Survival Models in SAS
Part 9: PROC PHREG – Part 4

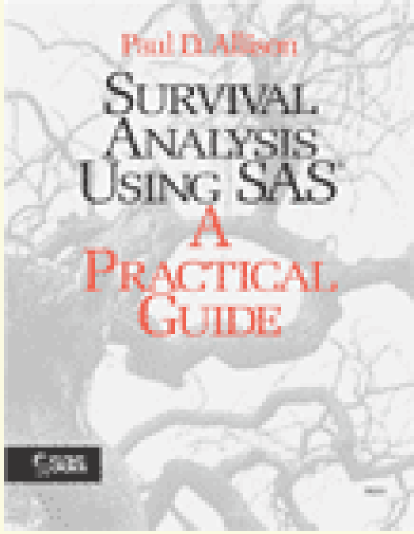
October 15, 2008

Charlie Hallahan

Chapter 5: Estimating Cox Regression Models with PROC PHREG

These talks are based on the book "**Survival Analysis Using the SAS System: A Practical Guide**" (1995) by Paul Allison.

The book is part of the SAS Books-by-Users series and can be found at http://www.sas.com/apps/pubscat/bookdetails.jsp?catid=1&pc=55233



Chapter 5: Estimating Cox Regression Models with PROC PHREG

This series of talks will cover

Chapter 5: Estimating Cox Regression Models with PROC PHREG

Topics in Chapter 5:

Introduction
The Proportional Hazards Model
Partial Likelihood
Tied Data
Time-Dependent Covariates
Cox Models with Nonproportional Hazards
Interactions with Time as Time-Dependent Covariates
Nonproportionality via Stratification
Left Truncation and Late Entry into the Risk Set
Estimating Survivor Functions
Residuals and Influence Statistics
Testing Linear Hypotheses with the TEST Statement
Conclusion

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Cox Models with Nonproportional Hazards

We’ve already seen a model with non-proportional hazards, namely, any model with time-varying covariates.

Such a model is still easily handled with Cox’s partial likelihood estimation method.

In general, though, how does one check the proportional hazards (PH) assumption?

(Note: the author thinks that too much emphasis is sometimes placed on this assumption to the detriment of other assumptions, such as having the correct covariates, measurement error, etc.)

Violations of the PH assumption are equivalent to interactions between one or more covariates and time.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Cox Models with Nonproportional Hazards

Estimating a model when the PH assumption is violated is akin to estimating a sort of *average effect* of that variable over the time range.

The author states that this, in and of itself, is not necessarily a bad thing.

In many types of regression models, interactions are omitted.

Three methods for checking the PH assumption are presented:

1. explicitly incorporating **interactions** in the model
2. *stratification* – subsuming the interactions into the arbitrary function of time, i.e. the baseline hazard
3. examining **covariate-wise residuals**

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Interactions with Time as Time-Dependent Covariates

Explicitly include an interaction with time in the model for the hazard:

For example, $\log\, h(t) = \alpha(t) + \beta_1 x + \beta_2 xt = \alpha(t) + \left(\beta_1 + \beta_2 t\right) x$

So, a positive value for β_2 implies the effect of x on the hazard increases over time.

One of the main goals of the recidivism study was to see if receiving financial aid had an effect on keeping a released prisoner from being rearrested. Financial aid, if it was given, was only provided for the 1st 13 weeks of the one-year observation period.

One way to check this would be to include an interaction between **`fin`** and **`week`** (the time variable in this example).

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Interactions with Time as Time-Dependent Covariates

```
proc phreg data=survival.recid;
   model week*arrest(0)=fin age race wexp mar paro prio
         finweek/ ties=efron;
   finweek=fin*week;
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.34350	0.42503	0.6532	0.4190	0.709
age	1	-0.05742	0.02200	6.8119	0.0091	0.944
race	1	0.31394	0.30799	1.0390	0.3081	1.369
wexp	1	-0.14967	0.21224	0.4973	0.4807	0.861
mar	1	-0.43381	0.38189	1.2904	0.2560	0.648
paro	1	-0.08496	0.19576	0.1883	0.6643	0.919
prio	1	0.09163	0.02868	10.2082	0.0014	1.096
finweek	**1**	**-0.00125**	**0.01321**	**0.0089**	**0.9247**	**0.999**

This results shows that there is no significant effect of receiving financial aid.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Interactions with Time as Time-Dependent Covariates

An alternative approach would take into account that aid was only provided for the 1st 13 weeks.

```
proc phreg data=survival.recid;
   model week*arrest(0)=fin age race wexp mar paro prio
                 fintime/ ties=efron;
   fintime=fin*(week>13);
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.19390	0.44995	0.1857	0.6665	0.824
age	1	-0.05739	0.02200	6.8029	0.0091	0.944
race	1	0.31410	0.30800	1.0400	0.3078	1.369
wexp	1	-0.14994	0.21222	0.4992	0.4799	0.861
mar	1	-0.43419	0.38190	1.2926	0.2556	0.648
paro	1	-0.08566	0.19577	0.1914	0.6617	0.918
prio	1	0.09195	0.02870	10.2656	0.0014	1.096
fintime	**1**	**-0.22546**	**0.49622**	**0.2064**	**0.6496**	**0.798**

Result no better. In fact, although highly insignificant, `fintime` has the wrong sign.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Interactions with Time as Time-Dependent Covariates

In summary, this method for checking the PH assumption for a particular variable involves creating an interaction of that variable with time and seeing if the interaction is significant or not.

If the interaction is not significant, then the PH assumption is not violated for that variable.

On the other hand, if the interaction is significant, then one can "cure the problem" by including the time-varying interaction in the model.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Stratification is another way to account for nonproportionality.

It works well with a variable that is **categorical** and not of direct interest, i.e., a **nuisance variable**.

Recall the myelomatosis example where the primary interest was in the effect of a treatment variable called **TREAT**.

The data contained another variable, **RENAL**, indicating whether or not there were renal problems.

Previously, we found that **RENAL** was strongly related to survival time, but the effect of **TREAT** was not significant.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

If we thought that the effect of RENAL varied with time, then, following the previous approach, we could include an interaction of RENAL with time in the model.

Alternatively, it is possible that the shape of the hazard function differs for those with and without renal problems.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Letting x represent the treatment indicator and z the renal functioning, the separate hazard functions could be modeled as:

$$\text{Impaired } (z=0): \quad \log h_j(t) = \alpha_0(t) + \beta x_j$$

$$\text{Not Impaired } (z=1): \quad \log h_j(t) = \alpha_1(t) + \beta x_j$$

Note that the coefficient of x, β, is the same, but the baseline hazard is allowed to be different.

These models could be combined into: $\log h_j(t) = \alpha_z(t) + \beta x_j$

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Partial likelihood estimation can be adapted to estimate a stratified model.

