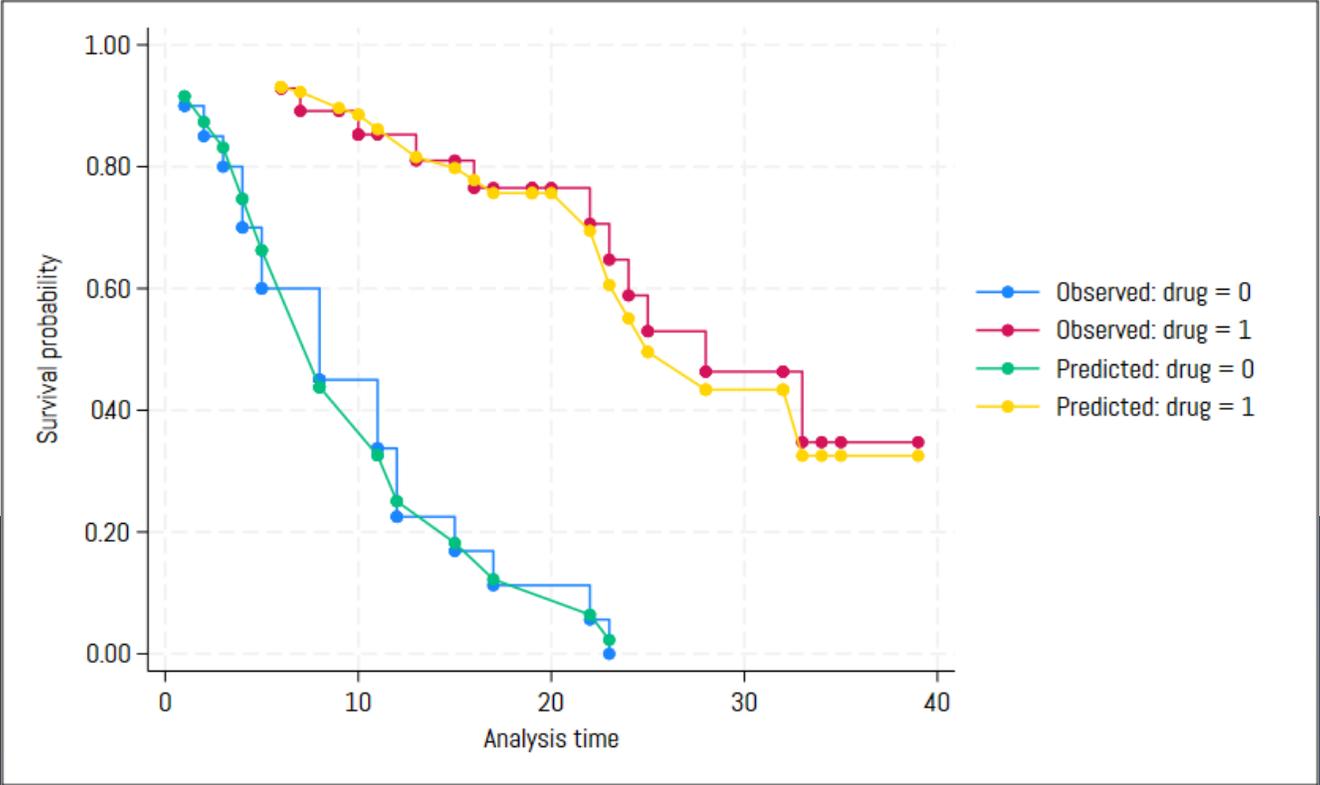


Survival analysis using Stata

Gabriela Ortiz
December 4, 2024



Overview

- Introduction to survival-time data
- Summary statistics
- Exploratory graphs
- Estimation
 - Semiparametric and parametric models
 - Predictions
- Diagnostics
 - Goodness-of-fit plots
 - Testing assumptions

Introduction to survival data

Survival-time data

- We measure time to an event of interest
- The occurrence of the event is typically called a failure
- An observation is censored if we don't know the exact time of failure
- Survival-time data is present in many fields
 - Health
 - Economics
 - Business
 - Criminology
- Stata's `st` suite of commands is designed for analyzing survival-time data

A look at survival data

One record per patient

Patient ID	Sex	Days	Died
1	Male	89	Yes
2	Female	91	No
3	Male	90	Yes

A look at survival data

One record per patient			
Patient ID	Sex	Days	Died
1	Male	89	Yes
2	Female	91	No ●
3	Male	90	Yes



The patient's time of death is right-censored if they survive until the end of the study.

Single- vs. multiple-record data

One record per patient			
Patient ID	Sex	Days	Died
1	Male	89	Yes
2	Female	91	No
3	Male	90	Yes

Two records per patient			
Patient ID	Sex	Days	Died
1	Male	33	No
1	Male	89	Yes
2	Female	33	No
2	Female	91	No
3	Male	32	No
3	Male	90	Yes

Final notes on survival data

- There are other varieties
 - A subject might be diagnosed before the study starts, meaning they are at risk before we observe them (delayed entry).
 - There might be a gap between the time the subject entered the study and the time the study ended. Suppose the patient was traveling and unable to be reached for a month in the middle of the study but returned before the study ended.
 - You might have multiple-failure data.
- We won't be focusing on these types of complications, but Stata's commands for analyzing survival-time data accommodate data with these features.

A first example

A look at survival-time data

```
. webuse stan3, clear  
(Heart transplant data)
```

```
. list id year age t1 surger transplant posttran died in 99/104, sepby(id)
```

	id	year	age	t1	surgery	transp~t	posttran	died
99.	61	71	52	2	0	0	0	1
100.	62	71	39	69	0	0	0	1
101.	63	71	32	27	0	1	0	0
102.	63	71	32	841	0	1	1	0
103.	64	72	48	32	1	1	0	0
104.	64	72	48	583	1	1	1	1

Before using Stata's `st` commands, we need to `stset` the data.

Declare data to be survival-time data

```
. stset t1, failure(died) id(id)
```

Survival-time data settings

```
      ID variable: id  
      Failure event: died!=0 & died<.  
Observed time interval: (t1[_n-1], t1]  
Exit on or before: failure
```

```
172 total observations  
  0 exclusions
```

```
172 observations remaining, representing  
103 subjects  
 75 failures in single-failure-per-subject data  
31,938.1 total analysis time at risk and under observation  
                                     At risk from t =           0  
Earliest observed entry t =           0  
Last observed exit t =           1,799
```

Describe survival-time data

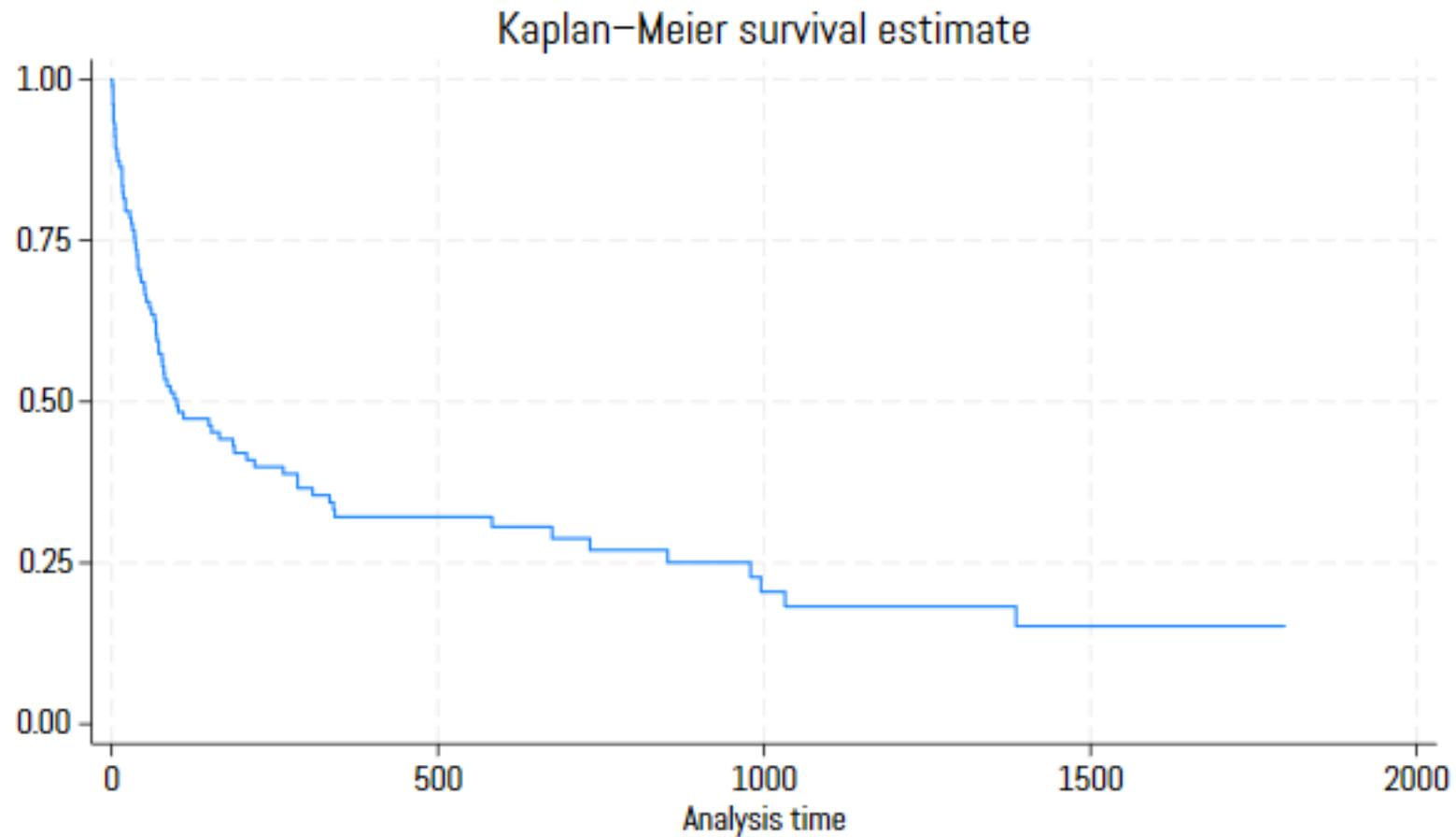
```
. stdescribe
```

```
      Failure _d: died  
Analysis time _t: t1  
      ID variable: id
```

Category	Total	Per subject			
		Mean	Min	Median	Max
Number of subjects	103				
Number of records	172	1.669903	1	2	2
Entry time (first)		0	0	0	0
Exit time (final)		310.0786	1	90	1799
Subjects with gap	0				
Time on gap	0
Time at risk	31938.1	310.0786	1	90	1799
Failures	75	.7281553	0	1	1

