‘rmap’ Package Documentation (v.02)

Gail Gong David Johnston

August 31, 2011

Contents

1 Grouped Analysis **2**

1.1 The problem . . . 2

1.2 Notation dictionary . . . 2

1.3 Two-stage sampling . . . 3

1.3.1 Simple random sampling . . . 3

1.3.2 Back to two-stage sampling . . . 3

1.4 The data . . . 4

1.5 Goals . . . 5

1.6 The likelihood . . . 5

1.7 u_n and V . . . 6

1.7.1 $u_n(\gamma)$ and $V(\gamma)$ . . . 6

1.7.2 $u_n(\lambda)$ and $V(\lambda)$ . . . 8

1.7.3 u_n and V . . . 10

1.8 The delta method . . . 11

1.9 The HT estimate $\tilde{\xi}^T = (\tilde{\gamma}, \tilde{\pi})^T$ . . . 12

1.10 Hosmer-Lemeshow statistic . . . 14

1.11 AUC . . . 14

1.11.1 Break up AUC(ξ) . . . 14

1.11.2 Calculating $f_1(\xi) = \sum_{k=1}^{K} \gamma_k^2 (1 - \pi_k)\pi_k$ . . . 15

1.11.3 Calculating $f_2(\pi) = \sum_{k'=1}^{K-1} \sum_{k''=k'+1}^{K} \gamma_{k'}\gamma_{k''}(1 - \pi_{k'})\pi_{k''}$ . . . 15

1.11.4 Calculating $g(\xi) = (1 - \pi)\pi$ . . . 16

1.12 SD of a Risk Model . . . 17

2 Ungrouped Analysis **18**

2.1 Introduction . . . 18

2.2 Estimation . . . 19

2.3 Calibration . . . 19

2.4 Discrimination . . . 19

1 Grouped Analysis

1.1 The problem

Assume that, during the time interval $[0, t*)$, a person can be diagnosed with a specific disease or die from other causes. Assume also that we have in hand a model that can calculate a probability of the person being diagnosed with the disease before dying from other causes and before time $t*$. We want to validate this model.

1.2 Notation dictionary

The notation used here is different slightly from that used in *Two-stage Sampling Designs for Validating Personal Risk Models* by Whittemore and Halpern, which has been submitted to *Biostatistics* in 2010.

WH	**rmap**	**to jog your memory**
l	k	risKgroup
L	K	total number of risKgroups
τ	e	Event
	θ	(γ, π)
θ		(γ, λ)
i	n	subject iNdex
N	N	total number of subjects
t_{lm}	τ_{km}	ordered event times in the kth risK group
$X_{li}(t_{lm})$	$N_{kn}(\tau_{km}) = N_{kmn}$	indicates whether kn person is at risk at time τ_{km}
	$D_{ken}(\tau_{km}) = D_{kemn}$	indicates whether kn person had event e at time τ_{km}
$n_{lm} = \sum_i a_i X_{li}(t_{lm})$	$N_{km} = \sum_n a_n N_{kn}(\tau_{km})$	Number in risK group k at risk at time τ_{km}
$d_{l\tau m} = \sum_i a_i X_{li}(t_{lm}) N_{l\tau i}(t_{lm})$	$D_{kem} = \sum_n a_n N_{kmn} D_{ken}(\tau_{km})$	Number in risK group k who has event e at time τ_{km}

1.3 Two-stage sampling

We will allow for the possibility that the people in the study are sampled according to two-stage sampling, and so we provide this tiny interlude.

1.3.1 Simple random sampling

For comparison, we begin this discussion with simple random sampling. Let $\{x_n\}_{n=1,...N}$ be a random sample from a population governed by the density $f(x, \theta)$. Introduce the notation

$$\begin{aligned}
\text{loglike}_n &= \log(f(x_n, \theta)) && (1)\\
u_n &= \frac{\partial \text{loglike}_n}{\partial \theta} && (2)\\
I_n &= -\frac{\partial u_n}{\partial \theta} && (3)\\
U(\theta) &= \sum_{n=1}^{N} u_n && (4)\\
A &= \frac{1}{N}\sum_{n=1}^{N} I_n && (5)\\
V &= A^{-1} && (6)
\end{aligned}$$

The MLE $\hat{\theta}$ is the solution to $U(\theta) = 0$; the asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta)$ is Normal with zero mean and variance V, and $\hat{\theta}$ has covariance matrix $\frac{1}{N}V$.

1.3.2 Back to two-stage sampling

We use two-stage sampling with bernoulli second stage sampling. In the first stage, screen N subjects; $\mathcal{S} = \{x_n\}_{n=1,...N}$ are the subjects in the first stage. Let $\mathcal{S}_c$ be those screened patients falling in the cth category, $Q_c = \{n | x_n \in \mathcal{S}_c\}$ be their subscripts, and $N_c = |Q_c|$ denote the number of people in the first stage who land in category c. (Interpret the term "screen" to mean get enough information on the nth subject to know what category c she falls in.) In the second stage, test each person in $\mathcal{S}_c$ with probability p_c, and let $\bar{\mathcal{S}}_c$ denote those people tested, $\overline{Q}_c$ denote their subscripts, and $\overline{N}_c = |\overline{Q}_c|$ denote the number of people who fall in category c and are tested. (Interpret the term "test" to mean get all the information on the nth subject.) The sets $\{\bar{\mathcal{S}}_c\}_{c=1,...C}$ contain all the observations we can get are hands on, the ones that make it into the data set we are going to analyze.

Define u_n and I_n as in simple random sampling, and

$$\begin{aligned}
\hat{\omega}_c &= \frac{N_c}{N} && (7)\\
\hat{p}_c &= \frac{\bar{N}_c}{N_c} && (8)\\
a_n &= \sum_c \frac{1}{\hat{p}_c} 1(n \in \bar{Q}_c) && (9)\\
U(\theta) &= \sum_{n=1}^{N} a_n u_n && (10)\\
A &= \frac{1}{N}\sum_{n=1}^{N} a_n I_n && (11)\\
B_1 &= \frac{1}{N}\sum_{n=1}^{N} a_n u_n u_n^T && (12)\\
V &= A^{-1} \text{ or } B_1^{-1} && (13)\\
\hat{\mu}_c &= \frac{1}{\bar{N}_c}\sum_{n\in\bar{Q}_c} u_n && (14)\\
\hat{\Phi}_c &= \frac{1}{\bar{N}_c}\sum_{n\in\bar{Q}_c} u_n u_n^T && (15)\\
B_2 &= \sum_c \hat{\omega}_c \frac{1-\hat{p}_c}{\hat{p}_c}(\hat{\Phi}_c - \hat{\mu}_c \hat{\mu}_c^T) && (16)\\
\text{V2Stage} &= V + V B_2 V && (17)
\end{aligned}$$

The solution $\tilde{\theta}$ to $U(\theta) = 0$ we call the Horvitz-Thompson estimate. Notice that the Horvitz-Thompson estimate maximizes the PSEUDO likelihood equation $\sum_n a_n \text{loglike}_n$. We have $\sqrt{N}(\tilde{\theta} - \theta)$ is Normal with zero mean and variance V2Stage, and $\tilde{\theta}$ has covariance matrix $\frac{1}{N}$V2Stage.

