1 Diagnostics for Cox models

Assumptions:

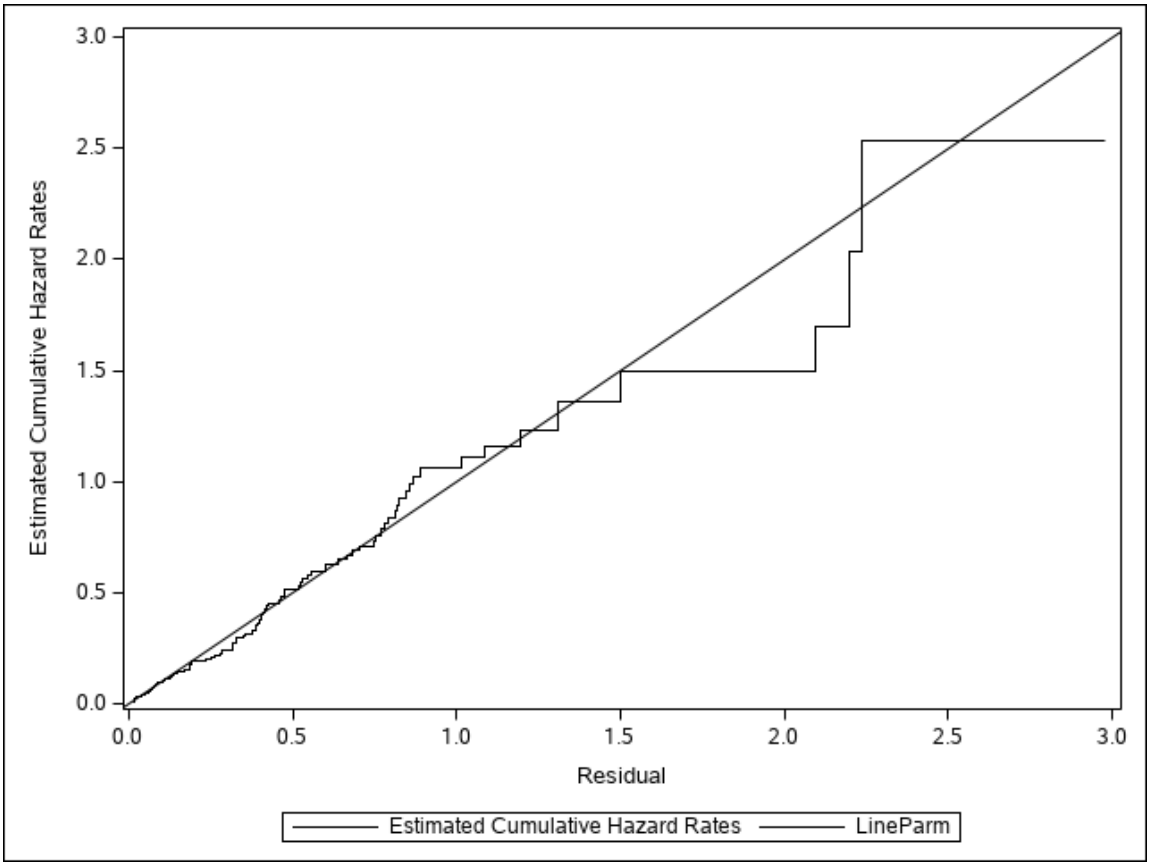
- (1) adequacy of proportional hazard assumption
- (2) for a given covariate, we would like to see the best functional form to explain the influence of the covariate on survival, adjusting for other covariates.
    - Cox–Snell Residuals – overall fit
    - Martingale Residuals – f(Z)
    - PH:
        - Log-log survival plot
        - Observed vs Expected plot
        - Schoenfeld residuals
        - time-dependent covariates

