Package: sta310r (via r-universe)

November 11, 2026

Title Data Sets and Computer Exercises for the STA310 Survival Analysis Course

Version 0.1.0

Description The data sets that accompany Collett (2023), ``Modelling Survival Data in Medical Research'' (4th edition), documented for the STA310 course at the University of Gothenburg, together with the computer exercises used in the course. The package focuses on the data sets, their documentation and the exercise vignettes.

License MIT + file LICENSE

URL https://github.com/eribul/sta310r, https://eribul.github.io/sta310r/

Encoding UTF-8

Roxygen list(markdown = TRUE)

Depends R (>= 4.1.0)

Suggests broom, broom.helpers, cmprsk, coxme, dplyr, eha, flexsurv, ggplot2, ggsurvfit, glmnet, gt, gtsummary, icenReg, knitr, mstate, pec, penalized, readxl, rmarkdown, rms, riskRegression, survival, survminer, testthat (>= 3.1.5), tibble, tidycmprsk, tidyr, tidyverse

VignetteBuilder knitr

LazyData true

Config/testthat/edition 3

Config/Needs/website pkgdown

Config/roxygen2/version 8.1.1

Repository https://eribul.r-universe.dev

Date/Publication 2026-11-11 15:48:01 UTC

RemoteUrl https://github.com/eribul/sta310r

RemoteRef main

RemoteSha 15ec297728952089064d7d2137099666dc15cff0

Contents

aortic . . . 2
bladder . . . 4
bmt . . . 5
bmt_leukaemia . . . 6
bmt_outcome . . . 8
breast . . . 9
breast_rfs . . . 11
cgd . . . 12
cirrhosis . . . 14
cirrhosis_cp . . . 15
cirrhosis_lbr . . . 16
dataset_by_exercise . . . 17
dialysis . . . 19
ducks . . . 20
gastric . . . 21
help . . . 23
hepatitis . . . 24
hyper . . . 25
iud . . . 27
kidney . . . 28
kidney_centres . . . 30
liver . . . 31
liver_wait . . . 33
lung_transplant . . . 34
melanoma . . . 35
mice . . . 37
myeloma . . . 38
numerical_illustration . . . 40
ovarian . . . 41
prostate . . . 43
pulmonary . . . 44
rats . . . 45
scottish_mbc_cohort . . . 46
tamoxifen . . . 48
ulcer . . . 49

Index **52**

`aortic` *Survival following aortic valve replacement*

Description

Longitudinal measurements of the left ventricular mass index (LVMI) and survival times for 256 patients who received an aortic valve replacement (Collett, 2023, Example 8.5).

Usage

```
aortic
```

Format

A data frame with 988 rows and 11 variables:

id Patient number (1-256).

futime Total follow-up time from the date of surgery, in years.

status Event indicator: 1 = death, 0 = right-censored.

time Time of the LVMI measurement after surgery, in years.

lvmi Standardised left ventricular mass index.

age Age of the patient in years.

sex Sex: 0 = male, 1 = female.

redo Previous cardiac surgery: 0 = no, 1 = yes.

emerg Operative urgency: 0 = elective, 1 = urgent or emergency.

dm Preoperative diabetes mellitus: 0 = no, 1 = yes.

type Type of valve: 1 = human tissue, 2 = porcine tissue.

Details

The file is in long format: one row per LVMI measurement, and the baseline and survival variables (`futime`, `status`, `age`, ...) are repeated on each row for the patient. There are 988 measurement rows for 256 patients; 54 of the patients died (`status` is 1, repeated on each of their rows). Patients who had a second procedure are censored at that time so that they do not appear twice in the survival data.

The data are used for joint modelling of a longitudinal and a survival outcome (Examples 8.5-8.9). The full data set is available as `heart.valve` in the joineR package.

Aliases

The data set is available under the aliases `d8.5`, `d8.6`, `d8.7`, `d8.8`, `d8.9`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 8.5 (Table 8.10). Data from Lim et al. (2008).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 8.5-8.9: LVMI as a time-dependent covariate, built with tmerge()
base <- aortic[!duplicated(aortic$id), c("id", "futime", "status", "age", "sex")]
long <- aortic[, c("id", "time", "lvmi")]
cp <- survival::tmerge(base, base, id = id, endpt = event(futime, status))
```

```
cp <- survival::tmerge(cp, long, id = id, lvmi = tdc(time, lvmi))
survival::coxph(
  survival::Surv(tstart, tstop, endpt) ~ lvmi + age + sex,
  data = cp
)
```

`bladder` *Recurrence of bladder cancer*

Description

Times to the first recurrence of a bladder tumour, in months, for 86 patients in a placebo-controlled trial of thiotepa (Collett, 2023, Appendix B.2).

Usage

```
bladder
```

Format

A data frame with 86 rows and 6 variables:

patient Patient number (1-86).

time Time to first recurrence, in months.

status Event indicator: 1 = recurrence, 0 = right-censored.

treat Treatment: 1 = placebo, 2 = thiotepa.

init Initial number of tumours.

size Diameter of the largest initial tumour, in cm.

Details

The trial was conducted by the Veterans Administration Cooperative Urological Research Group. Superficial tumours were removed transurethrally before randomisation. The original data set in Andrews and Herzberg (1985) contains the times to up to nine recurrences; the file here keeps only the time to the first recurrence (47 recurrences).

The appendix suggests a Cox model for `init` and `size`, an assessment of the treatment effect after adjustment, a test of interaction, model-checking diagnostics, a test of proportional hazards for treatment using a time-dependent variable, and a comparison with a Weibull proportional hazards model.

Aliases

This data set is not used in a numbered example, so it has no `dX.Y` alias. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Appendix B.2 (Table B.2). Data from Andrews and Herzberg (1985).

Examples

```
# Collett (2023), Appendix B.2 (no numbered example) ------------------
# Chapter 17 style: Cox model for the time to the first recurrence
survival::coxph(
  survival::Surv(time, status) ~ init + size + factor(treat),
  data = bladder
)

# Modern tools -------------------------------------------------------
if (requireNamespace("broom", quietly = TRUE)) {
  broom::tidy(
    survival::coxph(
      survival::Surv(time, status) ~ init + size + factor(treat),
      data = bladder
    ),
    exponentiate = TRUE, conf.int = TRUE
  )
}
```

---

bmt *Bone marrow transplantation in leukaemia patients*

Description

Survival times, in days, of 37 leukaemia patients who received a bone marrow transplant, together with donor and recipient variables (Collett, 2023, Appendix B.4).

Usage

```
bmt
```

Format

A data frame with 37 rows and 9 variables:

patient Patient number (1-37).

time Survival time in days.

status Event indicator: 1 = death, 0 = alive.

rage Age of the recipient in years.

dage Age of the donor in years.

type Type of leukaemia: 1 = AML, 2 = ALL, 3 = CML.

preg Whether the donor had previously been pregnant: 0 = no, 1 = yes.

index Index of mixed epidermal lymphocyte reactions.

gvhd Graft-versus-host disease: 0 = no, 1 = yes.

Details

The 37 patients were in complete remission from AML or ALL, or in the chronic phase of CML, and received a non-depleted allogeneic bone marrow transplant. There are 17 deaths. The data were also given in Altman (1991).

The appendix suggests fitting a Weibull accelerated failure time model, estimating and plotting the baseline survivor function, comparing with log-logistic and lognormal models and with a Cox model, and checking the adequacy of the preferred model. See bmt_leukaemia (a different study of 23 patients) and bmt_outcome (EBMT registry data).

Aliases

This data set is not used in a numbered example, so it has no `dX.Y` alias. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Appendix B.4 (Table B.4). Data from Bagot et al. (1988) and Altman (1991).

Examples

```
# Collett (2023), Appendix B.4 (no numbered example) ------------------
# Chapter 17 style: Weibull accelerated failure time model
survival::survreg(
  survival::Surv(time, status) ~ rage + dage + factor(type) + gvhd,
  data = bmt, dist = "weibull"
)

# The corresponding Cox model
survival::coxph(
  survival::Surv(time, status) ~ rage + dage + factor(type) + gvhd,
  data = bmt
)

# Modern tools ---------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(time, status) ~ rage + dage + factor(type) + gvhd,
    data = bmt, dist = "weibull"
  )
}
```

`bmt_leukaemia` *Bone marrow transplantation in the treatment of leukaemia*

Description

Survival times, in days, of 23 leukaemia patients who received a bone marrow transplant, with donor and recipient characteristics and an indicator of platelet recovery (Collett, 2023, Example 8.2).

Usage

```
bmt_leukaemia
```

Format

A data frame with 23 rows and 8 variables:

patient Patient number (1-23).

time Survival time in days from transplantation.

status Event indicator: 1 = death, 0 = alive at the end of follow-up.

group Disease group: 1 = acute lymphoblastic leukaemia (ALL), 2 = low-risk acute myeloid leukaemia (AML), 3 = high-risk AML.

page Age of the patient in years.

dage Age of the bone marrow donor in years.

precovery Platelet recovery: 0 = no, 1 = platelet count returned to a normal level of 40 x 10^9 per litre.

ptime Time in days until the platelet count returned to normal (`NA` when `precovery` is 0).

Details

This is the subset of the study from St Vincent's Hospital, Sydney, used by Klein and Moeschberger (2005). There are 12 deaths. Patients 7 and 21 died before their platelet count returned to normal, and so `ptime` is missing for them; these missing values are stored as `NA` in the package.

The data illustrate the use of a time-dependent covariate: the indicator of platelet recovery changes from 0 to 1 at `ptime`, so the survival data are expanded into the counting process format (Example 8.2, and Example 8.1 introduces the format itself).

Aliases

The data set is available under the alias `d8.2`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 8.2 (Table 8.2). Data from Klein and Moeschberger (2005).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 8.2: platelet recovery as a time-dependent covariate
base <- bmt_leukaemia[, c("patient", "time", "status", "group", "page", "dage")]
recovery <- subset(bmt_leukaemia, precovery == 1)[, c("patient", "ptime")]
cp <- survival::tmerge(base, base, id = patient, endpt = event(time, status))
cp <- survival::tmerge(cp, recovery, id = patient, preg = tdc(ptime))
survival::coxph(
  survival::Surv(tstart, tstop, endpt) ~ preg + factor(group) + page + dage,
  data = cp
)
```

`bmt_outcome` *Patient outcome following bone marrow transplantation*

Description

Data from 2204 leukaemia patients who received a bone marrow transplant, with times to platelet recovery and to relapse or death (Collett, 2023, Example 13.4).

