Analyzing interval-censored survival-time data in Stata

Xiao Yang

Senior Statistician and Software Developer
StataCorp LLC

2017 Stata Conference

Outline

- What is interval-censoring?
  - Motivating example
  - Introduction
- Parametric regression models
  - `stintreg` overview
  - Case I interval-censored data
  - Case II interval-censored data
- Postestimation for `stintreg`
  - Predictions
  - Survior function plots
  - Residuals and diagnostic measures
- Conclusion

Motivating example

Breast cancer study

- 94 patients with breast cancer
- Treated with either radiation therapy alone (`RT`), or radiation therapy plus adjuvant chemotherapy (`RCT`)
- Patients had different visit times and durations between visits
- Breast retraction (cosmetic deterioration) was measured at each visit
- The exact time of breast retraction was not observed and was known to fall in an interval between visits
- We want to study the effect of treatment on time (in months) to breast retraction

Motivating example cont.

id	treat	age	ltime	rtime
1	Radio	48	0	7
11	Radio	44	11	18
21	Radio	38	24	.
31	Radio	39	36	.
41	Radio	40	46	.
51	Radio+Chemo	37	5	8
61	Radio+Chemo	34	12	20
71	Radio+Chemo	29	16	24
81	Radio+Chemo	38	23	.
91	Radio+Chemo	37	35	.

What happens if interval censoring has been ignored or treated as right-censored data?

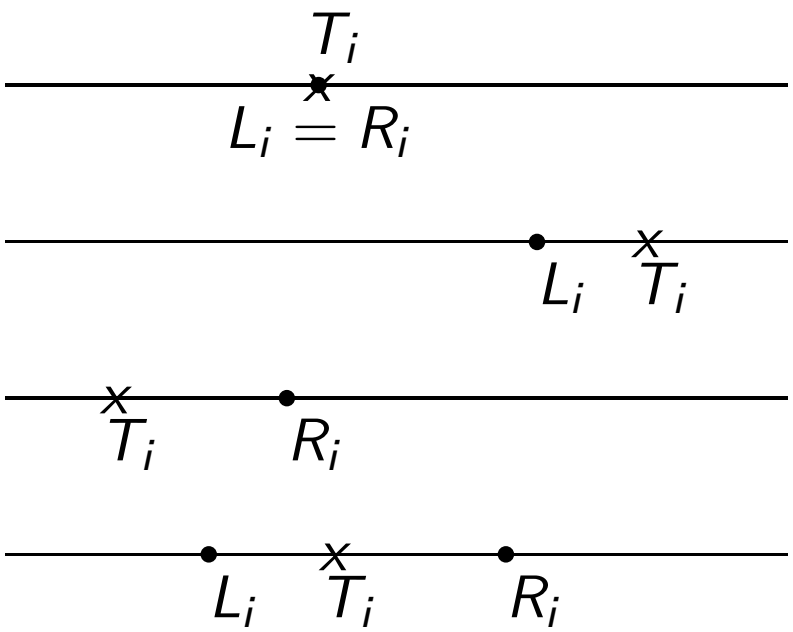
- Rucker and Messerer (1988) stated that assuming interval survival times as exact times can lead to biased estimates and underestimation of the true error variance, which may lead to false positive results.
- Law and Brookmeyer (1992) interpolated the failure time by the midpoint of the censored interval and showed that the statistical properties depend strongly on the underlying distributions and the width of the intervals. Therefore, the survival estimates may be biased and the variability of the estimates may be underestimated.

Introduction

- Suppose the event time T_i is an independent random variable with an underlying distribution function $f(t)$.
- The corresponding survival function is denoted as $S(t)$.
- Event time T_i is not always exactly observed.
- $(L_i, R_i]$ denotes the interval in which T_i is observed.
- There are three types of censoring: left-censoring, right-censoring, and interval-censoring.

Types of censoring

No censoring
$(L_i = T_i, R_i = T_i]$
Right-censoring
$(L_i, R_i = +\infty)$
Left-censoring
$(L_i = 0, R_i]$
Interval-censoring
$(L_i, R_i]$



Types of interval-censored data

- Case I interval-censored data (**current status data**):
occurs when subjects are observed only once, and we only know whether the event of interest occurred before the observed time. The observation on each subject is either left- or right-censored.
- Case II (**general**) interval-censored data:
occurs when we do not know the exact failure time T_i, but only know that the failure happened within a random time interval $(L_i, R_i]$, before the left endpoint L_i, or after the right endpoint R_i. The observation on each subject can be arbitrarily censored.

