ST227 Survival Models - Part III Estimation in the Markov Model

George Tzougas
Department of Statistics,
London School of Economics

January 31, 2019

1 Markov Models: Data and Estimation

1.1 Introduction

- We now have mathematical **models for various situations that arise in insurance**, e.g. a model of mortality with states $\{Alive, Dead\}$, a model for PHI with states $\{Well, Ill, Dead\}$.

- These **models depend on transition intensities** μ_x^{gh}, $g \neq h$ between states g and h.

- The values of these **transition intensities are unknown** - they are parameters in the models.

- Thus, **we need some data which we can use to estimate the values of the transition intensities μ_x^{gh}**.

- We will start with the 2-state model of mortality with states $\{Alive, Dead\}$, and work through this example in detail.

- It turns out that the **answers for the 2-state model give a very good guide to what happens in the multi-state case**.

1.2 Data and notation for the 2-state model

- Suppose we have **data on n independent lives all aged between x and** $x+1$. We suppose for life i:

$x + \alpha_i$ = age at which **observation begins**, $0 \le \alpha_i < 1$

$x + b_i$ = age at which **observation ends, if life i does not die**, $0 \le b_i < 1$

$x + t_i$ = age at which **observation stops, by death or censoring**

It will often be useful to work with the time spent under observation, so we define

$v_i = t_i - \alpha_i$, the observed **waiting time**

Also, we define

d_i = 1, **if life i is observed to die**

d_i = 0, **if life i is not observed to die**

Note:

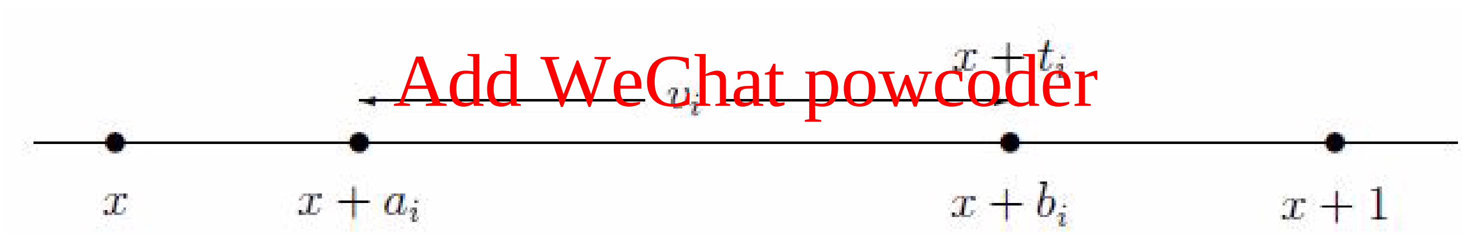
- *Censoring is a form of missing data problem in which time to event is not observed for reasons such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing an event.*

- *Censoring is the key feature of survival data (indeed survival analysis might be defined as the analysis of censored data) and the mechanisms which give rise to censoring play an important part in statistical inference.*

- *Censoring is present when we do not observe the exact length of a lifetime, but observe only that its length falls within some interval. This can happen in several ways which we are going to discuss in detail in the following lectures.*

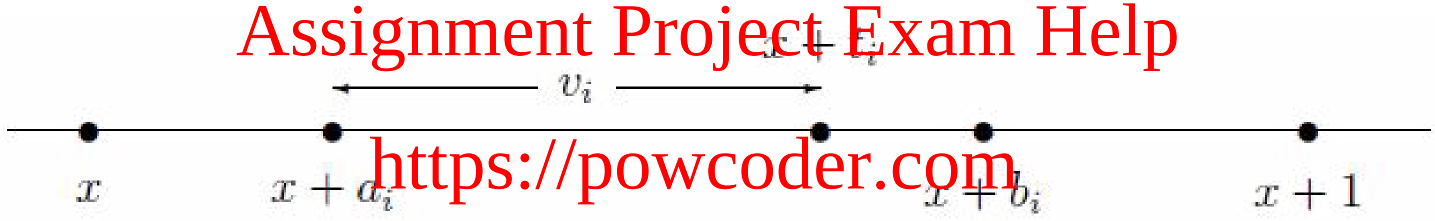
The following diagrams may be helpful.

Case 1. The ith **life does not die**, i.e. reaches the end of the observation period at $x + b_i$.



In this case, $d_i = 0$, $t_i = b_i$, $v_i = b_i - \alpha_i$.

Case 2. The ith **life dies** at $x + t_i$ before the end of the observation period.



In this case, $d_i = 1$, $\alpha_i < t_i < b_i$, $v_i = t_i - \alpha_i$.

1.3 The random variables D_i and V_i

In the simple two state model the observations on the ith life are: whether the life died or not, i.e. the value d_i, and the observed waiting time, i.e. the value v_i. We can think of the observations d_i and v_i as realisations of random variables D_i and V_i. Inference for the transition intensity μ_x will depend on the joint distribution of D_i and V_i. Before deriving this distribution we make two important points:

- D_i and V_i are not independent, since as it was previously shown

- V_i has a mixed distribution, i.e. the probability distribution has a **continuous density** for values of v_i which correspond to **death**, i.e. for

$\upsilon_i < b_i - \alpha_i$, and **a mass of probability (i.e. discrete distribution)** for the value of υ_i which corresponds to **survival**, i.e. for $\upsilon_i = b_i - \alpha_i$.

Assumption: Before deriving the joint distribution of D_i and V_i we will make a further assumption: we assume that $\mu_{x+t} = \mu_x = \mu$ for $0 \leq t \leq 1$, i.e. **the transition intensity μ_{x+t} is constant from age x to age $x+1$**. We can now prove the following result:

Result: The joint probability function of D_i and V_i can be written

$$f_i(d_i, \upsilon_i) = e^{-\mu \upsilon_i} \mu^{d_i} \tag{1}$$

where $\mu_{x+t} = \mu \quad$ for $0 \leq t \leq 1$.

Proof: We consider the two cases: $D_i = 0$ and $D_i = 1$.

