Lecture 9
Assessing the Fit of the Cox Model

The Cox (PH) model:

$$\lambda(t|\mathbf{Z}(t)) = \lambda_0(t)\ \exp\{\boldsymbol{\beta}'\mathbf{Z}(t)\}$$

Assumptions of this model:

(1) the regression effect $\boldsymbol{\beta}$ is constant over time (PH assumption)

(2) linear combination of the covariates (including possibly higher order terms, interactions)

(3) the link function is exponential

The PH assumption in (1) has received most attention in both research and application.

In order to check these model assumptions, we often make use of residuals.

Residuals for survival data are somewhat different than for other types of models, mainly due to the censoring.

What are the residuals for the Cox model?

(a) generalized (Cox-Snell)

(b) Schoenfeld

(c) martingale

We will first give the definition of these residuals, and their direct use in assessing model fit. Some residuals, in particular the martingale residuals, can be used in more sophisticated (and more powerful) ways, some of which we will talk about later.

First we need an important basic result -

Inverse CDF:

If T_i (the survival time for the i-th individual) has survivorship function $S_i(t)$, then the transformed random variable $S_i(T_i)$ should have a uniform distribution on $[0, 1]$, and hence $\Lambda_i(T_i) = -\log[S_i(T_i)]$ should have a unit exponential distribution.

That is,

$$\begin{aligned} \text{If} \quad T_i &\sim S_i(t) \\ \text{then} \quad S_i(T_i) &\sim \text{Uniform}(0,1) \\ \text{and} \quad \Lambda_i(T_i) &\sim \text{Exponential}(1) \end{aligned}$$

[Reading] (a) Generalized (Cox-Snell) Residuals:

The implication of the last result is that if the model is correct, the estimated cumulative hazard for each individual at the time of their death or censoring, $\hat{\Lambda}_i(X_i)$ $(i = 1, ..., n)$, should be like a censored sample from a unit exponential. $\hat{\Lambda}_i(X_i)$ is called the *generalized* or *Cox-Snell* residual.

Step 1: Previously, we had

$$\hat{S}(t; \mathbf{Z}_i) = [\hat{S}_0(t)]^{\exp(\boldsymbol{\beta}' \mathbf{Z}_i)}.$$

So let

$$\hat{\Lambda}_i(X_i) = -\log[\hat{S}(X_i; Z_i)]$$

Example: Nursing home data

Say we have (covariates = marital, gender)

- a single male
- with actual duration of stay of 941 days ($X_i = 941$)

We compute the entire distribution of survival probabilities for single males, and obtain $\hat{S}(941) = 0.260$.

$$-\log[\hat{S}(941, \text{single male})] \;=\; -\log(0.260) = 1.347$$

We repeat this for everyone in our dataset.

These should be like a censored sample from an Exponential(1) distribution if the model fits the data well.

How do we assess whether they are Exp(1)?

Step 2: Now suppose we have a censored sample $Y_i = \hat{\Lambda}_i(X_i)$, $i = 1, ..., n$, from an Exponential(1) distribution. **Recall:** we estimate the survival function by the KM estimate, denote $\hat{S}$, then

- plotting $-\log(\hat{S}(Y_i))$ vs Y_i should yield a straight line
- plotting $\log[-\log \hat{S}(Y_i)]$ vs $\log(Y_i)$ should yield a straight line through the origin with slope=1.

(Note: this of course does not necessarily mean that the underlying distribution of the original survival times is exponential!)

Caution notes

Allison states "Cox-Snell residuals... are not very informative for Cox models estimated by partial likelihood."

Encyclopedia of Biostatistics, Chapter on 'Goodness of Fit in Survival Analysis':
"Baltazar-Aban and Peña (1995) pointed out that the critical assumption of approximate unit exponentiality of the residual vector will often not be viable. Their analytical and Monte Carlo results show that the model diagnostic procedures thus considered can have serious defects when the failure time distribution is not exponential or when the residuals are obtained nonparametrically in the no-covariate model or semiparametrically in the Cox proportional hazards model. The difficulties stem from the complicated correlation structure arising through the estimation process of both the regression coefficients and the underlying cumulative hazard. It has also been argued that, even under quite large departures from the model, this approach may lack sensitivity (O'Quigley 1982, Crowley and Storer 1983)."

Summary. The main problem is caused by: although $\Lambda_i(T_i) \sim$ Exponential(1), we are using estimated $\hat{\Lambda}_i(T_i)$. The graphical part in step 2 is still a good way to check the assumption of unit exponentiality, but the overall procedure may not be that sensitive for checking the Cox model.

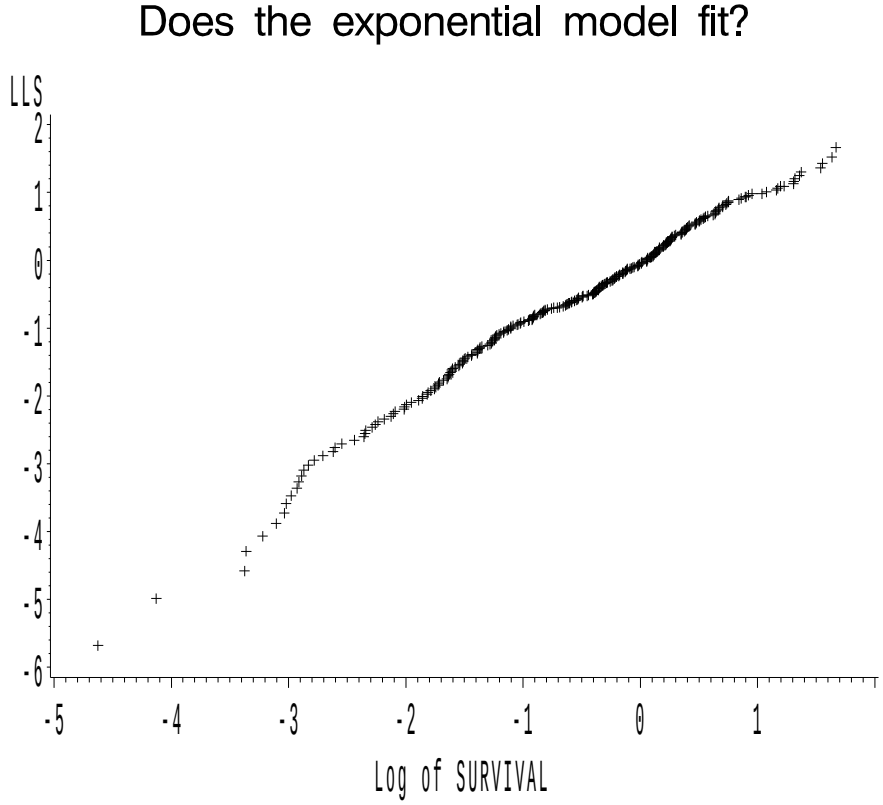


Figure 1: Example: Halibut data, using towdur, handling, length and logcatch as covariates.

