Enhancements to Proc PHReg for Survival Analysis in SAS 9.2

Brenda Gillespie, Ph.D.

University of Michigan

Presented at the

2010 Michigan SAS Users' Group

Schoolcraft College, Livonia, MI

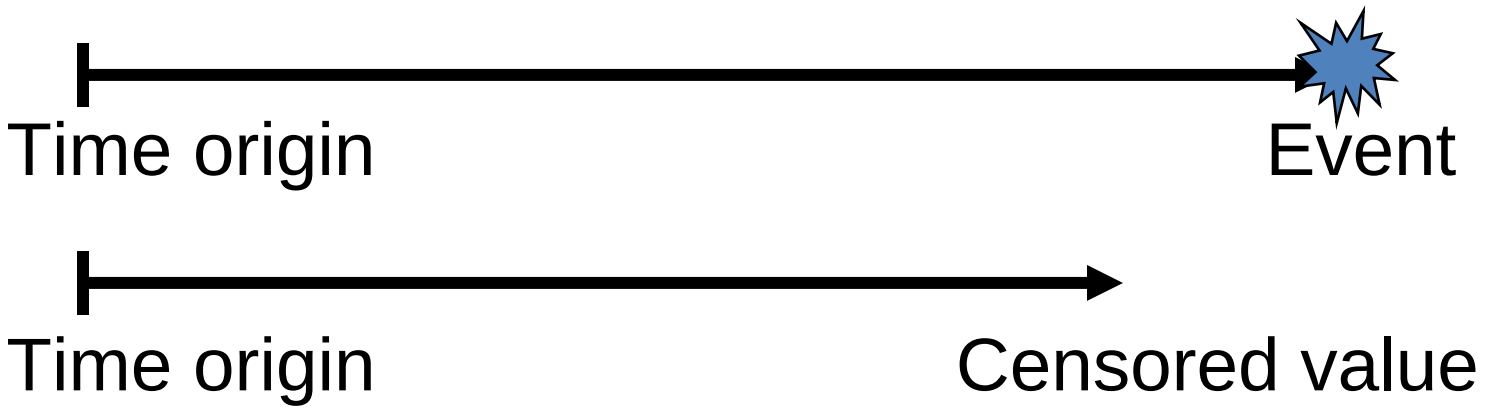
April 27, 2010

Outline

- Overview and introduction to the Cox model
- New features of the MODEL statement
- The CLASS statement (Yeah!!!!)
  - The CLASS statement with time*covariate interactions
- The HazardRatio (HR) statement
- The ASSESS statement
- The Bayes statement
- The Covsandwich option (not new, but still great)

Survival Analysis

- Methods to analyze “time to event” data.
- Useful for many different applications
  - Time to death from disease diagnosis
  - Length of hospital stay



Cox Regression Model

$$h(t;x) = h_o(t)\exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}$$

where $h(t ; x)$ is the hazard function at time t for a subject with covariate values x_1, … x_k,

$h_0(t)$ is the baseline hazard function, i.e., the hazard function when all covariates equal zero.

exp is the exponential function ($\exp(x)= e^x$),

x_i is the i^{th} covariate in the model, and

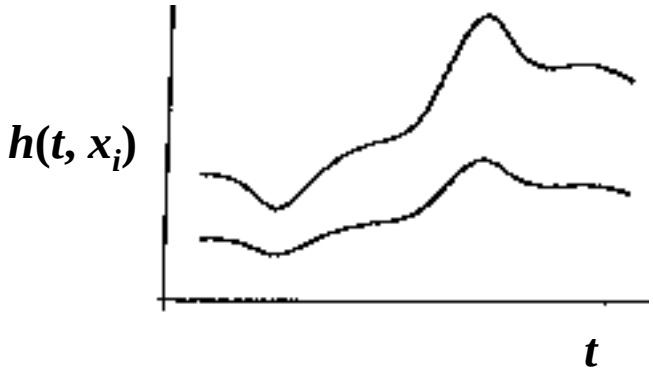
β_i is the regression coefficient for the i^{th} covariate, x_i.

Cox Regression (cont'd)

$$h(t;x) = h_o(t)\exp\{\beta_1 x_1 + \cdots + \beta_k x_k\}$$

- The Cox Model is different from ordinary regression in that the covariates are used to predict the hazard function, and not Y itself.
- The baseline hazard function can take any form, but it cannot be negative.
- The exponential function of the covariates is used to insure that the hazard is positive.
- There is no intercept in the Cox Model . (Any intercept could be absorbed into the baseline hazard.)

Cox Regression (cont'd)



- The basic Cox Model assumes that the hazard functions for two different levels of a covariate are proportional for all values of t.
- For example, if men have twice the risk of heart attack compared to women at age 50, they also have twice the risk of heart attack at age 60, or any other age.
- The underlying risk of heart attack as a function of age can have any form.

Proportional Hazards

To see the proportional hazards property analytically, take the ratio of h(t;*x*) for two different covariate values:

$$\frac{h(t;x_i)}{h(t;x_j)} = \frac{h_0(t)\exp\{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}\}}{h_0(t)\exp\{\beta_1 x_{j1} + \cdots + \beta_k x_{jk}\}}$$
$$= \exp\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk})\}$$

$h_o(t)$ cancels out => the ratio of those hazards is the same at all time points.

For a single dichotomous covariate, say with values 0 and 1, the **hazard ratio** is

$$\frac{h(t;\, x = 1)}{h(t;\, x = 0)} = \frac{h_0(t)e^{\beta * 1}}{h_0(t)e^{\beta * 0}} = \frac{e^{\beta}}{e^{0}} = e^{\beta}$$

Partial Likelihood

The partial likelihood function for one covariate is:

$$\mathrm{PL} = \prod_{i \in \text{events}} \frac{h(t_i)}{\sum_{j \in \mathrm{R}i} h(t_j)} = \prod_{i \in \text{events}} \frac{h_0(t_i)\exp\{\beta x_i\}}{\sum_{j \in \mathrm{R}i} h_0(t_i)\exp\{\beta x_j\}} = \prod_{i \in \text{events}} \left[\frac{e^{\beta xi}}{\sum_{j \in \mathrm{R}i} e^{\beta xj}} \right]$$

where t_i is the i^{th} death time, x_i is the associated covariate, and

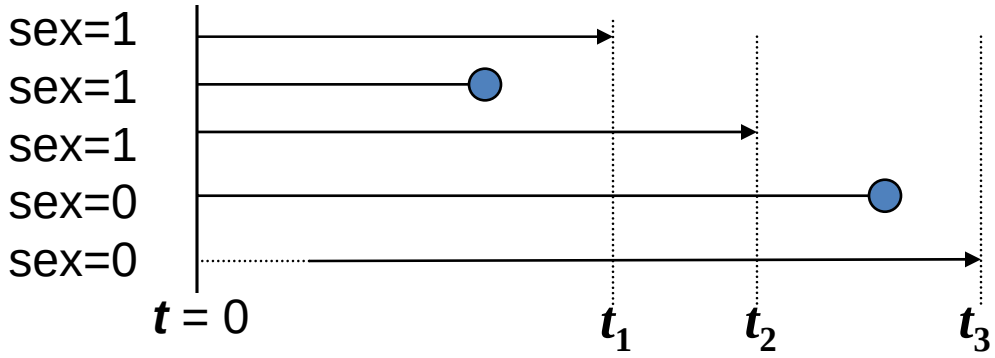
R_i is the risk set at time t_i , i.e., the set of subjects is still alive and uncensored just prior to time t_i.

The numerator is the hazard of death for the subject who died at time t_i.

The denominator is the sum of hazards of death for all subjects in the risk set at time t_i.

The ratio reflects the likelihood that the death occurred to subject *i*.

For the data below, with gender as the only covariate, the PL function is:



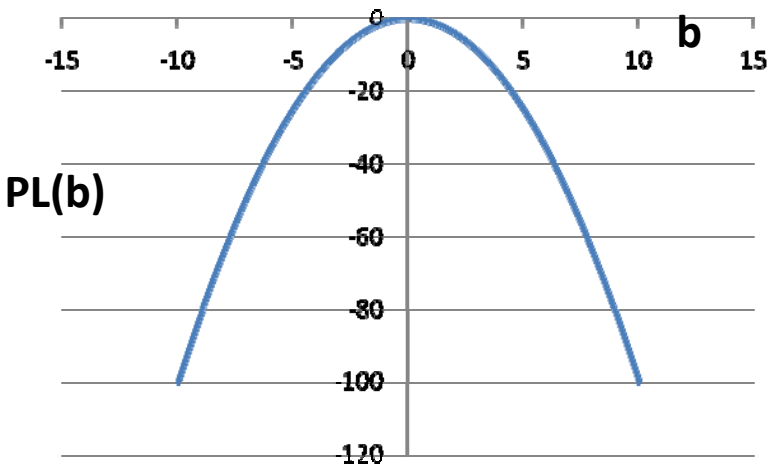
$$PL(\beta) = \frac{e^{1\beta}}{e^{1\beta} + e^{1\beta} + e^{0\beta} + e^{0\beta}} \times \frac{e^{1\beta}}{e^{1\beta} + e^{0\beta} + e^{0\beta}} \times \frac{e^{0\beta}}{e^{0\beta}}$$

- Note: The first censored observation does not enter the PL
- If β=0, PL=1/4 x 1/3 x 1
- See graph of PL(β) on next slide

