1. Анализ данных типа времени жизни

1.1. Кривая дожития

В анализе данных типа времени жизни исследуются объекты, для которых определено точечное событие, которое условно называют "отказ". Об этом событии известно, что оно происходит один раз и через некоторое случайное время $\tau > 0$. Например, 1) "отказ" – это наступление рефлекса выведения молока, τ – время до его наступления; 2) "отказ" – это инфаркт, или инсульт, или смерть, τ – временной интервал до его наступления.

Исходные данные представляют собой времена ожидания отказа $\tau_i, i = 1, \ldots, n$ для каждого из индивидов. Рассмотрим $n = 22$ наблюдения τ – времени ремиссии (в неделях).

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

По определению, кривая дожития – это $F(t) = P\{\tau > t\}$ – вероятность того, что отказ наступит после момента времени t. Обозначим через d_t количество отказов в момент времени t, r_t – количество объектов, у которых отказ наступает не раньше момента t, считаем, что $r_1 = n$. Тогда при полном наличии информации о временах отказа оценка кривой дожития вычисляется таким образом:

$$\widehat{F}_t = \frac{r_t - d_t}{n} = \frac{r_{t+1}}{n}\,. \tag{1}$$

t	r_t	d_t	$r_t - d_t$	$\widehat{F}_t$
0	22	0	22	1
1	22	3	19	$\frac{19}{22}$
2	19	2	17	$\frac{17}{22}$
3	17	1	16	$\frac{16}{22}$

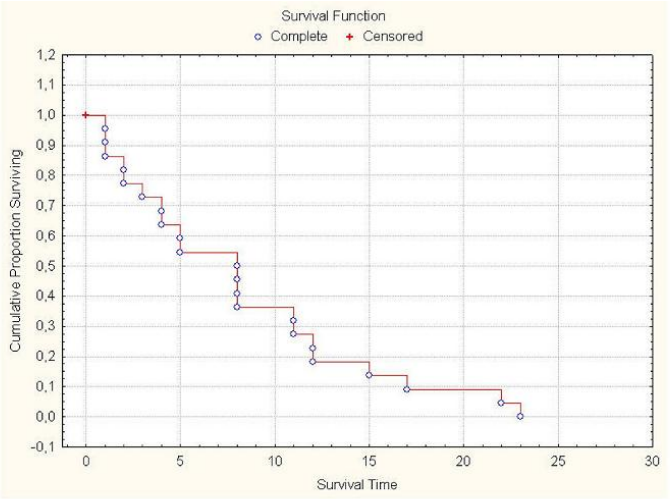


Рис. 1. Кривая дожития по времени ремиссии (без цензурирования).

Для фиксации данных требуется:

- четко установить начало отсчета,

- определить масштаб времени,

- определить понятие отказа.

Чаще всего стоит задача оценки вероятности дожития или сравниваются времена ожидания отказов в двух или нескольких выборках.

1.2. Функция риска

Кривую дожития иногда называют функцией надежности. Ее производная с обратным знаком называется плотностью и обозначается через $f(t)$.

$$f(t) = -F'(t) = -\lim_{\Delta\to 0}\frac{F(t+\Delta)-F(t)}{\Delta} = \lim_{\Delta\to 0}\frac{F(t)-F(t+\Delta)}{\Delta} =$$
$$= \lim_{\Delta\to 0}\frac{P\{\tau > t\} - P\{\tau > t+\Delta\}}{\Delta} = \lim_{\Delta\to 0}\frac{P\{t < \tau \le t+\Delta\}}{\Delta}$$

По смыслу, плотность $f(t)$ равна вероятности гибели в момент t.

Условная вероятность гибели в момент времени t выражается функ-

цией интенсивности $h(t)$ вида

$$h(t) = \lim_{\Delta \to 0} \frac{P\{t < \tau \leq t + \Delta | \tau > t\}}{\Delta} = \lim_{\Delta \to 0} \frac{P\{t < \tau \leq t + \Delta\}}{\Delta P\{\tau > t\}} = \frac{f(t)}{F(t)}, \quad (2)$$

$$\text{или} \quad h(t) = \frac{-F'(t)}{F(t)} = (-\ln F(t))'.$$

Минус логарифма от функции надежности называется *функцией риска:*

$$H(t) = -\ln F(t) \qquad F(t) = e^{-H(t)}.$$

Поскольку $F(0) = 1$, то $H(0) = 0$. Тогда учитывая, что $H'(t) = h(t)$, получаем:

$$H(t) = H(t) - H(0) = \int_0^t h(u)du \quad \text{и} \quad F(t) = e^{-\int_0^t h(u)du}.$$

Если риск (интенсивность) является постоянным $h(u) = \varrho$, то функция надежности убывает по экспоненте: $F(t) = e^{-\varrho t}$.