(b) Schoenfeld Residuals

The partial likelihood score equation

$$\sum_{\delta_i=1}\{\mathbf{Z}_i(X_i) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, X_i)\} = 0.$$

has the form of the sum of (observed covariate - expected covariate) at each failure time.

The Schoenfeld (1982) residuals are defined as

$$\mathbf{r}_i = \mathbf{Z}_i(X_i) - \bar{\mathbf{Z}}(\hat{\boldsymbol{\beta}}, X_i)$$

for each observed failure $(\delta_i = 1)$.

Component wise, it is

$$r_{ij} = Z_{ij}(X_i) - \bar{Z}_j(\hat{\boldsymbol{\beta}}, X_i)$$

for the jth component of $\mathbf{Z}$.

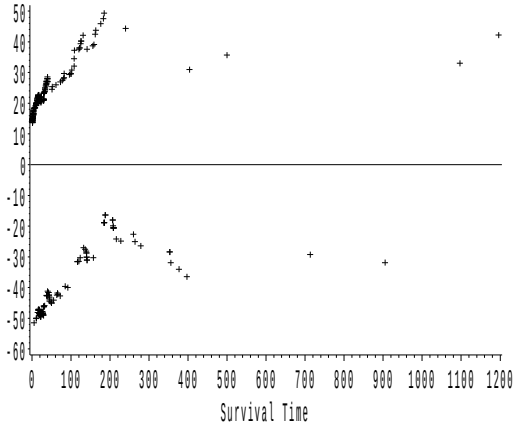
Notes:

- these represent the difference between the observed covariate and the expected given the risk set at that time
- calculated for each covariate
- not defined for censored failure times
- sum of the Schoenfeld residuals = 0. (why?)

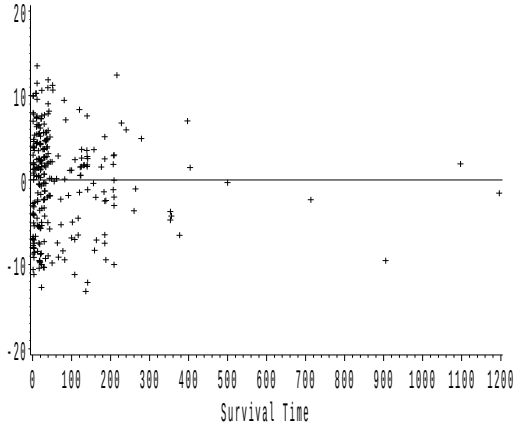
Schoenfeld (1982) showed that the $\mathbf{r}_i$'s are **asymptotically uncorrelated** and have **expectation zero** under the Cox model. Thus a plot of r_{ij} versus X_i should be centered about zero. On the other hand, non-PH in the effect of Z_j could be revealed in such a plot (why).

Ex. Show that if $\beta(t)$ decreases over time say, then the residuals should also show a monotone trend over time, e.g. $Z_{ij}(X_i) > \bar{Z}_j(\hat{\boldsymbol{\beta}}, X_i)$ early on, and $Z_{ij}(X_i) < \bar{Z}_j(\hat{\boldsymbol{\beta}}, X_i)$ later. For that you need to show:
1) $\hat{\beta}$ estimates some average value $\bar{\beta}$ of $\beta(t)$ over time; see Xu and O'Quigley (2000).
2) You are basically comparing $\frac{s^{(1)}(\beta(t),t)}{s^{(0)}(\beta(t),t)}$ with $\frac{s^{(1)}(\bar{\beta},t)}{s^{(0)}(\bar{\beta},t)}$.

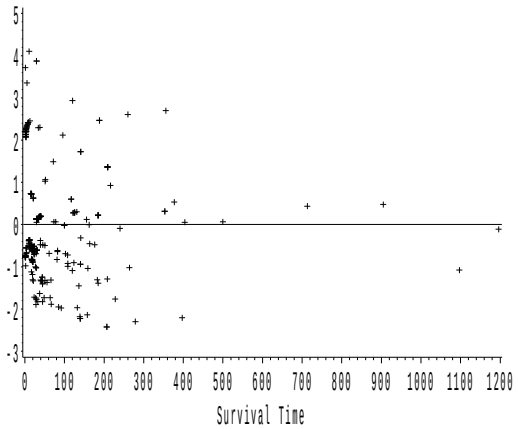
Schoenfeld resids for towing vs survival time



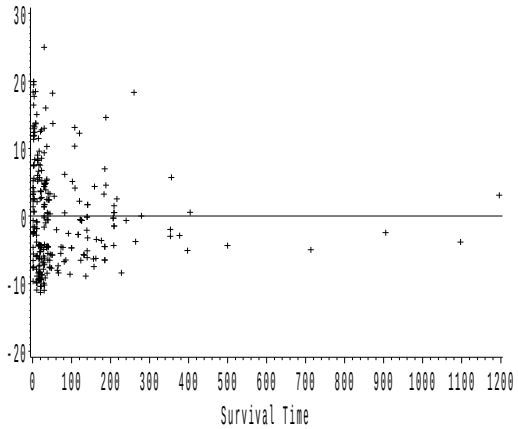
Schoenfeld resids for length vs survival time



