Censored and Truncated Data

Goals:

to understand the whole population by accounting for bias in the missing data and using what is known about the missing data to form better conclusions

Definitions:

Truncated: only some of the objects have a value because values outside a certain range are not observable (like a star is too dim) so you're not sure how many objects

Censored: an object is known to exist and some observation has been made, but it is undetected at a certain property or time frame

Upper-Limit (left-censored): the maximum value to which an object may have. The value is too low to be detected.

Functions Frequently used:

Survival Function:

$S(x) = P(X > x)$ → the probability that an object has a value above the specific "x" value

$= 1 - F(x)$

Cummulative Distribution Function

Hazard Rate:

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d \ln S(x)}{dx}$$

$f(x)$ - is the probability density distribution
$S(x)$ - survival function

↳ This function answers the question of "what is the chance that a person will die at a certain age"

Methods:

Cox Regression (proportional hazards)

used to determine the effects of different variables on "survival time".

time between a start time & an event

if the "event" doesn't occur in time, that data point is censored

$h(x) = h_0 e^{-\beta x}$ is the proportional hazards

β estimates are the cox regression

$h(x) = h_0 e^{(\beta_1 x + \beta_2 x + \beta_3 x + \dots)}$

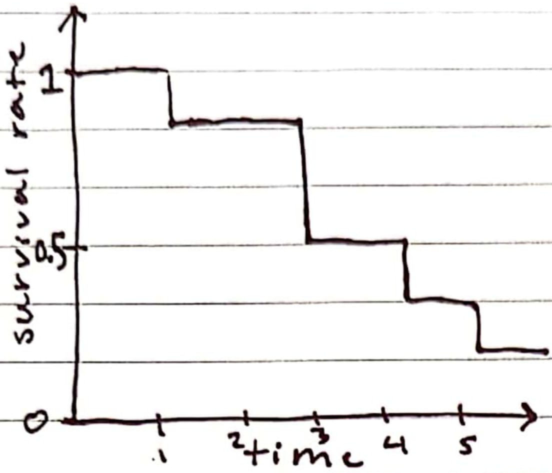
these β values are for each variable or covariate, is the log hazard ratio for the variable

Interpreting results:

covariate	coefficient	hazard ration	interpretation
X	2	7.4	↑ risk
Y	0.01	1.01	– risk
Z	-5	0.007	↓ risk

Coefficients near 0 indicate no change while coef. > 0 are increased "risk" and lower than 0 is less risk.

Kaplan Meier Curve:



Used to graphically represent the survival rate at various times, is used on censored data too

of "objects at x_i

$$S_{KM} = \prod_{x_j \geq x} \left(1 - \frac{d_i}{N_i}\right)$$

# of objects detected or undected

time to event	censored
3	1
4	1
4	1
4	0
6	1
7	1
7	1
8	1
9	0

time	detect.	cens.	n
0	0	0	9
3	1	0	9
4	2	1	8
6	1	0	5
7	2	0	4
8	1	1	2

time			S(t)
0	9/9	=	1
3	8/9 × 1	=	0.89
4	6/8 × 0.89	=	0.6675
6	4/5 × 0.66	=	0.528
7	2/4 × 0.528	=	0.264
8	1/2 × 0.264	=	0.132

these are now the survival times accounting for censoring as a function of time.

For KM to be effective censoring must be random.

Gehan's Test

Used to see if 2 survival curves are significantly different. It is very similar to wilcoxon test.

