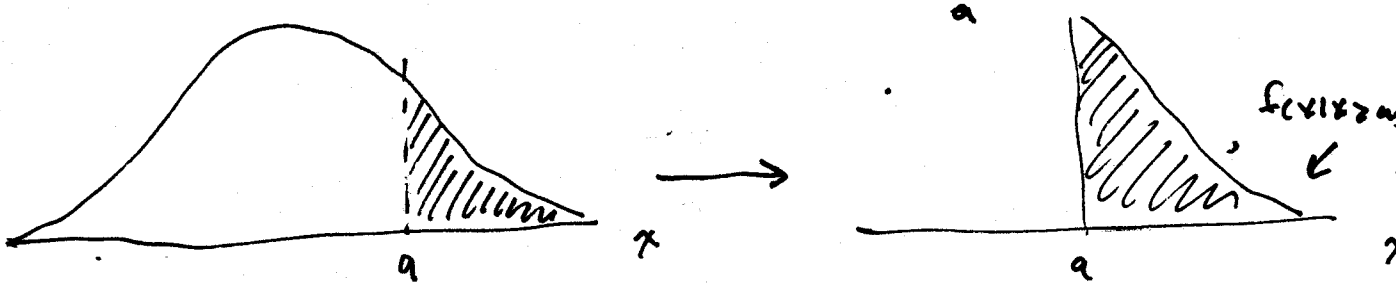
Q: What is the pdf of $y_i \mid x_i, y_i > a$?

In order to MLE we don't use the unconditional pdf of y_i, we must use the pdf conditional on $y_i > a$

Truncated Distributions

Thm (Greene pg 949 (4th)). Let x be a cts rv with density $f(x)$ and let a be a constant. Then

$$f(x \mid x > a) = \frac{f(x)}{P(x > a)} = \frac{f(x)}{\int_a^\infty f(x)\,dx} \qquad x > a.$$



Truncated Normal Distn $(TN(\mu, \sigma^2, a))$ ← a = truncation point

Let $X \sim N(\mu, \sigma^2)$. Then

$$f(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$= \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ = pdf for Std. Normal

Further,

$$\Pr(X > a) = \Pr\left(\frac{x-\mu}{\sigma} > \frac{a-\mu}{\sigma}\right)$$

$$= \Pr\left(z > \frac{a-\mu}{\sigma}\right), \quad z \sim N(0,1)$$

$$= 1 - \Pr\left(z \le \frac{a-\mu}{\sigma}\right)$$

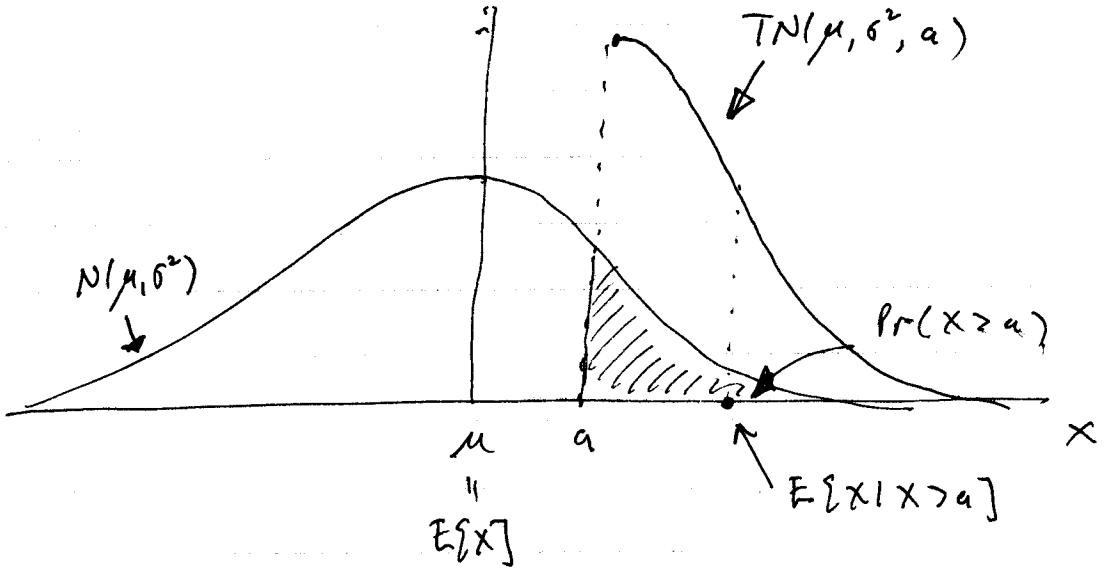
$$= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

where $\Phi(z)$ = CDF of $N(0,1)$ evaluated at z.

