Generalized Linear Models

Lecture 12. Censored and truncated models

Outline

- Censored/truncated count data models
- Survival models
- Tobit model

Censored/truncated count data models

- **Truncation** – some of the counts are completely omitted
  Usually truncated from left, analogous to zero-truncation, resulting model is a conditional model
- **Censoring** – the counts are restricted
  Censoring is considered as a *cutpoint*, censored observations are considered as potential outcomes

Log-likelihood generally consists of three parts:
left-censored part + uncensored part + right-censored part

Underlying distribution can be either Poisson or NB

Right-censoring

Typical situation in count data: corresponds to *'... or more'*

Instead of actual counts Y^* we consider **censored counts** Y

We assume that certain indicator C determines the observation process:

$$C := \mathbf{I}_{\{Y^* \le a\}} = \begin{cases} 1, & \text{if } Y^* \in \{0, \ldots, a\} \\ 0, & \text{if } Y^* \in \{a+1, a+2, \ldots\} \end{cases}$$

for some positive integer a

Censored variable $Y = \min(Y^*, a+1)$

$$\mathbf{P}\{C = 1\} = \mathbf{P}\{Y^* \le a\} = F_{Y^*}(a), \quad F_{Y^*}(a) = \sum_{j=0}^{a} p_{Y^*}(j)$$

$$\mathbf{P}\{C = 0\} = \mathbf{P}\{Y^* > a\} = 1 - F_{Y^*}(a)$$

Pmf of censored counts

$$p_Y(y) = \begin{cases} p_{Y^*}(y), & \text{if } \ y = 0, 1, \ldots, a \\ 1 - F_{Y^*}(a), & \text{if } \ y = a+1 \end{cases}$$

Models with right-censored data

Pmf of a censored count variable can be written as

$$p_Y(y) = C \cdot p_{Y^*}(y) + (1 - C)(1 - F_{Y^*}(a)) = C \cdot p_{Y^*}(y) + (1 - C)[1 - \sum_{j=0}^{a} p_{Y^*}(y)]$$

It has been proved that censored ML estimator is consistent and asymptotically normally distributed if the pmf $p_{Y^*}(y)$ is correctly determined

If the actual count Y^* is Poisson distributed, i.e. $p_{Y^*}(y; \mu) = \frac{\exp(-\mu)\mu^y}{y!}$, the censored pmf has the following form:

$$p_Y(y; \mu) = C\frac{\exp(-\mu)\mu^y}{y!} + (1 - C)[1 - \sum_{j=0}^{a} \frac{\exp(-\mu)\mu^j}{j!}]$$

One can also think of zero-modified censored models, e.g. right-censored zero-inflated Poisson model (CZIP)

Example. Right-censoring + Zero-inflation

Fish caught at a state park

Data on 250 groups that went to a park. Each group was questioned about how many fish they caught (`count`, 0–149, many zeros), how many children were in the group (`child`), how many people were in the group (`persons`), and whether or not they brought a camper to the park (`camper`).

- Applied models: 5 different CZIP (censored zero-inflated Poisson) models
- Censoring at $a = 4(18\%), 7(10\%), 10(7.2\%), 13(6.4\%), 16(4.8\%)$, i.e. the count is at least a
- The form of models:

$$\hat{\mu} = \exp(b_0 + b_1\texttt{camper} + b_2\texttt{persons} + b_3\texttt{child}), \quad \ln\frac{\hat{\pi}}{1-\hat{\pi}} = a_0 + a_1\texttt{child}$$

- Results: $a_1 > 0$, $b_3 < 0$; $b_1, b_2 > 0$
- Conclusion: increasing the number of censored values will increase the standard errors, but BIC decreases, i.e. the model is better

Source: Saffari, S.E., Adnan, R. Zero-Inflated Poisson regression models with right censored count data, *Matematika, UTM*, vol 27 (1), 21– 29

Left-truncation (zero-truncation)

Some of the counting results are omitted

Zero-truncation based on conditional distribution:

$$\mathbf{P}\{Y = y | Y > 0\} = \frac{\mathbf{P}\{Y = y\}}{1 - \mathbf{P}\{Y = 0\}}$$

Consider the Poisson distribution

$$p(y; \mu) = \frac{\exp(-\mu)\mu^y}{y!} = \frac{\mu^y}{\exp(\mu)y!}$$

Probability of zero $\exp(-\mu)$
Probability of non-zero $1 - \exp(-\mu)$

ZTP model (zero-truncated conditional model)

$$p(y; \mu | Y > 0) = \frac{\exp(-\mu)\mu^y}{(1 - \exp(-\mu))y!}$$

Left-truncation (further truncation)

Left-truncation (truncation from below) at $d > 0$:

$$\mathbf{P}\{Y = y | Y > d\} = \frac{\mathbf{P}\{Y = y\}}{1 - \mathbf{P}\{Y \leq d - 1\}}$$

Left-truncated Poisson model

$Y \in \{d, d+1, \ldots\}$,

$$\mathbf{P}\{Y = y | Y > d\} = \frac{\mu^y}{y![\exp(\mu) - \sum_{j=0}^{d-1} \mu^j / j!]}$$

In case of left-truncation $E(Y|Y > d) = EY + \delta,\ \delta > 0$, i.e. mean is greater than in non-truncated case

Right-truncation

Right-truncation (truncation from above) at c:

$$\mathbf{P}\{Y = y | Y \leq c\} = \frac{\mathbf{P}\{Y = y\}}{\mathbf{P}\{Y \leq c\}}$$

Right-truncated Poisson model

$Y \in \{0, 1, \ldots, c\}$,

$$\mathbf{P}\{Y = y | Y \leq c\} = \frac{\mu^y}{y! \sum_{j=0}^{c} \mu^j / j!}$$

In case of right-truncation $E(Y|Y \leq c) = EY + \delta^*,\ \delta^* < 0$, i.e. mean is smaller than in non-truncated case

Censoring vs truncation

Censoring and truncation are related but different concepts

Censoring

- Left: if b is the smallest observed value, it is not exact and means that $y \leq b$
- Right: if a is the largest observed value, it is not exact and means that $y \geq a$

Truncation

- Left: truncation at d means that only observations $y > d$ are used, smaller values, if they exist, are omitted
- Right: truncation at c means that only observations $y \leq c$ are used, larger values, if they exist, are omitted

Survival/Duration models

Duration data: response variable can be interpreted as **time to event**

Survival/Duration models is a widely used branch of GLM,
different fields use slightly different terminology:

- Demographics – life tables (since Halley's life table 1693), birth rates
- Medicine, biostatistics – survival analysis
- Insurance – hazard analysis, life table analysis
- Economics – duration analysis, transition analysis: labor markets, strike duration, 'survival' of companies, business failure prediction, government changes
- Sociology – event history analysis: length of marriages, analysis of consumer behaviour, recidivism analysis
- Engineering – reliability theory: analysis of reliability of a system (time to failure)
- Queuing theory, waiting time theory – optimization of service times (time until start/end of service)

Survival function

Let us denote

- T – a (continuous) nonnegative rv (lifetime of individuals in a population)
- $F(t) = \mathbf{P}(T \leq t)$ – (cumulative) distribution function (cdf) of T
- $f(t) = F'(t)$ – probability density function (pdf) of T

In survival analysis, the following function plays central role.

Definition (Survival function/survivor function/reliability function)

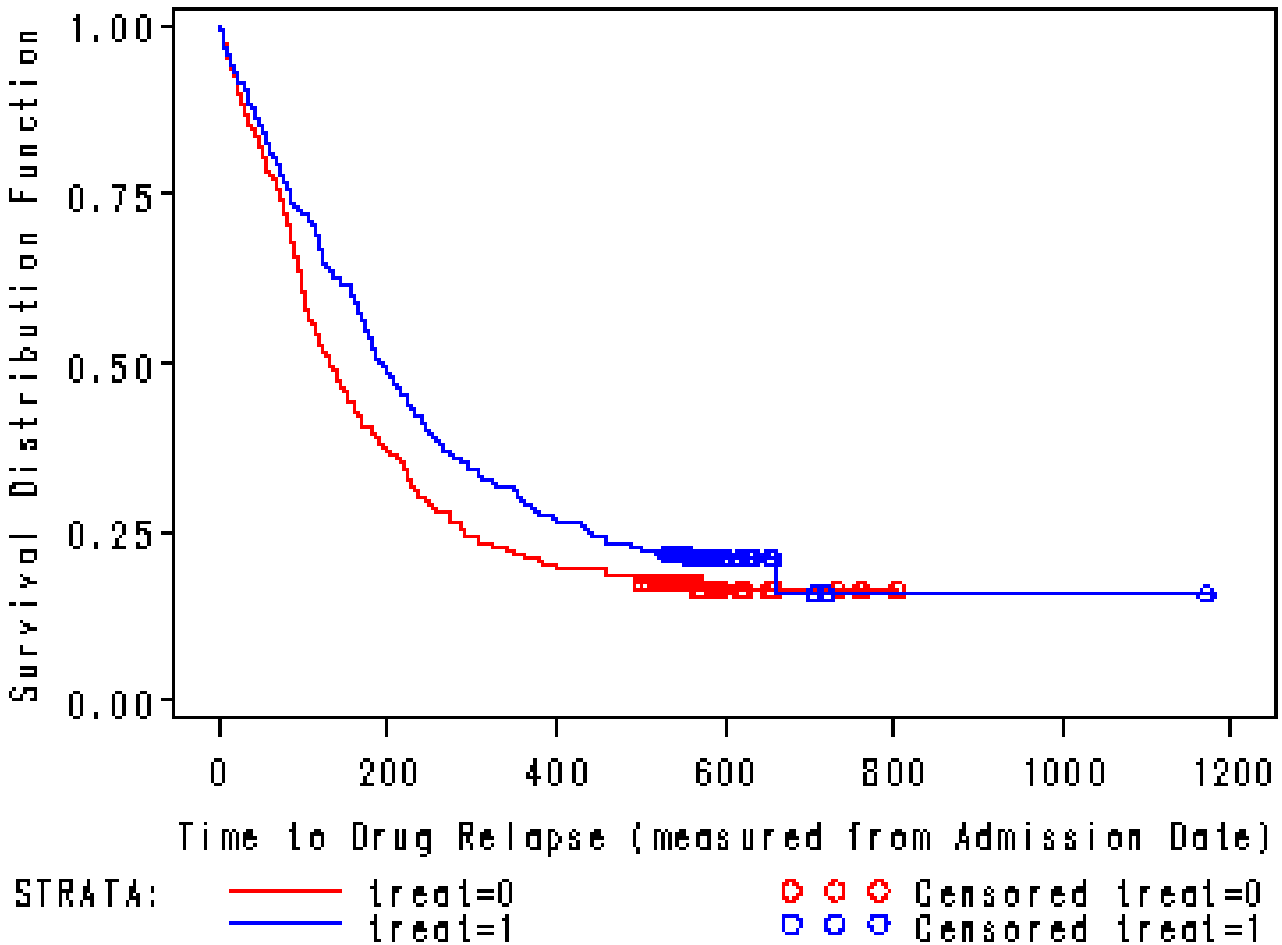
The probability that an individual survives to time t is given by the **survival function**:

