Censored data in Social Sciences & Survival in Healthcare

Tool : Tobit

The Analytics Edge

In some applications, we have only access to censored data. A censored variable has a large fraction of observations either at the minimum or maximum.

Examples of censored data include:

1) Demand of households for capital goods such as automobiles or major household appliances. When households report on whether they have purchased such a good in the past year, many might report zero expenditures but those who made expenditures, show significant variation

2) Number of extramarital affairs (This data has been collected by magazine surveys)

3) Expenditures on vacations

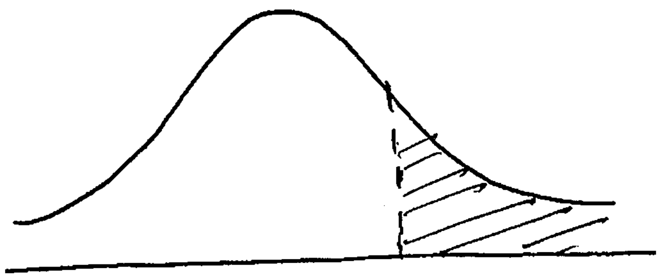
4) Educational testing: If an exam is too easy, lots of people with full marks while too hard, lots of people with zero marks

unemployme

5) Labor hours and relation to employee characteristics with higher

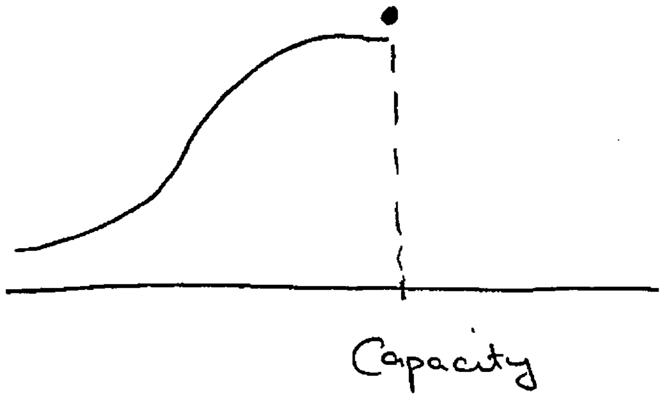
Censored dependent variable

Y^*



True Demand distribution for say seats for soccer games

Y



Demand is only observed up to the capacity and it is censored above that

Assume $Y^* \sim N(\mu, \sigma)$

Consider a censored value at C (say capacity)

Then, the variable we observe is:

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

In this example, the censored random variable is <u>right censored</u> and the new variable is a mixture of continuous and discrete parts.

$$P(Y=c) = P(Y^* > c) = 1 - \Phi\left(\frac{c-\mu}{\sigma}\right)$$

$Y^* \leq C$ then Y the df of Y^*

Censored regression (Tobit model)

James Tobin in 1958 proposed a model that deals with censored regression type problem where the model is referred to as the Tobit model (from Tobin + probit)

$$\underbrace{Y_i^*}_{\text{Latent variable (unobservable)}} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \quad \forall i = 1, \dots, n$$

$$\underbrace{Y_i}_{\text{observable}} = \begin{cases} Y_i^* & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \quad \text{left censored at } 0$$

The Tobit model assume $\epsilon_i \sim N(0, \sigma^2)$

Note in this model β_j is not the linear effect of X_{ij} on Y_i as is the case in linear regression. Here:

$$E(Y_i) = E(Y_i^* | Y_i^* > 0) P(Y_i^* > 0) + 0 \, P(Y_i^* \leq 0)$$

$$= E\left(\beta_0 + \beta' X_i + \epsilon_i \middle| \beta_0 + \beta' X_i + \epsilon_i > 0\right) P\left(\beta_0 + \beta' X_i + \epsilon_i >\right.$$

$$\frac{\partial E(Y_i | X_i)}{\partial X_i} = \beta \, P(Y_i^* > 0) = \beta \, P(\beta_0 + \beta' X_i + \epsilon_i > 0)$$