1.4 The data

For each person x_n, we record

Variable	Description	Range
e_n	event type	0 = censored, 1= disease, 2 = death from other causes
t_n	time of event	$[0, t*)$
r_n	probability of disease as predicted by the model	$(0, 1)$
k_n	risKgroup as defined by r_n	$1, ..., K$
c_n	two stage Category	$1, ..., C$
z_n	covariates used to calculate r_n (optional)	

The number of riskgroups K is chosen in advance by the user, and typically the riskgroups are defined by which K-tile each person's predicted probability r_n falls in.

The rmap package contains functions `df_randomSample` and `df_twoStage`, which randomly generate a sample dataset. This dataset is a data.frame with columns e, t, r, k, and c. Each row represents one subject.

1.5 Goals

Let $\lambda_{ke}(t)$ be the hazard for event type e of people in riskgroup k. The probability of disease in the interval $[0, t*)$ is π_k

$$\begin{aligned}
\pi_k &= \int_0^{t^*} \lambda_{k1}(t) S_{k1}(t) S_{k2}(t) dt && (18)\\
S_{ke} &= e^{-\Lambda_{ke}(t)} && (19)\\
\Lambda_{ke}(t) &= \int_0^t \lambda_{ke}(s) ds && (20)
\end{aligned}$$

We have the following goals for which we must derive appropriate formulas:

1. Estimate π_k
2. Obtain the estimated covariance matrix $\Sigma = \widehat{\text{cov}}(\hat{\gamma}_1, ..., \hat{\gamma}_{K-1}, \hat{\pi}_1, ..., \hat{\pi}_K)$
3. Calculate the Hosmer-Lemeshow Chi-squared goodness of fit statistic
4. Calculate the AUC and its estimated variance
5. Calculate SD, the standard deviation of the model and its estimated variance

1.6 The likelihood

For the nth person, we observe the data $x_n = (\varepsilon_n, t_n, k_n)$. We take her contribution to the likelihood to be

$$\begin{aligned}
f(x_n) &= P(k_n) \times P(\varepsilon_n, t_n | k_n) && (21)\\
&= \prod_{k=1}^{K} \Big(P(k_n = k)^{k_n = k} \times P(\varepsilon_n, t_n | k_n = k) \Big) && (22)
\end{aligned}$$

The first term of equation (21) we take to be a multinomial probability $P(k_n = k) = \gamma_k$, where $\sum_{k=1}^K \gamma_k = 1$.

The second term $P(\varepsilon_n, t_n | k_n = k)$ will be conditional on the failure times of riskgroup k. The failure times are times in which a subject either gets disease or dies. Order these failure times and denote them like this:

$$0 < \tau_{k1} < ... < \tau_{km} < ... < \tau_{kM_k} \le t* \quad (23)$$

$m \in \{1, ..., M_k\}$ indexes these unique failure times for one risk group.

Define

$$\begin{aligned}
\lambda_{kem} &= \lambda_{ke}(\tau_{km}) && (24)\\
\lambda_k &= ((\lambda_{k11}, \cdots, \lambda_{k1m}, \cdots, \lambda_{k1M}), (\lambda_{k21}, \cdots, \lambda_{k2m}, \cdots, \lambda_{k2M})) && (25)\\
\lambda &= (\lambda_1, ...\lambda_k, ...\lambda_K) && (26)\\
\lambda_{k\bullet m} &= \lambda_{k1m} + \lambda_{k2m} && (27)\\
L &= 2 \sum_{k=1}^{K} M_k && (28)
\end{aligned}$$

We call λ the vector of discrete hazards and L is the number of elements in λ. This is how we think about the second term $P(\varepsilon_n, t_n|k_n = k)$. Suppose t_n falls inside $[\tau_{k,m(n)}, \tau_{k,m(n)+1})$ for some $m(n) = 1, \cdots, M_k$. We assume that the only times when she can have an event is at times $\tau_{k1}, \cdots, \tau_{k,m(n)}$. We say she is at risk during these times. At time τ_{km}, the probability that she will have event $e = 1$ is λ_{k1m}, have event $e = 2$ is λ_{k2m}, and the probability that she will have neither is $1 - \lambda_{k\bullet m}$ In other words, at each failure time for which this person is at risk, she has a multinomial probability for the three outcomes, e = 0, 1 or 2. Define

$$
\begin{aligned}
N_{kmn} &= N_{kn}(\tau_{km}) = 1(k_n == k \text{ and } t_n \geq \tau_{km}) && (29)\\
D_{kemn} &= D_{ken}(\tau_{km}) = 1(k_n == k \text{ and } e_n == e \text{ and } t_n \leq \tau_{km}) && (30)\\
D_{k\bullet mn} &= D_{k1mn} + D_{k2mn} && (31)
\end{aligned}
$$

N_{kmn} indicates whether or not the nth person is at risk at time τ_{km}, and D_{kemn} indicates whether or not the nth person had event e at time τ_{km}. Now we can write the second term

$$
P(\varepsilon_n, t_n|k_n = k) = \prod_{m=1}^{M_k} \lambda_{k1m}^{N_{kmn}D_{k1mn}} \lambda_{k2m}^{N_{kmn}D_{k2mn}} (1 - \lambda_{k\bullet m})^{N_{kmn}(1-D_{k\bullet mn})} \tag{32}
$$

Putting together the first and second terms and then taking the log, the nth person's contribution to the loglikelihood is

$$
\begin{aligned}
\text{loglike}_n(\gamma, \lambda) &= \sum_{k=1}^{K} 1(k_n = k)\log(\gamma_k)\\
&+ \sum_{k=1}^{K}\sum_{m=1}^{M_k} N_{kmn}\Bigg(D_{k1mn}\log(\lambda_{k1m}) + D_{k2mn}\log(\lambda_{k2m}) + (1 - D_{k\bullet mn})\log(1 - \lambda_{k\bullet m})\Bigg) && (33)
\end{aligned}
$$

The first term is the first term in the equation that precedes (5) of Whittemore and Halpern 2010, and the second term is (10) in Whittemore and Halpern 2010.

