Model Selection in Survival Analysis

Suppose we have a censored survival time that we want to model as a function of a (possibly) set of covariates. Two important questions are:

- How to decide which covariates to use
- How to decide if the final model fits well

To address these topics, we'll consider a new example:

Survival of Atlantic Halibut - Smith et al

Obs #	*Survival* *Time* (min)	*Censor*ing Indicator	*Tow* *Dur*ation (min.)	Diff in *Depth*	*Length* of Fish (cm)	*Handling* Time (min.)	Total *log(catch)* ln(weight)
100	353.0	1	30	15	39	5	5.685
109	111.0	1	100	5	44	29	8.690
113	64.0	0	100	10	53	4	5.323
116	500.0	1	100	10	44	4	5.323
⋮							

Process of Model Selection

Collett (Section 3.6) has an excellent discussion of various approaches for model selection. In practice, model selection proceeds through a combination of

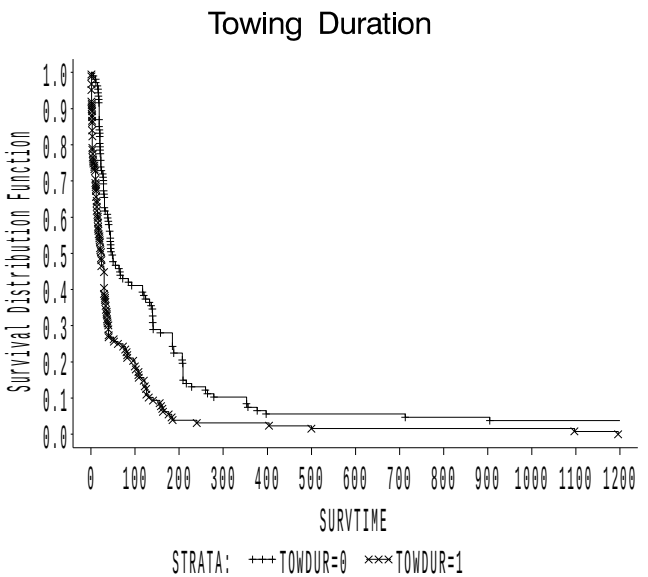
- knowledge of the science
- trial and error, common sense
- automatic variable selection procedures
  - forward selection
  - backward selection
  - stepwise seletion

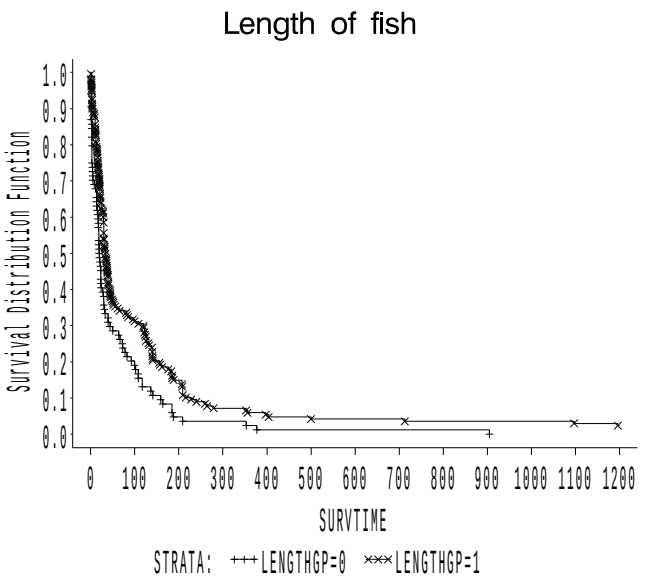
Many advocate the approach of first doing a univariate analysis to "screen" out potentially significant variables for consideration in the multivariate model (see Collett).

Let's start with this approach!

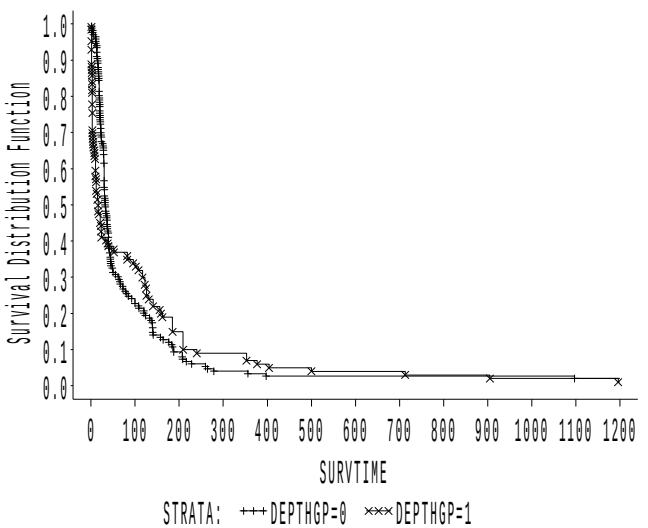
Univariate KM plots of Atlantic Halibut survival

(continuous variables have been dichotomized)