Usage

```
bmt_outcome
```

Format

A data frame with 2204 rows and 9 variables:

id Patient identifier.

leukaemia Disease type: `"AML"`, `"ALL"` or `"CML"`.

age Age group: `"<=20"`, `"21-40"` or `">40"`.

match Donor-recipient gender match: 0 = no, 1 = yes.

tcell T-cell depletion: 0 = no, 1 = yes.

ptime Time from transplantation to platelet recovery, in days.

pcens Platelet recovery indicator: 0 = censored, 1 = recovery.

rdtime Time from transplantation to relapse or death, in days.

rdcens Relapse-or-death indicator: 0 = censored, 1 = event.

Details

The data come from the European Group for Blood and Marrow Transplantation (EBMT) Registry, and are the version used by Putter, Fiocco and Geskus (2007). Following the transplant a patient may either recover platelet function and later relapse or die, or relapse or die before platelet recovery. The data therefore describe a three-state model with transition rates for transplantation to platelet recovery, transplantation to relapse or death, and platelet recovery to relapse or death.

Of the 2204 patients, 1169 experienced platelet recovery and 458 relapsed or died before their platelet count had returned to normal; a further 383 patients relapsed or died after platelet recovery. Note that the `ptime` of a patient who relapsed before recovering platelets is censored at the time of relapse when the first transition is modelled. To fit the model the data are expanded into one row per possible transition per patient (Example 13.4).

Aliases

The data set is available under the alias `d13.4`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 13.4 (Table 13.9). Data from the EBMT Registry, see Putter, Fiocco and Geskus (2007).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 13.4: the two endpoints of the three-state model
# 1 -> 2: platelet recovery
survival::coxph(
  survival::Surv(ptime, pcens) ~ leukaemia + age + match + tcell,
  data = bmt_outcome
)
# 1 -> 3: relapse or death before recovery
survival::coxph(
  survival::Surv(rdtime, rdcens) ~ leukaemia + age + match + tcell,
  data = bmt_outcome
)

# Modern tools ------------------------------------------------------
# Transition-specific hazards are fitted by stratifying on the transition
if (requireNamespace("mstate", quietly = TRUE)) {
  tmat <- mstate::transMat(
    list(c(2, 3), 3, c()),
    names = c("transplant", "recovery", "relapse")
  )
  tmat
}
```

---

breast *Prognosis for women with breast cancer*

Description

Survival times, in months, of 45 women who received a mastectomy for a Grade II, III or IV breast tumour, classified by whether the tumour stained positively or negatively with the lectin HPA (Collett, 2023, Example 1.2).

Usage

```
breast
```

Format

A data frame with 45 rows and 3 variables:

stain Result of HPA staining: 1 = negative, 2 = positive. Positive staining indicates a tumour with the potential to metastasise.

time Survival time in months from surgery.

status Event indicator: 1 = death from breast cancer, 0 = right-censored.

Details

The women were treated between January 1969 and December 1971 and the study was closed in July 1987. Women still alive at that date, as well as women who died from causes other than breast cancer, contribute censored times. There are 13 women in the negative-staining group and 32 in the positive-staining group, with 26 deaths in total.

This is the standard small example for comparing two survival curves: the Kaplan-Meier estimate by group (Examples 2.11-2.14), the log-rank and Wilcoxon tests (Section 2.4) and a Cox model with `stain` as the single covariate (Examples 3.1 and 3.3).

Aliases

The data set is available under the aliases `d1.2`, `d2.11`, `d2.12`, `d2.13`, `d2.14`, `d3.1`, `d3.3`, `d5.5`, `d5.6`, `d5.7`, `d5.8`, `d5.13`, `d5.14`, `d5.17`, `d10.1`, `d10.2`, `d16.4`, `d16.5`, `d16.6`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 1.2 (Table 1.2). Data from Leathem and Brooks (1987).

Examples

```
# Collett (2023), Chapter 17 style ------------------------------------
# Example 1.2: survival by HPA staining (1 = negative, 2 = positive)
survival::survfit(survival::Surv(time, status) ~ stain, data = breast)

# Examples 2.11-2.14: log-rank and Peto-Peto tests
survival::survdiff(survival::Surv(time, status) ~ stain, data = breast)
survival::survdiff(survival::Surv(time, status) ~ stain, data = breast, rho = 1)

# Examples 3.1 and 3.3: Cox model with the staining indicator
fit_breast <- survival::coxph(survival::Surv(time, status) ~ stain, data = breast)
summary(fit_breast)

# Examples 5.5-5.8, 5.13, 5.14 and 5.17: Weibull models
survival::survreg(survival::Surv(time, status) ~ stain, data = breast, dist = "weibull")

# Modern tools --------------------------------------------------------
if (requireNamespace("ggsurvfit", quietly = TRUE)) {
  ggsurvfit::survfit2(survival::Surv(time, status) ~ stain, data = breast) |>
    ggsurvfit::ggsurvfit() |>
    ggsurvfit::add_risktable()
}
if (requireNamespace("broom", quietly = TRUE)) {
  broom::tidy(fit_breast, exponentiate = TRUE, conf.int = TRUE)
}
if (requireNamespace("flexsurv", quietly = TRUE)) {
 flexsurv::flexsurvreg(survival::Surv(time, status) ~ stain, data = breast, dist = "weibull")
}
```

`breast_rfs` *Recurrence-free survival in breast cancer patients*

Description

Recurrence-free survival times, in days, of 686 patients with node-positive breast cancer from the German Breast Cancer Study Group (Collett, 2023, Example 6.5).

Usage

```
breast_rfs
```

Format

A data frame with 686 rows and 11 variables:

id Patient number.

treat Hormonal treatment: 0 = no tamoxifen, 1 = tamoxifen.

age Age of the patient in years.

men Menopausal status: 1 = premenopausal, 2 = postmenopausal.

size Tumour size in mm.

grade Tumour grade (1, 2 or 3).

nodes Number of positive lymph nodes.

prog Progesterone receptor status (femtomoles).

oest Oestrogen receptor status (femtomoles).

time Recurrence-free survival time in days.

status Event indicator: 1 = relapse or death, 0 = right-censored.

Details

The patients had primary, histologically proven, non-metastatic node-positive breast cancer treated with mastectomy; the study compared three with six cycles of chemotherapy and investigated additional hormonal treatment. Recurrence-free survival is the time from entry until a recurrence of the cancer or death, and there are 299 events.

The example introduces flexible parametric (Royston-Parmar) models with a treatment effect and several prognostic variables (Examples 6.5 and 6.6).

Aliases

The data set is available under the aliases `d6.5`, `d6.6`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 6.5 (Table 6.1). Data from Schumacher et al. (1994) and Sauerbrei and Royston (1999).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 6.5 and 6.6: log-logistic (proportional odds) model
survival::survreg(
  survival::Surv(time, status) ~ treat + age + men + size + grade +
    nodes + prog + oest,
  data = breast_rfs, dist = "loglogistic"
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvspline(
    survival::Surv(time, status) ~ treat + age + men + size + grade +
      nodes + prog + oest,
    data = breast_rfs, k = 1
  )
}
```

---

cgd *Chronic granulomatous disease*

Description

Time to the first serious infection for 128 patients with chronic granulomatous disease (CGD) in a multicentre trial (Collett, 2023, Appendix B.5).

Usage

```
cgd
```

Format

A data frame with 128 rows and 12 variables:

patient Patient number (1-128).

time Time to the first serious infection, in days.

status Event indicator: 1 = infection, 0 = censored.

centre Trial centre (1-14; centre 8 does not occur in these data).

treat Treatment: 0 = placebo, 1 = interferon.

age Age in years.

sex Sex: 1 = male, 2 = female.

height Height in cm.

weight Weight in kg.

pattern Pattern of inheritance: 1 = X-linked, 2 = autosomal recessive.

cort Use of corticosteroids at trial entry: 1 = used, 2 = not used.

anti Use of antibiotics at trial entry: 1 = used, 2 = not used.

Details

CGD is a group of inherited disorders that leave patients vulnerable to recurrent infections. The trial compared the antiviral drug interferon with placebo, and the endpoint is the time to the first serious infection after randomisation; there are 44 infections. The data were described by Therneau and Grambsch (2000); this is the 128-patient version used in the book.

The appendix suggests a wide range of analyses: model selection and model checking, consistency of the treatment effect across centres, a constructed interval-censored version of the data, shared frailty models with centre as a random effect, risk-adjusted centre-specific infection rates, and an inverse probability of censoring weighting analysis for dependent censoring. The infection data are also available, in a larger form, as `cgd` in the survival package.

Aliases

This data set is not used in a numbered example, so it has no `dX.Y` alias. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Appendix B.5. Data described by Therneau and Grambsch (2000); the trial was carried out by Genentech, Inc.

Examples

```
# Collett (2023), Appendix B.5 (no numbered example) ------------------
# Chapter 17 style: Cox model for the first serious infection
survival::coxph(
  survival::Surv(time, status) ~ factor(treat) + age + sex + height + weight,
  data = cgd
)

# A shared frailty for trial centre
survival::coxph(
  survival::Surv(time, status) ~ factor(treat) + age + frailty(centre),
  data = cgd
)

# Modern tools -------------------------------------------------------
if (requireNamespace("coxme", quietly = TRUE)) {
  coxme::coxme(
    survival::Surv(time, status) ~ factor(treat) + age + (1 | centre),
    data = cgd
  )
}
```

`cirrhosis` *Data from a cirrhosis study (baseline)*

Description

Survival times, in days, of 12 patients in an artificial study of the treatment of cirrhosis of the liver, with baseline age and bilirubin level (Collett, 2023, Example 8.4).

Usage

```
cirrhosis
```

Format

A data frame with 12 rows and 6 variables:

patient Patient number (1-12).

time Survival time in days.

status Event indicator: 1 = death, 0 = right-censored.

age Age of the patient in years.

treat Treatment: 0 = placebo, 1 = Liverol.

lbr Logarithm of the baseline bilirubin level (in micromol/l).

