

PROBLEM SET I, Part 1

Solutions

1. (a) Compare the two naive estimators of the hazard rate in the presence of censoring, $\tilde{\theta} = 1/\bar{t}$ and $\bar{\theta} = \sum d_i / \sum d_i t_i$. In the lecture notes it was shown that both overestimate the hazard rate. Which of these two is more severely biased?

The “ignore” estimator ignores the fact that some observations are censored, and estimates θ as

$$\hat{\theta}_{ignore} = \frac{N}{\sum_{i=1}^N t_i}.$$

The “discard” estimator discards censored observations and estimates θ as:

$$\hat{\theta}_{discard} = \frac{\sum_{i=1}^N d_i}{\sum_{i=1}^N d_i \cdot t_i}.$$

Both overestimate the hazard functions. Which is larger? Write the first estimator as

$$\hat{\theta}_{ignore} = \frac{N}{\sum_{i=1}^N t_i} = \frac{\sum_{i=1}^N d_i + (1 - d_i)}{\sum_{i=1}^N d_i \cdot t_i + (1 - d_i) \cdot c},$$

where c is the censoring time. Note that because $c \geq t_i$, it follows that $\sum(1 - d_i) \cdot c / \sum(1 - d_i) \geq (\sum d_i t_i / \sum d_i)$. Now let $a = \sum d_i$, $b = \sum d_i t_i$, $c = \sum(1 - d_i)$, and $d = \sum(1 - d_i)c$, so that $d/c > b/a$ and

$$\hat{\theta}_{ignore} = \frac{a + c}{b + d}, \quad \text{and} \quad \hat{\theta}_{discard} = \frac{a}{b}.$$

Then:

$$\hat{\theta}_{ignore} - \hat{\theta}_{discard} = \frac{a + c}{b + d} - \frac{a}{b} = \frac{ab + cb - ab - ad}{b(b + d)} = \frac{bc - ad}{b(b + d)} \leq 0,$$

since $bc - ad < 0$, which implies that $\hat{\theta}_{ignore} \leq \hat{\theta}_{discard}$.

- (b) Suppose we select individuals from the pool of unemployed persons and record how long they have been unemployed. Subsequently we interview them again h

periods later, and record only whether they found a job in between or not, but not the actual time they found a job. Write down the likelihood function if the hazard function for person i is $h_i(y)$.

For someone with incomplete duration s_i at the first interview, the probability of finding a job in the next h periods is $(F_Y(s_i + h) - F_Y(s_i))/(1 - F_Y(s_i))$. If d_i is an indicator for finding a job in that period, the likelihood function is

$$\mathcal{L} = \prod_{i=1}^N \left(\frac{F_Y(s_i + h) - F_Y(s_i)}{1 - F_Y(s_i)} \right)^{d_i} \cdot \left(\frac{1 - F_Y(s_i + h)}{1 - F_Y(s_i)} \right)^{1-d_i}.$$

See for example the likelihood function in the Lancaster paper.

2. Use the data in the ascii file DURATION.DAT. The file contains observations on unemployment durations for 4776 men. Each line contains the variables for one observation. The first variable is the duration of an unemployment spell in days (DUR). The second variable is the censoring indicator, equal to one if the spell is censored and equal to zero otherwise (CENS). The third variable is age in years (AGE). The fourth variable is education in years (ED). The fifth variable is an indicator for ethnicity, equal to one if the person is white and zero otherwise (WHITE). The sixth variable is the local unemployment rate (LOCRA). The third to the sixth variable are all measured at the start of the unemployment spell. **In the remainder of this problem set ignore the censoring indicator and the covariates.**

- (a) Calculate the minimum, maximum, mean and median of the durations.

The minimum is 1, the maximum is 1550, the mean is 310.8, and the median is 168 days.

- (b) The hazard rate is $\exp(-\beta)$. What is the value of the log likelihood function at $\beta = 5$.

The log likelihood function is

$$L(\beta) = \sum_{i=1}^N -\beta - Y_i \cdot \exp(-\beta).$$

At $\beta = 5$ this is $L(5) = -33881$.

- (c) Plot the log likelihood function for values of β between zero and ten.

See Figure 1.

- (d) Calculate the analytic first derivatives of the log likelihood function at $\beta = 5$.

The first derivative is

$$\frac{\partial L}{\partial \beta}(\beta) = \sum_{i=1}^N (-1 + Y_i \cdot \exp(-\beta)).$$

At $\beta = 5$ this is $\frac{\partial L}{\partial \beta}(5) = 5224.6$

- (e) Check the analytic value of the first derivative by calculating a numerical approximation to it as:

$$\frac{\partial L}{\partial \beta}(5) \approx \frac{L(5 + c) - L(5)}{c},$$

using $c = 0.00001$. Do the same using $c = 0.000001$.

$$\frac{\partial L}{\partial \beta}(5) \approx \frac{L(5 + 0.00001) - L(5)}{0.00001} = 5224.5,$$

and

$$\frac{\partial L}{\partial \beta}(5) \approx \frac{L(5 + 0.00001) - L(5)}{0.00001} = 5224.6.$$

Both are very close.

