# Package 'collett'

May 8, 2026

**Type** Package

**Title** Datasets from ``Modelling Survival Data in Medical Research'' by Collett

**Version** 0.1.2

**Maintainer** Mark Clements <mark.clements@ki.se>

**Description** Datasets for the book entitled ``Modelling Survival Data in Medical Research'' by Collett (2023) <doi:10.1201/9781003282525>. The datasets provide extensive examples of time-to-event data.

**URL** https://github.com/mclements/collett

**BugReports** https://github.com/mclements/collett/issues

**License** MIT + file LICENSE

**Depends** R (>= 3.5.0)

**Imports** utils, survival, tinyplot

**Suggests** splines

**Encoding** UTF-8

**RoxygenNote** 7.3.2

**LazyData** true

**NeedsCompilation** no

**Author** Mark Clements [aut, cre] (ORCID: <https://orcid.org/0000-0003-4518-5670>), Enoch Yi-Tung Chen [ctb] (ORCID: <https://orcid.org/0000-0003-2448-708X>)

**Repository** CRAN

**Date/Publication** 2026-01-20 21:30:02 UTC

## Contents

---

`active_hepatitis` *Chronic active hepatitis*

---

## Description

Clinical trial of 44 patients with chronic active hepatitis randomised to either the drug prednisolone or an untreated control group.

## Usage

```
active_hepatitis
```

## Format

A data frame with 44 rows and 3 variables:

`treatment` integer treatment (1=prednisolone, 2=control)

`time` integer survival time from admission to study (months)

`status` integer event indicator (1=event, 0=right censored)

## Details

See Collett (2023)

---

`bcancer` *Prognosis for women with breast cancer*

---

## Description

For female breast cancer patients from Middlesex Hospital. The dataset includes the result of staining using Helix pomatia agglutinin (HPA).

## Usage

```
bcancer
```

## Format

A data frame with 45 rows and 3 variables:

`stain` integer for negative staining (=1) or positive staining (=2)

`time` integer time in months for survival

`status` integer for status at end of follow-up (0=censored, 1=death)

## Details

For details about the study design, see Leathem and Brooks (1987).

The dataset is described in Example 1.2 and Table 1.2 (Collett, 2023, pages 6-7).

## References

Leathem AJ, Brooks S. Predictive value of lectin binding on breast-cancer recurrence and survival. The Lancet. 1987 May 9;329(8541):1054-6. doi:10.1016/S01406736(87)90482X

## Examples

```
library(survival)
plot(survfit(Surv(time,status)~stain, data=bcancer), col=1:2, xlab="Survival time (months)",
ylab="Survival")
legend("topright", legend=c("Negative staining","Positive staining"), col=1:2, lty=1,
bty="n")
```

## bladder *Recurrence of bladder cancer*

### Description

Placebo controlled trial of bladder cancer patients randomised to thiopeta or to placebo

### Usage

```
bladder
```

### Format

A data frame with 86 rows and 6 variables:

`patient` integer patient number (1-86)

`time` integer survival time in months

`status` integer status of patient (0=censored, 1=recurrence)

`treat` integer treatment group (1=placebo, 2=thiotepa)

`init` integer initial number of tumours

`size` integer diameter of larger initial tumour in cm

### Details

See Collett (2023)

## bone_marrow *Bone marrow transplantation*

### Description

A study of 37 patients with leukaemia in complete remission who received a non-depleted allogenic bone marrow transplant.

### Usage

```
bone_marrow
```

### Format

A data frame with 37 rows and 9 variables:

`patient` integer patient number (1-37)
`time` integer survival time in days
`status` integer status of patient (0=alive, 1=dead)
`rage` integer age of patient in years
`dage` integer age of donor in years
`type` integer type of leukaemia (1=AML, 2=ALL, 3=CML)
`preg` integer Donor pregnancy (0=no, 1=yes)
`index` double index of cell-lymphocyte reactions
`gvhd` integer graft-versus-host disease (0=no, 1=yes)

### Details

See Collett (2023)

---

## bone_marrow_tx *Patient outcome following bone marrow transplantation*

---

### Description

Patient outcome following bone marrow transplantation

### Usage

```
bone_marrow_tx
```

### Format

A data frame with 2204 rows and 9 variables:

`id` integer patient id
`leukaemia` character type of leukaemia (CML,ALL,AML)
`age` character age group of patient in years (<=20, 21-40, >40))
`match` integer indicator for whether there was a donor gender match (0=no, 1=yes)
`tcell` integer indicator for whether there was T-cell depletion (1=yes, n=no)
`ptime` integer time to platelet recovery (days)
`pcens` integer event indicator for platelet recovery (1=event, 0=censored)
`rdtime` integer time to relapse of death (days)
`rdcens` integer event indicator for relapse or death (1=event, 0=censored)

