

АНАЛИЗ ПРОПОРЦИОНАЛЬНЫХ РИСКОВ В ПРОГРАММНОЙ СРЕДЕ R: ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ ДОКТОРАНТОВ ПО СПЕЦИАЛЬНОСТИ «МЕДИЦИНА» И «ОБЩЕСТВЕННОЕ ЗДОРОВЬЕ»

В.Л. ЕГОШИН¹, Н.В. САВВИНА², А.М. ГРЖИБОВСКИЙ^{1,2}

¹Северный государственный медицинский университет, г. Архангельск, Россия

²Северо-Восточный федеральный университет, г. Якутск, Россия

Егошин В.Л. – <https://orcid.org/0000-0002-8407-3789>

Саввина Н.В. – <https://orcid.org/0000-0003-2441-6193>

Гржибовский А.М. – <https://orcid.org/0000-0002-5464-0498>

Citation/

библиографиялық сілтеме/
библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Proportional hazard analysis in R: practical guidelines for PhD students in medicine and public health. West Kazakhstan Medical Journal. 2020;62(2):95–103.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Анализ пропорциональных рисков в программной среде R: практические рекомендации для докторантов по специальности «Медицина» и «Общественное здоровье». West Kazakhstan Medical Journal. 2020;62(2):95–103.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортасында пропорционалды тәуекелді талдау: «Медицина» және «Қоғамдық денсаулық» мамандықтары бойынша докторанттарға арналған практикалық ұсынымдар. West Kazakhstan Medical Journal. 2020;62(2):95–103.

Proportional hazard analysis in R: practical guidelines for PhD students in medicine and public health

Egoshin VL¹, Savvina NV², Grjibovski AM^{1,2}

¹Northern State Medical University, Arkhangelsk, Russia

²North-Eastern Federal University, Yakutsk, Russia

In this paper we describe basic principles of using R package for constructing proportional hazard analysis models in survival analysis. We present step-by-step guidelines and syntax for the analysis using practical example with real and freely available data to simplify educational process. In addition to the syntax we present R outputs and their interpretation.

Keywords: R, survival analysis, proportional hazard analysis, Cox regression, syntax, listing.

R бағдарламалық ортасында пропорционалды тәуекелді талдау: «Медицина» және «Қоғамдық денсаулық» мамандықтары бойынша докторанттарға арналған практикалық ұсынымдар

В.Л. Егошин¹, Н.В. Саввина², А.М. Гржибовский^{1,2}

¹Солтүстік мемлекеттік медицина университеті, Архангельск, Ресей

²Солтүстік-Шығыс федералды университеті, Якутск, Ресей

Бұл жұмыста Кокс пропорционалды тәуекел әдісінің көмегімен өмір сүрудің зерттеуші-медиктер арасындағы белгілі анализі үшін R бағдарламалық орталығын қолданудың негізгі принциптері ұсынылған. Білім алушылардың практикалық жұмысы үшін еркін қол жетімділіктегілердің нақты мәліметтерін қолдана отырып практикалық мысал түрінде R-де кадамдық алгоритмі мен синтаксисі берілген. Синтаксисінен бөлек нәтижелері R, сондай-ақ олардың түсіндірмелері көрсеткендей түрде ұсынылған.

Негізгі сөздер: R, өмір сүру анализі, пропорционалды тәуекелділік талдауы, Кокс регрессиясы, синтаксис, листинг.

Анализ пропорциональных рисков в программной среде R: практические рекомендации для докторантов по специальности «Медицина» и «Общественное здоровье»

Егошин В.Л.¹, Саввина Н.В.², Гржибовский А.М.^{1,2}

¹Северный государственный медицинский университет, г. Архангельск, Россия

²Северо-Восточный федеральный университет, г. Якутск, Россия

В данной работе представлены основные принципы применения программной среды R для популярного среди исследователей-медиков анализа выживаемости с помощью метода пропорциональных рисков Кокса. Представлен пошаговый алгоритм и синтаксис в R для в виде практического примера с использованием реальных данных, находящихся в свободном доступе для практической работы обучающихся. Помимо синтаксиса представлены результаты в том виде, как их выдает R, а также их интерпретация.

Ключевые слова: R, анализ выживаемости, анализ пропорциональных рисков, регрессия Кокса, синтаксис, листинг.



Гржибовский А.М.
e-mail: andrej.grjibovski@gmail.com

Received/
Келіп түсті/
Поступила:
20.06.2020.

Accepted/
Басылымға қабылданды/
Принята к публикации:
22.06.2020.

ISSN 2707-6180 (Print)
© 2020 The Authors
Published by West Kazakhstan Marat Ospanov
Medical University

Модель пропорциональных рисков (Cox proportional hazards model), предложенная D. R. Cox [1], - наиболее часто встречающийся многомерный метод анализа выживаемости в медицинских исследованиях [2-8]. С помощью этого метода изучается зависимость времени дожития (survival time) от независимых переменных-предикторов (predictor variables, covariates). Это регрессионная модель, описывающая связь между частотой случаев и выражающаяся как функция риска и набор ковариат. Этот полупараметрический метод предполагает прогнозирование риска наступления события (hazard risk) для рассматриваемого объекта и оценивает влияние независимых переменных на этот риск. При этом риск наступления события является функцией, зависимой от времени, и выявляет вероятность наступления события для субъектов.