Kaplan–Meier survivor function

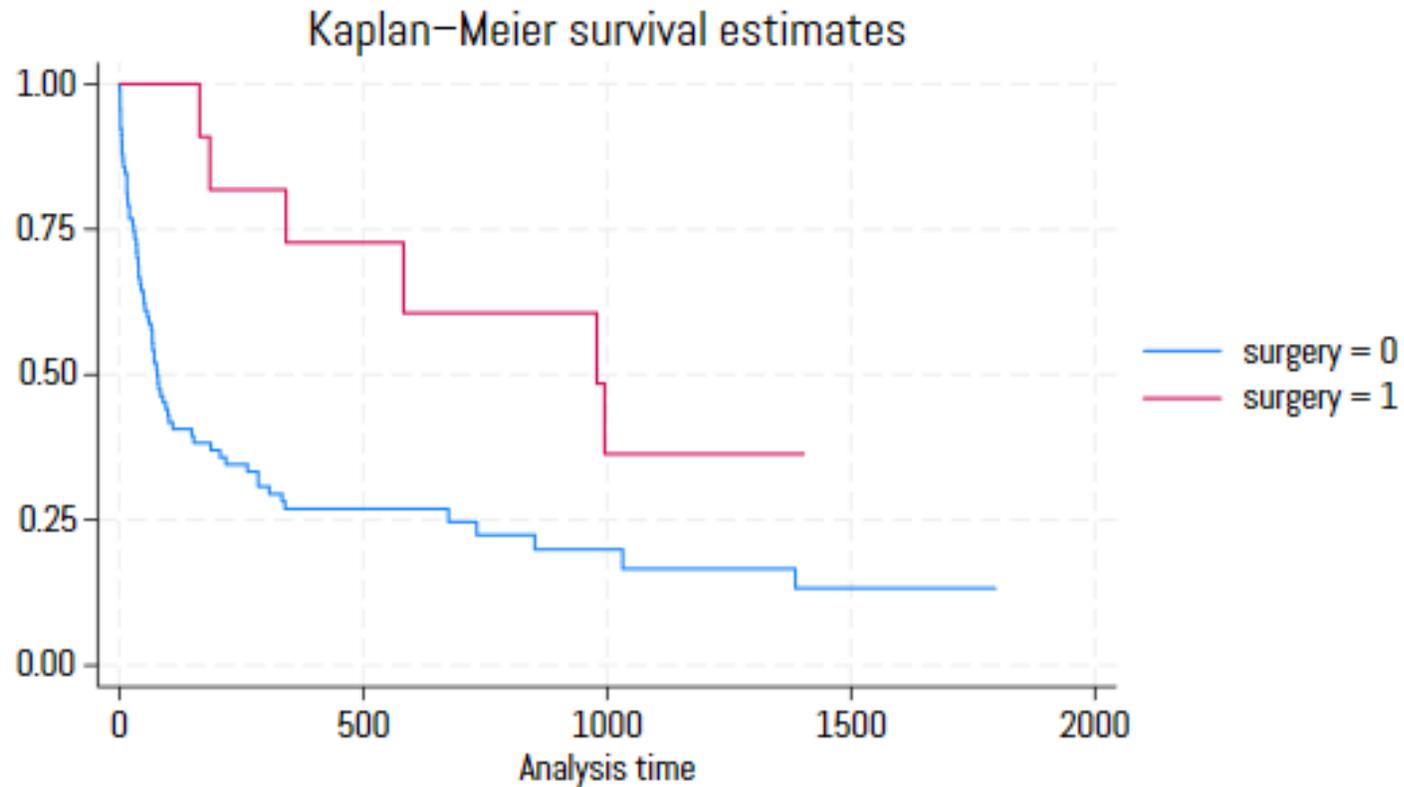
. sts graph



$$S(t) = \Pr(T > t)$$

Kaplan–Meier survivor function by group

```
. sts graph, by(surgery) risktable
```



Number at risk		0	500	1000	1500	2000
surgery = 0	91	16	6	3	0	0
surgery = 1	12	7	3	0	0	0

Confidence interval for median survival time

```
. stci
```

```
      Failure _d: died  
Analysis time _t: t1  
ID variable: id
```

	Number of subjects	50%	Std. err.	[95% conf. interval]	
Total	103	100	38.64425	69	219

Confidence interval by group

```
. stci, by(posttran)
```

```
Failure _d: died  
Analysis time _t: t1  
ID variable: id
```

posttran	Number of subjects	50%	Std. err.	[95% conf. interval]	
0	103	149	43.81077	69	340
1	69	96	58.71712	45	285
Total	103	100	38.64425	69	219

Summary statistics

```
. stsum, by(posttran)
```

```
Failure _d: died
```

```
Analysis time _t: t1
```

```
ID variable: id
```

posttran	Time at risk	Incidence rate	Number of subjects	Survival time		
				25%	50%	75%
0	5,936	.0050539	103	36	149	340
1	26,002.1	.0017306	69	39	96	979
Total	31,938.1	.0023483	103	36	100	979

Other statistics

- Incidence rates
 - Obtain estimates and confidence intervals for the incidence-rate ratio (IRR) and incidence-rate difference. See [\[ST\] stir](#).
 - Obtain person-time and incidence rate. Also, merge with standard-rate data to obtain SMRs. See [\[ST\] stptime](#).
- Failure rates
 - Tabulate failure rates by multiple categorical variables
 - Obtain stratified rate ratios
 - Carry out trend tests
 - See [\[ST\] strate](#).
- Life tables
 - Life, cumulative failure, and hazard tables
 - Graph survival rate and corresponding confidence interval
 - See [\[ST\] ltable](#).

Test equality of survivor functions

```
. sts test posttran
```

```
      Failure _d: died  
Analysis time _t: t1  
      ID variable: id
```

```
Equality of survivor functions  
Log-rank test
```

posttran	Observed events	Expected events
0	30	31.20
1	45	43.80
Total	75	75.00

```
chi2(1) = 0.13
```

```
Pr>chi2 = 0.7225
```

Cox proportional hazards model

Single-observation survival-time data

```
. webuse drugtr, clear  
(Patient survival in drug trial)
```

```
. describe studytime-age
```

Variable name	Storage type	Display format	Value label	Variable label
studytime	byte	%8.0g		Months to death or end of exp.
died	byte	%8.0g		1 if patient died
drug	byte	%8.0g		Drug type (0=placebo)
age	byte	%8.0g		Patient's age at start of exp.

Display survival-time settings

```
. stset
```

```
-> stset studytime, failure(died)
```

```
Survival-time data settings
```

```
Failure event: died!=0 & died<.
```

```
Observed time interval: (0, studytime]
```

```
Exit on or before: failure
```

```
48 total observations
```

```
0 exclusions
```

```
48 observations remaining, representing
```

```
31 failures in single-record/single-failure data
```

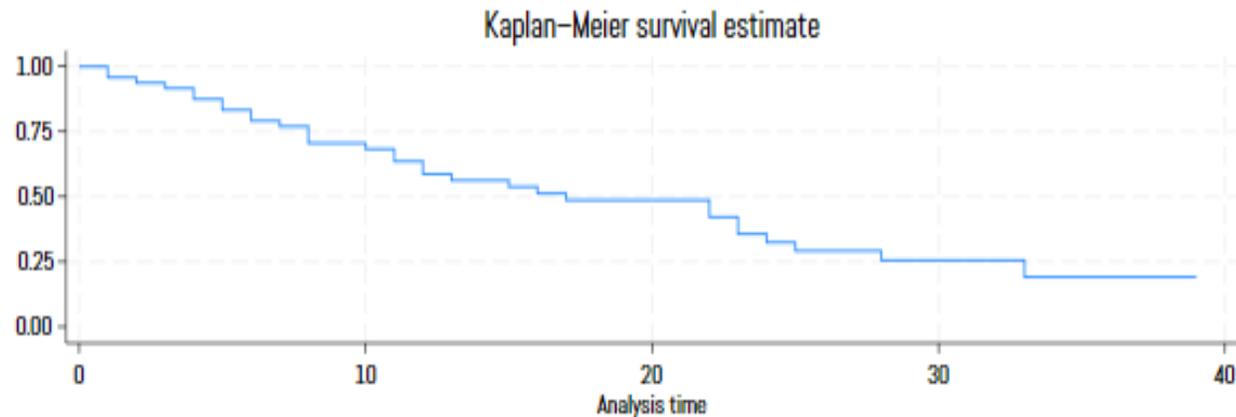
```
744 total analysis time at risk and under observation
```

```
At risk from t = 0
```

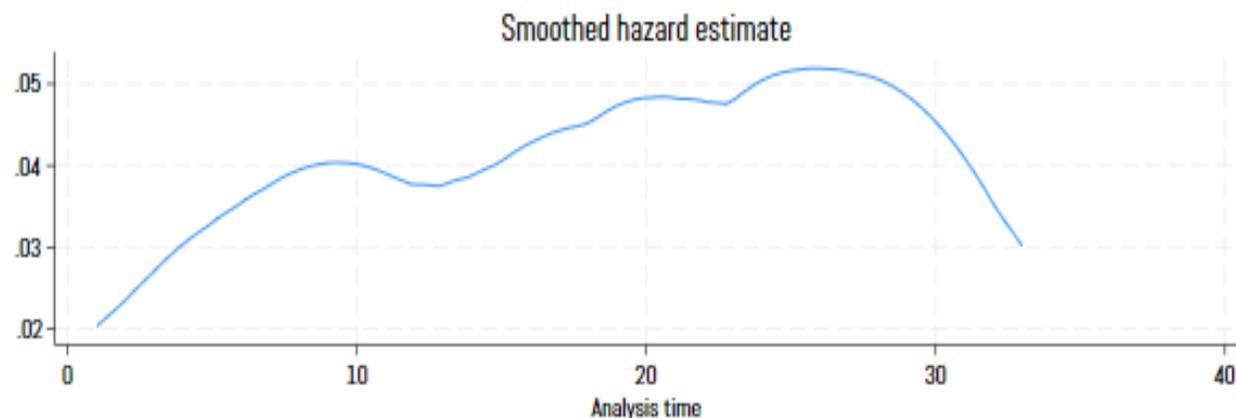
```
Earliest observed entry t = 0
```

```
Last observed exit t = 39
```

Survivor and hazard functions



Probability of surviving
beyond time t



Hazard of failing at time t

- . sts graph, surv saving(survival)
- . sts graph, hazard nob saving(hazard)
- . graph combine survival hazard

Cox proportional hazards model

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)$$

where $h_0(t)$ is the baseline hazard

- The hazard depends on the covariates; we estimate their coefficients (β_k).
- We assume the hazard ratio ($\exp(\beta_k)$) is fixed over time.

Cox proportional hazards model

```
. stcox drug age, nolog
```

```
Failure _d: died
```

```
Analysis time _t: studytime
```

Cox regression with Breslow method for ties

```
No. of subjects = 48
```

```
Number of obs = 48
```

```
No. of failures = 31
```

```
Time at risk = 744
```

```
LR chi2(2) = 33.18
```

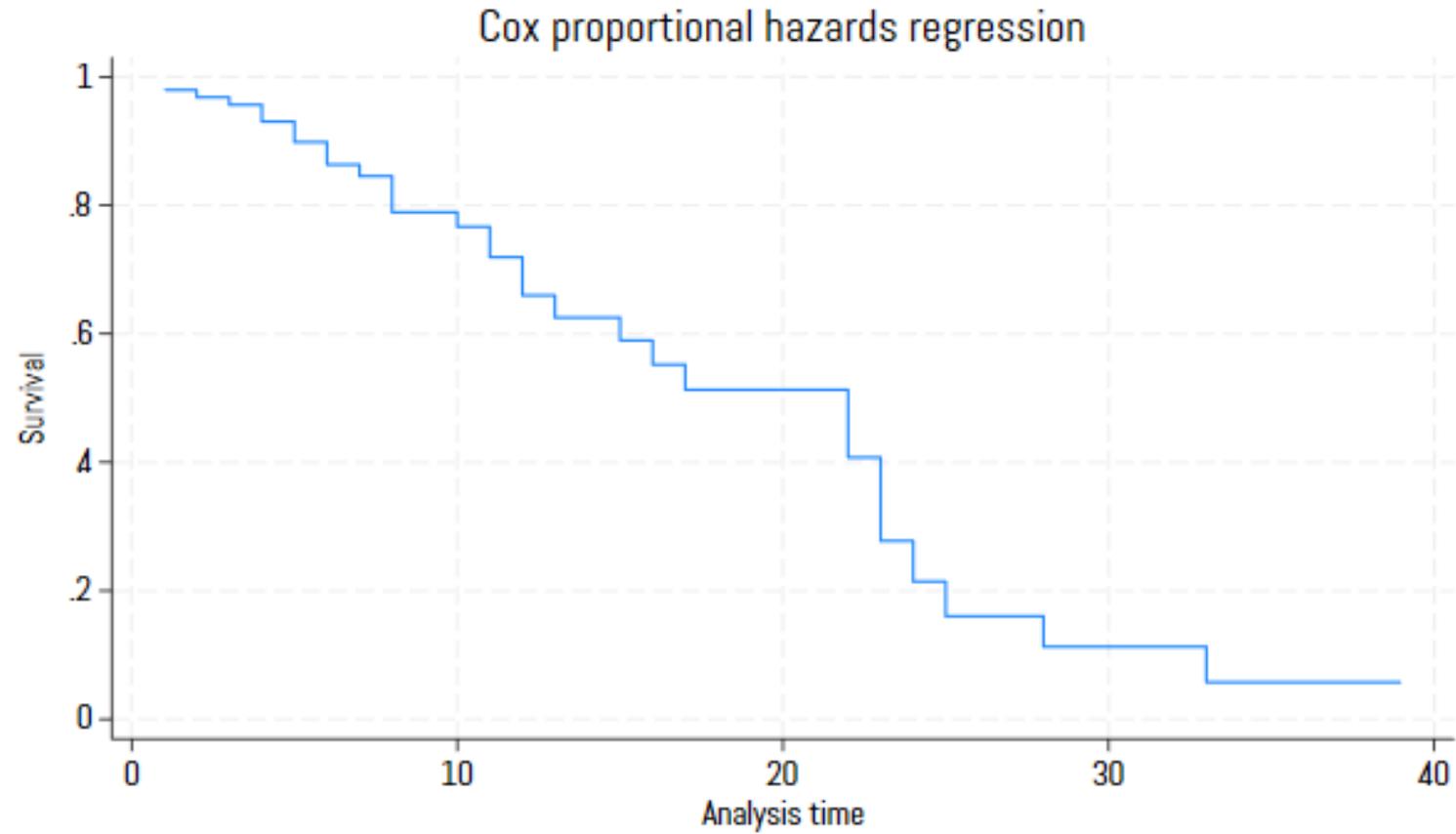
```
Log likelihood = -83.323546
```

```
Prob > chi2 = 0.0000
```