Case 1: $D_i = 0$. We need the probability the life survives from $x + \alpha_i$ to $x + b_i$. We have

$$
\begin{aligned}
P_r(D_i &= 0, V_i = b_i - \alpha_i) \\
&= P_r(\text{Life } i \text{ survives from } x + \alpha_i \text{ to } x + b_i) \\
&= \exp\left(-\int_0^{\upsilon_i} \mu_{x+\alpha_i+s} ds\right), \text{ where } \upsilon_i = b_i - \alpha_i \\
&= \exp(-\mu\upsilon_i), \text{ since } \mu_{x+\alpha_i+s} = \mu \text{ is constant} \\
&= e^{-\mu\upsilon_i}\mu^{d_i}, \text{ since } d_i = 0 \qquad (2)
\end{aligned}
$$

Case 2: $D_i = 1$. We need the probability density function that a life aged $x + \alpha_i$ dies at time $x + \alpha_i + \upsilon_i$. We have

$$
\begin{aligned}
P_r(D_i &= 1, \upsilon_i \leq V_i < \upsilon_i + d\upsilon_i) = {}_{\upsilon_i}p_{x+\alpha_i} \times_{du_i} q_{x+\alpha_i+\upsilon_i} \\
&= \exp\left(-\int_0^{\upsilon_i} \mu_{x+\alpha_i+s} ds\right) \times (\mu_{x+\alpha_i+\upsilon_i} d\upsilon_i + o(d\upsilon_i)) \\
&= \exp(-\mu\upsilon_i) \times (\mu d\upsilon_i + o(d\upsilon_i)) \text{ since } \mu_{x+t} = \mu \text{ is constant} \\
&= e^{-\mu\upsilon_i}\mu^{d_i} d\upsilon_i + o(d\upsilon_i), \text{ since } d_i = 1. \qquad (3)
\end{aligned}
$$

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Dividing Eq.(3) by $d\upsilon_i$ and taking the limit as $d\upsilon_i \longrightarrow 0$, we get:

$$f_i(d_i, \upsilon_i) = e^{-\mu\upsilon_i}\mu^{d_i}. \qquad (4)$$

- Thus, in **both cases, Eq.(2) and Eq.(4),** we can write the **probability function** as:

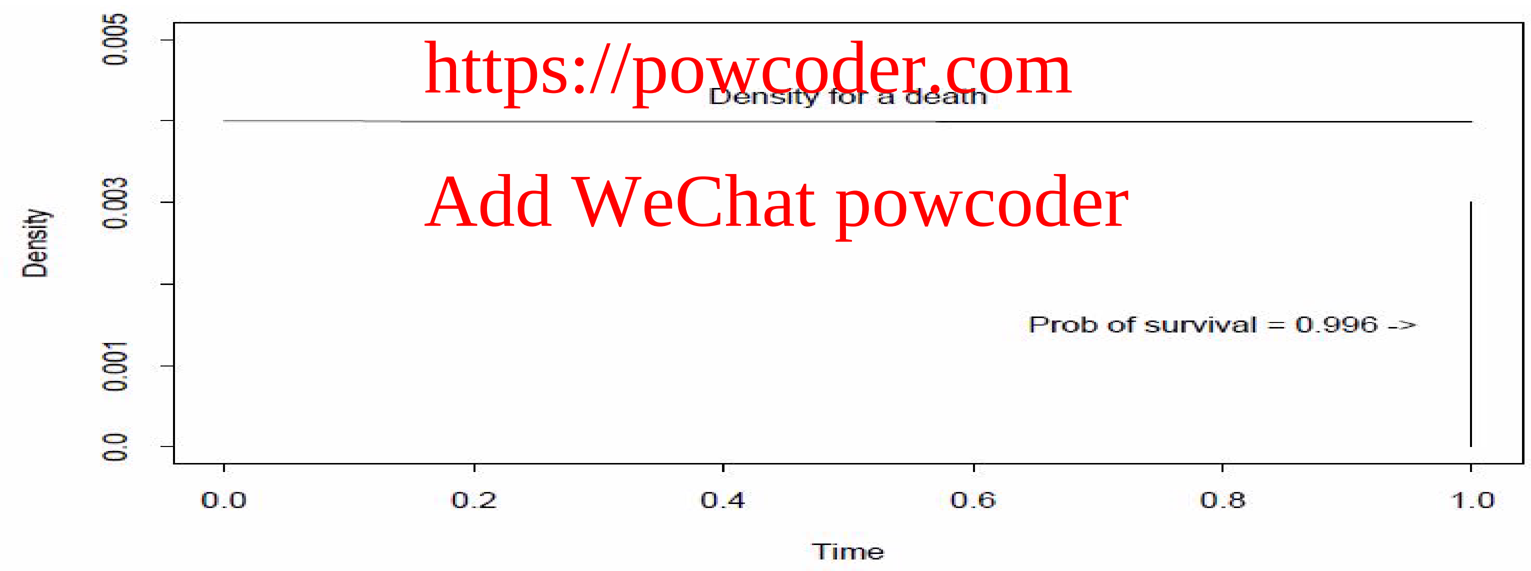
$$f_i(d_i, v_i) = e^{-\mu v_i} \mu^{d_i} \tag{5}$$

- But what does the mixed distribution of D_i and V_i look like? Here is a simple example: take $a_i = 0$, $b_i = 1$, i.e. the observation period is the full year of life, unless death occurs.

- Suppose $\mu = 0.004$, the approximate force of mortality for a male life aged 50.

- The probability of survival is $e^{-0.004} = 0.996$. The density function for a death is $\mu e^{-\mu t} = 0.004 e^{-0.004t}$, $0 \leq t \leq 1$.

- In the plot of the mixed distribution below note that the probability mass for survival at 1 is drawn on a different scale from the density function for death.



1.4 The maximum likelihood estimate of μ

Suppose that the waiting times of the n lives are $v' = (v_1, v_2, ..., v_n)$ and the record of deaths is $d' = (d_1, d_2, ..., d_n)$. The n lives are independent so, from (5), the total likelihood is

$$L(\mu; D, v) = \prod_{i=1}^{n} f_i(d_i, v_i) = \prod_{i=1}^{n} e^{-\mu v_i} \mu^{d_i} = e^{-\mu \sum v_i} \mu^{\sum d_i}. \qquad (6)$$

We are interested in maximizing $L(\mu; D, v)$ with respect to μ.

Firstly, we write $L(\mu; D, \upsilon)$ in more compact form as:

$$L(\mu) = e^{-\mu\upsilon}\mu^{d}, \tag{7}$$

where

$$\begin{aligned} \upsilon &= \sum \upsilon_i = \text{ Total waiting time}, \\ d &= \sum d_i = \text{ Total number of deaths.} \end{aligned} \tag{8}$$

Then, the log-likelihood will be:

$$
\begin{aligned}
l(\mu) &= \log\left(L(\mu)\right) = \log\left(e^{-\mu\upsilon}\mu^{d}\right) \\
&= \log\left(e^{-\mu\upsilon}\right) + \log(\mu^{d}) \\
&= -\mu\upsilon + d\log(\mu)
\end{aligned}
\tag{9}
$$

We need to find the maximum by finding the derivative:

$$
\frac{\partial l(\mu)}{\partial \mu} = -\upsilon + \frac{d}{\mu} = 0, \tag{10}
$$

The maximum likelihood estimate of μ follows from (10) as

$$\widehat{\mu} = \frac{d}{v} = \frac{\text{Total number of deaths}}{\text{Total waiting time}} \tag{11}$$

- We are interested in the **statistical properties of** $\widehat{\mu}$. For this purpose, we **think of** $L(\mu)$ **and** $\widehat{\mu}$, see Eqs (7 and 11) respectively, **as random variables.**

- Corresponding to Eq.(7) we define the random variable

$$L(\mu; D, V) = e^{-\mu V} \mu^{D}, \tag{12}$$

where $D = \sum D_i$ is the random total number of deaths and $V = \sum V_i$ is the random *total waiting time*.