Schoenfeld resids for log(catch) vs survival time



Schoenfeld resids for handling vs survival time



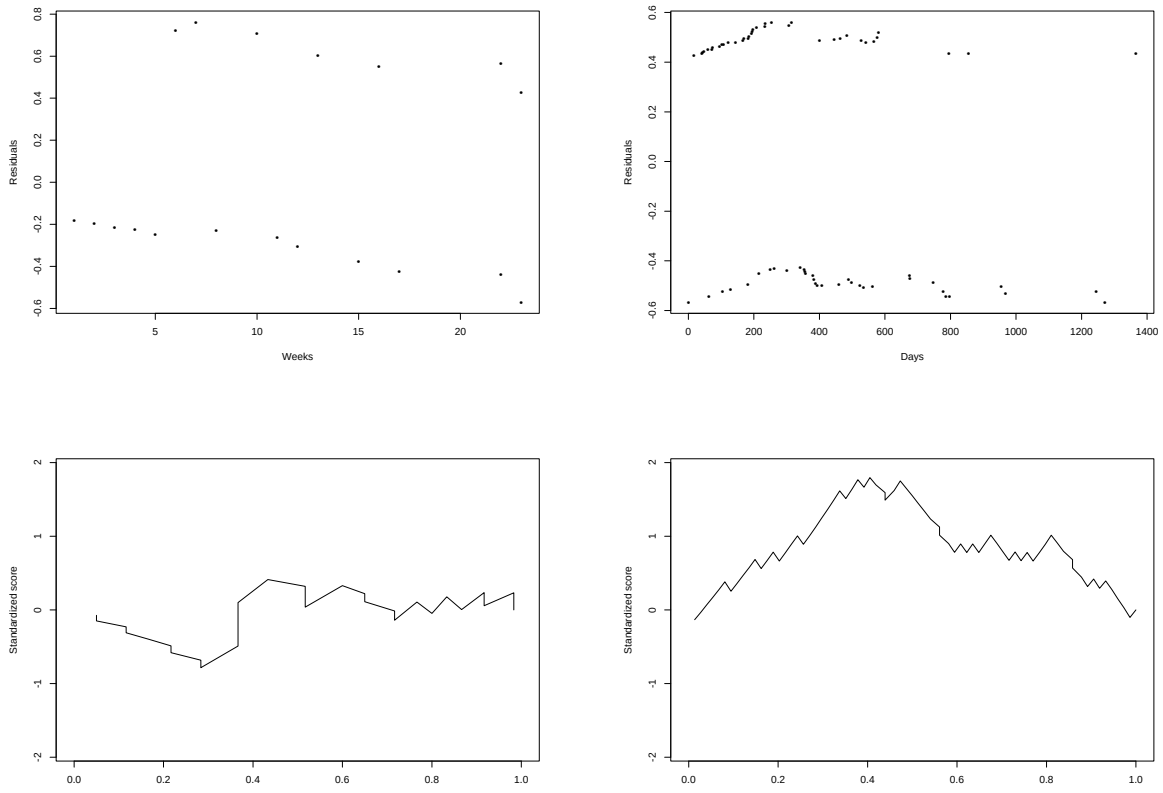


Figure 2: Schoenfeld residuals (top) and standardized cumulative Schoenfeld residuals (bottom) for Freireich data (left) and Stablein data (right).

Caution notes -

Encyclopedia of Biostatistics:

"...these direct residual plots are relatively insensitive to model departures. In practice it is more instructive to examine the cumulative residuals."

[Reading] Weighted Schoenfeld Residuals

These are defined as:

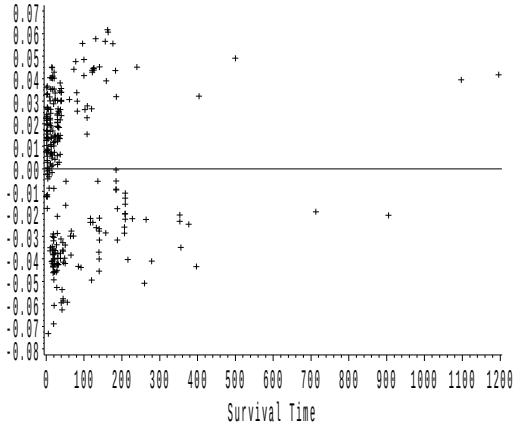
$$\mathbf{r}_i^w = n_e \widehat{V}\ \mathbf{r}_i$$

where n_e is the total number of events, $\widehat{V}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. The weighted residuals can be used in the same way as the unweighted ones to assess time trends and lack of proportionality.

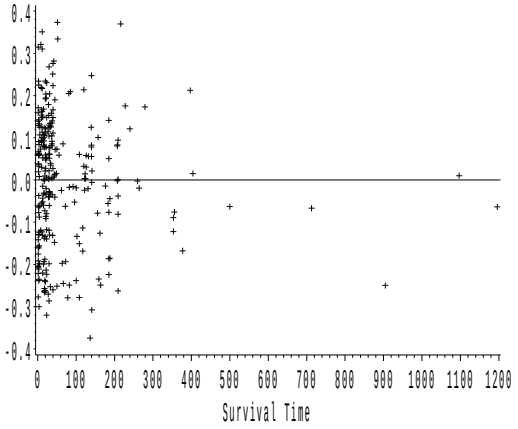
One advantage of the weighted residual is that they might look more normally distributed, especially when a covariate is binary.

Grambsch and Therneau (1993) showed that a smoothed plot of r_{ij}^w vs X_i roughly gives the shape of $\beta_j(t) - \hat{\beta}_j$.