### Details

See Collett (2023)

## breast_rfs *Recurrence free survival in breast cancer patients*

### Description

Recurrence free survival in breast cancer patients

### Usage

```
breast_rfs
```

### Format

A data frame with 686 rows and 11 variables:

`id` integer patient id

`treat` integer hormonal treatment (0=no tamoxifen, 1=tamoxifen)

`age` integer patient age (years)

`men` integer menopausal status (1=premenopausal, 2=postmenopausal)

`size` integer tumour size (mm)

`grade` integer tumour grade (1,2,3)

`nodes` integer number of positive pymph nodes

`prog` integer progesterone receptor status (femtomoles)

`oest` integer oestrogen receptor status (femtomoles)

`time` integer recurrence-free survival time (days)

`status` integer event indicator (0=censored, 1=relapse or death)

### Details

See Collett (2023)

## Datasets *Datasets*

## Description

The datasets are based on the official .zip file. A table for the dataset names and file names sorted by file name is here:

| Dataset name | File name |
|---|---|
| illustration | "A numerical illustration.dat" |
| leukaemia | "Bone marrow transplantation in the treatment of leukaemia.dat" |
| bone_marrow | "Bone marrow transplantation.dat" |
| ovarian | "Chemotherapy in ovarian cancer patients.dat" |
| active_hepatitis | "Chronic active hepatitis.dat" |
| granulomatous | "Chronic granulomatous disease.dat" |
| tamoxifen | "Clinical trial of tamoxifen in breast cancer patients.dat" |
| prostatic | "Comparison of two treatments for prostatic cancer.dat" |
| kidneytx | "Comparisons between kidney transplant centres.dat" |
| liverbase | "Data from a cirrhosis study (baseline).dat" |
| liver_counting | "Data from a cirrhosis study (in counting process format).dat" |
| lbrdata0 | "Data from a cirrhosis study (lbr data).dat" |
| HELP | "Health evaluation and linkage to primary care.dat" |
| dialysis | "Infection in patients on dialysis.dat" |
| bone_marrow_tx | "Patient outcome following bone marrow transplantation.dat" |
| bcancer | "Prognosis for women with breast cancer.dat" |
| pulmonary | "Pulmonary metastasis.dat" |
| breast_rfs | "Recurrence free survival in breast cancer patients.dat" |
| ulcer | "Recurrence of an ulcer.dat" |
| bladder | "Recurrence of bladder cancer.dat" |
| mammary | "Recurrence of mammary tumours in female rats.dat" |
| valve | "Survival following aortic valve replacement.dat" |
| tplant | "Survival following kidney transplantation.dat" |
| ducks | "Survival of black ducks.dat" |
| mice | "Survival of laboratory mice.dat" |
| liver | "Survival of liver transplant recipients.dat" |
| myeloma | "Survival of multiple myeloma patients.dat" |
| lung | "Survival of patients registered for a lung transplant.dat" |
| gcancer | "Survival of patients with gastric cancer.dat" |
| melanoma | "Survival times of patients with melanoma .dat" |
| livertx | "Time to death while waiting for a liver transplant.dat" |
| IUD | "Time to discontinuation of the use of an IUD.dat" |
| kidney | "Treatment of hypernephroma.dat" |

And now sorted by the dataset names:

| Dataset name | File name |
|---|---|
| active_hepatitis | "Chronic active hepatitis.dat" |
| bcancer | "Prognosis for women with breast cancer.dat" |
| bladder | "Recurrence of bladder cancer.dat" |

| | |
|---|---|
| bone_marrow | "Bone marrow transplantation.dat" |
| bone_marrow_tx | "Patient outcome following bone marrow transplantation.dat" |
| breast_rfs | "Recurrence free survival in breast cancer patients.dat" |
| dialysis | "Infection in patients on dialysis.dat" |
| ducks | "Survival of black ducks.dat" |
| gcancer | "Survival of patients with gastric cancer.dat" |
| granulomatous | "Chronic granulomatous disease.dat" |
| HELP | "Health evaluation and linkage to primary care.dat" |
| illustration | "A numerical illustration.dat" |
| IUD | "Time to discontinuation of the use of an IUD.dat" |
| kidney | "Treatment of hypernephroma.dat" |
| kidneytx | "Comparisons between kidney transplant centres.dat" |
| lbrdata0 | "Data from a cirrhosis study (lbr data).dat" |
| leukaemia | "Bone marrow transplantation in the treatment of leukaemia.dat" |
| liver | "Survival of liver transplant recipients.dat" |
| liver_counting | "Data from a cirrhosis study (in counting process format).dat" |
| liverbase | "Data from a cirrhosis study (baseline).dat" |
| livertx | "Time to death while waiting for a liver transplant.dat" |
| lung | "Survival of patients registered for a lung transplant.dat" |
| mammary | "Recurrence of mammary tumours in female rats.dat" |
| melanoma | "Survival times of patients with melanoma .dat" |
| mice | "Survival of laboratory mice.dat" |
| myeloma | "Survival of multiple myeloma patients.dat" |
| ovarian | "Chemotherapy in ovarian cancer patients.dat" |
| prostatic | "Comparison of two treatments for prostatic cancer.dat" |
| pulmonary | "Pulmonary metastasis.dat" |
| tamoxifen | "Clinical trial of tamoxifen in breast cancer patients.dat" |
| tplant | "Survival following kidney transplantation.dat" |
| ulcer | "Recurrence of an ulcer.dat" |
| valve | "Survival following aortic valve replacement.dat" |