1. Construct separate partial likelihood functions for each of the two renal functioning groups.

2. Multiply these two functions together.

3. Choose values of that maximize this function.

The **STRATA** statement **PHREG** carries out these steps.

```
proc phreg data=survival.myel;
      model dur*status(0) = treat / ties = efron;
      strata renal;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Summary of the Number of Event and Censored Values

Stratum	renal	Total	Event	Censored	Percent Censored
1	0	18	10	8	44.44
2	1	7	7	0	0.00
Total		25	17	8	32.00

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.0736	1	0.0137
Score	5.7908	1	0.0161
Wald	4.9254	1	0.0265

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
treat	**1**	**1.46398**	**0.65965**	**4.9254**	**0.0265**	**4.323**

With this specification, `treat` is significant. Recall that treat = 1 or 2 for the two treatments, so `treat` = 2 definitely increases the hazard of death.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Three other models were estimated:

- a model with **TREAT** as the only covariate. Coefficient estimate was 0.56 with a p-value of 0.27 (which is approximately the same as the *p*-value for the log-rank test from **LIFETEST**.

- a model with both **TREAT** and **RENAL** as covariates. The coefficient of **TREAT** was 1.22 with a *p*-value of 0.04.

- a model that included **TREAT**, **RENAL** and the time-varying product of **RENAL** and **DUR**. T he coefficient of **TREAT** was 1.34 with a *p*-value of 0.05. The interaction term was not significant.

So, it is important, in this case, to control for **RENAL** in order to obtain good estimates of the treatment effect.

It doesn't seem to matter whether this is done by stratification or including **RENAL** as a covariate. **Note:** you can't do both in the same model.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Advantages of stratification vs interactions

- Using an interaction requires a particular specification for the interaction, e.g., a linear interaction between RENAL and DUR forces the effect of RENAL on DUR to be linear. Stratification allows for a more general effect of RENAL on DUR.
- Stratification is easier to set up and is less computationally intensive, a possible consideration with large samples.

Disadvantages of stratification vs interactions

- No estimated effect of the stratifying variable is obtained. Most useful for nuisance variables.
- Cannot compare likelihoods of models with and without the stratifying variable (see the above point)
- A correctly-specified interaction provides more efficient estimates for the effects of the other covariates.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

A stratification variable can have any number of values, but, obviously, it is most useful for **categorical variables** with not too many possible values.

Stratification can be useful with **clustered data**, for example, data on patients from several hospitals. A variable identifying each hospital could be used for stratification if it was thought that each hospital had its own unique baseline hazard, and it was not of interest to estimate an effect for each hospital.

The above approach does assume that the observations within each stratum are **conditionally independent**, i.e., the survival time for one patient in a particular hospital does not have an effect on the survival time of other patients in the same hospital.

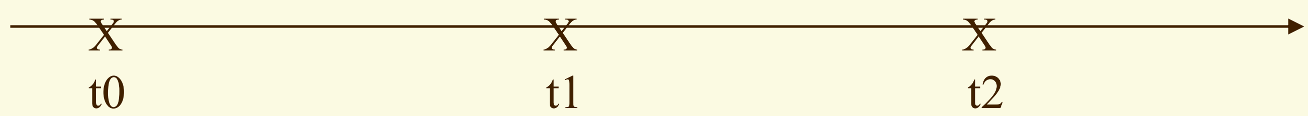
Stratifying also assumes that the **effect of the covariates** on survival is the same across strata.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

Left truncation, or **late entry**, can occur for many types of survival data.

- People are recruited into a medical study, say at time t1, and asked for the retrospective date of their diagnosis for some disease, say at t0. The purpose of the study is to model the time until death, say at time t2.

- Employees of a firm are asked at time t1 how low long they been employed by the firm, say they were hired at t0. The purpose of the study is to model how long they stayed employed at the firm, say until t2.



In each case, the duration time until the event or censoring for the subject is [t0, t2], but the subject cannot be in the risk set for the period [t0, t1] since they were either still alive or employed by the firm when recruited into the study.

The solution is to remove each subject from the risk set during periods they are not at risk.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

This is illustrated with the **Stanford Heart Transplant** data. The **duration time** is redefined as the **person's age** when accepted into the program in years rather than time since recruited into the study.

```
data stan2;
       set survival.stan;
       agels = (dls - dob)/365.25;
run;

proc phreg data=stan2;
       model (ageaccpt, agels)*dead(0) = surg plant ageaccpt /
             ties=efron;
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
surg	1	-1.04966	0.43934	5.7081	0.0169	0.350
ageaccpt	1	1.13190	0.27510	16.9285	<.0001	3.102

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

If a time-varying covariate is included in the model, it is common practice to assign a missing value during periods when the subject is not a risk.

For example, we could include the time-varying covariate of transplant status in the model.

```
proc phreg data=stan2;
       model (ageaccpt, agels)*dead(0) = surg plant ageaccpt /
               ties=efron;
       if agels < agetrans or agetrans = . then plant = 0;
              else plant = 1;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
surg	1	-1.07669	0.44177	5.9402	0.0148	0.341
plant	1	-0.47869	0.37353	1.6423	0.2000	0.620
ageaccpt	**1**	**0.99391**	**0.28224**	**12.4012**	**0.0004**	**2.702**

Compare these results with those previously obtained where the left truncation was ignored by measuring duration time as time since acceptance into the study.

```
proc phreg data=survival.stan;
      model surv1*dead(0) = surg plant ageaccpt / ties = exact;
      if wait > surv1 then plant = 0; else plant = 1;
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
surg	1	-0.77145	0.35961	4.6021	0.0319	0.462
plant	1	-0.04615	0.30276	0.0232	0.8788	0.955
ageaccpt	**1**	**0.03109**	**0.01391**	**4.9952**	**0.0254**	**1.032**

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

Note that dramatic change in the effect of **`ageaccpt`** on survival.

In the new model, age is held constant in the sense that each risk set consists of people of the same age, while in the earlier model, a risk set consists of people who have been in the study for the same length of time (people of different ages).

In the earlier model, the effects of time since acceptance are all captured by the baseline hazard function and not the **`ageaccpt`** parameter.

A related phenomena is when a subject leaves the risk set for a period of time after their initial entry time (drop out, but return).

This can be handled just as above. namely, define a time-varying covariate that is missing when a subject is not in the risk set.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

An alternative method for handling gaps in the risk set has been available since **SAS 6.10** (after the **Allison** book was published).

So far, all the datasets have been in a format where there is one observation per subject and time-varying covariates were created with programming statements in **`PHREG`**.

In the **counting-process syntax**, each subject could have multiple records, one for each period that they are at risk. The values for any time-varying covariates would then be allowed to change during each period.