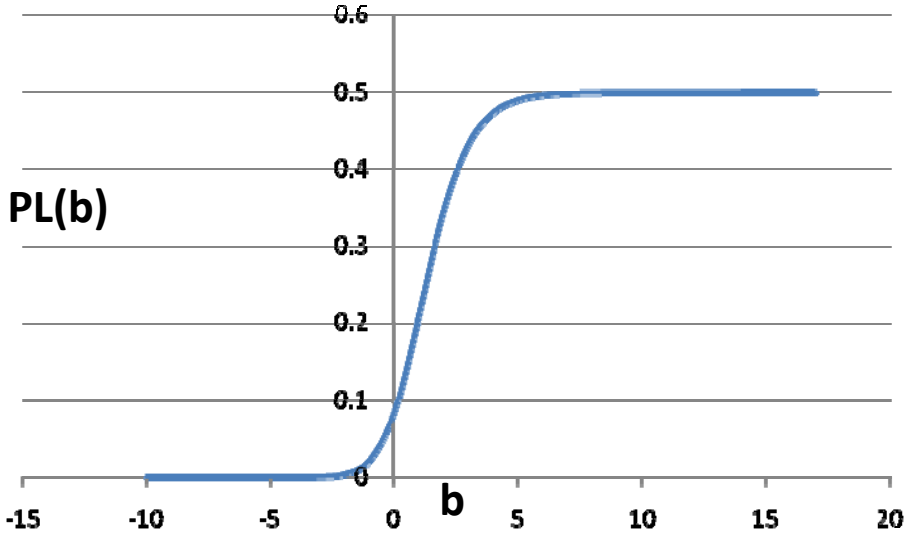
Graph of the Likelihood Function

- Many likelihood functions are approximately quadratic, concave functions (below, left)
  - β estimate is at the max
- In prev slide, PL increases indefinitely (below, right)
- β is maximized at $+\infty$ (occurs with "separation").
- Note early events for sex=1, later events for sex=0.

Common PL(β) shape



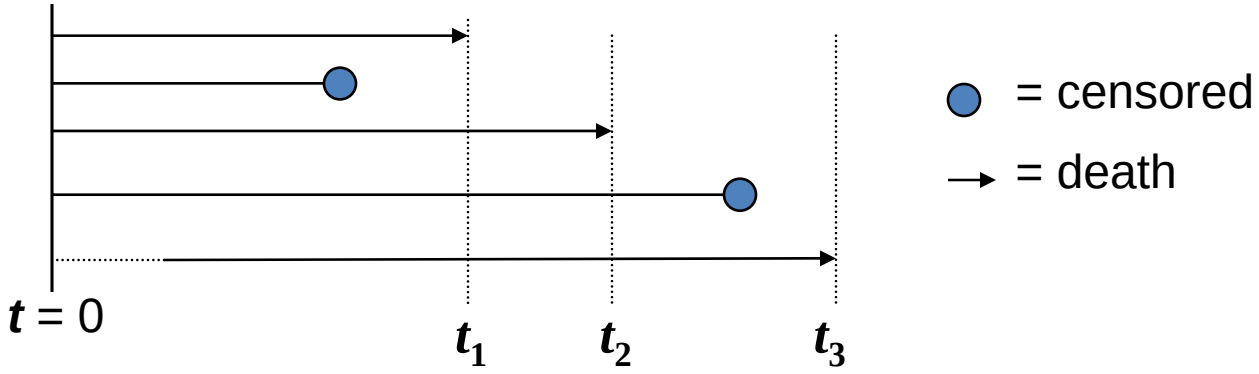
PL(b) with separation



Partial Likelihood (cont'd)

The beauty of the partial likelihood function is that for each event (death) there is one term, which only depends on the risk set at that point.

$$PL = L_1 \times L_2 \times \ldots \times L_d, \text{ where d = \# of deaths.}$$



This structure makes it easy to

•accommodate **right censoring**: A subject is in all risk sets up to the time of censoring, but not in any risk set after that point.

•incorporate **time-varying covariates**: A subject may be in many risk sets, but can have different covariate values in each.

•incorporate **left-truncated data**: A subject may not be in the first risk set(s), but may join later risk set(s) at any time.

www.york.ac.uk/.../ histstat/people/cox_d_r.gif

Syntax for Cox Regression using PHREG

- The time variable is “days”
- The censor code is “status” (1=dead, 0=alive)
- Underlined items are user-specified

```
proc phreg;
    model  days*status (0) = sex age;
    output out=temp resmart=Mresids
        resdev=Dresids ressch=Sresids;
    id  subj  group;
    run;
```

Options of the **MODEL** Statement

- **Risklimits (RL)**
  - to get confidence intervals for hazard ratios
- **NoDummyPrint (NODP)**
  - To suppress “Class Level Information” table
- **Firth**
  - Better estimation when you have complete separation for one variable
- **Selection=score Best=5**
  - An alternative to stepwise regression, where all covariate subsets are considered and several options are presented

Categorical Covariates

- Categorical covariates can be declared in a CLASS statement in SAS 9.2 (finally!!)
  - E.g., 3 levels of edema (None, Mild, and Severe)
  - By default, SAS uses Severe [=highest value] as the reference category

Dummy Variable Coding

Effect

	Design Matrix		
A	**A1**	**A2**	**A5**
1	**1**	0	0
2	0	**1**	0
5	0	0	**1**
7	**-1**	**-1**	**-1**

Estimates the difference in the effect of each level compared to the **average** effect over all four levels

Reference

	Design Matrix		
A	**A1**	**A2**	**A5**
1	**1**	0	0
2	0	**1**	0
5	0	0	**1**
7	0	0	0

Estimate the difference in the effect of each level compared to the **reference** level.

Time*Covariate Interactions with the Covariate in the CLASS statement

- To test of proportional hazard assumption for the design variables of A, use the following statements,

```
proc phreg data=Foo;
class A;
model T * Status(0)  =  A   X1   X2;
X1= T*(A=1);
X2= T*(A=2);
run;
```

The PHREG Procedure

Parameter		DF	Parameter Estimate	Standard Error	Pr > ChiSq	Hazard Ratio
A	1	1	-0.0076	1.6943	0.9964	0.992
A	2	1	-0.8813	1.6429	0.5917	0.414
X1		1	-0.1552	0.2017	0.4417	0.856
X2		1	0.0115	0.1885	0.9512	1.012

Example:
Time from Prison Release to Next Arrest

- n=432 male inmates released from Maryland state prisons in the 1970s
- The men were followed for one year.
- The dates of arrest were recorded.
- Out of several covariates (see Allison, p. 42-43), we will focus on two:
  - FIN = 1 if financial aid was received after release; FIN=0 if not (randomly assigned)
  - AGE = age in years at the time of release

The PHREG Procedure

Data Set: RECID

Dependent Variable: WEEK

Censoring variable: ARREST

Censoring value(s): 0

Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
432	114	318	73.61

Testing Global Null Hypothesis: BETA=O

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	1351.367	1318.241	33.126 with 7 DF (p=0.0001)
Score			33.383 with 7 DF (p=0.0001)
Wald			31.981 with 7 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
FIN	1	-0.379022	0.19136	3.92289	0.0476	0.685
AGE	1	-0.057246	0.02198	6.78122	0.0092	0.944
RACE	1	0.314130	0.30802	1.04008	0.3078	1.369
WEXP	1	-0.151115	0.21212	0.50750	0.4762	0.860
MAR	1	-0.432783	0.38179	1.28493	0.2570	0.649
PARO	1	-0.084983	0.19575	0.18848	0.6642	0.919
PRIO	1	0.091112	0.02863	10.12666	0.0015	1.095

Interpreting Covariate Effects

- Covariate effects are interpreted in terms of hazard ratios (HR), sometimes called risk ratios.

  HR=1 **=>** no effect. (i.e., $\beta = 0$)

  HR>1 **=>** increasing hazard with increasing x.

  HR<1 **=>** decreasing hazard with increasing x.

- For dichotomous variables, HR = exp(β)

  For FIN, HR = exp(-0.38) = 0.68

  Those receiving financial aid have only 68% the risk of arrest compared with those not receiving aid.

Categorical Variables with 3 or more levels

- For categorical variables with 3 or more levels, use a CLASS statement. The highest level is the reference category by default.
- To get the Hazard Ratios for other comparisons, e.g., RACE (3 levels):

```
proc phreg;   CLASS race;
  model time*cens(0) = age race;
  HAZARDRATIO   race / diff=all;  run;
```

Hazard Ratios for RACE

Description	Point Estimate	95% Wald Confidence Limits	
RACE Black vs Asian	0.563	0.332	0.955
RACE Black vs White	0.119	0.070	0.203
RACE Asian vs White	0.211	0.106	0.420