Details

The data are artificial but are used to mimic a real liver study. Six patients are allocated to Liverol and six to placebo. The baseline log(bilirubin) value is analysed first, and the treatment effect is estimated after adjustment for age and `lbr` (Example 8.4).

The same patients appear in three files in this package. `cirrhosis` holds the baseline values (one row per patient), `cirrhosis_lbr` holds the repeated log(bilirubin) measurements in long format (one row per visit) and `cirrhosis_cp` holds the survival data together with the time-dependent log(bilirubin) value in counting process format (one row per interval). See cirrhosis_lbr and cirrhosis_cp.

Aliases

The data set is available under the alias `d8.4`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 8.4 (Table 8.4). The hypothetical study is loosely based on the primary biliary cirrhosis data of Therneau and Grambsch (2000).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 8.4: the baseline model
survival::coxph(
  survival::Surv(time, status) ~ age + lbr + treat,
  data = cirrhosis
)

# Example 8.4: log(bilirubin) as a time-dependent covariate
# (two patients have a zero-length interval, hence the filter)
cp <- subset(cirrhosis_cp, stop > start)
survival::coxph(survival::Surv(start, stop, status) ~ lbrt + age, data = cp)
```

---

cirrhosis_cp *Data from a cirrhosis study (counting process format)*

Description

Survival data for the 12 patients in the cirrhosis study, with the log(bilirubin) level as a time-dependent covariate, in counting process format (Collett, 2023, Example 8.4).

Usage

```
cirrhosis_cp
```

Format

A data frame with 54 rows and 7 variables:

patient Patient number (1-12).

start Start of the interval, in days.

stop End of the interval, in days.

status Event indicator for the interval: 1 = death at `stop`, 0 = otherwise.

treat Treatment: 0 = placebo, 1 = Liverol.

age Age of the patient in years.

lbrt Time-dependent log(bilirubin) value in effect during the interval, i.e. the most recent measurement after `start` and up to `stop`.

Details

Each row is a time interval during which all covariates are constant, so a patient with many visits contributes several rows. Patient 1, for example, has log(bilirubin) 3.2 for t <= 47, 3.8 for 47 < t <= 184, 4.9 for 184 < t <= 251 and 5.0 for 251 < t <= 281. The data are used to fit a Cox model with a time-dependent covariate (`start`, `stop` and `status` form the response).

The baseline values are in cirrhosis and the repeated measurements in cirrhosis_lbr.

Aliases

Used in Example 8.4, but the alias `d8.4` is the baseline cirrhosis; this counting-process form has no alias of its own. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 8.4 (Tables 8.4 and 8.5).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 8.4: counting process form of the cirrhosis study
# (patient 5 and patient 8 have a zero-length interval, hence the filter)
cp <- subset(cirrhosis_cp, stop > start)
survival::coxph(survival::Surv(start, stop, status) ~ lbrt + age, data = cp)

# Without the time-dependent covariate, for comparison with the baseline data
survival::coxph(survival::Surv(start, stop, status) ~ treat + age, data = cp)
```

---

cirrhosis_lbr *Data from a cirrhosis study (log-bilirubin data)*

Description

Repeated measurements of log(bilirubin) for the 12 patients in the cirrhosis study, in long format (Collett, 2023, Example 8.4).

Usage

```
cirrhosis_lbr
```

Format

A data frame with 42 rows and 3 variables:

patient Patient number (1-12).

time Time of the measurement, in days from the start of treatment.

lbr Logarithm of the bilirubin level (in micromol/l) measured at `time`.

Details

Patients return to the clinic three, six and twelve months after the start of treatment and yearly thereafter. Values for a given patient form a step function that is constant between two adjacent visits, with the new value taking effect immediately after the reading. This file is the input to the counting process version of the data, cirrhosis_cp; the baseline values are in cirrhosis.

Aliases

Used in Example 8.4, but the alias `d8.4` is the baseline cirrhosis; this long-format form has no alias of its own. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 8.4 (Table 8.5).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 8.4: the repeated log(bilirubin) measurements in long format
head(cirrhosis_lbr)

# These are the measurements from which the time-dependent covariate lbrt
# in cirrhosis_cp is built: the value at each visit applies until the next.
merge(
  cirrhosis[c("patient", "time", "status")],
  cirrhosis_lbr,
  by = "patient", suffixes = c("_surv", "_visit")
) |> head()
```

`dataset_by_exercise` *Data sets by Collett (2023) example*

Description

An index of the data sets in the package, one row for each combination of a data set and a worked example in Collett (2023) that uses it. Each such combination also defines an alias `dX.Y`, where `X.Y` is the number of the example, and the alias is a copy of the data set. A data set that is used in several examples therefore has several aliases.

Usage

```
dataset_by_exercise
```

Format

A data frame with one row per data set and example, with columns:

dataset Name of the data set in the package.

dataset_title Descriptive title of the data set, as on its help page.

alias Alias of the data set for this example, `dX.Y`, or `NA` for data sets that are not used in a numbered example.

example Number of the example in Collett (2023), such as `"3.7"`, or `NA`.

chapter Number of the chapter that contains the example, or `NA`.

chapter_title Title of that chapter, or `NA`.

example_title Title of the example as it appears in the book, or `NA`.

n_rows Number of rows in the data set.

n_variables Number of columns in the data set.

note A note for data sets that are not used in a numbered example, for example that they come from an appendix of the book or from the course material.

Details

The index was compiled from the `Example X.Y` headings of the 4th edition. The aliases are the data sets themselves, saved under the name `dX.Y`, so `d3.7` is the data used in Example 3.7 (the `myeloma` data) and `d8.4` is the baseline cirrhosis data. Use dataset_by_exercise to look up which data set an example uses, and `?d3.7` to open the help page of that data set.

Four data sets from the appendices of the book (`bladder`, `ducks`, `bmt` and `cgd`) and the synthetic `scottish_mbc_cohort` are not used in a numbered example. They have no alias and their `example` is `NA`; the `note` column says where they come from. The cirrhosis study is the only example with more than one data set in the package: `cirrhosis` is the baseline form and has the alias `d8.4`, while `cirrhosis_lbr` and `cirrhosis_cp` are the long and counting-process forms used in the same example and have no alias of their own.

The table is generated by `data-raw/dataset_by_exercise.R`; re-run that script after adding a data set or changing the example mapping.

Source

The worked examples in Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC.

See Also

iud, breast, myeloma and the other data sets; the package vignette `vignette("datasets-by-exercise")` presents the same index.

Examples

```
head(dataset_by_exercise)
subset(dataset_by_exercise, dataset == "myeloma")
# The alias is a copy of the data set
identical(d3.7, myeloma)
```

`dialysis` *Infection in patients on dialysis*

Description

Times from insertion of a catheter to its removal following infection, for 13 kidney patients (Collett, 2023, Example 4.1).

Usage

```
dialysis
```

Format

A data frame with 13 rows and 5 variables:

patient Patient number (1-13).

time Time from insertion of the catheter to its removal, in days.

status Event indicator: 1 = catheter removed because of infection, 0 = removed for another reason.

age Age of the patient in years.

sex Sex: 1 = male, 2 = female.

Details

The data relate to the 13 patients with kidney disease coded as type 3 by McGilchrist and Aisbett (1991). In the original study each patient could experience an infection of both the first and the second catheter, but only the time to removal of the first catheter is given here. There is a single censored observation (patient 6).

With 13 observations and 12 events the data set is used to demonstrate model-checking diagnostics for the Cox model: Cox-Snell, martingale and deviance residuals and influence (Examples 4.1-4.4), a non-linear effect of age and a time-dependent effect of sex (Examples 4.10-4.12).

Aliases

The data set is available under the aliases `d4.1`, `d4.2`, `d4.3`, `d4.4`, `d4.6`, `d4.7`, `d4.10`, `d4.11`, `d4.12`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 4.1 (Table 4.1). Data from McGilchrist and Aisbett (1991).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 4.1-4.4: Cox model and the model-checking residuals
fit_dialysis <- survival::coxph(
  survival::Surv(time, status) ~ age + sex,
  data = dialysis
)
summary(fit_dialysis)
head(stats::residuals(fit_dialysis, type = "martingale"))
head(stats::residuals(fit_dialysis, type = "dfbeta"))

# Examples 4.10 and 4.11: a non-linear effect of age
survival::coxph(
  survival::Surv(time, status) ~ age + I(age^2) + sex,
  data = dialysis
)

# Example 4.12: a time-varying coefficient for sex
survival::coxph(
  survival::Surv(time, status) ~ age + sex + tt(sex),
  data = dialysis,
  tt = function(x, t, ...) x * t
)

# Modern tools ------------------------------------------------------
if (requireNamespace("rms", quietly = TRUE)) {
  rms::cph(survival::Surv(time, status) ~ age + sex, data = dialysis, x = TRUE, y = TRUE)
}
```

---

ducks *Survival of black ducks*

Description

Survival times, in days, of 50 female black ducks captured and fitted with radios (Collett, 2023, Appendix B.3).

Usage

```
ducks
```

Format

A data frame with 50 rows and 6 variables:

duck Duck number (1-50).

time Survival time in days.

status Event indicator: 1 = dead, 0 = alive or missing.

age Age group: 0 = hatch-year bird, 1 = bird aged at least one year.

weight Body weight in g.

length Wing length in mm.

Details

Fifty female ducks from two locations in New Jersey were captured between 8 November and 14 December 1983 and their status was recorded daily until the study ended on 15 February 1984; there are 18 deaths. The data are included in Hand et al. (1994).

The appendix suggests a proportional hazards model for age, weight and length, a test of whether the coefficients of weight and length differ between age groups, and an assessment of whether non-linear functions of weight and length are needed.

Aliases

This data set is not used in a numbered example, so it has no `dX.Y` alias. See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Appendix B.3 (Table B.3). Data from Hand et al. (1994); the study was conducted by the US Fish and Wildlife Service.