Maximum likelihood estimation

Given the predictor variables $x_{i1}, \ldots, x_{ip}$ for $i = 1, \ldots, n$ and the dependent variable Y_i which takes either a value 0 or some positive number, find β, σ that solves:

$$\max_{\beta,\sigma} LL(\beta) = \max_{\beta,\sigma} \sum_{i:Y_i=0} \ln\left(1 - \Phi\left(\frac{\beta' x_i}{\sigma}\right)\right) + \sum_{i:Y_i>0} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i - \beta' x_i)^2}{2\sigma^2}}\right)$$

The first term corresponds to $P(Y_i^* \leq 0)$ while the second term corresponds to $f(Y_i^*)$

Note here we can easily introduce β_0 by setting corresponding x_i parameter to be 1 thus can be easily included in the model.

$$\max_{\beta,\sigma} \sum_{i:Y_i=0} \ln\left(1 - \Phi\left(\frac{\beta' x_i}{\sigma}\right)\right) + \sum_{i:Y_i>0} -\frac{1}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{(Y_i - \beta' x_i)^2}{\sigma^2}\right]$$

Note we can transform variables as

$$\gamma = \beta/\sigma, \quad \theta = 1/\sigma$$

$$\max_{\gamma,\theta} \sum_{i:Y_i=0} \ln\left(1 - \Phi(\gamma' x_i)\right) + \sum_{i:Y_i>0} -\frac{1}{2}\left[\ln(2\pi) - \ln(\theta^2) + (\theta Y_i - \gamma' x_i)^2\right]$$

This is a mixture of continuous & discrete approaches. Yet the objective is concave and we can easily solve it.

Better predictors:

$$E[Y_i|x_i] = E\left(\beta' x_i + \epsilon_i \,\middle|\, x_i, \beta' x_i + \epsilon_i > 0\right) P\left(\beta' x_i + \epsilon_i > 0\right)$$

$$= \left(\beta' x_i + E\left(\epsilon_i \,\middle|\, \epsilon_i > -\beta' x_i\right)\right) \Phi\left(\frac{\beta' x_i}{\sigma}\right)$$

$$= \underbrace{\left(\beta' x_i + \sigma \frac{\phi(\beta' x_i/\sigma)}{\Phi(\beta' x_i/\sigma)}\right)}_{= E[Y_i | Y_i > 0, x_i]} \Phi\left(\frac{\beta' x_i}{\sigma}\right)$$

$$= \beta' x_i \Phi\left(\frac{\beta' x_i}{\sigma}\right) + \sigma \phi\left(\frac{\beta' x_i}{\sigma}\right)$$

The previous predictor is:

$$E[Y_i^* | x_i] = \beta' x_i$$

$$Y_i = \begin{cases} \beta' x_i & \text{if } \beta' x_i > 0 \\ 0 & \text{if } \beta' x_i \leq 0 \end{cases}$$

Here we assume that the intercept is include in the x_i with a 1 & the corresponding coefficient in β corresponds to β_0.

Analytics on censored data in R

```
exm <- read.csv("extramarital.csv")
str(exm)
```

This data consists of 6366 observations of 7 variables

marriage_rating (1 to 5)

age (age of woman)

yrs_married

religiosity (How religious the women is)

education

occupation (1 to 6)

time_in_affairs

This data was collected in a survey by Redbook.

```
table(exm$time_in_affairs > 0)
```

FALSE	TRUE
4313	2053

Out of the 6366 observations, 2053 observations are positive while 4313 observations are 0.

```
hist(exm$time_in_affairs, 100)
```

time-in-affairs: Amount of time spent in extramarital affairs

(Average = 0.7054)
Ranges from 0 to 57.6

marriage-rating (5 = very good, 1 = very poor)

age (17.5 = under 20, 22 = 20 to 24, ... 37 = 35 to 39, 42 = 40 or over)

yrs-married (0.5 = less than 1 year, 2.5 = 1 to 4 years, 6 = 5 to 7 years, ...)

religiosity (4 = strongly ... , 1 = not)

Occupation (6 = professional with advanced degree, 5 = managerial, administrative, business, ... 1 = student)

```
install.packages("survival")
library(survival)
```

Survival analysis package that can perform Tobit analysis
approach

Note that in this date, there is a large number of zeros for the affairs variable. For example if someone is not having an affair, a small decrese in one of the predictor variables may not change hear chances of having an affair in comparison to a person having an affair.