1.7 u_n and V

1.7.1 $u_n(\gamma)$ and $V(\gamma)$

We continue following the roadmap presented in equations (1) to (17). Here we get the partial derivatives of the a person's contribution to the loglikelihood with respect to γ.

$$
\begin{aligned}
u_n(\gamma_k) &= \frac{\partial \text{loglike}_n}{\partial \gamma_k} = \frac{1(k_n = k)}{\gamma_k} - \frac{1(k_n = K)}{1 - (\gamma_1 + \cdots + \gamma_{K-1})} && (34)\\
I_n(\gamma_k, \gamma_k) &= -\frac{\partial u_n(\gamma_k)}{\partial \gamma_k} = \frac{1(k_n = k)}{\gamma_k^2} + \frac{1(k_n = K)}{\Big(1 - (\gamma_1 + \cdots + \gamma_{K-1})\Big)^2} && (35)\\
I_n(\gamma_k, \gamma_{k'}) &= -\frac{\partial u_n(\gamma_k)}{\partial \gamma_{k'}} = \frac{1(k_n = K)}{\Big(1 - (\gamma_1 + \cdots + \gamma_{K-1})\Big)^2}, \text{ if } k \neq k' && (36)
\end{aligned}
$$

and we then sum over $\sum_{n=1}^{N} a_n$:

$$
\begin{aligned}
U(\gamma_k) &= \sum_{n=1}^{N} a_n u_n(\gamma_k) && (37)\\
&= \frac{\sum_{n=1}^{N} a_n 1(k_n = k)}{\gamma_k} - \frac{\sum_{n=1}^{N} a_n 1(k_n = K)}{1-(\gamma_1+\cdots+\gamma_{K-1})} && (38)\\
&= \frac{N_k}{\gamma_k} - \frac{N_K}{1-(\gamma_1+\cdots+\gamma_{K-1})} && (39)\\
N_k &= \sum_{n=1}^{N} a_n 1(k_n = k) && (40)\\
\sum_{n=1}^{N} a_n I_n(\gamma_k,\gamma_k) &= \frac{\sum_{n=1}^{N} a_n 1(k_n = k)}{\gamma_k^2} + \frac{\sum_{n=1}^{N} a_n 1(k_n = K)}{\Big(1-(\gamma_1+\cdots+\gamma_{K-1})\Big)^2} && (41)\\
&= \frac{N_k}{\gamma_k^2} + \frac{N_K}{\gamma_K^2} && (42)\\
\sum_{n=1}^{N} a_n I_n(\gamma_k,\gamma_{k'}) &= \frac{\sum_{n=1}^{N} a_n 1(k_n = K)}{\Big(1-(\gamma_1+\cdots+\gamma_{K-1})\Big)^2} && (43)\\
&= \frac{N_K}{\gamma_K^2} && (44)
\end{aligned}
$$

and solving $U(\tilde{\gamma}_k) = 0$ gives

$$
\begin{aligned}
\tilde{\gamma}_k &= N_k/N && (45)\\
N &= \sum_{k=1}^{K} N_k && (46)
\end{aligned}
$$

Substituting $\tilde{\gamma}_k$ into equations (42) and (44) gives:

$$
\begin{aligned}
NA(\gamma_k,\gamma_k) &= \frac{N}{\tilde{\gamma}_k} + \frac{N}{\tilde{\gamma}_K} && (47)\\
NA(\gamma_k,\gamma_{k'}) &= \frac{N}{\tilde{\gamma}_K} && (48)\\
\tilde{\gamma}_K &= 1-(\tilde{\gamma}_1+\cdots+\tilde{\gamma}_{K-1}) && (49)
\end{aligned}
$$

Using Mathematica we get

$$
\begin{aligned}
V(\gamma) &= A^{-1}(\gamma) && (50)\\
&= \frac{1}{N} \times \left(\begin{pmatrix} \gamma_1 & & 0 \\ & \ddots & \\ 0 & & \gamma_{K-1} \end{pmatrix} - \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{K-1} \end{pmatrix} \begin{pmatrix} \gamma_1 & \cdots & \gamma_{K-1} \end{pmatrix} \right) && (51)
\end{aligned}
$$

which matches (2) of Whittemore and Halpern 2010.

1.7.2 $u_n(\lambda)$ and $V(\lambda)$

Next, get the partial derivatives of a person's contribution to the loglikelihood with respect to λ.

$$
\begin{aligned}
u_n(\lambda_{kem}) &= \frac{\partial \text{loglike}_n}{\partial \lambda_{kem}} = N_{kmn}\left(\frac{D_{kemn}}{\lambda_{kem}} - \frac{1 - D_{k\bullet mn}}{1 - \lambda_{k\bullet m}}\right) && (52)\\
I_n(\lambda_{k1m}, \lambda_{k1m}) &= -\frac{\partial u_n(\lambda_{k1m})}{\partial \lambda_{k1m}} = N_{kmn}\left(\frac{D_{k1mn}}{\lambda_{k1m}^2} + \frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2}\right) && (53)\\
I_n(\lambda_{k1m}, \lambda_{k2m}) &= -\frac{\partial u_n(\lambda_{k1m})}{\partial \lambda_{k2m}} = N_{kmn}\left(\frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2}\right) && (54)\\
I_n(\lambda_{k2m}, \lambda_{k2m}) &= -\frac{\partial u_n(\lambda_{k2m})}{\partial \lambda_{k2m}} = N_{kmn}\left(\frac{D_{k2mn}}{\lambda_{k2m}^2} + \frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2}\right) && (55)
\end{aligned}
$$