Examples

```
# Collett (2023), Appendix B.3 (no numbered example) ------------------
# Chapter 17 style: proportional hazards model for age, weight and length
survival::coxph(
  survival::Surv(time, status) ~ age + weight + length,
  data = ducks
)

# Modern tools ---------------------------------------------------------
if (requireNamespace("broom", quietly = TRUE)) {
  broom::tidy(
    survival::coxph(survival::Surv(time, status) ~ age + weight + length, data = ducks),
    exponentiate = TRUE, conf.int = TRUE
  )
}
```

---

gastric *Survival of patients with gastric cancer*

Description

Survival times, in days, of 90 patients with locally advanced gastric cancer, randomised to chemotherapy alone or chemotherapy plus radiotherapy (Collett, 2023, Example 11.2).

Usage

```
gastric
```

Format

A data frame with 90 rows and 4 variables:

patient Patient number (1-90).

time Time from randomisation to death from gastric cancer, in days.

status Event indicator: 1 = death from gastric cancer, 0 = right-censored.

treat Treatment: 0 = chemotherapy alone, 1 = chemotherapy plus radiotherapy.

Details

The trial was carried out by the Gastrointestinal Tumor Study Group; the patients were followed for over eight years. There are 82 deaths. The estimated survivor functions cross, so the proportional hazards assumption is not tenable and the data are used to motivate non-proportional hazards: the log-cumulative hazard plot, the restricted mean survival time and time-dependent effects (Examples 11.2 and 11.3).

Aliases

The data set is available under the aliases `d11.2`, `d11.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 11.2 (Table 11.2). Data from Stablein and Koutrouvelis (1985).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 11.2 and 11.3: evidence against proportional hazards
survival::cox.zph(survival::coxph(survival::Surv(time, status) ~ treat, data = gastric))

# A treatment effect that differs between time periods
gastric_cp <- survival::survSplit(
  gastric,
  cut = c(360, 720, 1080),
  end = "time", event = "status", start = "tstart",
  episode = "period"
)
survival::coxph(
  survival::Surv(tstart, time, status) ~ treat + treat:strata(period),
  data = gastric_cp
)

# Modern tools ------------------------------------------------------
# The restricted mean survival time summarises the same non-proportional effect
```

```
rmst <- summary(
  survival::survfit(survival::Surv(time, status) ~ treat, data = gastric),
  rmean = 1000
)
rmst$table[, c("rmean", "se(rmean)")]
```

`help` *Health evaluation and linkage to primary care*

Description

Time from randomisation to linkage to primary care for 447 patients in a residential detoxification programme, in the HELP study (Collett, 2023, Example 16.7).

Usage

```
help
```

Format

A data frame with 447 rows and 7 variables:

subject Patient identifier.

days Time from randomisation to linkage to primary care, in days.

status Event indicator: 1 = linked to primary care, 0 = censored.

age Age of the patient in years.

gender Gender: 0 = female, 1 = male.

housing Housing status: 0 = homeless, 1 = housed.

linkage Assistance from the HELP clinic in linking to primary care: 0 = no, 1 = yes.

Details

Patients undergoing detoxification for alcohol, heroin or cocaine who had no primary care physician were randomised to help from a HELP clinic or to no such help. The response is the number of days until linkage to primary care, and there are 167 events. The file is the complete-case version of the `HELPmiss` data set in the mosaicData package.

The example is the running illustration of Bayesian survival analysis: a Bayesian Weibull model and a piecewise exponential model (Example 16.7).

Aliases

The data set is available under the alias `d16.7`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 16.7 (Tables 16.1 and 16.2). Data from Samet et al. (2003); see also the mosaicData package.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 16.7: a Bayesian Weibull and piecewise exponential model
# Chapter 17 fits these with rstanarm::stan_surv(); the maximum likelihood
# counterparts below give the same point estimates of the hazard ratios.
survival::survreg(
  survival::Surv(days, status) ~ age + gender + housing + linkage,
  data = sta310r::help, dist = "weibull"
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(days, status) ~ age + gender + housing + linkage,
    data = sta310r::help, dist = "weibull"
  )
}
```

---

hepatitis *Chronic active hepatitis*

Description

Survival times, in months, of 44 patients with chronic active hepatitis, randomised to prednisolone or to an untreated control group (Collett, 2023, Appendix B.1).

Usage

```
hepatitis
```

Format

A data frame with 44 rows and 3 variables:

treatment Treatment: 1 = prednisolone, 2 = control.

time Survival time in months from admission to the trial.

status Event indicator: 1 = death, 0 = right-censored.

Details

The data were originally reported by Kirk et al. (1980) and given in Pocock (1983); there are 27 deaths. The trial is used in Chapter 15 as the worked example for sample size calculation: with five-year survival of 0.41 under the standard treatment and 0.60 under the new treatment, the number of events required by Schoenfeld's formula is about 140, corresponding to roughly 400 patients (Examples 15.1-15.3). The appendix also suggests survival curves by treatment, the log-rank and Wilcoxon tests, Cox and Weibull models and a log-cumulative hazard plot.

Aliases

The data set is available under the aliases `d15.1`, `d15.2`, `d15.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Appendix B.1 (Table B.1) and Examples 15.1-15.3. Data from Kirk et al. (1980) via Pocock (1983).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 15.1-15.3: number of events for the log-rank test
# Schoenfeld's formula, as in Equation 15.1
hr <- 0.57
alpha <- 0.05
power <- 0.8
(stats::qnorm(1 - alpha / 2) + stats::qnorm(power))^2 /
  (0.25 * log(hr)^2)

# The survival times themselves are used for the exponential model
survival::survfit(survival::Surv(time, status) ~ treatment, data = hepatitis)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(time, status) ~ treatment,
    data = hepatitis, dist = "exponential"
  )
}
```

---

hyper *Treatment of hypernephroma*

Description

Survival times, in months, of 36 patients with hypernephroma (a malignant tumour of the kidney) treated with chemotherapy and immunotherapy, with or without a nephrectomy, and classified by age group (Collett, 2023, Example 3.4).

Usage

```
hyper
```

Format

A data frame with 36 rows and 4 variables:

nephrectomy Whether the kidney was surgically removed: 0 = no, 1 = yes.

age Age group at diagnosis: 1 = under 60 years, 2 = 60-70 years, 3 = over 70 years.

time Survival time in months.

status Event indicator: 1 = death, 0 = right-censored.

Details

The age of each patient has been grouped for the analysis (the underlying age in years is not included). The example is the running illustration of the Cox regression model: indicator variables for a factor with three levels, models with and without interaction, and the interpretation of the baseline hazard (Examples 3.4, 3.10, 3.11, 3.13 and 3.14), followed by the corresponding Weibull analysis (Examples 5.10, 6.1 and 6.4).

Aliases

The data set is available under the aliases `d3.4`, `d3.10`, `d3.11`, `d3.13`, `d3.14`, `d5.10`, `d6.1`, `d6.4`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 3.4 (Table 3.6). Data from Lee and Wang (2013).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 3.4, 3.10, 3.11, 3.13 and 3.14: Cox model with age group
fit_hyper <- survival::coxph(
  survival::Surv(time, status) ~ nephrectomy + factor(age),
  data = hyper
)
summary(fit_hyper)

# Example 5.10: the corresponding Weibull model
survival::survreg(
  survival::Surv(time, status) ~ nephrectomy + factor(age),
  data = hyper, dist = "weibull"
)

# Examples 6.1 and 6.4: piecewise exponential baseline hazard
# Chapter 17 fits this with phreg(dist = "pch", cuts = ...) in the eha package;
# pchreg() is the current name of the same fit.
```

```
eha::pchreg(
  survival::Surv(time, status) ~ factor(age) + nephrectomy,
  data = hyper,
  cuts = c(5, 6, 8, 9, 10, 12, 14, 15, 17, 18, 21, 26, 35, 36, 38,
           48, 52, 56, 68, 72, 84, 108, 115)
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvspline(
    survival::Surv(time, status) ~ nephrectomy + factor(age),
    data = hyper, k = 1
  )
}
```

---

iud *Time to discontinuation of an intrauterine device (IUD)*

Description

Number of weeks from the start of use of a Multiload 250 intrauterine device until discontinuation because of menstrual bleeding problems, for 18 women (Collett, 2023, Example 1.1). The data come from a World Health Organization study of contraceptive methods.

Usage

```
iud
```

Format

A data frame with 18 rows and 2 variables:

time Time from first use of the IUD to discontinuation, in weeks.

status Event indicator: 1 = discontinuation due to menstrual bleeding problems, 0 = right-censored.

Details

The time origin is the first day of IUD use and the endpoint is discontinuation because of bleeding problems. Women who stopped using the device for other reasons (desire for pregnancy, no further need of contraception, lost to follow-up) contribute censored observations, as do three women whose follow-up extended beyond the two-year study protocol. There are 9 events and 9 censored times among the 18 women, three of the censored times being greater than 104 weeks.

The data are small enough to illustrate the product-limit estimator by hand. In the book they are used for the Kaplan-Meier estimate and the median (Examples 2.3-2.10), for exponential, Weibull and log-logistic parametric models (Examples 5.1-5.4) and for Bayesian analyses (Examples 16.1-16.3).

Aliases

The data set is available under the aliases `d1.1`, `d2.3`, `d2.4`, `d2.5`, `d2.7`, `d2.8`, `d2.9`, `d2.10`, `d5.1`, `d5.2`, `d5.3`, `d5.4`, `d16.1`, `d16.2`, `d16.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 1.1 (Table 1.1). Data from the World Health Organization (1987).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 1.1: times to discontinuation of an IUD
survival::survfit(survival::Surv(time, status) ~ 1, data = iud)

# Examples 2.3-2.5 and 2.7-2.10: product-limit estimate and the median
km_iud <- survival::survfit(survival::Surv(time, status) ~ 1, data = iud)
summary(km_iud, times = c(12, 24, 48))
quantile(km_iud, probs = c(0.25, 0.5, 0.75))

# Examples 5.1-5.4: exponential, Weibull and log-logistic models
survival::survreg(survival::Surv(time, status) ~ 1, data = iud, dist = "exponential")
survival::survreg(survival::Surv(time, status) ~ 1, data = iud, dist = "weibull")
survival::survreg(survival::Surv(time, status) ~ 1, data = iud, dist = "loglogistic")

# Examples 16.1-16.3: conjugate gamma posterior for the exponential rate
# (a vague Gamma(0.001, 0.001) prior, as in the book)
a1 <- 0.001 + sum(iud$status)
b1 <- 0.001 + sum(iud$time)
c(mean = a1 / b1, stats::qgamma(c(0.025, 0.975), a1, b1))
# For a non-conjugate model Chapter 17 uses rstanarm::stan_surv().