As an alternative to using the R datasets, the `collett_data` function allows for reading from the original .dat files that are stored in the package.

## Usage

```
collett_data(name)
```

## Arguments

| | |
|---|---|
| `name` | Character string with the original filename |

## Value

A data-frame

### Author(s)

**Maintainer**: Mark Clements <mark.clements@ki.se> (ORCID)

Other contributors:

- Enoch Yi-Tung Chen <enoch.yitung.chen@ki.se> (ORCID) [contributor]

### Source

https://s3-eu-west-1.amazonaws.com/s3-euw1-ap-pe-ws4-cws-documents.ri-prod/9781032252858/Data%20sets%20from%20Modelling%20Survival%20Data%20in%20Medical%20Research%2C%204th%20edition.zip

### See Also

Useful links:

- https://github.com/mclements/collett
- Report bugs at https://github.com/mclements/collett/issues

### Examples

```
head(collett_data("A numerical illustration.dat"))
## which is equivalent to: head(illustration)
```

---

## dialysis *Infection in patients on dialysis*

---

### Description

Time from dialysis to infection for patients with diseases of the kidney.

### Usage

```
dialysis
```

### Format

A data frame with 13 rows and 5 variables:

`patient` integer patient id

`time` integer time to infection (days)

`status` integer event indicator (0=censored, 1=infection)

`age` integer age in years

`sex` integer sex of the patient (1=male, 2=female)

### Details

See Collett (2023)

## ducks *Survival of black ducks*

### Description

Black ducks, Anas rubripes, were followed the US Fish and Wildlife Service.

### Usage

```
ducks
```

### Format

A data frame with 50 rows and 6 variables:

`duck` integer duck indicator

`time` integer survival time in days

`status` integer status of bird (0=alive or missing, 1=dead)

`age` integer age group (0=hatch-year bird, 1=bird aged >= 1 year)

`weight` integer weight of bird in g

`length` integer length of wing in mm

### Details

See Collett (2023)

## gcancer *Survival of patients with gastric cancer*

### Description

Survival of patients with gastric cancer

### Usage

```
gcancer
```

### Format

A data frame with 90 rows and 4 variables:

`patient` integer patient id

`time` integer survival time in days

`status` integer event indicator (0=censored, 1=dead)

`treat` integer treatment arm (0=chemotherapy alone, 1=chemotherapy and radiotherapy)

## Details

See Collett (2023)

---

`granulomatous` *Chronic granulomatous disease*

---

## Description

Trial comparing interferon with a placebo.

## Usage

```
granulomatous
```

## Format

A data frame with 128 rows and 12 variables:

`patient` integer patient number (1-128)

`time` integer time to first infection (days)

`status` integer status of patient (0=censored, 1=infection)

`centre` integer treatment centre; see Collett (2023, page 504)

`treat` integer treatment group (0=placebo, 1=interferon)

`age` integer age in years

`sex` integer sex (1=male, 2=female)

`height` double height in cm

`weight` double weight in kg

`pattern` integer pattern of inheritance (1=X-linked, 2=autosomal recessive)

`cort` integer use of corticosteroids at trial entry (1=used, 2=not used)

`anti` integer Use of antibiotics at trial entry (1=used, 2=not used)

## Details

See Collett (2023)

## HELP *Health evaluation and linkage to primary care*

### Description

A clinical trial for patients in a residential detoxification programme. Patients were randomised to either get a referral to a HELP clinic or not.

### Usage

```
HELP
```

### Format

A data frame with 447 rows and 7 variables:

`subject` integer subject id

`days` integer time to linkage to primary care in days

`status` integer event indicator (0=no linkage, 1=linkage)

`age` integer age of patient in years

`gender` integer gender of the patient (0=female, 1=male)

`housing` integer Homelessness status (0=homeless, 1=housed)

`linkage` integer assistance to linking to healthcare (0=no, 1=yes)

### Details

Collett (2023) defines this dataset as "help", however that leads to issues with using R's help system. We have changed the dataset name to "HELP". Moreover, the book uses the variables "Time" an d"Help", whereas the dataset includes variables "days" and "linkage", respectively.