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
drug	.1048772	.0477017	-4.96	0.000	.0430057	.2557622
age	1.120325	.0417711	3.05	0.002	1.041375	1.20526

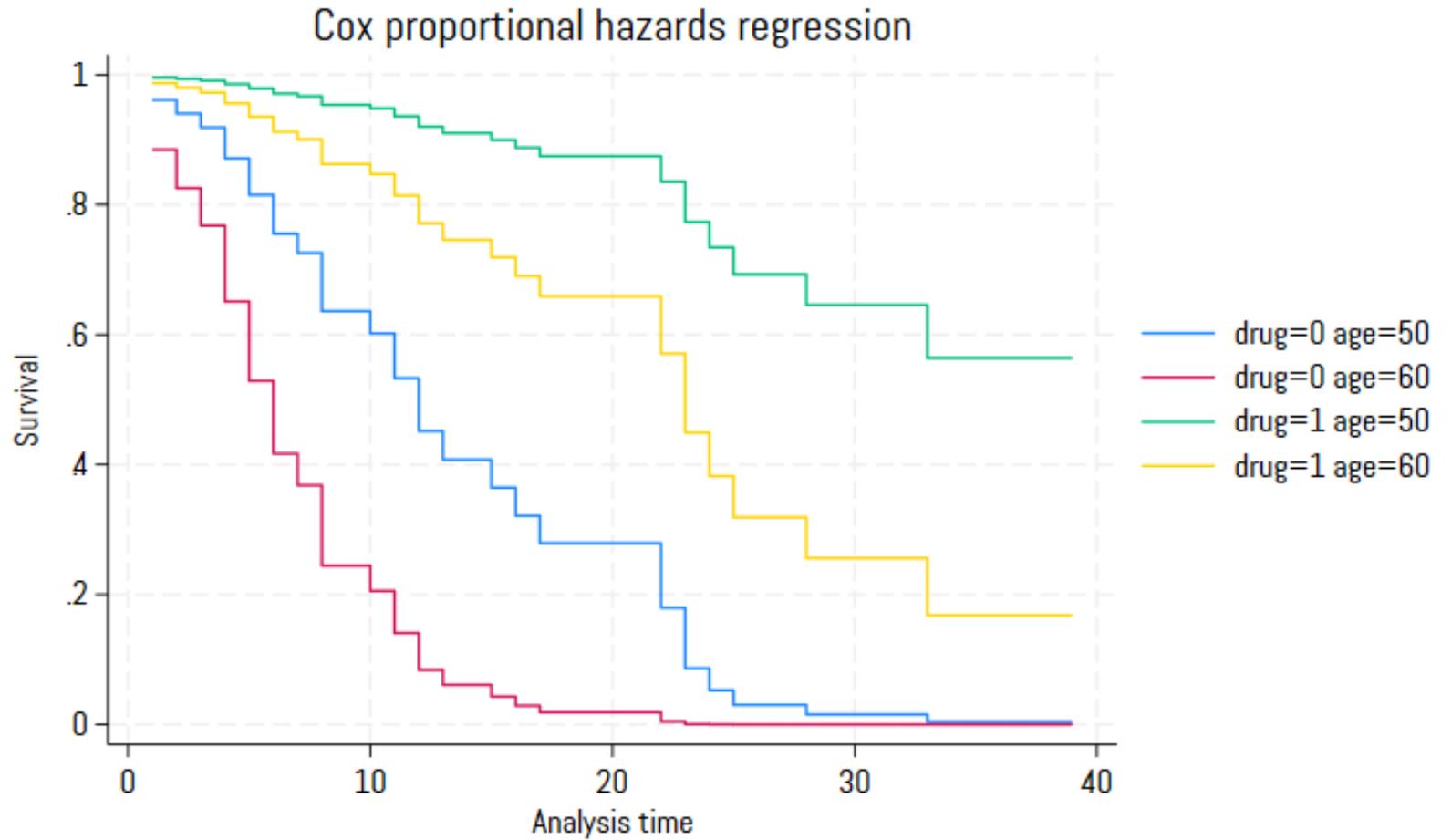
Survivor function

```
. stcurve, survival
```



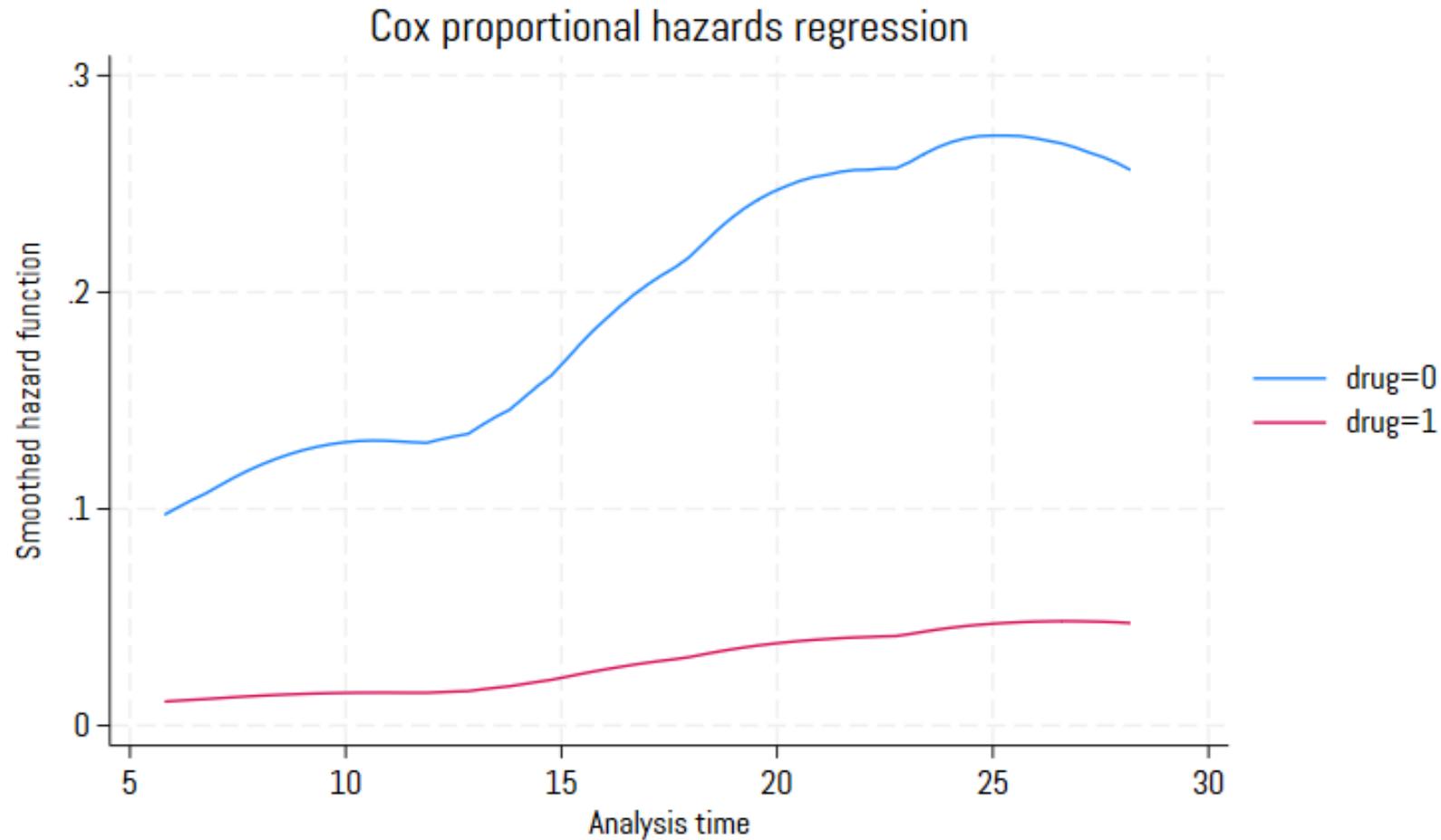
Survivor function

```
. stcurve, survival  
  at1 (drug=0 age=50)  
  at2 (drug=0 age=60)  
  at3 (drug=1 age=50)  
  at4 (drug=1 age=60)
```



Hazard function

```
. stcurve, hazard at(drug=(0 1))
```



Assessing our model

- Statistics
 - How well do our predictions agree with the outcomes?
 - Does the proportional-hazards assumption hold?
- Diagnostic plots
 - Plot of residuals versus time
 - Log-log plots
 - Comparison of the observed survival curve and the Cox predicted curve
 - Goodness-of-fit plot

Concordance probability

```
. estat concordance, gheller
```

```
Failure _d: died
```

```
Analysis time _t: studytime
```

```
Gonen and Heller's K concordance statistic
```

```
Number of subjects (N)      =      48
```

```
Gonen and Heller's K = 0.7748
```

```
Somers' D = 0.5496
```