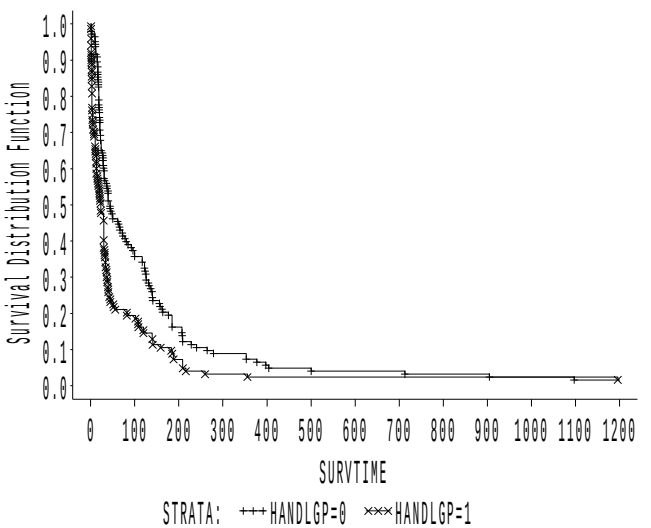


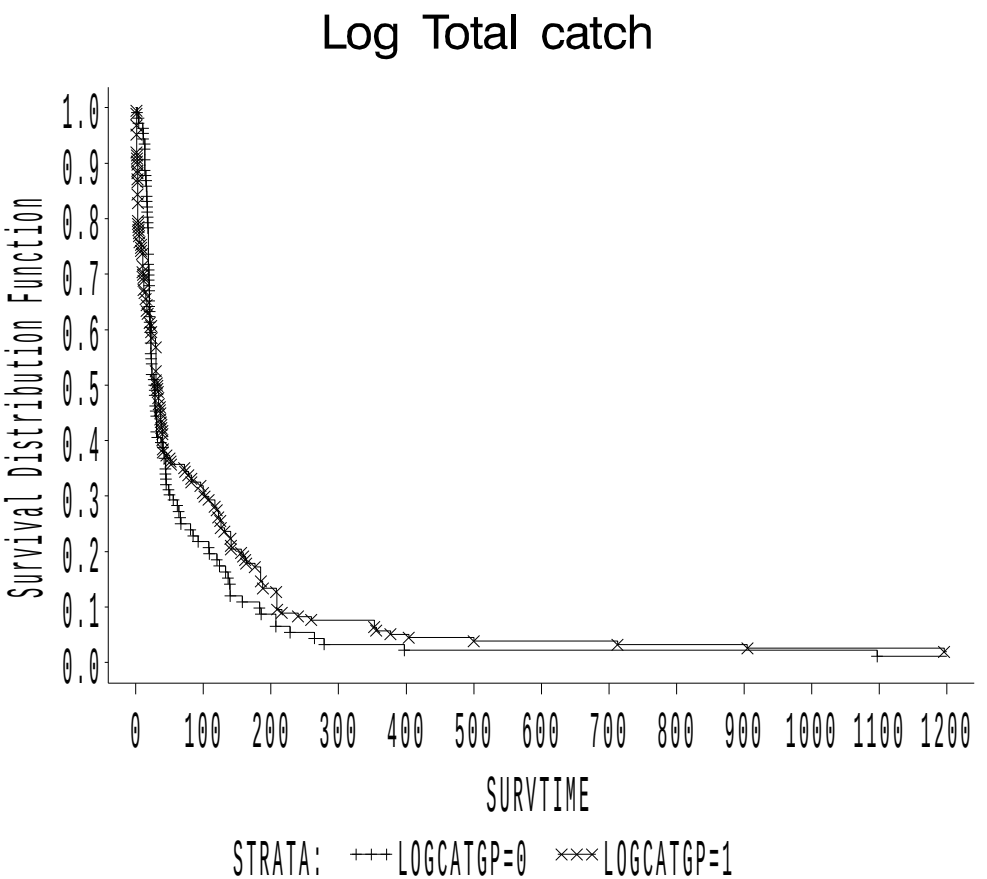


Difference in depth



Handling time





Which covariates look like they might be important?

Automatic Variable selection procedures in Stata and SAS

Statistical Software:

- Stata: `sw` command before `cox` command
- SAS: `selection=` option on model statement of `proc phreg`

Options:

(1) forward

(2) backward

(3) stepwise

(4) best subset (SAS only, using `score` option)

One drawback of these options is that they can only handle variables one at a time. When might that be a disadvantage?

Collett's Model Selection Approach
Section 3.6.1

This approach assumes that all variables are considered to be on an equal footing, and there is no *a priori* reason to include any specific variables (like treatment).

Approach:

(1) Fit a univariate model for each covariate, and identify the predictors significant at some level p_1, say 0.20.

(2) Fit a multivariate model with all significant univariate predictors, and use *backward* selection to eliminate non-significant variables at some level p_2, say 0.10.

(3) Starting with final step (2) model, consider each of the non-significant variables from step (1) using *forward* selection, with significance level p_3, say 0.10.

(4) Do final pruning of main-effects model (omit variables that are non-significant, add any that are significant), using *stepwise* regression with significance level p_4. At this stage, you may also consider adding interactions between any of the main effects currently in the model, under the hierarchical principle.

Collett recommends using a likelihood ratio test for all variable inclusion/exclusion decisions.

Stata Command for Forward Selection:

Forward Selection $\Longrightarrow$ use $pe(\alpha)$ option, where α is the significance level for entering a variable into the model.

```
. use halibut
. stset survtime censor
. sw cox survtime towdur depth length handling logcatch,
> dead(censor)  pe(.05)
                      begin with empty model
p = 0.0000 <  0.0500  adding   handling
p = 0.0000 <  0.0500  adding   logcatch
p = 0.0010 <  0.0500  adding   towdur
p = 0.0003 <  0.0500  adding   length
Cox Regression -- entry time 0                    Number of obs =    294
                                                  chi2(4)       =  84.14
                                                  Prob > chi2   = 0.0000
Log Likelihood = -1257.6548                       Pseudo R2     = 0.0324
------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
handling |   .0548994   .0098804      5.556   0.000     .0355341    .0742647
logcatch |  -.1846548    .051015     -3.620   0.000     .2846423   -.0846674
  towdur |   .5417745   .1414018      3.831   0.000     .2646321     .818917
  length |  -.0366503   .0100321     -3.653   0.000    -.0563129   -.0169877
------------------------------------------------------------------------------
```