One needs to account for possibly two types of people, those who are unlikely to be moved away from 0 by modest changes in predictors and those who are not 0 & likely to change by modest changes in predictors.

To do this, we use the Tobit model.

Create training & test set

```
set.seed(100)
spl <- sample(nrow(exm), 0.7*nrow(exm))
train <- exm[spl, ]
test <- exm[-spl, ]
```

Tobit analysis

```
model1 <- survreg(Surv(time_in_affairs,
                  time_in_affairs > 0, type = "left")
                  ~ ., data = train, dist = "gaussian")
```

This fits a Tobit model where the data is left censored at zero. The error terms are assumed to be Gaussian.

Summary(model1)

The model estimates β coefficients & scale parameter. Note that $\beta_{marriage\_rating}$ is negative, which indicates as the marriage rating increases the chances of affair & length of affair ↓.

Occupation has a positive coefficient indicating greater chances of affair with an advanced occupation. Number of years married ↑ indicates greater chances of affair while religiosity has a negative coefficient. The coefficients are all fairly significant.

The model on the training set has a log likelihood value of −5437.

predict1 <- predict(model1, newdata = test)

predict() provides the latent predicted values of the variables (positive and negative)

model2 <- lm(time_in_affairs ~ ., data = train)

Perform a linear regression for comparison

predict2 <- predict(model2, newdata = test)

table(predict1 <= 0, test$time_in_affairs == 0)

table(predict2 <= 0, test$time_in_affairs == 0)

	FALSE	TRUE
FALSE	92	43
TRUE	523	1252

Tobit

	FALSE	TRUE
FALSE	582	1175
TRUE	33	120

linear regression

The Tobit model has better accuracy in terms of accounting for the large number of individuals for whom the extramarital affairs variable was 0.

Note here we are simply comparing how many of the $\beta_0 + \sum_j \beta x_{ij}$ values are below & above zero. One can also more carefully evaluate the

$P(Y_i > 0) = P(Y_i^* > 0)$ & $P(Y_i = 0) = P(Y_i^* \leq 0)$

Duration (Survival data) in Healthcare

In certain types of data the duration of an event is observed. This is the length of time that elapses from the beginning of an event until its end or until the measurement is taken which might precede termination

Heart transplant

Observation
Patient might be dead
Patient might be alive } Censored data

Survival time for the patients alive is at least t but not exactly t.

Note that unlike conventional regression in this case, there are predictor variables X(t) that also evolves possibly from time 0 to time t and can be used to describe survival.

Other examples include for example governments studying the length of unemployment of citizens, the citizen might still be unemployed at the time data is collected but it is believed that this person will get the employment soon.

To model duration data is important for example for transport engineers in estimating the length of time until failure, for biomedical researchers in estimating survival times after an operation.

Note that durations times are nonnegative by definition.

Say T is the duration of an event (random variable) (assume continuous)

$$\underbrace{\text{Survival function}}_{\text{Probability duration is at least } t} \quad S(t) = P(T \geq t) = 1 - F(t)$$

One important question in analysis of duration data is, given that an event has lasted until t what is the chances that it will end in a short interval of time Δt?