- Corresponding to Eq.(11) we define the estimator

$$\tilde{\mu} = \frac{D}{V}. \tag{13}$$

Assignment Project Exam Help

Comment: The previous **maximum likelihood estimation** procedure concerned **lives all aged between** x **and** $x + 1$, e.g. from 20 to 21. An obvious question is: "How do we extend the estimation process to a wider range of ages, e.g. from 20 to 70?" One way of doing this is to split the age range into a number of short age ranges in such a way that the μ_x can be assumed constant within each sub-range. Unfortunately, this solution creates another problem: there is no reason to expect that the large number of estimates of μ_x for different x-values will be consistent. For example, we expect mortality to increase, at least from age 20, say. We will be forced to smooth (or **graduate)** these raw rates.

https://powcoder.com

Add WeChat powcoder

1.5 The Maximum Likelihood Theorem (MTL)

- As was previously mentioned we are interested in investigating the **statistical properties of** $\widehat{\mu}$, such as **asymptotic unbiasedness, asymptotic efficiency and asymptotic normality.** For this purpose, we defined $L(\mu; D, V)$ and $\widetilde{\mu}$ as the random variables corresponding to $L(\mu)$ and $\widehat{\mu}$.

- In order to be able to carry on with our investigation, we have make use of two important general results from mathematical statistics:
  - properties of the **score function**, and
  - the **Maximum Likelihood Theorem**.

Here we give a brief review of this material. First, we consider the one parameter case, we maximize the log-likelihood with respect to the unknown parameter θ.

Definition: The **score function**. Suppose $\mathbf{X}' = (X_1, X_2, ..., X_n)$ is a vector of random variables where the X_i are independent random variables with probability function $f_i\,(x, \theta)$ which depends on a parameter μ

Let

$$l = l(\theta) = l(\theta, \mathbf{X}) = \sum_1^n \log f_i\,(X_i, \theta) \tag{14}$$

be the **log-likelihood function (viewed as a random variable)**. Then

$$U = U(\theta) = \frac{\partial l}{\partial \theta} \tag{15}$$

is called the **score function**.

Assignment Project Exam Help

Note: The score function is **additive** as follows: let

https://powcoder.com

Add WeChat powcoder

$$U_i = U_i(\theta) = \frac{\partial l_i}{\partial \theta} = \frac{\partial}{\partial \theta} \log f_i\left(X_i, \theta\right) \tag{16}$$

be the score function for the ith observation, $i = 1, ..., n$. Then

$$U = \sum_{1}^{n} U_i \tag{17}$$

Result: The score function satisfies:

(a) $E(U) = 0$, and

(b) $Var(U) = -E\left(\frac{\partial U}{\partial \theta}\right) = -E(\frac{\partial^2 l}{\partial \theta^2})$.

Proof: First, **take n = 1** and without loss of generality $f_i(x,\theta) = f(x,\theta)$ (or for simplicity's sake f).

(a) Consider the identity

$$\int f(x,\theta)dx = 1. \tag{18}$$

Differentiate (18) with respect to θ. We get

$$\int \frac{\partial f}{\partial \theta} dx = 0.$$

Rewrite this equation as

$$\int \left(\frac{\partial}{\partial \theta} \log f \right) f dx = 0. \tag{19}$$

Note that, $U = \frac{\partial}{\partial \theta} \log f$ and $E(U) = E\left(\frac{\partial}{\partial \theta} \log f\right)$. Thus, the left hand side of this equation gives $E(U) = 0$ for $n = 1$. Hence $E(U) = 0$ for all n by the additivity property, see Eq. (17).

(b) The proof uses the same trick. In particular we differentiate Eq. (19), with respect to θ and we get

$$\frac{\partial}{\partial\theta}\left\{\int\left(\frac{\partial}{\partial\theta}\log f\right) f\,dx\right\} = 0 \tag{20}$$

Hence, using the product rule we have that

$$\int \left\{ \frac{\partial}{\partial\theta} \left(\frac{\partial}{\partial\theta} \log f \right) f + \left(\frac{\partial}{\partial\theta} \log f \right) \frac{\partial f}{\partial\theta} \right\} dx = 0$$

$$\Longrightarrow \int \frac{\partial}{\partial\theta} \left(\frac{\partial}{\partial\theta} \log f \right) f\,dx + \int \left(\frac{\partial}{\partial\theta} \log f \right) \left(\frac{\partial}{\partial\theta} \log f \right) f\,dx = 0$$

$$\Longrightarrow E\left(\frac{\partial U}{\partial\theta} \right) + E(U^2) = 0, \text{ since } U = \frac{\partial}{\partial\theta} \log f$$

$$\Longrightarrow E(U^2) = -E(\frac{\partial U}{\partial\theta})$$

$$\Longrightarrow Var(U) = E(U^2) - E^2(U) = -E(\frac{\partial U}{\partial\theta}), \text{ since } E(U) = 0$$

$$\text{or } Var(U) = -E(\frac{\partial^2 l}{\partial\theta^2}), \text{since } \frac{\partial U}{\partial\theta} = \frac{\partial^2}{\partial\theta^2} \log f = \frac{\partial^2 l}{\partial\theta^2}$$

and **(b)** is proved for $n = 1$. The general case follows by the additivity of the score function, see Eq. (17), and the independence of the observations.

- It is important to note that the score function U is a fundamental quantity in statistics and it is connected with the properties of the maximum likelihood estimate of θ. In particular,
  - The MLE satisfies $\partial l/\partial\theta = 0$ i.e. $U = 0$
  - The asymptotic variance of the MLE is found from $Var(U)$.

- In what follows we will define the **Fisher information function** (sometimes simply called information function) based on U.

- The **Fisher information function** is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Formally, **it is the variance of the score** $Var(U)$, or $-E\left(\frac{\partial U}{\partial \theta}\right)$, since as was proved in **(b)** $Var(U) = -E\left(\frac{\partial U}{\partial \theta}\right)$.

Definition: **Fisher's information function** $I(\theta)$ is

$$I(\theta) = -E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E\left(\frac{\partial U}{\partial \theta}\right) = Var(U) \tag{21}$$

We usually just call $I(\theta)$ the information function. We can now state (without proof) the **Maximum Likelihood Theorem (MLT)**.

Theorem: Suppose $\widehat{\theta}$ satisfies $U = 0$ and that θ_0 is the true (but unknown) value of θ. If $\widetilde{\theta}$ is the random variable corresponding to $\widehat{\theta}$, then $\widetilde{\theta}$ is

(i) asymptotically unbiased, i.e. $E(\widetilde{\theta}) \to \theta_0$,

Assignment Project Exam Help

(ii) asymptotically efficient, i.e. $Var(\widetilde{\theta}) \to I(\theta_0)^{-1}$,

https://powcoder.com

Add WeChat powcoder

(iii) asymptotically normal, i.e.

$$\widetilde{\theta} \approx \mathcal{N}(\theta_0, I(\theta_0)^{-1}). \tag{22}$$

Of course, θ_0 is unknown so how do we calculate $I(\theta_0)^{-1}$? There are two main approximations.