Stata Command for Backward Selection:

Backward Selection $\Longrightarrow$ use $pr(\alpha)$ option, where α is the significance level for a variable to remain in the model.

```
. sw cox survtime towdur depth length handling logcatch,
> dead(censor) pr(.05)
                      begin with full model

p = 0.1991 >= 0.0500  removing depth

Cox Regression -- entry time 0                    Number of obs =    294
                                                  chi2(4)       =  84.14
                                                  Prob > chi2   = 0.0000
Log Likelihood = -1257.6548                       Pseudo R2     = 0.0324
------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000     .2646321      .818917
logcatch |  -.1846548    .051015     -3.620   0.000    -.2846423    -.0846674
  length |  -.0366503   .0100321     -3.653   0.000    -.0563129    -.0169877
handling |   .0548994   .0098804      5.556   0.000     .0355341     .0742647
------------------------------------------------------------------------------
```

Stata Command for Stepwise Selection:

Stepwise Selection $\Longrightarrow$ use both $pe(.)$ and $pr(.)$ options, with $pr(.) > pe(.)$

```
. sw cox survtime towdur depth length handling logcatch,
> dead(censor) pr(0.10) pe(0.05)
                      begin with full model
p = 0.1991 >= 0.1000  removing depth
Cox Regression -- entry time 0                    Number of obs =    294
                                                  chi2(4)       =  84.14
                                                  Prob > chi2   = 0.0000
Log Likelihood = -1257.6548                       Pseudo R2     = 0.0324
------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000  .2646321     .818917
handling |   .0548994   .0098804      5.556   0.000  .0355341    .0742647
  length |  -.0366503   .0100321     -3.653   0.000 -.0563129   -.0169877
logcatch |  -.1846548    .051015     -3.620   0.000 -.2846423   -.0846674
------------------------------------------------------------------------------
```

It is also possible to do forward stepwise regression by including both $pr(.)$ and $pe(.)$ options with `forward` option

Notes:

- When the halibut data was analyzed with the forward, backward and stepwise options, the same final model was reached. However, this will not always be the case.

- Variables can be forced into the model using the `lockterm` option in Stata and the `include` option in SAS. Any variables that you want to force inclusion of must be listed first in your model statement.

- Stata uses the Wald test for both forward and backward selection, although it has an option to use the likelihood ratio test instead (`lrtest`). SAS uses the score test to decide what variables to add and the Wald test for what variables to remove.

- If you fit a range of models manually, you can apply the AIC criteria described by Collett:

$$\text{minimize AIC} \quad = \quad -2 \ \log(\hat{L}) + (\alpha * q)$$

where q is the number of unknown parameters in the model and α is typically between 2 and 6 (they suggest $\alpha = 3$).

The model is then chosen which minimizes the AIC (similar to maximizing log-likelihood, but with a penalty for number of variables in the model)

Assessing overall model fit

How do we know if the model fits well?

- Always look at univariate plots (Kaplan-Meiers) Construct a Kaplan-Meier survival plot for each of the important predictors, like the ones shown at the beginning of these notes.
- Check proportionality assumption (this will be the topic of the next lecture)
- **Check residuals!**
  (a) generalized (Cox-Snell)
  (b) martingale
  (c) deviance
  (d) Schoenfeld
  (e) weighted Schoenfeld

Residuals for survival data are slightly different than for other types of models, due to the censoring. Before we start talking about residuals, 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 $S_i(T_i)$ (i.e., the survival function evaluated at the actual survival time T_i) should be from a uniform distribution on $[0, 1]$, and hence $-\log[S_i(T_i)]$ should be from a unit exponential distribution

More mathematically:

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

(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 should be like a censored sample from a unit exponential. This quantity is called the *generalized* or *Cox-Snell* residual.

Here is how the generalized residual might be used. Suppose we fit a PH model:

$$S(t;Z) = [S_0(t)]^{\exp(\beta Z)}$$

or, in terms of hazards:

$$\begin{aligned} \lambda(t;Z) &= \lambda_0(t)\exp(\beta Z) \\ &= \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k) \end{aligned}$$

After fitting, we have:

- $\hat{\beta}_1, \ldots, \hat{\beta}_k$
- $\hat{S}_0(t)$

So, for each person with covariates $\mathbf{Z}_i$, we can get

$$\hat{S}(t; \mathbf{Z}_i) = [\hat{S}_0(t)]^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)}$$

This gives a predicted survival probability at each time t in the dataset (see notes from the previous lecture).

Then we can calculate

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

In other words, first we find the predicted survival probability at the actual survival time for an individual, then log-transform it.

Example: Nursing home data

Say we have

- 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 $\hat{S}(941) = 0.260$.

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

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.

Based on the properties of a unit exponential model

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

To convince yourself of this, start with $S(t) = e^{-\lambda t}$ and calculate $\log[-\log S(t)]$. What do you get for the slope and intercept?