## illustration *A numerical illustration*

### Description

Artificial data on patient survival classified according to factors a and b

### Usage

```
illustration
```

## Format

A data frame with 37 rows and 4 variables:

`a` integer factor a

`b` integer factor b

`time` integer event time

`status` integer event status (1=event, 0=right censored)

## Details

See Collett (2023).

---

`IUD` *Time to discontinuation of the use of an IUD*

---

## Description

A very simple dataset showing potential right censoring for time to discontinuation of the use of an IUD.

## Usage

```
IUD
```

## Format

A data frame with 18 rows and 2 variables:

`time` integer Time in weeks to discontinuation of the use of an IUD

`status` integer Indicator for whether the IUD was discontinued: 0=No, 1=Yes

## Details

These data are reported in Table 1.1 (Collett, 2023, page 6).

## kidney *Treatment of hypernephroma*

### Description

This study was undertaken at the University of Oklahoma Health Sciences Center to investigate survival among 36 patients with a kidney tumour (hypernephroma). Standard tangent included chemotherapy and immunotherapy, with some patients also having a nephrectomy, or surgical removal of the kidney. For further details, see Lee and Wang (2013).

### Usage

```
kidney
```

### Format

A data frame with 36 rows and 4 variables:

`nephrectomy` integer indicator for nephrectomy (0=No; 1=Yes)

`age` integer age group (1=<60; 2=60-70; 3=>70

`time` integer for the follow-up time in months

`status` integer for status at the end of follow-up (1=died; 0=censored)

### References

Lee ET, Wang J. Statistical Methods for Survival Data Analysis. New York, NY: John Wiley & Sons; 2013, fourth edition. https://www.wiley.com/en-sg/Statistical+Methods+for+Survival+Data+Analysis%252C+4th+Edition-p-9781118095027

## kidneytx *Comparisons between kidney transplant centres*

### Description

Transplant survival rates by recipients of organs from deceased donors. No event was defined as being alive with a functioning graft at the last known follow-up.

### Usage

```
kidneytx
```

## Format

A data frame with 1439 rows and 9 variables:

`patient` integer patient id

`centre` integer transplant centre (1-8)

`tsurv` integer transplant survival time (days)

`tcens` integer event indicator (0=censored, 1=transplant failure)

`dage` integer donor age (years)

`dtype` integer donor type (0=deceased following brain death, 1=circulatory death)

`rage` integer recipient age (years)

`diab` integer diabetic status (0=absent, 1=present)

`cit` double cold ischaemic time (hours)

## Details

See Collett (2023). Thirty-five patients had tsurv==0 (that is, the transplanted kidney did not function).

---

`lbrdata0` *Data from a cirrhosis study (lbr data)*

---

## Description

DATASET_DESCRIPTION

## Usage

```
lbrdata0
```

## Format

A data frame with 42 rows and 3 variables:

`patient` integer patient id

`time` integer date of measurement (days)

`lbr` double log bilirubin level

## Details

See Collett (2023)

## leukaemia *Bone marrow transplantation in the treatment of leukaemia*

### Description

Bone marrow transplantation in the treatment of leukaemia

### Usage

```
leukaemia
```

### Format

A data frame with 23 rows and 8 variables:

`patient` integer patient id

`time` integer survival time in days

`status` integer event indicator (0=alive, 1=dead)

`group` integer disease group (1=ALL, 2=low-risk AML, 3=high-risk AML)

`page` integer age of patient in years

`dage` integer age of donor in years

`precovery` integer platelet recovery indicator (0=no, 1=yes)

`ptime` character time in days to return of platelets to normal level (if precovery=1)

### Details

See Collett (2023). Note that ptime will need conversion:).

## liver *Survival of liver transplant recipients*

### Description

Survival of liver transplant recipients

### Usage

```
liver
```

## Format

A data frame with 1761 rows and 7 variables:

`patient` integer patient id

`age` integer patient age in years

`gender` integer patient gender (1=male, 2=female)

`disease` integer primary disease (1=PBC, 2=PSC, 3=ALD)

`time` integer time to event (days)

`status` integer cof>0

`cof` integer cause of graft failure (0=functioning graft, 1=rejection, 2=thrombosis, 3=recurrent disease, 4=other)

## Details

See Collett (2023)

---

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

---

## Description

Articial data

## Usage

```
liverbase
```

## Format

A data frame with 12 rows and 6 variables:

`patient` integer patient id

`time` integer survival time in days

`status` integer event indicator (0=censored, 1=uncensored)

`age` integer age of the patient (years)

`treat` integer treatment group (0=placebo, 1=Liverol)

`lbr` double logarithm of bilirubin level

## Details

See Collett (2023)

## livertx *Time to death while waiting for a liver transplant*

### Description

Investigate the time on the liver transplantation list.