This is given by

$$P(t \le T \le t + \Delta t)$$

Hazard rate $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t+\Delta t \mid T \ge t)}{\Delta t}$

Rate at which events are completed after duration t given that they last at least until t

$$= \lim_{\Delta t \to 0} \frac{P(t \le T \le t+\Delta t)}{P(T \ge t)\, \Delta t}$$

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

$$= \frac{f(t)}{S(t)}$$ ($f(t)$ is density function)

Note $\lambda(t) = -\frac{d \ln S(t)}{dt}$ (alternative definition)

For example if you assume $\lambda(t) = \lambda$ (a constant over time) then

$$S(t) = K e^{-\lambda t}$$

Given $S(0) = 1 \Rightarrow S(t) = e^{-\lambda t}$ (exponential)

Kaplan - Meier estimator

Kaplan and Meier in 1958 proposed a non-parametric estimate of survival function from lifetime data. Their paper is one of the heavily cited works (34000 times) due to its applicability.

Say you observe duration of events of N people (say time until death)

$$t_1 \le t_2 \le \dots \le t_N \quad \text{(No censoring)}$$

For each time t_i, there is a number of people who are at risk just prior to this instant (n_i)

Let d_i be number of deaths at time t_i.

Estimate of survival function & hazard rate

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \qquad \lambda(t) = \frac{d_i}{n_i}$$

$$= \frac{n_i - d_i}{n_1} \quad \text{(No censoring)}$$

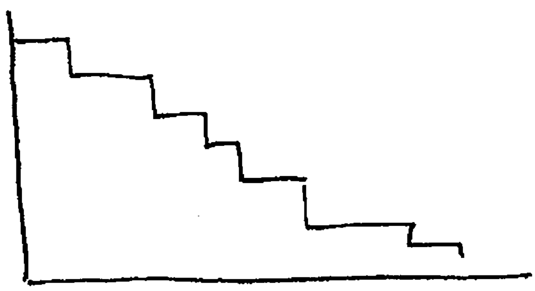
With censoring, n_i is the number of people at risk minus the number of losses (censored data)

Example :

Consider remission times for 21 patients in weeks suffering from leukemia

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

t	$S(t)$
$t < 1$	$21/21 = 1$
$1 \le t < 2$	$\frac{21-2}{21} = 0.905$
$2 \le t < 3$	$\frac{19-2}{21} = 0.81$



Kaplan - Meier Curve

(Series of declining horizontal steps)

The Kaplan - Meier curve can deal with censored data easily. For example right censored data (patient withdraws from study, alive at last ~~[illegible]~~ follow up)

Cox proportional hazard model (Over 40000 citations for original paper)

This is a semi-parametric approach where the effects of the ~~(covariates, independent variables)~~ variables are modeled on the dependent variable.

In fact, the Framingham Heart Study used this model.

Model assumes

$$\lambda(t, x) = \underbrace{\lambda_0(t)}_{\text{baseline hazard function}} e^{\beta' \underset{\sim}{x}} \longrightarrow \text{predictor variables}$$

Let t_i be the time at which individual i exits.

The probability of this event $= \dfrac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}}$

Here $R(t_i)$ are all the individuals at risk at time t_i

To find β, assuming no censoring where n is no. of observations:

$$\max_{\beta} \sum_{i=1}^{n} \left(\beta' x_i - \ln \left(\sum_{j \in R(t_i)} e^{\beta' x_j} \right) \right)$$

This model can be extended to censored data and instances where multiple individuals leave at the same time. Also x might depend on time t.

Heart transplant survival data

In heart transplant, a donor heart matched on blood type is sought. This data comes from Stanford Heart Transplantation Program.

The goal is to estimate the survival of patients from the data and understand the effect of other explanatory variables on whether transplantation helps.
Note that in some cases the appropriate heart for transplant might not be available and patients need to wait for it.

This study was conducted in April 1,1974.
The data provides survival information from early heart transplants at Stanford.

The survival time is censored if the patient drops out of the program (no follow up information) or the patient is alive at the time of the end of the study.