1.1 Cox–Snell Residuals

- X true death time
- $H\left(X_j \mid \mathbf{Z}_j\right) \sim Exp(1)$
- if the proportional hazards model $h\left(t \mid \mathbf{Z}_j\right) = h_o(t) \exp\left(\Sigma \beta_k Z_{jk}\right)$
- $r_j = \hat{H}_o\left(T_j\right) \exp\left(\sum_{k=1}^{p} Z_{jk} b_k\right), j = 1, \ldots, n$
    - $\hat{H}_o(t)$ is Breslow's estimator of the baseline hazard rate
    - $\boldsymbol{\beta}$ from Cox model
- r_j 's should look like a censored sample from a unit exponential distribution
- We know: $H_{\mathrm{E}}(t) = t$
- then, Nelson-Aalen estimator of the cumulative hazard rate of the r_j 's,$\hat{H}_r\left(r_j\right)$
- if the overall model fit is good, $\hat{H}_r\left(r_j\right)$ versus r_j should be a straight line through the origin with slope 1

```
* A Cox regression model is fit to the data with these eight covariates
* residuals are computed;
* the default method of estimating a survivor function is Breslow (1972) estimator, i.e..method
proc phreg data = bone_marrow1;
  model t2*dfree(0) = z1 z2 z1xz2 g2 g3 z7c z8 z10 ;
  output out = plot1_1 LOGSURV = logsurv1 /method = ch;  /*-logsurv is the cox-snell residual*/
run;

data plot1_1;
  set plot1_1;
```

```
  snell = -logsurv1;
  cons = 1;
run;
proc phreg data = plot1_1;
  model  snell*dfree(0) = cons;
  output out = plot1_2 logsurv = logsurv2 /method = ch;
run;
data plot1_2;
  set plot1_2;
    cumhaz = - logsurv2;
run;
proc sort data = plot1_2;
 by snell;
run;

proc sgplot data = plot1_2;
step y=cumhaz x=snell /MARKERFILLATTRS=(color="red");
lineparm x=0 y=0 slope=1; /** intercept, slope **/
  label cumhaz = "Estimated Cumulative Hazard Rates";
  label snell = "Residual";
run;
```



- not too bad!
- The Cox-Snell residuals are most useful for examining the overall fit of a model. A drawback of their use is they do not indicate the type of departure from the model.
- departures from the exponential distribution may be partly due to the uncertainty in estimating β and $H.$, esp in small samples and in right-hand tail of the distribution.

1.2 Martingale Residuals

- determining the functional form to be used for a given covariate
- $\hat{M}_j = \delta_j - \hat{H}_o(T_j) \exp(\sum_{k=1}^{p} Z_{jk} b_k) = \delta_j - r_j, j = 1, \ldots, n$
- difference over time of the observed number of events minus the expected number of events under the assumed cox model
- martingale residuals: estimate of the excess number of events seen in the data but not predicted by the model

- $H\left(t \mid \mathbf{Z}^*, Z_1\right) = H_o(t) \exp\left(\boldsymbol{\beta}^* \mathbf{Z}^*\right) \exp\left[f\left(Z_1\right)\right]$
- To find f,
  - fit a cox model to the data based on $\mathbf{Z}^*$ and compute the martingale residuals, $\hat{M}_j$
  - plot resuduals against Z_1
  - smoothed-fitted curve gives an indication of the function f.
    * If the plot is linear, then, no transformation of Z_1 is needed.
    * ...

```
* Martingale residual;
* leave out patient age;
proc phreg data = bone_marrow1;
  model t2*dfree(0) = z2 z1xz2 g2 g3 z7c z8 z10 ;
  output out = plot2_1 RESMART = mgale ;
run;

ods listing gpath='/folders/myfolders/Lab8/';
ods graphics / imagename="p2" imagefmt=png;

proc loess data=plot2_1;
  model mgale = z1 / smooth=0.6 direct;
run;
```