### Usage

```
livertx
```

### Format

A data frame with 281 rows and 7 variables:

`patient` integer patient id

`time` integer time on the list

`status` integer event indicator (0=censored, including having a transplant, 1=died on the list)

`age` integer patient age in years

`gender` integer patient gender (1=male, 0=female)

`bmi` double body mass index (kg/m^2)

`ukeld` integer UK endstage liver disease score

### Details

See Collett (2023). A higher UKELD is associated with worse disease severity.

## liver_counting *Data from a cirrhosis study (in counting process format)*

### Description

Artificial data

### Usage

```
liver_counting
```

## Format

A data frame with 54 rows and 7 variables:

`patient` integer patient id

`start` integer start time (days)

`stop` integer stop time (days)

`status` integer event indicator (0=censored, 1=uncensored)

`treat` integer treatment group (0=placebo, 1=Liverol)

`age` integer age of the patient at start of study (years)

`lbrt` double logarithm of bilirubin level

## Details

See Collett (2023). Note that the variable for log of bilirubin differs to that for "liverbase".

---

# lung *Survival of patients registered for a lung transplant*

---

## Description

Survival of patients registered for a lung transplant

## Usage

```
lung
```

## Format

A data frame with 196 rows and 7 variables:

`patient` integer patient id

`time` integer time from registration to the earlist of removal from list, last known follow-up date, 30 April 2012, or death (days)

`status` integer event indicator (0=censored, 1=dead)

`age` integer age in years

`gender` integer gender (1=male, 2=female)

`bmi` double body mass index

`disease` integer disease (1=COPD, 2=fibrosis, 3=suppurative, 4=other)

## Details

See Collett (2023)

## mammary *Recurrence of mammary tumours in female rats*

### Description

This is an animal experiment to compare the use of retinyl acetate (related to vitamin A) across the study (treatment) to treatment with retinyl acetate to 60 days and then no further treatment (control). The female rats all had mammary tumours.

### Usage

```
mammary
```

### Format

A data frame with 254 rows and 4 variables:

`rat` integer id for each rat

`treatment` integer treatment arm indicator (1=treatment, 0=control)

`time` double follow-up time (days)

`status` integer recurrence indicator (0=no, 1=yes)

### Details

See Collett (2023)

## melanoma *Survival times of patients with melanoma*

### Description

Comparing two immunotherapy treatments for patients with melanoma

### Usage

```
melanoma
```

### Format

A data frame with 30 rows and 4 variables:

`age` integer age group (1=21-44, 2=41-60, 3=61+)

`treatment` integer treatment arm (1=BCG, 2=C. parvum)

`time` integer survival time (months)

`status` integer event indicator (0=censored, 1=dead)

## Details

See Collett (2023)

---

`mice` *Survival of laboratory mice*

---

## Description

Laboratory study of survival for two groups of mice exposed to radiation.

## Usage

```
mice
```

## Format

A data frame with 181 rows and 3 variables:

`environment` integer type of environment (1=standard, 2=germ-free)

`causeofdeath` integer cause of death (1=thymic lymphoma, 2=reticulum cell sarcoma, 3=other causes)

`time` integer survival time (days)

## Details

See Collett (2023). Note that are no censored event times.

---

`myeloma` *Survival of multiple myeloma patients*

---

## Description

Patients diagnosed with multiple myeloma who were diagnosed and treated with alkylating agents at West Virginia University Medical Center for ages 50-80 years.

## Usage

```
myeloma
```

## Format

A data frame with 48 rows and 10 variables:

`patient` integer for a patient identifier

`time` integer survival time in months

`status` integer for status at follow-up (0=Alive, 1=Dead)

`age` integer age at diagnosis in years

`sex` integer for sex of the patient (1=male, 2=female)

`bun` integer level of blood urea nitrogen at diagnosis (unit assumed to be mg/dL based on the normal range for adults reported by https://en.wikipedia.org/wiki/Blood_urea_nitrogen)

`ca` integer serum calcium at diagnosis in mg/dL

`hb` double for serum hemoglobin level at diagnosis in g/dL (equivalently, grams per 100 mL)

`pcells` integer percent of plasma cells in the bone marrow at diagnosis

`protein` integer indicator for whether or not the Bence-Jones protein was present in the urine at diagnosis (0=absent, 1=present)

## Details

Krall et al (1975) did not provide the units for all of these measurements. In their analyses, they used some data transformations: log(bun). Collett (2023) converted data from Krall et al (1975): BUN is reported by Krall and colleagues as X1=log(BUN), however the log base and unit is unclear; Krall and colleagues reported for 65 individuals, including those younger than 50 and older than 80.