Математически модель пропорциональных рисков Кокса может быть записана как

$$h(t) = h_0(t) + \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p),$$

где функция риска $h(t)$ зависит от (определяется) набором p ковариат (x_1, x_2, \dots, x_p) , чье влияние зависит от величины коэффициентов $(\beta_1, \beta_2, \dots, \beta_p)$.

Особое значение в модели пропорциональных рисков Кокса имеет отношение рисков (HR - hazard ratio) для ковариаты, являющееся экспоненцированной величиной -коэффициента. HR показывает во сколько раз изменяется риск события при сравнении с референтным значением. Если HR меньше 1, то риск события уменьшается, если больше - увеличивается. При этом надо обращать внимание на доверительный интервал. Если доверительный интервал включает 1 (единицу), то можно говорить об отсутствии статистически значимого влияния этого фактора на исход.

Использование модели пропорциональных рисков для анализа выживаемости в R

Как и другие статистические программы, программная среда R часто используется для применения анализа выживаемости [9]. Наиболее часто в R используется пакет `survival` [10]. В данной работе при выполнении анализа выживаемости также применены пакет `survminer` [11] для создания графиков, пакеты `broom` [12], `finalfit` [13], `gtsummary` [14] - для вывода данных в табличной форме. В работе использованы базовые пакеты R 3.6.2 [15], а также пакеты `dplyr` [16], `ggplot2` [17], `tidyr` [18], входящие в пакет `tidyverse` [19], `arsenal` [20], `knitr` [21]. Мы рекомендуем все эти пакеты для практической работы для решения задач, связанных с определением выживаемости пациентов в медицинских исследованиях. Работа выполнена в IDE RStudio ver. 1.2.5033.

Используемые данные

Для практического обучающего примера использован набор данных из пакета `survival`. Это реальные данные, полученные в ходе изучения адьювантной

химиотерапии при лечении рака толстого кишечника. Результаты этой работы были опубликованы Laurie et al. [22] и Moertel et al. [23]. В набор включены данные о методах лечения, поле и возрасте пациентов, характере опухоли (степень дифференциации, распространение, связь с соседними органами, наличие обструкции, перфорации), состоянии лимфатических узлов, времени наступления события или цензурирования и другие. Пациенты были разделены на три группы по терапевтическому подходу (переменная `rx`): наблюдение (Obs), адьювантная терапия левамизолом (Lev), или левамизол + 5-фторурацил (Lev+5FU). В листинге 1 представлены используемые пакеты, выполнена загрузка и модификация данных. Для анализа отобраны пациенты в возрасте от 40 до 60 лет без обструкции, без перфорации, прорастания в соседние органы. Данные об отобранных для изучения участников представлены в таблице 1. В таблице также указаны статистические методы, использованные для сравнения между группами.

Листинг 1

использованные пакеты

```
library(tidyverse)
library(broom)
library(survival)
library(survminer)
library(finalfit)
```

загрузка данных

```
data(«colon»)
```

модификация набора данных

```
colon_mod <- colon %>%
  filter(obstruct == 0, perfor == 0, adhere == 0, age >= 40, age <
  60) %>%
  mutate(
    sex = factor(sex, labels = c("female", "male")),
    differ = factor(differ, labels = c("well", "moderate", "poor")),
    extent = factor(extent, labels = c("submuc.", "muscle",
    "serosa", "contig.)),
    node4 = factor(node4, labels = c("<= 4", "More 4"))
  ) %>%
  select(c(3:12, 14, 15))
summary(tableby(rx ~ status + time + differ + extent + node4 +
sex + age, data = colon_mod),
  pfootnote = TRUE,
  title = "Таб. 1 Данные об отобранных для изучения
участниках исследования")
```

Анализ пропорциональных рисков Кокса в R

На первом этапе создается Survival Object с использованием функции `Surv` из пакета `survival`, в дальнейшем он будет использоваться как переменная отклика в формуле модели. Для объекта (в случае цензурирования справа) необходимы два аргумента: `time` - время до наступления события или цензурирования, `status`

- принимает значение 1 в случае наступления события и 0 при его отсутствии (цензурировании). Формат функции - `Surv(time, status)`.

Для модели регрессии Кокса используется функция `coxph` в формате `coxph(formula = Surv(time, status) ~ covariates, data = , ...)`.

В R имеется много возможностей увидеть данные о модели (листинг 2): с использованием функции `summary` базового пакета и функций `tidy`, `glance`, `augment` из пакета `broom`, функции `finalfit` одноимённого пакета, функции `tbl_regression` пакета `gtsummary`, функция `ggforest` пакета `survminer` даёт графическое представление о коэффициентах риска. Функция `kable` пакета `knitr` используется для вывода результатов в виде таблицы.