This approach is also known as "**episode splitting**".

This is the syntax used by **`Stata`**.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

Associated with each record are a beginning (t0) and ending time (t1) for that period of the risk set.

The interval (t0, t1] is open on the left and closed on the right.

The value for any covariate is the value it has throughout the interval and any events are assumed to happen at the end of the interval.

The following example represents four observations for a subject.

The variable **`event`** is 1 if the event occurred at t1, otherwise **`event`** = 0.

`Trt` is a time-fixed covariate representing a treatment and **`test`** is a time-varying covariate representing a test result appropriate for that interval.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Left Truncation and Late Entry into the Risk Set

t0	t1	event	trt	test
0	3	1	1	31.3
3	5	1	1	37.2
8	11	1	1	30.9
11	20	0	1	29.5

Note the gap between 5 and 8.

The syntax for such a dataset would be:

```
proc phreg;
   model (t0, t1) * event(0) = trt test;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions

Even though the baseline hazard function is factored out of the partial likelihood and plays no role in estimating the covariate parameters, it is still possible to obtain estimates for both baseline and specific survivor functions.

The survivor function for a proportional hazard model is: $S(t) = \left[S_0(t)\right]^{\exp(\beta x)}$.

This follows from $h(t) = h_0(t)e^{\beta x}$ and $h(t) = -\dfrac{d \ \ln S(t)}{dt}$ or $S(t) = e^{-\int_0^t h(u)du}$.

Thus, $S(t) = e^{-\int_0^t h_o(u)e^{\beta x}du} = e^{-e^{\beta x}\int_0^t h_o(u)du} = \left[e^{-\int_0^t h_o(u)du}\right]^{e^{\beta x}} = \left[S_0(t)\right]^{\exp(\beta x)}$.

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions

The **BASELINE** statement in **PHREG** is used to produce a dataset from which various plots can be created.

The **default** is to assign **mean values** to all the **covariates**. Specific covariate values can be specified with the **COVARIATES=** option naming a **SAS** dataset with covariate values.

```
proc phreg data=survival.recid;
       model week*arrest(0)= fin age prio / ties = efron;
       baseline out=a survival=s logsurv=ls loglogs=lls;
run;
```

SURVIVAL= requests the survivor probabilities, **LOGSURV=** requests the logarithm of the survivor probabilities (recall that *– ln(S(t)) = H(T)*, the **cumulative hazard function**) and **LOGLOGS=** requests *ln(– ln(S(t)))*.

Recall that for the **Weibull** distribution, *ln(– ln(S(t)))* is linear in *ln(t)* resulting in a straight line when *ln(– ln(S(t)))*. is plotted against *ln(t)*.

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions

```
title "Output dataset from the BASELINE statement";
proc print data=a(obs=20); run;
```

Obs	fin	age	prio	week	s	ls	lls
1	0.5	24.5972	2.98380	0	1.00000	0.00000	.
2	0.5	24.5972	2.98380	1	0.99801	-0.00200	-6.21671
3	0.5	24.5972	2.98380	2	0.99601	-0.00399	-5.52281
4	0.5	24.5972	2.98380	3	0.99402	-0.00600	-5.11672
5	0.5	24.5972	2.98380	4	0.99203	-0.00800	-4.82813
6	0.5	24.5972	2.98380	5	0.99004	-0.01001	-4.60379
7	0.5	24.5972	2.98380	6	0.98804	-0.01203	-4.41998
8	0.5	24.5972	2.98380	7	0.98604	-0.01406	-4.26429
9	0.5	24.5972	2.98380	8	0.97601	-0.02428	-3.71817
10	0.5	24.5972	2.98380	9	0.97200	-0.02840	-3.56146
11	0.5	24.5972	2.98380	10	0.96999	-0.03047	-3.49105
12	0.5	24.5972	2.98380	11	0.96593	-0.03467	-3.36194
13	0.5	24.5972	2.98380	12	0.96184	-0.03891	-3.24646
14	0.5	24.5972	2.98380	13	0.95979	-0.04104	-3.19320
15	0.5	24.5972	2.98380	14	0.95363	-0.04748	-3.04744
16	0.5	24.5972	2.98380	15	0.94951	-0.05181	-2.96011
17	0.5	24.5972	2.98380	16	0.94538	-0.05616	-2.87948
18	0.5	24.5972	2.98380	17	0.93918	-0.06275	-2.76860
19	0.5	24.5972	2.98380	18	0.93293	-0.06942	-2.66754
20	0.5	24.5972	2.98380	19	0.92876	-0.07391	-2.60493

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions

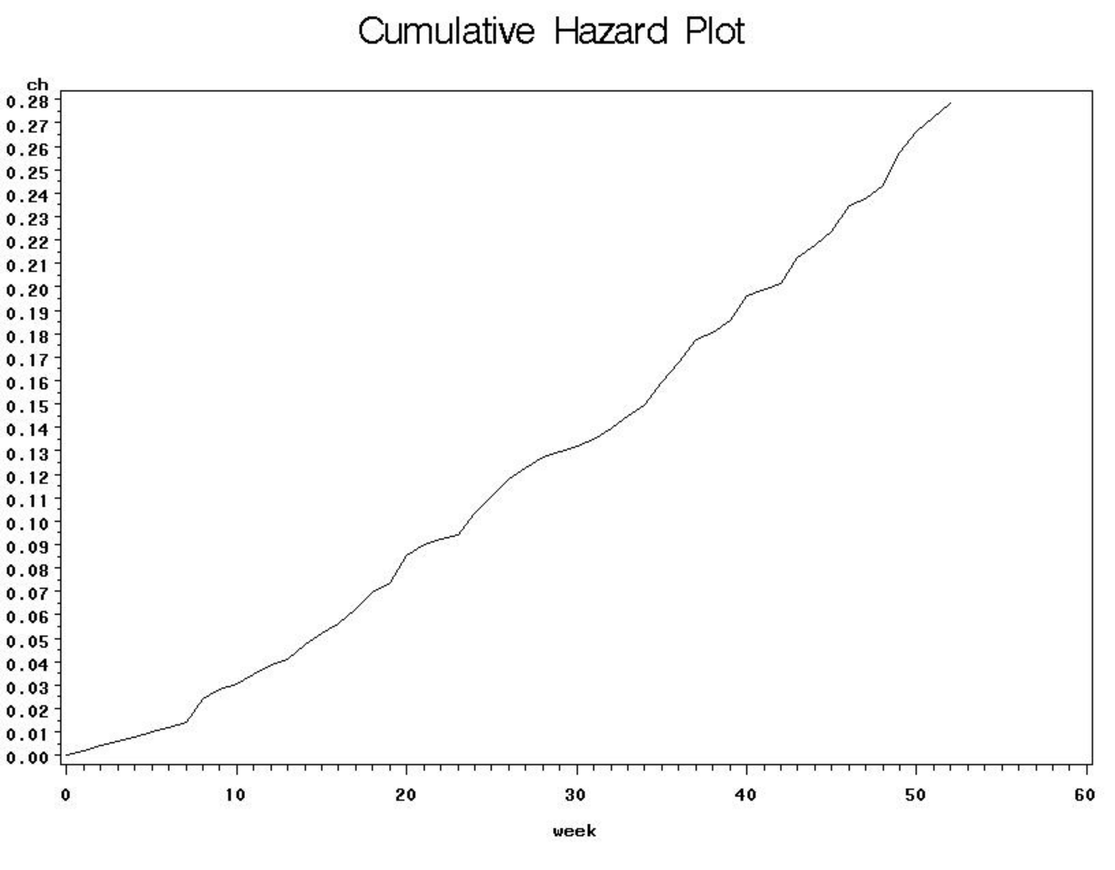
These output variables can now be **plotted** to check various hypotheses about the **shape of the hazard function** for this data.
For example, if a plot of the **cumulative hazard** against time is a **straight line**, then this implies a constant hazard (or an **exponential distribution** for time to failure).

On the other hand, an upward bow implies an increasing hazard while a downward bow implies a decreasing hazard (recall that the hazard function if the first derivative of the cumulative hazard).