## References

Krall JM, Uthoff VA, Harley JB. A step-up procedure for selecting variables associated with survival. Biometrics. 1975 Mar 1:49-57. doi:10.2307/2529709

## Examples

```
## To be completed.
```

---

# ovarian *Chemotherapy in ovarian cancer patients*

---

## Description

Trial for treatment of ovarian cancer patients comparing cyclophosphamide alone with cyclophosphamide combined with adriamycin.

## Usage

```
ovarian
```

## Format

A data frame with 26 rows and 7 variables:

`patient` integer identifer

`time` integer survival time from randomisation in days

`status` integer event indicator (0=right censored, 1=event)

`treat` integer treatment (1=single, 2=combined)

`age` integer age of patients in years

`rdisease` integer extent of residual disease (1=incomplete, 2=complete)

`perf` integer performance status (1=good, 2=poor)

## Details

See Collett (2023)

---

# prostatic *Comparison of two treatments for prostatic cancer*

---

## Description

Randomised controlled trial from the Veteran's Administration Cooperative Urological Research Group. Includes patients who had stage III cancers and were randomised to placebo or daily oral treatment with 1.0 mg of diethylstilbesterol (DES).

## Usage

```
prostatic
```

## Format

A data frame with 38 rows and 8 variables:

`patient` integer patient identifier

`treatment` integer treatment indicator (1=placebo; 2=daily treatment with 1.0 mg of diethylstilbesterol (DES))

`time` integ er survival time from trial entry to end of follow-up in months

`status` integer for follow-up status (0=alive or died from other causes, 1=died from prostate cancer

`age` integer age at trial entry in years

`shb` double serum hemoglobin at trial entry in g/dL

`size` integer size of the primary tumour in cm^3

`index` integer Gleason index based on histopathology

## Details

TBC.

## References

Andrews DF, Herzberg AM. Data: a collection of problems from many fields for the student and research worker. Springer Series in Statistics; Springer New York, NY; 1985. doi:10.1007/9781-461250982

---

`pulmonary` *Pulmonary metastasis*

---

## Description

A very simple dataset with no censoring

## Usage

```
pulmonary
```

## Format

A data frame with 11 rows and 1 variables:

`time` integer survival time from pulmonary metastasis to death in months

## Details

See Collett (2023)

---

`simdata` *Simulated data*

---

## Description

Simulated data with left truncated follow-up and potentially right censored outcomes.

## Usage

```
simdata
```

## Format

A data frame with 30 observations and 8 variables:

`id` integer index each individual

`trt` numeric for whether treated (1=treated; 0=not treated)

`age` integer for age in years

`entry_time` numeric for year of entry

`observed_duration` numeric for years that an individual was observed

`status` integer for status at the end of follow-up (1=event, 0=censored)

`event_calendar_time` numeric hypothetical event time in calendar time

`stop_calendar_time` numeric end of follow-up in calendar time

## Examples

```
## Simulate 30 individuals survival based on Weibull distribution
set.seed(13579)
n <- 30

## Randomly assign treatment groups (15 each)
trt <- sample(c(1, 0), n, replace = TRUE)

## Randomly assign integer age 50-80 to each individual
age <- sample(50:80, n, replace = TRUE)

## Simulate true event times based on Weibull distribution
true_shape <- 3
true_scale <- 8
true_times <- rweibull(n, shape = true_shape, scale = true_scale)

## Simulate right censoring times based on exponential distribution
censoring_rate <- 0.1
censoring_times <- rexp(n, rate = censoring_rate)

## Random entry times
entry_time <- runif(n, min = 2000, max = 2010)

## Convert durations (time-on-study) to calendar times
event_calendar_time <- entry_time + true_times
censor_calendar_time <- entry_time + censoring_times

## Study end time (Administrative Censoring)
study_end_time <- 2012
study_censor_calendar_time <- rep(study_end_time, n)

## Determine the calendar time when observation ends
stop_calendar_time <- pmin(event_calendar_time, censor_calendar_time,
                           study_censor_calendar_time)

## Calculate the observed duration
```

```
observed_duration <- stop_calendar_time - entry_time

## Create tied data
observed_duration <- round(observed_duration * 2) / 2

## Determine the final status (1 if event, 0 if censored)
status <- as.integer(event_calendar_time <= pmin(censor_calendar_time, study_censor_calendar_time))

## Create a data frame
simdata <- data.frame(
  id = 1:n,
  trt, ## Treatment group
  age, ## Age at diagnosis
  entry_time, ## When they entered (Calendar time)
  observed_duration, ## Time-on-study
  status, ## Event status (1=event, 0=censored)
  event_calendar_time, ## Hypothetical event time (Calendar time)
  stop_calendar_time ## When observation ended (Calendar time)
)
## Save the data frame
## save(simdata, file = "~/src/R/collett/data/simdata.rda")
```

---

`step_s` *Utilities*

---

## Description

Given a data-frame with an "s" step function, expand the data-frame to include the steps.