- Replace θ_0 with $\widehat{\theta}$, i.e.

$$I(\theta_0) \approx I(\widehat{\theta}) = -E\left(\frac{\partial U}{\partial \theta}\right)]_{\theta=\widehat{\theta}} = -E\left(\frac{\partial^2 l}{\partial \theta^2}\right)]_{\theta=\widehat{\theta}}$$

- If we can't work out the expectation in $E\left(\frac{\partial U}{\partial \theta}\right)$ then try

$$I(\theta_0) \approx -\left(\frac{\partial^2 l}{\partial \theta^2}\right)]_{\theta=\widehat{\theta}} \tag{23}$$

1.6 Properties of $\tilde{\mu} = D/V$ in the 2-state model

The maximum likelihood theorem is a most powerful and general result and we apply it now to the problem of determining the properties of the MLE of the force of mortality in the 2-state model.

We know from section 1.4 that

$$\begin{aligned} L(\mu) &= e^{-\mu V} \mu^D \\ &\Rightarrow l(\mu) = -\mu V + D \log \mu \\ &\Rightarrow U(\mu) = -V + \frac{D}{\mu} \\ &\Rightarrow \tilde{\mu} = \frac{D}{V} \end{aligned} \tag{24}$$

Now apply the Maximum Likelihood Theorem. Suppose that μ_0 **is the true**, but unknown, value of the unknown **force of mortality**.

(i) First, $\widetilde{\mu}$ is asymptotically unbiased so

$$E(\widetilde{\mu}) \longrightarrow \mu_0 \tag{25}$$

(ii) Second, $\widetilde{\mu}$ is asymptotically efficient so

$$
\begin{aligned}
Var\left(\tilde{\mu}\right) &\longrightarrow I(\mu_0)^{-1} \\
&= \left\{ -E\left(\frac{\partial U}{\partial \mu}\right) \right\}^{-1}]_{\mu=\mu_0} \\
&= -\left\{ -E\left(-\frac{D}{\mu^2}\right) \right\}^{-1}]_{\mu=\mu_0}, \qquad \text{by differentiating (24)} \\
&= \left(\frac{E(D)}{\mu^2}\right)]_{\mu=\mu_0} \\
&= \frac{\mu^2}{E(D)}]_{\mu=\mu_0}
\end{aligned}
$$

Hence,

$$Var\left(\tilde{\mu}\right) \longrightarrow I(\mu_0)^{-1} = \frac{\mu^2}{E(D)}]_{\mu=\mu_0}. \tag{26}$$

Now $E(U) = 0$ so from (24) we get

$$E\left(\frac{D}{\mu} - V\right) = 0 \Rightarrow \frac{E\left(D\right)}{\mu} = E\left(V\right) \Rightarrow E(D) = \mu E(V). \tag{27}$$

Thus, from Eqs(26 and 27) we get

$$Var(\tilde{\mu}) \longrightarrow \frac{\mu}{E(V)}]_{\mu=\mu_0} = \frac{\mu_0}{E(V)}. \tag{28}$$

(iii) Third, $\widetilde{\mu}$ is asymptotically normal.

Our final result is

$$\widetilde{\mu} \approx N\left(\mu_0, \frac{\mu_0}{E(V)}\right). \tag{29}$$

Comments: We note the following

- The observed waiting time $v = \sum v_i$ is usually called the **central exposed to risk** E_x^c in actuarial language.

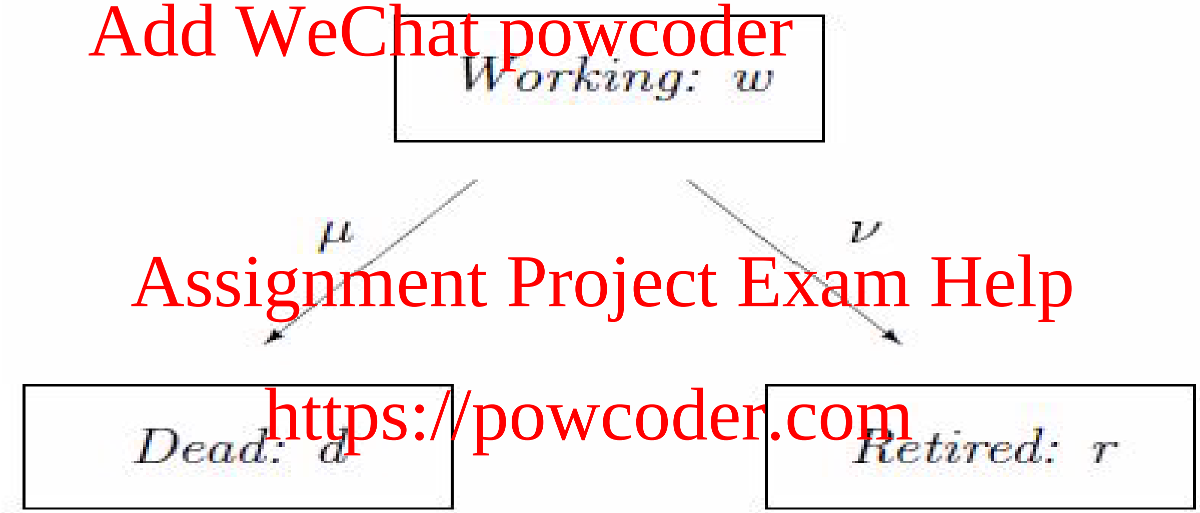
- In practice, we evaluate $Var(\widetilde{\mu}) = \mu_0/E(V)$ by

$$Var(\widetilde{\mu}) = \frac{\mu_0}{E(V)} \approx \frac{\widehat{\mu}}{v} = \frac{d/v}{v} = \frac{d}{v^2} \tag{30}$$

How does the estimation result (29) for the 2-state model extend to more complicated multi-state models such as Illness/Death models?

1.7 Estimation in the 3-state model

In Part II we considered the following 3-state model for working, retiring and dying.



Suppose we have data on n lives and for the ith life we observe

$$
\begin{aligned}
d_i &= \begin{cases} 0 & \text{when } i\text{th life does not die} \\ 1 & \text{when } i\text{th life does die} \end{cases} \qquad (31a) \\
r_i &= \begin{cases} 0 & \text{when } i\text{th life does not retire} \\ 1 & \text{when } i\text{th life does retire} \end{cases} \\
v_i &= t_i - a_i, \quad \text{the } i\text{th waiting time}
\end{aligned}
$$

- Our observation is the triple (v_i, d_i, r_i) with corresponding underlying random variable (V_i, D_i, R_i). The distribution of (V_i, D_i, R_i) is given by

$$
P_r(D_i = d_i, R_i = r_i, v_i < V_i < v_i + dv_i). \qquad (32)
$$

- Notice that once more this is a mixed distribution with a mass of probability corresponding to remaining in the working state and a continuous density function corresponding to retirement or death.