```
data b;
        set a;
        *define the cumulative hazard;
        ch = -ls;
run;
title "Cumulative Hazard Plot";
proc gplot data=b;
        symbol1 value=none interpol=join;
        plot ch*week;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions



The plot appears to curve upward, implying an increasing hazard.

This is consistent with earlier conclusions in Chapters 3 and 4.

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

Estimating Survivor Functions

Recall that the proportional hazards assumption is that $h_1(t) = \lambda h_2(t)$ for any two subjects. This implies that $H_1(t) = \lambda H_2(t) \Rightarrow S_1(t) = \left[S_2(t)\right]^{\lambda}$.

Hence, $\log[\text{-}\log S_1(t)] = \log\lambda + \log[-\log S_2(t)]$. So a plot of the two log-log survival curves should differ by a constant amount.

We could test this assumption for the covariate **`fin`** by stratifying on **`fin`** and plotting the two curves.

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

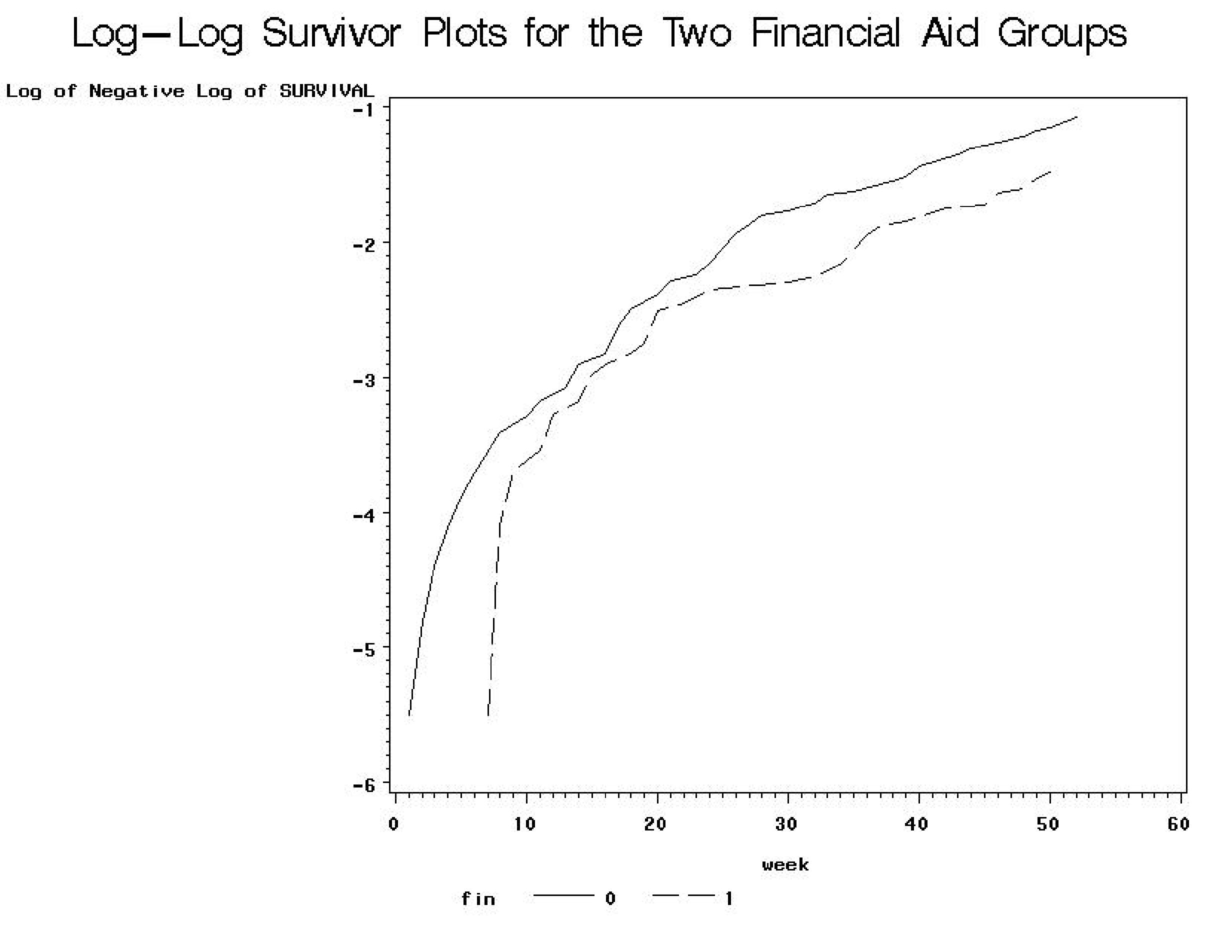
Estimating Survivor Functions

```
proc phreg data=survival.recid;
      model week*arrest(0)=  age prio / ties = efron;
      strata fin;
      baseline out=a survival=s loglogs=lls;
run;

title "Log-Log Survivor Plots for the Two Financial Aid
Groups";
proc gplot data=a;
      symbol1 interpol=join color=black line=1;
      symbol2 interpol=join color=black line=2;
      plot lls*week=fin;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG:

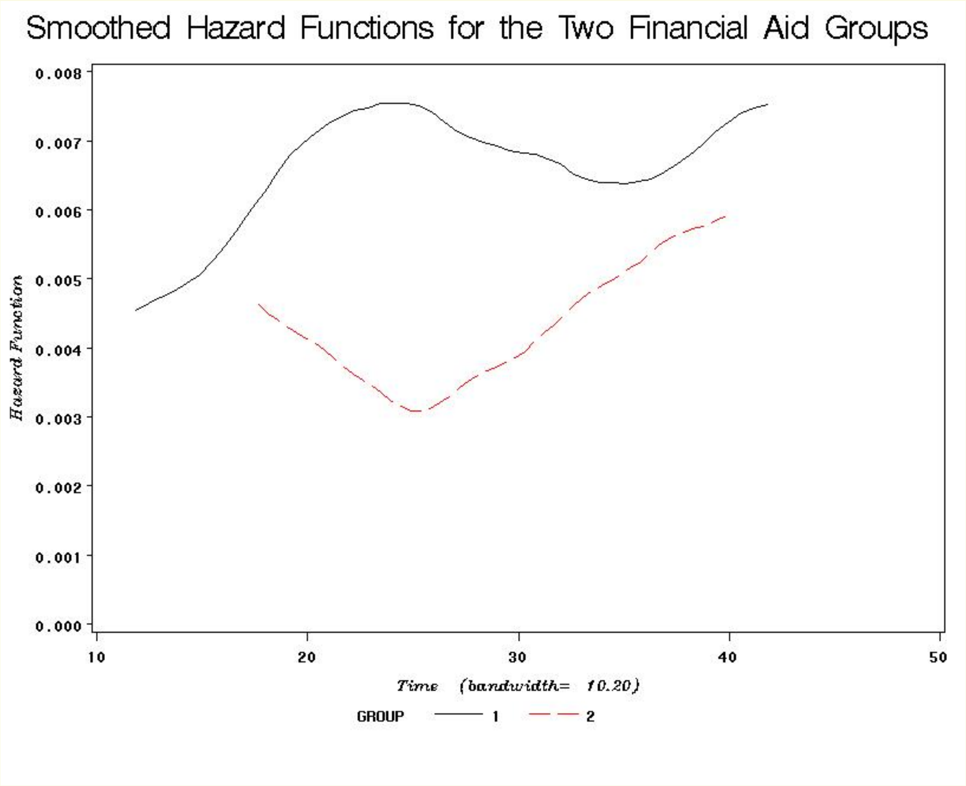
Estimating Survivor Functions



Distance between the two curves does vary a little.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Perhaps a smoothed plot of the two estimated hazard functions will more clearly demonstrate if the hazards are proportional.



Hazards rapidly diverge after week 15 but get closer near the end of the year. So it appears that the effect of financial aid wears off eventually.

In this study, those getting financial aid got it during the first 15 weeks after release.

The graph suggests an **interaction** between **`fin`** and **`week`** between the weeks of 20 and 30.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