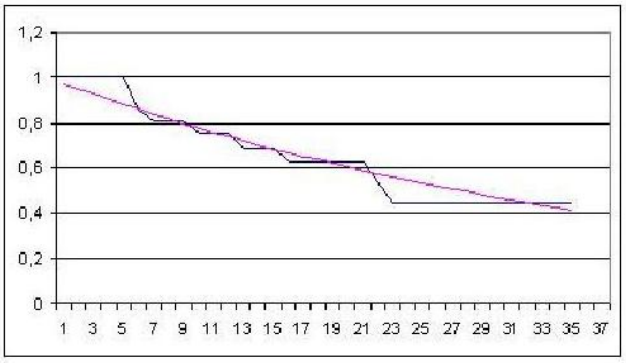


Рис. 2. Кривые дожития времени ремиссии с цензурированием и ее экспоненциальная модель).

1.3. Цензурирование и оценка Каплана-Мейера

Цензурирование – это неполное наблюдение за временем ожидания отказа. Некоторые объекты не могут наблюдаться в течение полного времени до отказа. Известным оказывается только то, что до некоторого времени

t отказ еще не наступил. Например, больной состоял на медицинском учете, а потом уехал в другой город, и о нем больше ничего не известно, или у животного фиксируемый рефлекс не наступил за установленное время эксперимента. При этом наблюдения записываются в виде пары признаков: в одном указывается или время наступления отказа, или время цензурирования, второй признак является индикатором цензурирования.

Если обозначить через τ_i время ожидания отказа, c_i – момент цензурирования, то исходные данные можно представить в виде:

$$(X_i, V_i), \quad \text{где } X_i = \min(\tau_i, c_i), \quad i = 1, \ldots, n,$$

$$V_i = \begin{cases} 0, & \text{если } \tau_i \le c_i \quad (\text{отказ }), \\ 1, & \text{если } \tau_i > c_i \quad (\text{цензурирование}). \end{cases}$$

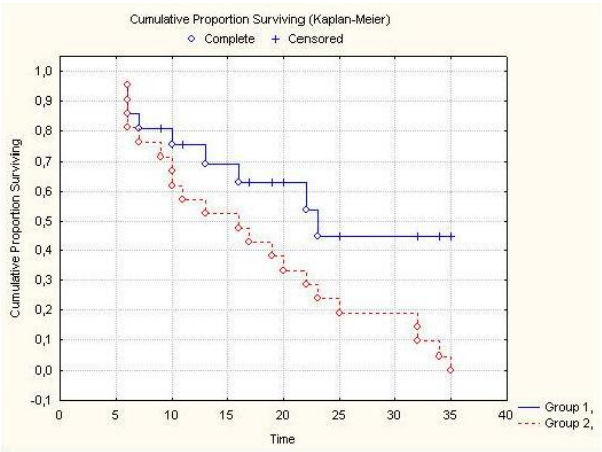


Рис. 3. Кривые дожития времени ремиссии с цензурированием и без цензурирования.

При наличии цензурирования в оценке кривой дожития используется другой принцип. Обозначим через p_i вероятность того, что при условии дожить до момента i отказ в этот момент так и не наступил. В качестве оценки вероятности p_i будем рассматривать отношение

$$\widehat{p}_i = \frac{r_i - d_i}{r_i} = 1 - \frac{d_i}{r_i}, \tag{3}$$

где r_i число наблюдений с отказом не ранее момента i, d_i – число отказов в момент i. Тогда

$$P\{\tau > t\} = P\{\tau > t-1\}P\{\tau > t|\tau > t-1\} = P\{\tau > t-1\}p_t$$

равна произведению вероятностей $P\{\tau > t\} = p_1 p_2 \dots p_t$, так как $P\{\tau > 0\} = 1$. Следовательно,

$$\widehat{P}\{\tau > t\} = \frac{r_1 - d_1}{r_1}\frac{r_2 - d_2}{r_2} \dots \frac{r_t - d_t}{r_t} \tag{4}$$

Обозначим через $h_i = \frac{d_i}{r_i}$ функцию интенсивности или условную вероятность погибнуть в момент i. Оценка кривой дожития, выраженная через интенсивности, носит название функции дожития Каплана-Мейера:

$$\widehat{P}\{\tau > t\} = \prod_{i=1}^{t} (1 - h_i) \ . \tag{5}$$

Для построения функции Каплана-Мейер рассмотрим данные с цензурированием, представленные в виде пары признаков:

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
X_j	6	6	6	6	7	9	10	10	11	13	16	17	19	20	22	23	25	32	32	34	35
V_j	1	0	0	0	0	1	1	0	1	0	0	1	1	1	0	0	1	1	1	1	1

i	d_i отказы	z_i цензурирование	$r_i = r_{i-1} -$ $-d_{i-1} - z_{i-1}$ оставшиеся	$\widehat{p}_i = 1 - h_i$ усл.вер-ть	$\widehat{F}_i$ $\prod_{j=1}^{i}(1-h_j)$
1	0	0	21	1	1
2	0	0	21	1	1
3	0	0	21	1	1
4	0	0	21	1	1
5	0	0	21	1	1
6	3	1	21	0.857	0.857
7	1	0	17	0.941	0.807
8	0	0	16	1	0.807
9	0	1	16	1	0.807
10	1	1	15	0.933	0.753
11	0	1	13	1	0.753
12	0	0	12	1	0.753
13	1	0	12	0.917	0.690
14	0	0	11	1	0.690
15	0	0	11	1	0.690
16	1	0	11	0.909	0.627
17	0	1	10	1	0.627
18	0	0	9	1	0.627
19	0	1	9	1	0.627
20	0	1	8	1	0.627
21	0	0	7	1	0.627
22	1	0	7	0.857	0.538
23	1	0	6	0.833	0.448
24	0	0	5	1	0.448
25	0	1	5	1	0.448
26	0	0	4	1	0.448
27	0	0	4	1	0.448
28	0	0	4	1	0.448
29	0	0	4	1	0.448
30	0	0	4	1	0.448
31	0	0	4	1	0.448
32	0	2	4	1	0.448
33	0	0	2	1	0.448
34	0	1	2	1	0.448
35	0	1	1	1	0.448