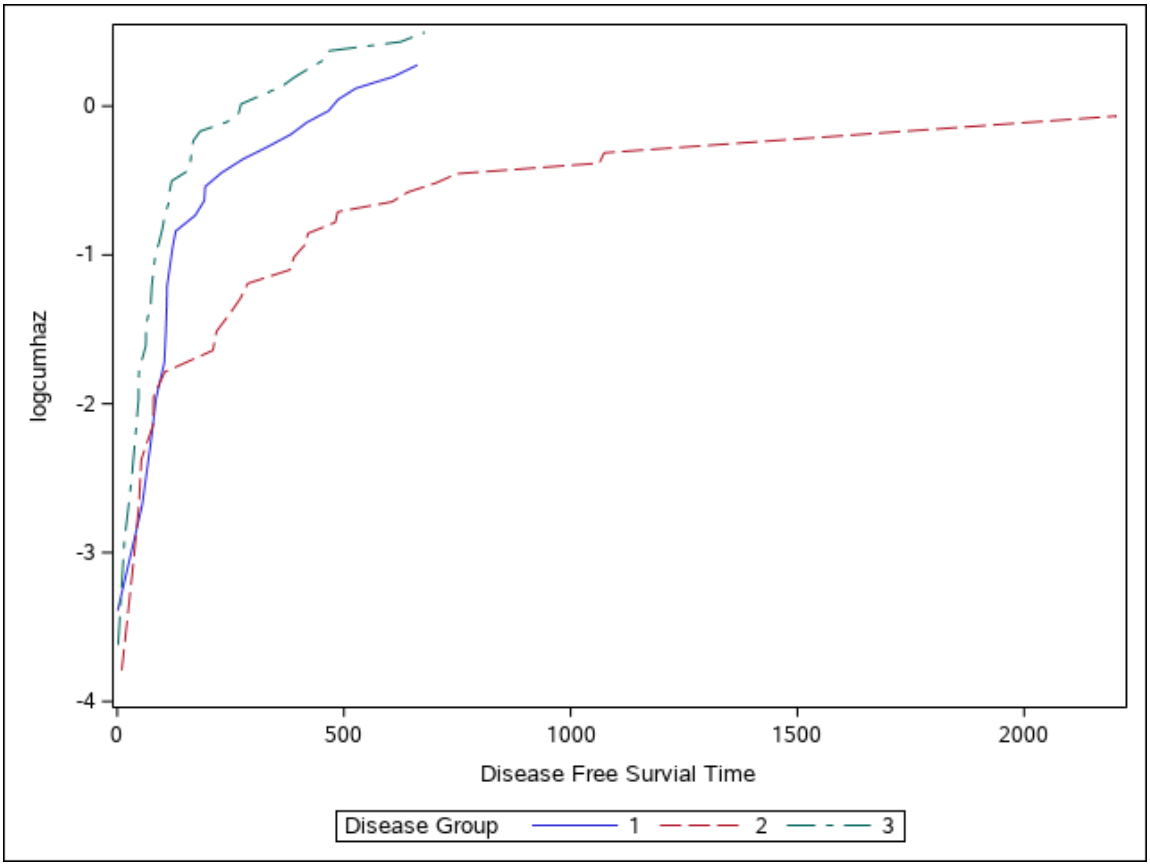
- smoothed plot of $\hat{M}_j$ versus patient age suggest correct functional form is linear.

2 Log-log Survival Plot

Z_1 is/be discretized

- fit a Cox model stratified on the discrete values of Z_1
- let $\hat{H}_{go}(t)$ be the estimated cumulative baseline hazard rate in the gth stratum.
- If the proportional hazards model holds, baseline cumulative hazard should exhibit proportionality property $H_{go}(t) = \exp(\gamma_g) H_{10}(t)$
  - plot $\ln\left[\hat{H}_{1o}(t)\right], \ldots, \ln\left[\hat{H}_{Ko}(t)\right]$ versus t. If the assumption holds, then, these should be approximately parallel, by a constant

```
data bone_marrow ;
  set bone_marrow ;
  cons = 1;
run;

* The baseline cumulative hazards are estimated using
* Breslow's estimator for each stratu;
proc phreg data = bone_marrow  ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =cons/rl ; * risk limit: provide CI for harzard ratio;
 strata g;
 output out = base logsurv = ls /method = ch;

 run;

data base;
  set base ;
    logH = log (-ls);
        if g= 1 then logH1 = logH;
    else if g= 2 then logH2 = logH;
    else if g= 3 then logH3 = logH;
    proc sort;by g t2  ;
    proc print;var g t2 logH logH1 logH2 logH3;
run;

ods listing gpath='/folders/myfolders/Lab8/';
ods graphics / imagename="p3" imagefmt=png;
proc sgplot data =base;
where logH ne .;
series x=t2 y=logH /group=g ;
run;
```