```
proc phreg data=survival.recid;
   model week*arrest(0)=  fin finmid age prio / ties = efron;
   mid=(20<week<30);
   finmid=fin*mid;
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.15807	0.20504	0.5943	0.4408	0.854
finmid	**1**	**-1.45568**	**0.66470**	**4.7961**	**0.0285**	**0.233**
age	1	-0.06696	0.02084	10.3265	0.0013	0.935
prio	1	0.09675	0.02726	12.5929	0.0004	1.102

The results show a significant effect for the interaction. In fact, there is about a 77% reduction in the hazard of being re-arrested during those middle weeks for those who received financial aid.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

To get predicted values for specific covariate values, you need to create a separate dataset with the appropriate covariate values.

For example, to get a prediction for a 40-year old male with three prior convictions who did not receive financial aid:

```
data covals;
   input fin age prio;
   cards;
0 40 3
run;

proc phreg data=survival.recid;
   model week*arrest(0)=fin age prio / ties=efron;
   baseline out=a covariates=covals survival=s lower=lcl
         upper=ucl / nomean;
run;

proc print data=a;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Nonproportionality Via Stratification

Obs	fin	age	prio	week	s	lcl	ucl
1	0	40	3	0	1.00000	.	.
2	0	40	3	1	0.99916	0.99739	1.00000
3	0	40	3	2	0.99831	0.99566	1.00000
4	0	40	3	3	0.99746	0.99405	1.00000
5	0	40	3	4	0.99662	0.99249	1.00000
6	0	40	3	5	0.99577	0.99095	1.00000
7	0	40	3	6	0.99492	0.98942	1.00000
8	0	40	3	7	0.99407	0.98790	1.00000
9	0	40	3	8	0.98980	0.98042	0.99927
10	0	40	3	9	0.98808	0.97742	0.99885
:							
:							
42	0	40	3	43	0.91432	0.85416	0.97871
43	0	40	3	44	0.91219	0.85073	0.97809
44	0	40	3	45	0.91005	0.84729	0.97747
45	0	40	3	46	0.90575	0.84039	0.97620
46	0	40	3	47	0.90467	0.83865	0.97588
47	0	40	3	48	0.90250	0.83518	0.97524
48	0	40	3	49	0.89703	0.82648	0.97361
49	0	40	3	50	0.89373	0.82124	0.97261
50	0	40	3	52	0.88926	0.81419	0.97126

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

A number of different kinds of residuals are associated with survival models.

Three options are available with **PHREG**:

`LOGSURV` for Cox-Snell residuals
`RESMART` for martingale residuals
`RESDEV` for deviance residuals

Deviance residuals are transformations of martingale residuals which are transformations of Cox-Snell residuals.

Allison recommends using the deviance residuals since they behave much like **OLS** residuals in that they are symmetrically distributed around 0 and have an approximation standard deviation of 1.0.

They are negative for observations which have a longer survival time than expected and positive for observations that have a shorter survival time than expected.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

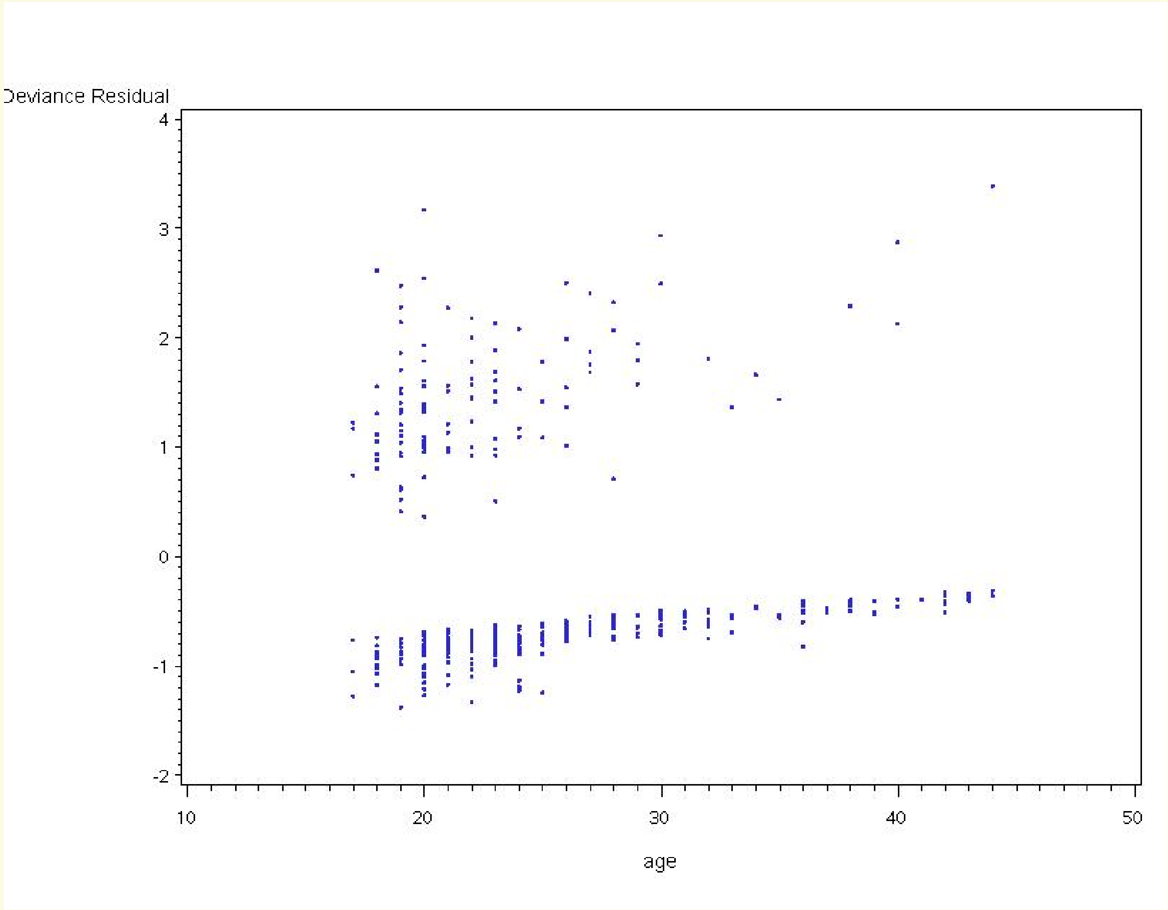
As with **OLS** residuals, deviance residuals can be plotted against various covariates to look for outliers.

Below we plot the deviance residuals against **age**.