- The key result is ${}_tp_x^{ww} = {}_tp_x^{\overline{ww}}$ (occupation probability) since in this case state w cannot be re-entered once it has been left).

$$
{}_tp_x^{ww} = \exp\left(-\int_0^t (\mu+\nu)\, ds \right) = e^{-(\mu+\nu)t} \tag{33}
$$

We discuss the three cases: (i) remaining working, (ii) dying, (iii) retiring.

Case 1: Remaining working. In this case $d_i = 0$, $r_i = 0$ and the life remains in the working state from $x+\alpha_i$ to $x+b_i$ so we have

$$
\begin{aligned}
P_r(D_i = 0, R_i = 0, V_i = b_i - \alpha_i) &\\
&= P_r(\text{Neither died nor retired during } x + \alpha_i \longrightarrow x + b_i) \\
&= {}_{b_i - \alpha_i}\, p_x^{ww} \\
&= e^{-(\mu+\nu)(b_i - \alpha_i)}, \text{ from (33) for } t = b_i - \alpha_i \\
&= e^{-(\mu+\nu)\upsilon_i}
\end{aligned}
$$

since $\upsilon_i = b_i - \alpha_i$ when $d_i = 0, r_i = 0$.

Hence, the probability density function (pdf) in this case is

$$
f_i(0, 0, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}. \tag{34}
$$

Case 2: Dying. In this case $d_i = 1$, $r_i = 0$, $\upsilon_i < V_i < \upsilon_i + d\upsilon_i$ so we have

$$
\begin{aligned}
P_r(D_i &= 1, R_i = 0, \upsilon_i < V_i < \upsilon_i + d\upsilon_i) \\
&= P_r(\text{Neither died nor retired during } (x + \alpha_i, x + \alpha_i + \upsilon_i)\ldots \\
&\quad \ldots \textbf{ and } \text{then died in } (x + \alpha_i + \upsilon_i, x + \alpha_i + \upsilon_i + d\upsilon_i)) \\
&= {}_{\upsilon_i}p^{ww}_{x+\alpha_i} \times {}_{d\upsilon_i}q_{x+\alpha_i+\upsilon_i} \\
&= e^{-(\mu+\nu)\upsilon_i} \times (\mu d\upsilon_i + o(d\upsilon_i))
\end{aligned}
$$

Dividing by $d\upsilon_i$ and taking the limit we find the pdf is

$$f_i(1, 0, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}\mu. \tag{35}$$

.

Case 3: Retiring. In this case $d_i = 0$, $r_i = 1$, $\upsilon_i < V_i < \upsilon_i + d\upsilon_i$ so by symmetry with case 2 we have

$$P_r(D_i = 0, R_i = 1, \upsilon_i < V_i < \upsilon_i + d\upsilon_i) = e^{-(\mu+\nu)\upsilon_i} \times (\nu d\upsilon_i + o(d\upsilon_i))$$

and the pdf in this case is

$$f_i(0, 1, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}\nu. \tag{36}$$

In summary, see Eqs(34,35 and 36) we have

- Case 1: $d_i = 0$, $r_i = 0 \Rightarrow f_i(0, 0, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}$,

- Case 2: $d_i = 1$, $r_i = 0 \Rightarrow f_i(1, 0, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}\mu$,

- Case 3: $d_i = 0$, $r_i = 1 \Rightarrow f_i(0, 1, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i}\nu$.

All three cases can conveniently be packaged as

$$f_i(d_i, r_i, \upsilon_i) = e^{-(\mu+\nu)\upsilon_i} \times \mu^{d_i} \times \nu^{r_i} \tag{37}$$

and so, assuming that all n lives are independent, we obtain the likelihood as

$$
\begin{aligned}
L(\mu,\nu) &= \prod_{1}^{n} e^{-(\mu+\nu)\upsilon_i} \times \mu^{d_i} \times \nu^{r_i} \\
L(\mu,\nu) &= e^{-\mu\upsilon}\mu^{d} \times e^{-\nu\upsilon}\nu^{r} \qquad (38)
\end{aligned}
$$

where $\upsilon = \sum \upsilon_i$ is the total waiting time, $d = \sum d_i$ is the total number of deaths and $r = \sum r_i$ is the total number of retirals.

Notice that $L(\mu,\nu)$ **factorises** into a product of two functions, one of μ and one of ν. As a result the log-likelihood has the particularly simple form

$$l(\mu,\nu) = -\mu\upsilon + d\log\mu - \nu\upsilon + r\log\nu \qquad (39)$$

and the estimators of μ and ν are easily obtained by computing the partial derivatives of the log-likelihood with respect to μ and ν and setting them equal to 0:

$$\widetilde{\mu} = \frac{D}{V}, \ \widetilde{\nu} = \frac{R}{V}. \tag{40}$$

Assignment Project Exam Help

https://powcoder.com

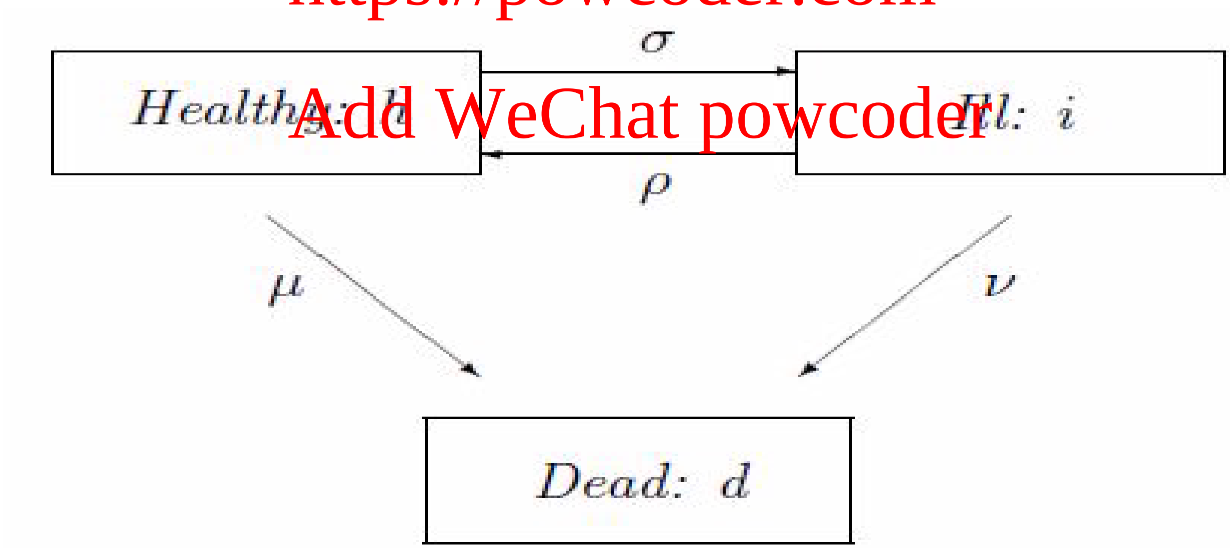
Add WeChat powcoder

Comment: There are obvious parallels between

- (37) and the likelihood $L(\mu) = e^{\mu v}\mu^d$ in the 2-state model, and
- (39) and the estimator $\widetilde{\mu} = D/V$ in the 2-state model.