$$S(t) = \mathbf{P}(T > t) = 1 - F(t).$$

From the properties of cdf it immediately follows that

- $S(0) = 1$, $S(\infty) = \lim_{t\to\infty} S(t) = 0$
- $S(t)$ is continuous and non-increasing

Example. Survival function (in medicine)



Elukestusfunktsioon

Example. Survival function (Strike duration)

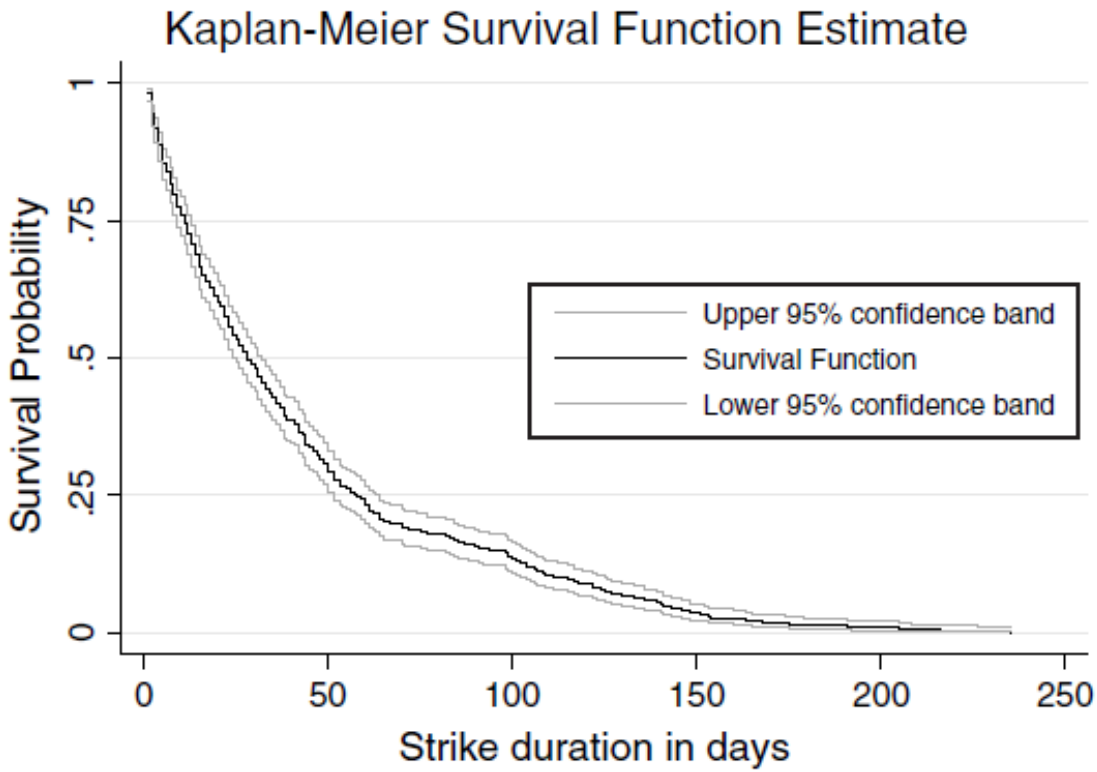


Figure 17.1: Strike duration: Kaplan-Meier estimate of survival function. Data on completed spells for 566 strikes in the U.S. during 1968–76.

Source: Cameron & Trivedi (2005). Microeconometrics Methods and Applications, Cambridge University Press, p. 575

Example. Survival function ('survival' of companies)

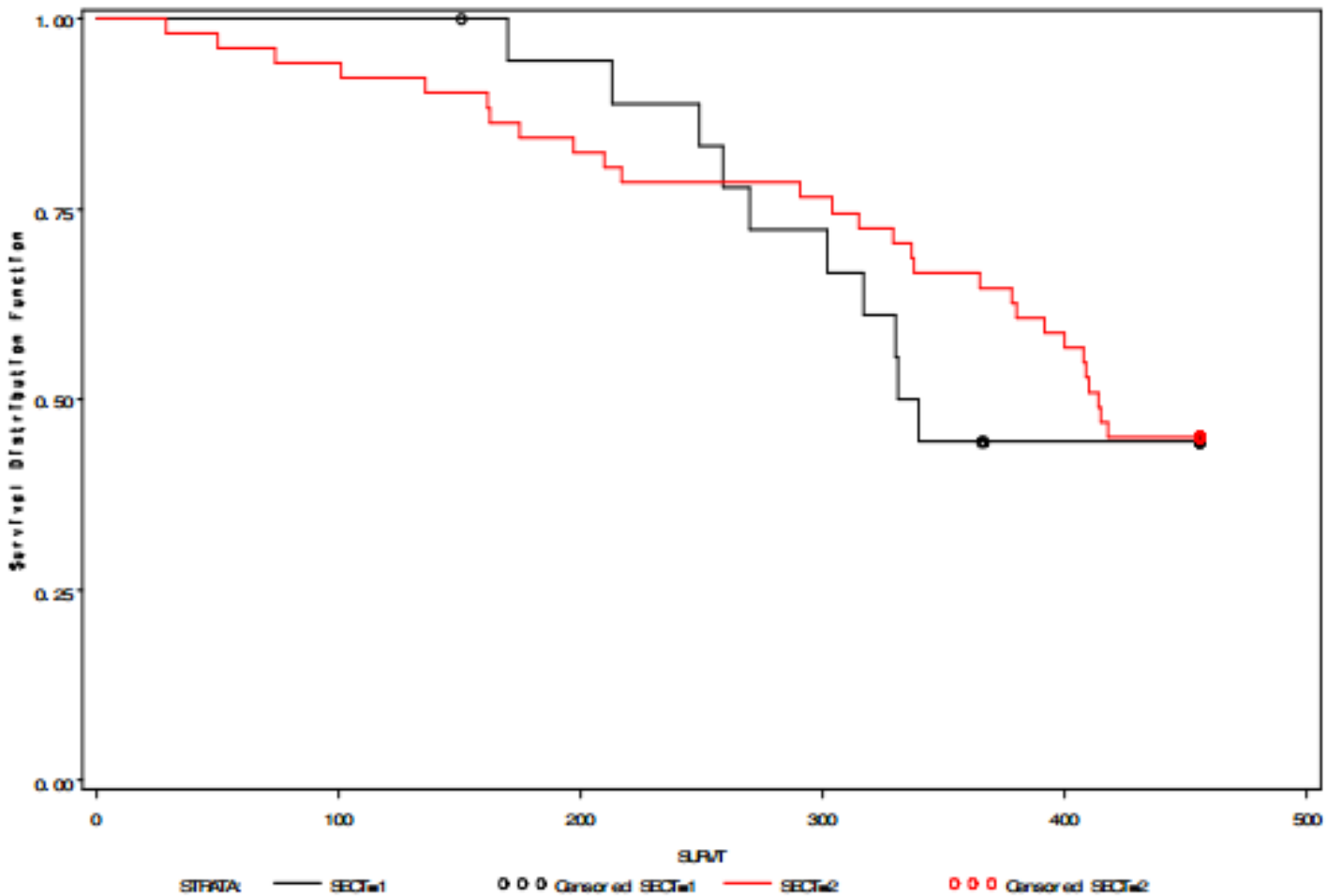


Figure 1. Estimated survivor curves for the two sectors

Source: Oki Kim. Developing Business Failure Models using SAS software. *Business failure prediction*

Hazard function

Definition (Hazard function)

The **hazard function** specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t