```
proc phreg data=survival.recid;
   model week*arrest(0)=fin age prio / ties=efron;
   output out=c resdev=dev;
run;

proc gplot data=c;
   symbol1 value=dot h=.2;
    plot dev*age;
run;
```

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics



A few residuals with values larger than 3 might bear further scrutiny as potential outliers.

The lower portion of the graph corresponds to the censored observations and the positive slope corresponds to survival increasing with age.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

Another set of residuals associated with survival models differ with each covariate.

Three options are available with **PHREG**:

RESSCH for Schoenfeld residuals
WTRRESSCH for weighted Schoenfeld residuals
RESSCO for score residuals

Each set of residuals approximately sums to zero.

The **Schoenfeld residuals** are **not defined for the censored observations**.

Allison states that plots of these sets of residuals are about equally informative, and he recommends using the **Schoenfeld** residuals.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

Here's how to interpret the **Schoenfeld** residuals.

Suppose an observation fails at time t_i and at that time there were 30 subjects still at risk for failure, indexed by $j = 1, \ldots, 30$.

Let p_j be the estimated probablilty from the Cox model of subject j failing at time t_i.

Imagine **randomly selecting** one of these 30 subjects with probability p_j.

For each covariate x_k, we can calculate the **expected value of that covariate** as

$$\bar{x}_k = \sum_{j=1}^{30} p_j x_{kj}$$

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

The **Schoenfeld residual** is then defined as the covariate value for the subject who actually failed at time t_i, x_{ki}, minus the expected value.

Since **Schoenfeld residuals** are, in principle, **independent of time**, a plot of these residuals against time should not show any relationship.

Any functional relationship between the residuals and time is evidence of failure of the proportional hazards assumption.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

```
proc phreg data=survival.recid;
   model week*arrest(0)=fin age prio / ties=efron;
  output out=b ressch=schfin schage schprio;
run;

proc sort data=b ;
      by descending week ;