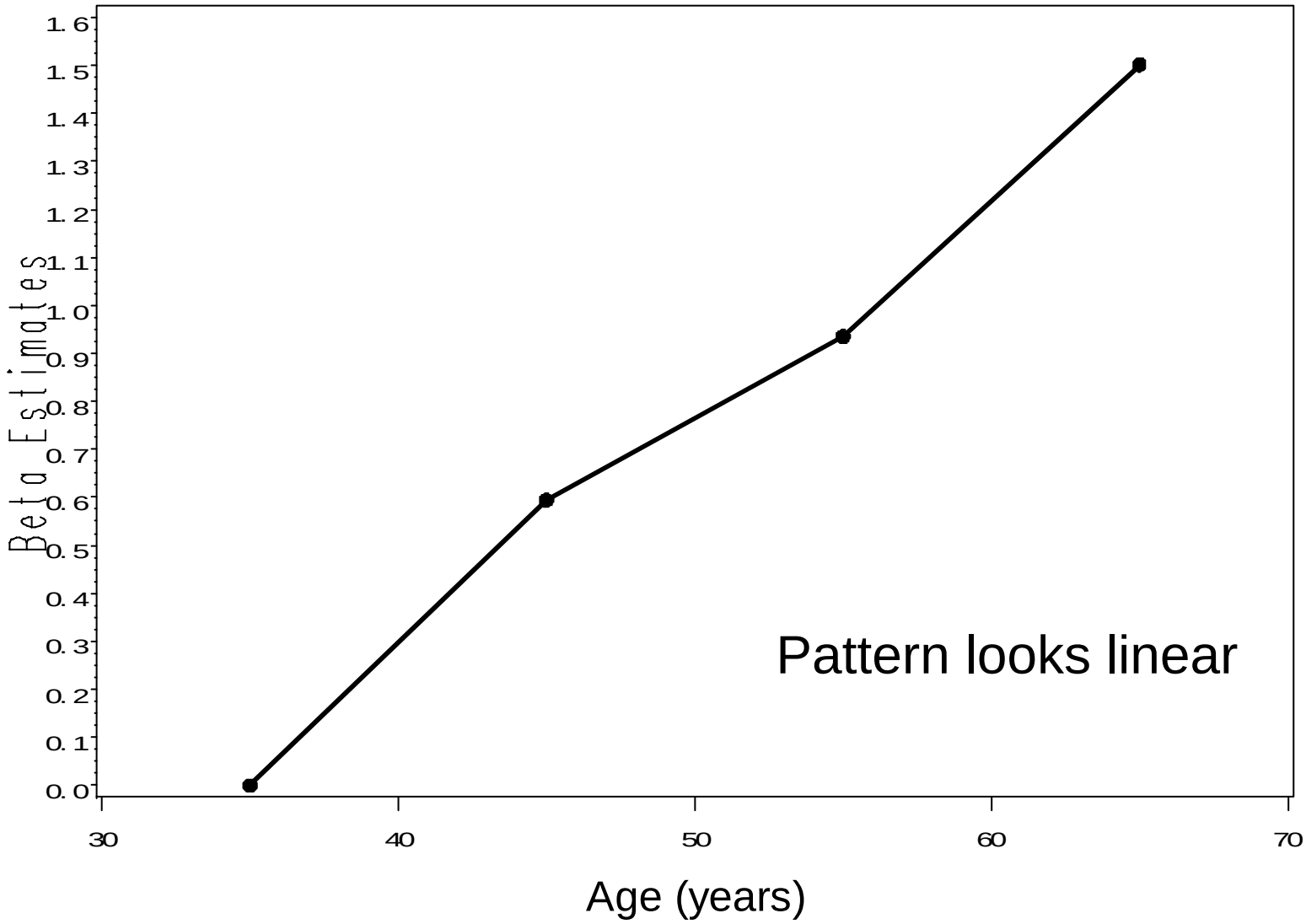
Interpreting Covariate Effects (cont'd)

- For continuous variables, HR = exp(β) reflects the hazard ratio for a one-unit increase in x. For a 10-unit increase in x, HR = exp(10β).
  - For age, HR = exp(-0.06) = 0.94 for a one-year increase, but HR= exp(-0.6) = 0.56 for a 10-year increase in age. Older men have lower hazard of arrest.
  - In SAS, use the HAZARDRATIO statement:
    HAZARDRATIO AGE / units=10;
- The betas for continuous variables are often smaller than the betas for dichotomous variables, because they reflect a smaller increment.

Assessing Functional Form of Continuous Covariates

- Often we assume continuous covariates have a linear form. However, this assumption should always be checked. We give 3 ways to check:
- **Method 1 - Try X categorical:**
  - Categorize X into $\geq$4 intervals, say by quantiles.
  - Create dummy variables for the categories and fit a model with these dummy variables.
  - Plot β estimates by X interval midpoints, with β=0 for the reference category.
  - Look at the shape, and model X accordingly (e.g., linear, quadratic, threshold).

Plot of Beta Estimates by Age Category Midpoints

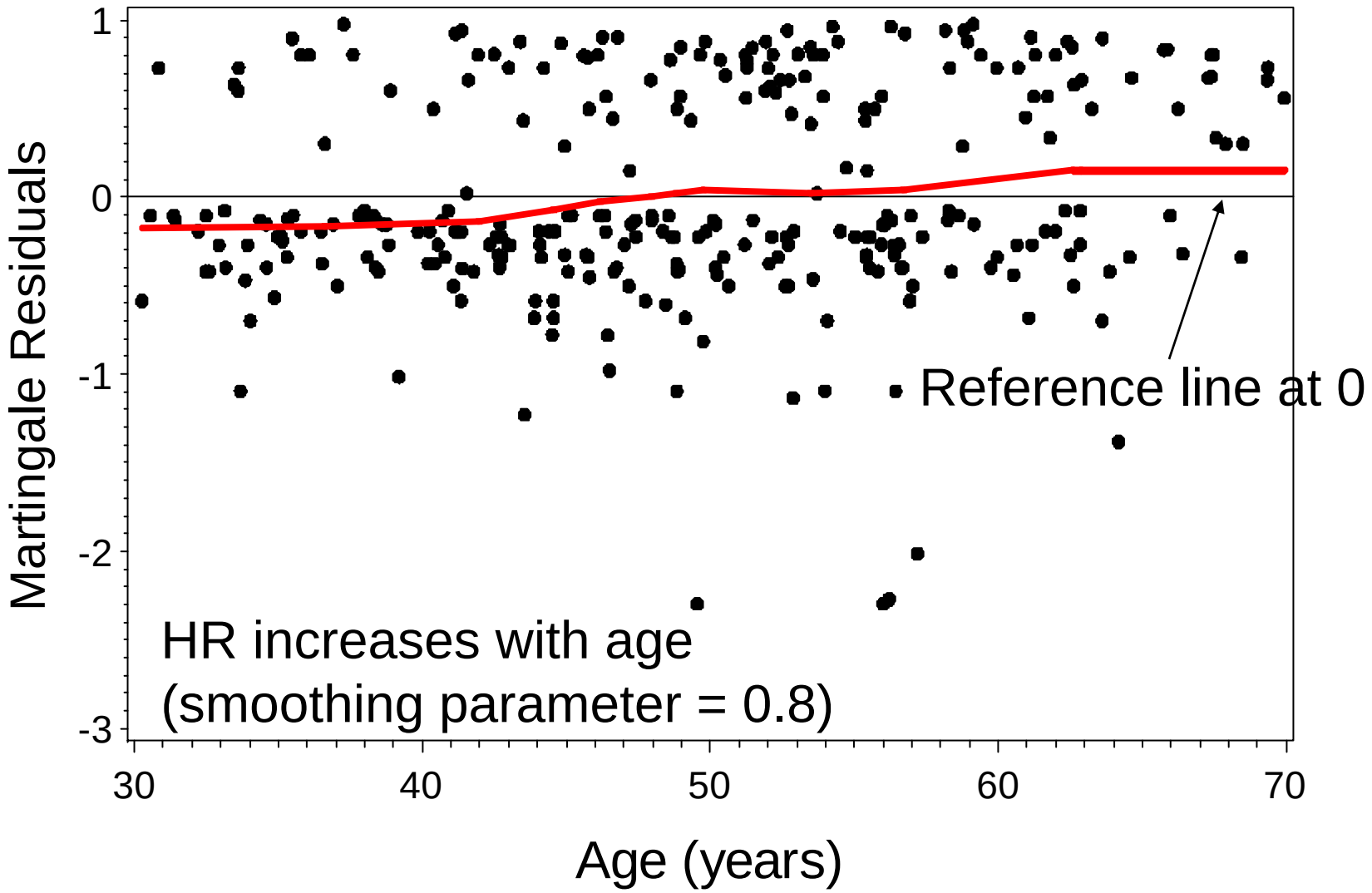


Assessing Functional Form (cont'd)