The first equation in the above display checks with (16) of Whittemore and Halpern. Next, sum over $\sum_{n=1}^{N} a_n$:

$$
\begin{aligned}
U(\lambda_{kem}) &= \sum_{n=1}^{N} a_n N_{kmn} \left(\frac{D_{kemn}}{\lambda_{kem}} - \frac{1 - D_{k\bullet mn}}{1 - \lambda_{k\bullet m}} \right) && (56) \\
&= \frac{\sum_{n=1}^{N} a_n N_{kmn} D_{kemn}}{\lambda_{kem}} - \frac{\sum_{n=1}^{N} a_n N_{kmn} - \sum_{n=1}^{N} a_n N_{kmn} D_{k\bullet mn}}{1 - \lambda_{k\bullet m}} && (57) \\
&= \frac{D_{kem}}{\lambda_{kem}} - \frac{N_{km} - D_{k\bullet m}}{1 - \lambda_{k\bullet m}} && (58) \\
D_{kem} &= \sum_{n=1}^{N} a_n N_{kmn} D_{kemn} && (59) \\
D_{k\bullet m} &= \sum_{n=1}^{N} a_n N_{kmn} D_{k\bullet mn}) && (60) \\
N_{km} &= \sum_{n=1}^{N} a_n N_{kmn} && (61) \\
\sum_{n=1}^{N} a_n I_n(\lambda_{k1m}, \lambda_{k1m}) &= \sum_{n=1}^{N} a_n N_{kmn} \left(\frac{D_{k1mn}}{\lambda_{k1m}^2} + \frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2} \right) && (62) \\
\sum_{n=1}^{N} a_n I_n(\lambda_{k1m}, \lambda_{k2m}) &= \sum_{n=1}^{N} a_n N_{kmn} \left(\frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2} \right) && (63) \\
\sum_{n=1}^{N} a_n I_n(\lambda_{k2m}, \lambda_{k2m}) &= \sum_{n=1}^{N} a_n N_{kmn} \left(\frac{D_{k2mn}}{\lambda_{k2m}^2} + \frac{1 - D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2} \right) && (64) \\
\sum_{n=1}^{N} a_n I_n(\lambda_{kem}, \lambda_{kem}) &= \frac{\sum_{n=1}^{N} a_n N_{kmn} D_{kemn}}{\lambda_{k1m}^2} + \frac{\sum_{n=1}^{N} a_n N_{kmn} - \sum_{n=1}^{N} a_n N_{kmn} D_{k\bullet mn}}{(1 - \lambda_{k\bullet m})^2} && (65) \\
&= \frac{D_{kem}}{\lambda_{kem}^2} + \frac{N_{km} - D_{k\bullet m}}{(1 - \lambda_{k\bullet m})^2} && (66) \\
&= N_{km} \left(\frac{D_{kem}/N_{km}}{\lambda_{kem}^2} + \frac{(N_{km} - D_{k\bullet m})/N_{km}}{(1 - \lambda_{k\bullet m})^2} \right) && (67)
\end{aligned}
$$

and solving $0 = U(\lambda)$ gives

$$
\tilde{\lambda}_{kem} = \frac{D_{kem}}{N_{km}} \qquad (68)
$$

Since λ_{k1m} and λ_{k2m} both appear in the equations for $0 = U(\lambda_{k1m})$ and $0 = U(\lambda_{k2m})$, we need to consider this system of two equations and two unknowns. Simple substitution of $\tilde{\lambda}_{k1m}$ and $\tilde{\lambda}_{k2m}$ into these equations show that they are the required solutions. Substituting $\tilde{\lambda}_{kem}$ into appropriate sum over $\sum_{n=1}^{N} a_n$ equations,

$$
\begin{aligned}
\sum_{n=1}^{N} a_n I_n(\tilde{\lambda}_{kem}, \tilde{\lambda}_{kem}) &= N_{km}\left(\frac{\tilde{\lambda}_{kem}}{\tilde{\lambda}_{kem}^2} + \frac{1-\tilde{\lambda}_{k\bullet m}}{(1-\tilde{\lambda}_{k\bullet m})^2}\right) && (69)\\
&= \frac{N_{km}}{\tilde{\lambda}_{kem}} + \frac{N_{km}}{1-\tilde{\lambda}_{k\bullet m}} && (70)\\
\sum_{n=1}^{N} a_n I_n(\tilde{\lambda}_{k1m}, \tilde{\lambda}_{k2m}) &= \frac{N_{km}}{1-\tilde{\lambda}_{k\bullet m}} && (71)
\end{aligned}
$$

We can build a two-by-two matrix using equations (70) and (71). Setting $e = 1$ or $e = 2$ in equation (70) fills the diagonal elements of the matrix, and equation (71) fills the off-diagonal elements.

$$
(NA)_{km} = \frac{N_{km}}{\tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m}(1-\tilde{\lambda}_{k1m}-\tilde{\lambda}_{k2m})}\begin{pmatrix} \tilde{\lambda}_{k2m}(1-\tilde{\lambda}_{k2m}) & \tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} \\ \tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} & \tilde{\lambda}_{k1m}(1-\tilde{\lambda}_{k1m}) \end{pmatrix} \tag{72}
$$

We can put the matrix into Mathematica and get

$$
\begin{aligned}
(NA)_{km}^{-1} &= \frac{1}{N_{km}}\begin{pmatrix} \tilde{\lambda}_{k1m}(1-\tilde{\lambda}_{k1m}) & -\tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} \\ -\tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} & \tilde{\lambda}_{k2m}(1-\tilde{\lambda}_{k2m}) \end{pmatrix} && (73)\\
V_{km} &= A_{km}^{-1} = \frac{N}{N_{km}}\begin{pmatrix} \tilde{\lambda}_{k1m}(1-\tilde{\lambda}_{k1m}) & -\tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} \\ -\tilde{\lambda}_{k1m}\tilde{\lambda}_{k2m} & \tilde{\lambda}_{k2m}(1-\tilde{\lambda}_{k2m}) \end{pmatrix} && (74)
\end{aligned}
$$

This checks with (13) of Whittemore and Halpern 2010. The implied order for V_{km} defined in equation (74) is different than the order we see in the `rmap` package. Equation (75) better accommodates the order of the data structure in the `rmap` package.

$$
V_{k,e_1,e_2,m} = \begin{cases} \lambda_{ke_1m}(1-\lambda_{ke_1m}) & \text{if } e_1 = e_2 \\ \lambda_{k1m}\lambda_{k2m} & \text{if } e_1 \neq e_2 \end{cases} \tag{75}
$$

1.7.3 u_n and V

Write

$$
\begin{aligned}
\theta &= \begin{pmatrix} \gamma \\ \lambda \end{pmatrix} && (76)\\
u_n &= \begin{pmatrix} u_n(\gamma) \\ u_n(\lambda) \end{pmatrix} && (77)
\end{aligned}
$$

where $u_n(\gamma)$ is the $K-1$ dimensional vector of derivatives with respect to $\gamma_1, \cdots \gamma_{K-1}$ and $u_n(\lambda)$ is the L dimensional vector of derivatives with respect to all the components of λ. Remember that $\lambda = (\lambda_1, ..., \lambda_k, ..., \lambda_K)$, where $\lambda_k = ((\lambda_{k11}, \cdots, \lambda_{k1m}, \cdots, \lambda_{k1M}), (\lambda_{k21}, \cdots, \lambda_{k2m}, \cdots, \lambda_{k2M}))$. Also remember that $L = 2\sum_{k=1}^{K} M_k$. Also write

$$A = \begin{pmatrix} A(\gamma) & 0 \\ 0 & A(\lambda) \end{pmatrix} \tag{78}$$

$$V = A^{-1} = \begin{pmatrix} V(\gamma) & 0 \\ 0 & V(\lambda) \end{pmatrix} \tag{79}$$