Methods for analyzing interval-censored data

- Imputation-based methods
- **Parametric regression models**
- Nonparametric maximum-likelihood estimation
- Semiparametric regression models
- Bayesian analysis
- ...

stintreg overview

stintreg fits parametric models to survival-time data, which can be uncensored, right-censored, left-censored, or interval-censored.

- Supports different distributions and parameterizations
- Fits models to two types of interval-censored data:
  - Case I interval-censored data (current status data)
  - Case II interval-censored data (general interval-censored data)
- Supports ancillary parameters and stratification
- Supports postestimation commands

Basic syntax

`stintreg` [*indepvars*], `interval`(t_l t_u) `distribution`(*distname*)

- `interval()` specifies two time variables that contain the endpoints of the censoring interval.
- `distribution()` specifies the survival model to be fit.
- `stseting` the data is not necessary and will be ignored.

Interval-censored data setup

Each subject should contain two time variables, t_l and t_u, which are the left and right endpoints of the time interval.

Type of data		t_l	t_u
uncensored data	$a = [a, a]$	a	a
interval-censored data	$(a, b]$	a	b
left-censored data	$(0, b]$	.	b
left-censored data	$(0, b]$	0	b
right-censored data	$[a, \infty)$	a	.
missing		.	.
missing		0	.

Maximum likelihood estimation

`stintreg` estimates parameters via maximum likelihood:

$$\log L = \sum_{i \in UC} \log f_i(t_{li}) + \sum_{i \in RC} \log S_i(t_{li}) + \sum_{i \in LC} \{1 - \log S_i(t_{ui})\} + \sum_{i \in IC} \{\log S_i(t_{li}) - \log S_i(t_{ui})\}$$

Supported distributions and parameterizations

`stintreg` supports six different parametric survival distributions and two parameterizations: proportional hazards (PH) and accelerated failure-time (AFT).

Distribution	Metric
Exponential	PH, AFT
Weibull	PH, AFT
Gompertz	PH
Lognormal	AFT
Loglogistic	AFT
Generalized gamma	AFT

Example of Case II interval-censored data

Time to resistance to zidovudine

- 31 AIDS patients enrolled in four clinical trials
- Resistance assays were very expensive; few assessments were performed on each patient
- Covariates of interest:
  - The stage of the disease, `stage`
  - The dose level of the treatment, `dose`
- Time interval, in months, is stored in variables `t_l` and `t_r`
- We want to investigate whether `stage` has any effect on time to drug resistance

Fit Weibull model

```
. stintreg i.stage, interval(t_l t_r) distribution(weibull)
Weibull PH regression                           Number of obs     =         31
                                                  Uncensored      =          0
                                                  Left-censored   =         15
                                                  Right-censored  =         13
                                                  Interval-cens.  =          3
                                                LR chi2(1)        =      10.02
Log likelihood =  -13.27946                     Prob > chi2       =     0.0016
```

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.stage	6.757496	4.462932	2.89	0.004	1.851897	24.65783
_cons	.0003517	.0010552	-2.65	0.008	9.82e-07	.1259497
/ln_p	1.036663	.3978289	2.61	0.009	.2569325	1.816393
p	2.819791	1.121795			1.292958	6.149638
1/p	.3546362	.1410845			.1626112	.7734204

```
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline hazard.
```

Model ancillary parameters

Assume that the hazards for different dosage levels have different shape parameters.

```
. stintreg i.stage, interval(t_l t_r) distribution(weibull) ancillary(i.dose)
note: option nohr is implied if option strata() or ancillary() is specified
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t_l						
1.stage	2.795073	1.167501	2.39	0.017	.5068139	5.083332
_cons	-10.8462	4.233065	-2.56	0.010	-19.14286	-2.549547
ln_p						
1.dose	.1655302	.0874501	1.89	0.058	-.0058689	.3369292
_cons	1.252361	.4143257	3.02	0.003	.4402972	2.064424

$\widehat{ln(p)}_{low} = 1.25$ and $\widehat{ln(p)}_{high} = 1.25 + 0.17 = 1.42$.
Thus, $\hat{p}_{low} = 3.49$ and $\hat{p}_{high} = 4.14$

Fit stratified model

A stratified model means that the coefficients on the covariates are the same across strata, but the intercept and ancillary parameters are allowed to vary for each level of the stratum variable.

You can fit the stratified model using