# Modern tools ------------------------------------------------------
if (requireNamespace("ggsurvfit", quietly = TRUE)) {
  ggsurvfit::survfit2(survival::Surv(time, status) ~ 1, data = iud) |>
    ggsurvfit::ggsurvfit() +
    ggsurvfit::add_confidence_interval()
}
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(survival::Surv(time, status) ~ 1, data = iud, dist = "weibull")
}
```

---

kidney *Survival following kidney transplantation*

Description

Transplant survival times of 434 kidney recipients, where the two kidneys of a donor may have been transplanted into two different recipients (Collett, 2023, Example 10.4).

Usage

```
kidney
```

Format

A data frame with 434 rows and 7 variables:

patient Patient number (1-434).

donor Donor number (1-270). Recipients of kidneys from the same donor share this value.

time Transplant survival time in days; the earlier of graft failure and death with a functioning graft.

status Event indicator: 1 = transplant failure, 0 = right-censored.

age Age of the recipient in years.

diabetes Diabetes status: 0 = absent, 1 = present.

cit Cold ischaemic time, in hours, between retrieval of the kidney and transplantation.

Details

The transplants used organs from deceased donors following circulatory death. Of the 270 donors, 106 gave one kidney and 164 gave both, so many recipients are clustered within donors and their survival times are likely to be correlated; there are 71 failures. The example fits Weibull models with and without a shared frailty effect for the donor (Example 10.4).

See kidney_centres for transplant survival in eight centres, a centre-level analogue of the same problem.

Aliases

The data set is available under the alias `d10.4`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 10.4 (Table 10.5). Data from the UK Transplant Registry.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 10.4: shared gamma frailty for kidneys from the same donor
survival::coxph(
  survival::Surv(time, status) ~ age + cit + diabetes +
    frailty(donor, distribution = "gamma"),
  data = kidney
)

# Modern tools ------------------------------------------------------
if (requireNamespace("coxme", quietly = TRUE)) {
  coxme::coxme(
    survival::Surv(time, status) ~ age + cit + diabetes + (1 | donor),
```

```
    data = kidney
  )
}
```

`kidney_centres` *Comparisons between kidney transplant centres*

Description

Transplant survival times of 1439 kidney recipients in eight UK transplant centres (Collett, 2023, Example 11.4).

Usage

```
kidney_centres
```

Format

A data frame with 1439 rows and 9 variables:

patient Patient identifier.

centre Transplant centre (1-8).

tsurv Transplant survival time in days.

tcens Event indicator: 1 = transplant failure, 0 = right-censored.

dage Age of the donor in years.

dtype Donor type: 0 = donation after brain death (DBD), 1 = donation after circulatory death (DCD).

rage Age of the recipient in years.

diab Diabetic status of the recipient: 0 = absent, 1 = present.

cit Cold ischaemic time, in hours.

Details

The data cover transplants with organs from deceased donors performed between January 2009 and December 2011. Transplant survival is the earlier of graft failure and patient death, and there are 206 failures. One recipient (patient 7) had a kidney that never functioned, so `tsurv` is 0.

The data are used to compare centre-specific one-year failure rates, both unadjusted and risk-adjusted through a Cox model, and to estimate centre-specific rates with fixed and random effects (Examples 11.4-11.8).

Aliases

The data set is available under the aliases `d11.4`, `d11.5`, `d11.6`, `d11.7`, `d11.8`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 11.4 (Table 11.4). Data from the UK Transplant Registry.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 11.4-11.8: comparison of transplant centres
survival::coxph(
  survival::Surv(tsurv, tcens) ~ strata(centre) + dage + dtype + rage + diab + cit,
  data = kidney_centres
)

# A random (frailty) effect for centre instead of stratification
survival::coxph(
  survival::Surv(tsurv, tcens) ~ dage + dtype + rage + diab + cit +
    frailty(centre, distribution = "gamma"),
  data = kidney_centres
)

# Modern tools ------------------------------------------------------
if (requireNamespace("coxme", quietly = TRUE)) {
  coxme::coxme(
    survival::Surv(tsurv, tcens) ~ dage + dtype + rage + diab + cit +
      (1 | centre),
    data = kidney_centres
  )
}
```

`liver` *Survival of liver transplant recipients*

Description

Graft survival times of 1761 adult liver transplant recipients, with the cause of graft failure (Collett, 2023, Example 12.1).

Usage

```
liver
```

Format

A data frame with 1761 rows and 7 variables:

patient Patient number (1-1761).

age Age of the patient in years.

gender Gender: 1 = male, 2 = female.

disease Primary liver disease: 1 = primary biliary cirrhosis (PBC), 2 = primary sclerosing cholangitis (PSC), 3 = alcoholic liver disease (ALD).

time Time from transplantation to graft failure, death or last follow-up, in days.

status Event indicator: 1 = graft failure or death, 0 = right-censored.

cof Cause of graft failure: 0 = functioning graft (censored), 1 = rejection, 2 = thrombosis, 3 = recurrent disease, 4 = other.

Details

The patients had a first elective transplant with an organ from a deceased donor between January 2000 and December 2010 and were followed up to the end of 2012. In the book there are 261 graft failures among the 1761 patients, and an additional 211 patients die with a functioning graft; the failure times of the latter group are censored at death. In the distributed file `status` is 1 for 265 patients while `cof` is non-zero for 260, so the two indicators are not perfectly consistent.

The data are the main illustration of competing risks: the cumulative incidence of each cause of graft failure, cause-specific hazard models and the subdistribution approach (Examples 12.1-12.6). See liver_wait for the waiting-list data from the same registry.

Aliases

The data set is available under the aliases `d12.1`, `d12.2`, `d12.3`, `d12.4`, `d12.5`, `d12.6`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 12.1 (Table 12.1). Data from the UK Transplant Registry.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 12.1-12.6: cumulative incidence by cause of graft failure
# (cof: 1 = rejection, 2 = thrombosis, 3 = recurrent disease, 4 = other)
ci <- cmprsk::cuminc(liver$time, liver$cof, liver$disease)
plot(ci, curvlab = c("PBC", "PSC", "ALD"), xlab = "Time (days)")

# A Fine-Gray model for one cause of failure (cause 2, thrombosis)
cmprsk::crr(
  liver$time, liver$cof,
  cov1 = as.matrix(liver[c("age", "gender")]),
  failcode = 2
)

# Modern tools ------------------------------------------------------
if (requireNamespace("tidycmprsk", quietly = TRUE)) {
  tidycmprsk::cuminc(survival::Surv(time, factor(cof)) ~ disease, data = liver)
  tidycmprsk::crr(
    survival::Surv(time, factor(cof)) ~ age + gender,
    data = liver, failcode = "2"
```

```
  )
}
```

`liver_wait` *Time to death while waiting for a liver transplant*

Description

Time from registration to death on the waiting list for 281 adults with primary biliary cirrhosis (Collett, 2023, Example 14.1).

Usage

```
liver_wait
```

Format

A data frame with 281 rows and 7 variables:

patient Patient identifier.

time Time from registration to death on the waiting list, in days.

status Event indicator: 1 = died while waiting, 0 = censored (usually at transplantation).

age Age of the patient in years.

gender Gender: 0 = female, 1 = male.

bmi Body mass index, in kg/m squared.

ukeld UK end-stage liver disease score; higher values indicate more severe disease.

Details

The patients were first registered for a liver transplant in the five years from 1 January 2006, and there are 64 deaths. The time of a patient who receives a transplant is censored at transplantation. Livers are allocated on the basis of need, so the time to transplantation (censoring) is associated with the time to death; the data are used to illustrate sensitivity analysis for dependent censoring and inverse probability of censoring weighting (Examples 14.1 and 14.3). The rows are ordered by survival time.

See liver for the post-transplant graft survival data from the same registry.

Aliases

The data set is available under the aliases `d14.1`, `d14.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 14.1 (Table 14.1). Data from the UK Transplant Registry.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 14.1 and 14.3: dependent censoring (transplant) modelled with
# inverse probability of censoring weights
liver_wait$cens <- 1 - liver_wait$status
cens_fit <- survival::survreg(
  survival::Surv(time, cens) ~ age + gender + bmi + ukeld,
  data = liver_wait
)
lp <- cens_fit$linear.predictors
sigma <- cens_fit$scale

# Probability of still being uncensored at each observed time
survival_uncensored <- exp(-exp((log(liver_wait$time) - lp) / sigma))
weights <- 1 / survival_uncensored
summary(weights)

survival::coxph(
  survival::Surv(time, status) ~ age + gender + bmi + ukeld,
  data = liver_wait, weights = weights
)
```

---

lung_transplant *Survival of patients registered for a lung transplant*

Description

Survival times from registration for a lung transplant for 196 UK patients (Collett, 2023, Example 10.3).

Usage

```
lung_transplant
```

Format

A data frame with 196 rows and 7 variables:

patient Patient identifier.

time Survival time in days from registration for a transplant.

status Event indicator: 1 = death, 0 = right-censored.

age Age at registration in years.

gender Gender: 1 = male, 2 = female.

bmi Body mass index at registration.

disease Primary disease: 1 = COPD, 2 = fibrosis, 3 = suppurative disease, 4 = other.

Details

The patients were registered in 2004 and survival is measured from registration to death regardless of any intervening transplantation; 123 of the 196 patients died. Times are censored at removal from the list, at the last known follow-up date, or at 30 April 2012 for patients still alive. The disease codings above follow the values in the distributed file (they can be checked against Table 10.2 of the book).

The example compares a Weibull model with and without an individual lognormal frailty term (Example 10.3). See kidney for a shared frailty example and liver_wait for a similar waiting-list cohort.

Aliases

The data set is available under the alias `d10.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 10.3 (Table 10.2). Data from the UK Transplant Registry.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 10.3: an individual gamma frailty for each patient
survival::coxph(
  survival::Surv(time, status) ~ age + bmi + factor(gender) + factor(disease) +
    frailty(patient, distribution = "gamma"),
  data = lung_transplant
)

# Modern tools ------------------------------------------------------------
if (requireNamespace("coxme", quietly = TRUE)) {
  coxme::coxme(
    survival::Surv(time, status) ~ age + bmi + factor(gender) +
      factor(disease) + (1 | patient),
    data = lung_transplant
  )
}
```

---

melanoma *Survival times of patients with melanoma*

Description

Survival times, in months, of 30 patients with melanoma, from two immunotherapy treatment groups and stratified by age group (Collett, 2023, Example 2.15).