Result: $f(x \mid x > a, \mu, \sigma^2) = \dfrac{f(x; \mu, \sigma^2)}{\Pr(X > a)}$

$$= \frac{\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

Note: $\int_a^{\infty} f(x \mid x > a, \mu, \sigma^2)\, dx = 1$



Remarks

(1) The mean of the truncated distribution is always larger than the mean of the untruncated distribution

(2) The variance of the truncated distn is always smaller than the untruncated distribution

Moments of a truncated Normal

Let $X \sim N(\mu, \sigma^2)$ and let a be a constant

Then

$$(i)\quad E[X \mid X > a] = \mu + \sigma \cdot \lambda\left(\frac{a-\mu}{\sigma}\right)$$

$$(ii)\quad \text{var}(X \mid X > a) = \sigma^2\left(1 - \delta\left(\frac{a-\mu}{\sigma}\right)\right)$$

where

$$\lambda(x) = \frac{\varphi(x)}{1 - \Phi(x)} = \text{inverse mills ratio}$$

$$= \text{Normal hazard function}$$

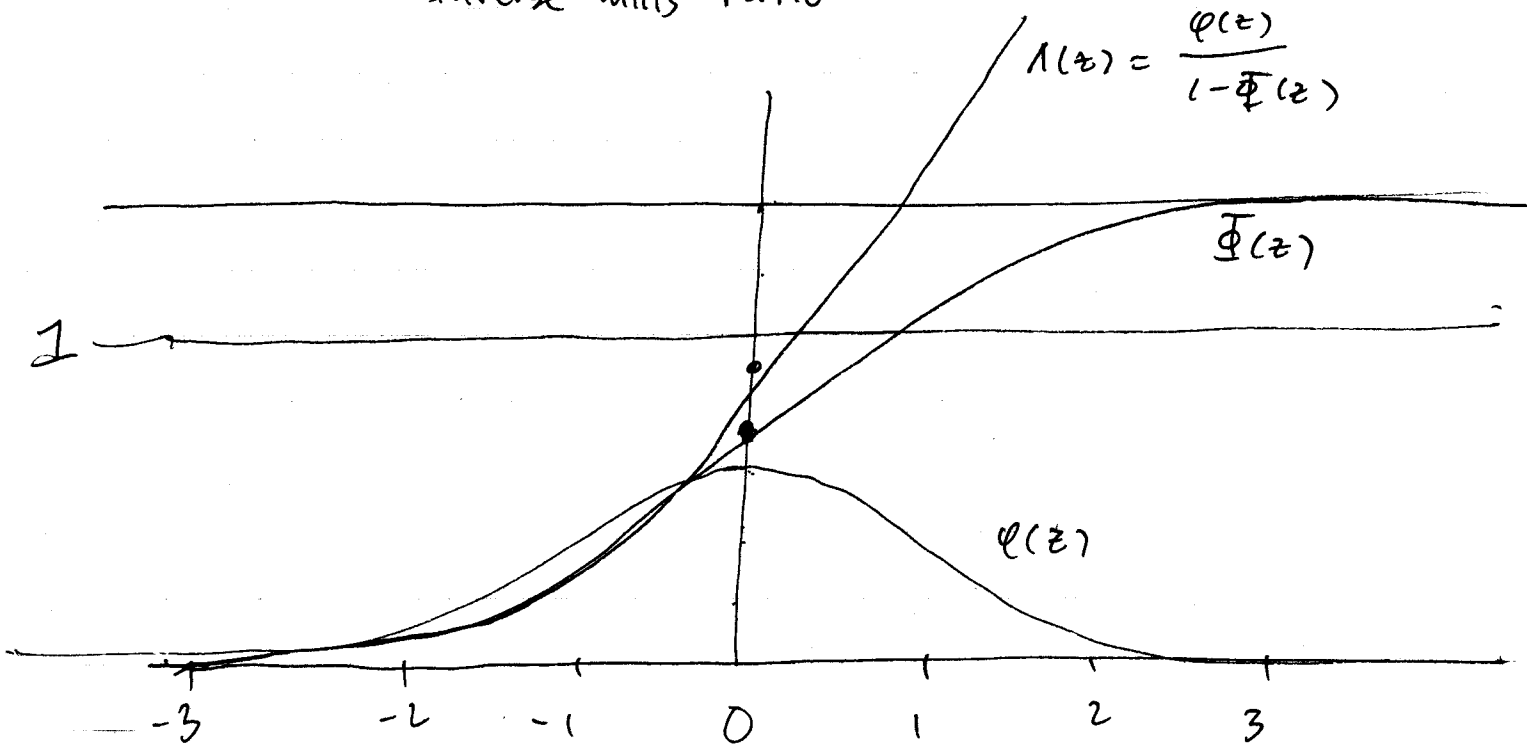
$$\delta(x) = \lambda(x)\left(\lambda(x) - x\right), \quad 0 < \delta(x) < 1$$