1.8 More estimation in the 3-state model

Our second example of the 3-state model is the model for PHI discussed in Part II. Suppose that the transition intensities are constant for the n lives in the investigation and consider the diagram:

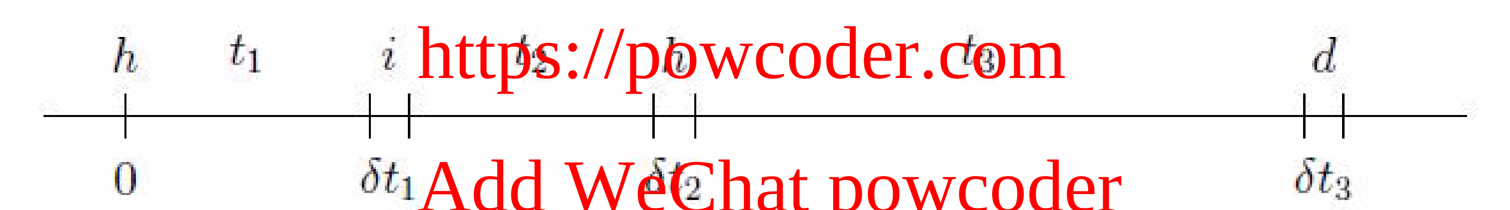


Suppose that the observation on the ith life is

$$
\begin{aligned}
v_i &= \text{total waiting time in state } h \qquad (41)\\
w_i &= \text{total waiting time in state } i\\
s_i &= \text{number of transitions from } h \text{ to } i\\
r_i &= \text{number of transitions from } i \text{ to } h\\
d_i &= 1 \text{ if } h \longrightarrow d, 0 \text{ otherwise}\\
u_i &= 1 \text{ if } i \longrightarrow d, 0 \text{ otherwise.}
\end{aligned}
$$

A straightforward extension of the argument in the previous example for the ith life gives the likelihood. An example may help.

Example: Consider a life that is initially healthy, falls ill after time t_1, recovers after a further time t_2 and then dies after a further time t_3. As always we consider the events falling ill, recovering and dying as occurring in a short interval of time, as in the diagram



The probability of observing this sequence of events is

$$e^{-(\mu+\sigma)t_1}\left[\sigma\delta t_1 + 0(\delta t_1)\right]\times e^{-(\rho+\nu)t_2}\left[\rho\delta t_2 + o(\delta t_2)\right]\times e^{-(\mu+\sigma)t_3}\left[\mu\delta t_3 + 0(\delta t_3)\right].$$

Dividing by $\delta t_1 \delta t_2 \delta t_3$ and taking the limit as $\delta t_i \to 0$, $i = 1, 2, 3$ we find the contribution to the likelihood from this life is

$$e^{-(\mu+\sigma)(t_1+t_3)} e^{-(\rho+\nu)t_2} \mu\sigma\rho\nu^0. \tag{42}$$

The general formula for the likelihood $L(\mu, \nu, \sigma, \rho)$ follows by multiplying up over all lives:

$$\begin{aligned} &\prod_1^n e^{-(\sigma+\mu)v_i} \times e^{-(\rho+\nu)\omega_i} \times \sigma^{s_i} \times \rho^{r_i} \times \mu^{d_i} \times \nu^{u_i} \\ = \;& e^{-(\sigma+\mu)v} e^{-(\rho+\nu)\omega} \sigma^s \rho^r \mu^d \nu^u. \end{aligned}$$

Once more the likelihood factorises into separate functions and so the estimators are easily obtained by computing the partial derivatives of the log-likelihood with respect to μ and ν and setting them equal to 0:

$$\tilde{\mu} = \frac{D}{V}, \ \tilde{\nu} = \frac{U}{W}, \ \tilde{\sigma} = \frac{S}{V}, \ \tilde{\rho} = \frac{R}{W} \tag{43}$$

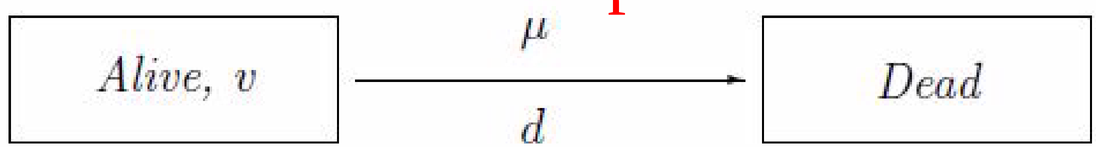
The estimates in all these examples have the form

$$\begin{array}{c}\text{MLE of transition}\\ \text{intensity } g \to h\end{array} = \frac{\text{No.transfers from } g \text{ to } h}{\text{Total waiting time in } g}. \tag{44}$$

1.9 The likelihood in multi-state models

It is important to write down the likelihood for a multi-state model. Thus, to emphasise this we collect the results of all our examples.

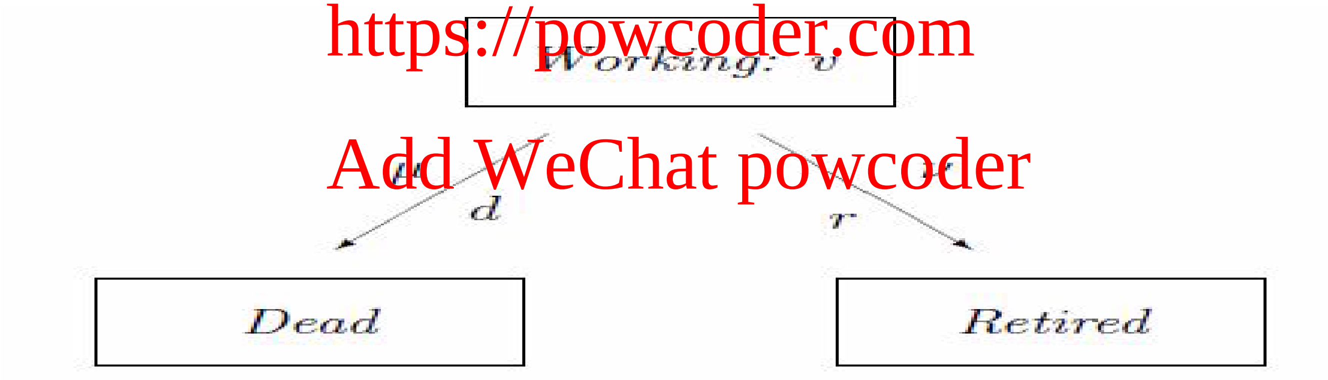
Example 1: Simple two state model; see 1.3 and 1.4.