Usage

```
melanoma
```

Format

A data frame with 30 rows and 4 variables:

age Age group: 1 = 21-40 years, 2 = 41-60 years, 3 = 61 years and over.

treatment Treatment: 1 = BCG vaccine, 2 = Corynebacterium parvum vaccine.

time Survival time in months.

status Event indicator: 1 = death, 0 = right-censored.

Details

The tumour was removed surgically before randomisation. The data are used to illustrate the log-rank test for comparing two treatment groups within each age stratum (Examples 2.15 and 2.16) and the log-rank test for trend.

Aliases

The data set is available under the aliases `d2.15`, `d2.16`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Examples 2.15 and 2.16 (Table 2.9). Data from Lee and Wang (2013).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 2.15 and 2.16: log-rank test within each age stratum
survival::survfit(
  survival::Surv(time, status) ~ treatment + strata(age),
  data = melanoma
)
survival::survdiff(
  survival::Surv(time, status) ~ treatment + strata(age),
  data = melanoma
)

# Modern tools ------------------------------------------------------
if (requireNamespace("ggsurvfit", quietly = TRUE)) {
  ggsurvfit::survfit2(
    survival::Surv(time, status) ~ treatment + strata(age),
    data = melanoma
  ) |>
    ggsurvfit::ggsurvfit()
}
```

`mice` *Survival of laboratory mice with competing causes of death*

Description

Survival times and causes of death for 181 irradiated mice, reared in a standard or a germ-free environment (Collett, 2023, Example 12.7).

Usage

```
mice
```

Format

A data frame with 181 rows and 3 variables:

environment Rearing environment: 1 = standard, 2 = germ-free.

causeofdeath Cause of death: 1 = thymic lymphoma, 2 = reticulum cell sarcoma, 3 = other causes.

time Survival time in days.

Details

The mice were RFM strain males, a strain susceptible to tumour development after exposure to ionising radiation, and were given 300 rads at five or six weeks of age. An autopsy established the cause of death for every mouse. There is no censoring. There are 99 mice in the standard environment (22 deaths from thymic lymphoma, 38 from reticulum cell sarcoma and 39 from other causes) and 82 in the germ-free environment (29, 15 and 38).

The data illustrate estimation of the cumulative incidence function for each cause and cause-specific Cox regression (Example 12.7).

Aliases

The data set is available under the alias `d12.7`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 12.7 (Table 12.7). Data first described by Hoel (1972).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 12.7: three competing causes of death
# (1 = thymic lymphoma, 2 = reticulum cell sarcoma, 3 = other causes)
ci_mice <- cmprsk::cuminc(mice$time, mice$causeofdeath, mice$environment)
plot(ci_mice, curvlab = c("Standard", "Germ-free"), xlab = "Time (days)")
```

```
# Modern tools ------------------------------------------------------
if (requireNamespace("tidycmprsk", quietly = TRUE)) {
  tidycmprsk::cuminc(
    survival::Surv(time, factor(causeofdeath)) ~ environment,
    data = mice
  )
}
```

`myeloma` *Survival of multiple myeloma patients*

Description

Survival times, in months, from diagnosis until death for 48 patients with multiple myeloma, together with seven prognostic variables (Collett, 2023, Example 1.3).

Usage

```
myeloma
```

Format

A data frame with 48 rows and 10 variables:

patient Patient number (1-48).

time Survival time in months from diagnosis.

status Event indicator: 1 = death from multiple myeloma, 0 = right-censored.

age Age at diagnosis in years (50-80).

sex Sex: 1 = male, 2 = female.

bun Blood urea nitrogen level.

ca Serum calcium level.

hb Haemoglobin level.

pcells Percentage of plasma cells in the bone marrow.

protein Bence-Jones protein in the urine: 0 = absent, 1 = present.

Details

There are 36 deaths among the 48 patients. The time origin is the date of diagnosis and the endpoint is death from multiple myeloma; patients alive at the end of the study contribute censored times.

The data are used throughout the book: life-table estimation (Examples 2.2 and 2.6), Cox regression with several covariates, variable selection and interaction (Examples 3.2-3.20), model checking and influence diagnostics (Examples 4.5, 4.8 and 4.9) and a Weibull model for a prognostic index (Example 5.16).

Aliases

The data set is available under the aliases `d1.3`, `d2.2`, `d2.6`, `d3.2`, `d3.5`, `d3.7`, `d3.8`, `d3.9`, `d3.15`, `d3.17`, `d3.19`, `d3.20`, `d4.5`, `d4.8`, `d4.9`, `d5.16`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 1.3 (Table 1.3). Original data from Krall, Uthoff and Harley (1975); the study was carried out at the Medical Center of the University of West Virginia, USA.

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 2.2 and 2.6: the survivor function for a single sample
survival::survfit(survival::Surv(time, status) ~ 1, data = myeloma)

# Examples 3.2, 3.5, 3.8, 3.9, 3.15, 3.17, 3.19 and 3.20: Cox model
fit_myeloma <- survival::coxph(
  survival::Surv(time, status) ~ age + sex + bun + ca + hb + pcells + protein,
  data = myeloma
)
summary(fit_myeloma)

# Example 3.7: lasso variable selection (the book uses the penalized package)
# The fitting routine prints its progress; it is captured here so that the
# selected coefficients are easy to read.
progress <- utils::capture.output(
  lass <- penalized::penalized(
    survival::Surv(myeloma$time, myeloma$status),
    penalized = ~ age + sex + bun + ca + hb + pcells + protein,
    lambda1 = 30,
    data = myeloma
  )
)
round(lass@penalized, 3)

# Examples 4.5, 4.8 and 4.9: residuals and the proportional hazards assumption
head(stats::residuals(fit_myeloma, type = "martingale"))
survival::cox.zph(fit_myeloma)

# Example 5.16: Weibull model with a prognostic index
survival::survreg(
  survival::Surv(time, status) ~ age + bun + hb,
  data = myeloma, dist = "weibull"
)

# Modern tools ------------------------------------------------------
if (requireNamespace("broom", quietly = TRUE)) {
  broom::tidy(fit_myeloma, exponentiate = TRUE, conf.int = TRUE)
}
if (requireNamespace("glmnet", quietly = TRUE)) {
```

```
  x <- as.matrix(myeloma[c("age", "sex", "bun", "ca", "hb", "pcells", "protein")])
 cv <- glmnet::cv.glmnet(x, survival::Surv(myeloma$time, myeloma$status), family = "cox")
  stats::coef(cv, s = "lambda.min")
}
if (requireNamespace("rms", quietly = TRUE)) {
  rms::cph(
    survival::Surv(time, status) ~ age + sex + bun + ca + hb + pcells + protein,
    data = myeloma, x = TRUE, y = TRUE
  )
}
```

`numerical_illustration` *A numerical illustration of two factors*

Description

Artificial survival times of 37 patients classified by the levels of two two-level factors, A and B (Collett, 2023, Example 5.9).

Usage

```
numerical_illustration
```

Format

A data frame with 37 rows and 4 variables:

a Level of factor A: 1 or 2.

b Level of factor B: 1 or 2.

time Survival time.

status Event indicator: 1 = event, 0 = right-censored.

Details

The data are not real. They are constructed so that factor A has a proportional effect while the effect of factor B, and the interaction between A and B, are more complicated, and are used to compare Weibull proportional hazards and accelerated failure time models fitted to the same data (Example 5.9).

Aliases

The data set is available under the alias `d5.9`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 5.9 (Table 5.1).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 5.9: Weibull models for two two-level factors and their interaction
survival::survreg(
  survival::Surv(time, status) ~ factor(a) * factor(b),
  data = numerical_illustration, dist = "weibull"
)

# The proportional hazards form of the same model
survival::coxph(
  survival::Surv(time, status) ~ factor(a) * factor(b),
  data = numerical_illustration
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(time, status) ~ factor(a) * factor(b),
    data = numerical_illustration, dist = "weibull"
  )
}
```

---

ovarian *Chemotherapy in ovarian cancer patients*

Description

Survival times, in days, of 26 women with ovarian cancer randomised to one of two chemotherapy regimens (Collett, 2023, Example 5.11).

Usage

```
ovarian
```

Format

A data frame with 26 rows and 7 variables:

patient Patient number (1-26).

time Survival time in days from randomisation.

status Event indicator: 1 = death, 0 = right-censored.

treat Treatment: 1 = cyclophosphamide alone (single), 2 = cyclophosphamide combined with adriamycin.

age Age of the patient in years.

rdisease Extent of residual disease: 1 = incomplete excision, 2 = complete excision.

perf Performance status: 1 = good, 2 = poor.

Details

All women had minimal residual disease after surgical excision of tumours larger than 2 cm. There are 12 deaths among the 26 patients. The factors `treat`, `rdisease` and `perf` each have two levels and are fitted as indicator variables, so the baseline hazard does not correspond to a patient with all covariates zero (see Example 5.11).

The data are used for Weibull proportional hazards models and variable selection (Examples 5.11 and 5.12), for model checking of parametric models (Examples 7.1-7.4) and for a time-dependent covariate (Example 8.3).