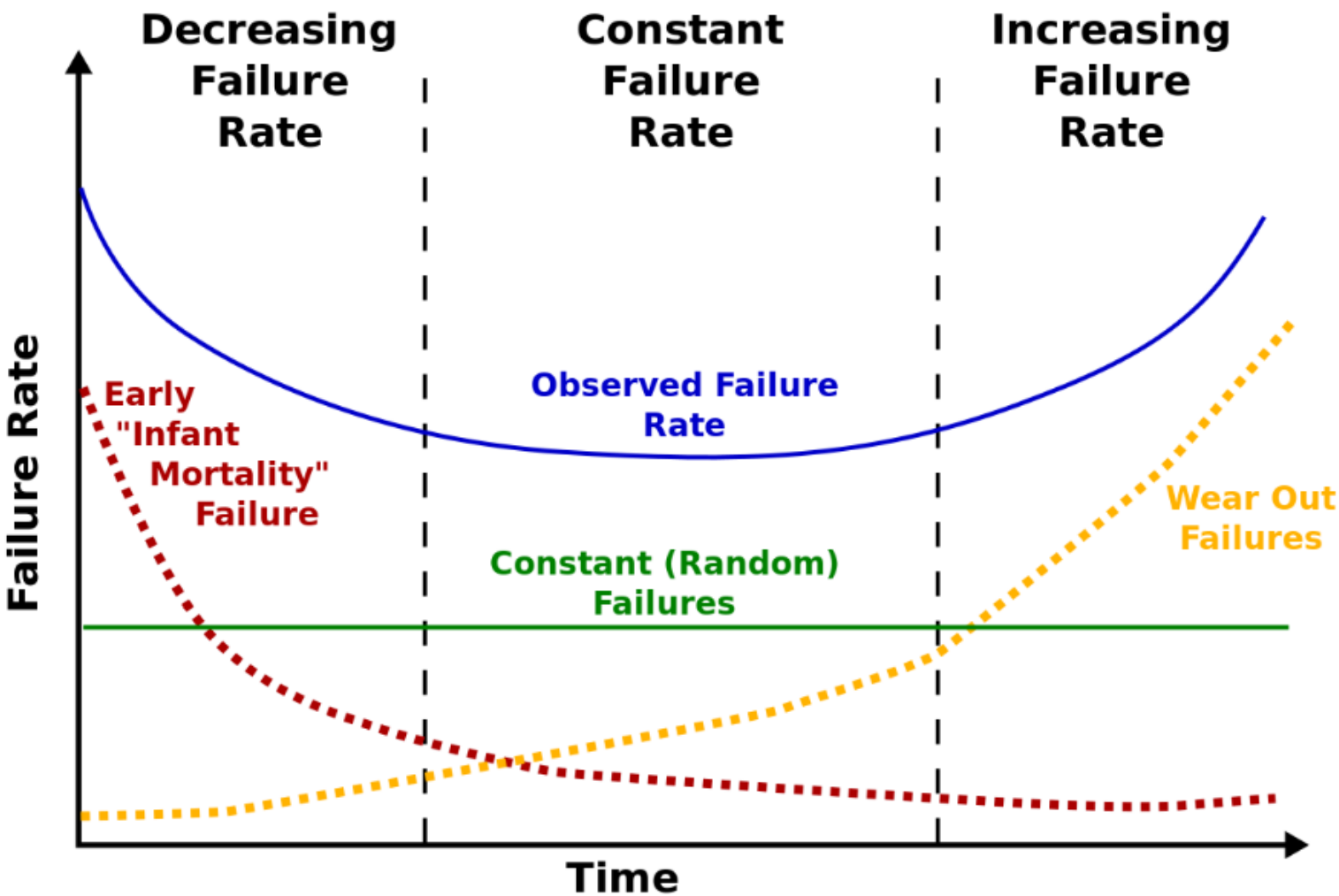
$$h(t) = \lim_{\Delta t \to 0+} \frac{\mathbf{P}\{t \leqslant T < t + \Delta t | T \geqslant t\}}{\Delta t} = \frac{f(t)}{S(t)}$$

In terms of survival function, we can write

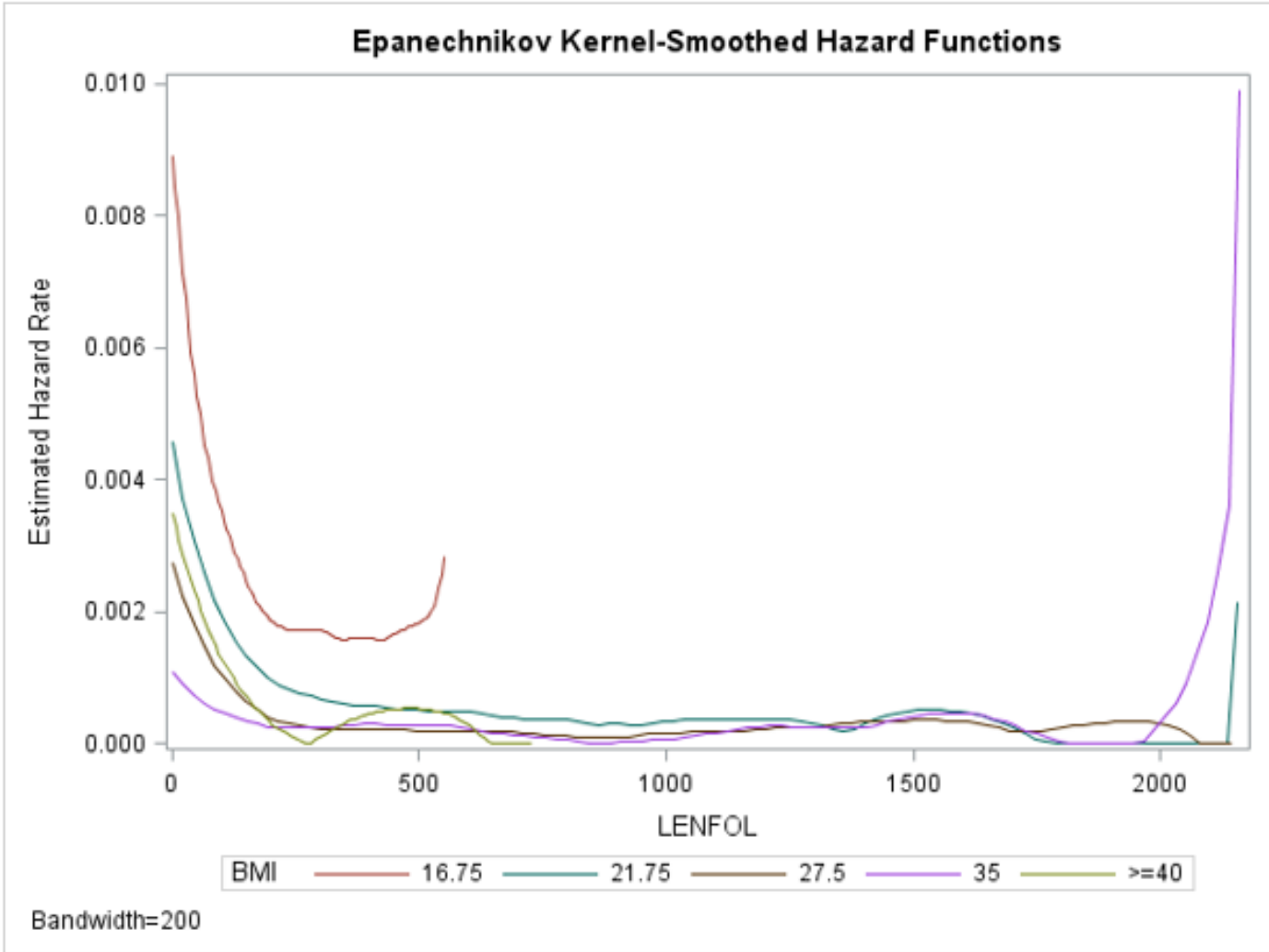
$$h(t) = \lim_{\Delta t \to 0+} \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)}$$

The hazard function is not a density or a probability. We can think of it as the probability of failure in an infinitesimally small time period between t and $t + \Delta t$ given that the subject has survived up till time t. In this sense, the hazard is a measure of risk: the greater the hazard between times t_1 and t_2, the greater the risk of failure in this time interval