```
. stintreg i.stage i.dose, interval(t_l t_r)
distribution(weibull) ancillary(i.dose)
```

or, more conveniently, using

```
. stintreg i.stage, interval(t_l t_r) distribution(weibull)
strata(i.dose)
```

Fit stratified model

```
. stintreg i.stage, interval(t_l t_r) distribution(weibull) strata(dose)
note: option nohr is implied if option strata() or ancillary() is specified
Weibull PH regression                           Number of obs     =         31
                                                   Uncensored     =          0
                                                   Left-censored  =         15
                                                   Right-censored =         13
                                                   Interval-cens. =          3
                                                LR chi2(2)        =      12.40
Log likelihood = -11.115197                     Prob > chi2       =     0.0020
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t_l						
1.stage	2.711532	1.084146	2.50	0.012	.5866456	4.836419
1.dose	-2.661872	5.883967	-0.45	0.651	-14.19424	8.870492
_cons	-9.143003	4.930789	-1.85	0.064	-18.80717	.5211664
ln_p						
1.dose	.453894	.670098	0.68	0.498	-.8594739	1.767262
_cons	1.051935	.6190537	1.70	0.089	-.1613879	2.265258

Example of Case I interval-censored data

Nonlethal lung tumor

- 144 male mice in a tumorigenicity experiment
- two groups: conventional environment (CE) or germ-free environment (GE)
- Lung tumors are known to be nonlethal for the mice
- Consists of the death time and indicator of lung tumor presence
- Time to tumor onset is of interest but not directly observed

Data setup

- Conventional storage: observation times and an indicator of whether the event of interest occured by the observation time.

```
. list in 26/30
```

	group	status	death
26.	CE	With tumor	811
27.	CE	With tumor	839
28.	CE	No tumor	45
29.	CE	No tumor	198
30.	CE	No tumor	215

Data setup

- `stintreg` requires two time variables:

```
. generate ltime = death
. generate rtime = death
. replace ltime = . if status == 1
(62 real changes made, 62 to missing)
. replace rtime = . if status == 0
(82 real changes made, 82 to missing)
. list in 26/30
```

	group	status	death	ltime	rtime
26.	CE	With tumor	811	.	811
27.	CE	With tumor	839	.	839
28.	CE	No tumor	45	45	.
29.	CE	No tumor	198	198	.
30.	CE	No tumor	215	215	.

Fit exponential PH model

```
. stintreg i.group, interval(ltime rtime) distribution(exponential)
Exponential PH regression                     Number of obs     =        144
                                                 Uncensored     =          0
                                                 Left-censored  =         62
                                                 Right-censored =         82
                                                 Interval-cens. =          0
                                              LR chi2(1)        =      16.09
Log likelihood = -81.325875                   Prob > chi2       =     0.0001
```

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group						
GE	2.90202	.7728318	4.00	0.000	1.721942	4.890828
_cons	.0005664	.0001096	-38.63	0.000	.0003876	.0008277

Note: _cons estimates baseline hazard.

The estimated hazard for the mice in GE is approximately three times the hazard for the mice in CE.

Fit exponential AFT model

```
. stintreg i.group, interval(ltime rtime) distribution(exponential) time
Exponential AFT regression                      Number of obs     =        144
                                                   Uncensored     =          0
                                                   Left-censored  =         62
                                                   Right-censored =         82
                                                   Interval-cens. =          0
                                                LR chi2(1)        =      16.09
Log likelihood = -81.325875                     Prob > chi2       =     0.0001
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group						
GE	-1.065407	.2663082	-4.00	0.000	-1.587362	-.5434525
_cons	7.476278	.1935597	38.63	0.000	7.096908	7.855648

The survival time for the mice in GE is 66% ($e^{-1.07} = 0.34$) shorter than the survival time for the mice in CE.

Postestimation overview

`stintreg` provides several postestimation features after estimation:

- Predictions of survival time, hazard, and scores
- Plots for survivor, hazard, and cumulative hazard function
- Prediction of residuals and diagnostic measures

Returning to our motivating example