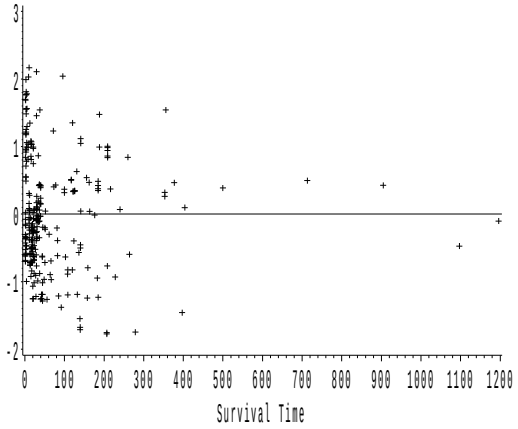
Weighted Schoenfeld resids for towing vs time



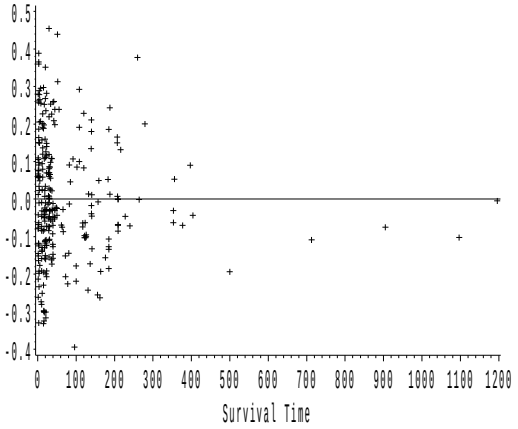
Schoenfeld resids for length vs time



Schoenfeld resids for log(catch) vs time



Schoenfeld resids for handling vs time



(c) Martingale Residuals

(see Fleming and Harrington, p.164)

Martingale residuals are defined for the i-th individual as:

$$M_i = \delta_i - \hat{\Lambda}_i(X_i)$$

Interpretation: - the residual M_i can be viewed as the difference between the observed number of deaths (0 or 1) for subject i between time 0 and X_i, and the expected numbers based on the fitted model (why).

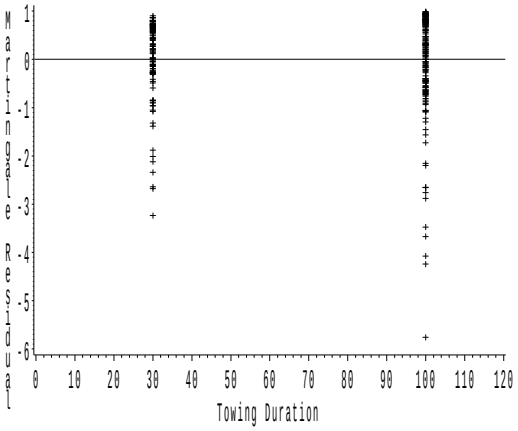
Properties:

- M_i's have mean 0
- range of M_i's is between $-\infty$ and 1
- approximately uncorrelated (in large samples)

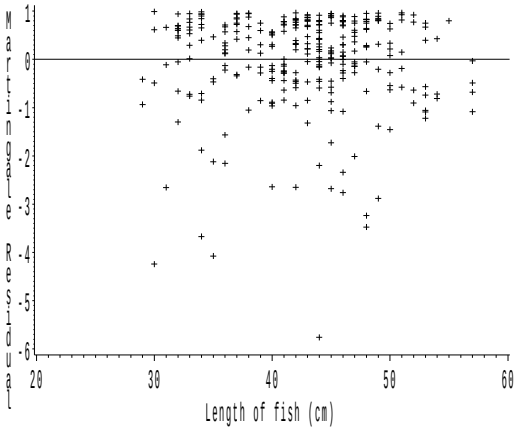
How to use?

You can plot it versus the predicted prognostic index (i.e., $\boldsymbol{\beta}'\mathbf{Z}_i$, also called the linear predictor) to check the exponential form of the link function, or any of the individual covariates to check the functional form of a covariate.

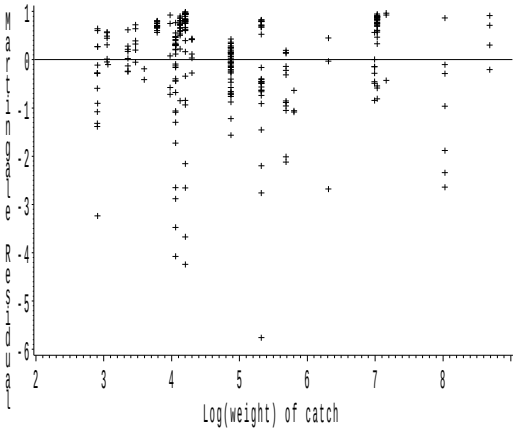
Martingale residuals vs towing duration



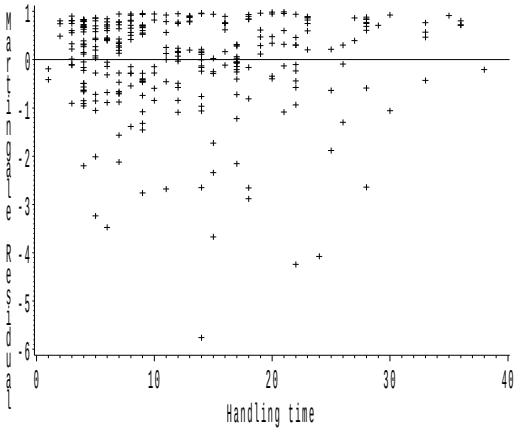
Martingale residuals vs length of fish



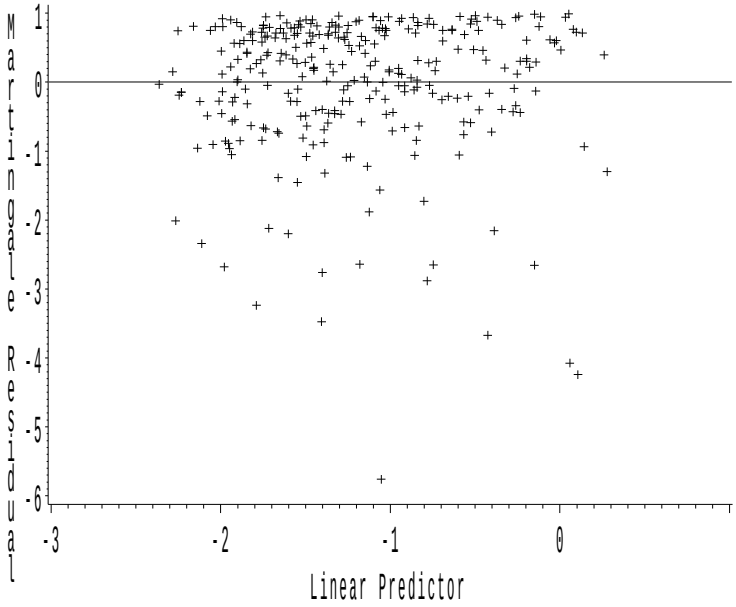
Martingale residuals vs log(catch)



Martingale residuals vs handling



Martingale residuals vs predicted values



Deviance Residuals

One problem with the martingale residuals is that they tend to be asymmetric.