where v is the total waiting time in the alive state and d is the total number of deaths. Then

$$L(\mu) = e^{-\mu v}\mu^{d} \tag{45}$$

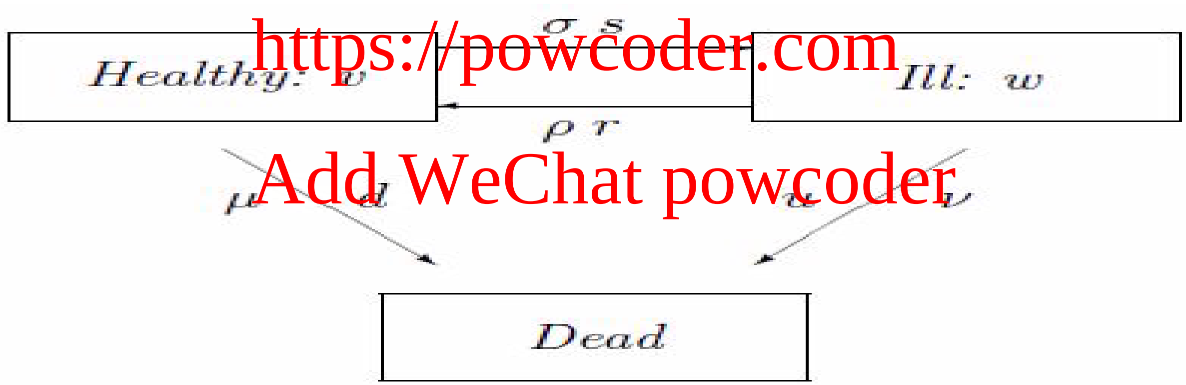
Example 2: Retiral; see 1.7.



where v is the total waiting time in the working state, d is the total number of deaths and r is the total number of retirals. Then

$$L(\mu, \nu) = e^{-(\mu+\nu)v}\mu^d\nu^r. \tag{46}$$

Example 3: PHI; see 1.8.



where v and ω are total waiting times and r, s, d and u are total numbers of transfers between states as shown. Then

$$L(\mu,\nu) = e^{-(\mu+\sigma)\omega} \times e^{-(\nu+\rho)\omega} \times \mu^d \times \nu^u \times \sigma^s \times \rho^r. \qquad (47)$$

1.10 The likelihood methodology

These examples allow us to formulate a general methodoloy as follows:

(a) Each **waiting time** ω_i in state i contributes a factor

$$e^{-\left(\sum_j \mu_{ij}\right)\omega_i}, \qquad (48)$$

where $\sum_j \mu_{ij}$ is the sum of the transition intensities **out of state** i.

(b) Each **transfer number** t_{ij} from state i to state j contributes a factor

$$\mu_{ij}^{t_{ij}}, \tag{49}$$

where μ_{ij} is the transition intensity from state i to state j

(c) The total likelihood is the product of all the above contributions.

A tip for the exam: You should be able to apply this rule and derive the likelihood for simple examples, as done in lectures/classes.

1.11 Properties of the MLE's

We saw in the simple two state model

$$\boxed{Alive} \xrightarrow{\mu} \boxed{Dead}$$

that

$$\tilde{\mu} = \frac{D}{V} = \frac{\text{No.deaths}}{\text{Total waiting time}} \tag{50}$$

and that asymptotically

$$\widetilde{\mu} \approx \mathcal{N}(\mu_0, \frac{\mu_0}{E(V)}). \tag{51}$$

The proof consisted of applying the Maximum Likelihood Theorem (MLT) together with some properties of the score function.

For the model for retiral we have the likelihood $L(\mu, \nu)$

$$L(\mu, \nu) = e^{-\mu v} \mu^d \times e^{-\nu v} \nu^r \tag{52}$$

where v is the total waiting time in the working state, d is the number of deaths and r is the number of retirals.

What is the **joint distribution** of $\widetilde{\mu}$ and $\widetilde{\nu}$? We need to **extend the MLT to the multiparameter case.**

The following parallels the one parameter case for the simple two state model in 1.5.

1.12 Multi-parameter version of the Maximum Likelihood Theorem (MTL)

Definition: The **score function**. Suppose $\mathbf{X}' = (X_1, X_2, ..., X_n)$ is a vector of random variables where the X_i are independent random variables with probability function $f_i(x, \boldsymbol{\theta})$ which depends on a parameter $\boldsymbol{\theta}' = (\theta_1, ..., \theta_r)$ with r components. Let

$$l = l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}, \mathbf{X}) = \sum_{1}^{n} \log f_i(X_i, \boldsymbol{\theta}) \tag{53}$$

be the log-likelihood function (viewed as a random variable). Also, let

$$U_i = U_i(\boldsymbol{\theta}) = \frac{\partial l}{\partial \theta_i},\ i = 1, ..., r. \tag{54}$$

Then $\mathbf{U}' = (U_1, ..., U_r)$ is called the **score function**.

Result: The score function satisfies:

(a) $E(\mathbf{U}) = 0$, and

(b) The variance-covariance matrix is $Var(\mathbf{U}) = -E\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\theta}}\right) = (v_{ij})_{r\times r}$, where $v_{ij} = -E(\frac{\partial^2 l}{\partial\theta_i \partial\theta_j})$ are its elements.

Assignment Project Exam Help

For example, when we have two parameters θ_1 and θ_2 , we get

https://powcoder.com

Add WeChat powcoder

$$Var(\mathbf{U}) = \begin{bmatrix} -E\left(\frac{\partial^2 l}{\partial\theta_1^2}\right) & -E\left(\frac{\partial^2 l}{\partial\theta_1\,\partial\theta_2}\right) \\ -E\left(\frac{\partial^2 l}{\partial\theta_2\,\partial\theta_1}\right) & -E\left(\frac{\partial^2 l}{\partial\theta_2^2}\right) \end{bmatrix}$$

Proof: The proof for **(a)** and **(b)** is pretty much the same as the one-parameter case. Firstly, we take $n = 1$ and write $f_i = f$.

(a) Consider the identity

$$\int f(x, \boldsymbol{\theta}) dx = 1. \tag{55}$$

Differentiate with respect to θ_i. We get

$$\int \frac{\partial f}{\partial \theta_i} dx = 0. \tag{56}$$

Rewrite this equation as

$$\int \left(\frac{\partial}{\partial \theta_i} \log f \right) f dx = 0 \tag{57}$$

The left hand side of this equation involves $E(\mathrm{U}_i)$, so $E(\mathbf{U}) = 0$ as required.

(b) The proof uses the same trick as in the one-parameter case. Differentiate (57) with respect to θ_j and use the product rule. We get

$$\implies \int \left\{ \frac{\partial}{\partial \theta_j} \left(\frac{\partial}{\partial \theta_i} \log f \right) f + \left(\frac{\partial}{\partial \theta_i} \log f \right) \frac{\partial f}{\partial \theta_j} \right\} dx = 0$$

$$\implies \int \frac{\partial}{\partial \theta_j} \left(\frac{\partial}{\partial \theta_i} \log f \right) f dx + \int \left(\frac{\partial}{\partial \theta_i} \log f \right) \left(\frac{\partial}{\partial \theta_j} \log f \right) f dx = 0$$

$$\implies \int \frac{\partial U_i}{\partial \theta_j} f dx + \int U_i U_j f dx = 0$$

$$\implies E(U_i U_j) = -E\left(\frac{\partial U_i}{\partial \theta_j} \right), \text{ since } U_i = \frac{\partial}{\partial \theta_i} \log f$$

$$\implies Cov(U_i, U_j) = E(U_i U_j) - E(U_i)E(U_j) = -E\left(\frac{\partial^2 l}{\partial \theta_i \theta_j} \right), \text{ since } E(U_i) = 0$$

Hence, **(b)** is proved.