Examples of hazard functions

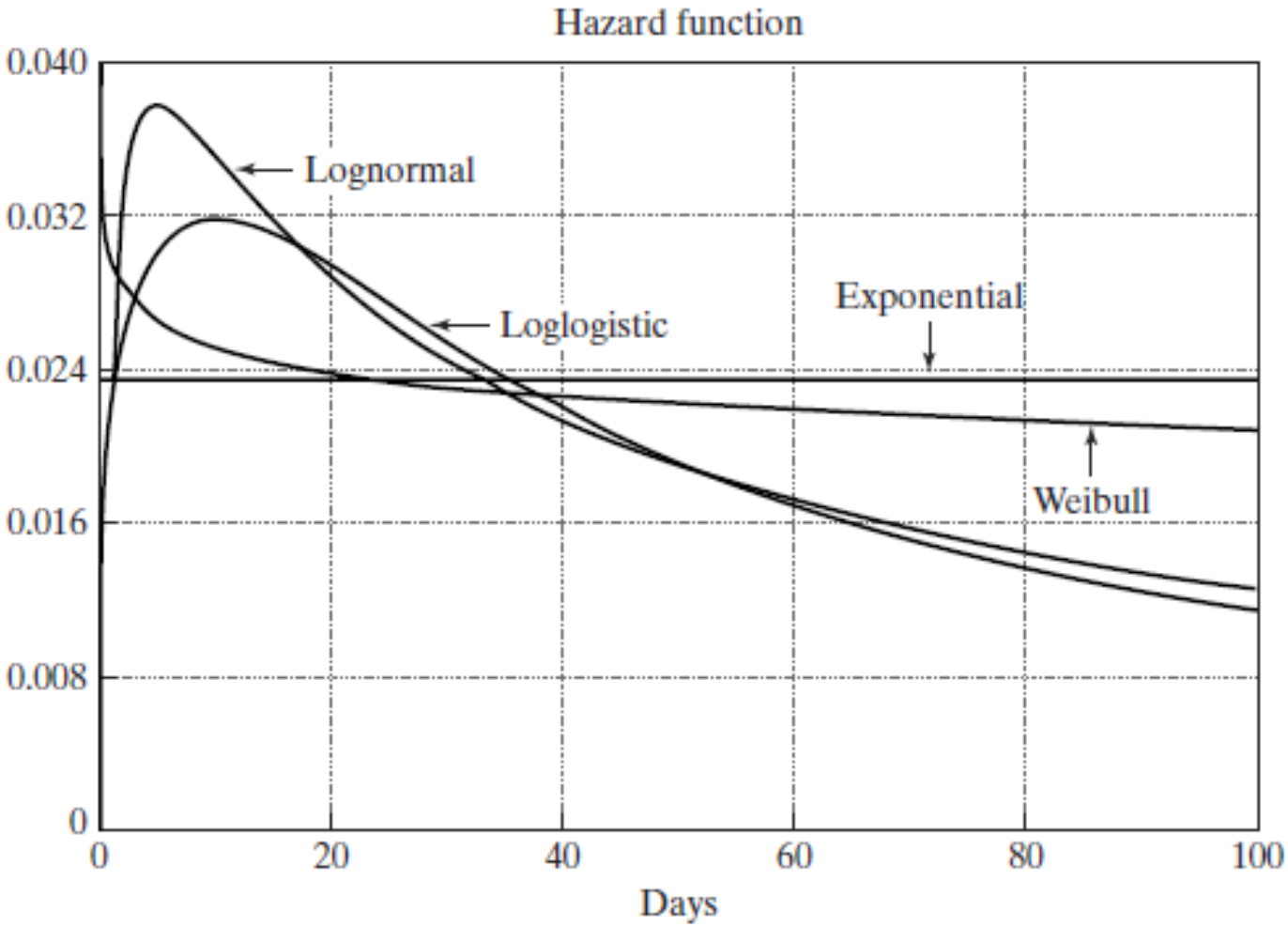


U-shaped hazard



Survival after heart attacks (WHAS500 data)

Hazard functions of different distributions



Source: W.G. Greene (2003). Econometric analysis. New York University, Prentice Hall.

Cumulative hazard function

Definition (Cumulative hazard function/integrated hazard function)

Integrating the hazard function $h(t)$ over interval $(0, t)$ gives the **cumulative hazard function**

$$H(t) = \int_0^t h(u)\, du$$

Similarly to the hazard function, the cumulative hazard function is not a probability but is a measure of risk: the greater the value of $H(t)$, the greater the risk of failure by time t.

NB!

Both hazard function and cumulative hazard function define the corresponding distribution uniquely, thus they can be considered as alternative tools for defining a distribution.

Location-scale family. Log-linear models

One common modelling option for survival data is the class log-linear models for lifetime T

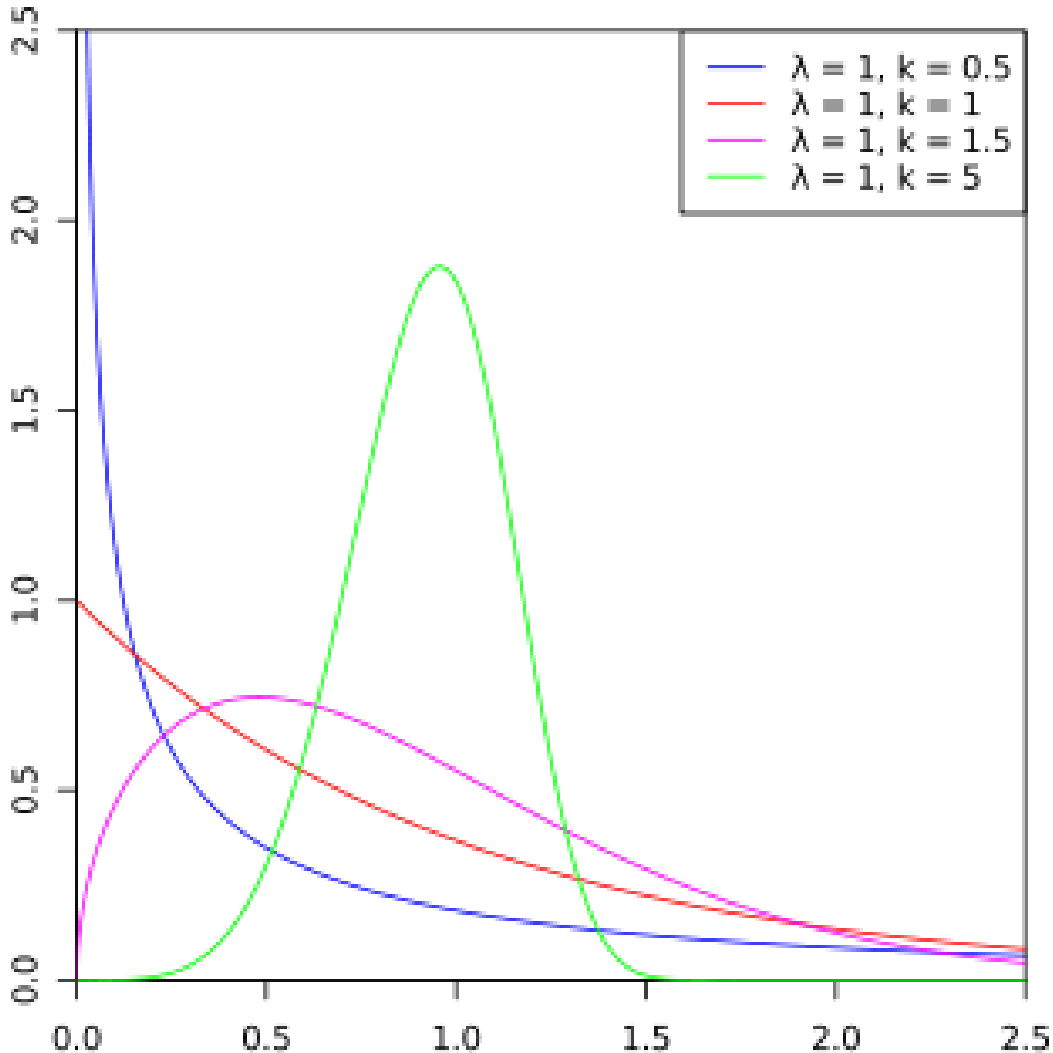
More specifically, we are interested in the models where the log-transform $Y = \ln(T)$ belongs to **location-scale family**, i.e. each rv Y can be expressed as

$$Y = \mu + \sigma Z,$$

where Z is the standard member ($\mu = 0$ and $\sigma = 1$):

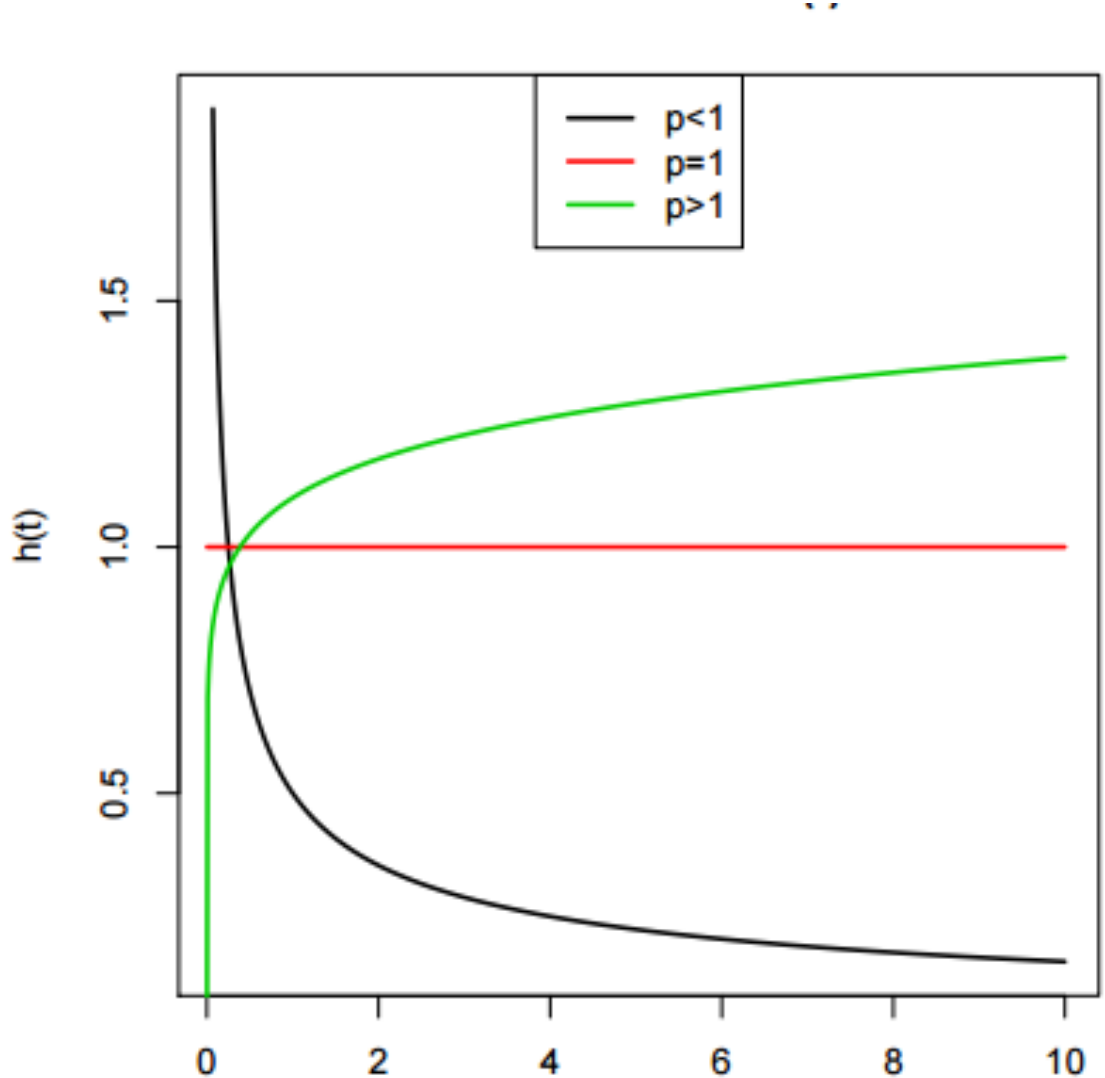
T	$Y = \ln(T)$
Weibull	extreme minimum value
log-normal	normal
log-logistic	logistic