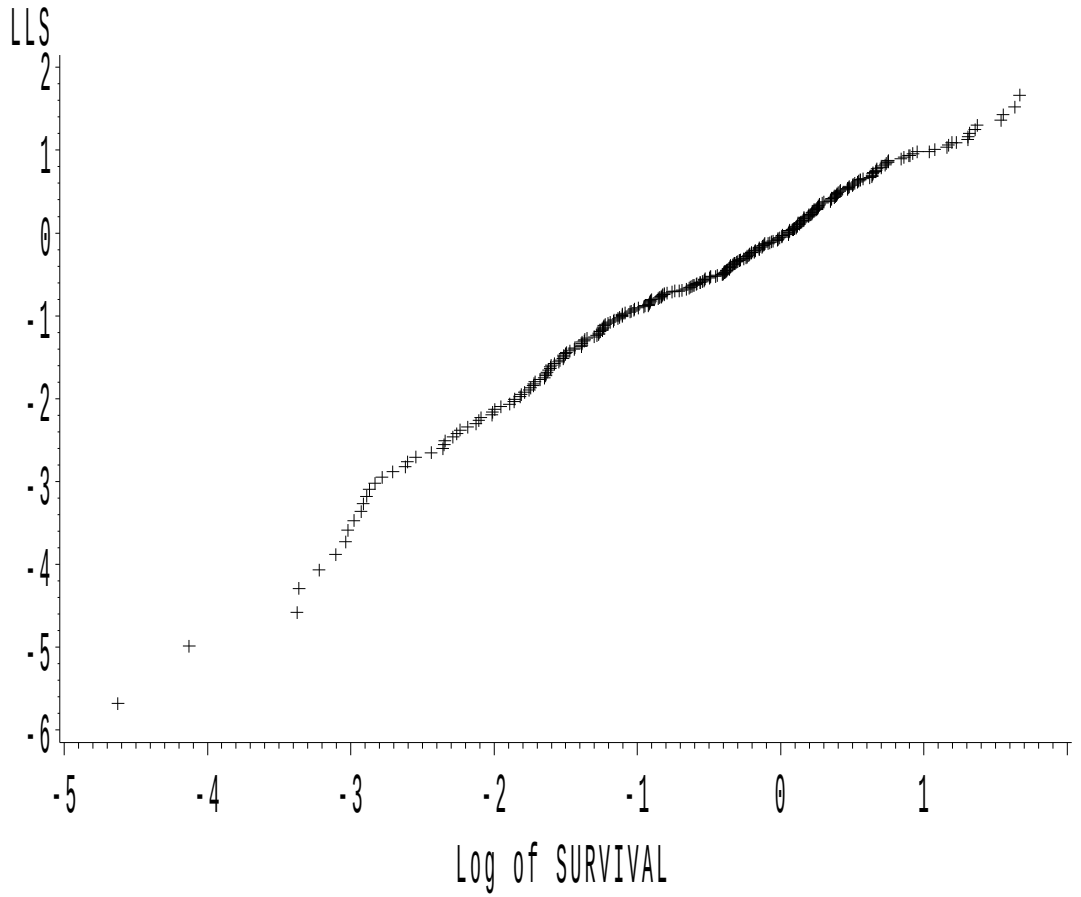
(Note: this does not necessarily mean that the underlying distribution of the original survival times is exponential!)

Obtaining the generalized residuals from Stata

- Fit a Cox PH model with `stcox` and the `mgale(`*`newvar`*`)` option
- Use the `predict` command with the `csnell` option
- Define a survival dataset using the Cox-Snell residuals as the "pseudo" failure times
- Calculate the estimated KM survival
- Take the $\log[-\log(S(t))]$ based on the above
- Generate the log of the Cox-Snell residuals
- Graph $\log[-\log S(t)]$ vs $\log(t)$

```
. stcox towdur handling length logcatch, mgale(mg)
. predict csres, csnell
. stset csres censor
. sts list
. sts gen survcs=s
. gen lls=log(-log(survcs))
. gen loggenr=log(csres)
. graph lls loggenr
```

Does the exponential model fit?



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

(b) Martingale Residuals

(see Fleming and Harrington, p.164)

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

$$r_i = \delta_i - \hat{\Lambda}(T_i)$$

Properties:

- r_i's have mean 0
- range of r_i's is between $-\infty$ and 1
- approximately uncorrelated (in samples)
- **Interpretation:** - the residual r_i can be viewed as the difference between the observed number of deaths (0 or 1) for subject i between time 0 and T_i, and the expected numbers based on the fitted model.

The **martingale residuals** can be obtained from Stata

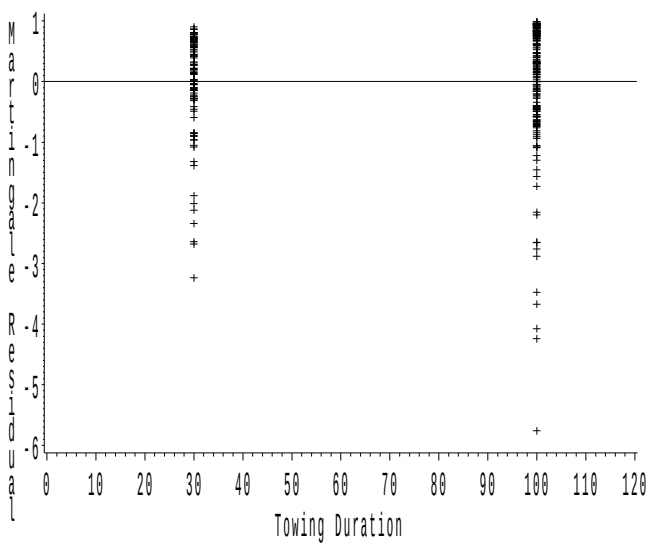
using the `mgale` option shown previously.

Once the martingale residual is created, you can plot it versus the predicted log HR (i.e., $\boldsymbol{\beta}\mathbf{Z}_i$), or any of the individual covariates.