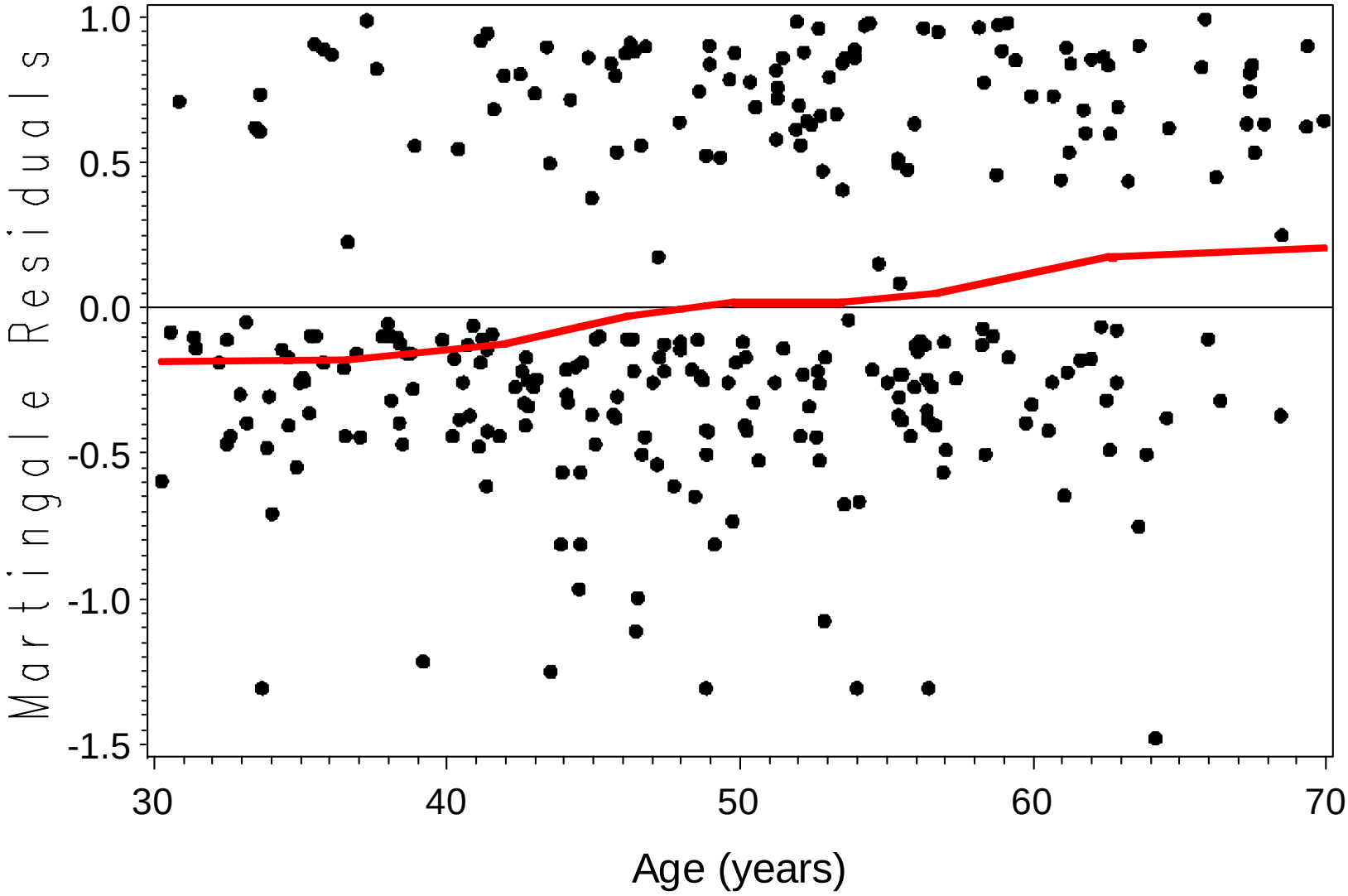
Method 2 - loess line through martingale residuals:

- Output martingale residuals from a model WITHOUT X.
  - proc phreg; model …; output out=temp resmart=mresids;
- Fit a loess line through the martingale residuals, as a function of X, and plot (several ways to do this in SAS):
  - **proc sgplot** data=temp;
  - loess y=mresids x=X / smooth=0.6; run;
- Or
  - ODS GRAPHICS ON;   /* Gives default plots */
  - **proc loess** data=temp; model mresids=X; run;
  - ODS GRAPHICS OFF;
- Look for pattern, then model X as appropriate (e.g., linear, quadratic, threshold), and re-check martingale residuals.

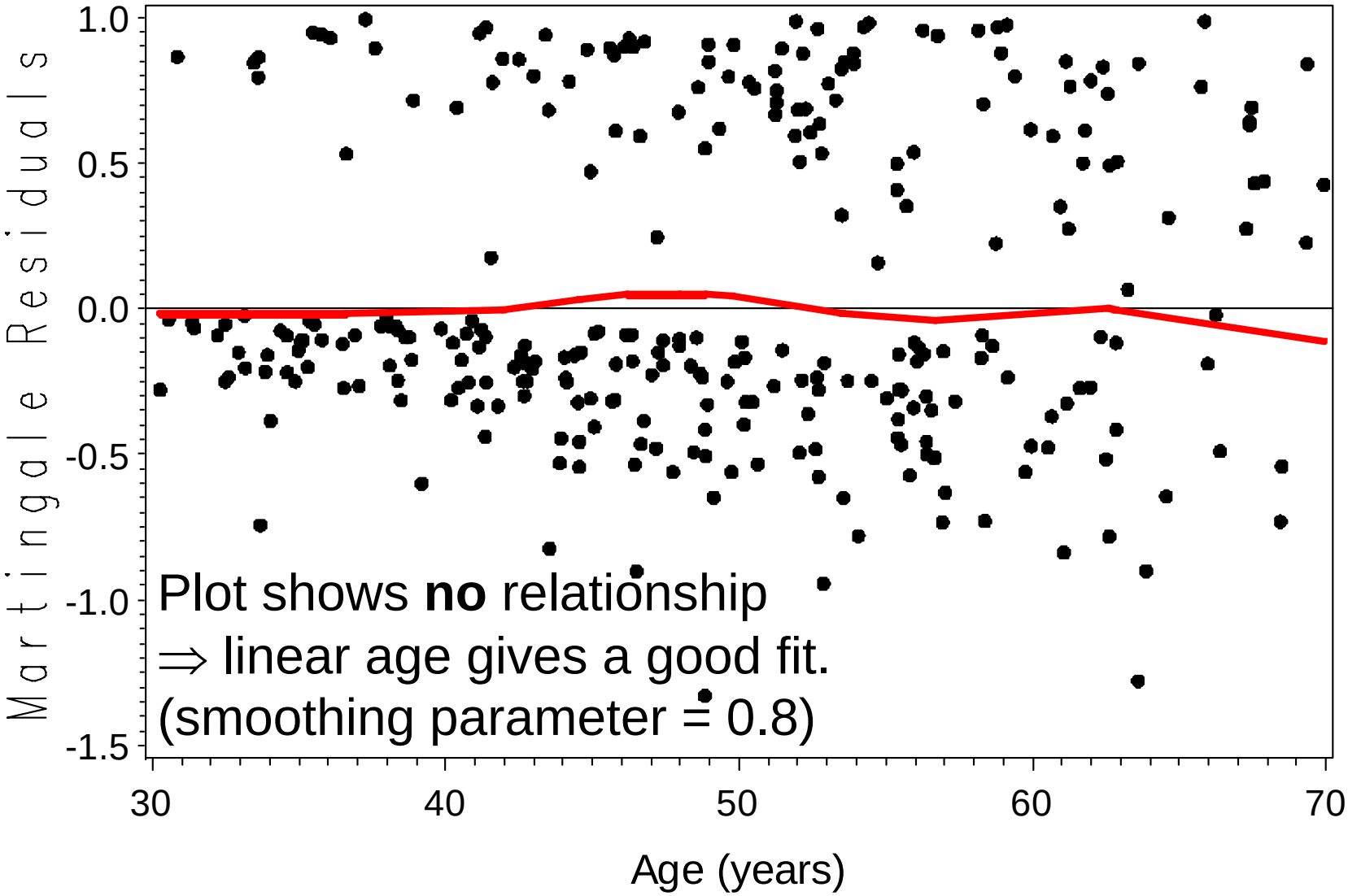
Plot of Martingale Residuals by Age, with Loess Line (Age **not** in model)



Same as previous graph, but y-axis scale starting at -1.5 instead of -6



Plot of Martingale Residuals by Age, with Loess Line (**Age in model as linear**)

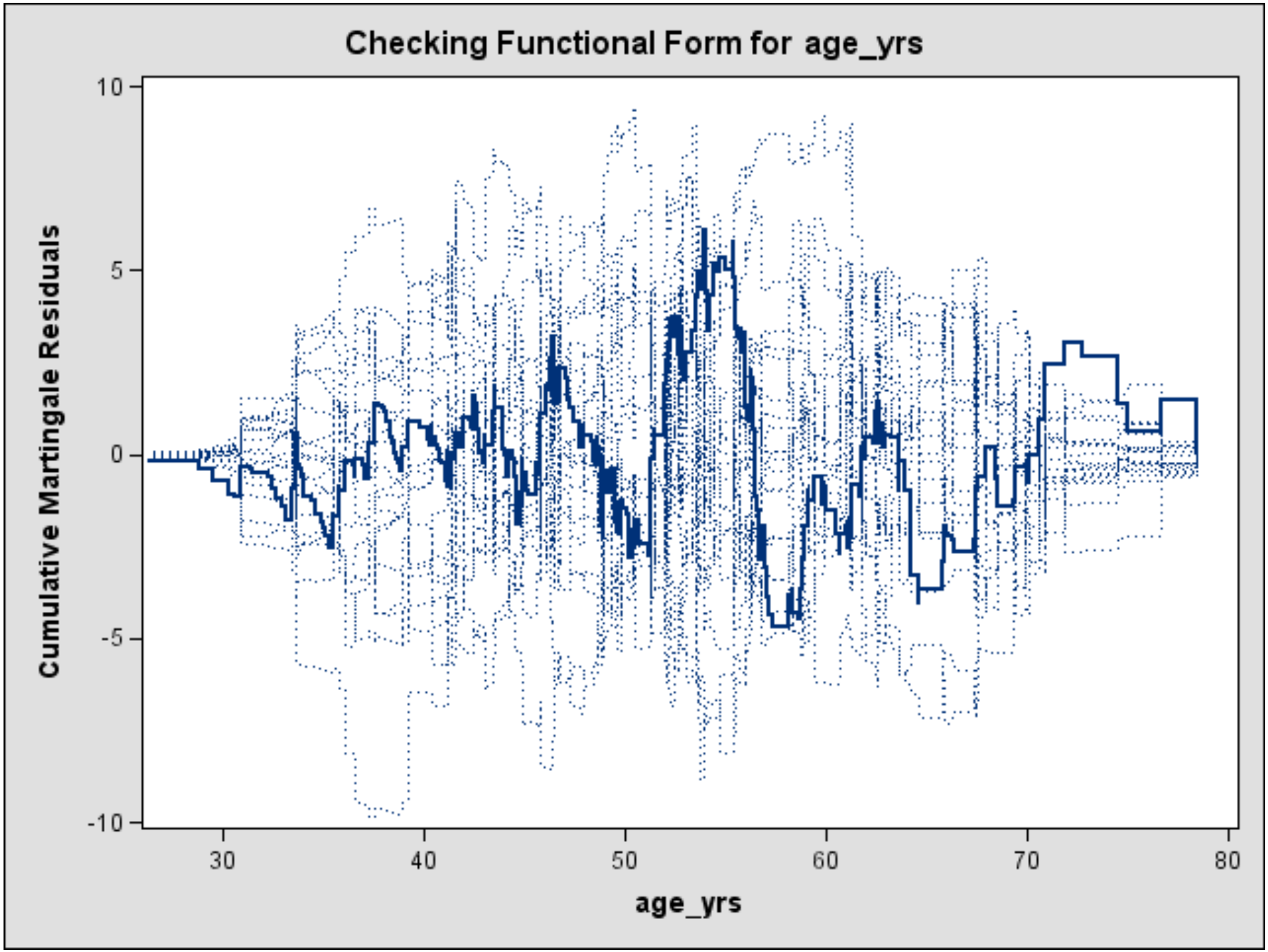


Plot of Martingale Residuals by Age, with Loess Line (Age in model as linear)