Two-stage sample theory says $\tilde{\theta}$ has covariance matrix V2Stage.

$$\text{V2Stage} = V + V\ B_2\ V \tag{80}$$

$$\hat{\mu}_c = \frac{1}{\bar{N}_c} \sum_{n \in \bar{Q}_c} u_n \tag{81}$$

$$\hat{\Phi}_c = \frac{1}{\bar{N}_c} \sum_{n \in \bar{Q}_c} u_n u_n^T \tag{82}$$

$$B_2 = \sum_c \hat{\omega}_c \frac{1 - \hat{p}_c}{\hat{p}_c} \frac{\bar{N}_c}{\bar{N}_c - 1} (\hat{\Phi}_c - \hat{\mu}_c \hat{\mu}_c^T) \tag{83}$$

In the `rmap` package, `B2Fn` divides the calculation of equation (83) into two parts: `PhiHatPart` and `muHatPart`. To follow the logic of the `rmap` it is useful to write B_2 as follows:

$$B_2 = \sum_c \hat{\omega}_c \frac{1 - \hat{p}_c}{\hat{p}_c} \frac{\bar{N}_c}{\bar{N}_c - 1} \hat{\Phi}_c - \sum_c \hat{\omega}_c \frac{1 - \hat{p}_c}{\hat{p}_c} \frac{\bar{N}_c}{\bar{N}_c - 1} \hat{\mu}_c \hat{\mu}_c^T \tag{84}$$

The first term of equation (84) is calculated as `PhiHatPart`, and the second term is calculated as `muHatPart`.

1.8 The delta method

Suppose X is an I-dimensional random vector with distribution

$$X \sim \text{Normal}(\mu, \Sigma) \tag{85}$$

(Therefore μ is also an I dimensional vector and Σ is an $I \times I$ dimensional matrix.) Define the J dimensional random vector $Y = f(X)$. To make things very explicit write

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_J \end{pmatrix} = \begin{pmatrix} f_1(X_1, \cdots, X_I) \\ \vdots \\ f_J(X_1, \cdots, X_I) \end{pmatrix} \tag{86}$$

Then

$$Y \sim \text{Normal}(f(\mu), \Delta^T \Sigma \Delta) \tag{87}$$

where

$$\Delta^T_{ji} = \frac{\partial f_j}{\partial \mu_i} \tag{88}$$

Again to make things really explicit we can write out the covariance of Y like this:

$$\text{cov}(Y) = \begin{pmatrix} \frac{\partial f_1}{\partial \mu_1} & \cdots & \frac{\partial f_1}{\partial \mu_I} \\ \vdots & & \vdots \\ \frac{\partial f_J}{\partial \mu_1} & \cdots & \frac{\partial f_J}{\partial \mu_I} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1I} \\ \vdots & & \vdots \\ \sigma_{II} & \cdots & \sigma_{II} \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial \mu_1} & \cdots & \frac{\partial f_J}{\partial \mu_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial \mu_I} & \cdots & \frac{\partial f_J}{\partial \mu_I} \end{pmatrix} \tag{89}$$

1.9 The HT estimate $\tilde{\xi}^T = (\tilde{\gamma}, \tilde{\pi})^T$

Apply the delta method to $X = \tilde{\theta}$, $Y = \tilde{\xi} = \begin{pmatrix} \tilde{\gamma} \\ \tilde{\pi} \end{pmatrix}$, $\mu = \theta$, and $\xi = \begin{pmatrix} \gamma \\ \pi \end{pmatrix} = f(\mu) = f\begin{pmatrix} \gamma \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma \\ g(\lambda) \end{pmatrix}$ where

$$\tilde{\pi}_k = g_k(\tilde{\lambda}_k) = \sum_{m=1}^{M_k} \tilde{\lambda}_{k1m} \prod_{m'=1}^{m-1} (1 - \tilde{\lambda}_{k\bullet m'}) \tag{90}$$

From the fact that $\tilde{\theta} \sim \text{Normal}(\theta, \frac{\text{V2Stage}}{N})$ and from the delta method, we get

$$\tilde{\xi} = \begin{pmatrix} \tilde{\gamma} \\ \tilde{\pi} \end{pmatrix} \sim \text{Normal}(\begin{pmatrix} \tilde{\gamma} \\ \tilde{\pi} \end{pmatrix}, \frac{\Sigma}{N}) \tag{91}$$

where

$$D = \begin{pmatrix} I_{K-1} & 0 \\ 0 & D(\lambda) \end{pmatrix} \tag{92}$$

$$D(\lambda) = \begin{pmatrix} \frac{\partial g_1}{\partial \lambda_1} & \cdots & \frac{\partial g_J}{\partial \lambda_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial \lambda_L} & \cdots & \frac{\partial g_J}{\partial \lambda_L} \end{pmatrix} \tag{93}$$

$$\Sigma = D^T \text{ V2Stage } D \tag{94}$$

The derivatives can be gotten in closed form. From equations (95) to (102) we drop the subscript k and the $\tilde{}$ from the notation so Equation (90) becomes

$$\pi = g(\lambda) = \sum_{m=1}^{M} \lambda_{1m} \prod_{m'=1}^{m-1} (1 - \lambda_{\bullet m'}) \tag{95}$$

$$\lambda_{\bullet m} = \lambda_{1m} + \lambda_{2m} \tag{96}$$

We are going to write out the gory details for $M = 5$. Here is a list of all the elemnts inside λ. Remember that we are dropping the subscipt k.

$$\lambda = \left(\begin{array}{cccccc} & m=1 & m=2 & m=3 & m=4 & m=5 \\ e=1 & \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} & \lambda_{15} \\ e=2 & \lambda_{21} & \lambda_{22} & \lambda_{23} & \lambda_{24} & \lambda_{25} \end{array} \right) \tag{97}$$