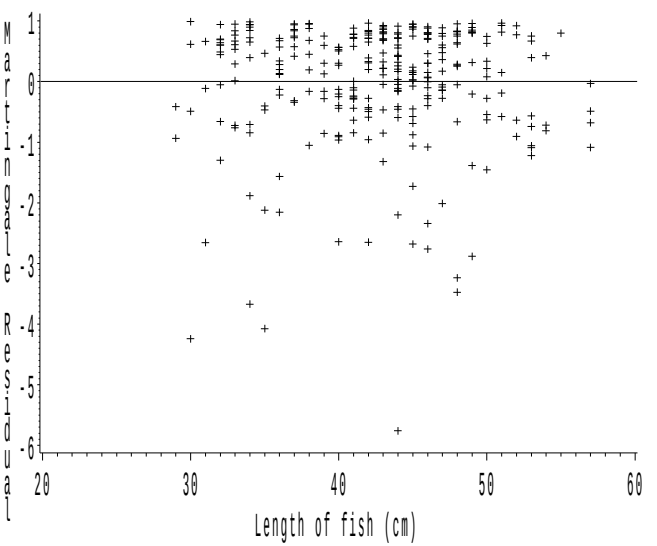
```
. stcox towdur handling length logcatch, mgale(mg)

. predict betaz=xb

. graph mg betaz

. graph mg logcatch

. graph mg towdur

. graph mg handling

. graph mg length
```

Martingale Residuals

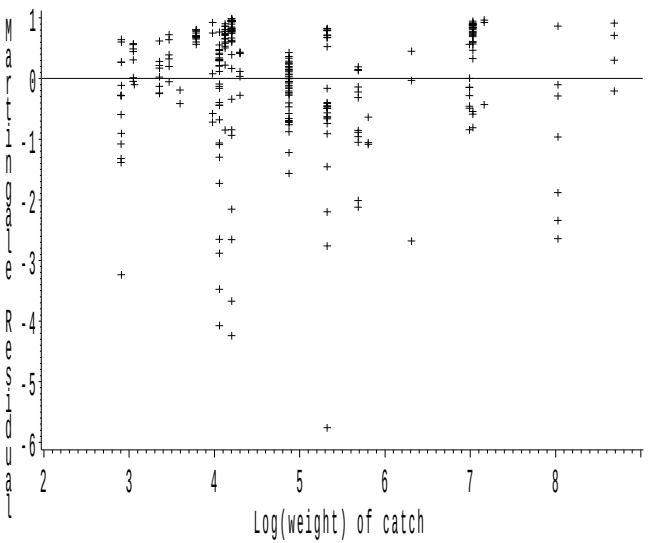
Martingale residuals vs towing duration



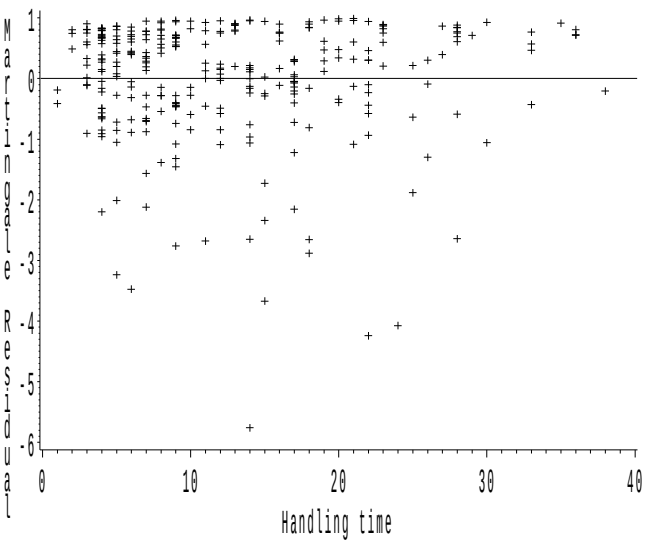
Martingale residuals vs length of fish



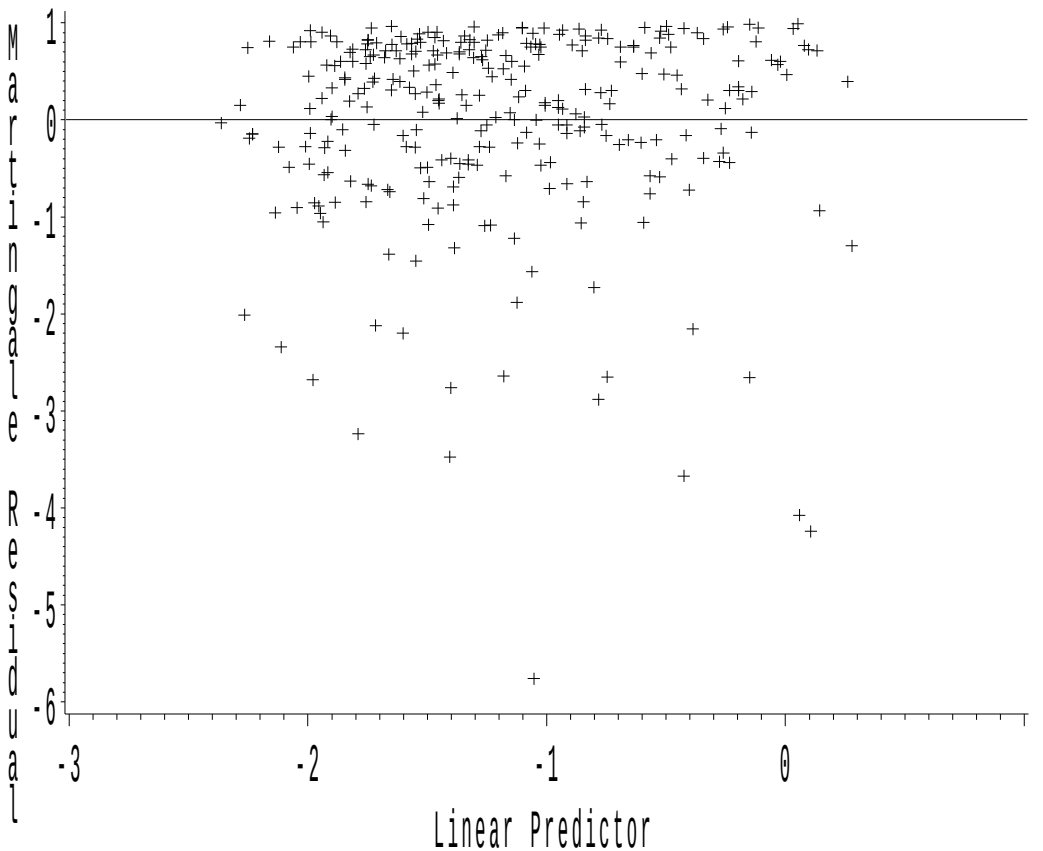
Martingale residuals vs log(catch)