- (f) Calculate the analytic second derivative of the log likelihood function at $\beta = 5$.

The second derivative is

$$\frac{\partial^2 L}{\partial \beta^2}(\beta) = \sum_{i=1}^N -Y_i \cdot \exp(-\beta).$$

At $\beta = 5$ this is $\frac{\partial^2 L}{\partial \beta^2}(5) = -10001$.

- (g) Compare this to a numerical second derivative calculated as

$$\frac{\partial^2 L}{\partial \beta^2}(5) \approx \frac{\frac{\partial L}{\partial \beta}(5 + c) - \frac{\partial L}{\partial \beta}(5)}{c},$$

using $c = 0.00001$.

$$\frac{\partial^2 L}{\partial \beta^2}(5) \approx \frac{L(5 + 0.00001) - L(5)}{0.00001} = -10001.$$

- (h) Calculate the maximum likelihood estimate for β using Newton-Raphson. Use analytic first and second derivatives. Start at $\beta = 0$. Report at the beginning of each iteration the current value, the first derivative at this point, the direction, the step, and the criterion for determining convergence. The criterion for determining convergence is the sum of the absolute value of the first derivative and the absolute value of the change of the parameter value. (So, for the first iteration, report the current value (zero), the first derivative at zero, the step (the new value minus the old one), and the criterion).

See Table 1

Table 1: NEWTON-RAPHSON ITERATIONS

iteration	current value	first der	direction	step	criterion
1	0	-1479400	0.9968	0.4992	1479400
2	0.9968	-543000	0.9913	0.4978	543000
3	1.9881	-198500	0.9765	0.4941	198500
4	2.9646	-71784	0.9376	0.4839	71785
5	3.9022	-25202	0.8407	0.4567	25203
6	4.7429	-8156.9	0.6307	0.3868	8157.5
7	5.3736	-2107.1	0.3061	0.2344	2107.4
8	5.6797	-291.98	0.0576	0.0545	292.03
9	5.7373	-8.2514	0.0017	0.0017	8.2531
10	5.7390	-0.0071	0.0000	0.0000	0.0071
11	5.7390	-0.0000	0.0000	0.0000	0.0000

- (i) Calculate the maximum likelihood estimate for β using the golden section approach. We will do this in a couple of steps. The starting value is $\beta_l = 0$, and the initial high value is $\beta_h = 1$. (You can assume that we know the maximum likelihood estimate is greater than zero).
- i. First we determine initial values for the interval. Calculate the value of minus the log likelihood at β_l and β_h . If the value at β_h is higher than at β_l you are done. If not, (*) calculate the value of minus the log likelihood

at $\beta_l + 3(\beta_h - \beta_l)$. Now if this last value is higher than the value at β_h , set $\beta_h = \beta_l + 3(\beta_h - \beta_l)$ and you are done again. If not, replace (β_l, β_h) by $(\beta_h, \beta_l + 3(\beta_h - \beta_l))$ and go back to (*).

In the end you are left with values (β_l, β_h) such that the maximum likelihood estimate must be in this interval. Report these values and the value of minus the log likelihood function at these points

- ii. Now we will use the golden section method to determine the optimum within this range. Iterate till the width of the interval is less than 0.0000001. Report for each iteration the lower and upper bounds and the distance between them. See Table 2

Table 2: GOLDEN SECTION ITERATIONS

iteration	β_l	β_h	$\beta_h - \beta_l$
0	3.0000	15.0000	12.0000
1	3.0000	10.4164	7.4164
2	3.0000	7.5836	4.5836
3	4.7508	7.5836	2.8328
4	4.7508	6.5016	1.7508
5	5.4195	6.5016	1.0820
6	5.4195	6.0883	0.6687
7	5.4195	5.8328	0.4133
8	5.5774	5.8328	0.2554
9	5.6749	5.8328	0.1579
10	5.6749	5.7725	0.0976
11	5.7122	5.7725	0.0603
12	5.7122	5.7495	0.0373
13	5.7265	5.7495	0.0230
14	5.7352	5.7495	0.0142
15	5.7352	5.7440	0.0088
16	5.7352	5.7407	0.0054
17	5.7373	5.7407	0.0034
18	5.7386	5.7407	0.0021
19	5.7386	5.7399	0.0013
20	5.7386	5.7394	0.0008
21	5.7389	5.7394	0.0005
22	5.7389	5.7392	0.0003
23	5.7389	5.7391	0.0002
24	5.7390	5.7391	0.0001
25	5.7390	5.7391	0.0001
26	5.7390	5.7391	0.0000
27	5.7390	5.7391	0.0000
28	5.7390	5.7390	0.0000
29	5.7390	5.7390	0.0000
30	5.7390	5.7390	0.0000
31	5.7390	5.7390	0.0000
32	5.7390	5.7390	0.0000
33	5.7390	5.7390	0.0000
34	5.7390	5.7390	0.0000
35	5.7390	5.7390	0.0000
36	5.7390	5.7390	0.0000
37	5.7390	5.7390	0.0000
38	5.7390	5.7390	0.0000
39	5.7390	5.7390	0.0000