Aliases

The data set is available under the aliases `d5.11`, `d5.12`, `d7.1`, `d7.2`, `d7.3`, `d7.4`, `d8.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 5.11 (Table 5.6). Data from Edmunson et al. (1979) via Therneau (1986).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 5.11 and 5.12: Weibull proportional hazards model
fit_ovarian <- survival::survreg(
  survival::Surv(time, status) ~ treat + age,
  data = ovarian, dist = "weibull"
)
summary(fit_ovarian)

# Examples 7.1-7.4: residual checks for the parametric model
head(stats::residuals(fit_ovarian, type = "deviance"))

# Example 8.3: a time-dependent effect of treatment
survival::coxph(
  survival::Surv(time, status) ~ treat + tt(treat),
  data = ovarian,
  tt = function(x, t, ...) x * log(t)
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(time, status) ~ treat + age,
    data = ovarian, dist = "weibull"
  )
}
```

`prostate` *Comparison of two treatments for prostatic cancer*

Description

Survival times, in months, of 38 men with Stage III prostatic cancer, randomised to placebo or to diethylstilbestrol (DES) (Collett, 2023, Example 1.4).

Usage

```
prostate
```

Format

A data frame with 38 rows and 8 variables:

patient Patient number (1-38).

treatment Treatment group: 1 = placebo, 2 = diethylstilbestrol (DES).

time Survival time in months from randomisation.

status Event indicator: 1 = death from prostatic cancer, 0 = right-censored.

age Age at trial entry, in years.

shb Serum haemoglobin level, in g/100 ml.

size Size of the primary tumour, in cm squared.

index Combined index of tumour stage and grade (the Gleason index); higher values indicate a more advanced tumour.

Details

The trial was run by the Veteran's Administration Cooperative Urological Research Group and only patients with a local tumour extension but without elevated serum prostatic acid phosphatase are included. Deaths from other causes and losses to follow-up are censored. There are only 6 deaths in this subset of 38 patients.

With so few events the example is used to show the behaviour of the partial likelihood and of the estimated hazard ratios (Examples 3.6, 3.12, 3.16 and 5.15), and the fitted Cox model is re-used for an accelerated failure time comparison (Example 5.15).

Aliases

The data set is available under the aliases `d1.4`, `d3.6`, `d3.12`, `d3.16`, `d5.15`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 1.4 (Table 1.4). The full trial is given in Andrews and Herzberg (1985).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 1.4: the trial data
survival::survfit(survival::Surv(time, status) ~ treatment, data = prostate)

# Examples 3.6, 3.12 and 3.16: Cox model for treatment and prognostic variables
fit_prostate <- survival::coxph(
  survival::Surv(time, status) ~ treatment + age + shb + size + index,
  data = prostate
)
summary(fit_prostate)

# Example 5.15: the corresponding Weibull accelerated failure time model
survival::survreg(
  survival::Surv(time, status) ~ treatment + age + shb + size + index,
  data = prostate, dist = "weibull"
)

# Modern tools ------------------------------------------------------
if (requireNamespace("flexsurv", quietly = TRUE)) {
  flexsurv::flexsurvreg(
    survival::Surv(time, status) ~ treatment + age + shb + size + index,
    data = prostate, dist = "weibull"
  )
}
```

`pulmonary` *Pulmonary metastasis in patients with osteosarcoma*

Description

Survival times, in months, of eleven male patients with pulmonary metastasis arising from osteosarcoma (Collett, 2023, Example 2.1).

Usage

```
pulmonary
```

Format

A data frame with 11 rows and 1 variable:

time Survival time in months.

Details

There is no censoring in these data and all survival times are distinct, so the survivor function can be estimated directly from the observed proportions surviving. The estimated values of S(t) at 11, 13, 14, 15 and 17 months are 1.000, 0.909, 0.455, 0.273 and 0.091, and S(t) is zero after 17 months. The example is used in the book to introduce the empirical survivor function before censoring is considered.

Aliases

The data set is available under the alias `d2.1`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 2.1. Data from Burdette and Gehan (1970).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 2.1: no censoring, so Surv() needs only the time
survival::survfit(survival::Surv(time) ~ 1, data = pulmonary)

# Modern tools ------------------------------------------------------
if (requireNamespace("ggsurvfit", quietly = TRUE)) {
  ggsurvfit::survfit2(survival::Surv(time) ~ 1, data = pulmonary) |>
    ggsurvfit::ggsurvfit()
}
```

---

rats *Recurrence of mammary tumours in female rats*

Description

Times from randomisation to the development of mammary tumours for 48 female rats, treated with a retinoid or a control, in recurrent-event form (Collett, 2023, Example 13.2).

Usage

```
rats
```

Format

A data frame with 254 rows and 4 variables:

rat Rat number (1-48).

treatment Treatment: 0 = control, 1 = retinoid prophylaxis.

time Time from randomisation to a detected tumour, in days, or to censoring.

status Event indicator: 1 = tumour detected, 0 = right-censored.

Details

All rats were injected with a carcinogen and the 48 rats that were tumour free after 60 days were randomised to retinoid or control. Rats were examined twice weekly and the observation period ended 122 days after randomisation. Because several tumours could be found at the same assessment, tied times have been broken by adding 0.1 and 0.2 days, so `time` is not always a whole number. There are 23 treated and 25 control rats, 212 detected tumours in total, and only two rats never developed a tumour.

The data are used for multiple-event and event-history models: the Andersen-Gill, PWP and WLW models, with the data expanded into counting process format (Example 13.2; see also cirrhosis_cp for a simpler counting process illustration).

Aliases

The data set is available under the alias `d13.2`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 13.2 (Table 13.2). Data from Gail, Santner and Brown (1980).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 13.2: recurrent tumours; Andersen-Gill model with robust SEs
# The start of each interval is the previous tumour time for that rat.
rats <- rats[order(rats$rat, rats$time), ]
rats$start <- unlist(lapply(split(rats$time, rats$rat), function(t) c(0, t[-length(t)])))
survival::coxph(
  survival::Surv(start, time, status) ~ treatment + cluster(rat),
  data = rats
)

# Modern tools ------------------------------------------------------
# The Prentice-Williams-Peterson model stratifies by the number of prior events
rats$event_number <- unlist(lapply(split(rats$rat, rats$rat), seq_along))
survival::coxph(
  survival::Surv(start, time, status) ~ treatment + strata(event_number),
  data = rats
)
```

`scottish_mbc_cohort` *Synthetic Scottish metastatic breast cancer cohort*

Description

A cohort of 850 patients diagnosed with metastatic breast cancer.

Usage

```
scottish_mbc_cohort
```

Format

A data frame with 850 rows and 21 variables:

patient_id Patient identifier, `MBC0001` to `MBC0850`.

health_board NHS Health Board of the treating centre, one of the 14 Scottish boards.

age_at_diagnosis Age at diagnosis of metastatic disease, in years.

sex Sex of the patient.

simd_quintile Scottish Index of Multiple Deprivation quintile, 1 (most deprived) to 5 (least deprived).

ecog_ps ECOG performance status at baseline: 0 (best) to 3 (worst).

er_status Oestrogen receptor status: `Positive` or `Negative`.

pr_status Progesterone receptor status: `Positive` or `Negative`.

her2_status HER2 status: `Positive` or `Negative`.

n_metastatic_sites Number of metastatic sites at diagnosis.

bone_mets, liver_mets, lung_mets, brain_mets Metastatic involvement of bone, liver, lung and brain: `Yes` or `No`.

presentation `De novo metastatic` or `Recurrent` disease.

diagnosis_date Date of diagnosis of metastatic disease.

first_line_treatment First-line treatment.

observed_time Follow-up time in months from diagnosis to death or censoring.

status Event indicator: 1 = death, 0 = right-censored.

age_group Age group at diagnosis: `<50`, `50-64`, `65-74` or `75+`.

hormone_receptor Hormone receptor status: `HR-positive` or `HR-negative`, derived from `er_status` and `pr_status`.

Details

The time origin is the date of diagnosis of metastatic breast cancer and the time scale is time since diagnosis, in months. The event of interest is death from any cause; patients who were alive at the end of follow-up are right-censored.

The cohort is **entirely synthetic**. It was generated for teaching purposes and contains no real patient data, but it was designed to reproduce realistic patterns of metastatic breast cancer in Scotland.

Source

The file `scottish_mbc_cohort.xlsx` in the `onco-stats` repository of the Edinburgh Cancer Informatics group, https://github.com/Edinburgh-Cancer-Informatics/onco-stats/tree/main/docs/data, redistributed under the MIT licence of that repository (Copyright (c) 2025 Edinburgh-Cancer-Informatics).

`tamoxifen` *Clinical trial of tamoxifen in breast cancer patients*

Description

Time-to-event data for five different endpoints in 641 women from a randomised trial of tamoxifen alone versus tamoxifen plus radiotherapy (Collett, 2023, Example 13.3).

Usage

```
tamoxifen
```

Format

A data frame with 641 rows and 18 variables:

id Patient identifier.

treat Treatment: 0 = tamoxifen plus radiotherapy, 1 = tamoxifen alone.

age Age of the patient in years.

size Tumour size in cm.

hist Tumour histology: 1 = ductal, 2 = lobular, 3 = medullary, 4 = mixed, 5 = other.

hr Hormone receptor level: 0 = negative, 1 = positive.

hb Haemoglobin level in g/l.

andis Axillary node dissection: 0 = no, 1 = yes.

lsurv Time to local relapse or last follow-up.

ls Local relapse: 0 = no, 1 = yes.

asurv Time to axillary relapse or last follow-up.

as Axillary relapse: 0 = no, 1 = yes.

dsurv Time to distant relapse or last follow-up.

ds Distant relapse: 0 = no, 1 = yes.

msurv Time to second malignancy or last follow-up.

ms Second malignancy: 0 = no, 1 = yes.

tsurv Time from randomisation to death or last follow-up.

ts Status at last follow-up: 0 = alive, 1 = dead.

Details

The women were randomised to tamoxifen alone or with radiotherapy; 320 received tamoxifen plus radiotherapy and 321 tamoxifen alone. Each of the five endpoints has its own time and indicator, and the times are censored at the last date the woman was known to be alive. For women who relapsed after the start of follow-up the relevant indicator switches from 0 to 1.

The example was used by Pintilie (2006) to illustrate the WLW model for multiple events. Because a woman can experience more than one of the five event types, the data are rearranged into one row per possible event per woman before fitting the model (Example 13.3).