Test the proportional hazards assumption

```
. estat phtest, detail
```

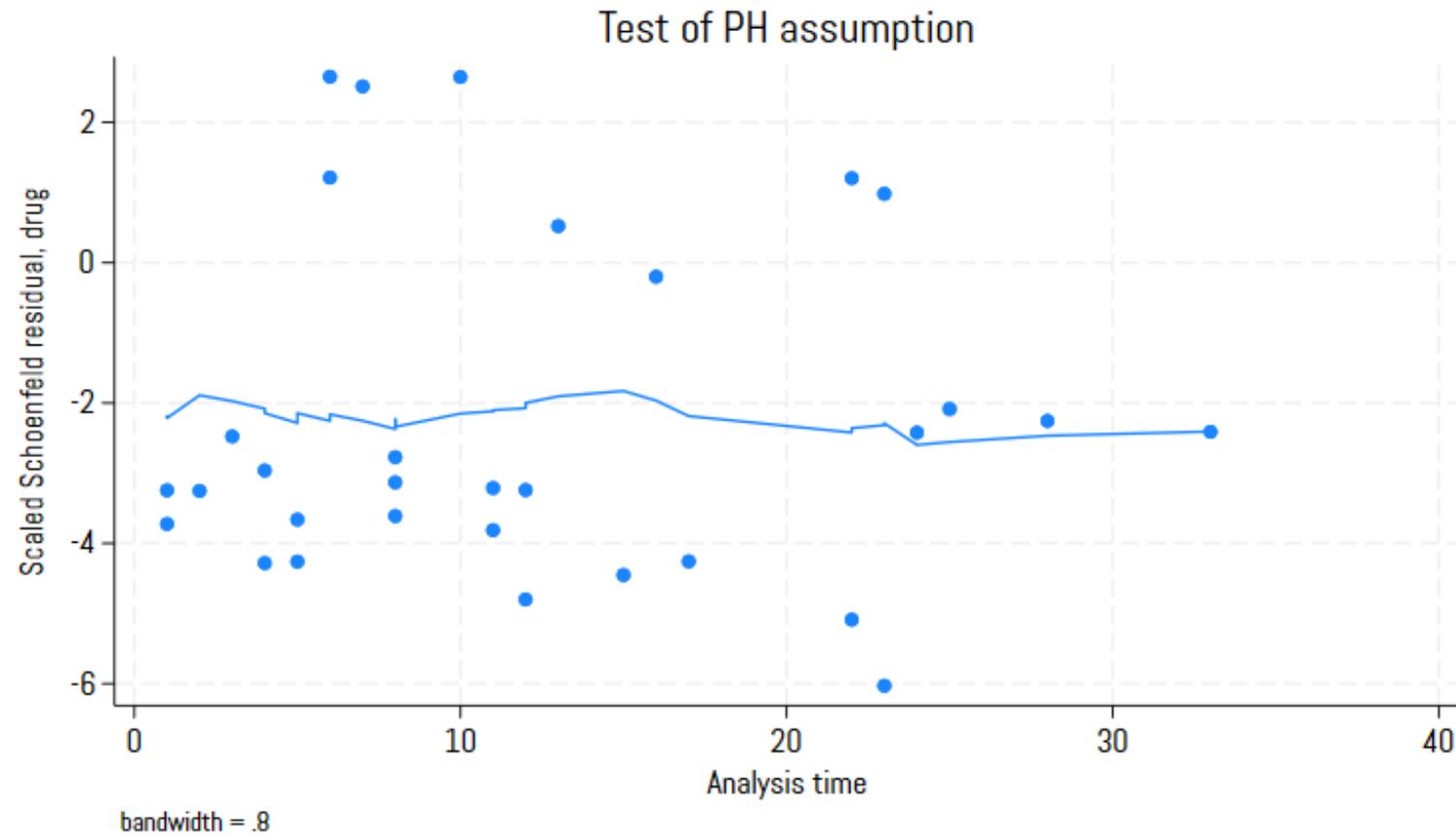
```
Test of proportional-hazards assumption
```

```
Time function: Analysis time
```

	rho	chi2	df	Prob>chi2
drug	0.00949	0.00	1	0.9603
age	-0.11758	0.42	1	0.5168
Global test		0.43	2	0.8064

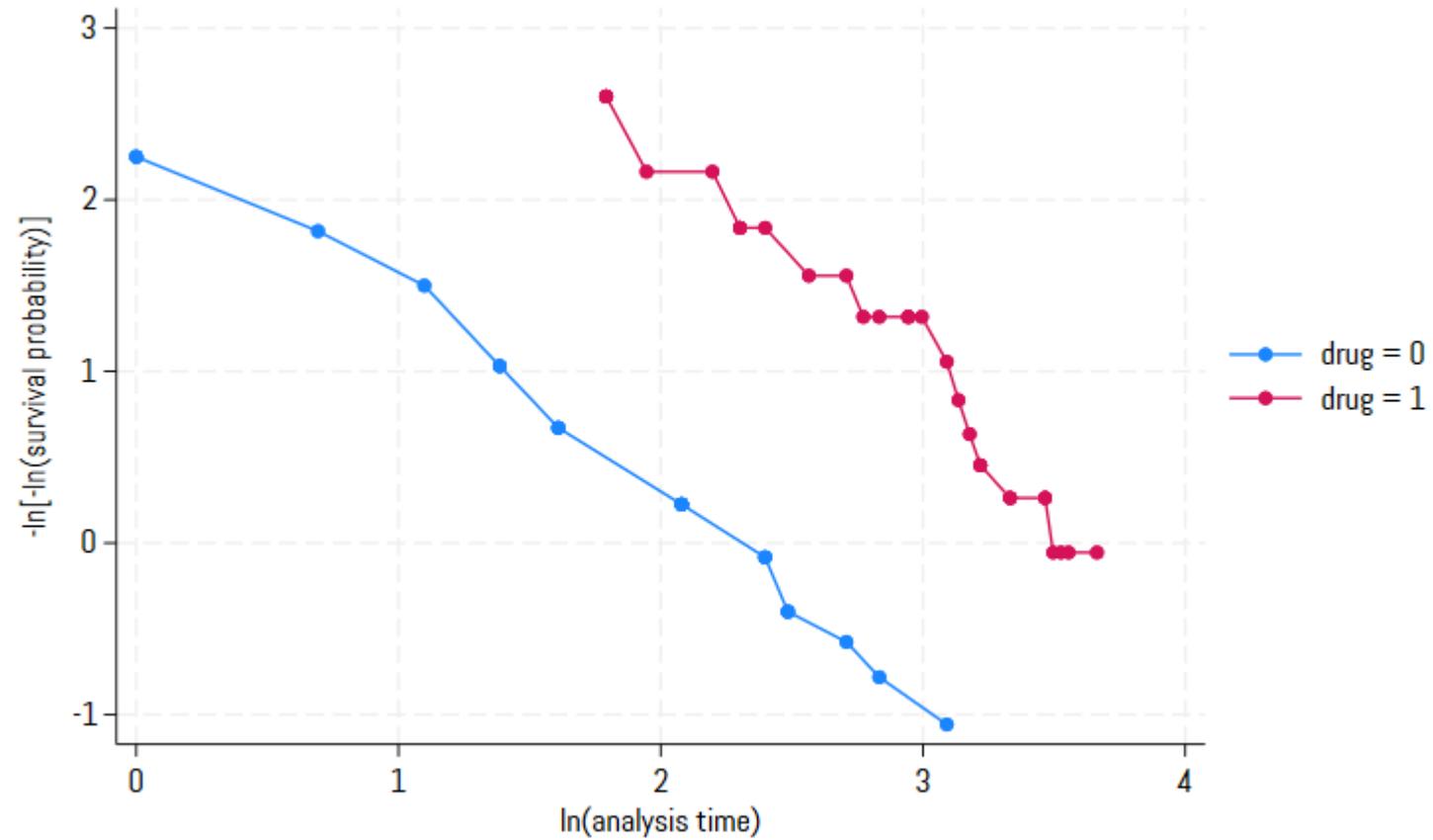
Plotting Schoenfeld residuals versus time

```
. estat phtest, plot(drug)
```



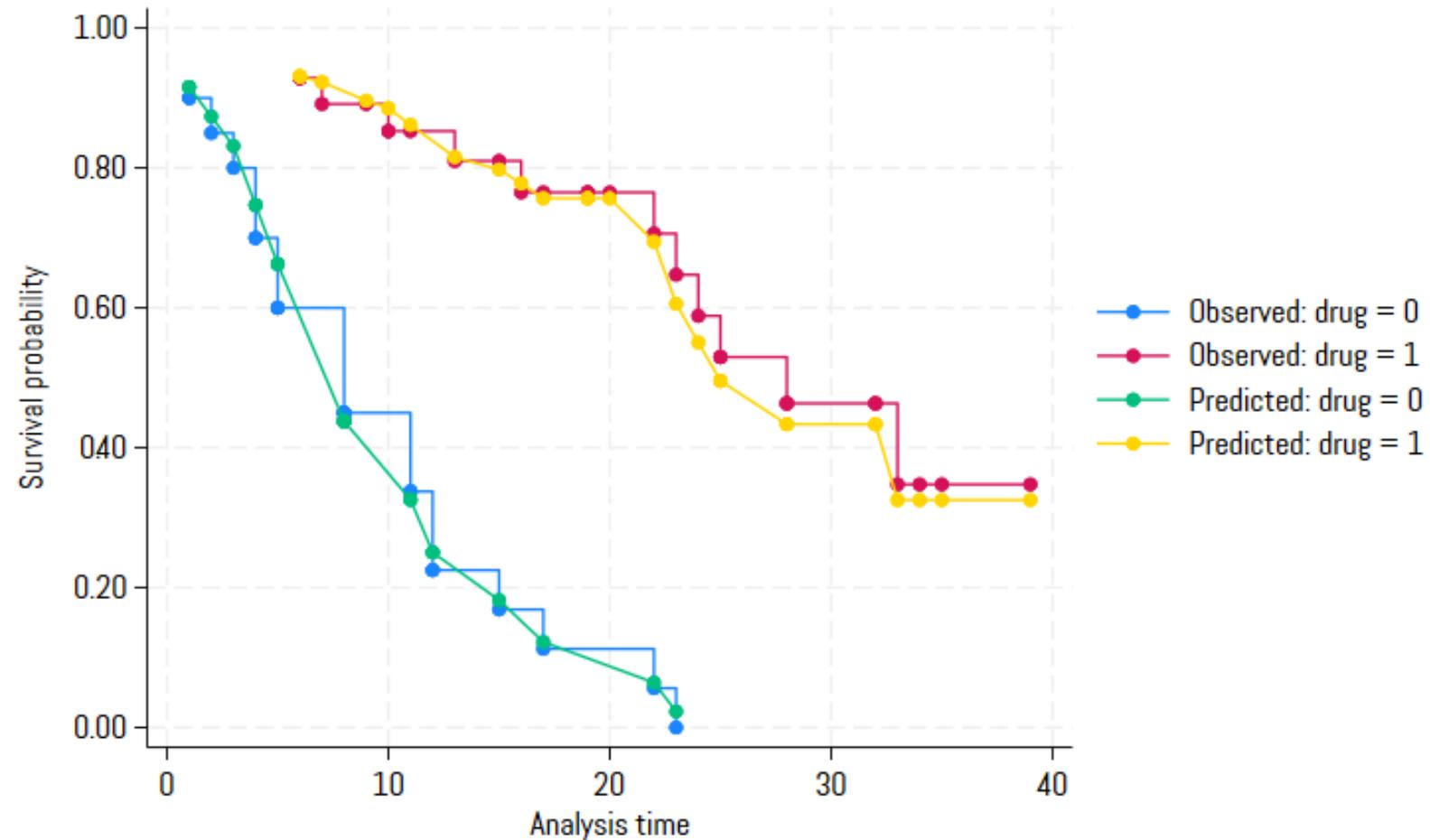
Log-log plot

```
. stpplot, by(drug)
```