Листинг 2

```
# создание модели Кокса регрессии
coxph_fit <- coxph(Surv(time, status) ~ rx + extent + differ +
node4 + sex, data = colon_mod)

summary(coxph_fit)
Call:
coxph(formula = Surv(time, status) ~ rx + extent + differ +
node4 +
sex, data = colon_mod)

n= 462, number of events= 203
(14 observations deleted due to missingness)

      coef exp(coef) se(coef)  z Pr(>|z|)
rxLev  -0.4996  0.6067  0.1841 -2.714  0.00665
rxLev+5FU -0.4150  0.6603  0.1680 -2.470  0.01350
```

Таб. 1 Данные об отобранных для изучения участниках исследования

| | Obs (N=182) | Lev (N=132) | Lev+5FU (N=162) | Total (N=476) | p value |
|---------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| status | | | | | 0.030 ¹ |
| Mean (SD) | 0.511 (0.501) | 0.379 (0.487) | 0.395 (0.490) | 0.435 (0.496) | |
| Range | 0.000 - 1.000 | 0.000 - 1.000 | 0.000 - 1.000 | 0.000 - 1.000 | |
| time | | | | | 0.040 ¹ |
| Mean (SD) | 1563.786 (892.095) | 1771.000 (939.095) | 1791.438 (914.899) | 1698.727 (917.358) | |
| Range | 45.000 - 3192.000 | 38.000 - 2951.000 | 23.000 - 3309.000 | 23.000 - 3309.000 | |
| differ | | | | | 0.001 ² |
| N-Miss | 6 | 4 | 4 | 14 | |
| well | 6 (3.4%) | 20 (15.6%) | 18 (11.4%) | 44 (9.5%) | |
| moderate | 150 (85.2%) | 88 (68.8%) | 112 (70.9%) | 350 (75.8%) | |
| poor | 20 (11.4%) | 20 (15.6%) | 28 (17.7%) | 68 (14.7%) | |
| extent | | | | | 0.054 ² |
| submuc. | 10 (5.5%) | 2 (1.5%) | 12 (7.4%) | 24 (5.0%) | |
| muscle | 20 (11.0%) | 18 (13.6%) | 18 (11.1%) | 56 (11.8%) | |
| serosa | 146 (80.2%) | 106 (80.3%) | 132 (81.5%) | 384 (80.7%) | |
| contig. | 6 (3.3%) | 6 (4.5%) | 0 (0.0%) | 12 (2.5%) | |
| node4 | | | | | 0.156 ² |
| <= 4 | 136 (74.7%) | 86 (65.2%) | 110 (67.9%) | 332 (69.7%) | |
| More 4 | 46 (25.3%) | 46 (34.8%) | 52 (32.1%) | 144 (30.3%) | |
| sex | | | | | 0.492 ² |
| female | 84 (46.2%) | 56 (42.4%) | 80 (49.4%) | 220 (46.2%) | |
| male | 98 (53.8%) | 76 (57.6%) | 82 (50.6%) | 256 (53.8%) | |
| age | | | | | 0.096 ¹ |
| Mean (SD) | 52.945 (5.231) | 52.197 (5.134) | 51.691 (5.765) | 52.311 (5.408) | |
| Range | 40.000 - 59.000 | 41.000 - 59.000 | 40.000 - 59.000 | 40.000 - 59.000 | |

1. Linear Model ANOVA
2. Pearson's Chi-squared test

```
extentmuscle 1.9132 6.7744 1.0363 1.846 0.06486
extentserosa 2.4584 11.6857 1.0064 2.443 0.01457
extentcontig. 2.1909 8.9435 1.1230 1.951 0.05106
differmoderate 0.4481 1.5653 0.3200 1.400 0.16147
differpoor 0.8522 2.3448 0.3488 2.443 0.01456
node4More 4 0.8019 2.2298 0.1458 5.499 3.82e-08
sexmale -0.3127 0.7315 0.1429 -2.188 0.02865
```

```
rxLev **
rxLev+5FU *
extentmuscle .
extentserosa *
extentcontig. .
differmoderate
differpoor *
node4More 4 ***
sexmale *
```

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
exp(coef) exp(-coef) lower .95 upper .95
rxLev 0.6067 1.64813 0.4230 0.8704
rxLev+5FU 0.6603 1.51436 0.4751 0.9179
extentmuscle 6.7744 0.14761 0.8888 51.6343
extentserosa 11.6857 0.08557 1.6257 83.9995
extentcontig. 8.9435 0.11181 0.9900 80.7950
differmoderate 1.5653 0.63884 0.8360 2.9311
differpoor 2.3448 0.42647 1.1836 4.6454
node4More 4 2.2298 0.44848 1.6754 2.9675
```

```
sexmale 0.7315 1.36707 0.5528 0.9679
```

Concordance= 0.676 (se = 0.019)

Likelihood ratio test= 76.1 on 9 df, p=1e-12

Wald test = 62.24 on 9 df, p=5e-10

Score (logrank) test = 71.33 on 9 df, p=8e-12

```
glance(coxph_fit) %>%
```

```
select(1:8) %>%
```

```
knitr::kable(digits = 3, caption = "Таб.2 Общие данные о модели (начало)")
```