Example. Weibull pdf



k – shape parameter ($k = 1$ means exponential dist.)

Example. Weibull hazard function



p – shape parameter ($p = 1$ means exponential dist.)

Parametric regression models (1)

Suppose that arguments $X = (X_1, \ldots, X_p)$ influence the time variable T
$\Rightarrow$ parameters of dist. of T depend on those arguments, i.e. we have conditional functions $f(t|\boldsymbol{x})$, $h(t|\boldsymbol{x})$, $S(t|\boldsymbol{x})$

Two main models that are used

1. **Accelerated failure time model** – parametric model, (usually) the effect of arguments to *survival time* is estimated
2. **Proportional hazards model** – semiparametric model, (usually) the effect of arguments to *hazard function* is estimated

In more general case, the arguments can change in time

Parametric regression models (2)

Table 17.5. *Standard Parametric Models and Their Hazard and Survivor Functions*[a]

Parametric Model	Hazard Function	Survivor Function	Type
Exponential	γ	$\exp(-\gamma t)$	PH, AFT
Weibull	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$	PH, AFT
Generalized Weibull	$\gamma\alpha t^{\alpha-1}S(t)^{-\mu}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$	PH
Gompertz	$\gamma\exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$	PH
Log-normal	$\frac{\exp(-(\ln t-\mu)^2/2\sigma^2)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t-\mu)/\sigma)]}$	$1-\Phi((\ln t-\mu)/\sigma)$	AFT
Log-logistic	$\alpha\gamma^{\alpha}t^{\alpha-1}/[(1+(\gamma t)^{\alpha})]$	$1/[1+(\gamma t)^{\alpha}]$	AFT
Gamma	$\frac{\gamma(\gamma t)^{\alpha-1}\exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha,\gamma t)]}$	$1-I(\alpha,\gamma t)$	AFT

[a] All the parameters are restricted to be positive, except that $-\infty < \alpha < \infty$ for the Gompertz model.

γ – scale, α – shape

Source: A.C. Cameron, P.K. Trivedi (2005). Microeconometrics Methods and Applications, Cambridge University Press.

Accelerated failure time (AFT) model (1)

Accelerated failure time (AFT) model

AFT model is characterized by the following property of survival function:

$$S(t|\boldsymbol{x}) = S_0(\exp(-\boldsymbol{x}^T\beta)t) = S_0(t^*),$$

where $S_0(\cdot)$ denotes the baseline survival function

Thus

- if $\boldsymbol{x}^T\beta$ decreases then t^* increases, which means the time to failure accelerates – accelerated failure time model
- if $\boldsymbol{x}^T\beta$ increases then t^* decreases, which means the time to failure decelerates – decelerated failure time model

Accelerated failure time (AFT) model (2)

AFT model is also called location-scale model, as it can be written as a log-linear model for failure time T such that

$$Y = \ln T = \boldsymbol{x}^T \boldsymbol{\beta} + Z,$$

where Z has a distribution from location-scale family

For example, the following regression models belong to the class of AFT models:

- exponential
- Weibull
- log-logistic
- log-normal

Cox proportional hazards model

For Cox PH model (Cox, 1972), the hazard function is

$$h(t|\boldsymbol{x}) = h_0(t) \cdot \exp(\boldsymbol{x}^T \beta),$$

where $h_0(t)$ is a baseline hazard function (does not depend on the covariates $\boldsymbol{x}$)

Exponential and Weibull models are special cases of this model

Definition (Proportional hazards property)

For two different observations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the **hazard ratio**

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)} = \frac{\exp(\boldsymbol{x}_1^T \beta)}{\exp(\boldsymbol{x}_2^T \beta)} = \exp\left((\boldsymbol{x}_1^T - \boldsymbol{x}_2^T)\beta\right)$$

is constant with respect to time t.

Estimates for β are based on *marginal likelihood* (Kalbfleisch (1973)) or *partial likelihood* (Cox (1972, 1975), Prentice (1983))

Example. Survival analysis of leukemia patients

Survival times of leukemia patients (in weeks)

Treatment: 6 6 6 6* 7 9* 10 10* 11* 13 16 17* 19* 20* 22 23 25* 32* 32* 34* 35*
Placebo: 1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23

Estimated survival function (*Kaplan-Meier*) (* means censoring)

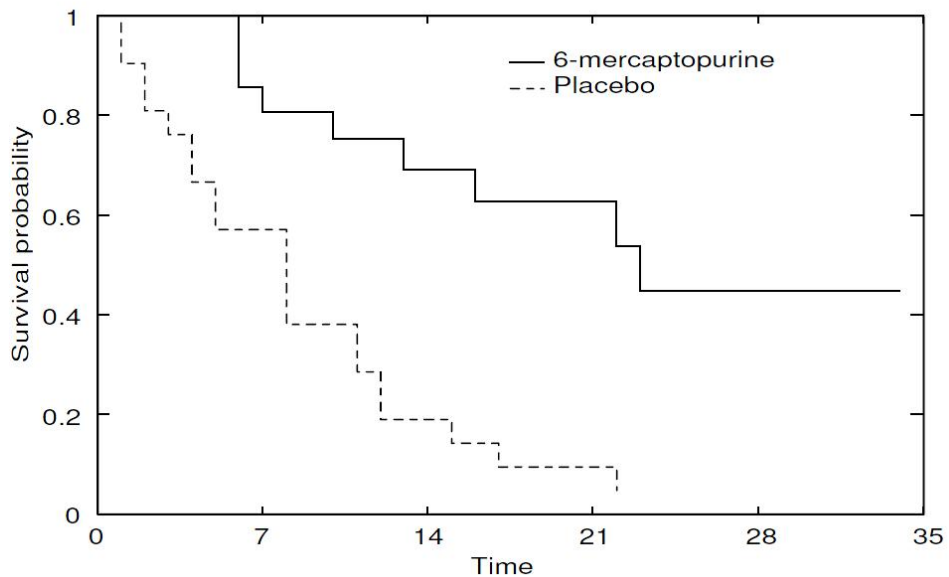


FIGURE 6.2. Kaplan–Meier curves for the survival data of Table 6.1.

Source: Gehan (1965) data. Lindsay (2007). *Applying Generalized Linear Models*, p 114

Example continued

Exponential and Weibull AFT models and Cox PH model were fitted (using treatment as an argument)

Results:

1. Exponential distribution
   - null model: $AIC = 235.5$
   - two groups: $AIC = 221.0$
   - estimated (constant) hazard ratio $\exp(1.527) = 4.604$ (since $\hat{\beta} = 1.527$)
2. Weibull distribution
   - null model: $AIC = 236.5$
   - two groups: $AIC = 218.9$
   - estimated (constant) hazard ratio $\exp(1.726) = 5.618$ (since $\hat{\beta} = 1.726$)
3. Cox PH model
   - null model: $AIC = 255.8$
   - two groups: $AIC = 242.4$
   - estimated (constant) hazard ratio $\exp(1.521) = 4.577$ (since $\hat{\beta} = 1.521$)

Conclusion?