Now we can write out π

$$\begin{aligned} \pi &= \lambda_{11} \\ &+ \lambda_{12}(1-\lambda_{\bullet 1}) \\ &+ \lambda_{13}(1-\lambda_{\bullet ,1})(1-\lambda_{\bullet 2}) \\ &+ \lambda_{14}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2})(1-\lambda_{\bullet 3}) \\ &+ \lambda_{15}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2})(1-\lambda_{\bullet 3})(1-\lambda_{\bullet 4}) \end{aligned} \tag{98}$$

Now think about taking the partial derivatives $\frac{\partial \pi}{\partial \lambda_{13}}$ and $\frac{\partial \pi}{\partial \lambda_{23}}$. Notice that in the equation for π, λ_{13} shows up only in the terms that begins λ_{13}, λ_{14}, λ_{15}, and exactly one time in each term. Also, π, λ_{23} shows up only in the terms that begin with λ_{14}, λ_{15}, and again exactly one time in each term. And one more thing, $\frac{\partial(1-\lambda_{\bullet 3})}{\partial \lambda_{13}} = \frac{\partial(1-\lambda_{13}-\lambda_{23})}{\partial \lambda_{13}} = -1$ and $\frac{\partial(1-\lambda_{\bullet 3})}{\partial \lambda_{23}} = -1$. Now it is time to take derivatives.

$$\begin{aligned} \frac{\partial \pi}{\partial \lambda_{13}} &= (1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2}) \\ &- \lambda_{14}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2}) \\ &- \lambda_{15}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2})(1-\lambda_{\bullet 4}) && (99) \\ \frac{\partial \pi}{\partial \lambda_{23}} &= -\lambda_{14}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2}) \\ &- \lambda_{15}(1-\lambda_{\bullet 1})(1-\lambda_{\bullet 2})(1-\lambda_{\bullet 4}) && (100) \end{aligned}$$

And we see that in general,

$$\frac{\partial \pi}{\partial \lambda_{2m}} = -\sum_{m''=m+1}^{M} \lambda_{1m''} \prod_{m'=1, m'\neq m}^{m''-1} (1-\lambda_{\bullet m'}) \tag{101}$$

$$\frac{\partial \pi}{\partial \lambda_{1m}} = \prod_{m'=1}^{m-1} (1-\lambda_{\bullet m'}) + \frac{\partial \pi}{\partial \lambda_{2m}} \tag{102}$$

Now we can write

$$D = \Delta(D_1, \cdots, D_K) \tag{103}$$

$$D_k = \begin{pmatrix} D_{k1} \\ D_{k2} \end{pmatrix} \tag{104}$$

$$D_{k1m} = \prod_{m'=1}^{m-1} (1-\lambda_{\bullet m'}) + D_{k2m} \tag{105}$$

$$D_{k2m} = -\sum_{m'=m+1}^{M} \lambda_{1m'} \prod_{m''=1, m''\neq m}^{m'-1} (1-\lambda_{\bullet m''}) \tag{106}$$

1.10 Hosmer-Lemeshow statistic

To evaluate the validity of the model that predicts risks r_n, we use the Hosmer-Lemeshow chi-square statistic

$$\chi^2_K = \sum_{k=1}^{K} \frac{(\hat{\pi}_k - \tilde{r}_k)^2}{\hat{\sigma}_k^2} \tag{107}$$

where $\tilde{r}_k$ is a central measure of r_n for subjects in risk group k.

1.11 AUC

The AUC is defined

$$\begin{aligned} \mathrm{AUC}(\xi) &= \frac{\sum_{k=1}^{K} \gamma_k^2 (1-\pi_k)\pi_k + \sum_{k=1}^{K} \sum_{k'>k} \gamma_k \gamma_{k'} (1-\pi_k)\pi_{k'}}{2(1-\pi)\pi} && (108) \\ \pi &= \sum_{k=1}^{K} \gamma_k \pi_k && (109) \end{aligned}$$

We will want to calculate a confidence interval for the AUC. Since the values of the AUC fall inside the unit interval, we will define $B = \mathrm{logit}(\mathrm{A}) = \log(\mathrm{AUC}/(1-\mathrm{AUC})) = \log(\mathrm{AUC}) - \log(1-\mathrm{AUC})$ and we will approximate the distribution of $\tilde{B}$ to be Normal with mean logit(AUC)) and variance

$$\begin{aligned} \mathrm{Var}\big(\tilde{B}\big) &= \frac{D_B D_{\mathrm{AUC}}^T \, \Sigma \, D_{\mathrm{AUC}} D_B}{N} && (110) \\ D_B &= \frac{\partial B}{\partial \mathrm{AUC}} = \frac{\partial}{\partial \mathrm{AUC}} \big(\log(\mathrm{AUC}) - \log(1-\mathrm{AUC})\big) && (111) \\ &= \frac{1}{\mathrm{AUC}} - \frac{1}{1-\mathrm{AUC}} = \frac{1}{\mathrm{AUC}(1-\mathrm{AUC})} && (112) \\ D_{\mathrm{AUC}}^T &= \begin{pmatrix} \frac{\partial}{\partial \gamma_1} & \cdots & \frac{\partial}{\partial \gamma_{K-1}} & \frac{\partial}{\partial \pi_1} & \cdots & \frac{\partial}{\partial \pi_K} \end{pmatrix} \mathrm{AUC}(\xi) && (113) \end{aligned}$$

Then we form a 95 percent confidence interval for B: $\big[\tilde{B} - 1.96\sigma, \tilde{B} + 1.96\sigma\big]$ where $\sigma = \sqrt{\mathrm{Var}(\tilde{B})}$. and then a 95 percent confidence interval for AUC is $\big[\mathrm{logistic}(\tilde{B} - 1.96\sigma), \mathrm{logistic}(\tilde{B} + 1.96\sigma)\big]$.

1.11.1 Break up AUC(ξ)

(We are going to use the delta method again. This time, we are going to transform the random variable $\hat{\xi}$ to AUC. The transformation will be written in terms of f and g, which are different from the f and g from section (1.9).)