```
glance(coxph_fit) %>%
```

```
select(9:15) %>%
```

```
knitr::kable(digits = 3, caption = "Таб.3 Общие данные о модели (продолжение)")
```

```
tidy(coxph_fit, exponentiate = TRUE) %>%
```

```
knitr::kable(digits = 3, caption = "Таб.4 Коэффициенты HR")
```

```
# с использованием пакета finalfit
```

```
explanatory <- c("rx", "extent", "differ", "node4", "sex")
```

```
dependent <- "Surv(time, status)"
```

```
colon_mod %>%
```

```
finalfit(dependent, explanatory) %>%
```

```
knitr::kable(
```

```
row.names = FALSE, align = c(rep("l", 2), rep("r", 4)),
```

```
caption = "Таб.5 Коэффициенты HR (уни- и мультивариантные)")
```

```
)
```

```
gtsummary::tbl_regression(coxph_fit, exponentiate =
```

Таб.2 Общие данные о модели (начало)

| n | nevent | statistic.log | p.value.log | statistic.sc | p.value.sc | statistic.wald | p.value.wald |
|-----|--------|---------------|-------------|--------------|------------|----------------|--------------|
| 462 | 203 | 76.097 | 0 | 71.329 | 0 | 62.24 | 0 |

Таб.3 Общие данные о модели (продолжение)

| r.squared | r.squared.max | concordance | std.error.concordance | logLik | AIC | BIC |
|-----------|---------------|-------------|-----------------------|-----------|----------|----------|
| 0.152 | 0.994 | 0.676 | 0.019 | -1145.817 | 2309.634 | 2339.453 |

Таб.4 Коэффициенты HR

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|----------------|----------|-----------|-----------|---------|----------|-----------|
| rxLev | 0.607 | 0.184 | -2.714 | 0.007 | 0.423 | 0.870 |
| rxLev+5FU | 0.660 | 0.168 | -2.470 | 0.014 | 0.475 | 0.918 |
| extentmuscle | 6.774 | 1.036 | 1.846 | 0.065 | 0.889 | 51.634 |
| extentserosa | 11.686 | 1.006 | 2.443 | 0.015 | 1.626 | 84.000 |
| extentcontig. | 8.943 | 1.123 | 1.951 | 0.051 | 0.990 | 80.795 |
| differmoderate | 1.565 | 0.320 | 1.400 | 0.161 | 0.836 | 2.931 |
| differpoor | 2.345 | 0.349 | 2.443 | 0.015 | 1.184 | 4.645 |
| node4More 4 | 2.230 | 0.146 | 5.499 | 0.000 | 1.675 | 2.967 |
| sexmale | 0.731 | 0.143 | -2.188 | 0.029 | 0.553 | 0.968 |

Таб.5 Коэффициенты HR (уни- и мультивариантные)

| Dependent: Surv(-time, status) | | all | HR (univariable) | HR (multivariable) |
|--------------------------------|----------|-------------|------------------------------|-----------------------------|
| rx | Obs | 182 (100.0) | - | - |
| | Lev | 132 (100.0) | 0.69 (0.49-0.97, p=0.033) | 0.61 (0.42-0.87, p=0.007) |
| | Lev+5FU | 162 (100.0) | 0.70 (0.51-0.97, p=0.030) | 0.66 (0.48-0.92, p=0.014) |
| extent | submuc. | 24 (100.0) | - | - |
| | muscle | 56 (100.0) | 7.35 (0.97-55.64, p=0.053) | 6.77 (0.89-51.63, p=0.065) |
| | serosa | 384 (100.0) | 15.89 (2.23-113.43, p=0.006) | 11.69 (1.63-84.00, p=0.015) |
| | contig. | 12 (100.0) | 16.22 (1.95-134.72, p=0.010) | 8.94 (0.99-80.80, p=0.051) |
| differ | well | 44 (100.0) | - | - |
| | moderate | 350 (100.0) | 1.99 (1.08-3.67, p=0.027) | 1.57 (0.84-2.93, p=0.161) |
| | poor | 68 (100.0) | 3.11 (1.59-6.11, p=0.001) | 2.34 (1.18-4.65, p=0.015) |
| node4 | <= 4 | 332 (100.0) | - | - |
| | More 4 | 144 (100.0) | 2.38 (1.81-3.14, p<0.001) | 2.23 (1.68-2.97, p<0.001) |
| sex | female | 220 (100.0) | - | - |
| | male | 256 (100.0) | 0.86 (0.66-1.13, p=0.288) | 0.73 (0.55-0.97, p=0.029) |

TRUE)

| N = 462 | HR | 95% CI | p-value |
|----------|------|------------|---------|
| rx | | | |
| Obs | — | — | |
| Lev | 0.61 | 0.42, 0.87 | 0.007 |
| Lev+5FU | 0.66 | 0.48, 0.92 | 0.014 |
| extent | | | |
| submuc. | — | — | |
| muscle | 6.77 | 0.89, 51.6 | 0.065 |
| serosa | 11.7 | 1.63, 84.0 | 0.015 |
| contig. | 8.94 | 0.99, 80.8 | 0.051 |
| differ | | | |
| well | — | — | |
| moderate | 1.57 | 0.84, 2.93 | 0.2 |
| poor | 2.34 | 1.18, 4.65 | 0.015 |
| node4 | | | |
| <= 4 | — | — | |
| More 4 | 2.23 | 1.68, 2.97 | <0.001 |
| sex | | | |
| female | — | — | |
| male | 0.73 | 0.55, 0.97 | 0.029 |

ggforest(coxph_fit, data = colon_mod)

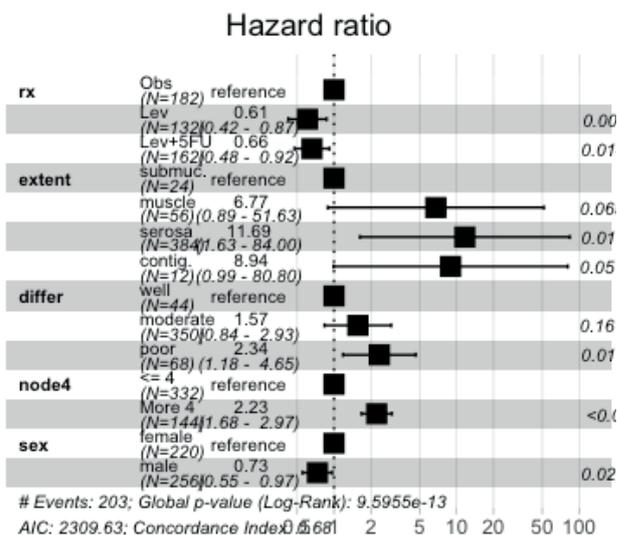


Рис.1 Коэффициенты регрессии Кокса

Изменение риска со временем может быть продемонстрировано графически