Censoring in survival data

Censoring can arise in different situations and can be classified in several ways:

- **right censoring** – a data point is above a certain value but it is unknown by how much, e.g., indemnity limits in non-life insurance
- **left censoring** – a data point is below a certain value but it is unknown by how much, e.g., it is fixed when a diagnose is set, but the exact starting time of illness is not known
- **interval censoring** – a data point is somewhere in an interval between two values, e.g., time between visits to a doctor

Censoring is also classified into Type I, Type II and random censoring

Likelihood for censored data (1)

Let us denote

- T_i – lifetime of individual i, $i = 1, \ldots, n$ (with cdf F, pdf f, and survival function S)
- C_i – random censor time with survival function S_C and pdf f_C

Thus we can model the observed time Y_i as

$$Y_i = \min\{T_i, C_i\},$$

Let δ_i denote the **failure indicator** for observation i. Then the likelihood

- for uncensored observations ($\delta_i = 1$) is

$$L_{\delta_i=1} = f(t_i)S_C(t_i)$$

- for censored observations ($\delta_i = 0$) is

$$L_{\delta_i=0} = f_C(t_i)S(t_i)$$

Likelihood for censored data (2)

Now

$$L = \prod_{i=1}^{n} [f(t_i)(S_C(t_i))]^{\delta_i} [S(t_i) f_C(t_i)]^{1-\delta_i},$$

which (assuming C does depend on any arguments of interest) implies

$$\ln L = \sum_{i=1}^{n} [\delta_i \ln f(t_i) + (1 - \delta_i) \ln S(t_i)] \quad (+\text{const.})$$

or, equivalently,

$$\ln L = \sum_{i=1}^{n} [\ln S(t_i) + \delta_i \ln h(t_i)] \quad (+\text{const.})$$

Censored data in proportional hazards model

Proportional hazards model: $h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^T\beta)$

Cumulative hazard

$$H(t|\boldsymbol{x}) = \int_0^t h(u|\boldsymbol{x})du = H_0\exp(\boldsymbol{x}^T\beta),$$

where H_0 is cumulative baseline hazard: $H_0(t) = \int_0^t h_0(u)du$

Survival function:

$$S(t|\boldsymbol{x}) = \exp(-H_0(t)\exp(\boldsymbol{x}^T\beta))$$

Now the expressions for $S(t|\boldsymbol{x})$ and $h(t|\boldsymbol{x})$ imply

$$\ln S(t|\boldsymbol{x}) = -H_0(t)\exp(\eta), \qquad \ln h(t|\boldsymbol{x}) = \ln h_0(t) + \eta, \quad \text{where } \eta = \boldsymbol{x}^T\beta$$

$\Rightarrow$ we have both required components for log-likelihood:

$$\ln L = \sum_{i=1}^{n}[\delta_i(\ln h_0(t_i) + \eta_i) - H_0(t_i)\exp(\eta_i)]$$

Maximizing the log-likelihood gives us estimates for β

Proportional hazards model as Poisson model

Denote $\mu_i = H_0(t_i)\exp(\eta_i)$, then $\ln \mu_i = \ln H_0(t_i) + \eta_i$

Recall the expression of log-likelihood
$\ln L = \sum_{i=1}^{n}[\delta_i(\ln h_0(t_i) + \eta_i) - H_0(t_i)\exp(\eta_i)]$ and rewrite it as

$$\ln L = \sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i) + \sum_{i=1}^{n}\delta_i \ln \frac{h_0(t_i)}{H_0(t_i)}$$

Notice that the first summand is similar to Poisson log-likelihood and second does not depend on arguments $\Rightarrow$ the failure indicator δ_i can be considered a Poisson r.v. with mean μ_i

Parameter vector $\boldsymbol{\beta}$ is estimated using standard methods,
where $\ln H_0(t_i)$ can be considered as offset (fixed additional intercept)

Thus, the model for failure indicator is a Poisson model with log-link:

$$\ln \mu_i = \ln H_0(t_i) + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Example. Cox proportional hazards model

Recidivism analysis

Persons released from Maryland's prison ($n = 432$) were under surveillance the following year. Weekly data was collected: 52 weeks, censoring occurred if there were no arrests during this period (`week` – week of arrest)

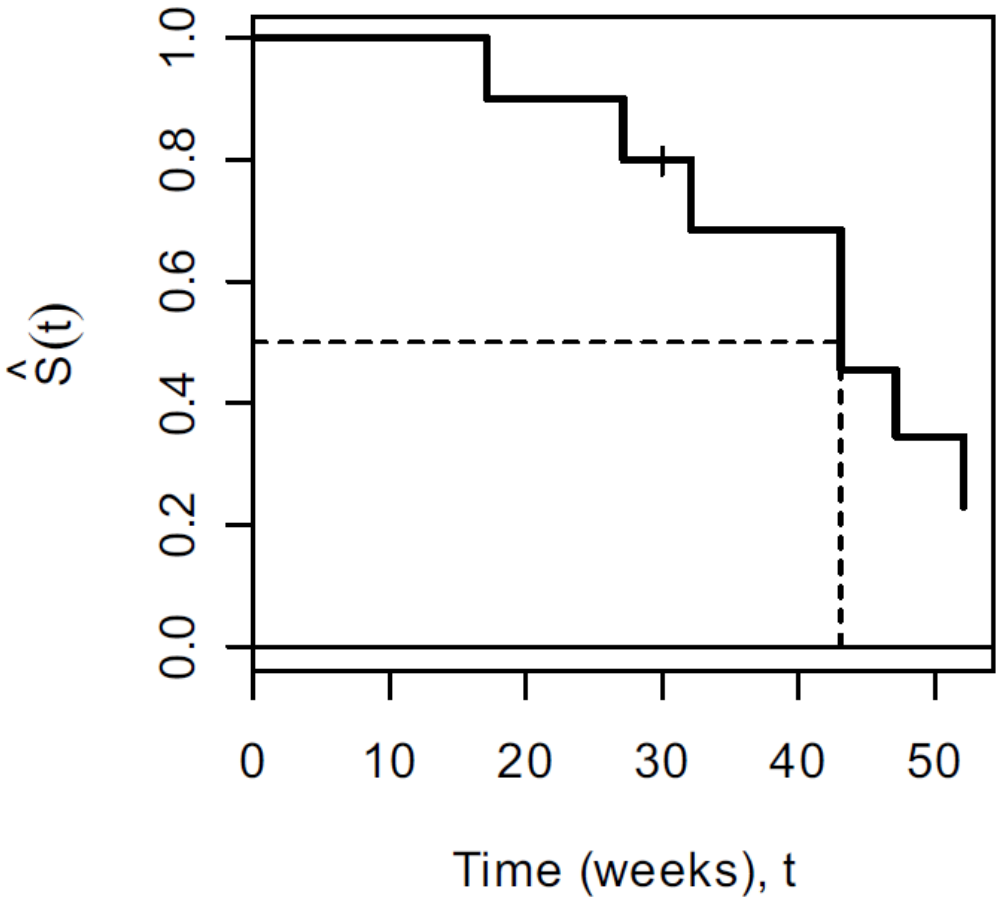
Prisoner	week	arrest
174	17	1
999	30	0
77	43	1
168	52	0
...		

Available arguments: `fin` (1–support, 0–no support), `age`, `wexp` (1–working experience, 0–no exp.), `mar` (1–married), `education`, `prio` (number of prior convictions)
Cox proportional hazards model was fitted: $h(t) = h_0 \exp(b_1 x_1 + \ldots + b_k x_k)$
Results: `fin`: $b = -0.38$, $\exp(-0.38) = 0.68$, `age`: $b = -0.057$, $\exp(-0.057) = 0.94$, `prio`: $b = 0.09$, $\exp(0.09) = 1.09$, coef.-s for education, marriage, working experience were not significant

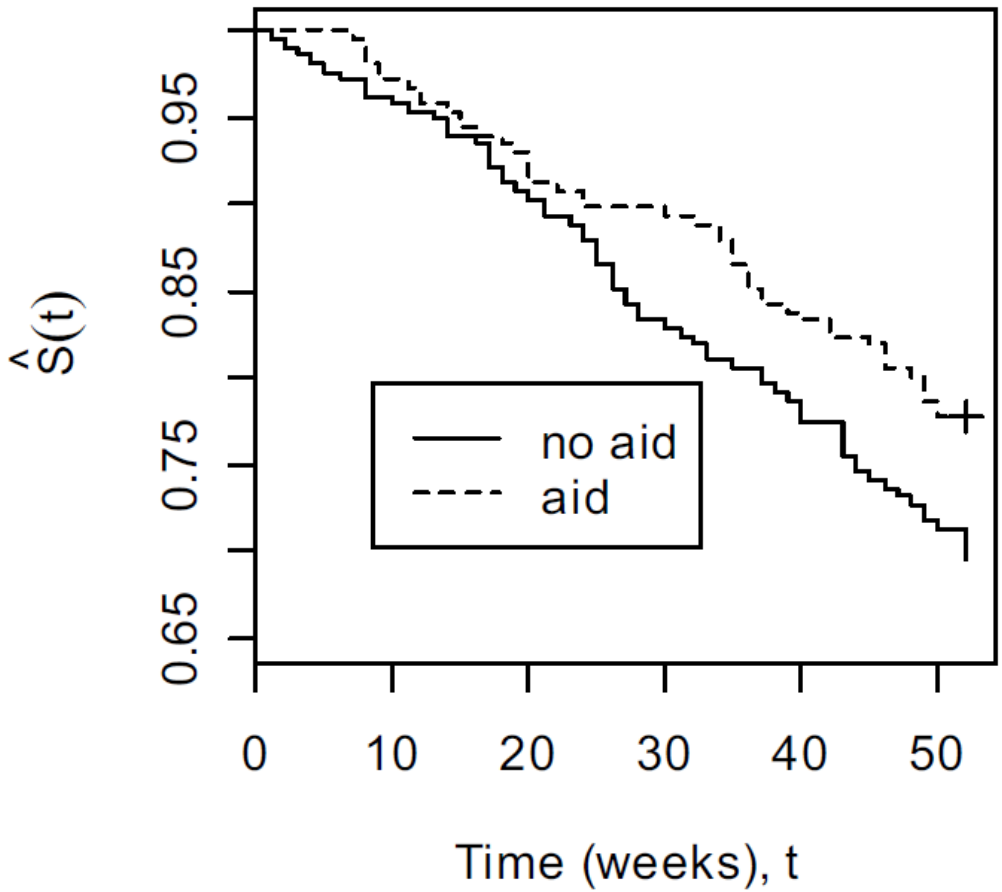
Data: Rossi et al. 1980, analyzed by Allison 1984, Fox 2002 (in R)