Given a survfit object, return a data-frame

Given a summary.survfit object, return a data-frame

Calculates lp=tt(x). Typically, the tt function should include an intercept term (see the examples below). Note that spline terms assume that the x argument is multiplicative; moreover, the additional arguments are not passed. For other types of tt terms, the x is passed directly to the tt function together with other arguments.

## Usage

```
step_s(
  data,
  x,
  y,
  ymin,
  ymax,
  group,
  add_origin = TRUE,
  x_origin = 0,
  y_origin = 1
)
```

```
## S3 method for class 'survfit'
as.data.frame(x, row.names, optional, type = c("expanded", "plain"), ...)

## S3 method for class 'summary.survfit'
as.data.frame(x, row.names, optional, type = c("expanded", "plain"), ...)

predict_coxph_tt(object, times, type = "lp", se.fit = FALSE, x = 1, ...)

predict_coxph_tv(object, data, id)

plot_coxph_functional(
  formula,
  data,
  x = NULL,
  pch = 19,
  ylab = "Martingale residual for null model",
  smoother = c("loess", "lm"),
  smoother.formula = resi ~ xi,
  smoother.args = list(),
  points.args = list(),
  ...
)
```

## Arguments

| | |
|---|---|
| data | a dataset for evaluation of the coxph model |
| x | a numeric vector for the smoother (defaults to the 401 values between the range) |
| y | name of the y-variable (required) |
| ymin | name of the ymin variable (optional) |
| ymax | name of the ymax variable (optional) |
| group | name of a grouping variable (optional) |
| add_origin | logical for whether to add an origin to the start of each group |
| x_origin | double for the value of x at the origin |
| y_origin | double for the value of y, ymin and ymax at the origin |
| row.names | not used (in generic signature) |
| optional | not used (in generic signature) |
| type | a character for the type of prediction (currently only the linear predictor for the tt argument) |
| ... | other arguments to pass to the plot function |
| object | a coxph object |
| times | a numeric vector of times to evaluate the linear predictor |
| se.fit | a logical for whether to return the standard errors |
| id | a character for the subject id |

| | |
|---|---|
| formula | a formula with a Surv on the lhs and a single variable on the rhs |
| pch | an integer for the pch argument in the plot for the residuals |
| ylab | a character for the ylab argument in the plot |
| smoother | a character for the name of the smoother |
| smoother.formula | a formula for the smoother in terms of resi and xi |
| smoother.args | a list of arguments to pass to the smoother function |
| points.args | a list of arguments to pass to the points function |

## Value

expanded data-frame with the same names

a vector of fitted values (when se.fit=FALSE) or a data-frame with fitted and se.fit columns (when se.fit=TRUE)

an update of data with survival probabilities

invisible plot return

## Examples

```
step_s(data.frame(g=c(1,1), a=1:2, b=4:5), a, b)
step_s(data.frame(g=c(2,2), a=3:4, b=6:7), a, b)
step_s(data.frame(g=c(1,1,2,2), a=1:4, b=4:7), a, b, group=g)
library(survival)
library(tinyplot)
sfit1 = survfit(Surv(time,status)~rx, data=survival::colon,
                subset=etype==1)
with(as.data.frame(sfit1),
     tinyplot::plt(surv~time|strata,ymin=lower,ymax=upper,type="ribbon"))
library(survival)
library(tinyplot)
sfit1 = survfit(Surv(time,status)~rx, data=survival::colon,
                subset=etype==1)
with(as.data.frame(sfit1, type="expanded"),
     tinyplot::plt(surv~time|strata,ymin=lower,ymax=upper,type="ribbon"))
library(splines)
library(tinyplot)
fit1 = coxph(Surv(time,status)~tt(treat),data=breast_rfs,
             tt=function(x,t,...) x*cbind(1,t))
fit2 = coxph(Surv(time,status)~tt(treat),data=breast_rfs,
             tt=function(x,t,...) x*ns(t,df=4,intercept=TRUE))
times = seq(0,2500,len=301L)
df1 = transform(predict_coxph_tt(fit1,times,se.fit=TRUE),
                lower=exp(fitted-1.96*se.fit),
                upper=exp(fitted+1.96*se.fit),
                fitted=exp(fitted),
                model="linear",times=times)
df2 = transform(predict_coxph_tt(fit2,times,se.fit=TRUE),
                lower=exp(fitted-1.96*se.fit),
                upper=exp(fitted+1.96*se.fit),
```

```
                fitted=exp(fitted),
                model="ns",times=times)
 with(rbind(df1,df2),
      plt(fitted ~ times | model, ymin=lower, ymax=upper, type="ribbon",
          xlab="Time since diagnosis (days)",
          ylab="Hazard ratio comparing treated with untreated"))
 with(subset(breast_rfs,status==1), rug(time))
library(survival)
liver = transform(collett::liverbase, lbr=NULL)
liver = tmerge(liver, liver, id=patient, status=event(time,status))
liver = tmerge(liver,
               rbind(with(collett::liverbase, data.frame(patient,tstart=0,lbr)),
                     with(collett::lbrdata0, data.frame(patient,tstart=time,lbr))),
           id=patient, lbr = tdc(tstart,lbr))
fit3 = coxph(Surv(tstart,tstop,status)~lbr+treat,liver)
predict_coxph_tv(fit3,data=subset(liver,patient %in% c(1,7)),
                 id="patient")
library(survival)
par(mfrow=c(2,2))
plot_coxph_functional(Surv(time,status)~hb, data=collett::myeloma,
                      xlab="Value of Hb")
plot_coxph_functional(Surv(time,status)~bun, data=collett::myeloma,
                      xlab="Value of Bun")
plot_coxph_functional(Surv(time,status)~log(bun), data=collett::myeloma,
                      xlab="Value of log Bun")
```