При отсутствии цензурирования количество выживших в момент $i+1$ равно $r_{i+1} = r_i - d_i$, поэтому в формуле (5) элементы в соседних дробях сокращаются, и остается выражение

$$\widehat{P}\{\tau > t\} = \frac{r_t - d_t}{r_1} = \frac{r_{t+1}}{r_1},$$

соответствующее оценке (1).

Если в интенсивности $h_i = d_i/r_i$ числитель и знаменатель разделить на n, то получим выражение интенсивности как отношение вероятности погибнуть в момент i к вероятности дожить до этого момента (2):

$$h_i = \frac{d_i/n}{r_i/n} = \frac{\widehat{f}_i}{\widehat{F}_{i-1}} = \frac{\hat{f}_i}{\hat{f}_i + \hat{f}_{i+1} + \ldots}.$$

По аналогии для генеральных дискретных характеристик получаем

$$h_1 = f_1, \quad \text{так как} \quad h_1 = \frac{f_1}{f_1 + f_2 + \ldots}, \quad \Longrightarrow \quad F_0 = \frac{f_1}{h_1} = 1;$$

$$h_2 = \frac{f_2}{f_2 + f_3 + \ldots} = \frac{f_2}{1 - f_1} = \frac{f_2}{1 - h_1}, \quad \Longrightarrow \quad f_2 = h_2(1 - h_1),$$

$$\Longrightarrow \quad F_1 = \frac{f_2}{h_2} = (1 - h_1),$$

$$f_3 + f_4 + \ldots = 1 - f_1 - f_2 = 1 - h_1 - h_2(1 - h_1) = (1 - h_1)(1 - h_2),$$

$$\Longrightarrow \quad f_3 = h_3(1 - h_1)(1 - h_2), \quad F_2 = (1 - h_1)(1 - h_2),$$

далее аналогично.

В *StatSoft* анализ данных типа времени жизни осуществляется при помощи программы *Statistics - Advanced Linear - Nonlinear Models - Survivel Analysis*.

Выбираем *Kaplan and Meier product – limit method*. Данные должны быть представлены по крайней мере в двух столбцах. В одном из них должны быть указаны наработки до отказа или до цензурирования, обозначим его, например, через X, во втором – индикатор C цензурирования. В *Variables* указываем требуемые переменные. После активизации признака C появляются окна, в которых нужно указать коды цензурирования. Для получения графика, подобного рис.1, нужно во вкладке *Kaplan – Meier plots* выбрать *Survivel times vs.cum. proption surviving*.

1.4. Оценка параметра постоянного риска

Функцией правдоподобия $L(x_1, \ldots, x_n)$ называется вероятность появления данной выборки $x_i = \min(t_i, c_i),\ i = 1, \ldots, n$.

$$L(x_1, \ldots, x_n) = \prod_{t_i} f(t_i, \varrho) \prod_{c_i} F(c_i, \varrho)\,,$$
$$\ln L(x_1, \ldots, x_n) = \sum_{t_i} \ln f(t_i, \varrho) + \sum_{c_i} \ln F(c_i, \varrho)\,.$$

Поскольку $f(t) = h(t)F(t)$, то

$$\ln L(x_1, \ldots, x_n) = \sum_{t_i} \ln h(t_i, \varrho) + \sum_{t_i} \ln F(t_i, \varrho) + \sum_{c_i} \ln F(c_i, \varrho) \ =$$

$$= \sum_{t_i} \ln h(t_i, \varrho) + \sum_{x_i} \ln F(x_i, \varrho) = \sum_{t_i} \ln h(t_i, \varrho) - \sum_{x_i} H(x_i, \varrho)\,.$$

Если риск постоянный, то $h(t_i, \varrho) = \varrho$, $H(t) = \varrho t$. Следовательно,

$$\ln L(x_1, \ldots, x_n) = \ln \varrho d - \varrho \sum_{x_i} t\,,$$

где d означает общее число отказов, $\sum\limits_{x_i} t$ – суммарную наработку. Оценки максимального правдоподобия находятся дифференцированием функции правдоподобия $\ln L(x_1, \ldots, x_n)$ по параметру ϱ.

$$\frac{\partial \ln L}{\partial \varrho} = \frac{d}{\varrho} - \sum_{x_i} t = 0.$$

Таким образом, параметр экспоненциальной модели кривой дожития вычисляется как отношения числа отказов к суммарной наработке. В нашем примере $\widehat{\varrho} = \frac{9}{359} = 0.025$.