Assessing Functional Form (cont'd)

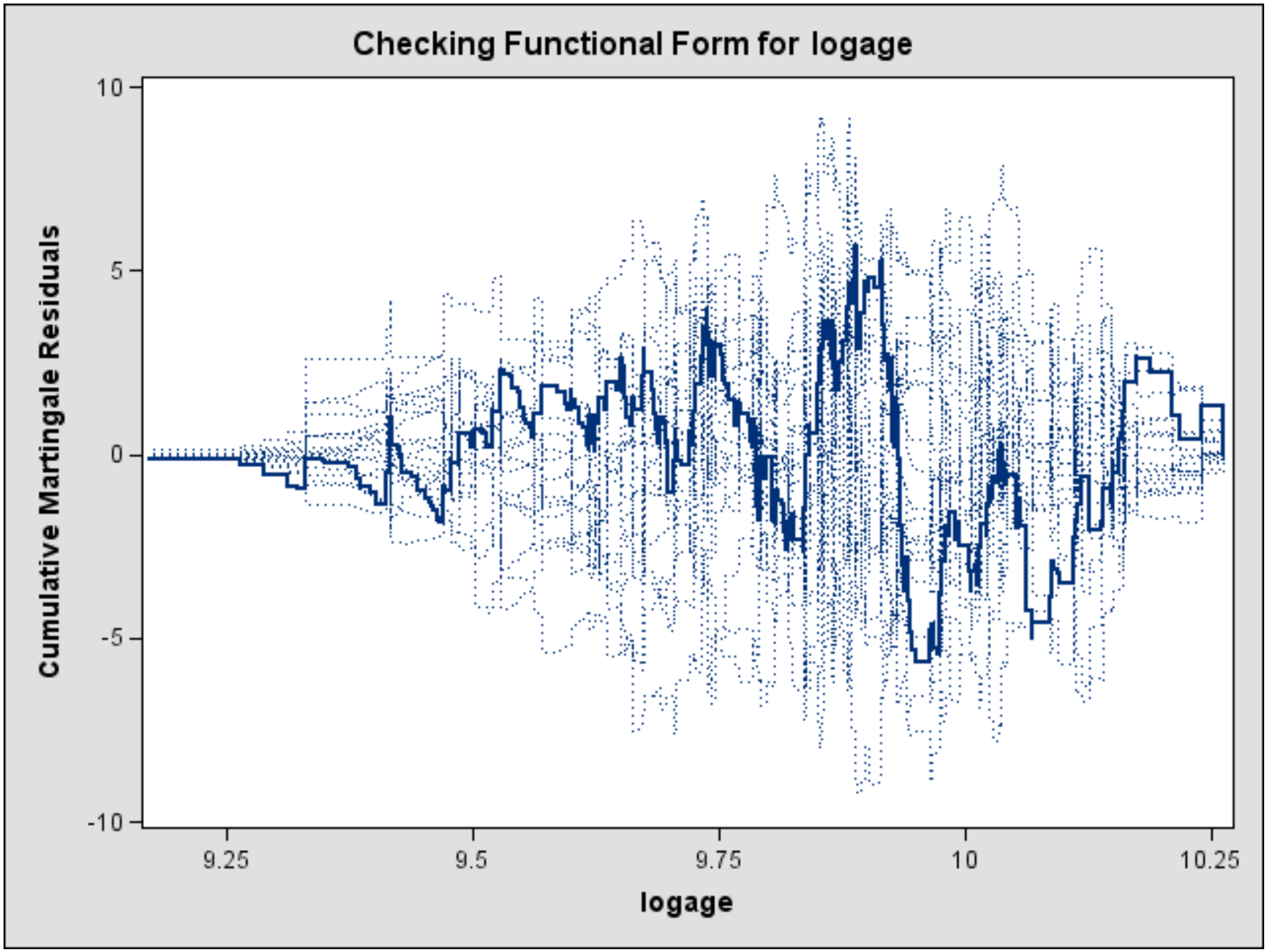
- Method 3 (ASSESS option of **proc phreg** plots cumulative sums of martingale residuals against X (to check functional form) or the observed score process against Time (to check PH):
- The following code checks Age for functional form, (and all variables (Sex, Age, Hepatom) for PH).

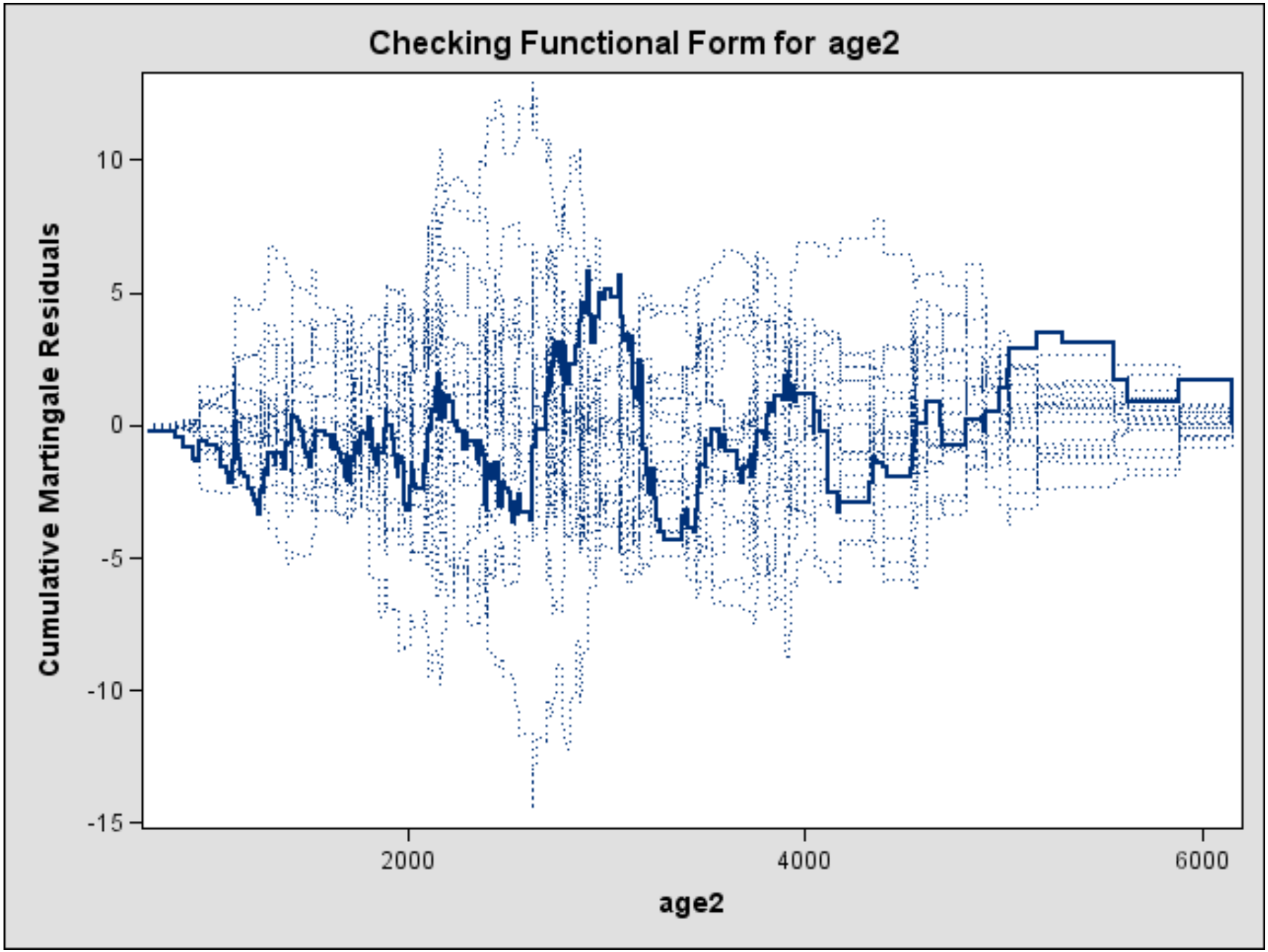
```
ods html; ods graphics on;
proc phreg data=pbc;
 model logfuday*status(0) = sex age_yrs
     hepatom;
 assess var=(age_yrs) PH / npaths=50;
run;
ods graphics off; ods html close;
```