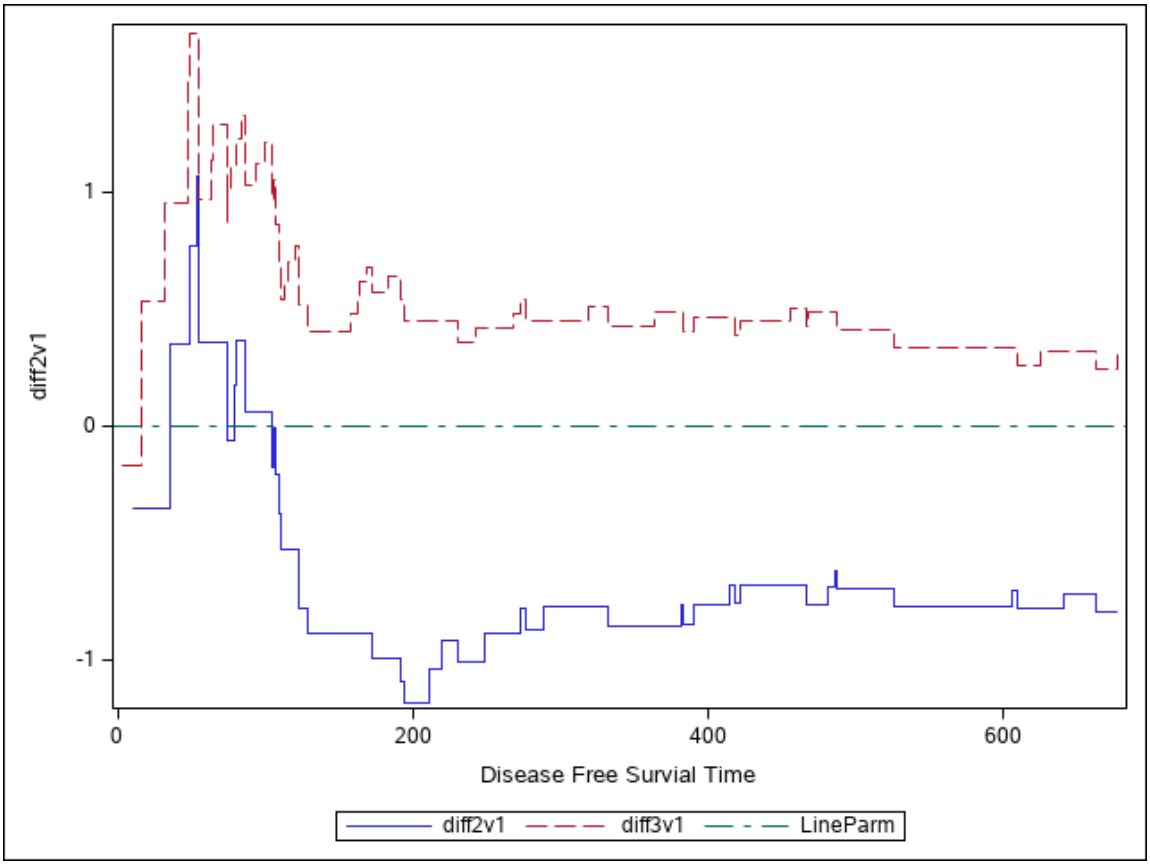


  - $*$ seem to cross at the beginning, look at time less than 700

- Or, plot $\ln\left[\hat{H}_{go}(t)\right] - \ln\left[\hat{H}_{1o}(t)\right]$ versus t for $g = 2, \ldots, K$. each curve should be roughly constant.

```
* or;
proc sort data = base;
 by t2;
run;
data base;
  set base;
  retain temp1 temp2 temp3;
   if logH1 ~= . then temp1 = logH1;
   if logH2 ~= . then temp2 = logH2;
   if logH3 ~= . then temp3 = logH3;
   diff2v1 = temp2 - temp1;
   diff3v1 = temp3 - temp1;
   proc print;
   var t2 g diff2v1 diff3v1;
```