Example. Analysis of recidivism data (1)

Recidivism data survival function, median $\hat{T}_{0.5} = 43$



Example. Analysis of recidivism data (2)



Censoring of a continuous response. Motivating examples

Censoring of a continuous response variable is a common problem in economic data. Several examples are available in economics and sociology:

1. Household expenditures on durable goods (Tobin, 1958)
2. Analysis of extramarital affairs (Fair, 1977, 1978)
3. The number of hours worked by a woman in the labor force (Quester, Green, 1982)
4. Household expenditures on various commodity groups (Jarque, 1987)
5. Vacation expenditures (Melenberg, Soest, 1996)

Censoring and truncation for continuous response

Recall that

- **Truncated data** – some observations are missing (in response and in arguments)
- **Censored data** – some observations are missing in response

In case of truncated data, more information is lost

The truncated/censored variable Y is usually given conditionally through a (partially observed) latent variable Y^* and the corresponding realizations are denoted by y and y^*

In case of truncation/censoring from below (left), Y^* is observed if it exceeds certain threshold (often 0)

Let us have $Y = Y^*$ if $Y^* > 0$

Example

Model for annual hours worked by hourly wage w (generated data)
Model $y^* = -2500 + \ln(w)$, censored below from 0 hours (35% of observations)

Source: Cameron & Trivedi (2005). Chap. 16: *Tobit and selection models*, 529 – 572

Example. Model for hours worked

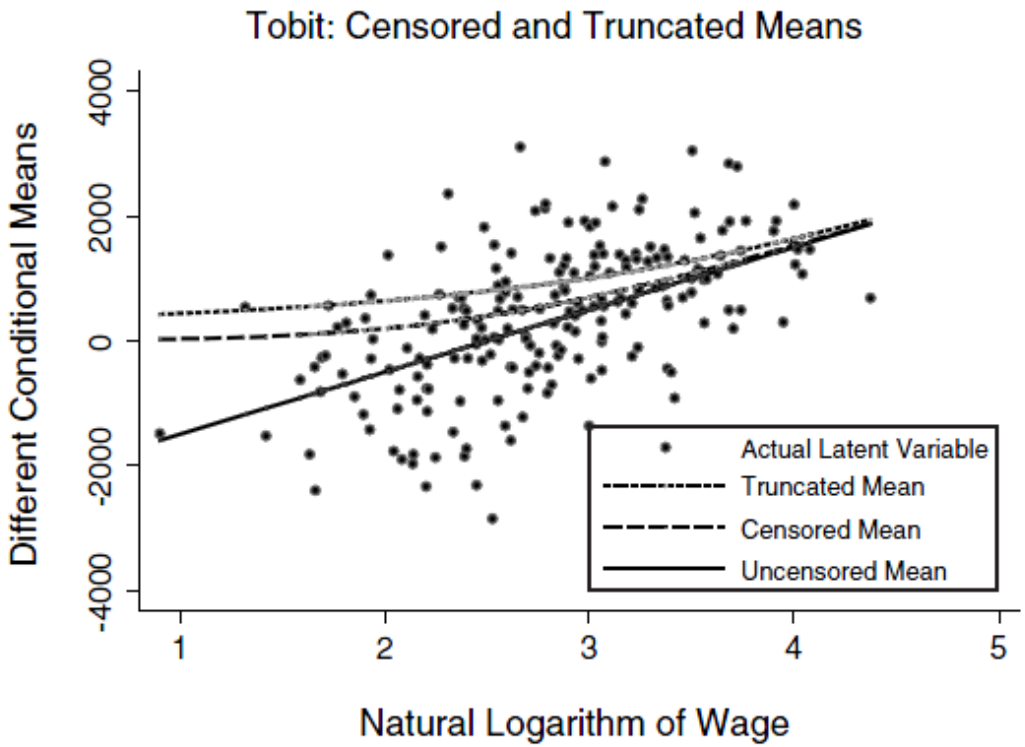


Figure 16.1: Tobit regression of hours on log wage: uncensored conditional mean (bottom), censored conditional mean (middle), and truncated conditional mean (top) for censoring/truncation from below at zero hours. Data are generated from a classical linear regression model.

Censoring (1)

Censoring from below/left:

$$Y = \begin{cases} Y^*, & \text{if } Y^* > b \\ b, & \text{if } Y^* \leq b \end{cases}$$

In other words, we have a continuous-discrete mixture:

- If $Y > b$, we have a continuous distribution with pdf $f(y|\boldsymbol{x}) = f^*(y|\boldsymbol{x})$
- There is point mass at b, $\mathbf{P}\{Y = b|\boldsymbol{x}\} = \mathbf{P}\{Y^* \leq b|\boldsymbol{x}\} = F^*(b|\boldsymbol{x})$

Thus, in case of left-censoring

$$f(y|\boldsymbol{x}) = \begin{cases} f^*(y|\boldsymbol{x}), & \text{if } y > b \\ F^*(b|\boldsymbol{x}), & \text{if } y = b \end{cases}$$

Censoring (2)

Defining the indicator D by $D = 1$ for $Y > b$ and $D = 0$ for $Y = b$ (recall the construction from survival models!) and denoting the observed value of D by d, we obtain the conditional density:

$$f(y|\boldsymbol{x}) = f^*(y|\boldsymbol{x})^d F^*(b|\boldsymbol{x})^{1-d} \tag{1}$$

and the corresponding sample log-likelihood is

$$\ln L(\theta) = \sum [d_i \ln f^*(y_i|\boldsymbol{x}_i, \theta) + (1 - d_i) \ln F^*(b_i|\boldsymbol{x}_i, \theta)]$$

where θ is the parameter of corresponding distribution

A common option is to choose equal threshold: $b_i = b$

Truncation

Truncation from below/left: $Y = Y^*$, if $Y^* > b$

Conditional density under truncation:

$$f(y|\boldsymbol{x}) = f^*(y^*|\boldsymbol{x}; (Y^*|\boldsymbol{x}) > b) = \frac{f^*(y|\boldsymbol{x})}{\mathbf{P}\{(Y^*|\boldsymbol{x}) > b\}} = \frac{f^*(y|\boldsymbol{x})}{1 - F^*(b|\boldsymbol{x})}$$

Corresponding sample log-likelihood:

$$\ln L(\theta) = \sum [\ln f^*(y_i|\boldsymbol{x}_i, \theta) - \ln(1 - F^*(b_i|\boldsymbol{x}_i, \theta))] \tag{2}$$

ML estimates obtained using truncated or censored log-likelihood are consistent and asymptotically normal if the distribution of the latent variable is correctly determined

Tobit model

Tobit model: **censored normal regression model**

Censoring from below (at $b = 0$):

Tobit model

Model $Y^* = \boldsymbol{x}^T\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$

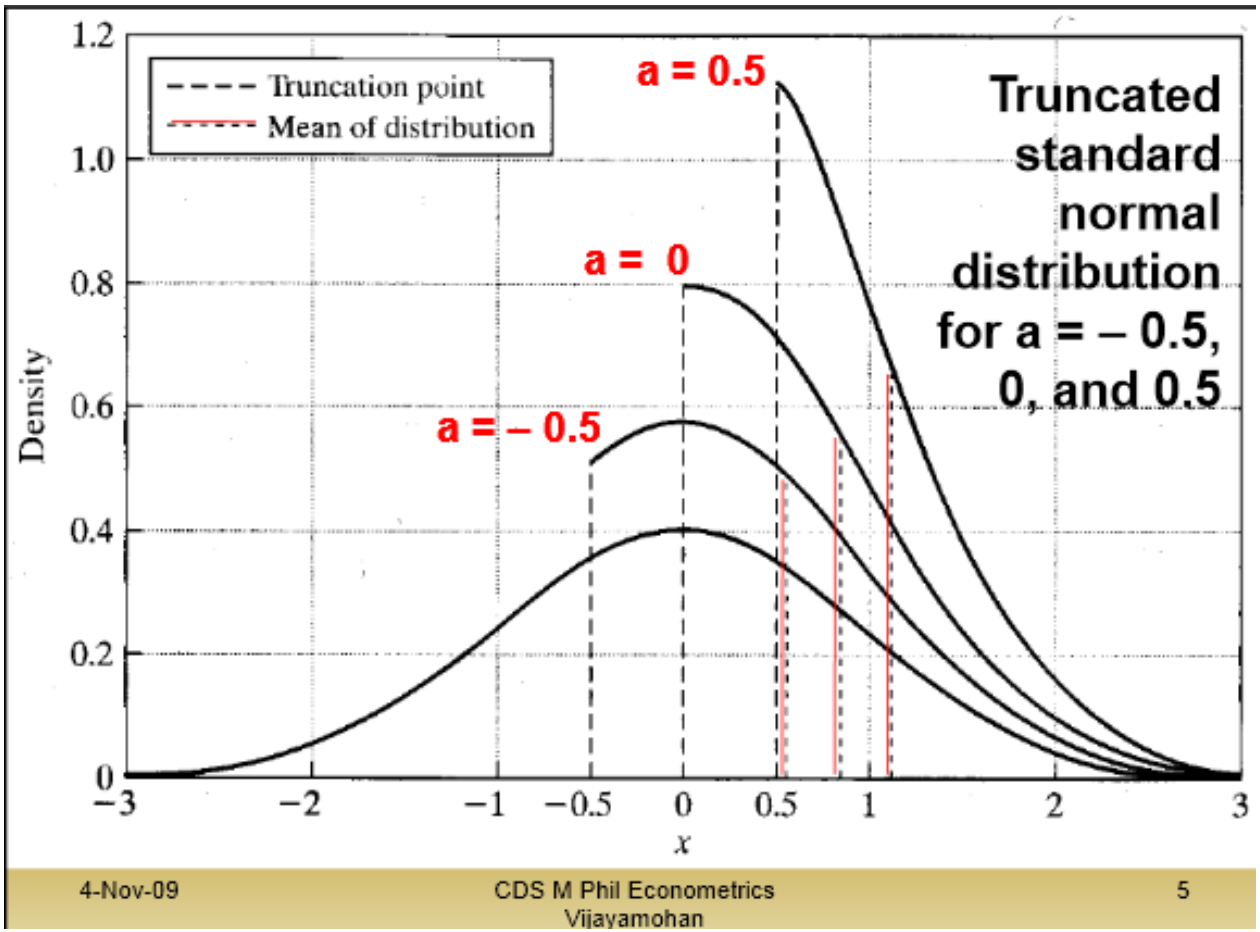
$$Y = \begin{cases} Y^*, & \text{if } Y^* > 0 \\ -, & \text{if } Y^* \leq 0 \end{cases}$$