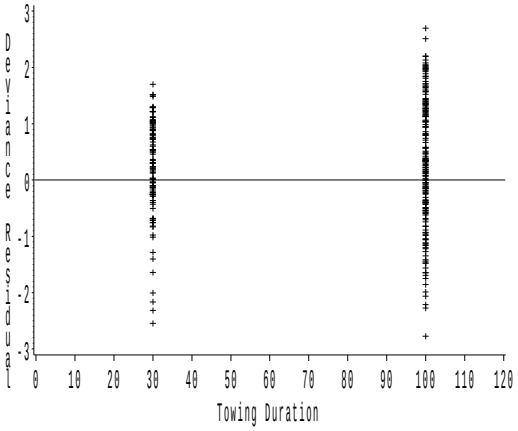
One solution is to use **deviance residuals**. For subject i, it is defined as a function of the martingale residual (M_i):

$$\hat{D}_i = \text{sign}(M_i)\sqrt{-2[M_i + \delta_i \log(\delta_i - M_i)]}$$

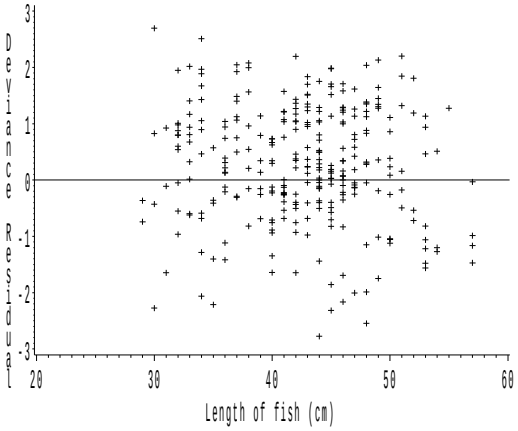
They can be plotted versus the prognostic index or the individual covariates, the same as for the Martingale residuals.

Deviance residuals are more like residuals from OLS regression (i.e. roughly mean=0, s.d.=1). But in practice they have not been "as useful as anticipated".

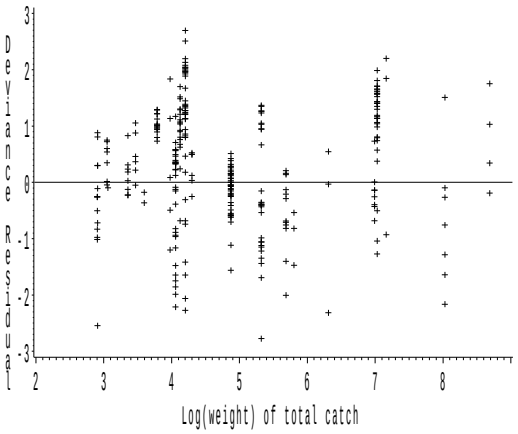
Deviance residuals vs towing duration



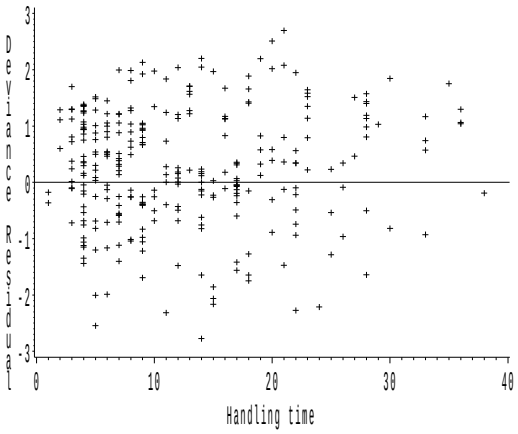
Deviance residuals vs length of fish

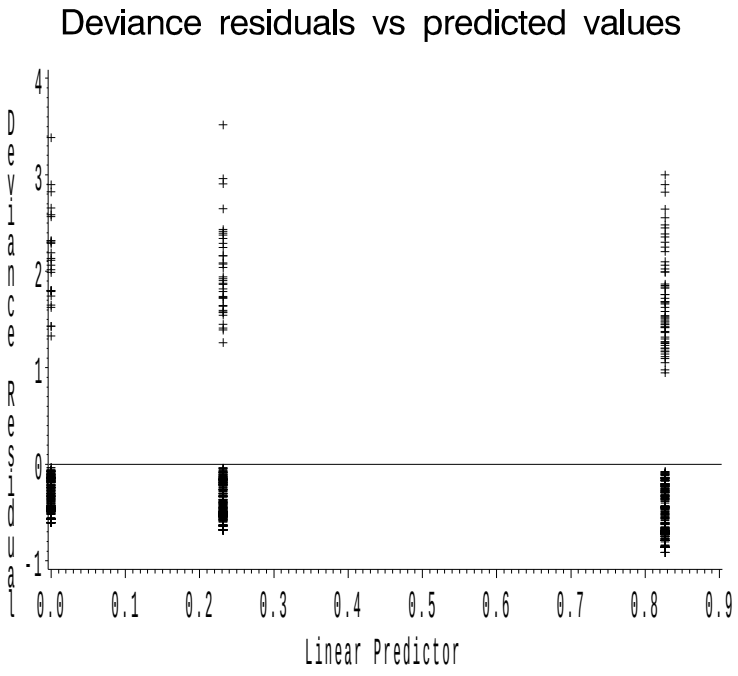


Deviance residuals vs log(catch)



Deviance residuals vs handling




Deviance residuals vs predicted values
Deviance Residual
4
3
2
1
0
-1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Linear Predictor

(d) Cumulative Martingale Residuals

Lin et al. (1993) developed powerful methods using cumulative sums of martingale residuals. Their procedure gives both graphical check and formal tests, and is sensitive to various departures from the Cox model.