```
ggurvplot(survfit(Surv(time, status) ~ rx, data = colon_
mod),
palette = "grey",
fun = "cumhaz",
pval = TRUE
)
```

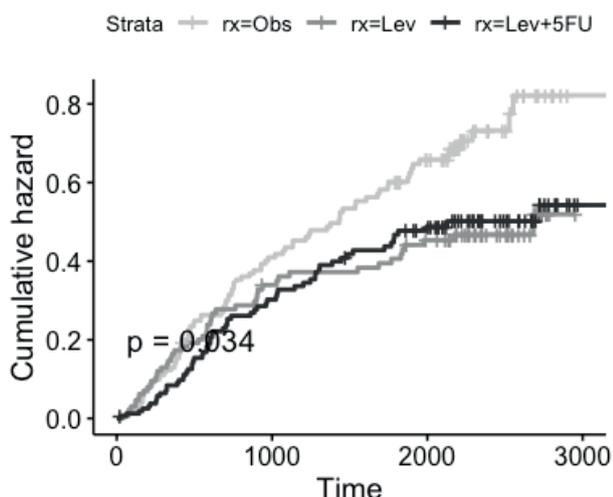


Рис.2 Кумулятивный риск

Модель пропорциональных рисков - это регрессионная модель, показатели которой можно посмотреть при использовании функции `summary` или `broom::glance` (таб. 2 и 3). Полученные результаты, представленные в таблицах и рисунках, позволяют оценить значения HR, показывающими во сколько раз изменится риск события при наличии этого фактора. Наиболее наглядно значение этих коэффициентов представлено на рис. 1. Коэффициенты, доверительные интервалы которых не пересекают вертикальную линию (со значением 1) прежде всего подлежат рассмотрению. Значимыми для отобранной группы пациентов факторами, снижающими риск события, являются адъювантная терапия (HR = 0,61 и 0,66), мужской пол участника (HR = 0,73); факторами, повышающими риск события, являются: распространение на серозную оболочку (HR = 11,7), недифференцированные опухоли (HR = 2,34), наличие более четырёх вовлечённых лимфатических узлов (HR = 2,23). Рис.2 указывает на лучший прогноз при применении адъювантной терапии.

Диагностика модели пропорциональных рисков

Прежде, чем признать результаты полученной модели валидными (достоверными), необходимо ответить на вопросы: является ли принятие пропорциональных рисков допустимым, имеются ли выбросы и влияющие наблюдения [24]. Часто используемыми являются методы оценки пригодности (goodness-of-fit) модели. Большинство диагностических процедур для модели Кокса основываются на её остатках.

Оценка пропорциональности рисков

Schoenfeld [25] предложил использовать тест для оценки регрессионной модели Кокса, основанный на изучении остатков в форме “Ожидаемые - Наблюдаемые”. В дальнейшем он дал им формальное определение [26]. Развитие темы получило продолжение

у Moreau, O’Quigley и Mesbah [27] и Grambsch & Therneau [28]. В R для оценки пропорциональности рисков регрессии Кокса используется функция `sox.zph` пакета `survival`, аргумент этой функции `transform` определяет, как может быть изменено время выживания перед выполнением теста. Возможные значения `km` (по умолчанию), `rank`, `identity`. Очень малые значения `p` в результатах теста указывают на зависящие от времени коэффициенты.

В листинге 3 использована модель Кокса-регрессии с такими ковариатами как способ лечения, дифференцированность опухоли и возраст. Возможна графическая оценка допущений пропорциональных рисков. В R для этого могут быть использованы функции `ggcoxzph` и `ggcoxdiagnostics` из пакета `survminer`. Первым аргументом в этих функциях является объект класса `sox.zph`. При использовании `ggcoxzph` будут выведены графики для каждой ковариаты, включенной в объект `sox.zph`, остатков Schoenfeld против трансформированного времени.

Используя функцию `ggcoxdiagnostics()` можно графически представить пригодность модели пропорциональных рисков. Важным является аргумент `type` этой функции, он может применять такие значения: «martingale», «deviance», «score», «schoenfeld», «dfbeta», «dfbetas», «scaledsch», «partial». На рис.5 отображена log-log кривая выживаемости. Выполнение кода из листинга 3 позволит оценить допущения пропорциональных рисков для выбранной модели.

Листинг 3