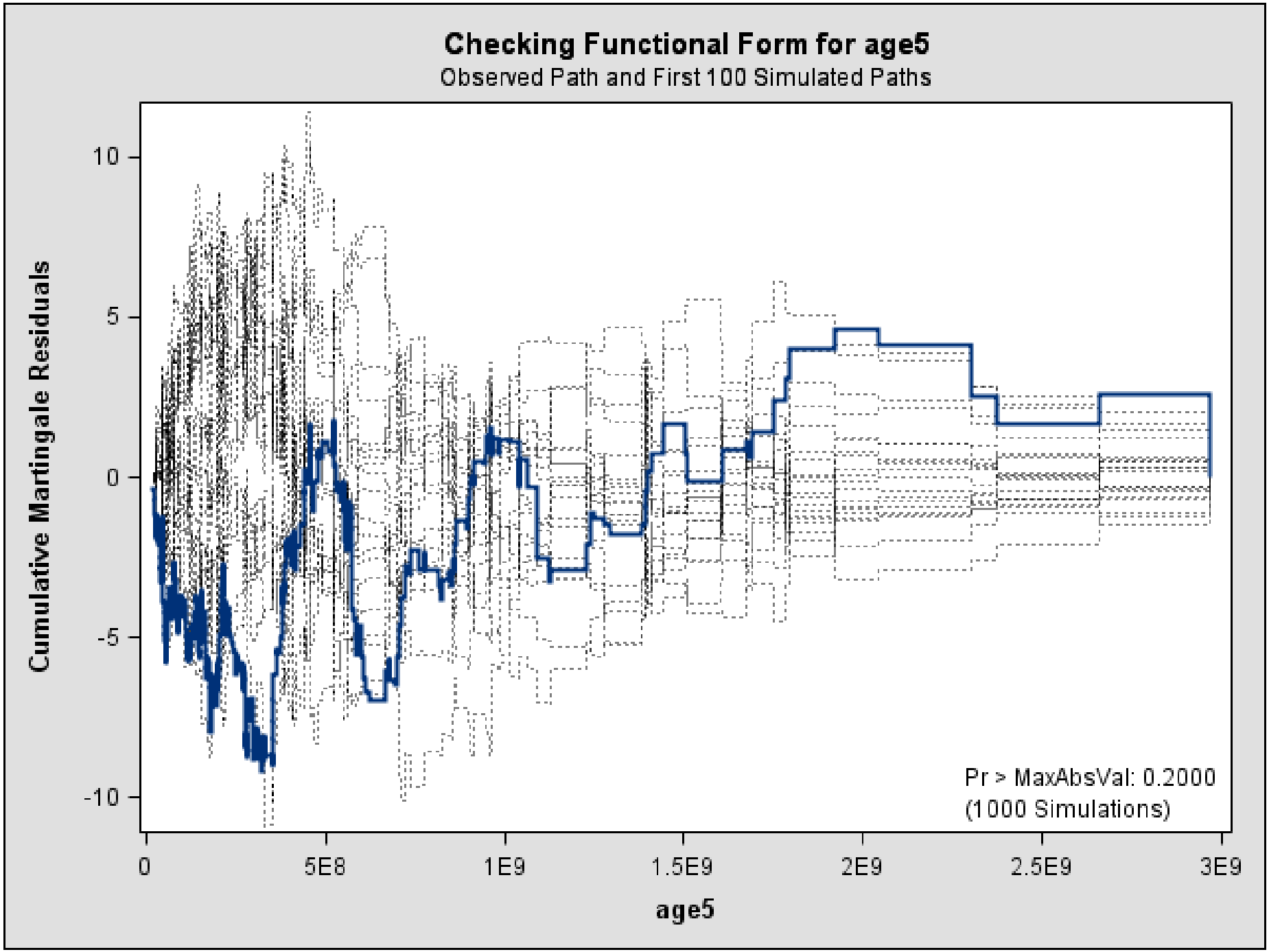

Checking Functional Form for age_yrs
Cumulative Martingale Residuals
10
5
0
-5
-10
30
40
50
60
70
80
age_yrs

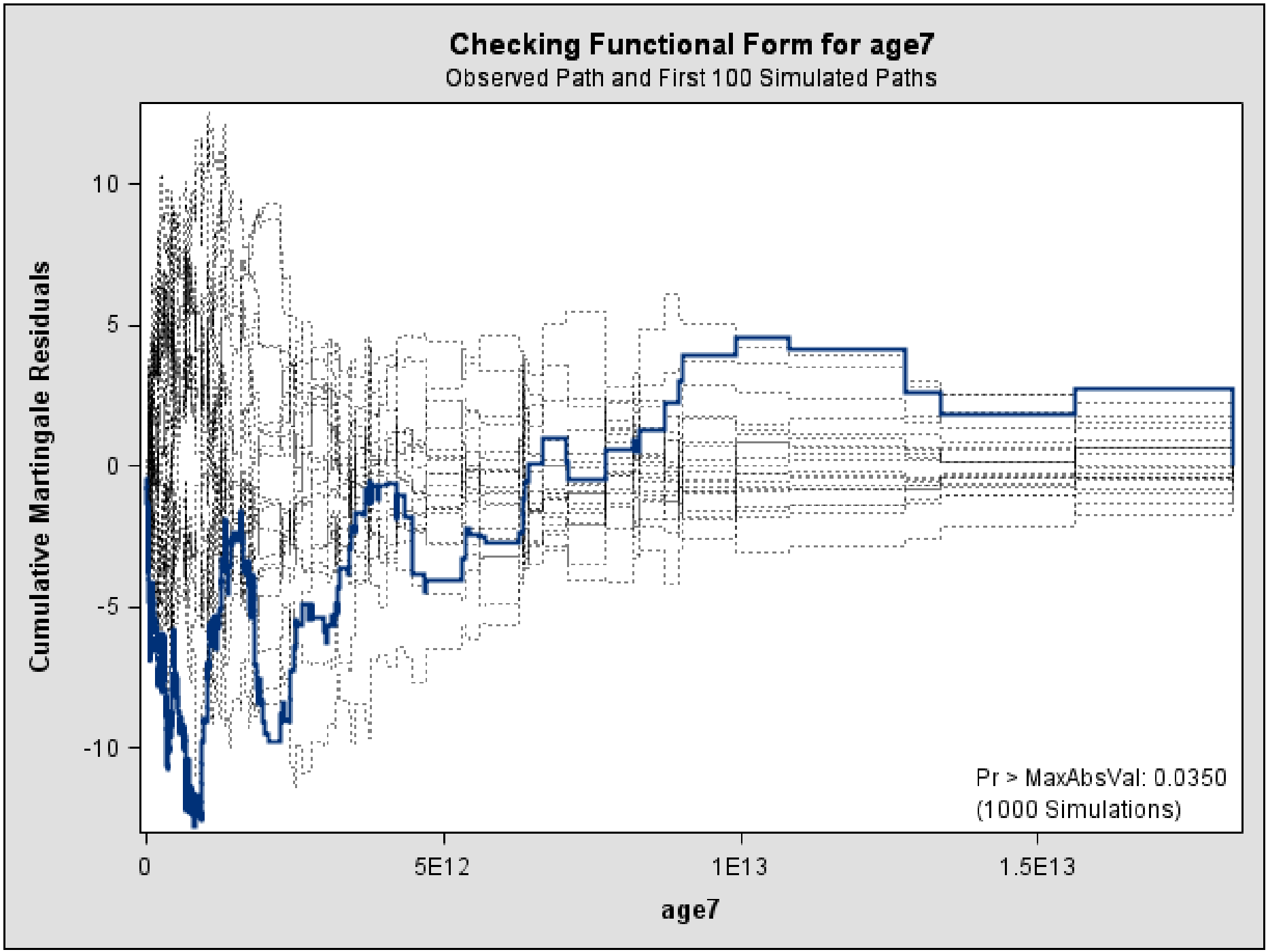
Assessing the Cumulative Martingale Residual Plot

- The first plot shows the observed curve for Age to be within the distribution of the simulated cumulative martingale residual curves, indicating acceptable fit.
- Note that we cannot check functional form with a variable out of the model. It must be included in the model in some form.
- To try to illustrate a bad fit, we try log(Age), Age^2, and Age^5. Only Age^5 shows poor fit.


Checking Functional Form for logage
Cumulative Martingale Residuals
10
5
0
-5
-10
9.25
9.5
9.75
10
10.25
logage


Checking Functional Form for age2
Cumulative Martingale Residuals
10
5
0
-5
-10
-15
2000
4000
6000
age2


Checking Functional Form for age5
Observed Path and First 100 Simulated Paths
Cumulative Martingale Residuals
10
5
0
-5
-10
0
5E8
1E9
1.5E9
2E9
2.5E9
3E9
age5
Pr > MaxAbsVal: 0.2000
(1000 Simulations)


Checking Functional Form for age7
Observed Path and First 100 Simulated Paths
Cumulative Martingale Residuals
10
5
0
-5
-10
0
5E12
1E13
1.5E13
age7
Pr > MaxAbsVal: 0.0350
(1000 Simulations)

The Resample option of ASSESS

- The Resample option of ASSESS gives
  - a test of the functional form
  - A test of PH
- Tests are based on a Kolmogorov-type supremum test using 1000 simulated patterns.
- A significant p-value indicates poor fit.
- ASSESS var=(age_yrs) PH / resample;

Supremum Test for Functional Form

Variable	Maximum Absolute Value	Pr > MaxAbsVal
Age	6.0466	0.6390
Log(Age)	6.2909	0.5960
Age^2	5.7657	0.7170
Age^5	9.1995	0.1960
Age^6	11.0837	0.0820
Age^7	12.8065	0.0350

Next, Checking Proportional Hazards for Each Variable

The checks of the PH assumption indicate PH problems only for Edema.