Analytics on heart transplant data set

```
heart <- read.csv("heart.csv")
```

str(heart)   172 observations of 7 variables

start, stop, event } Entry & exit times & status for this time of interval (in days)
event   1 = dead, 0 = alive

age : Age at the start of the program

surgery : Prior bypass surgery
1 = yes, 0 = No

transplant : Received transplant
1 = yes, 0 = No

id : Patient id

unique(heart$id) : A total of 103 patients data is provided in this set

subset(heart, id == 4)

start	stop	event	age	surgery	transplant	id
0	36	0	40.00	0	0	4
36	39	1	40.00	0	1	4

Patient id = 4 waited for 36 days for a heart transplant and then died on the day 39.

subset(heart, id == 1)

start	stop	event	age	surgery	transplant	id
0	50	1	31.00	0	0	1

Patient id = 50 waited for 50 days & died.

Subset (heart, id == 25)

Start	stop	event	age	surgery	transplant	id
0	25	0	33.00	0	0	25
25	1800	0	33.00	0	1	25

This indicates a patient who was alive at the time of the end of the study (roughly 5 years). This patient had a transplant on day 25.

? survfit
? Surv

} Details on survfit(·) and Surv(·) object

km <- survfit (Surv (start, stop, event) ~ 1,
data = heart)

Fits a Kaplan Meier estimator where the survfit(·) and Surv(·) commands are used

plot (km)

Plots the Kaplan-Meier curve along with 95 % confidence intervals

Summary(km, censored = TRUE)

Provides details on the fit for the Kaplan-Meier curve

subset (heart, stop == 1)　　subset (heart, stop == 3)

? Summary.survfit　　Details of summary

The summary calculations are roughly as follows

<u>Time 1</u> : Prob of survival $= \frac{102}{103}$

<u>Time 2</u> : Prob of survival $= \frac{99}{102} \; \frac{102}{103} = 0.961$

$\vdots$

<u>Time 9</u> : Prob of survival $= \frac{90}{91} \; \frac{91}{92} \; \frac{92}{94} \; \frac{94}{96} \; \frac{96}{99} \; \frac{99}{102} \; \frac{102}{103}$

$= 0.874$

<u>Time 11</u> : We have one censored observation without death

Subset (heart, stop = = 11)

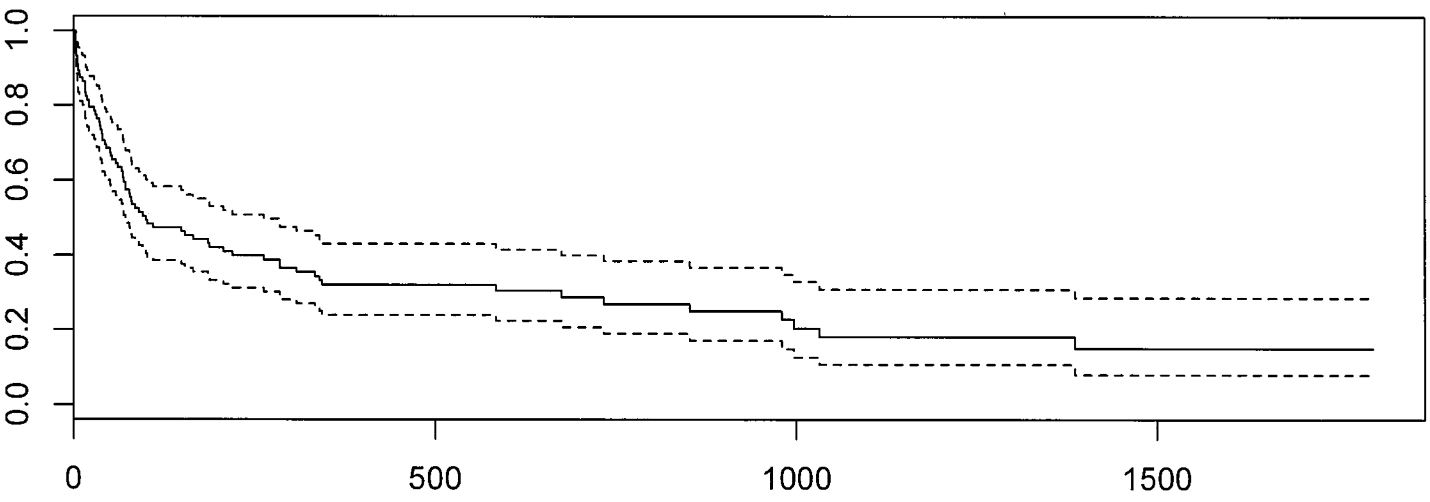
Subset (heart, id = = 102)