```
coxph_fit_m <- coxph(Surv(time, status) ~ rx + differ + age,
data = colon_mod)
```

```
gtsummary::tbl_regression(coxph_fit_m, exponentiate = TRUE)
```

| | N = 462 | HR | 95% CI | p-value |
|----------|---------|------|------------|---------|
| rx | | | | |
| Obs | | — | — | |
| Lev | | 0.68 | 0.47, 0.97 | 0.031 |
| Lev+5FU | | 0.69 | 0.50, 0.96 | 0.028 |
| differ | | | | |
| well | | — | — | |
| moderate | | 1.77 | 0.95, 3.30 | 0.070 |
| poor | | 2.98 | 1.51, 5.86 | 0.002 |
| age | | 0.98 | 0.96, 1.01 | 0.2 |

```
cox.zph(coxph_fit_m, transform = "rank")
```

```
chisq df p
rx 2.974 2 0.226
differ 11.019 2 0.004
age 0.668 1 0.414
GLOBAL 13.555 5 0.019
```

```
ggcoxdiagnostics(coxph_fit_m, type = "schoenfeld")
```

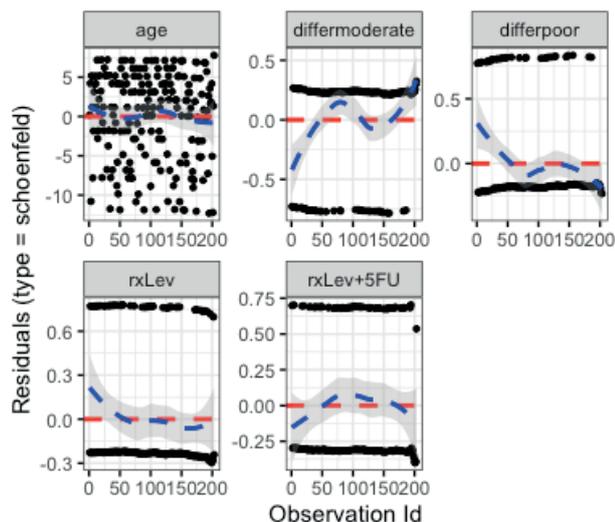


Рис.3 Оценка пропорциональности рисков - функция `ggcoxdiagnostics`

```

ggsurvplot(survfit(Surv(time, status) ~ rx, data = colon_
mod),
palette = "grey",
fun = "cloglog",
pval = TRUE
)
    
```

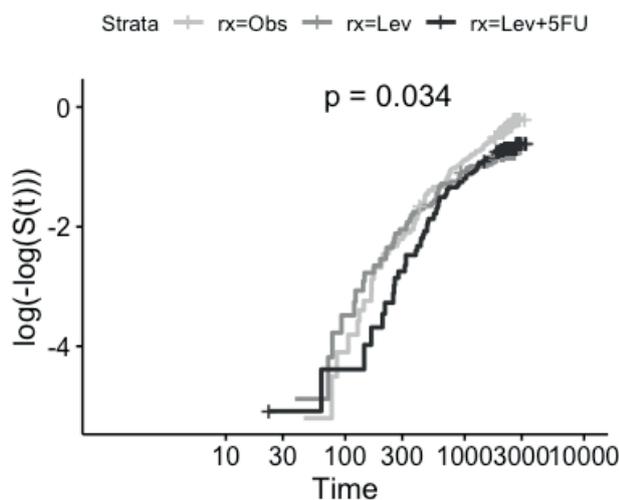


Рис.4 log-log кривая выживаемости при оценке метода лечения

```

ggsurvplot(survfit(Surv(time, status) ~ differ, data = colon_
mod),
palette = "grey",
fun = "cloglog",
pval = TRUE
)
    