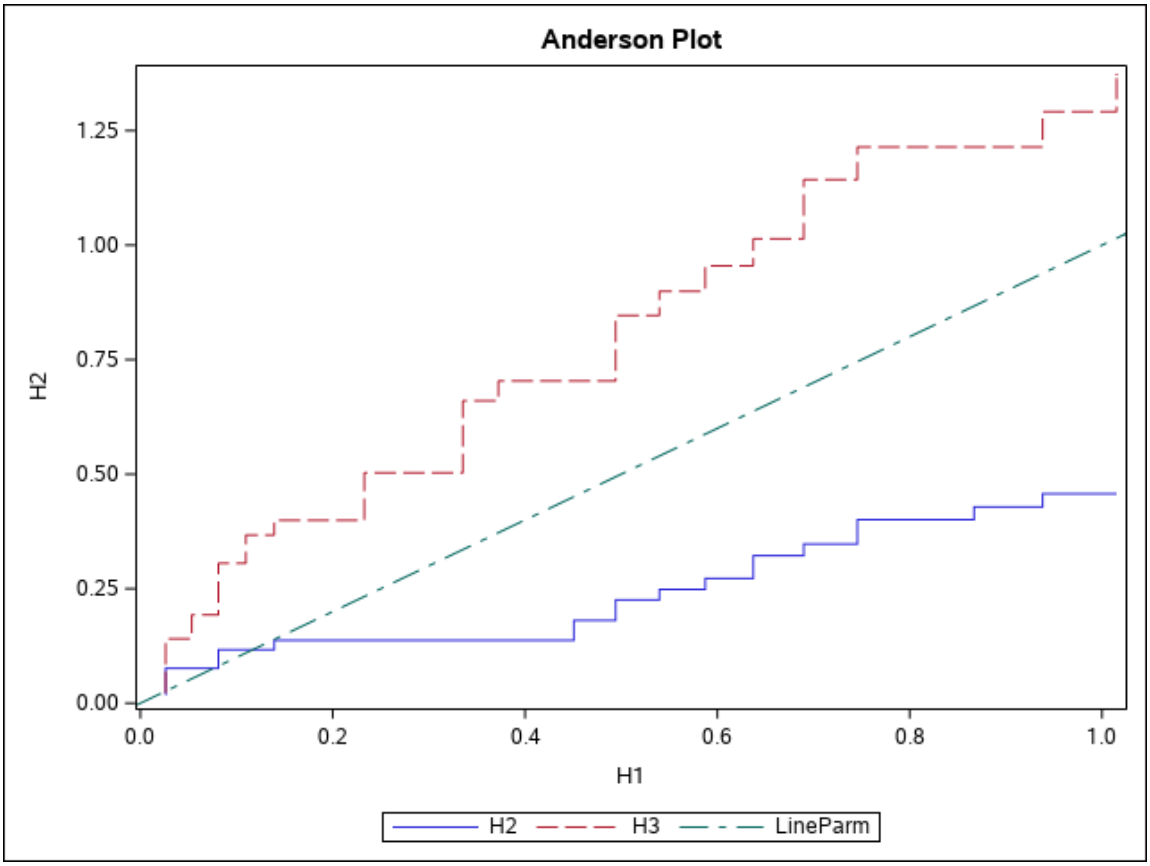
```
run;

ods listing gpath='/folders/myfolders/Lab8/';
ods graphics / imagename="p4" imagefmt=png;
proc sgplot data =base;
where t2<=700;
step x=t2 y=diff2v1  ;
step x=t2 y=diff3v1  ;
lineparm x=0 y=0 slope=0; /** intercept, slope **/

run;
```



- proportional hazards assumption is rejected because the plotted curve is not roughly constant over time.
- This figure shows an early advantage for stage 1; After 6-month, advantage still hold for stage 3 versus stage 1, but stage 2 risk is lower than stage 1.
- guess what beta should be if cox model was applied?

– Or, Anderson plot, should be straight lines, $\hat{H}_{go}(t)$ versus $\hat{H}_{10}(t)$ for $g = 2, \ldots, K$. slope of these lines is a crude estimate of $\exp(\gamma_g)$

```
* or;
proc phreg data = bone_marrow  ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =cons/rl ; * risk limit: provide CI for harzard ratio;
 strata g;
 output out = base logsurv = ls /method = ch;

 run;
proc sort data = base;
 by t2;
run;
data base;
  set base;
  retain H1 H2 H3;
  *retain its value from one iteration of the DATA step to the next;
  if g=1 then H1 = -ls;
  if g=2 then H2 = -ls;
  if g=3 then H3 = -ls;
   proc print;
   var t2 g ls H1 H2 H3 ;
run;

ods listing gpath='/folders/myfolders/Lab8/';
ods graphics / imagename="p5" imagefmt=png;
proc sgplot data =base;
title "Anderson Plot";
where t2<=700;
series x=H1 y=H2 ;
series x=H1 y=H3;
lineparm x=0 y=0 slope=1; /** intercept, slope **/
run;
```



∗ If the model held, we would have expected a linear plot through the origin, but this is not the case in this plot.