При отсутствии цензурирования $\frac{2d\varrho}{\hat{\varrho}}$ имеет распределение хи-квадрат с $2d$ степенями свободы. Действительно, суммарная наработка имеет гамма

распределения

$$\sum_{i=1}^{d} x_i \sim \gamma(\varrho, d), \quad \frac{1}{d}\sum_{i=1}^{d} x_i \sim \gamma(d\varrho, d),$$

$$2d\varrho\left(\frac{1}{d}\sum_{i=1}^{d} x_i\right) = \frac{2d\varrho}{\hat{\varrho}} \sim \gamma(d\varrho\left(\frac{1}{2d\varrho}\right), d) = \gamma\left(\frac{1}{2}, \frac{2d}{d}\right) = \chi^2(2d)\,.$$

1.5. Доверительные интервалы для кривых дожития

При незначительном влиянии цензурированных данных используется формула для вычисления дисперсии

$$\mathbf{D}(\hat{F}_j) = \frac{\hat{F}_j^2(1-\hat{F}_j)}{r_{j+1}}\,.$$

Пусть n_0 число нецензурированных данных, $\hat{F}_j = \frac{r_{j+1}}{n_0}$, значение $n_0 = \frac{r_{j+1}}{\hat{F}_j}$ подставляем в выражение для дисперсии биномиальной частоты $\hat{F}_j$ в n_0 испытаниях:

$$\mathbf{D}(\hat{F}_j) = \frac{\hat{F}_j(1-\hat{F}_j)}{n_0} = \frac{\hat{F}_j^2(1-\hat{F}_j)}{r_{j+1}}\,.$$

Обозначим через $s_j = r_j - d_j$ число выживших при $t = j$. Если p_j - вероятность выжить в этот момент, то s_j имеет биномиальное распределение $\mathcal{B}(p_j, r_j)$. Формула Гринвуда:

$$\mathbf{D}(\hat{F}_j) = \hat{F}_j^2 \sum_{j=1}^{t} \frac{d_j}{r_j(r_j - d_j)}$$

Обозначим через $F_j = \mathbf{E}\hat{F}_j$, где $\hat{F}_j = \prod\limits_{j=1}^{t} \frac{s_j}{r_j}$. Очевидно, что $\mathbf{E}\left(\frac{s_j}{r_j}\right) = p_j$,

$$\frac{\hat{F}_j}{F_j} = \prod_{j=1}^{t} \frac{s_j}{r_j p_j} = \prod_{j=1}^{t}\left(1 + \frac{s_j - r_j p_j}{r_j p_j}\right) \approx 1 + \sum_{j=1}^{t} \frac{s_j - r_j p_j}{r_j p_j}$$

$$\mathbf{D}\left(\frac{\hat{F}_j}{F_j}\right) = \sum_{j=1}^{t} \frac{r_j p_j (1-p_j)}{(r_j p_j)^2} = \sum_{j=1}^{t} \frac{1-p_j}{r_j p_j} \approx \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j},$$

$$\mathbf{D}\hat{F}_j = {F_j}^2 \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j} \approx \hat{F}_j^2 \sum_{j=1}^{t} \frac{1-\hat{p}_j}{r_j \hat{p}_j} = \hat{F}_j^2 \sum_{j=1}^{t} \frac{d_j}{r_j(r_j-d_j)}\,.$$

1.6. Регрессионная модель Кокса

Рассматривается модель пропорциональных интенсивностей

$$h(t,z) = \psi(z,\beta) h_0(t) = e^{\beta^T z} h_0(t)\,,$$

в которой для любого объекта вектор поясняющих переменных z не зависит от времени. Будем считать, что базовая функция интенсивности $h_0(t)$ не известна. Обозначим через $\tau_1 < \tau_2 < \ldots < \tau_n$ упорядоченные моменты наработок до отказов у n объектов, T_j – номер объекта, отказавшего в момент τ_j, то есть

$$T_j = i \;\Leftrightarrow\; t_i = \tau_j\,.$$

Введем обозначение для множества риска – множества объектов, действующего непосредственно перед j-м отказом: $R(\tau_j) = \{i : t_i \geq \tau_j\}$, r_j – размер этого множества.

Пусть известна полная история до момента τ_j

$$H_{j-1} = \{\tau_1, \tau_2, \ldots, \tau_{j-1}; i_1, i_2, \ldots, i_{j-1}\}\,,$$

и один из объектов множества риска $R(\tau_j)$ отказал в момент τ_j. Условная вероятность того, что это был i-й объект, равна

$$p_j(i|i_1, i_2, \ldots, i_{j-1}) = \frac{h_i(\tau_j)}{\sum\limits_{k\in R(\tau_j)} h_k(\tau_j)} = \frac{\psi(i)}{\sum\limits_{k\in R(\tau_j)} \psi(k)}\,,$$

где $\psi(i) = \psi(z_i, \beta)$ есть множитель, влияющий на интенсивность i-го объекта. По цепному правилу для условных вероятностей получаем выражение для функции правдоподобия или вероятность истории H_n

$$p(i_1, \ldots, i_n) = \prod_{j=1}^{n} p_j(i_j | i_1, i_2, \ldots, i_{j-1}) = \prod_{j=1}^{n} \frac{\psi(i_j)}{\sum\limits_{k \in R(\tau_j)} \psi(k)} .$$