Rewrite

$$\begin{aligned}
\mathrm{AUC}(\xi) &= \frac{\frac{1}{2}f_1(\xi) + f_2(\xi)}{g(\xi)} && (114)\\
f_1(\xi) &= \sum_{k=1}^{K} \gamma_k^2(1-\pi_k)\pi_k && (115)\\
f_2(\xi) &= \sum_{k'=1}^{K-1}\sum_{k''=k'+1}^{K} \gamma_{k'}\gamma_{k''}(1-\pi_{k'})\pi_{k''} && (116)\\
g(\pi) &= (1-\pi)\pi && (117)
\end{aligned}$$

1.11.2 Calculating $f_1(\xi) = \sum_{k=1}^{K} \gamma_k^2(1-\pi_k)\pi_k$

Calculate the partial derivative with respect to γ_k. Ignoring the constraint on the γ_ks,

$$\frac{\partial f_1}{\partial \gamma_k} = 2\gamma_k(1-\pi_k)\pi_k \tag{118}$$

and then imposing the constraint,

$$\frac{\partial f_1}{\partial \gamma_k} = 2\Big(\gamma_k(1-\pi_k)\pi_k - \gamma_K(1-\pi_K)\pi_K\Big) \tag{119}$$

Also,

$$\frac{\partial f_1}{\partial \pi_k} = \gamma_k^2\big(1-2\pi_k\big) \tag{120}$$

And putting them all together,

$$\begin{aligned}
\frac{\partial f_1}{\partial \gamma_k} &= 2\Big(\gamma_k(1-\pi_k)\pi_k - \gamma_K(1-\pi_K)\pi_K\Big) && (121)\\
\frac{\partial f_1}{\partial \pi_k} &= \gamma_k^2\big(1-2\pi_k\big) && (122)
\end{aligned}$$

1.11.3 Calculating $f_2(\pi) = \sum_{k'=1}^{K-1}\sum_{k''=k'+1}^{K} \gamma_{k'}\gamma_{k''}(1-\pi_{k'})\pi_{k''}$

To see how to proceed, it helps to imagine differentiating with respect to γ_3 or π_3. Here are the possible values of (k', k'')

```
12 13 14 15
   23 24 25
      34 35
         45
```

All the places where 3 shows up are $k' = 1, 2$ and so $k'' = 3$ and $k' = 3$ and $k'' = 4, 5$. Ignoring the constraints on the γ_ks,

$$\begin{aligned}
\frac{\partial f_2}{\partial \gamma_k} &= \sum_{k'=1}^{k-1} \gamma_{k'}(1 - \pi_{k'})\pi_k + \sum_{k''=k+1}^{K} \gamma_{k''}(1 - \pi_k)\pi_{k''} && (123) \\
\frac{\partial f_2}{\partial \gamma_K} &= \sum_{k'=1}^{K-1} \gamma_{k'}(1 - \pi_{k'})\pi_K && (124)
\end{aligned}$$

and then imposing the constraint,

$$\frac{\partial f_2}{\partial \gamma_k} = \sum_{k'=1}^{k-1} \gamma_{k'}(1 - \pi_{k'})\pi_k + \sum_{k''=k+1}^{K} \gamma_{k''}(1 - \pi_k)\pi_{k''} - \sum_{k'=1}^{K-1} \gamma_{k'}(1 - \pi_{k'})\pi_K \tag{125}$$

for $k = 1, \cdots, K - 1$. Using the same reasoning as for the unsconstrained calculation for the γ_k, we get a similar expression for the derivative with respect to π_k, and putting them together,

$$\begin{aligned}
\frac{\partial f_2}{\partial \gamma_k} &= \sum_{k'=1}^{k-1} \gamma_{k'}(1 - \pi_{k'})\pi_k + \sum_{k''=k+1}^{K} \gamma_{k''}(1 - \pi_k)\pi_{k''} - \sum_{k'=1}^{K-1} \gamma_{k'}(1 - \pi_{k'})\pi_K && (126) \\
\frac{\partial f_2}{\partial \pi_k} &= \sum_{k'=1}^{k-1} \gamma_{k'}\gamma_k(1 - \pi_{k'}) - \sum_{k''=k+1}^{K} \gamma_k\gamma_{k''}\pi_{k''} && (127)
\end{aligned}$$

1.11.4 Calculating $g(\xi) = (1 - \pi)\pi$

Before differentiating g, first calculate without regard to the constraint

$$\frac{\partial \pi}{\partial \gamma_k} = \frac{\partial}{\partial \gamma_k} \sum_{k=1}^{K} \gamma_k \pi_k = \pi_k \tag{128}$$

and then imposing the constraint,

$$\frac{\partial \pi}{\partial \gamma_k} = \pi_k - \pi_K \tag{129}$$

Also,

$$\frac{\partial \pi}{\partial \pi_k} = \gamma_k \tag{130}$$

Next, write

$$g(\xi) = (1-\pi)\pi = \pi - \pi^2 \tag{131}$$

$$\frac{\partial g(\xi)}{\partial \gamma_k} = (1-2\pi)\frac{\partial \pi}{\partial \gamma_k} = (1-2\pi)(\pi_k - \pi_K) \tag{132}$$

$$\frac{\partial g(\xi)}{\partial \pi_k} = (1-2\pi)\frac{\partial \pi}{\partial \pi_k} = (1-2\pi)\gamma_k \tag{133}$$

Finally, use the quotient rule to calculate

$$\frac{\partial}{\partial \xi_i}(\mathrm{AUC}) = \frac{\partial}{\partial \xi_i}\left(\frac{\frac{f_1}{2}+f_2}{g}\right) = \frac{\frac{\partial(\frac{f_1}{2}+f_2)}{\partial \xi_i}g - (\frac{f_1}{2}+f_2)\frac{\partial g}{\partial \xi_i}}{g^2} \tag{134}$$

1.12 SD of a Risk Model

The standard deviation of outcome probabilities across the risk groups is defined as

$$\mathrm{SD} = \sqrt{\sum_{k=1}^{K} \gamma_k (\pi_k - \pi)^2} \tag{135}$$

where $\pi = \sum_{k=1}^{K} \gamma_k \pi_k$ is defined by equation (109). This formula for SD is equation (4) of Whittemore and Halpern 2010.