```

Результаты оценки пропорциональности рисков с использованием функции `cox.zph` позволяют сказать, что риск, определяемый дифференцированностью опухоли, изменяется со временем ($p\text{-value} = 0,004$), риски, определяемые методом терапии и возрастом, не меняются. Об этом так же свидетельствуют графики на рис. 3. log-log кривые выживаемости на рис. 4 (оценивались методы терапии) представлены почти

параллельными линиями в отличие от кривых на рис. 5 (оценивалась дифференцированность опухоли).

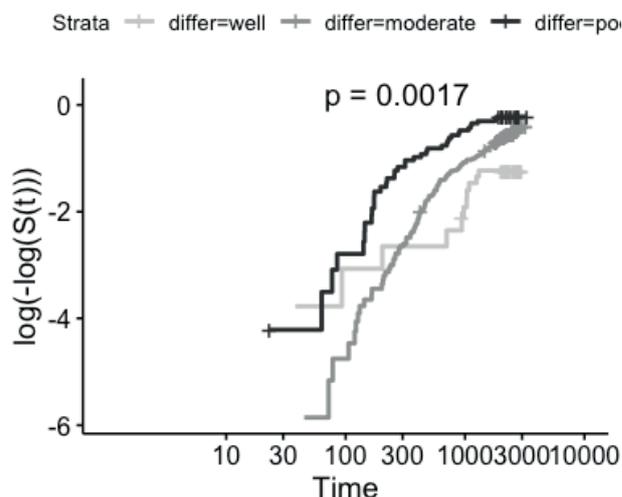


Рис.5 log-log кривая выживаемости при оценке дифференцированности

Влияющие случаи

Графически влияющие случаи можно представить (листинг 4), используя функцию `ggcoxdiagnostics`, указав в качестве аргумента `type = «dfbeta»`

Листинг 3

```

ggcoxdiagnostics(coxph_fit_m, type = "dfbeta", point.size
= 0, hline.col = "black", sline.col = "black") + geom_bar(stat =
"identity")
    
```

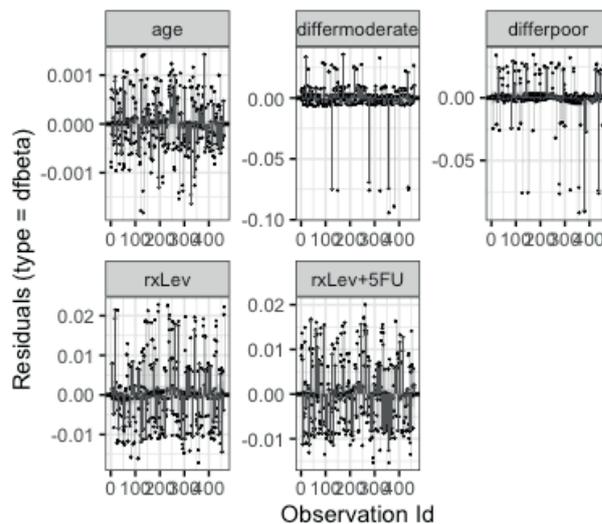


Рис.6 Оценка влияющих случаев

Выбросы (выскакивающие случаи)

Визуальная оценка выбросов возможна при использовании функции `ggcoxdiagnostics` с использованием значения аргумента `type martingale` и/или `deviance` (листинг 5).

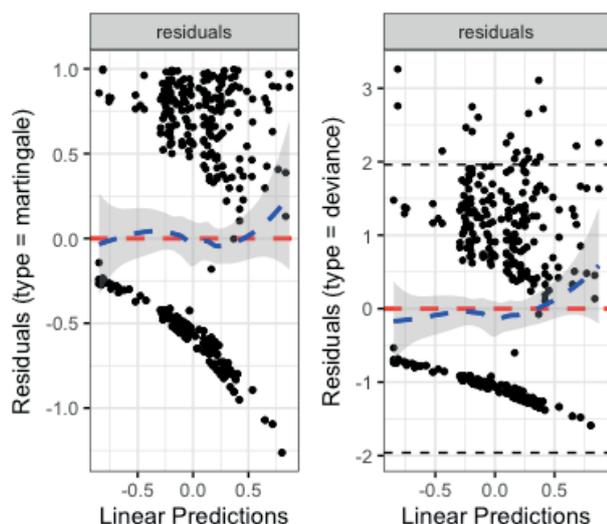


Рис.7 Выявление выбросов

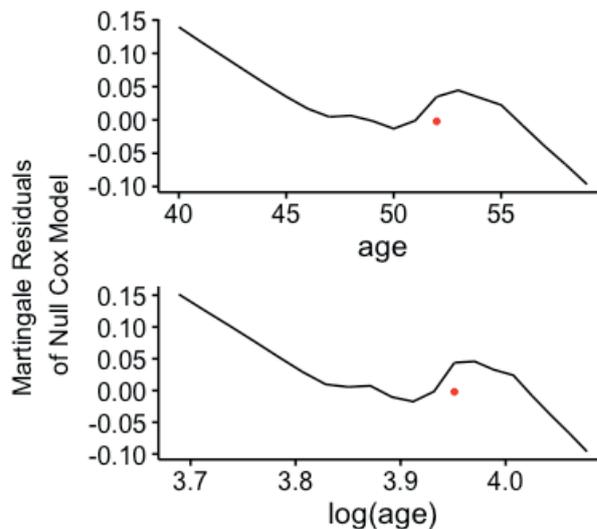


Рис.8 Оценка линейности

Листинг 5