Aliases

The data set is available under the alias `d13.3`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 13.3 (Table 13.5). Data from Fyles et al. (2004) and Pintilie (2006).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Example 13.3: four event types, each analysed as a cause-specific hazard
# local relapse
survival::coxph(survival::Surv(lsurv, ls) ~ treat + age + size + hr, data = tamoxifen)
# axillary relapse
survival::coxph(survival::Surv(asurv, as) ~ treat + age + size + hr, data = tamoxifen)
# distant relapse
survival::coxph(survival::Surv(dsurv, ds) ~ treat + age + size + hr, data = tamoxifen)
# second malignancy
survival::coxph(survival::Surv(msurv, ms) ~ treat + age + size + hr, data = tamoxifen)

# Modern tools ------------------------------------------------------
# A competing-risk view of the first event of any of the four types
if (requireNamespace("tidycmprsk", quietly = TRUE)) {
  first <- data.frame(
    time = pmin(tamoxifen$lsurv, tamoxifen$asurv, tamoxifen$dsurv, tamoxifen$msurv),
    treat = tamoxifen$treat
  )
  first$cause <- factor(
    c("local", "axillary", "distant", "second")[
      apply(cbind(tamoxifen$lsurv, tamoxifen$asurv, tamoxifen$dsurv, tamoxifen$msurv),
            1, which.min)
    ]
  )
  tidycmprsk::cuminc(survival::Surv(time, cause) ~ treat, data = first)
}
```

---

ulcer *Recurrence of an ulcer*

Description

Interval-censored times to recurrence of an ulcer among 43 patients in a double-blind clinical trial of two treatments (Collett, 2023, Example 9.1).

Usage

```
ulcer
```

Format

A data frame with 43 rows and 6 variables:

patient Patient number (1-43).

age Age of the patient in years.

duration Duration of verified disease: 1 = less than five years, 2 = five years or more.

treatment Treatment group: 1 = A, 2 = B.

time Time of the last visit, in months.

result Result of the endoscopy at the last visit: 1 = no ulcer detected, 2 = ulcer detected.

Details

Endoscopies were scheduled 6 and 12 months after randomisation, and patients who presented with symptoms between visits had an unscheduled endoscopy. The pair (`time`, `result`) is converted to an interval-censored recurrence time: for example, a visit at 7 months with `result` 2 means recurrence in the interval (6, 7], while `result` 1 at 12 months means the recurrence time is right-censored at 12 months. Because 6 months is the earliest possible interval limit, the combination (`time` = 6, `result` = 2) becomes (0, 6]. There are 11 detected recurrences.

The data are used for interval-censored analyses (Examples 9.1-9.7) and for the restricted mean survival time computed from interval-censored data (Example 11.1). The full multinational study is described by Whitehead (1989); the data here are the Belgian subset.

Aliases

The data set is available under the aliases `d9.1`, `d9.4`, `d9.5`, `d9.6`, `d9.7`, `d11.1`; `dX.Y` is the alias for the data set used in Example X.Y of Collett (2023). See dataset_by_exercise for the complete index.

Source

Collett, D. (2023) *Modelling Survival Data in Medical Research*, 4th ed. Chapman & Hall/CRC, Example 9.1 (Tables 9.1 and 9.2). Data from Whitehead (1989).

Examples

```
# Collett (2023), Chapter 17 style ----------------------------------
# Examples 9.1 and 9.4: interval-censored recurrence times
# (Table 9.2 of the book gives the interval for each visit and result)
ulcer$left  <- ifelse(ulcer$result == 2, ifelse(ulcer$time <= 6, 0, 6), ulcer$time)
ulcer$right <- ifelse(ulcer$result == 2, ulcer$time, Inf)
survival::Surv(ulcer$left, ulcer$right, type = "interval2")[1:5]
# Chapter 17 fits the Turnbull estimate with icfit() in the interval package;
# icenReg provides the same estimator.
icenReg::ic_np(cbind(ulcer$left, ulcer$right))

# Examples 9.5-9.7: a proportional hazards model for interval-censored data
icenReg::ic_sp(
  survival::Surv(left, right, type = "interval2") ~ treatment,
```

```
  data = ulcer
)

# Modern tools ------------------------------------------------------------
if (requireNamespace("icenReg", quietly = TRUE)) {
  plot(icenReg::ic_np(cbind(ulcer$left, ulcer$right)))
}
```

Index

∗ **datasets**
aortic, 2
bladder, 4
bmt, 5
bmt_leukaemia, 6
bmt_outcome, 8
breast, 9
breast_rfs, 11
cgd, 12
cirrhosis, 14
cirrhosis_cp, 15
cirrhosis_lbr, 16
dataset_by_exercise, 17
dialysis, 19
ducks, 20
gastric, 21
help, 23
hepatitis, 24
hyper, 25
iud, 27
kidney, 28
kidney_centres, 30
liver, 31
liver_wait, 33
lung_transplant, 34
melanoma, 35
mice, 37
myeloma, 38
numerical_illustration, 40
ovarian, 41
prostate, 43
pulmonary, 44
rats, 45
scottish_mbc_cohort, 46
tamoxifen, 48
ulcer, 49

aortic, 2

bladder, 4
bmt, 5
bmt_leukaemia, 6, *6*
bmt_outcome, *6*, 8
breast, 9, *18*
breast_rfs, 11

cgd, 12
cirrhosis, 14, *15–17*
cirrhosis_cp, *14*, 15, *16*, *46*
cirrhosis_lbr, *14, 15*, 16

d1.1 *(*iud*)*, 27
d1.2 *(*breast*)*, 9
d1.3 *(*myeloma*)*, 38
d1.4 *(*prostate*)*, 43
d10.1 *(*breast*)*, 9
d10.2 *(*breast*)*, 9
d10.3 *(*lung_transplant*)*, 34
d10.4 *(*kidney*)*, 28
d11.1 *(*ulcer*)*, 49
d11.2 *(*gastric*)*, 21
d11.3 *(*gastric*)*, 21
d11.4 *(*kidney_centres*)*, 30
d11.5 *(*kidney_centres*)*, 30
d11.6 *(*kidney_centres*)*, 30
d11.7 *(*kidney_centres*)*, 30
d11.8 *(*kidney_centres*)*, 30
d12.1 *(*liver*)*, 31
d12.2 *(*liver*)*, 31
d12.3 *(*liver*)*, 31
d12.4 *(*liver*)*, 31
d12.5 *(*liver*)*, 31
d12.6 *(*liver*)*, 31
d12.7 *(*mice*)*, 37
d13.2 *(*rats*)*, 45
d13.3 *(*tamoxifen*)*, 48
d13.4 *(*bmt_outcome*)*, 8
d14.1 *(*liver_wait*)*, 33
d14.3 *(*liver_wait*)*, 33
d15.1 *(*hepatitis*)*, 24

d15.2 (hepatitis), 24
d15.3 (hepatitis), 24
d16.1 (iud), 27
d16.2 (iud), 27
d16.3 (iud), 27
d16.4 (breast), 9
d16.5 (breast), 9
d16.6 (breast), 9
d16.7 (help), 23
d2.1 (pulmonary), 44
d2.10 (iud), 27
d2.11 (breast), 9
d2.12 (breast), 9
d2.13 (breast), 9
d2.14 (breast), 9
d2.15 (melanoma), 35
d2.16 (melanoma), 35
d2.2 (myeloma), 38
d2.3 (iud), 27
d2.4 (iud), 27
d2.5 (iud), 27
d2.6 (myeloma), 38
d2.7 (iud), 27
d2.8 (iud), 27
d2.9 (iud), 27
d3.1 (breast), 9
d3.10 (hyper), 25
d3.11 (hyper), 25
d3.12 (prostate), 43
d3.13 (hyper), 25
d3.14 (hyper), 25
d3.15 (myeloma), 38
d3.16 (prostate), 43
d3.17 (myeloma), 38
d3.19 (myeloma), 38
d3.2 (myeloma), 38
d3.20 (myeloma), 38
d3.3 (breast), 9
d3.4 (hyper), 25
d3.5 (myeloma), 38
d3.6 (prostate), 43
d3.7 (myeloma), 38
d3.8 (myeloma), 38
d3.9 (myeloma), 38
d4.1 (dialysis), 19
d4.10 (dialysis), 19
d4.11 (dialysis), 19
d4.12 (dialysis), 19
d4.2 (dialysis), 19
d4.3 (dialysis), 19
d4.4 (dialysis), 19
d4.5 (myeloma), 38
d4.6 (dialysis), 19
d4.7 (dialysis), 19
d4.8 (myeloma), 38
d4.9 (myeloma), 38
d5.1 (iud), 27
d5.10 (hyper), 25
d5.11 (ovarian), 41
d5.12 (ovarian), 41
d5.13 (breast), 9
d5.14 (breast), 9
d5.15 (prostate), 43
d5.16 (myeloma), 38
d5.17 (breast), 9
d5.2 (iud), 27
d5.3 (iud), 27
d5.4 (iud), 27
d5.5 (breast), 9
d5.6 (breast), 9
d5.7 (breast), 9
d5.8 (breast), 9
d5.9 (numerical_illustration), 40
d6.1 (hyper), 25
d6.4 (hyper), 25
d6.5 (breast_rfs), 11
d6.6 (breast_rfs), 11
d7.1 (ovarian), 41
d7.2 (ovarian), 41
d7.3 (ovarian), 41
d7.4 (ovarian), 41
d8.2 (bmt_leukaemia), 6
d8.3 (ovarian), 41
d8.4 (cirrhosis), 14
d8.5 (aortic), 2
d8.6 (aortic), 2
d8.7 (aortic), 2
d8.8 (aortic), 2
d8.9 (aortic), 2
d9.1 (ulcer), 49
d9.4 (ulcer), 49
d9.5 (ulcer), 49
d9.6 (ulcer), 49
d9.7 (ulcer), 49
dataset_by_exercise, *3, 4*, *6–8*, *10, 11*, *13, 14*, *16*, 17, *17–19*, *21–23*, *25, 26*,

28–30, *32*, *33*, *35–37*, *39*, *40*, *42*, *43*, *45*, *46*, *49*, *50*
dialysis, 19
ducks, 20

gastric, 21

help, 23
hepatitis, 24
hyper, 25

iud, *18*, 27

kidney, 28, *35*
kidney_centres, *29*, 30

liver, 31, *33*
liver_wait, *32*, 33, *35*
lung_transplant, 34

melanoma, 35
mice, 37
myeloma, *18*, 38

numerical_illustration, 40

ovarian, 41

prostate, 43
pulmonary, 44

rats, 45

scottish_mbc_cohort, 46

tamoxifen, 48

ulcer, 49