Martingale residuals vs handling



Martingale residuals vs predicted values



(c) Deviance Residuals One problem with the martingale residuals is that they tend to be asymmetric.

A solution is to use **deviance residuals**. For person i, these are defined as a function of the martingale residuals (r_i):

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

In Stata, the deviance residuals are generated using the same approach as the Cox-Snell residuals.

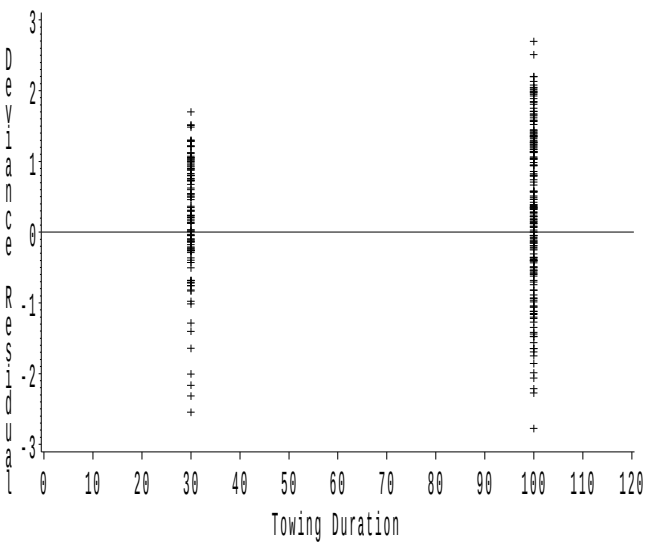
```
. stcox towdur handling length logcatch, mgale(mg)
. predict devres, deviance
```

and then they can be plotted versus the predicted log(HR) or the individual covariates, as shown for the Martingale residuals.

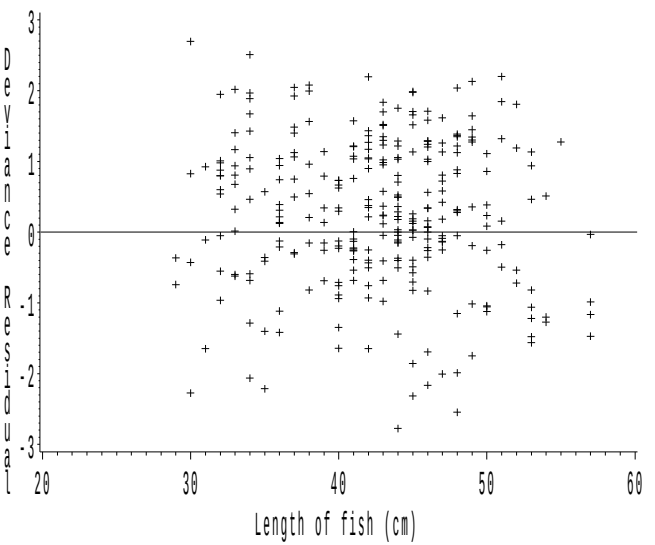
Deviance residuals behave much like residuals from OLS regression (i.e., mean=0, s.d.=1). They are negative for observations with survival times that are smaller than expected.