Observed values y are equal to observed values y^* if $y^* > 0$ and missing (or 0) if $y^* \leq 0$

In general, data may be censored either from below or below or from both sides

Tobin (1958) was the first to apply this model (modelling household expenditures on durable goods)

Truncation of normal distribution



Source: Vijayamohanan, Pillai N. (2009)

Estimation of Tobit model

The parameters of $Y^* = \boldsymbol{x}^T\beta + \varepsilon$ are estimated using MLE

Let us find the likelihood, starting from (1), where y^* is a realization of $Y^* \sim N(\boldsymbol{x}^T\beta, \sigma^2)$

For $F^*(0)$ we write

$$F^*(0) = \mathbf{P}\{Y^* \le 0\} = \mathbf{P}\{\boldsymbol{x}^T\beta + \varepsilon \le 0\} = \Phi\left(-\frac{\boldsymbol{x}^T\beta}{\sigma}\right) = 1 - \Phi\left(\frac{\boldsymbol{x}^T\beta}{\sigma}\right) \quad (3)$$

Thus, using the indicator d (with $b = 0$), the relation (1) can be rewritten

$$f(y|\boldsymbol{x}) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y - \boldsymbol{x}^T\beta)^2\right\}\right]^d \left[1 - \Phi\left(\frac{\boldsymbol{x}^T\beta}{\sigma}\right)\right]^{1-d}$$

Sample log-likelihood:

$$\ln L(\beta, \sigma^2) = \sum \left\{ d_i \left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\beta)^2\right] + (1 - d_i)\ln\left[1 - \Phi\left(\frac{\boldsymbol{x}_i^T\beta}{\sigma}\right)\right]\right\} \quad (4)$$

Estimation of truncated model

Tobit model can be applied to *truncated* data as well (although originally meant for censored data)

Let us start with log-likelihood (2):

$$\ln L(\theta) = \sum [\ln f^*(y_i|\boldsymbol{x}_i, \theta) - \ln(1 - F^*(L_i|\boldsymbol{x}_i, \theta))]$$

Now, applying formula (3) for $F^*(0)$, we obtain:

$$\ln L(\beta, \sigma^2) = \sum \left[-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - \boldsymbol{x}_i^T \beta)^2 - \ln \Phi \left(\frac{\boldsymbol{x}_i^T \beta}{\sigma} \right) \right]$$

Problems with Tobit model

Tobit model is **very sensitive to assumptions about the distribution**:

if the model errors are not normally distributed or the variance is not constant, **MLE estimates are not consistent**

One possibility to overcome this problem is to also estimate a model for variance: $\sigma_i^2 = \boldsymbol{x}_i^T \boldsymbol{\gamma}$ and use this estimate in (4)

Variations of Tobit model

In case of censoring we assume that the same process is defining measuring and censoring

In general, these processes can be different

Two approaches:

1. *Two-step (or two-part) model*
2. *Sample selection model*

Two-part model

Model proposed by Cragg (1971) as a generalization of Tobit model

We define a participation model, where

- **participant**, indicator $D = 1$, fully measured, $Y > 0$
- **non-participant**, indicator $D = 0$, not measured, $Y = 0$

Two-part model

$$f(y|\boldsymbol{x}, \boldsymbol{w}) = \begin{cases} \mathbf{P}\{D = 0|\boldsymbol{w}\}, & \text{if } y = 0 \\ \mathbf{P}\{D = 1|\boldsymbol{w}\} f(y|D = 1, \boldsymbol{x}), & \text{if } y > 0 \end{cases}$$

Main choices for participation model are *logit* and *probit* model

Does this construction ring a bell? Do you recall any two-part models we have studied already?

Known also as *hurdle model*, since participation starts if a "hurdle" is crossed. For example, ZAP and ZANB models are two-part models

Example of a two-part model

Analysis of health expenditure

Data from health insurance. Two questions:

- was there any expenditure during a year
- if there is some expenditure, how to model it

Two-part solution:

- *Probit* model: $\mathbf{P}\{D = 1|\boldsymbol{w}\} = \Phi(\boldsymbol{w}^T\beta_1)$
- Log-normal model (provided that there is some expenditure):
  $\ln Y|D = 1, \boldsymbol{x} \sim N(\boldsymbol{x}^T\beta_2, \sigma_2^2)$

Source: Duan *et al* (1983)

Sample selection model

The assumption of random sample is not always fulfilled, **selection may be related to the response variable** (*self-selection*)!

In extreme case, only **participants** are selected. This needs to be taken into account while estimating the parameters

Depending on what the selection depends, we may apply different models: Tobit model, Tobit 2 model (Heckman's model), Roy model

Most of these models were developed by Heckman in 1970s

James Joseph Heckman (s. 1944), US economist, Nobel prize in 2000 (together with McFadden)

Two-dimensional sample selection (Tobit 2)

Let Y_2^* be the response var. of interest. In case of regular Tobit model we only observe $Y_2^* > 0$

In a more general case, we may include a latent var. Y_1^* so that Y_2^* is observed if $Y_1^* > 0$

Two-dimensional sample selection model (Tobit 2 / Heckman's model)

1. Participation

$$Y_1 = \begin{cases} 1, & \text{if } Y_1^* > 0 \\ 0, & \text{if } Y_1^* \leq 0 \end{cases}$$

2. Measuring

$$Y_2 = \begin{cases} Y_2^*, & \text{if } Y_1^* > 0 \\ -, & \text{if } Y_1^* \leq 0 \end{cases}$$

In other words, the model assumes that Y_2 is observed if $Y_1^* > 0$

Models for latent variables:

- $Y_1^* = \boldsymbol{w}^T \beta_1 + \varepsilon_1$,
- $Y_2^* = \boldsymbol{x}^T \beta_2 + \varepsilon_2$

Reduces to regular Tobit model if $Y_1^* = Y_2^*$

The Roy model (Roy, 1951)

So far we assumed that Y_2 is not observed if $Y_1 = 0$

In general, Y_2 can be observed, but it has two possible states

In that case we have three latent responses Y_1^*, Y_2^*, Y_3^*:
Y_1^* determines whether Y_2^* or Y_3^* is observed

The Roy model

$$Y_1 = \begin{cases} 1, & \text{if } Y_1^* > 0 \\ 0, & \text{if } Y_1^* \leq 0 \end{cases}$$

$$Y = \begin{cases} Y_2^*, & \text{if } Y_1^* > 0 \\ Y_3^*, & \text{if } Y_1^* \leq 0 \end{cases}$$

Models for latent variables:

- $Y_1^* = \boldsymbol{w}^T \beta_1 + \varepsilon_1$,
- $Y_2^* = \boldsymbol{x}^T \beta_2 + \varepsilon_2$,
- $Y_3^* = \boldsymbol{x}^T \beta_3 + \varepsilon_3$

Typology (from Berinsky/Breene)

Sample	Y Variable	X Variable	Example
Censored	y is known exactly only if some criterion defined. in terms of y is met.	x variables are observed for the entire sample, regardless of whether y is observed exactly	Determinants of income; income is measured exactly only if it above the poverty line. All other incomes are reported at the poverty line
Sample Selected	y is observed only if a criteria defined. in terms of some other random variable (Z) is met.	x and w (the determinants of whether Z =1) are observed for the entire sample, regardless of whether y is observed or not	Survey data with item or unit non-response
Truncated	y is known only if some criterion defined in terms of y is met.	x variables are observed only if y is observed.	Donations to political campaigns.

Source: Hopkins, D. (2005). *Heckman Selection Models*