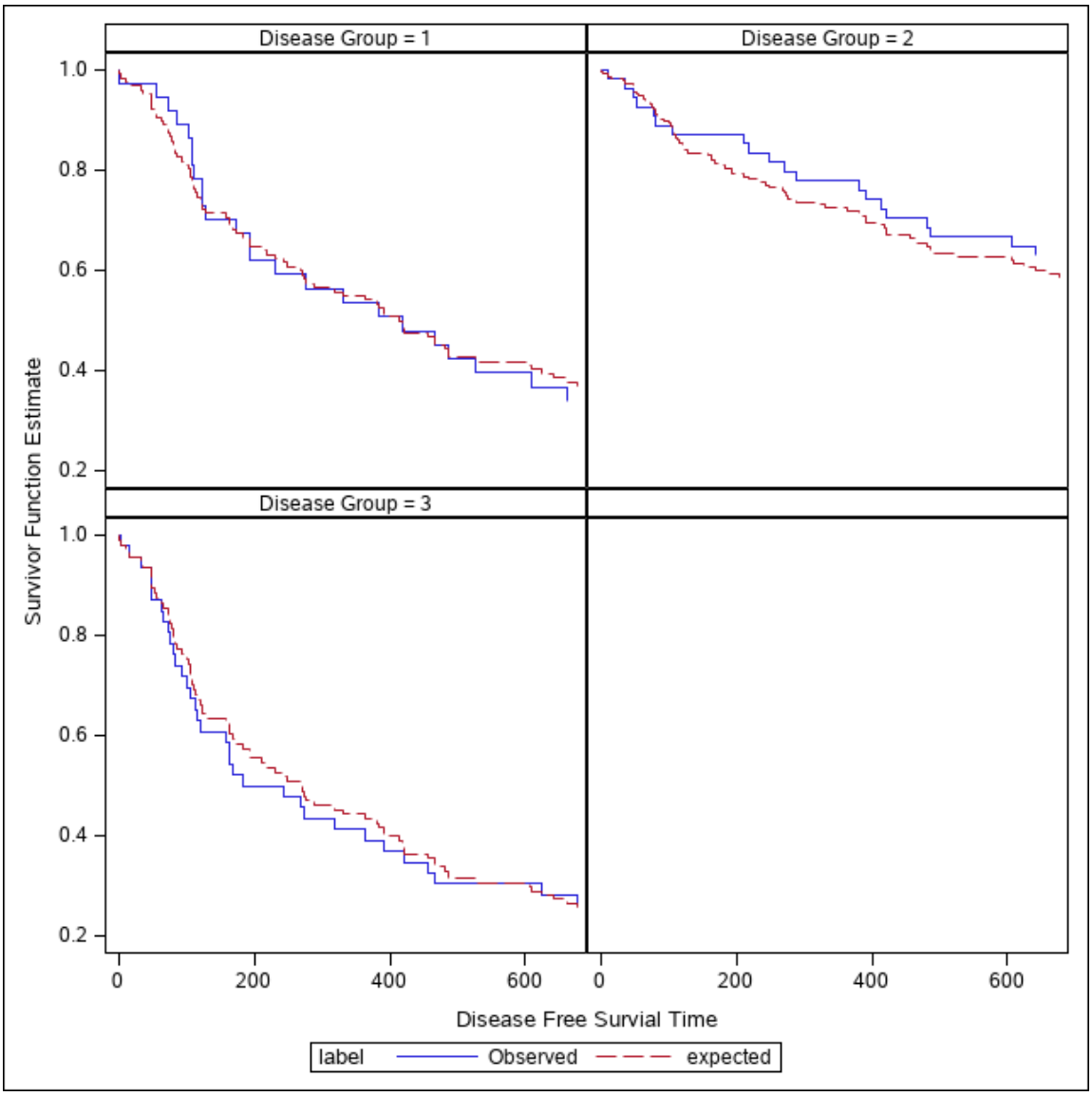
3 Observed vs Expected plot

Graphical GOF test

- observed plot: estimate KM curves for each group of interest, can adjust for other variables by cox model
- expected plot: fit another cox model including the group variable of interest, get the baseline survival curve for each group by setting other covariates to some fixed level, such as means or other prespecified levels

```
* observed versus expected, by setting other covariates to the same level;

data cov;
input g z1;
cards;
1 30
2 30
3 30
;
run;

* observed;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =z1/rl ;
 strata g;
 baseline covariates=cov out=observed LOGLOGS=loglogs survival=survival cumhaz=cumhaz/DIRADJ  ;
run;
* expected;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =z1 g/rl ;
 baseline covariates=cov  out=expected LOGLOGS=loglogs survival=survival cumhaz=cumhaz/DIRADJ g
run;

data all;
set observed(in=in1) expected(in=in2);
if in1=1 then label="Observed";
else if in2=1 then label="expected";
proc sort;by label t2;
run;


ods listing gpath='/folders/myfolders/Lab9/';
ods graphics / imagename="p6" imagefmt=png;
proc sgpanel data =all ;
panelby g/columns=3;
```

```
where t2<=700;
step x=t2 y=survival /group=label  ;
run;
```