При наличии цензурирования произведение осуществляется по отказам:

$$lik = \prod_{i \in D} \frac{\psi(i)}{\sum\limits_{k \in R_i} \psi(k)},$$

где через D обозначено множество отказавших объектов, через R_i соответствующее i множество риска. При $\psi(z; \beta) = e^{\beta^T z}$ для оценивания параметров вектора β логарифм функции правдоподобия

$$\ln(lik) = \sum_{i \in D} \left(\ln \psi(i) - \ln \sum_{k \in R(\tau_j)} \psi(k) \right) = \sum_{i \in D} \left(\beta^T z_i - \ln \sum_{k \in R(\tau_j)} e^{\beta^T z_k} \right)$$

дифференцируют по параметрам, приравнивают производные к нулю и решают систему уравнений:

$$\sum_{i \in D} (z_{i\nu} - A_{i\nu}(\beta)) = 0, \ \nu = 1, 2, \ldots, K,$$

где $z_{i\nu}$ – значение ν-й из K компонент поясняющей переменной z для i-го объекта,

$$A_{i\nu}(\beta) = \frac{\sum\limits_{k \in R_i} z_{k\nu} e^{\beta^T z_k}}{\sum\limits_{k \in R_i} e^{\beta^T z_k}}$$

взвешенное среднее переменной $z_{i\nu}$ по соответствующему множеству риска.

В случае двух выборок имеется одна $(K = 1)$ поясняющая переменная z, которая принимает два значения: 0, если объект принадлежит нулевой выборке, и 1, если принадлежит первой выборке.

$$\psi(i) = \begin{cases} e^{\beta}, & \text{i-ый объект из группы 1} \\ 1, & \text{i-ый объект из группы 0} \end{cases}$$

Обозначим через d_1 количество отказов в первой выборке, через r_{0i} и r_{1i} объемы множеств риска для каждой выборки в момент i. Логарифм функции правдоподобия равен

$$\ln(lik) = \sum_{i\in D}\left(\ln\psi(i) - \ln(\sum_{k\in R_i}\psi(k))\right) = \beta d_1 - \sum_{i\in D}\ln(r_{1i}e^{\beta} + r_{0i}),$$

$$\frac{d\ln(lik)}{d\beta} = d_1 - \sum_{i\in D}\frac{r_{1i}e^{\beta}}{r_{1i}e^{\beta} + r_{0i}} = 0.$$

Из последнего уравнения находится оценка параметра $\widehat{\beta}$.

1.7. Сравнение двух групп по кривым дожития

Статистики, используемые для проверки гипотезы однородности, можно получить из одной таблицы. Для упрощения рассмотрим нецензурированные данные. Обозначим через d_{0i} и d_{1i} количество отказов в группе 0 и 1 в момент времени i, через $d_i = d_{0i} + d_{1i}$ число отказов в обеих группах. Пусть r_{0i} и r_{1i} – количество доживших до момента i.

время	группа 0		группа 1		всего		*Peto Cox*		Cox-Mantel	Log-rank
i	r_{0i}	d_{0i}	r_{1i}	d_{1i}	r_i	d_i	$\frac{d_i r_{0i}}{r_i}$	$\frac{d_i r_{1i}}{r_i}$	$\frac{d_i r_{0i} r_{1i}}{r_i^2}$	$\frac{d_i r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i-1)}$
1	3	1	3	0	6	1	0.5	0.5	0.25	0.25
2	2	1	3	0	5	1	0.4	0.6	0.24	0.24
3	1	0	3	1	4	1	0.25	0.75	0.1875	0.1875
4	1	0	2	1	3	1	$\frac{1}{3}$	$\frac{2}{3}$	0.22	0.25
5	1	1	1	0	2	1	0.5	0.5	0	1.15
6	0	0	1	1	1	1	0	1		
сумма		3		3			1.98	4.02	1.15	1.15
		d_0		d_1			E_0	E_1	D_{cox}	D

Обозначим через E_0 и E_1 математические ожидания числа отказов

$$d_0 = \sum_{i=1}^{n} d_{0i}, \quad d_1 = \sum_{i=1}^{n} d_{1i}.$$

При $\beta = 0$ ожидаемые числа отказов как суммы математических ожиданий

гипергеометрических случайных величин равны

$$\widehat{E}_0 = \sum_{i=1}^{n} \frac{d_i r_{0i}}{r_i}, \quad \widehat{E}_1 = \sum_{i=1}^{n} \frac{d_i r_{1i}}{r_i}\,.$$

В каждый момент времени общее множество рисков имеет мощность $r_i = r_{0i} + r_{1i}$, всего отказов d_i. Если ξ случайное число отказов из 0-группы подчиняется гипергеометрическому распределению:

$$P\{\xi = d_{0i}\} = \frac{C_{r_{0i}}^{d_{0i}} C_{r_{1i}}^{d_{1i}}}{C_{r_i}^{d_i}}$$

с математическим ожиданием $\mathbf{E}\xi = \frac{d_i r_{0i}}{r_i}$ и дисперсией $\mathbf{D}\xi = d_i \frac{r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i - 1)}$. Лог-ранговый критерий проверки однородности основан на асимптотической нормальности числа отказов d_0, которое в случае однородных выборок имеет нормальное распределение со средним E_0 и дисперсией

$$D = \sum_{i=1}^{n} d_i \frac{r_{0i} r_{1i}(r_i - d_i)}{r_i^2(r_i - 1)}.$$

Статистика лог-рангового критерия вида

$$\chi^2 = \frac{(d_0 - E_0)^2}{D} = 0.904,\ p = 0.34$$

в случае однородных кривых дожития имеет распределение хи-квадрат с одной степенью свободы $\chi^2(1)$.