proc print data=b;
      id week;
      var arrest fin schfin age schage prio schprio;
      where week <= 12;
run;
```

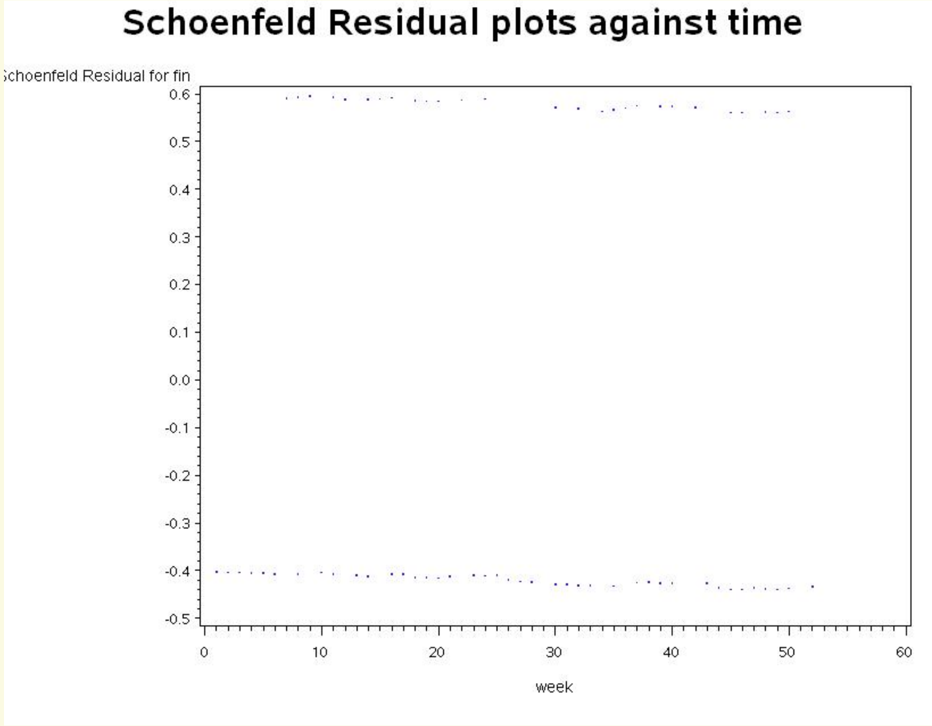
Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

week	arrest	fin	schfin	age	schage	prio	schprio
12	1	1	0.58848	27	4.3963	0	-3.9626
12	1	1	0.58848	22	-0.6037	2	-1.9626
11	1	1	0.59347	19	-3.5397	2	-2.1748
11	1	0	-0.40653	19	-3.5397	18	13.8252
10	1	0	-0.40284	21	-1.5257	14	9.7360
9	1	1	0.59573	30	7.4612	3	-1.2575
9	1	1	0.59573	26	3.4612	0	-4.2575
8	**1**	**1**	**0.59318**	**40**	**17.4559**	**1**	**-3.2912**
8	1	0	-0.40682	23	0.4559	5	0.7088
8	1	1	0.59318	20	-2.5441	11	6.7088
8	1	0	-0.40682	28	5.4559	4	-0.2912
8	1	1	0.59318	21	-1.5441	4	-0.2912
7	1	1	0.59193	20	-2.5388	2	-2.2863
6	1	0	-0.40617	19	-3.5222	6	1.7056
5	1	0	-0.40475	19	-3.5099	3	-1.2898
4	1	0	-0.40351	18	-4.4961	1	-3.2798
3	1	0	-0.40284	30	7.4914	3	-1.2777
2	1	0	-0.40261	44	21.4788	2	-2.2763
1	1	0	-0.40163	20	-2.5151	0	-4.2660

The first person arrested after 8 weeks was 40 years old with 1 prior arrest and who did receive financial aid. According to the model, a person arrested after just 8 weeks was expected to be about 17 ½ years younger with over 4 prior arrests and to have only about a 40% chance of receiving financial aid.

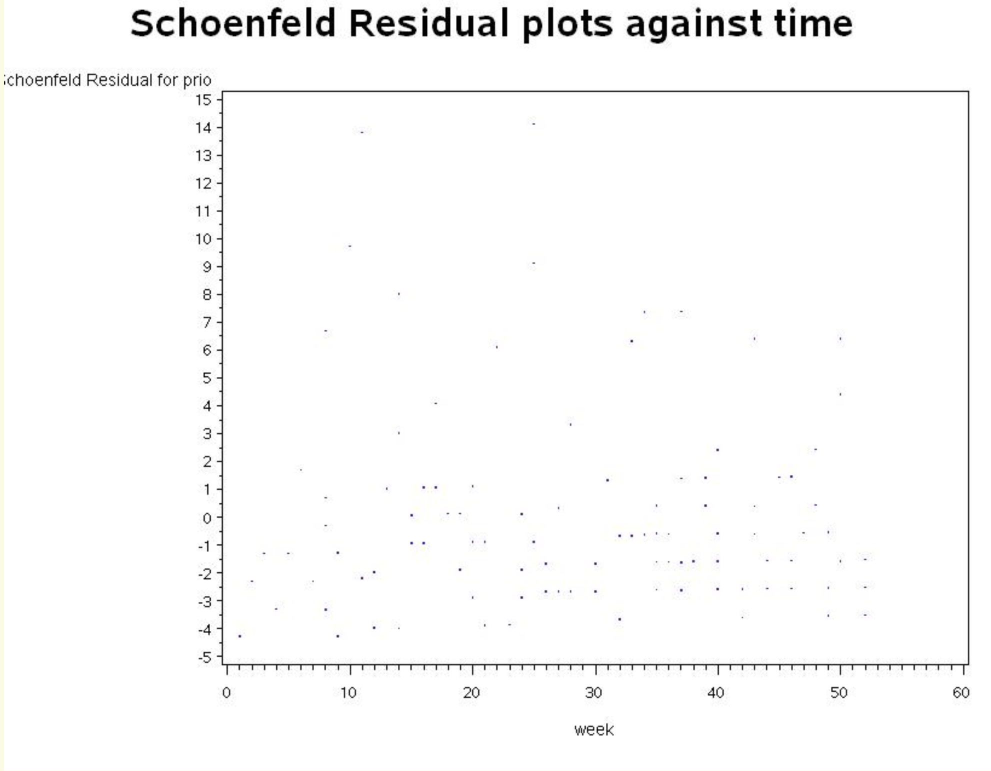
Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics

```
title "Schoenfeld Residual plots against time";
proc gplot data=b;
   plot schfin*week schprio*week schage*week;
   symbol1 value=dot h=.02;
run;
```



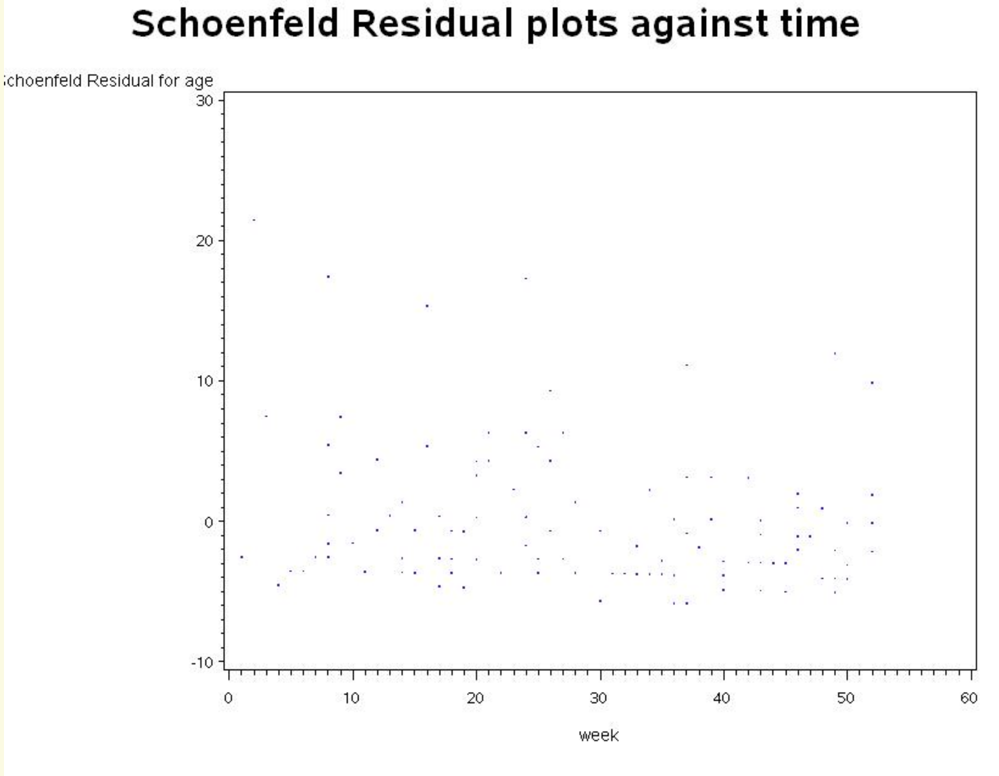
No pattern apparent with **`fin`**, not surprising with a binary variable.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics



Plot with the continuous variable **`prio`** seems to be a random scatter.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Residuals and Influence Statistics



Residual plot with **`age`** does appear to have a slightly increasing slope.

Allison regresses the residuals against all three covariates and finds *p*-values of 0.94 and 0.33 for **`fin`** and **`prio`** and a value of 0.02 for **`age`** – indicating that there might be some departure from proportionality for **`age`**.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Influence Diagnostics

As with **OLS** models, influence diagnostics can be calculated for the **Cox** model.

Two such statistics available in **PHREG** are:

1. the **likelihood displacement (LD) statistic** which measure the influence of an observation on the overall fit of the model. It approximates how much twice the log likelihood changes when an observation is dropped.

2. The **DFBETA** statistic measures how much each coefficient estimate changes when an observation is dropped.

The statistics will be demonstrated using the **myelomatosis** dataset.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Influence Diagnostics