Note: prove using brute force manipulations; e.g.,

$$E[X \mid X > a] = \int_a^\infty x f(x \mid x > a, \mu, \sigma^2)\, dx$$

$$= \int_a^\infty x \frac{\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)} \, dx$$

Relationship b/w degree of truncation & Inverse Mills Ratio



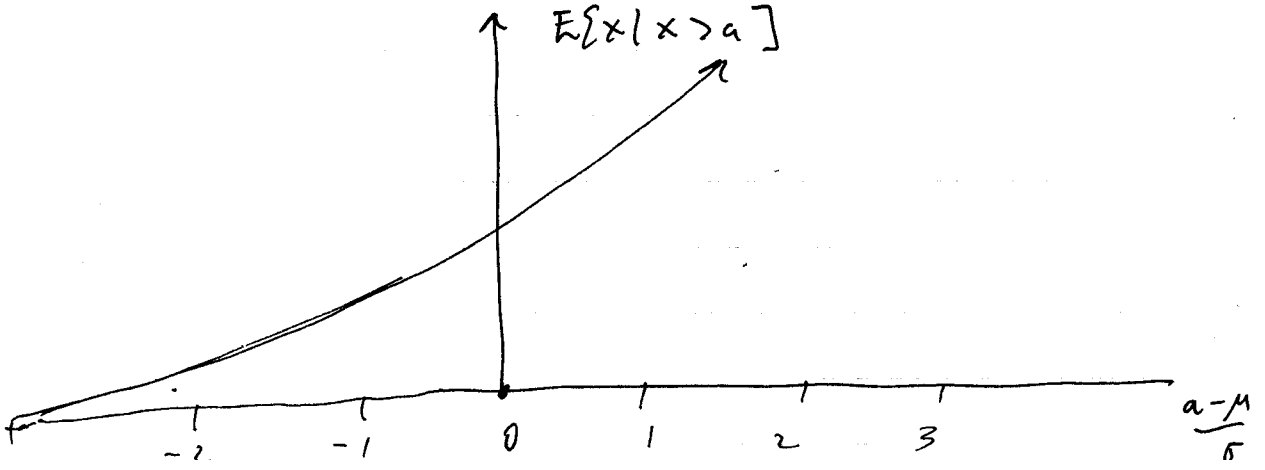
less truncation

more truncation

$$z = \frac{a-\mu}{\sigma}$$

= standardized truncation point.

⇒ Inverse Mills ratio increases with the amount of truncation!