```
. stintreg i.treat, interval(ltime rtime) distribution(weibull)
Weibull PH regression                           Number of obs     =         94
                                                  Uncensored      =          0
                                                  Left-censored   =          5
                                                  Right-censored  =         38
                                                  Interval-cens.  =         51
                                                LR chi2(1)        =      10.93
Log likelihood = -143.19228                     Prob > chi2       =     0.0009
```

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
treat						
Radio+Chemo	2.498526	.7069467	3.24	0.001	1.434961	4.350383
_cons	.0018503	.0013452	-8.66	0.000	.000445	.007693
/ln_p	.4785787	.1198973	3.99	0.000	.2435843	.713573
p	1.613779	.1934877			1.275814	2.041272
1/p	.6196635	.074296			.4898907	.7838134

Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline hazard.

Using predict after stintreg

- What is the median survival time?

```
. predict time, median time
. tabulate treat, summarize(time) means freq
```

Treatment	Summary of Predicted median for (ltime,rtime] Mean	Freq.
Radio	39.332397	46
Radio+Che	22.300791	48
Total	30.635407	94

Obtain survivor probabilities

- Estimates of survivor probabilities (as well as hazard estimates and Cox-Snell residuals) are intervals.
- We need to specify two new variable names in `predict`.

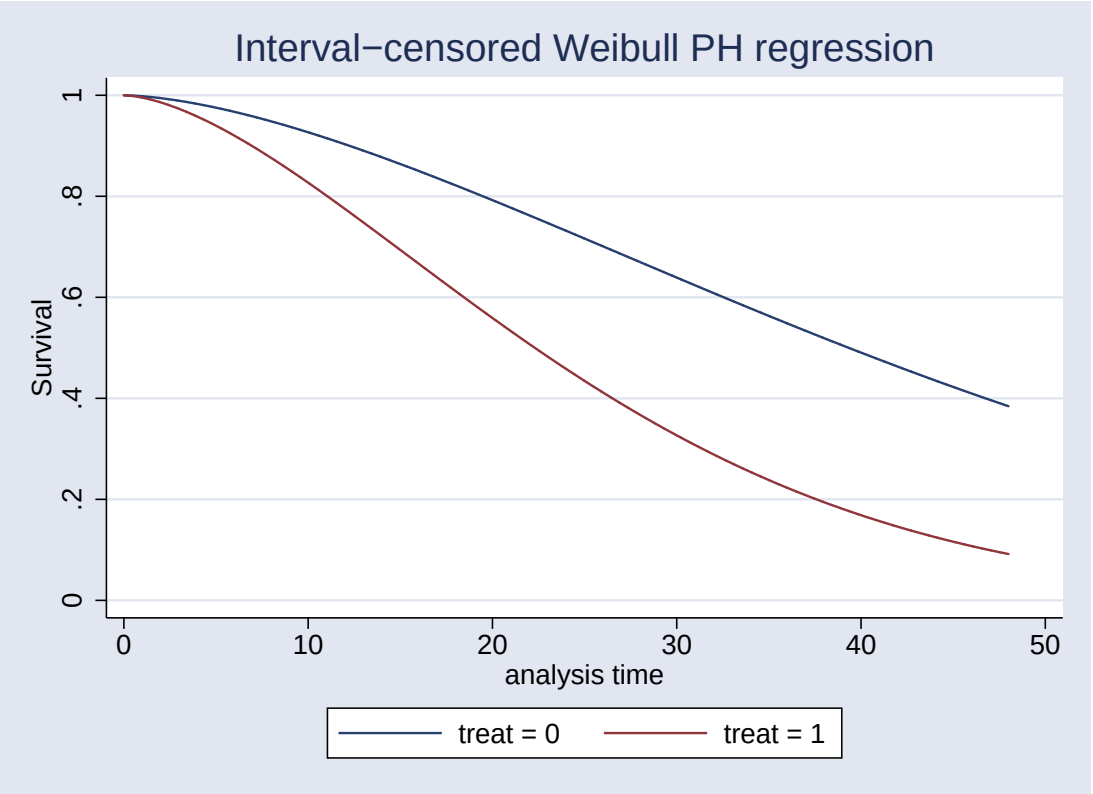
```
. predict surv_l surv_u, surv
. list surv_l surv_u in 1/5
```

	surv_l	surv_u
1.	1	.95814
2.	1	.948338
3.	1	.9754614
4.	.9828176	.9151379
5.	.9754614	.9029849

Plot survivor function

- Do RCT (`treat` = 1) patients experience breast retraction earlier than RT (`treat` = 0) patients?