- The case of general n follows by the additivity of the score function and the independence of the X_i

- The information function $I(\boldsymbol{\theta})$ extends as follows:

Definition: Fisher's **information function** $I(\boldsymbol{\theta})$ is

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l}{\partial \theta_i \theta_j}\right) = -E\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\theta}}\right) = Var(\mathbf{U}). \tag{58}$$

As before we usually call $I(\boldsymbol{\theta})$ the information function. We can now state (without proof) the multi-parameter version of the **MLT**.

Theorem: Suppose $\widehat{\boldsymbol{\theta}}$ satisfies $\mathbf{U} = \mathbf{0}$ and that $\boldsymbol{\theta}_0$ is the true (but unknown) value of $\boldsymbol{\theta}$. If $\widetilde{\boldsymbol{\theta}}$ is the random variable corresponding to $\widehat{\boldsymbol{\theta}}$, then $\widetilde{\boldsymbol{\theta}}$ is

(i) asymptotically unbiased, i.e. $E(\widetilde{\boldsymbol{\theta}}) \longrightarrow \boldsymbol{\theta}_0$,

(ii) asymptotically efficient, i.e. $Var(\widetilde{\boldsymbol{\theta}}) \longrightarrow I(\boldsymbol{\theta}_0)^{-1}$,

(iii) asymptotically normal, i.e we have

$$\widetilde{\theta} \approx \mathcal{N}(\boldsymbol{\theta}_0, I(\boldsymbol{\theta}_0)^{-1}) \tag{59}$$

1.13 Examples of MLE's

Example 1: Retiral/Death model. We found the likelihood for this model in 1.7 as

$$L(\mu, \nu) = e^{-\mu \upsilon} \mu^d \times e^{-\nu \upsilon} \nu^r \tag{60}$$

which gives

$$l(\mu, \nu) = -\mu \upsilon + d \log \mu - \nu \upsilon + r \log \nu \tag{61}$$

as the log-likelihood and

$$\widetilde{\mu} = \frac{D}{V}, \quad \widetilde{\nu} = \frac{R}{V} \tag{62}$$

as the estimators of μ and ν. Now the asymptotic distribution of $\widetilde{\mu}$ and $\widetilde{\nu}$ is determined by the information matrix $I(\mu, \nu)$.

Observe that in this case is given by

$$I(\mu, \nu) = \begin{bmatrix} -E\left(\frac{\partial^2 l}{\partial \mu^2}\right) & -E\left(\frac{\partial^2 l}{\partial \mu \partial \nu}\right) \\ -E\left(\frac{\partial^2 l}{\partial \nu \partial \mu}\right) & -E\left(\frac{\partial^2 l}{\partial v^2}\right) \end{bmatrix}$$

$$\mathbf{U}_{\mu} = \frac{\partial l}{\partial \mu} = -\upsilon + \frac{d}{\mu}, \text{ thus} \tag{63}$$

$$\mu E(\upsilon) = E(d), \text{ since } E(\mathbf{U}_{\mu}) = 0,$$

$$\text{from (63)} \quad \frac{\partial^2 l}{\partial \mu^2} = -\frac{d}{\mu^2}, \text{ thus}$$

$$-E\left(\frac{\partial^2 l}{\partial \mu^2}\right) = -E\left(-\frac{d}{\mu^2}\right) = \frac{E(d)}{\mu^2} = \frac{\mu E(\upsilon)}{\mu^2} = \frac{E(\upsilon)}{\mu}, \tag{64}$$

$$\text{from (63)} \quad \frac{\partial^2 l}{\partial \mu \partial \nu} = 0, \text{ thus}$$

$$-E\left(\frac{\partial^2 l}{\partial \mu \partial \nu}\right) = 0. \tag{65}$$

Similarly, since

$$\mathbf{U}_v = \frac{\partial l}{\partial v} = -v + \frac{r}{v}, \text{ we have} \tag{66}$$

$$-E\left(\frac{\partial^2 l}{\partial v^2}\right) = \frac{E(v)}{v} \text{ and} \tag{67}$$

$$-E\left(\frac{\partial^2 l}{\partial v \partial \mu}\right) = 0 \tag{68}$$

Hence, using Eqs(64, 65 and 67, 68) we find that

$$I(\mu, \nu) = \begin{bmatrix} \frac{E(V)}{\mu} & 0 \\ 0 & \frac{E(V)}{v} \end{bmatrix} \tag{69}$$

Thus, $I(\mu, \nu)$ is **diagonal** and so $\widetilde{\mu}$ and $\widetilde{\nu}$ are **asymptotically independent**.

N.B. The **estimators will always be asymptotically independent** whenever the **log-likelihood can be expressed as the sum of separate functions of the parameters** as in Eq.(61).

Therefore,

$$\widetilde{\mu} \approx \mathcal{N}\left(\mu_0, \frac{\mu_0}{E(V)}\right) \text{ and} \tag{70}$$

$$\widetilde{\nu} \approx \mathcal{N}\left(\nu_0, \frac{\nu_0}{E(V)}\right) \text{, thus} \tag{71}$$

$\widetilde{\mu}$ and $\widetilde{\nu}$ are asymptotically independent

Example 2: PHI model. We found the likelihood for this model in 1.8 as

$$\tilde{L}(\mu, \nu, \rho, \sigma) = e^{-(\sigma+\mu)v} e^{-(\rho+\nu)\omega} \sigma^s \rho^r \mu^d \nu^u \tag{72}$$

which gives

$$\tilde{\mu} = \frac{D}{V}, \quad \tilde{\nu} = \frac{U}{W}, \ \tilde{\sigma} = \frac{S}{V}, \ \tilde{\rho} = \frac{R}{W}. \tag{73}$$

as the estimators of μ, ν, σ and ρ .

Once more the likelihood factorises into separate functions and so the asymptotic distribution of the estimators is easily obtained. First, since $I(\mu, \nu, \rho, \sigma)$ is diagonal, $\tilde{\mu}, \tilde{\nu}, \tilde{\sigma}$ and $\tilde{\rho}$ are asymptotically independent. Moreover,

$$\tilde{\mu} \approx \mathcal{N}\left(\mu_0, \frac{\mu_0}{E(V)}\right)$$

with corresponding results for $\tilde{\nu}$, $\tilde{\sigma}$ and $\tilde{\rho}$.

Assignment Project Exam Help

Conclusion: The Markov model gives simple estimates of the transition intensities with simple properties. We need to keep track of only two things:

https://powcoder.com

Add WeChat powcoder

- the number of transfers between states, and

- the total time spent in each state.

This simplicity is part of the appeal of this class of models.