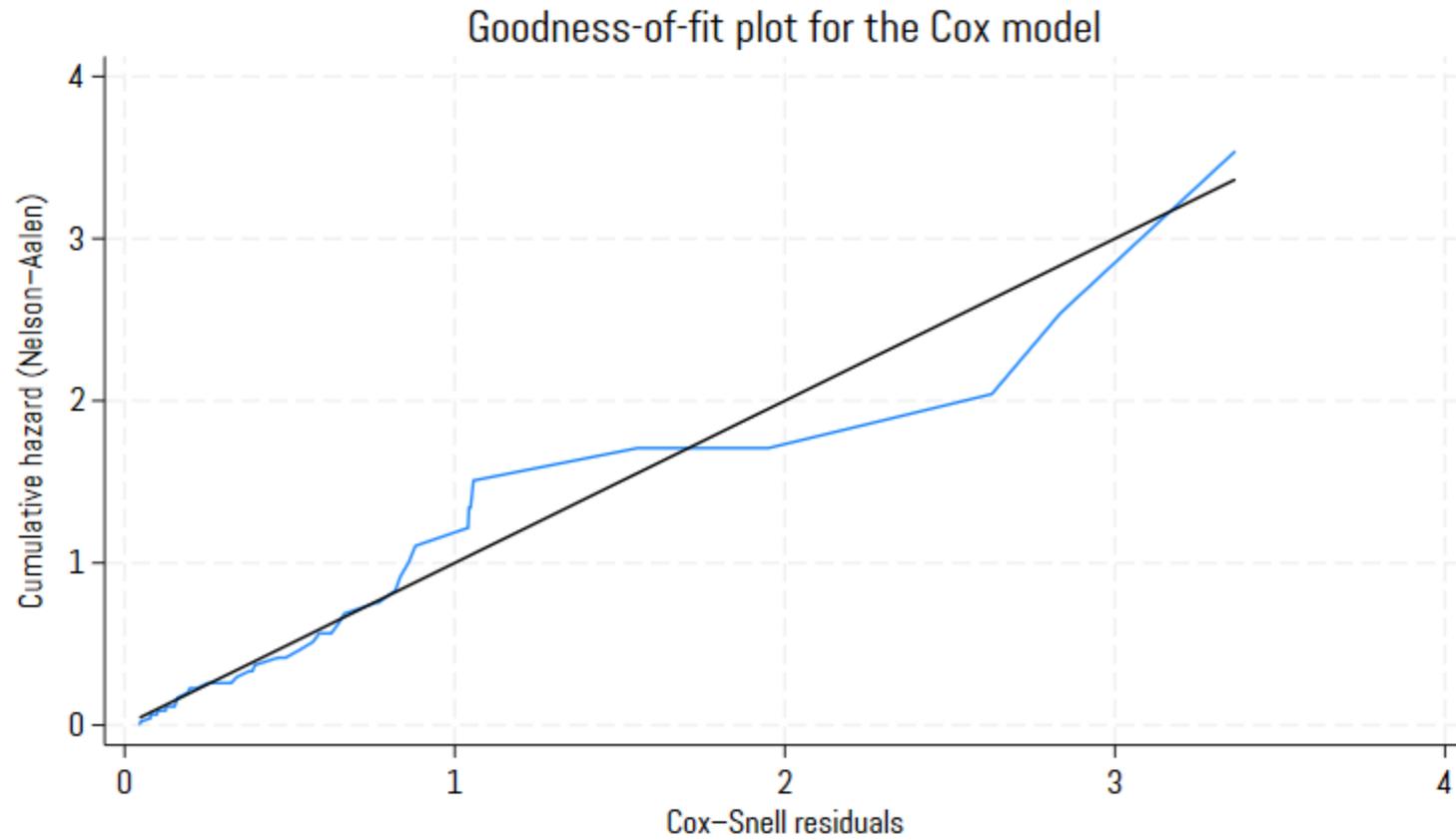
Kaplan–Meier and predicted survival plots

```
. stcoxkm, by (drug)
```



Goodness-of-fit plot

```
. estat gofplot
```



More on the proportional-hazards assumption

Graphical assessment of the proportional-hazards assumption

- Log-log plots
 - Adjust the estimates to average values of specified variables
- Kaplan–Meier and predicted survival plots
 - Specify the method to handle tied failures

Test the proportional-hazards assumption

- Test using Schoenfeld residuals
 - Choose from other time-scale functions or specify your own function of time
- To learn more, see [\[ST\]stcox PH-assumption tests](#).

Interaction between a covariate and analysis time

```
. stcox drug age, tvc(age) nolog
```

```
      Failure _d: died  
      Analysis time _t: studytime
```

Cox regression with Breslow method for ties

No. of subjects = 48

Number of obs = 48

No. of failures = 31

Time at risk = 744

LR chi2(3) = 33.63

Log likelihood = -83.095036

Prob > chi2 = 0.0000

	_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
main	drug	.1059862	.0478178	-4.97	0.000	.0437737	.2566171
	age	1.156977	.07018	2.40	0.016	1.027288	1.303037
tvc	age	.9970966	.0042415	-0.68	0.494	.988818	1.005445

Note: Variables in **tvc** equation interacted with **_t**.

Shared-frailty models

Shared-frailty models

$$h_{ij}(t) = h_0(t) \exp(x_{ij}\beta + v_i)$$

where v_i is the effect of being in group i

- Observations within a group share the same frailty and are thus correlated
- Frailties are unobserved and can be predicted after fitting the model
- Analogous to regression models with random effects

Shared-frailty data

```
. webuse catheter, clear
```

```
(Kidney data, McGilchrist and Aisbett, Biometrics, 1991)
```

```
. sort patient time
```

```
. list patient time infect age female in 1/6, sep(2)
```

	patient	time	infect	age	female
1.	1	8	1	28	0
2.	1	16	1	28	0
3.	2	13	0	48	1
4.	2	23	1	48	1
5.	3	22	1	32	0
6.	3	28	1	32	0

Declare data to be survival-time data

```
. stset time, fail(infect)
```

Survival-time data settings

```
      Failure event: infect!=0 & infect<.  
Observed time interval: (0, time]  
Exit on or before: failure
```

```
  76 total observations  
   0 exclusions
```

```
  76 observations remaining, representing  
  58 failures in single-record/single-failure data  
7,424 total analysis time at risk and under observation  
                                     At risk from t =           0  
Earliest observed entry t =           0  
Last observed exit t =           562
```

Cox regression with shared frailty

```
. stcox age female, shared(patient) noshow nolog
```

Cox regression with Breslow method for ties

Gamma shared frailty

Group variable: **patient**

Number of obs = **76**

Number of groups = **38**

Obs per group:

min = **2**

avg = **2**

max = **2**

No. of subjects = **76**

No. of failures = **58**

Time at risk = **7,424**

Wald chi2(2) = **11.66**

Prob > chi2 = **0.0029**

Log likelihood = **-181.97453**

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.006202	.0120965	0.51	0.607	.9827701	1.030192
female	.2068678	.095708	-3.41	0.001	.0835376	.5122756
theta	.4754497	.2673108				

LR test of theta=0: **chibar2(01) = 6.27**

Prob >= chibar2 = **0.006**

Note: Standard errors of hazard ratios are conditional on theta.

Estimates of log frailties

```
. predict nu, effects
```

```
. sort nu
```

```
. list patient nu in 1/2
```

	patient	nu
1.	21	-2.448707
2.	21	-2.448707

```
. list patient nu in 75/L
```

	patient	nu
75.	7	.5187159
76.	7	.5187159

Estimates of log frailties

```
. predict nu, effects
```

```
. sort nu
```

```
. list patient nu in 1/2
```

	patient	nu
1.	21	-2.448707
2.	21	-2.448707

```
. list patient nu in 75/L
```

	patient	nu
75.	7	.5187159
76.	7	.5187159

$$h_{ij}(t) = h_0(t) \times \exp(x_{ij}\beta) \times \exp(v_i)$$

```
. display exp(-2.448707)  
.08640524
```

```
. display exp(0.5187159)  
1.6798691
```

Other variations of the Cox model

- Stratified Cox regression

- Group specific baseline hazard

```
. stcox x1 x2, strata(svar)
```

- Select another method to handle tied failures

- Efron, exact marginal-likelihood, or exact partial-likelihood

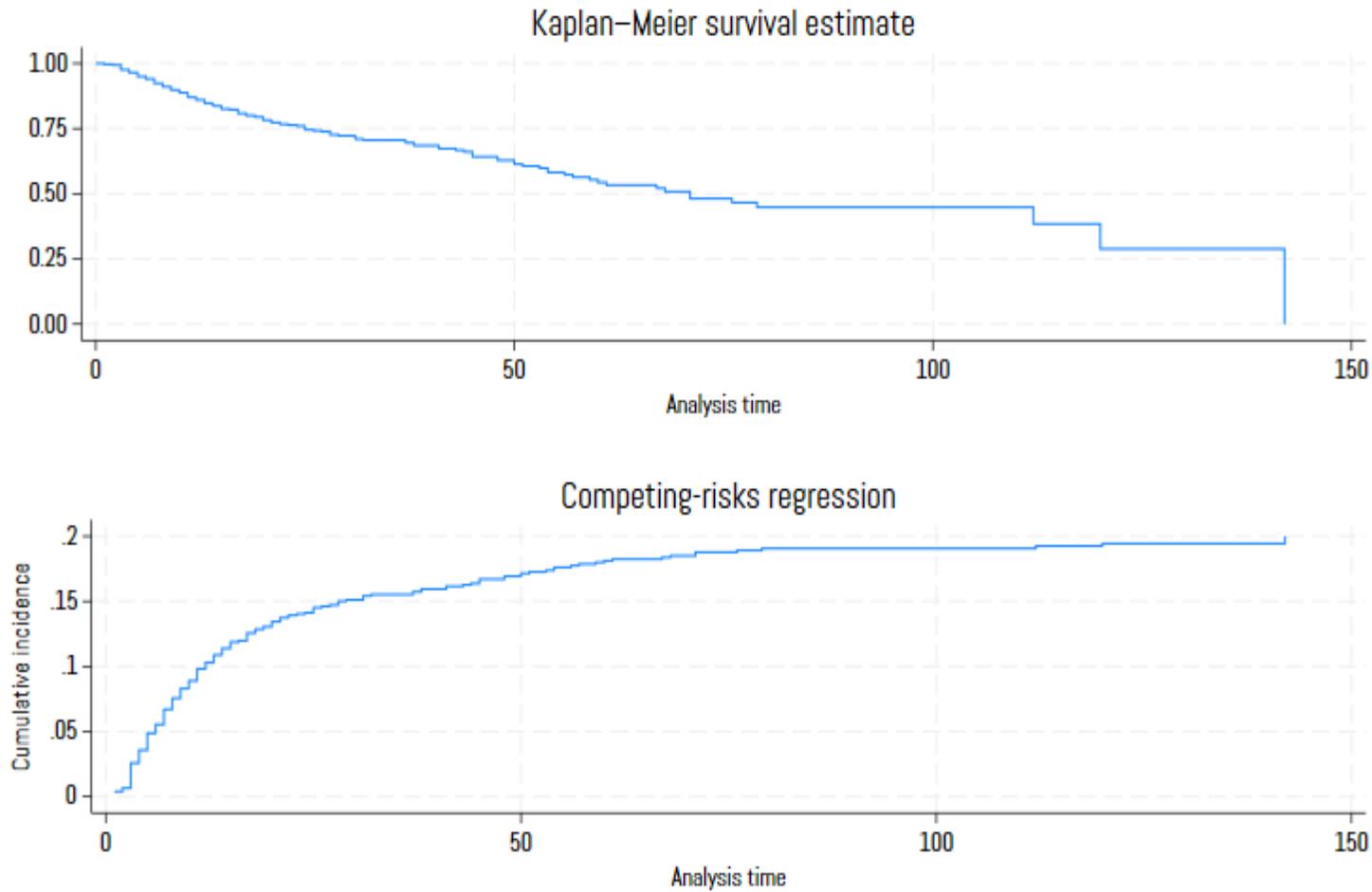
- Learn more about fitting a Cox proportional hazards model in [\[ST\]stcox](#).