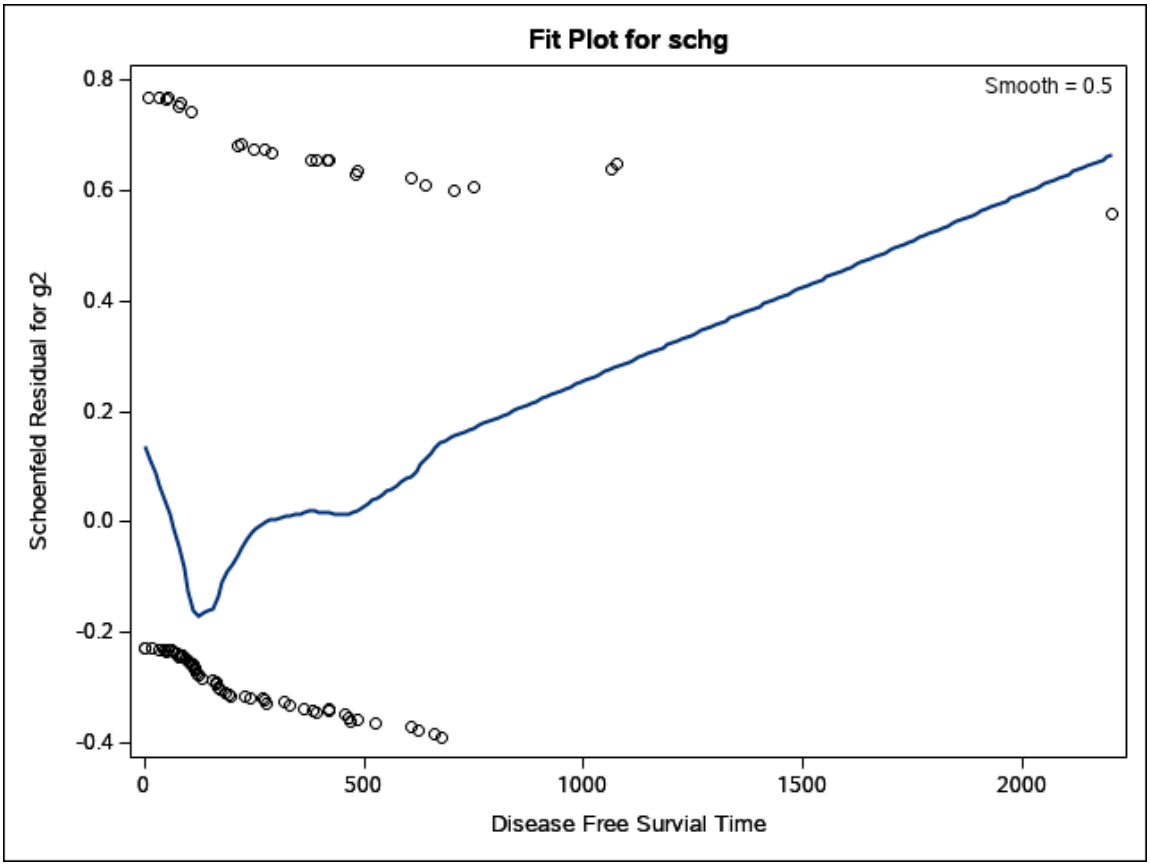
4 Schoenfeld residual

- The Schoenfeld residual for observation j and covariate p is defined as the difference between covariate p for observation j and the weighted average of the covariate values for all subjects still at risk when observation j experiences the event.
- a scaled version of the Schoenfeld residual at time k for a particular covariate p will approximate the change in the regression coefficient at time k :

$$E\left(s_{kp}^{\star}\right)+\hat{\beta}_{p} \approx \beta_{j}\left(t_{k}\right)$$

- In the relation above, $s_{kp}^{\star}$ is the scaled Schoenfeld residual for covariate p at time k, β_{p} is the time-invariant coefficient, and $\beta_{j}\left(t_{k}\right)$ is the time-variant coefficient.
- In other words, the average of the Schoenfeld residuals for coefficient p at time k estimates the change in the coefficient at time k.
- Thus, if the average is 0 across time, then that suggests the coefficient p does not vary over time and that the proportional hazards assumption holds for covariate p.
  - SAS provides Schoenfeld residuals for each covariate, and they are output in the same order as the coefficients are listed in the "Analysis of Maximum Likelihood Estimates" table. Only as many residuals are output as names are supplied on the ressch= option.

```
* Shoenfeld residual;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =z1 z2 g/rl ;
 output out=schoen
 ressch=schz1 schz2 schg  ;
run;

ods listing gpath='/folders/myfolders/Lab9/';
ods graphics / imagename="p7" imagefmt=png;
proc loess data = schoen plots=FITPLOT;
model schg=t2 / smooth=(0.5);
run;
```


Fit Plot for schg
Smooth = 0.5
Schoenfeld Residual for g2
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
2000
Disease Free Survial Time