Patient data dropped from time 11 (censored)

<u>Time 12</u> : Prob of survival $= \frac{88}{89} \; \frac{90}{91} \; \cdots \; \frac{102}{103}$

$= 0.863$

Note denominator here is 89 & numerator is 88 $\because$ we dropped the patient 102 to compute survival probability here.



Kaplan - Meier survival curve

? coxph   Details of Cox proportional hazard model.

cox <- coxph (Surv (start, stop, event) ~
age + surgery + transplant,
data = heart)

This fits a Cox proportional hazard model where the age, surgery & transplat variables are used to explain survival rates.

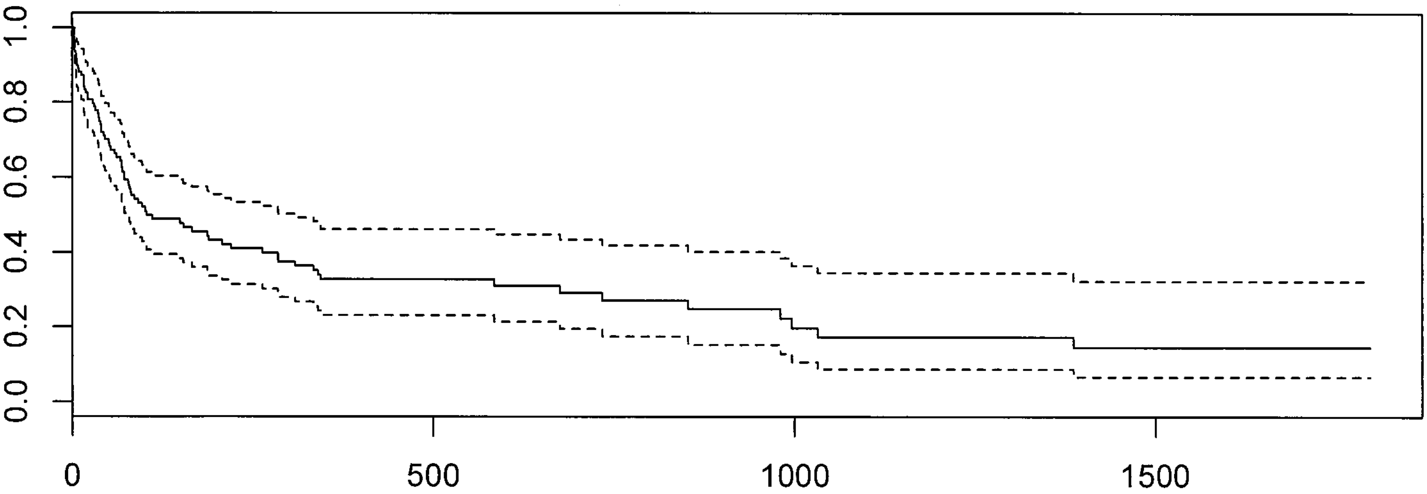
Summary (cox)

	β	e^{β}	
age	0.030	1.03	*
surgery	-0.77	0.46	*
transplant	0.019	1.01	

Signs indicate that the hazard rate increases in age & transplat variables & decreases in surgery variable. You can build models to take joint effects of surgery & transplant. (interaction terms)

Plot (Survfit (cox))

Plots survival fit for Cox PH model



Cox model