Competing risks regression models

Competing failure events

- Consider patients in an ICU after having a heart attack
- Model the time until a cardiac arrest
- If a patient dies, they are no longer at risk for cardiac arrest
- The event of death competes with our event of interest
- With this type of data, we want to focus on the cumulative incidence function

Cumulative incidence function



$$CIF(t) = \Pr(T \leq t \text{ and event of interest})$$

Hazards for competing risks

- Hazard for a cardiac arrest: $h_1(t)$
- Hazard for death: $h_2(t)$
- Total hazard: $h(t) = h_1(t) + h_2(t)$
- Probability of the event being a cardiac arrest: $\frac{h_1(T)}{h_1(T) + h_2(T)}$
- Subhazard for cardiac arrest: $\overline{h_1}(t)$

Subhazard

- Cumulative subhazard: $\overline{H}_1(t) = \int_0^t \overline{h}_1(t) dt$
- $\text{CIF}_1(t) = 1 - \exp\{-\overline{H}_1(t)\}$
 - This accounts for the fact that the cumulative incidence is a function of both hazards
- Model: $\overline{h}_1(t|x) = \overline{h}_{1,0}(t) \times \exp(x\beta)$

Data with competing failure events

```
. use cardiac, clear  
(Fictional cardiac arrest data)
```

```
. describe
```

```
Contains data from cardiac.dta
```

```
Observations:          957           Fictional cardiac arrest data  
Variables:             7           1 Dec 2024 22:56
```

Variable name	Storage type	Display format	Value label	Variable label
id	int	%9.0g		Patient ID
age	byte	%9.0g		Age at admission
ndays	int	%9.0g		Days in ICU
carrest	byte	%9.0g		1 if cardiac arrest
censored	byte	%9.0g		1 if alive and in ICU at the end of the study
death	byte	%9.0g		1 if died
pneumonia	byte	%9.0g		1 if pneumonia

Declare data to be survival-time data

```
. stset ndays, id(id) failure(carrest)
```

Survival-time data settings

```
      ID variable: id  
      Failure event: carrest!=0 & carrest<.  
Observed time interval: (ndays[_n-1], ndays]  
      Exit on or before: failure
```

```
  957 total observations  
    0 exclusions
```

```
  957 observations remaining, representing  
  855 subjects  
  178 failures in single-failure-per-subject data  
16,901 total analysis time at risk and under observation  
                                     At risk from t =      0  
      Earliest observed entry t =      0  
                                     Last observed exit t =    142
```

Competing risks regression

```
. stcrreg age pneumonia, compete(death) noshow nolog
```

```
Competing-risks regression                No. of obs      =      957
                                           No. of subjects =      855
Failure events:   carrest nonzero, nonmissing  No. failed      =      178
Competing events: death nonzero, nonmissing    No. competing   =      641
                                           No. censored    =       36

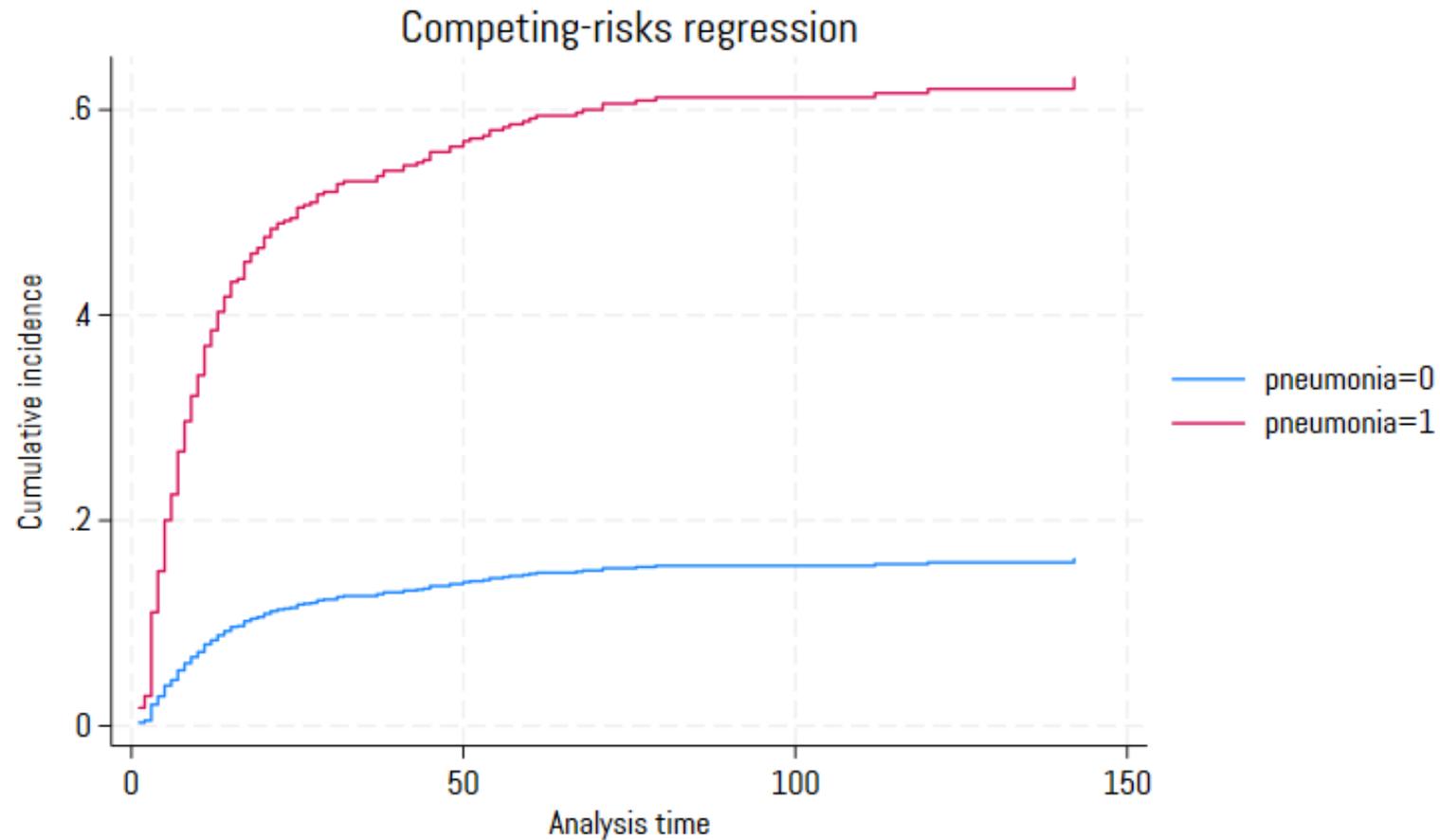
                                           Wald chi2(2)    =     121.21
Log pseudolikelihood = -1128.6096           Prob > chi2     =     0.0000
```

(Std. err. adjusted for **855** clusters in **id**)

_t	SHR	Robust std. err.	z	P> z	[95% conf. interval]	
age	1.021612	.0076443	2.86	0.004	1.006739	1.036705
pneumonia	5.587052	.9641271	9.97	0.000	3.983782	7.835558

Graph of cumulative incidence function

```
. stcurve, cif at(pneumonia=(0 1))
```



Parametric survival models

Parametric survival models

With Stata, you can fit

- Accelerated failure-time (AFT) models
 - $\log t_j = x_j\beta + z_j$
 - Change the time scale by a factor of $\exp(-x_j\beta)$

Parametric survival models

With Stata, you can fit

- Accelerated failure-time (AFT) models

- $\log t_j = x_j\beta + z_j$

- Change the time scale by a factor of $\exp(-x_j\beta)$

- Proportional hazards (PH) models

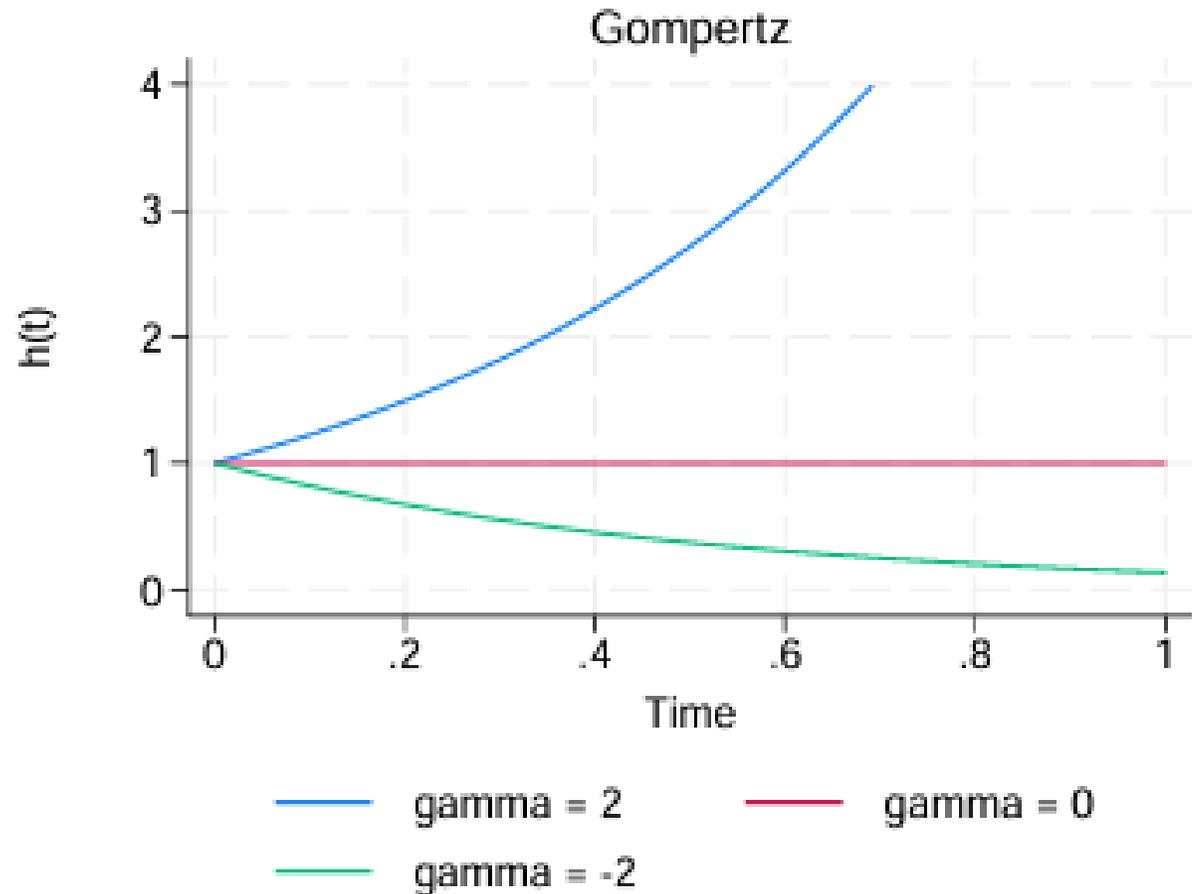
- $h(t_j) = h_0(t) \times g(x_j)$

- Covariates have a multiplicative effect on the hazard function

Stata supports multiple parametric survival distributions when fitting either of these types of models.