Пусть $D_{cox} = \sum\limits_{i=1}^{n} \frac{d_i r_{0i} r_{1i}}{r_i^2}$. Первая приближенная статистика Cox-Mantel вида:

$$\chi^2 = \frac{(d_0 - E_0)^2}{D_{cox}} = 0.904,\ p = 0.34$$

также имеет в случае однородных кривых дожития распределение $\chi^2(1)$. Вторая приближенная статистика Peto вида:

$$\chi^2 = \frac{(d_0 - E_0)^2}{E_0} + \frac{(d_1 - E_1)^2}{E_1} = 0.784\,,\ p = 0.376,$$

имеет в случае однородных кривых дожития распределение $\chi^2(1)$. При построении статистики Кокса используется отношение

$$F = \frac{\widehat{E}_0/(2d_0)}{\widehat{E}_1/(2d_1)} = 2.03,\ p = 0.2,$$

которое имеет распределение Фишера с $2d_0$ и $2d_1$ степенями свободы.

Рис. 4. Сравнение кривых дожития.

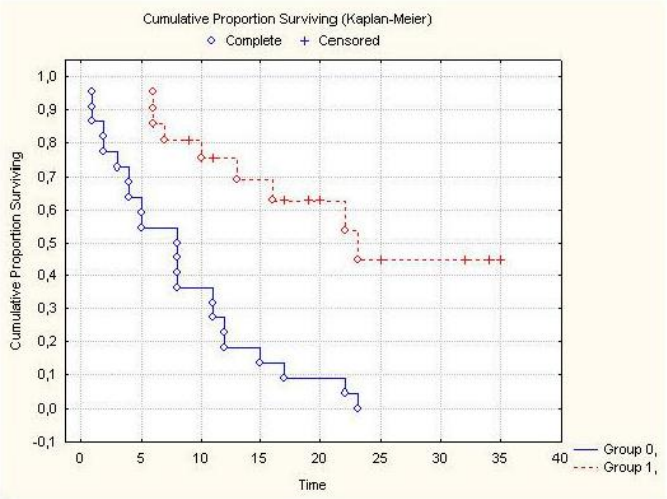


Рис. 5. Кривые дожития по времени ремиссии для двух групп.

2. Анализ кривых дожития в R

2.1. Кривая Каплана-Мейера

Рассмотрим для примера данные о больных глиомой из BMA.

$data <- read.csv('base\_short.xls', sep =';')[, -1]$

$colnames(data)$

diagnosis тяжесть заболевания: 2 - диффузная астроцитома, 3 - анапластическая астроцитома, 4 - мультиформная глиобластома.

karnovsky индекс Карновского: чем ниже, тем тяжелее состояние.

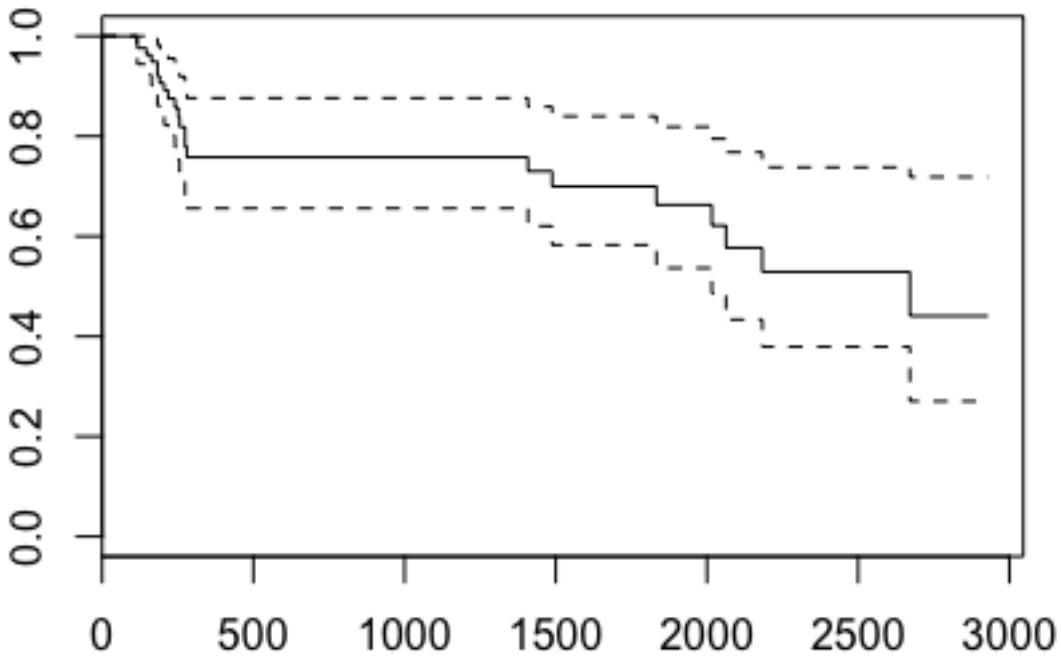
c цензурирование: 0 - больной умер, 1 - выбыл из наблюдения.

time время в днях от дня операции либо до момента выбывания из наблюдений, либо до смерти больного.