5 formally test by time-dependent covariates

- To perform the test, we define $Z_2(t) = Z_1 \times \ln t$ and fit the Cox model with covariates Z_1 and $Z_2(t)$.
- is a constant only if $\beta_2 = 0$. This is the rationale for testing the local hypothesis H_o : $\beta_2 = 0$ to check the proportional hazards assumption.

```
data bone_marrow ;
set bone_marrow ;
where g in (1,3);
 if g=3 then g3=1;else if g=1 then g3=0;
 g3t=log(t2)*g3;
 proc print ;var g t2 g3 g3t;
 run;

proc phreg data = bone_marrow  ;
class g3(ref="0")/param=ref;
 model t2*dfree(0) =g3 g3t/rl ; * risk limit: provide CI for harzard ratio;
 run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
g3	1	1	13.00004	1.82773	50.5902	<.0001	442428.9	12304.92	15907725	g3 1
g3t		1	-2.12939	0.31941	44.4431	<.0001	0.119	0.064	0.222	

$*$ not proportional!

6 Stratified Cox PH Model

- when the proportional hazards assumption is violated for some covariate.
- it may be possible to stratify on that variable and employ the proportional hazards model within each stratum for the other covariates.
- subjects in the j th stratum have an arbitrary baseline hazard function $h_{oj}(t)$
- regression coefficients are assumed to be the same in each stratum although the baseline hazard functions may be different and completely unrelated.
- Example: patients who where given MTX as a graft-versus-host prophylactic did not have hazard rates proportional to those patients not given MTX.
- A key assumption in using a stratified proportional hazards model is that the covariates sharing for all stratum. This can be tested by using either a likelihood ratio test or a Wald test.
- LRT:
  - fit stratified cox model, obtain the log partial likelihood, $LL(\mathbf{b})$.
  - Using only data from the j th stratum, fit a Cox model, get the log partial likelihood $LL_j(\mathbf{b}_j)$
  - $LRT = -2[LL(\mathbf{b}) - \sum_{j=1}^{s} LL_j(\mathbf{b}_j)] \sim \chi^2((s-1)p)$

```
** Stratified cox model;

* assume PH for z10;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1") z10/param=ref;
 model t2*dfree(0) =z1 z2 g z10/rl ;
run;

* not assume PH for z10;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
 model t2*dfree(0) =z1 z2 g/rl ;
strata z10  ;
run;

* z10=0;
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
where z10=0;
 model t2*dfree(0) =z1 z2 g/rl ;
run;
* z10=1;
```

```
proc phreg data = bone_marrow plots=survival ;
class g(ref="1")/param=ref;
where z10=1;
 model t2*dfree(0) =z1 z2 g/rl ;
run;
```

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	746.719	731.154
AIC	746.719	741.154
SBC	746.719	753.248

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.5653	5	0.0082
Score	16.1614	5	0.0064
Wald	15.3436	5	0.0090

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
z1	1	0.0074	0.9314
z2	1	0.0606	0.8055
g	2	11.5993	0.0030
z10	1	1.6659	0.1968

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
z1		1	0.00158	0.01832	0.0074	0.9314	1.002	0.966	1.038	Patient Age
z2		1	0.00429	0.01741	0.0606	0.8055	1.004	0.971	1.039	Donor Age
g	2	1	-0.52584	0.31373	2.8092	0.0937	0.591	0.320	1.093	Disease Group 2
g	3	1	0.39325	0.30291	1.6855	0.1942	1.482	0.818	2.683	Disease Group 3
z10	0	1	-0.32401	0.25103	1.6659	0.1968	0.723	0.442	1.183	MTX 0

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	644.652	632.002
AIC	644.652	640.002
SBC	644.652	649.677

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.6500	4	0.0131
Score	13.2086	4	0.0103
Wald	12.5170	4	0.0139

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
z1	1	0.0204	0.8865
z2	1	0.0001	0.9922
g	2	11.7189	0.0029

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
z1		1	0.00263	0.01842	0.0204	0.8865	1.003	0.967	1.039	Patient Age
z2		1	0.0001707	0.01746	0.0001	0.9922	1.000	0.967	1.035	Donor Age
g	2	1	-0.51790	0.31340	2.7309	0.0984	0.596	0.322	1.101	Disease Group 2
g	3	1	0.41119	0.30290	1.8429	0.1746	1.509	0.833	2.732	Disease Group 3

- LRT=632.002-460.358-164.981=663,p-value=0.8451945
- effect sharing holds and stratified cox model is appropriate.