Gompertz distribution

- Suitable for data with monotone hazard rates
- $h(t) = \exp(x\beta) \times \exp(\gamma t)$
- Used heavily to model mortality data
- Gompertz models can only be fit in the proportional hazards metric



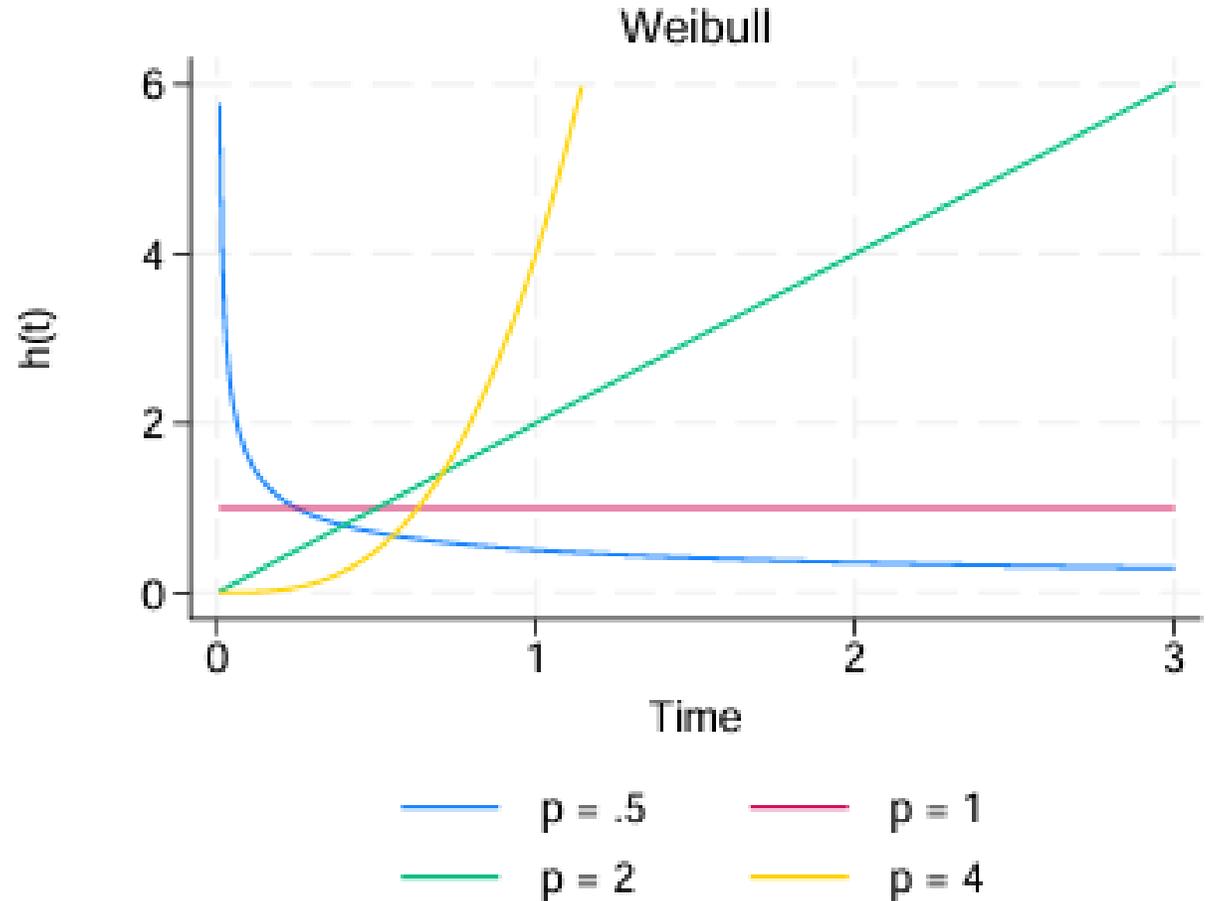
Weibull and exponential distributions

Weibull:

- Suitable for data that exhibit monotone hazard rates
- $h(t) = p \times \exp(x_j\beta) \times t^{p-1}$

Exponential:

- Suitable for data that exhibit a constant hazard
- $h(t) = \exp(x_j\beta)$

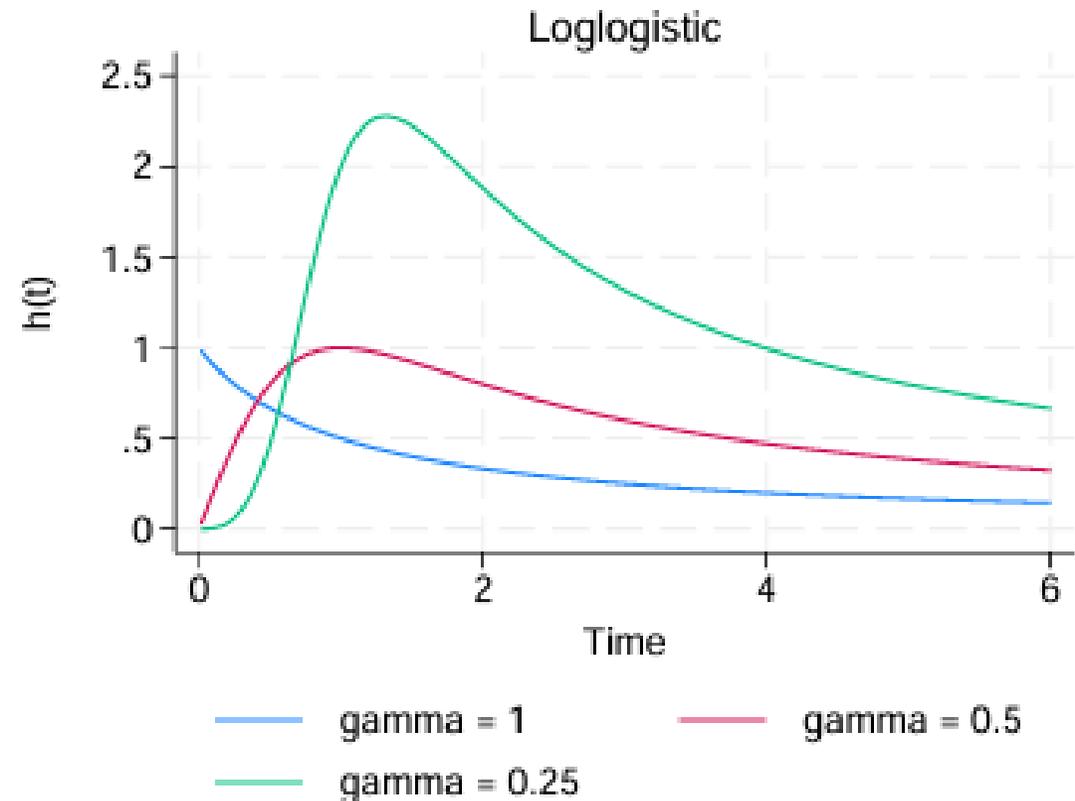


Loglogistic distribution

- Suitable for modeling data with nonmonotonic hazard rates

- $S(t) = \{1 + (\lambda t)^{1/\gamma}\}^{-1}$

- $\ln(t)$ follows a logistic distribution



Fictional data from a drug trial

```
. use cancer2  
(Patient survival in drug trial)
```

```
. describe studytime-age
```

Variable name	Storage type	Display format	Value label	Variable label
studytime	byte	%8.0g		Months to death or end of exp.
died	byte	%8.0g	diedlbl	Patient died
drug	byte	%8.0g		Drug type
age	byte	%8.0g		Patient's age at start of exp.

Declaring data to be survival-time data

```
. stset studytime, failure(died)
```

Survival-time data settings

```
      Failure event: died!=0 & died<.  
Observed time interval: (0, studytime]  
Exit on or before: failure
```

```
 48 total observations  
  0 exclusions
```

```
 48 observations remaining, representing  
 31 failures in single-record/single-failure data  
744 total analysis time at risk and under observation  
                                     At risk from t =           0  
Earliest observed entry t =           0  
Last observed exit t =           39
```

Parametric survival model

```
. streg age i.drug, distribution(llogistic) tratio noshow nolog
```

Loglogistic AFT regression

No. of subjects = **48**

Number of obs = **48**

No. of failures = **31**

Time at risk = **744**

LR chi2(2) = **35.14**

Log likelihood = **-43.21698**

Prob > chi2 = **0.0000**

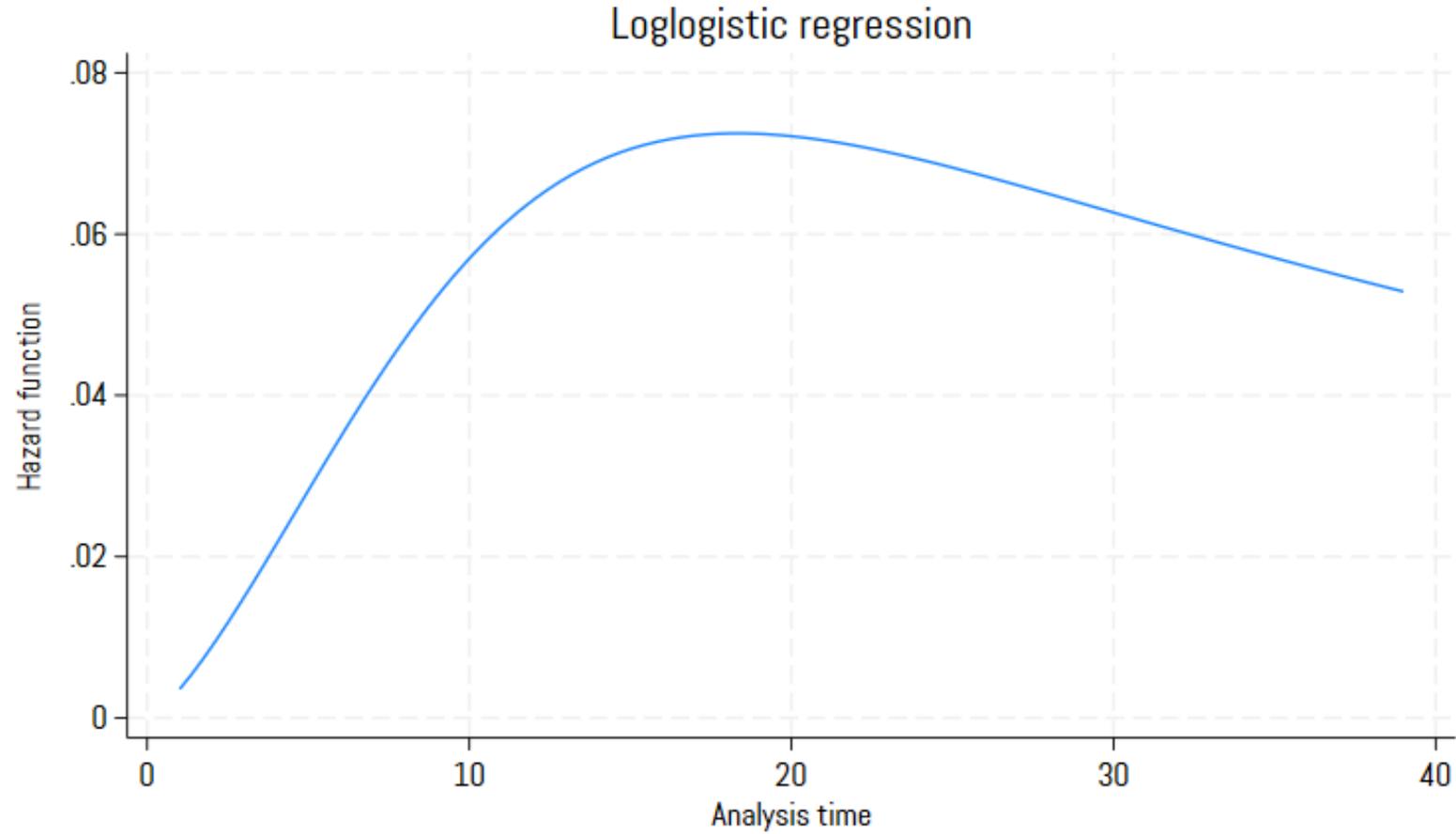
_t	Time ratio	Std. err.	z	P> z 	[95% conf. interval]	
age	.9228128	.0204494	-3.62	0.000	.8835906	.9637759
1.drug	4.138101	1.035414	5.68	0.000	2.534066	6.757473
_cons	630.6247	776.8754	5.23	0.000	56.38505	7053.068
/lgamma	-.8456552	.1479337	-5.72	0.000	-1.1356	-.5557105
gamma	.429276	.0635044			.3212293	.5736646

Note: [Estimates are transformed](#) only in the first equation to time ratios.

Note: **_cons** estimates baseline time.