Sample 1	Sample 2
X_1	X_1
X_2	X_2
X_3	X_3
$\vdots$	$\vdots$
X_n	X_n

This test looks at the values of the samples to see which are larger/smaller. Data can be censored.

$$U_{ij} = \begin{cases} +1 & \text{if } X_i^1 < X_j^2 \\ -1 & \text{if } X_i^1 > X_j^2 \\ 0 & \text{if } X_i^1 = X_j^2 \end{cases}$$

or if ill-determined due to censoring

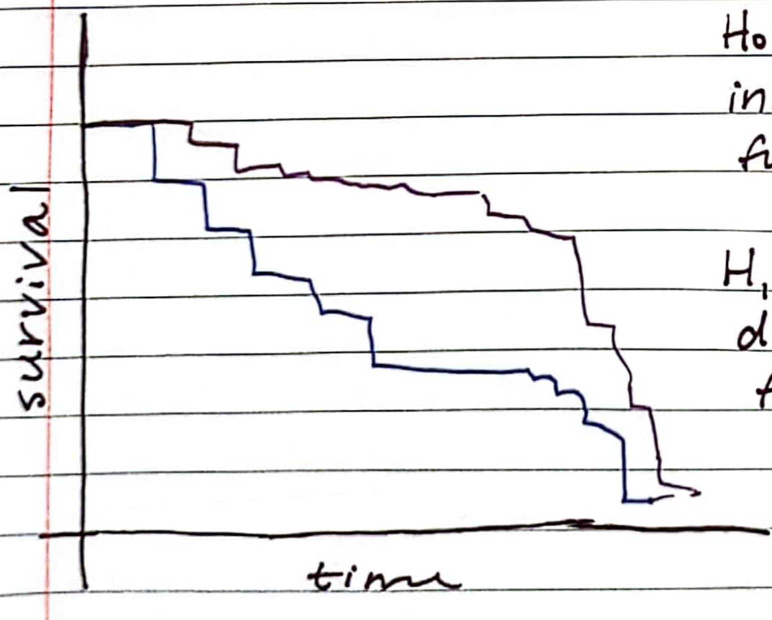
The test statistic is:

$$W_{Gehan} = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

For significance use $z = \dfrac{W_{Gehan}}{\sqrt{W_{Gehan}}}$ to compare to normal distribution

Logrank Test

this test is commonly used to determine if 2 survival curves are significantly different.



H_0: both groups have indentical survival functions

H_1: the group have different survival functions

You could make a chart that looks like

time	Group 1 # at risk $N_{1,j}$	Group 2 # at risk $N_{2,j}$	Group 1 observed $O_{1,j}$	Group 2 Observed $O_{2,j}$

For each event times you can calculate the expected value for the group:

$$E_{i,j} = O_j \frac{N_{i,j}}{N_j} \quad \text{or like} \quad E_{1,j} = O_j \frac{N_{1,j}}{N_j}$$

N_j is total from both

From hear you can calculate the variance:

$$V_{i,j} = E_{i,j} \left(\frac{N_j - O_j}{N_j} \right) \times \left(\frac{N_j - N_{i,j}}{N_j - 1} \right)$$

test statistic:

$$\chi_i^2 = \frac{\left(\sum_{j=1}^{J} O_{i,j} - \sum_{j=1}^{J} E_{i,j} \right)^2}{\sum_{j=1}^{J} V_{i,j}}$$

Correlation

Correlation can be measured with methods like spearman ρ or Kendall τ (discussed in non-parametrics), but w/ censoring the tests change slightly.

$$\tau_H = \frac{n_c - n_d}{\sqrt{\left(\frac{n(n-1)}{2} - n_{t,x} \right) \left(\frac{n(n-1)}{2} - n_{t,y} \right)}}$$

where n_t are the points with unknown relationships.

Lynden-Bell-Woodroofe (LBW) Estimator

used for getting the survival curve for truncated data

$$\frac{S(x)}{S(u_{min})} = P(X \geq x \mid X \geq u_{min})$$

we can only do the survival curve of a specific range where u_i is the sensitivity limit

$$S_{LBW} = \prod_{i: x_i < x} \left(1 - \frac{d_i}{n_i}\right)$$

n_i : is number of points in the set $u_i \leq x \leq x_i$
d_i : # of points at n_i

This is very similar to the KM estimator

Other potential Regression Models

- Accelerated failure-time
- Iterative least squares
- Buckley-James
- Tobit
- Akritas-Thiel-Sen