```
. stcurve, survival at1(treat = 0) at2(treat = 1)
```



Residuals and diagnostic measures

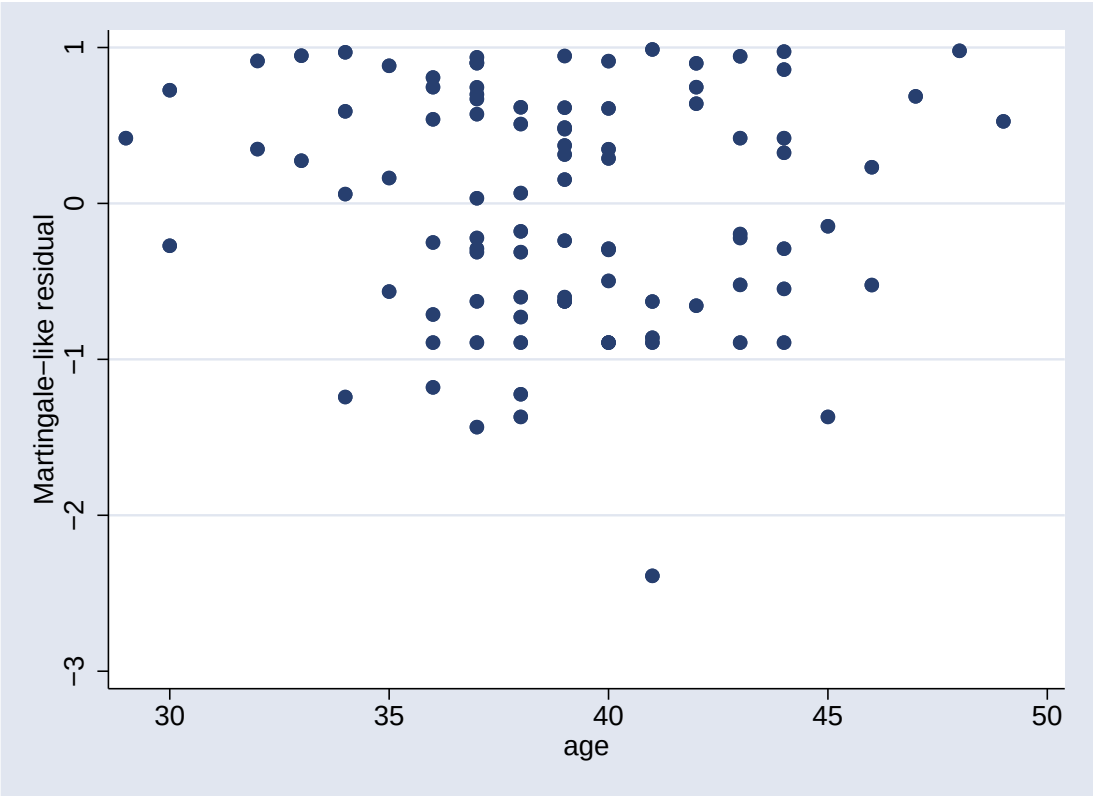
`stintreg` provides two types of residuals to assess the appropriateness of the fitted models.

- Martingale-like residuals:
  - to examine the functional form of covariates
  - to assess whether additional covariates are needed
  - to identify outliers
- Cox-Snell residuals: to assess the overall model fit

Check whether additional covariates are needed

- Should the patient's age be included in the model?