```
gr1 <- ggcoxdiagnostics(coxph_fit_m, type = "martingale")
gr2 <- ggcoxdiagnostics(coxph_fit_m, type = "deviance") +
  geom_hline(yintercept = c(-1.96, 1.96), lty = 2)
cowplot::plot_grid(gr1, gr2)
```

К выбросам можно отнести случаи, выходящие за пределы пунктирной линии.

Оценка линейности

Может выполняться в отношении непрерывных ковариат и Martingale отстатков (листинг 6).

Листинг 6

```
fit_age <- coxph(Surv(time, status) ~ age + log(age), data =
  colon_mod)
```

```
ggcoxfunctional(fit_age, data = colon_mod)
```

На рис.8 выполнена оценка линейности в отношении возраста и натурального логарифма возраста.

Работа с R

Программная среда R является свободно распространяемым кросс-платформенным программным средством, используемым для статистических вычислений и визуализации данных. Дистрибутивы R доступны на сайтах The Comprehensive R Archive Network, <https://cran.r-project.org>, Microsoft R Application Network, <https://mran.microsoft.com/download>. Удобным IDE (integrated development environment, интегрированная среда разработчика) для программы R является программа RStudio, свободно распространяемый дистрибутив может быть загружен на сайте RStudio IDE, <https://www.rstudio.com/products/rstudio/>. В наших более ранних публикациях мы уже касались вопросов применения программной среды R в биомедицинских исследованиях. Используемый в работе файл с набором данных и скрипт с кодом доступны на сайте https://github.com/valegoshin/Paper_Scripts.

Список литературы:

1. Cox DR. Models and Life-Tables Regression. Journal of the Royal Statistical Society. Series B (Methodological) 1972;34(2):187–220.
2. Vijver MJ, He YD, Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N Engl J Med. 2002;347(25):1999–2009.
3. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. 2003;89(2):232–38.
4. Clark TG, Bradburn MJ, Love SB, Altman DG. "Survival analysis part IV: Further concepts and methods in survival analysis." British Journal of Cancer 2003;89(5):781–86.
5. Bradburn MJ, Clark TJ, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis: An introduction to concepts and methods. British Journal of Cancer. 2003;89(3):431–36.
6. Moyé L. Statistical Methods for Cardiovascular Researchers. Circ Res. 2016;118(3):439–53.
7. Austin PC. "A tutorial on multilevel survival analysis: Methods, models and applications." International Statistical Review. 2017;85(2):185–203.
8. Li H. 2017. Survival Analysis for a Breast Cancer Data Set. Advances in Breast Cancer Research. 2017;6(1):1–15.
9. Fox J, Weisberg S. Cox Proportional-Hazards Regression for Survival Data in R. Most. 2011;2008:1–18.
10. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer, 2000.
11. Kassambara A, Kosinski M, Biecek P. Survminer: Drawing Survival Curves Using 'Ggplot2'. <https://CRAN.R-project.org/package=survminer>.
12. Robinson D, Hayes A. Broom: Convert Statistical Analysis Objects into Tidy Tibbles. <https://CRAN.R-project.org/package=broom>.
13. Harrison E, Drake T, Ots R. Finalfit: Quickly Create Elegant Regression Results Tables and Plots When Modelling. <https://CRAN.R-project.org/package=finalfit>.
14. Sjoberg DD, Hannum M, Whiting K, Zabor EC. Gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. <https://CRAN.R-project.org/package=gtsummary>.
15. R Core Team. R: A Language and Environment for Statistical

- Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
16. Wickham H, François R, Henry L, Müller K. Dplyr: A Grammar of Data Manipulation. <https://CRAN.R-project.org/package=dplyr>.
 17. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
 18. Wickham H, Henry L. Tidy: Tidy Messy Data. <https://CRAN.R-project.org/package=tidy>.
 19. Wickham H, Averick M, Bryan J, Chang W, D'Agostino , François R, Golemund G. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4(43):1686.
 20. Heinzen E, Sinnwell J, Atkinson E, Gunderson T, Dougherty G. Arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries. <https://CRAN.R-project.org/package=arsenal>.
 21. Xie Y. Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. 2015.
 22. Laurie JA, Moertel CG, Fleming TR, Wieand HS, Leigh JE, Rubin J, McCormack GW, Gerstner JB, Krook JE, Malliard J. Surgical Adjuvant Therapy of Large-Bowel Carcinoma: An Evaluation of Levamisole and the Combination of Levamisole and Fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *J Clin Oncol*. 1989;7(10):1447–56.
 23. Moertel CG, Fleming TR, Macdonald TJ, Haller DG, Laurie JA, Goodman PJ, Ungerleider JS, Emerson WA, Tormey DC, Glick JH. Levamisole and Fluorouracil for Adjuvant Therapy of Resected Colon Carcinoma. *N Engl J Med*. 1990;322(6):352–58.
 24. Xuea Y, Schifano ED. Diagnostics for the Cox Model. *Communications for Statistical Applications and Methods*. 2017;24(6):583–604.
 25. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*. 1980;67(1):145–53.
 26. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika*. 1982;69(1):239–41.
 27. Moreau T, Quigley JO', Mesbah M. "A Global Goodness-of-Fit Statistic for the Proportional Hazards Model." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 1985;34(3):212–18.
 28. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515–26.