Recall the martingale residuals

$$M_i = \delta_i - \hat{\Lambda}_i(X_i)$$

We may form the following two types of cumulative residuals ('cumulate' by value of a covariate component or of the prognostic index)

$$W_k(z) = \sum_{i=1}^{n} I(Z_{ik} \leq z) M_i,$$

plotting this vs. z checks the functional form of the kth covariate, Z_k;

and

$$W_r(r) = \sum_{i=1}^{n} I(\hat{\beta}' \mathbf{Z}_i \leq r) M_i,$$

plotting this vs. r checks the (exponential) link function of the Cox model.

More generally we may use the martingale process over time

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) e^{\beta_0' Z_i} \lambda_0(u) du$$

and

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) e^{\hat{\beta}' Z_i} d\hat{\Lambda}_0(u),$$

$i = 1, ..., n.$

The more general form of cumulative residuals are processes:

$$\begin{aligned} W_z(t, z) &= \sum_{i=1}^{n} f(\mathbf{Z}_i) I(\mathbf{Z}_i \leq \mathbf{z}) \hat{M}_i(t), \\ W_r(t, r) &= \sum_{i=1}^{n} f(\mathbf{Z}_i) I(\hat{\beta}' \mathbf{Z}_i \leq r) \hat{M}_i(t), \end{aligned}$$

The empirical score process is in fact $U(\hat{\beta}, t) = \sum_{i=1}^{n} Z_i \hat{M}_i(t)$, and the cumulative Schoenfeld residuals is also a special case of cumulative residuals, for checking the PH assumption.

One can further standardize the above to a standardized score process (bottom panel of Figure 2 on page 11), which has the asymptotic distribution of a Brownian Bridge, so that p-value can be obtained from its maximum absolution value (Fleming and Harrington book, p. 191-4).

These powerful cumulative residuals are implemented in the R package 'timereg' (Martinussen and Scheike).

Assessing the PH assumption

There are several options for checking the assumption of proportional hazards:

I. **Graphical**

(a) Plots of survival estimates for two subgroups;
(Indications of non-PH: *estimated survival curves are fairly separated, then converge or cross (why?)*)

(b) Plots of $\log[-\log(\hat{S}(t))]$ for two subgroups;
(*unparallel; see later*)

(c) Plots of (weighted, cumulative) Schoenfeld residuals vs time;
(*without cumulating: increase or decrease over time, may fit a OLS regression line to see the trend*)

(d) Some (like Kleinbaum, ch.4) also suggest plots of observed survival probabilities (estimated using KM) versus expected under PH model, but survival curves tend not to be sensitive (they all decrease monotonically between [0,1]).

II. **Formal goodness of fit tests** - Many such tests have been developed in the literature (it was a hot research area, see *Encyclopedia of Biostatistics* for some of them), unfortunately many of them are not available in common softwares. We will talk about a couple of these (including interaction terms between a covariate and t we talked about).

If PH doesn't exactly hold for a particular covariate but we fit the PH model anyway, then what we are getting is sort of an average log HR, averaged over the event times (more later).

Implications of proportional hazards

Consider a PH model with covariate $\mathbf{Z}$:

$$\lambda(t; Z) = \lambda_0(t) e^{\boldsymbol{\beta}'\mathbf{Z}}$$

Then,

$$S(t|\mathbf{Z}) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{Z})},$$

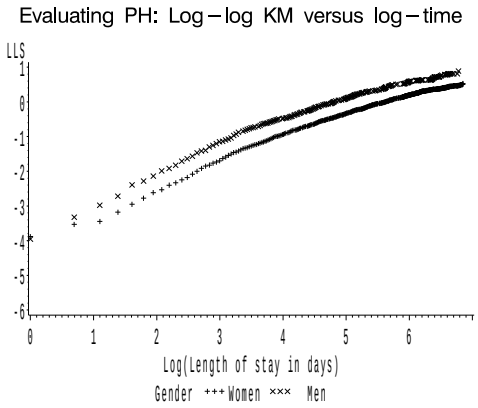
so

$$\log[-\log[S(t|\mathbf{Z})]] = \log[-\log[S_0(t)]] + \boldsymbol{\beta}'\mathbf{Z}.$$

Thus, to assess if the hazards are actually proportional to each other over time (using graphical option I(b))

- calculate Kaplan Meier Curves for various levels of Z (can we use the Brewlow's estimate here?)
- compute $\log[-\log(\hat{S}(t; Z))]$ (i.e., log cumulative hazard) from the KM
- plot vs time to see if they are parallel (lines or curves)

Note: If Z is continuous, break into categories.



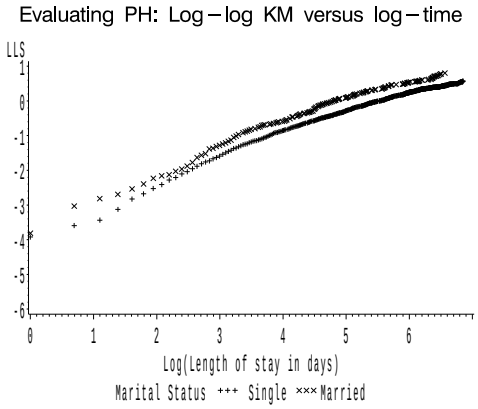


Figure 3: Log cumulative hazard plots are easier to view and tend to give more stable estimates.