```
. predict mg, mgale
. scatter mg age
```

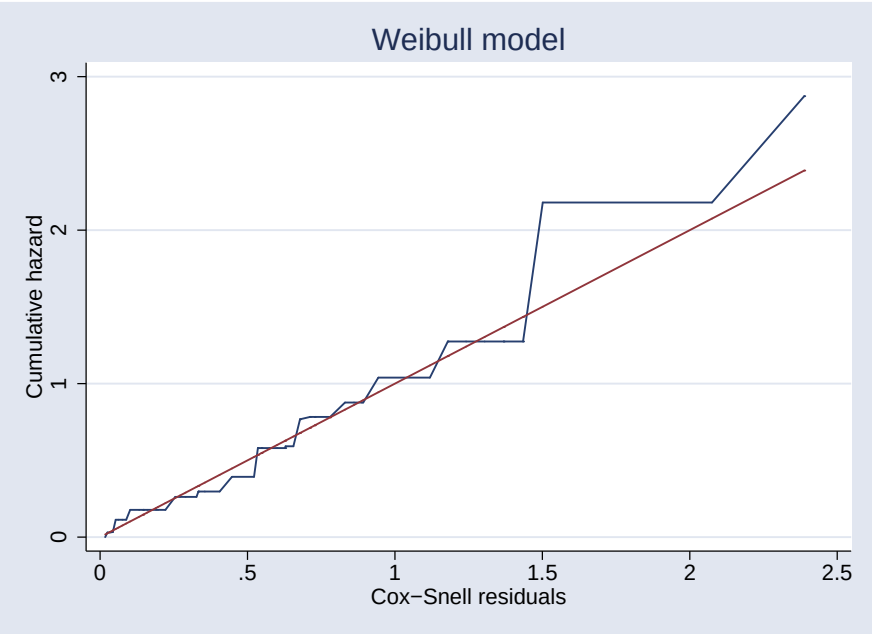


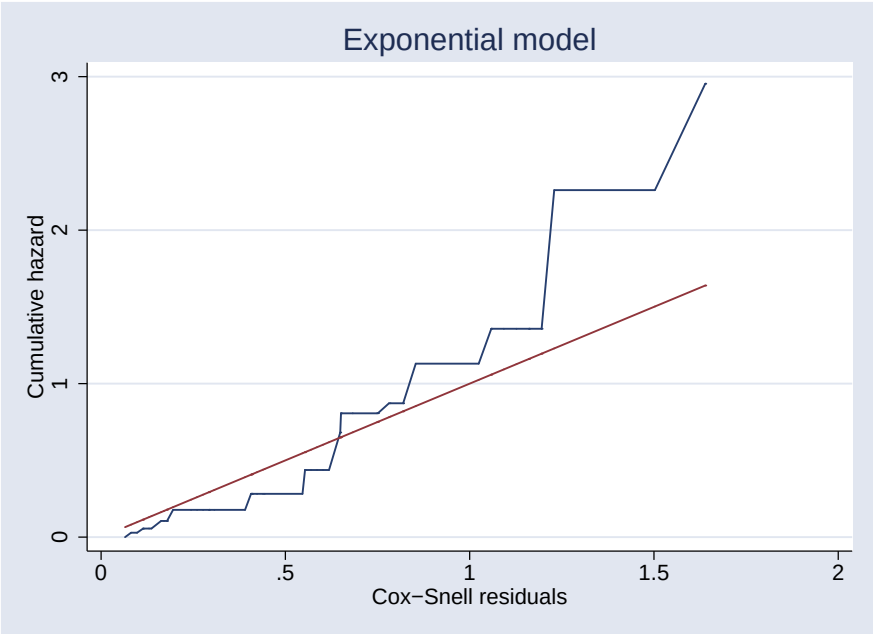
Goodness-of-fit plot

- `estat gofplot` is used to assess the goodness-of-fit of the model visually; available as of the 20170720 update.
- It plots the Cox-Snell residuals versus the estimated cumulative hazard function corresponding to these residuals.
- The estimated cumulative hazards are calculated using the self-consistency algorithm proposed by Turnbull (1976).
- The Cox-Snell residuals form the 45° reference line. If the model fits the data well, the plotted estimated cumulative hazards should be close to the reference line.

Goodness-of-fit plot

- Does the Weibull model fit the data better than the exponential model?





Conclusions

- The models fit by `stintreg` are generalizations of the models fit by `streg` to support interval-censored data.
- A main advantage of parametric approaches is that their implementation is straightforward and standard maximum likelihood theory generally applied.
- They provide attractive choices in particular if censored intervals are very wide and/or sample sizes are small, resulting in very limited information about survival variables of interest.

References

[1] C. C. Law and R. Brookmeyer. "Effects of mid-point imputation on the analysis of doubly censored data". In: *Statistics in Medicine* 11 (1992), pp. 1569–1587.

[2] G. Rucker and D. Messerer. "Remission duration: an example of interval-censored observations". In: *Statistics in Medicine* 7 (1988), pp. 1139–1145.

[3] B. W. Turnbull. "The empirical distribution function with arbitrarily grouped censored and truncated data". In: *Journal of the Royal Statistical Society, Series B* 38 (1976), pp. 290–295.