---

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

---

## Description

Clinical trial for breast cancer patients comparing combined tamoxifen and radiotherapy with tamoxifen alone.

## Usage

```
tamoxifen
```

## Format

A data frame with 641 rows and 18 variables:

`id` integer patient identifier

`treat` integer treatment group (0=tamoxifen+radiotherapy, 1=tamoxifen)

`age` integer patient age at study entry (years)

`size` double tumour size (cm)

`hist` integer tumour histology (1=ductal, 2=lobular, 3=medullary, 4=mixed, 5=other)

`hr` integer hormone receptor level (0=negative, 1=positive)

`hb` integer Haemoglobin level (g/l)

`andis` integer axillary relapse (0=no, 1=yes)

`lsurv` integer time to local relapse or last follow-up (days)

`ls` integer local relapse (0=no, 1=yes))

`asurv` integer time to axillary relapse or last follow-up (days)

`as` integer axillary relapse (0=no, 1=yes)

`dsurv` integer Time to distant relapse or last follow-up (days)

`ds` integer distant relapse (0=no, 1=yes)

`msurv` integer time to second malignancy or last follow-up (days)

`ms` integer second malignancy (0=no, 1=yes)

`tsurv` integer time from randomisation to death or last follow-up (days)

`ts` integer status at last follw-up (0=alive, 1=dead)

## Details

See Collett (2023)

---

`tplant` *Survival following kidney transplantation*

---

## Description

Survival following kidney transplantation

## Usage

```
tplant
```

## Format

A data frame with 434 rows and 7 variables:

`patient` integer patient id

`donor` integer donoe id

`time` integer survival time in days

`status` integer event indicator (0=censored, 1=graft failure or death with a functioning graft)

`age` integer patient age (years)

`diabetes` integer diabetes status (0=absent, 1=present)

`cit` double cold ischaemic time, the time in hours between retrieval of the kidney from the donor and the transplantation

## Details

See Collett (2023)

## ulcer *Recurrence of an ulcer*

### Description

A double-blind trial comparing two treatments for ulcers. Data from Belgium.

### Usage

```
ulcer
```

### Format

A data frame with 43 rows and 6 variables:

`patient` integer patient id

`age` integer age at the end of the trial in years

`duration` integer duration of verified disease (1: <5 years, 2: >=5 years

`treatment` integer treatment arm (1=A,2=B)

`time` integer time since last visit (months)

`result` integer result of the last visit (1=no ulcer detected, 2=ulcer detected)

### Details

See Collett (2023)

## valve *Survival following aortic valve replacement*

### Description

Patients following an aortic valve replacement are measured for left ventricular mass index (LVMI).

### Usage

```
valve
```

## Format

A data frame with 988 rows and 11 variables:

`id` integer patient id

`futime` double total follow-up time from date of surgery (years)

`status` integer event indicator (0=censored, 1=death)

`time` double time of LVMI measurement after surgery (years)

`lvmi` double standardised LVMI

`age` integer age of patient in years

`sex` integer sex of patient (0=male, 1=female)

`redo` integer previous cardiac surgery (0=no, 1=yes)

`emerg` integer operative urgency (0=elective, 1=urgent or emergency)

`dm` integer preoperative diabetes mellitus (0=no, 1=yes)

`type` integer type of valve (1=human tissue, 2=porcine tissue)

## Details

See Collett (2023)

# Index