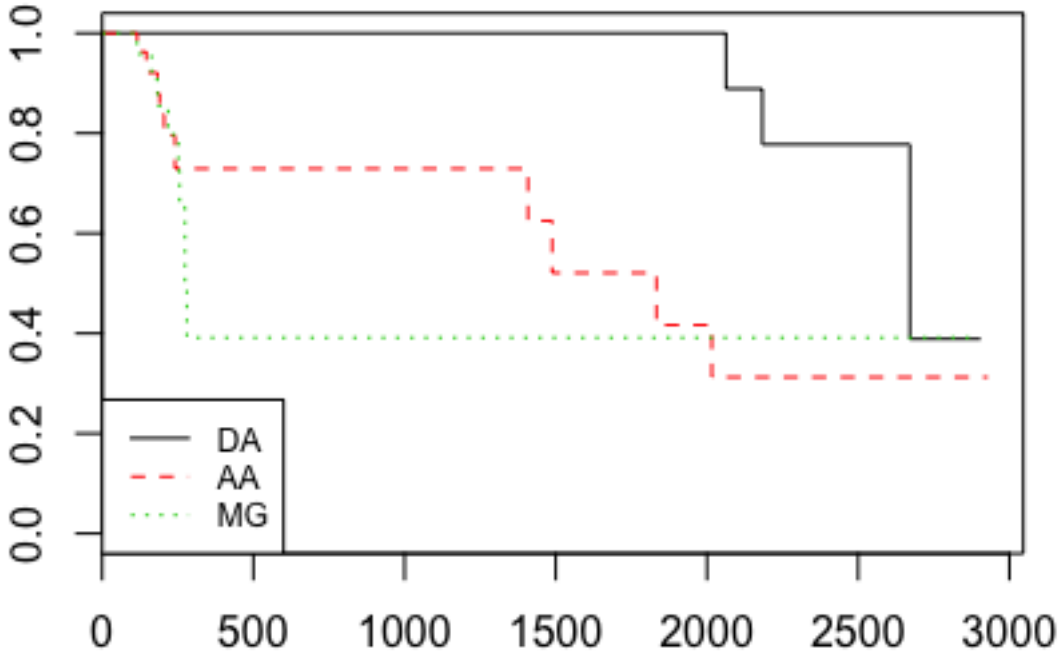
Для построения кривой дожития Каплана-Мейера воспользуемся следующей функцией.

$fitKM <- survfit(Surv(time, event = 1 - c) \sim 1, data = data)$

$plot(fitKM, mark.time = FALSE, col = 1, conf.int = TRUE)$


0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000

2.2. Сравнение по группам



Для сравнения кривых дожития используются характеристики медианы дожития и среднего времени дожития. Как видно из следующей таблицы, наименьшее время дожития наблюдается при мультиформной глиобластоме. Известно, что при безоперационном течении заболевания дожитие не превышает трех месяцев.

$print(fitKM, print.rmean = TRUE)$

	n	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
diagnosis=2	51	3	2640	113	2671	2671	NA
diagnosis=3	41	10	1661	263	1833	1408	NA
diagnosis=4	40	9	1281	368	274	256	NA

Для того чтобы определить значимость различия по группам, исполь-

зуем следующую функцию.

$survdiff(Surv(time, event = 1-c) \sim diagnosis, data = data, rho = 0)$

В результате получаем, что различие кривых дожития значимо.

$Chisq = 13.5 on 2 degrees of freedom, p = 0.00115$

Параметр $rho = 0$ означает, что использовался лог-ранговый критерий. Для попарных сравнений можно использовать критерий Гехана-Вилкоксона. Например, следующим образом.

$$with(na.omit(subset(data, data\$diagnosis! = 2, \\ select = c(time, diagnosis, c))), \\ gehan.test(time, 1 - c, diagnosis))$$

откуда получаем, что $p_{3,4} = 0.6936468$. Аналогично, $p_{2,4} = 0.0001231596$, $p_{2,3} = 0.0007377914$, что свидетельствует о том, что имеет место значимое отличие группы с диагнозом DA от двух других.