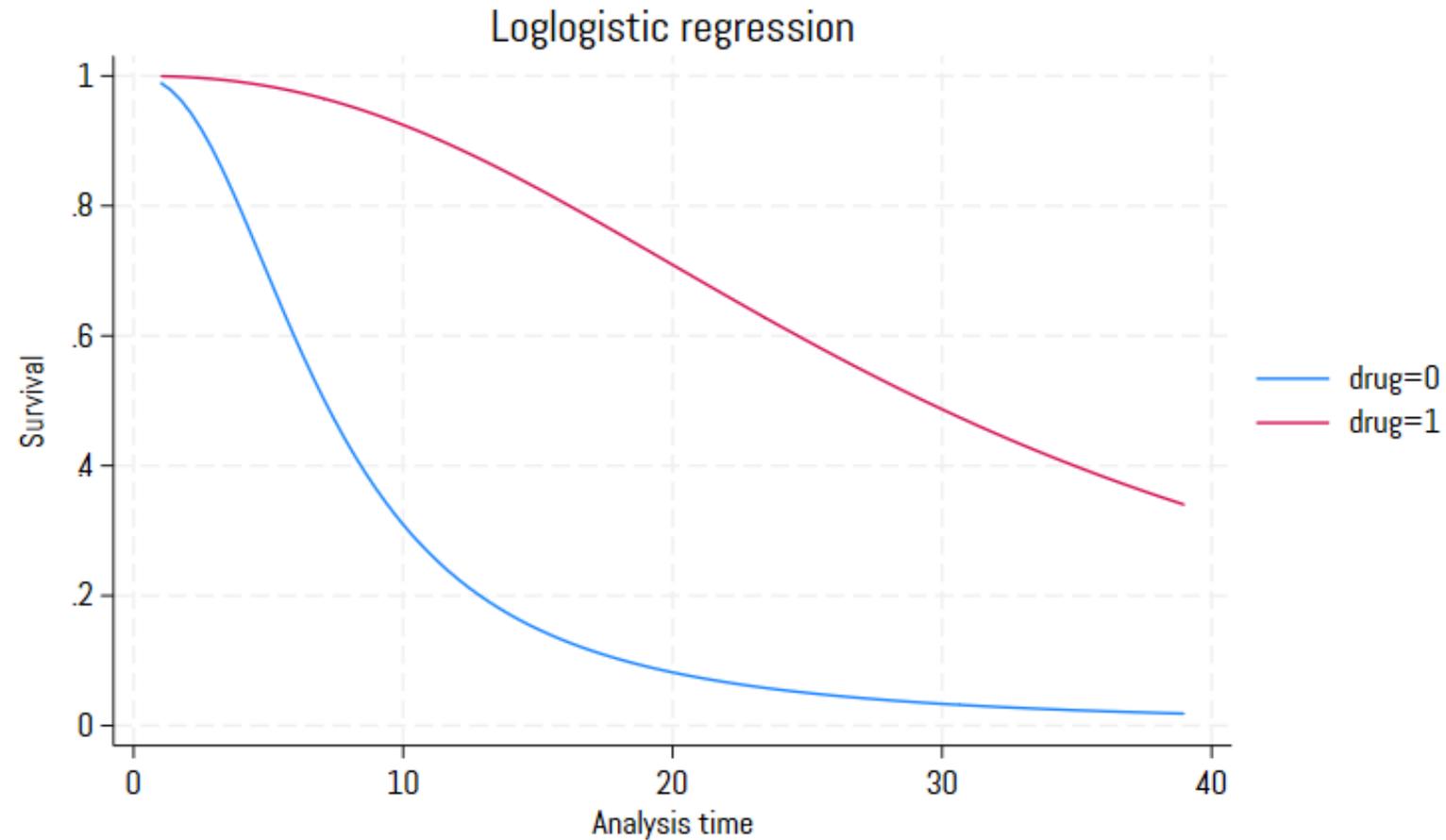
Graph of the hazard function

```
. stcurve, hazard
```



Graphs of survivor functions

```
. stcurve, survival at(drug=(0 1))
```



Expected median survival time

```
. margins drug, at(age=(50(1)65)) noatlegend
```

Adjusted predictions

Number of obs = 48

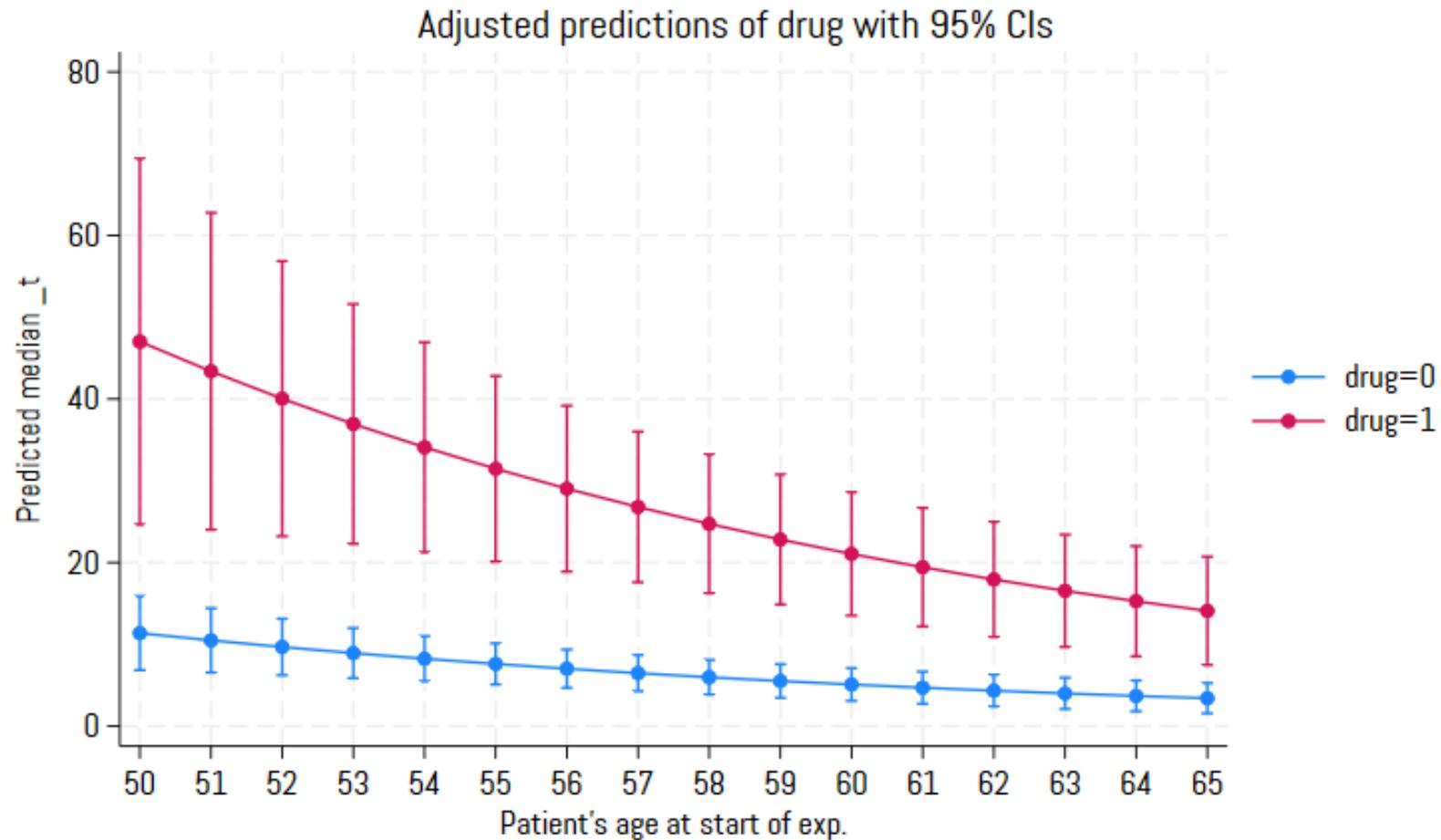
Model VCE: OIM

Expression: Predicted median _t, predict()

	Delta-method					[95% conf. interval]
	Margin	std. err.	z	P> z		
_at#drug						
1 0	11.36189	2.304356	4.93	0.000	6.845438	15.87835
1 1	47.01666	11.41323	4.12	0.000	24.64715	69.38617
2 0	10.4849	2.007197	5.22	0.000	6.550865	14.41893
2 1	43.38757	9.889598	4.39	0.000	24.00432	62.77083
3 0	9.675599	1.761507	5.49	0.000	6.223108	13.12809
3 1	40.03861	8.584062	4.66	0.000	23.21415	56.86306
4 0	8.928766	1.56249	5.71	0.000	5.866343	11.99119
4 1	36.94814	7.477533	4.94	0.000	22.29244	51.60383
5 0	8.239579	1.405181	5.86	0.000	5.485475	10.99368
5 1	34.09621	6.552483	5.20	0.000	21.25358	46.93884
6 0	7.603589	1.284239	5.92	0.000	5.086527	10.12065
6 1	31.46442	5.792208	5.43	0.000	20.1119	42.81694

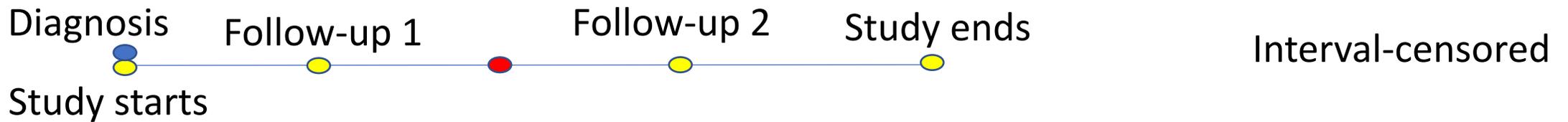
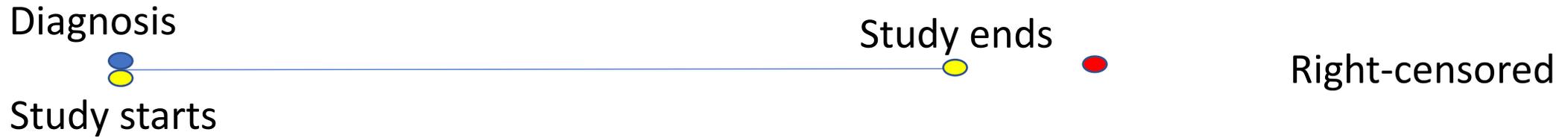
Plot of expected median survival times

```
. marginsplot
```



Interval-censored survival-time data

Interval censoring



Interval-censored survival-time data

- Fit a Cox proportional hazards model for interval-censored data with `[ST]stintcox`
- Fit a parametric model for interval-censored data with `[ST]stintreg`
- Observations can be uncensored, right-censored, left-censored, or interval-censored
- Unlike with other `st` commands, data do not need to be `stset`

What else can Stata
do with survival data?

Data transformations

Convert

- Count-time data to survival-time data; see [\[ST\] cttost](#)
- Snapshot data to time-span data; see [\[ST\] snapspan](#)
- Survival-time data to case-control data; see [\[ST\] sttocc](#)
- Survival-time data to count-time data; see [\[ST\] sttoct](#)
- Manipulate
 - Generate variables reflecting entire histories; see [\[ST\] stgen](#)
 - Split or join time-span records; see [\[ST\] stsplitt](#)
 - Report variables that vary over time; see [\[ST\] stvary](#)

Other models with survival data

- Models with multilevel/panel data
 - Random-effects parametric survival models; see [\[XT\] xtstreg](#)
 - Multilevel mixed-effects parametric survival models; see [\[ME\] mestreg](#)
- Finite mixtures of parametric survival models; see [\[FMM\] fmm: streg](#)
- Bayesian analysis
 - See [\[BAYES\] bayes: streg](#)
 - See [\[BAYES\] bayes: mestreg](#)
- Structural equation models with survival data; see [\[SEM\] Intro 5](#)
- Treatment-effects estimation; see [\[TE\] stteffects](#)

Designing a study for survival analysis

- Sample size, power, and effect size for the Cox proportional hazards model; see [\[PSS\] power cox](#)
- Sample size and power for the exponential test; see [\[PSS\] power exponential](#)
- Sample size, power, and effect size for the log-rank test; see [\[PSS\] power logrank](#)

Where to learn more

- Overview of Stata's [survival analysis features](#)
- Video tutorials on working with [survival-time data in Stata](#)
- FAQs on working with [survival-time models in Stata](#)

References

- Sun, J. 2006. *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer
- Finkelstein, D. M., and R. A. Wolfe. 1985. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 41: 933–945.
- McGilchrist, C. A., and C. W. Aisbett. 1991. Regression with frailty in survival analysis. *Biometrics* 47: 461–466.

Thank you