```
proc phreg data=survival.myel;
   model dur*status(0)=treat renal;
   id id;
   output out=c ld=ldmyel dfbeta=dtreat drenal;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
treat	**1**	**1.24308**	0.59932	4.3021	0.0381	3.466
renal	**1**	**4.10540**	1.16451	12.4286	0.0004	60.667

The coefficient estimates for the full model are shown above.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Influence Diagnostics

id	dur	status	treat	renal	**dtreat**	**drenal**	**ldmyel**
1	8	1	1	1	-0.21768	-0.08058	0.13393
12	**8**	**1**	**1**	**0**	**-0.46468**	**-1.44581**	**1.71228**
13	13	1	2	1	0.01809	0.02970	0.00120
16	18	1	2	1	0.00851	0.01759	0.00033
25	23	1	2	1	-0.01950	-0.01780	0.00109
5	52	1	1	1	0.08899	0.18409	0.03600
8	63	1	1	1	0.09915	0.04170	0.02762
21	63	1	1	1	0.09915	0.04170	0.02762
11	70	1	2	0	0.05023	0.11302	0.01263
10	76	1	2	0	0.04974	0.11273	0.01250
2	180	1	2	0	0.04908	0.11234	0.01232
9	195	1	2	0	0.04819	0.11180	0.01208
20	210	1	2	0	0.04694	0.11106	0.01174
7	220	1	1	0	-0.19650	-0.09357	0.10797
24	365	0	1	0	0.05934	0.05943	0.01027
3	632	1	2	0	0.00971	0.08879	0.00587
17	700	1	2	0	0.00699	0.08717	0.00575
4	852	0	1	0	0.08582	0.07526	0.02097
18	1296	0	1	0	0.10383	0.08603	0.03048
23	1296	1	2	0	-0.01244	0.07554	0.00604
22	1328	0	1	0	0.10383	0.08603	0.03048
19	1460	0	1	0	0.10383	0.08603	0.03048
15	1976	0	1	0	0.10383	0.08603	0.03048
14	1990	0	2	0	-0.11223	0.01587	0.04067
6	2240	0	2	0	-0.11223	0.01587	0.04067

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Influence Diagnostics

`ID = 12` has unusually large values for the influence diagnostics. For example, dropping that observation **increases** the parameter estimate for **`RENAL`** by 1.44585.

(Note: **`DFRENAL`**$_i$ = estimate from full model – estimate dropping i^{th} observation.)

A listing of the dataset (next page) shows that `id = 12` plays an unusual role in the dataset. Namely, except for `id = 12`, all observations with `renal = 1` have shorter duration times than observations with `renal = 0`.

In fact, when the **Allison** book was written in 1995 using **SAS 6.10**, dropping `id=12` from the estimation dataset resulted in non-convergence.

I tried estimating without `id=12` using **SAS 9.2** and the output does report convergence achieved, however, there are dramatic changes in coefficient estimates. Allison states in his text that there is a dichotomous covariate in the model with the property that duration times for one of the covariate's values are always less than duration times for the other value.

Chapter 5: Estimating Cox Regression Models with PROC PHREG: Influence Diagnostics

id	dur	renal
12	**8**	**0**
1	8	1
13	13	1
16	18	1
25	23	1
5	52	1
8	63	1
21	63	1
11	70	0
10	76	0
2	180	0
9	195	0
20	210	0
7	220	0
24	365	0
3	632	0
17	700	0
4	852	0
18	1296	0
23	1296	0
22	1328	0
19	1460	0
15	1976	0
14	1990	0
6	2240	0

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Hazard Ratio
treat	1	1.90588	0.77672	6.725
renal	**1**	**20.35342**	**2426**	**6.9084E8**

Estimation without `id=12`. Note the very large Standard Error for renal.

Chapter 5: Testing Linear Hypotheses with the TEST Statement

The **TEST** statement works in **PHREG** just as it does in other SAS PROCS, such as **PROC REG**.

Since **PHREG** does not have a **CLASS** statement, the output does not report an ***F*-test** that all dummies associated with a discrete covariate are insignificant.

Below, we create dummy variables associated with the continuous variable **ed** in the **recidivism dataset** and illustrate the **test** statement..

```
data recid2;
   set survival.recid;
   ed3=(educ=3);
   ed4=(educ=4);
   ed5=(educ=5);
   ed6=(educ=6);
run;
```

Chapter 5: Testing Linear Hypotheses with the TEST Statement

```
proc phreg data=recid2;
   model week*arrest(0)=fin age prio ed3-ed6 /ties=efron;
   No_Educ: test ed3,ed4,ed5,ed6;
   Ed3_Ed6: test ed3=ed6;
   Ed4_Ed6: test ed4=ed6;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.36382	0.19129	3.6171	0.0572	0.695
age	1	-0.06005	0.02096	8.2070	0.0042	0.942
prio	1	0.08383	0.02839	8.7174	0.0032	1.087
ed3	**1**	**0.54767**	**0.51869**	**1.1149**	**0.2910**	**1.729**
ed4	**1**	**0.31568**	**0.54082**	**0.3407**	**0.5594**	**1.371**
ed5	**1**	**-0.18082**	**0.67309**	**0.0722**	**0.7882**	**0.835**
ed6	**1**	**-0.46189**	**1.12059**	**0.1699**	**0.6802**	**0.630**

Note that **ed2** is used as the **reference category**, and that none of the other categories are significantly different from it.

Chapter 5: Testing Linear Hypotheses with the TEST Statement

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
No_Educ	**4.5496**	**4**	**0.3367**
Ed3_Ed6	**0.9992**	**1**	**0.3175**
Ed4_Ed6	**0.5817**	**1**	**0.4456**

The labels are not required, but serve to identify the various tests on the output.