The truncated Regression Model

$$y_i = \underset{1\times K}{x_i'} \underset{K\times 1}{\beta} + \epsilon_i \quad , \epsilon_i \sim iid\ N(0, \sigma^2)$$

x_i is treated as fixed

$$\Rightarrow \quad y_i | x_i \sim N(x_i'\beta, \sigma^2)$$

Truncation: only observe y_i, x_i for $y_i > a$
where a is some constant

Using the results Now

- $E[y_i | x_i] = x_i'\beta$

$$\frac{\partial E[y_i | x_i]}{\partial x_i} = \beta \quad = \text{marginal effect for entire population}$$

Using the results on truncated distributions

$$E[y_i | x_i, y_i > a] = x_i'\beta + \sigma * \lambda\left(\frac{a - x_i'\beta}{\sigma}\right)$$

and the marginal effects for the subpopulation are

$$\frac{\partial E[y_i | x_i, y_i > a]}{\partial x_i} = \beta + \sigma * \frac{d\lambda\left(\frac{a - x_i'\beta}{\sigma}\right)}{dx_i}$$

which, by the chain rule, reduces to

$$= \beta\left(1 - \delta\left(\frac{a - x_i'\beta}{\sigma}\right)\right)$$

$$< \beta$$

since $\delta(x)$ lies b/w 0 and 1.

Maximum Likelihood Estimation

$y_i | x_i, y_i > a \sim$ truncated Normal so that

$$f(y_i; x_i, y_i > a, \beta, \sigma^2) = \frac{\frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right)}$$

Given an iid sample, the joint density is

$$L(\beta, \sigma^2 | y, X) = \prod_{i=1}^{n} \frac{\frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right)}$$

and the log-likelihood is

$$\ln L(\beta, \sigma^2 | y, X) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{1}^{n}(y_i - x_i'\beta)^2 - \sum_{i=1}^{n}\ln\left(1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right)\right)$$

Example: Hausman & Wise (1977) "Social Experimentation, truncated distributions, and efficient Estimation", Econometrica

Earnings regression $y_i = x_i'\beta + \epsilon_i$, ϵ_i iid $N(0,\sigma^2)$, $i = 1,\ldots,684$

y_i: log earnings

Truncation: y_i, x_i observed if $y_i < 1.5 *$ poverty level = \$5002

Variables	$\hat{\beta}_{OLS}$	$\hat{\beta}_{MLE}$	$2E[y_i \mid y_i < 5002]/\sigma$
Const	8.20	9.10	
	(0.09)	(0.03)	
Education	0.010	0.015	0.0036
	(0.006)	(0.007)	
IQ	0.002	0.006	0.0014
	(0.002)	(0.005)	
Training	0.002	0.007	0.0014
	(0.002)	(0.003)	
Union	0.090	0.246	
	(0.031)	(0.090)	
Illness	-0.076	-0.226	
	(0.040)	(0.104)	
Age	-0.003	-0.016	-0.0039
	(0.002)	(0.005)	

Remark

The bias e inconsistency of OLS in the truncated regression model can be explained in terms of omitted variables bias. For the truncated regression model, the regression function takes the form

$$y_i = E\{y_i \mid x_i, y_i > a\} + \epsilon_i$$

where $\epsilon_i = y_i - E\{y_i \mid x_i, y_i > a\}$ is the regression error. Using the results on the mean of a truncated normal r.v.

inverse mills ratio

$$E\{y_i \mid x_i, y_i > a\} = x_i'\beta + \sigma \cdot \lambda\left(\frac{a - x_i'\beta}{\sigma}\right)$$

So that the correct regression function is

$$y_i = x_i'\beta + \sigma \cdot \lambda\left(\frac{a - x_i'\beta}{\sigma}\right) + \epsilon_i$$

Standard OLS estimates

$$y_i = x_i'\beta + u_i$$

and which omits the term $\sigma \cdot \lambda\left(\frac{a - x_i'\beta}{\sigma}\right)$

So that we may think of u_i as

$$u_i = \sigma \cdot \lambda\left(\frac{a - x_i'\beta}{\sigma}\right) + \epsilon_i$$

Consequently, x_i in the OLS regression is correlated with u_i $\left(x_i \text{ is correlated with } \lambda\left(\frac{a - x_i'\beta}{\sigma}\right)\right)$ due to the correlation between x_i and the omitted variable $\lambda\left(\frac{a - x_i'\beta}{\sigma}\right)$.