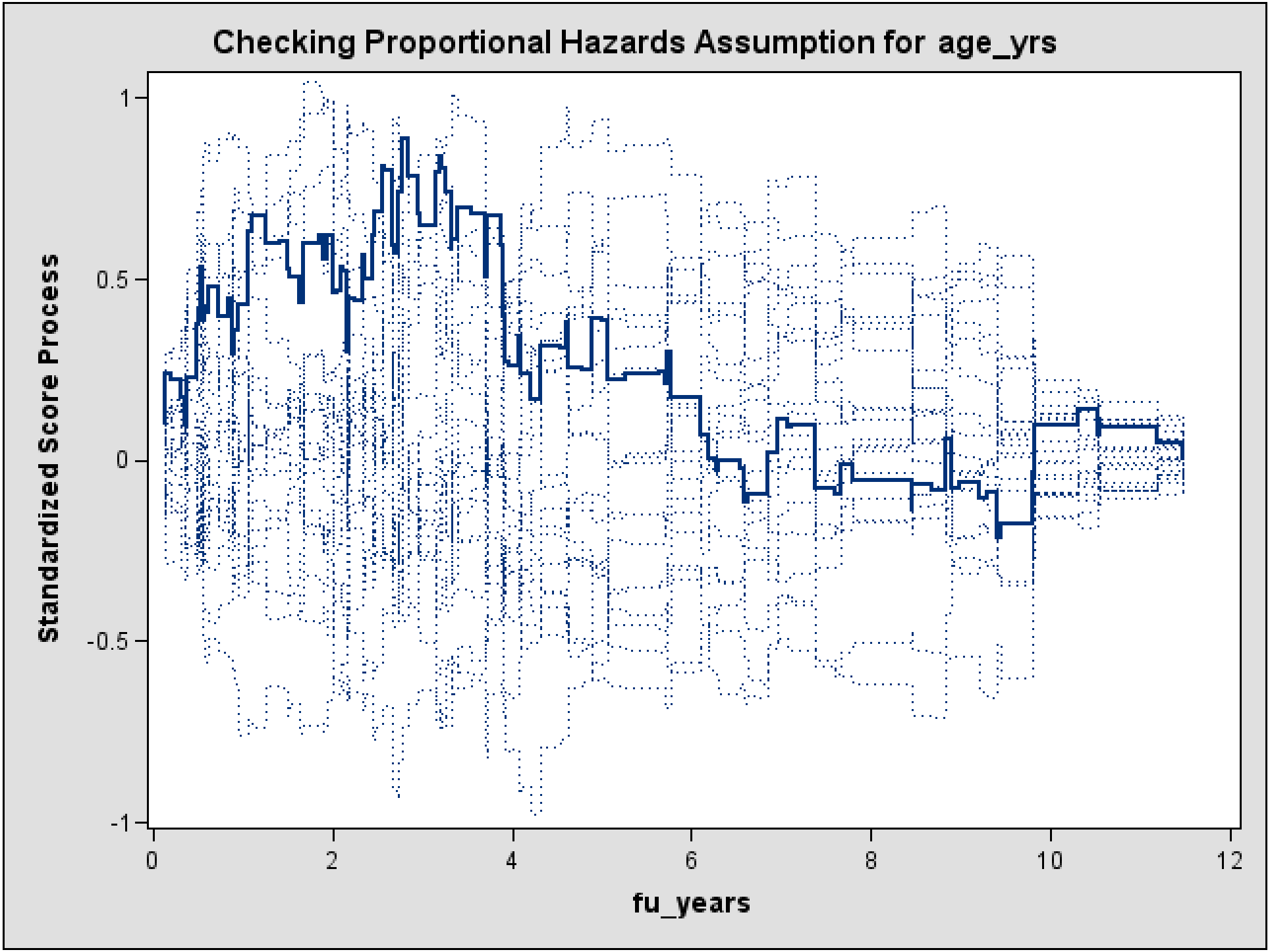
Checking PH

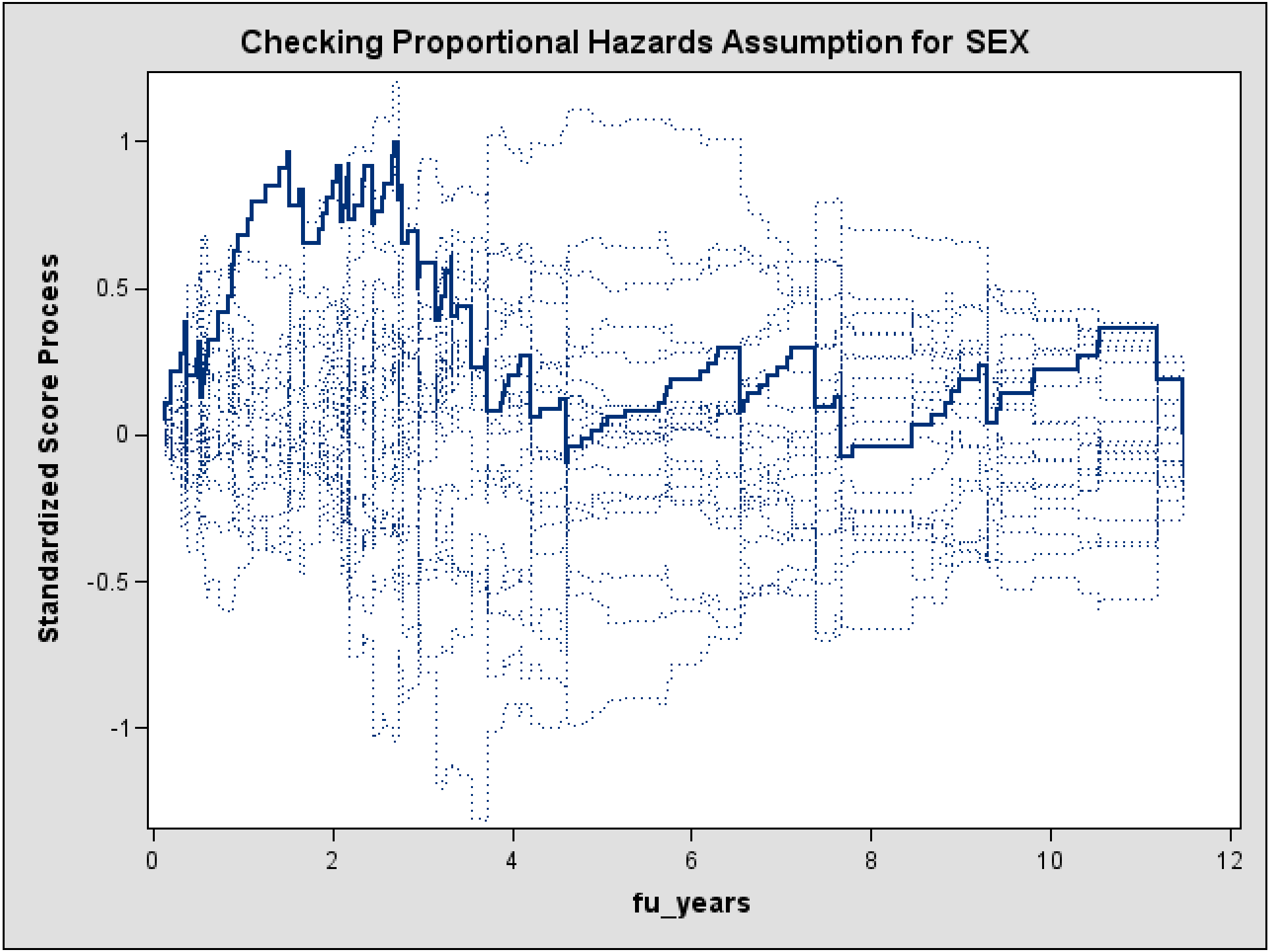
- Check PH for each covariate separately.
- If interactions are present, check PH over all interaction subgroups (e.g., Males, A; Females, A; Males, B; Females, B)
- If collinearity (confounding, treatment imbalance) is present among covariates: To check PH for x_1, estimate $S_i(t)$ for the levels of x_1 based on a Cox model stratified by x_1, with other covariates in the model. Plot