2.3. Регрессионная модель Кокса

Для анализа кривых дожития возможно привлечение ковариат. Используем, например, индекс Карновского, и выясним, насколько значимо влияет этот индекс на дожитие. Доверительный уровень вероятности 0.075 указывает на значимый характер влияния индекса Карновского с уровнем значимости 0.1.

$$fit < -coxph(Surv(time, event = 1 - c) \sim diagnosis + karnovsky, \\ data = data)$$

$summary(fit)$

В результате (табл. 1) и

$Score(logrank)test = 19.86 on 3 df, p = 0.0001818$

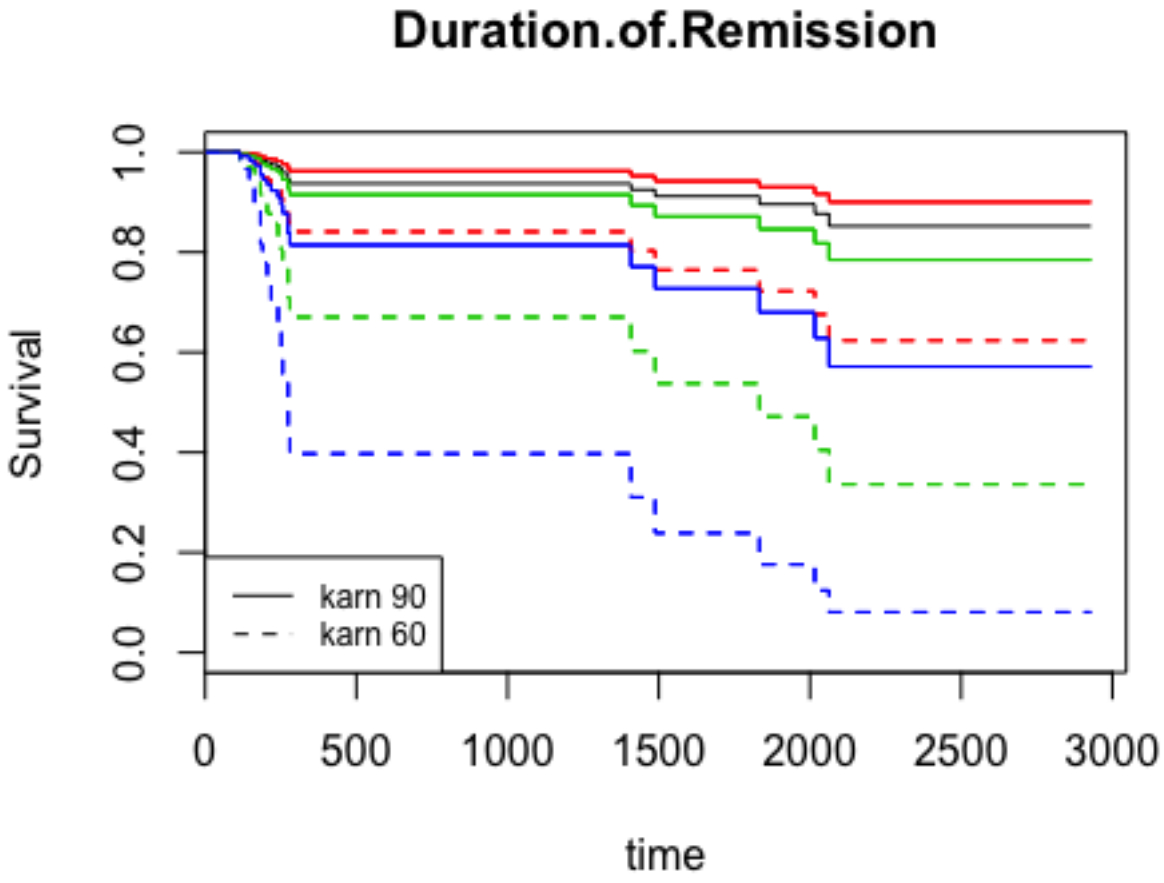
	coef	exp.coef.	se.coef.	z	Pr...z..
diagnosis	0.837	2.309	0.339	2.472	0.013
karnovsky	-0.050	0.951	0.028	-1.783	0.075

Таблица 1. Регрессионная модель Кокса

По знакам коэффициентов $coef$ (табл.1) можно сделать вывод, что при более слабом диагнозе и при более высоком индексе Карновского показатели дожития лучше.

Для построения кривых дожития при разных значениях параметров возможен следующий вариант.

$$fit1 < -coxph(Surv(time, event = 1 - c) \sim karnovsky, data = na.omit(data))$$
$$plot(survfit(fit1, newdata = data.frame(karnovsky = 90)), type =' n',$$
$$xscale = 1, xlab =' time', ylab =' Survival', conf.int = FALSE)$$
$$for(i\ in\ c(2, 3, 4))$$
$$\{$$
$$lines(survfit(fit, newdata = data.frame(diagnosis = i, karnovsky = 90)),$$
$$conf.int = FALSE, col = i)$$
$$lines(survfit(fit, newdata = data.frame(diagnosis = i, karnovsky = 60)),$$
$$conf.int = FALSE, lty = 2, col = i)$$
$$\}$$
$$legend('bottomleft', c('karn90',' karn60'), lty = seq(2), cex = 0.75)$$
$$title(main =' Duration.of.Remission')$$


Duration.of.Remission
Survival
0.0
0.2
0.4
0.6
0.8
1.0
karn 90
karn 60
0
500
1000
1500
2000
2500
3000
time