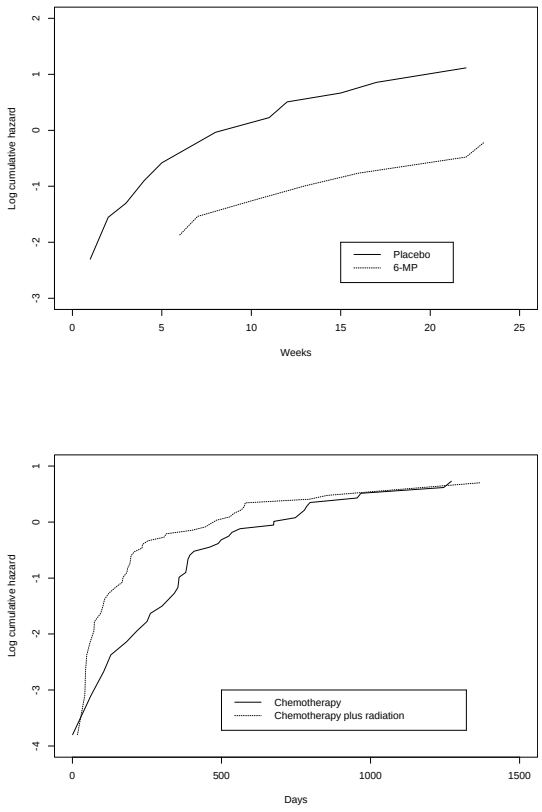
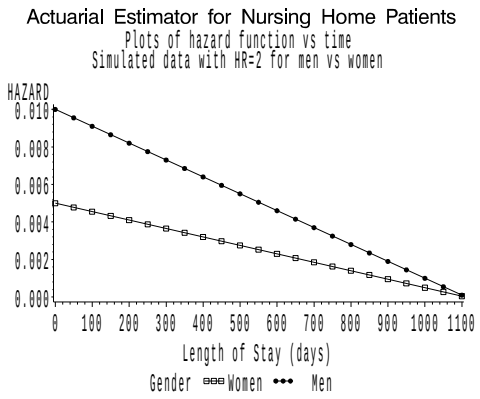


Figure 4: Assessment of proportional hazards for Freireich data and Stablein data.

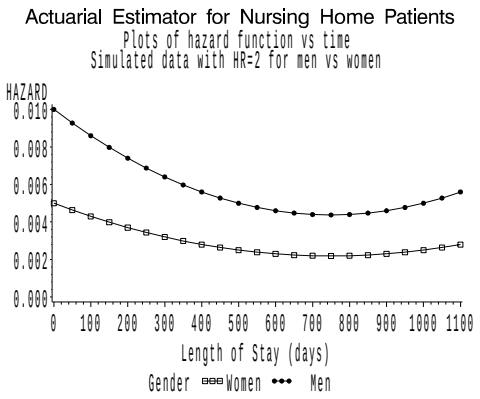
[Reading] Question: Why not just compare the underlying hazard rates to see if they are proportional?

Reason 1: It's hard to eyeball two curves and see if they are proportional - it would be easier to look for a constant shift between lines, i.e. if they are parallel.

Reason 2: It is not so easy to estimate hazard functions (we will not talk about in this class), and the estimated hazard rates tend to be more unstable than the cumulative hazard rates.

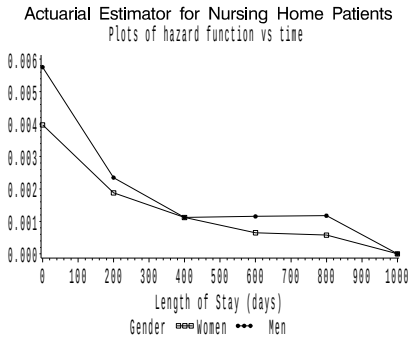


Weibull-type hazard:

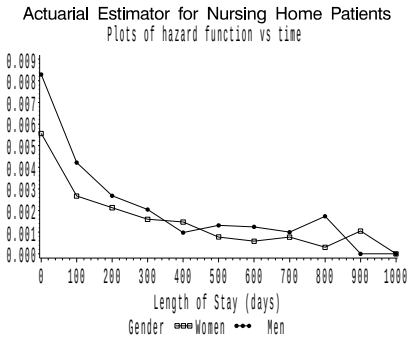


U-shaped hazard:

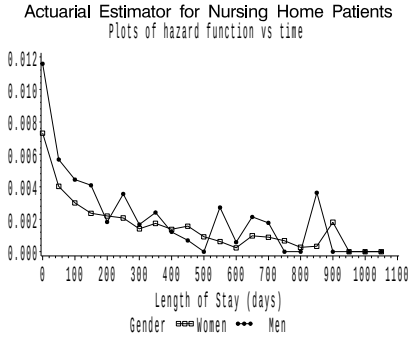
Figure 5: Not easy to eyeball if two curve are proportional to each other.



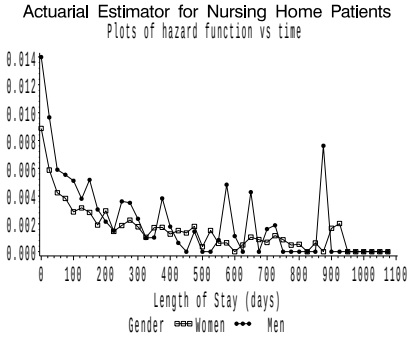
200 day intervals



100 day intervals



50 day intervals



25 day intervals

Figure 6: Estimated hazard by grouping data into intervals.

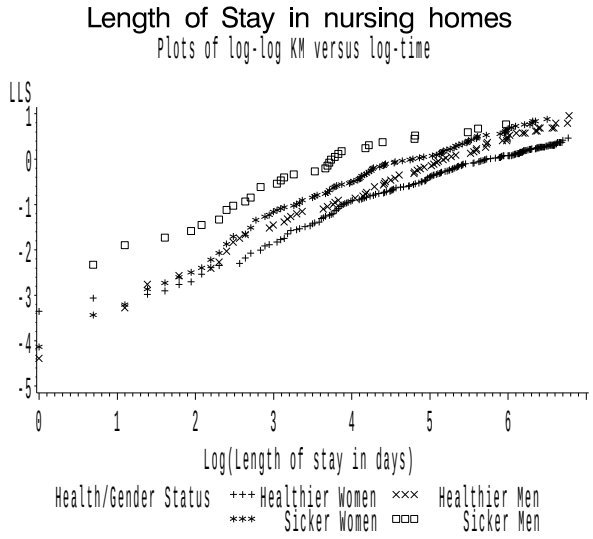


Figure 7: Log[-log(survival)] Plots for Health status*gender.

Assessing proportionality with several covariates

If there is enough data and you only have a couple of categorical covariates, create a new covariate that takes a different value for every combination of covariate values (prognostic index).