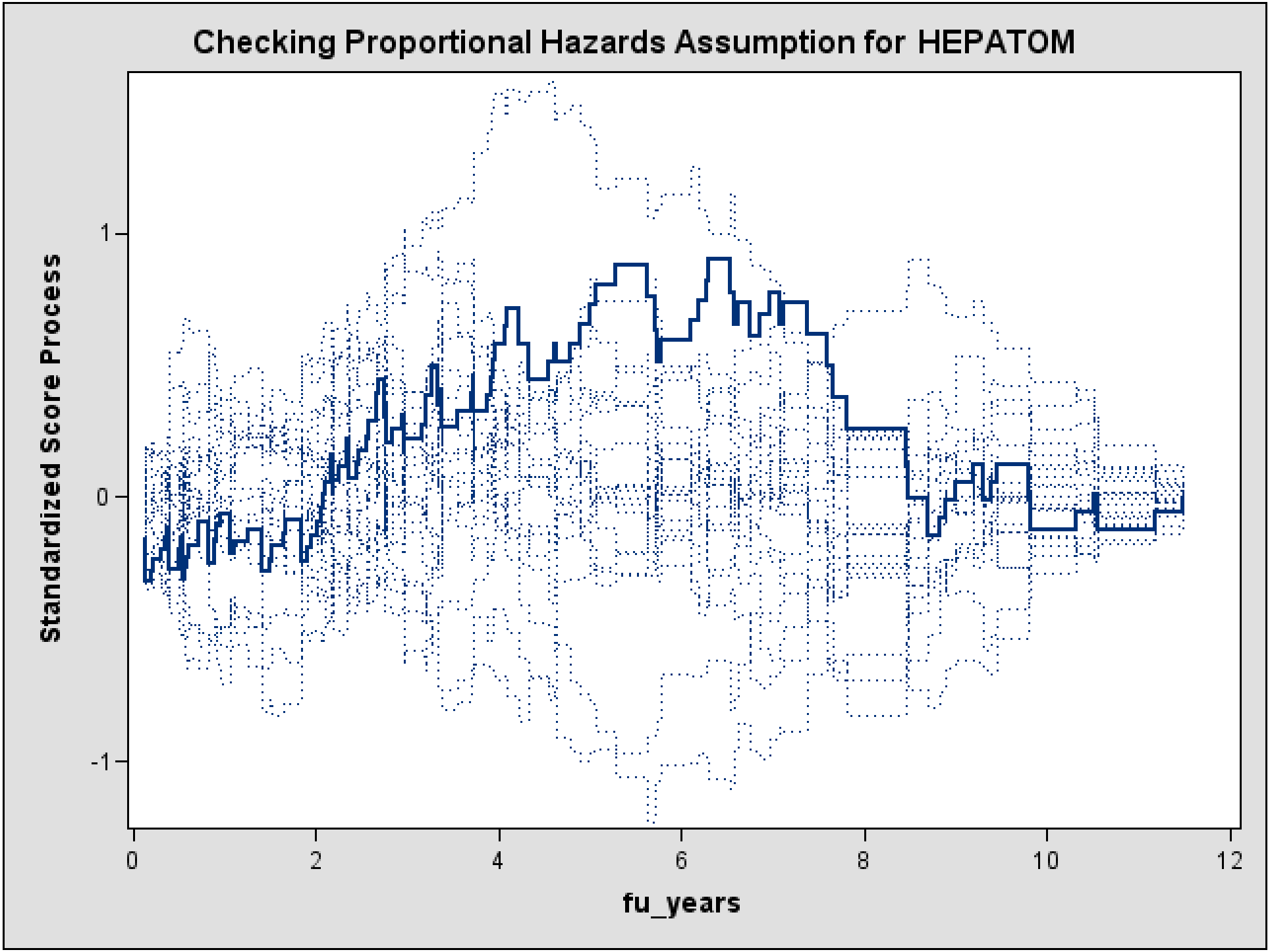
$$\log(-\log \hat{S}_i(t)) \text{ vs. } \log(t).$$

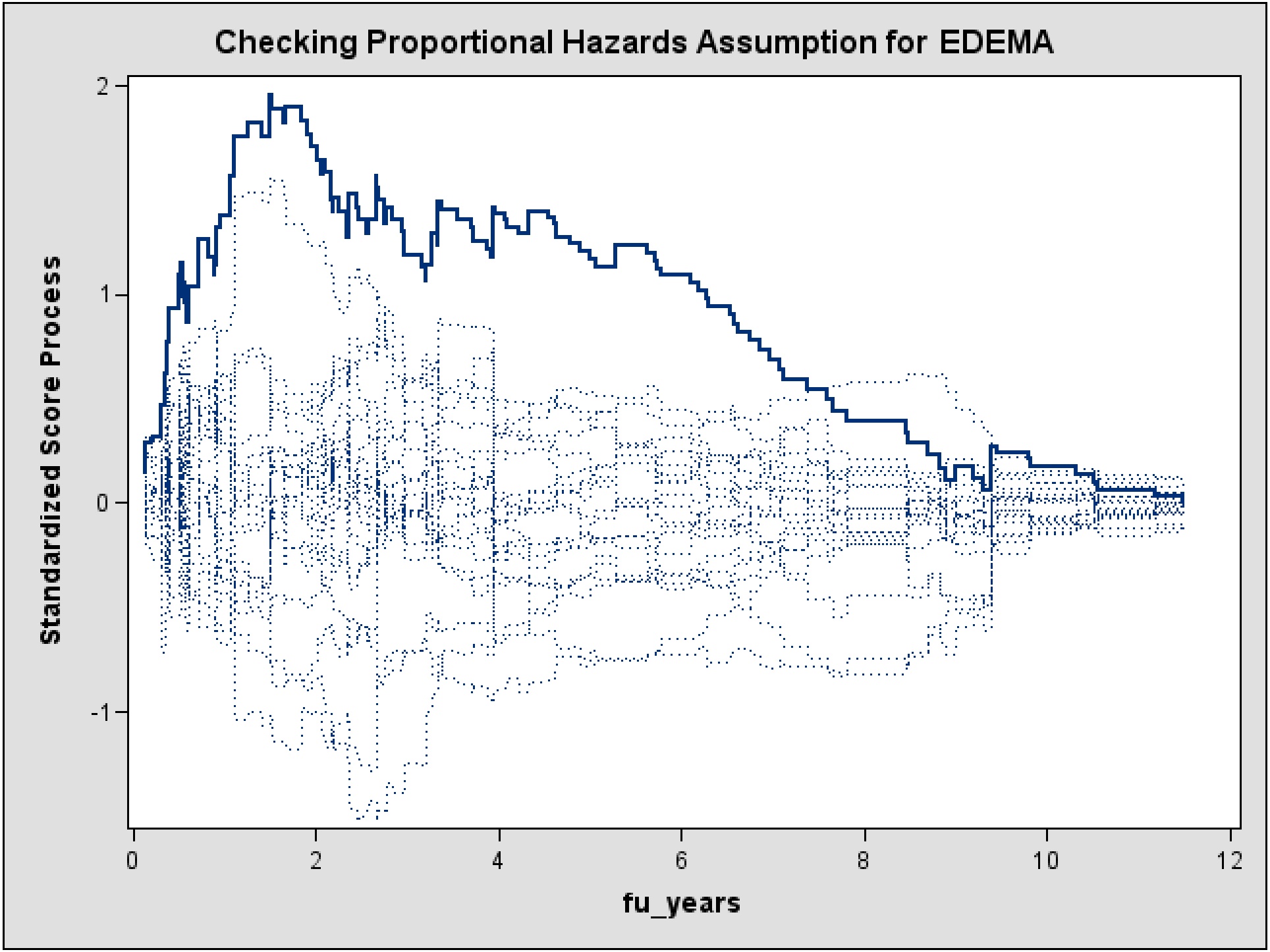
Difficulty of Checking PH

- In checking each covariate, we assume PH holds for the other covariates. Which covariate do we start with?
- If PH fails for a covariate, we should go back and re-check the others after adjusting for the non-PH of the first.
- **A wrong functional form or a missing covariate can look like non-PH.**
- Checking PH can be a difficult process.
- See Kleinbaum for more details.


Checking Proportional Hazards Assumption for age_yrs
Standardized Score Process
1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
fu_years


Checking Proportional Hazards Assumption for SEX
Standardized Score Process
1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
fu_years


Checking Proportional Hazards Assumption for HEPATOM
Standardized Score Process
1
0
-1
0
2
4
6
8
10
12
fu_years


Checking Proportional Hazards Assumption for EDEMA
Standardized Score Process
2
1
0
-1
0
2
4
6
8
10
12
fu_years

Supremum Test for Proportionals Hazards Assumption

Variable	Maximum Absolute Value	Pr > MaxAbsVal
Age	0.6153	0.7220
SEX	0.9933	0.2200
HEPATOM	0.8982	0.3310
Edema	**1.9528**	**<0.0001**

Summary of ASSESS Option

- The ASSESS option is a useful tool, but should be used in conjunction with other checks for functional form and PH.

- The cumulative martingale residual plots are not very sensitive for fine-tuning functional form. They can only detect grossly incorrect functional forms.

Bayesian Estimates are Easy Now!

```
proc phreg;   class edema;
  model time*cens(0) = age hepatom edema;
  bayes;
run;
```

The default options give posterior means and medians of estimates (in addition to ML estimated), and lots of other output.

There are MANY options for the BAYES statement. Enjoy exploration.

Comparison of ML and Bayes Estimates

		ML	Posterior	
Parameter	**DF**	**Estimate**	**Mean**	**Median**
HEPATOM	1	0.9817	0.9893	0.9872
Age	1	0.0404	0.0404	0.0404
EDEMA 0	1	-2.1350	-2.1135	-2.1211
EDEMA0 0.5	1	-1.5875	-1.5922	-1.5928

Estimates are very similar! (But the Bayes analysis may be useful to detect anomolies.)

Correlated Data

- Examples of correlated data
  - Family or facility clusters
  - repeated events for the same person (e.g., transplants)
- With correlated data at the facility level, the standard errors of β estimates will be incorrect.
  - The s.e.'s for facility-level covariates will be under-estimated, in general.
  - The s.e.'s for patient-level covariates will be over-estimated, in general.

SAS Syntax for Correlated Data

- To adjust for correlated data, proc phreg will compute **robust variances** based on the sandwich estimator with an independence working correlation matrix.
- The StdErrRatios are also given – the ratios of the robust to the model-based s.e. estimates.
- Let the facility id variable be called “facid”.

```
Proc phreg covsandwich(aggregate);
  model fu_days*status(0) = age sex drug;
  id facid;
run;
```

Conclusions

- Thank you, SAS, for making our lives easier (and more fun!)
- The many new options for PHReg are a big step forward.
- We look forward to more new options in the future.
  - Frailty models
  - Interaction plots??

www.fields.utoronto.ca/.../ DLSS/cox/Cox3.JPG