To get an estimate for SD, substitute estimates for true values

$$\tilde{\mathrm{SD}} = \sqrt{\sum_{k=1}^{K} \tilde{\gamma}_k (\tilde{\pi}_k - \tilde{\pi})^2} \tag{136}$$

where $\tilde{\pi} = \sum_{k=1}^{K} \tilde{\gamma}_k \tilde{\pi}_k$

To obtain confidence intervals, we again use the delta method. From Equation (91), $\tilde{\xi} = \begin{pmatrix} \tilde{\gamma} \\ \tilde{\pi} \end{pmatrix} \sim \mathrm{Normal}(\begin{pmatrix} \tilde{\gamma} \\ \tilde{\pi} \end{pmatrix}, \frac{\Sigma}{N})$
In this section, define the function $f(\xi)$ to be

$$f(\xi) = \mathrm{VAR} = \sum_{k=1}^{K} \gamma_k (\pi_k - \pi)^2 \tag{137}$$

By the delta method,

$$\mathrm{Var(VAR)} = D^T \frac{\Sigma}{N} D \tag{138}$$

$$D^T = \left(\frac{\partial}{\partial \gamma_1}, \cdots, \frac{\partial}{\partial \gamma_{K-1}}, \frac{\partial}{\partial \pi_1}, \cdots, \frac{\partial}{\partial \pi_K}\right) f(\xi) \tag{139}$$

First take derivatives of π.

$$\frac{\partial \pi}{\partial \gamma_k} = \pi_k - \pi_K, \ k = 1, \cdots K-1 \tag{140}$$

$$\frac{\partial \pi}{\partial \pi_k} = \gamma_k \tag{141}$$

Next, take derivatives of $f(\xi)$ with respect to γ_k

$$\frac{\partial f(\xi)}{\partial \gamma_k} = \sum_{k'=1}^{K} \frac{\partial}{\partial \gamma_k}\left(\gamma_{k'}\big(\pi_{k'}-\pi\big)^2\right) \tag{142}$$

$$= \sum_{k'=1}^{K}\left(\frac{\partial \gamma_{k'}}{\partial \gamma_k}\big(\pi_{k'}-\pi\big)^2 + \gamma_{k'}\frac{\partial}{\partial \gamma_k}\big(\pi_{k'}-\pi\big)^2\right) \tag{143}$$

$$= \big(\pi_k-\pi\big)^2 - \big(\pi_K-\pi\big)^2 - 2\sum_{k'=1}^{K}\left(\gamma_{k'}\big(\pi_{k'}-\pi\big)\big(\gamma_k-\gamma_K\big)\right. \tag{144}$$

$$= \big(\pi_k-\pi\big)^2 - \big(\pi_K-\pi\big)^2 \tag{145}$$

And then take derivatives of $f(\xi)$ with respect to π_k

$$\frac{\partial f(\xi)}{\partial \pi_k} = \sum_{k'=1}^{K} \frac{\partial}{\partial \pi_k}\left(\gamma_{k'}\big(\pi_{k'}-\pi\big)^2\right) \tag{146}$$

$$= \sum_{k'=1}^{K} \gamma_{k'}\frac{\partial}{\partial \pi_k}\left(\big(\pi_{k'}-\pi\big)^2\right) \tag{147}$$

$$= \sum_{k'=1}^{K} \gamma_{k'}2\big(\pi_{k'}-\pi\big)\frac{\partial}{\partial \pi_k}\big(\pi_{k'}-\pi\big) \tag{148}$$

$$= \sum_{k'=1}^{K} \gamma_{k'}2\big(\pi_{k'}-\pi\big)\left(\delta_{k,k'}-\gamma_k\right) \tag{149}$$

$$= 2\gamma_k\big(\pi_k-\pi\big) \tag{150}$$

Finally, write $\text{SD} = g(\text{VAR}) = \sqrt{\text{VAR}}$. Using the delta method again, $\text{Var}(\text{SD}) = g'(\text{VAR})^2\text{Var}(\text{VAR})$. where $g'(\text{VAR}) = \frac{1}{2\text{SD}}$. We get

$$\text{Var}(\text{SD}) = \frac{1}{4\text{VAR}} D^T \frac{\Sigma}{N} D. \tag{151}$$

2 Ungrouped Analysis

2.1 Introduction

Suppose a risk model is used to assign risks to N subjects at entry to a cohort study. We follow the subjects until time t^* and record for each subject a followup time T, an event status E, and an assigned risk R where

$$T = \min(t^*, U, C) \tag{152}$$

$$U = \text{time to disease or death} \tag{153}$$

$$C = \text{time to censoring} \tag{154}$$

$$E = \begin{cases} 0 & \text{if censored} \\ 1 & \text{if disease} \\ 2 & \text{if death from other causes} \end{cases} \tag{155}$$

The joint probability that an individual is assigned risk $R = r$ and experiences event $E = j$, $j = 1, 2$ in the period $(0, t^*)$ is

$$P\big(U \le t \text{ and } E = j \text{ and } R = r\big) = f_R(r)F_{jr}(t) \tag{156}$$

$$f_R(r) = \text{the probability density function of } R \tag{157}$$

$$F_{jr}(t) = P\big(U \le t \text{ and } E = j\big|R = r\big) \tag{158}$$

The quantity $F_{jr}(t)$ is the event-specific cumulative incidence function among those assigned risk r. Our goals are to estimate the probabilites $g(r) = f_R(r)$ and $\pi(r) = F_{1r}(t^*)$ and use functions of the estimates to assess model calibration (how well assigned risks agree with subsequent outcomes) and discrimination (how well the risks distinguish those who do and do not develop the outcome in the risk period).

Our previous work corresponds to the special case in which individual risks have been grouped into K bins or risk groups and summarized by means or medians $0 \le r_1 < \cdots < r_K \le 1$ with $\gamma_k = g(r_k)$ and nonparametric estimates obtained for the group-specific outcome probabilities π_k, $k = 1, \cdots, K$. Here we generalize this approach by using the nearest neighbor estimates (NNEs) proposed by Akritos (1994) and Saha and Heagerty (2010).

2.2 Estimation

Let $\mathcal{R}$ denote the set of distinct assigned risks. For each $\rho \in \mathcal{R}$, estimate $g(\rho)$ by the empirical pdf

$$\hat{g}(\rho) = \frac{\big|\{n : r_n = \rho\}\big|}{N} \tag{159}$$

and estimate $\hat{\pi}(\rho)$ by (1) Obtain a ε kernel nearest neighborhood of ρ

$$\mathcal{NN}(\rho) = \Big\{n : \big|\hat{G}(r_n) - \hat{G}(\rho)\big| < \varepsilon\Big\}. \tag{160}$$

(2) Considering all observations in $\mathcal{NN}(\rho)$ to be one bin or risk group, use our previous methodology to estimate the group-specific outcome probability $\pi_{\mathcal{NN}(\rho)}$.

For two-stage sampling, replace Equation (159) with

$$\hat{g}(\rho) = \frac{\sum_n a_n 1(r_n = \rho)}{\sum_n a_n} \tag{161}$$

2.3 Calibration

We compute a calibration curve or an individualized attribute diagram, defined to be a scatterplot of points $\{\rho, \hat{\pi}(\rho) : \rho \in \mathcal{R}\}$ with line segments connecting adjacent points. 95 percent (nonsimultaneous) confidence bands for this curve are obtained by calculating bootstrap estimates of the standard deviation of $\hat{\pi}(\rho)$.