Example: Health status and gender for nursing home

Assessing PH Assumption for Several Covariates

Suppose we have several covariates ($\mathbf{Z} = Z_1, Z_2, \ldots Z_p$), and we want to know if the following PH model holds:

$$\lambda(t; \mathbf{Z}) = \lambda_0(t)\ e^{\beta_1 Z_1 + \ldots + \beta_p Z_p}$$

To start, we fit a model which stratifies by Z_k (discretize first if continuous):

$$\lambda(t; \mathbf{Z}) = \lambda_{0Z_k}(t)\ e^{\beta_1 Z_1 + \ldots + \beta_{k-1} Z_{k-1} + \beta_{k+1} Z_{k+1} + \ldots + \beta_p Z_p}$$

We can estimate the baseline survival function, $S_{0Z_k}(t)$, for each level of Z_k.

Then we compute $\log[-\log S_{0Z_k}(t)]$ for each level of Z_k, and graphically check whether the log cumulative hazards are parallel across strata levels (why?).

Ex: PH assumption for `gender` (nursing home data):

- include `married` and `health` as covariates in a Cox PH model, but *stratify* by `gender`.
- calculate the baseline survival function for each level of the variable `gender` (i.e., males and females)
- plot the log stratum-specific baseline cumulative hazards for males and females and evaluate whether the lines (curves) are parallel

In the above example, we make the PH assumption for `married` and `health`, but not for `gender`.

This is like getting a KM ('observed') survival estimate for each gender without assuming PH, but is more flexible since we can control for other covariates.

We would repeat the stratification for each variable for which we wanted to check the PH assumption.

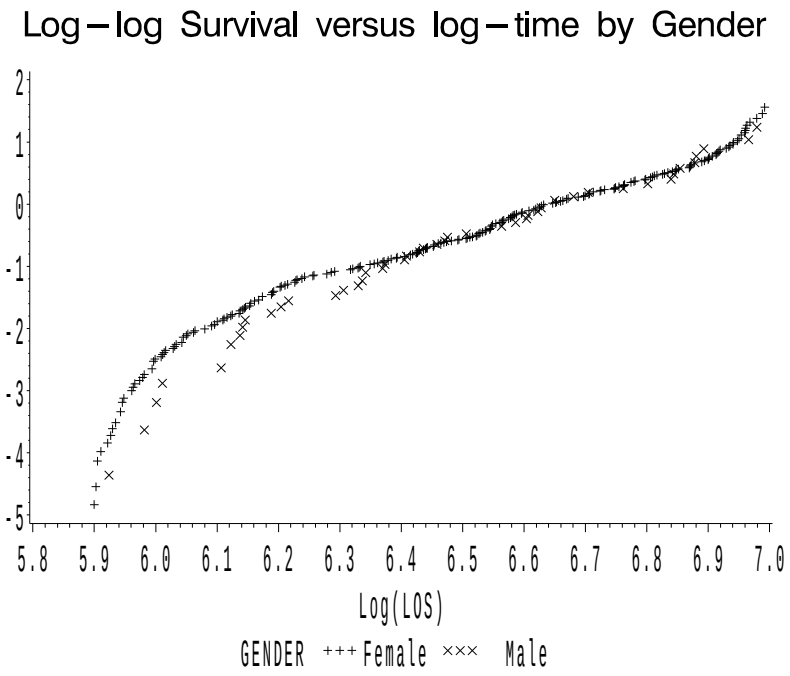


Figure 8: Log[-log(survival)] Plots for for Gender Controlling for Marital and Health Status.

Tests Using Time-Covariate Interactions

The above tests are mainly graphical (some did go further and developed formal tests associated with the graphical ones) Here we show one type of formal tests using time-covariate interactions.

Consider a PH model with two covariates Z_1 and Z_2. The standard PH model assumes

$$\lambda(t; Z) = \lambda_0(t)\ e^{\beta_1 Z_1 + \beta_2 Z_2}$$

If we want to test the proportionality of the effect of Z_2, we can try adding an interaction with time:

$$\lambda(t; Z) = \lambda_0(t)\ e^{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_2 * Q(t)}$$

A test of the coefficient β_3 would be a test of the proportional hazards assumption for Z_2.

Examples we've seen are $Q(t) = t$, $e^{-t/c}$, it can also be a step function (Moreau et al. 1985). In the latter case there are more parameters associated with $Q(t)$ to test the PH assumption.

What if proportional hazards fails?

Why is it important to assess the PH assumption?

We said before that if the truth is non-PH and we fit a PH model, then we are estimating some sorta average log HR. The trouble is that, if there is censoring, this average is affected by the nuisance censoring mechanism in a complex way, therefore the interpretation is difficult (Xu and O'Quigley, 2000). Note that under the PH model, even if there is censoring, we are still estimating the true log HR. But when we are outside the PH model, the censoring distribution comes into play.

What can we do?

- try transformations on the covariates, higher order terms, and interaction among the covariates
- do a stratified analysis
- fit a $\boldsymbol{\beta}(t)$ model
- try other models

Stratified Analyses

Suppose:

- we are happy with the proportionality assumption on Z_1
- proportionality simply does not hold between various levels of a second variable Z_2.

If Z_2 is discrete (with a levels) and there is enough data, fit the following **stratified model**:

$$\lambda(t; Z_1, Z_2) = \lambda_{Z_2}(t)e^{\beta Z_1}$$

(Recall the interpretion is something like, a new treatment might lead to a 50% decrease in hazard of death over the standard treatment, but the hazard for standard treatment might be different for each stratum.)

A stratified model can be useful both for primary analysis and for checking the PH assumption.

More on Cox model diagnostics

Using Residual plots to explore relationships

If you calculate martingale residuals without certain covariates in the model and then plot against these covariates, you obtain a graphical impression of the relationship between the covariate and the log hazard (the derivation involves martingale theory for the Cox model and is omitted here).

In other words, suppose that $\mathbf{Z}$ is a vector of covariates already included in the model, and we want to decide the funtional form of an additional covariate Z_2.

If Z_2 is not strongly correlated with $\mathbf{Z}$, then fit a model with $\mathbf{Z}$ only and a smoothed plot of the martingale residuals against Z_2 will show approximately the correct funtional form for Z_2.

If the smoothed plot appears linear, then Z_2 may enter linearly into the Cox model, i.e. $\lambda_(t|\mathbf{Z}, Z_2) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z} + \beta_2 Z_2$.

We will use this in the case study later.