Deviance Residuals

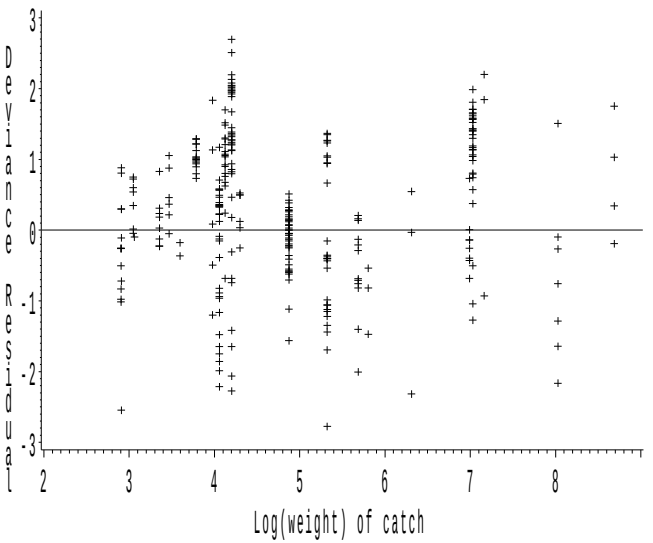
Deviance residuals vs towing duration



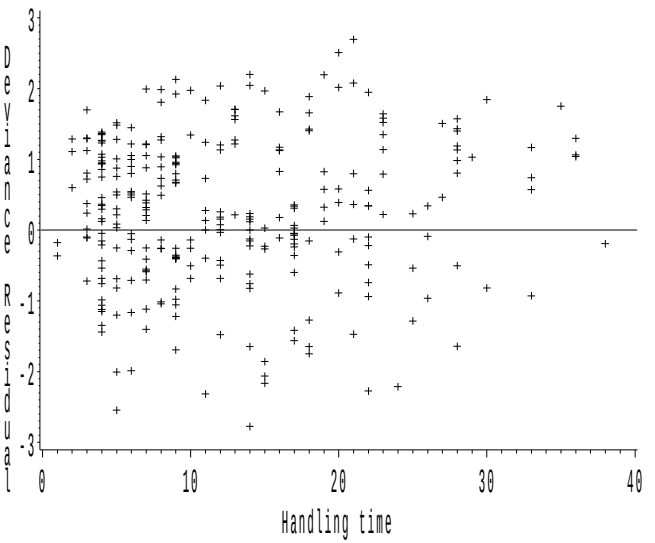
Deviance residuals vs length of fish



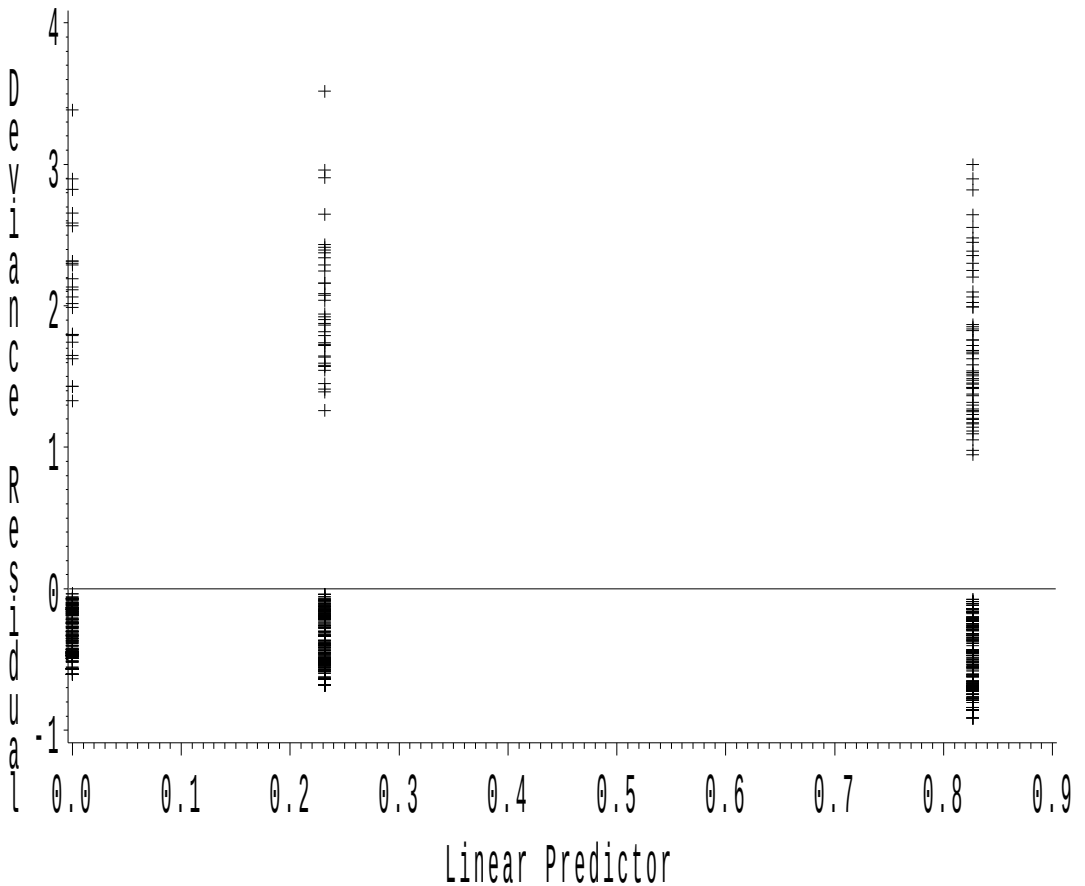
Deviance residuals vs log(catch)



Deviance residuals vs handling



Deviance residuals vs predicted values



(d) Schoenfeld Residuals

These are defined at each observed failure time as:

$$r_{ij}^{s} = Z_{ij}(t_i) - \bar{Z}_j(t_i)$$

Notes:

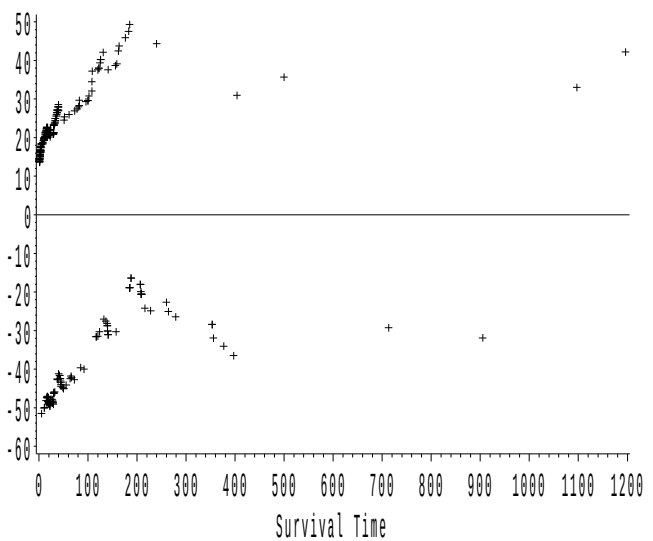
- represent the difference between the observed covariate and the average over the risk set at that time
- calculated for each covariate
- not defined for censored failure times.
- useful for assessing time trend or lack or proportionality, based on plotting versus event time
- sum to zero, have expected value zero, and are uncorrelated (in samples)

In Stata, the Schoenfeld residuals are generated in the `stcox` command itself, using the `schoenf(`*`newvar(s)`*`)` option:

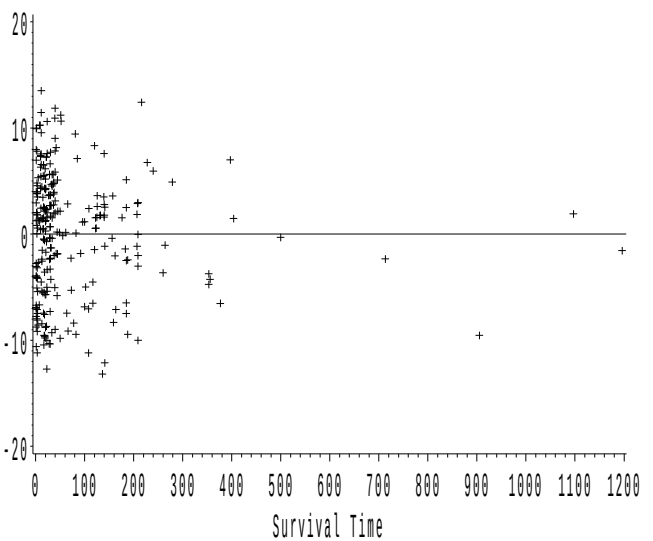
```
. stcox towdur handling length logcatch, schoenf(towres handres lenres
logres)

. graph towres survtime
```

Schoenfeld Residuals

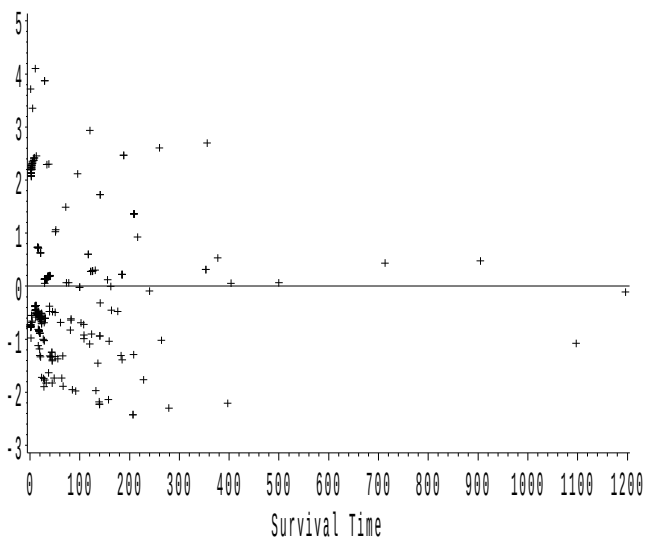
Schoenfeld resids for towing vs survival time



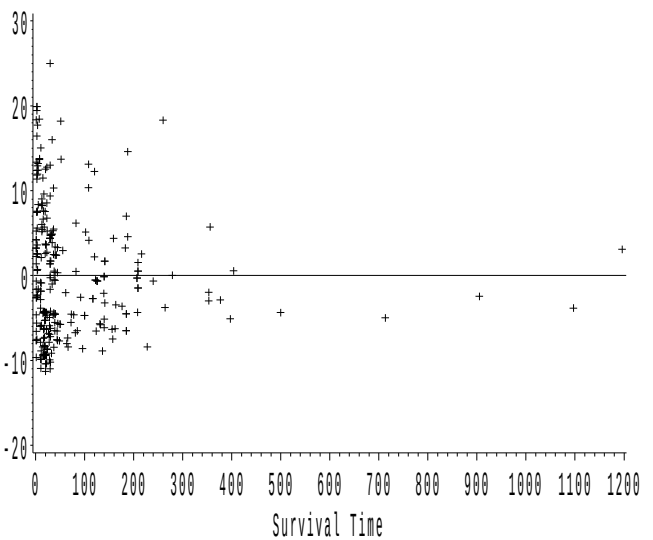
Schoenfeld resids for length vs survival time



Schoenfeld resids for log(catch) vs survival time



Schoenfeld resids for handling vs survival time



(e) Weighted Schoenfeld Residuals

These are actually used more often than the previous unweighted version, because they are more like the typical OLS residuals (i.e., symmetric around 0).

They are defined as:

$$r_{ij}^{w} \quad = \quad n\widehat{V}\, r_{ij}^{s}$$

where $\widehat{V}$ is the estimated variance 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.

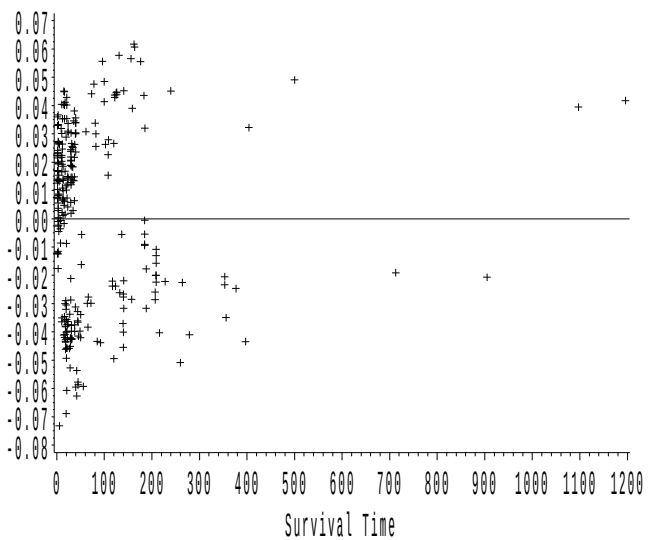
In Stata, use the command:

```
. stcox towdur length logcatch handling depth, scaledsch(towres2
> lenres2 logres2 handres2 depres2)
. graph logres2 survtime
```

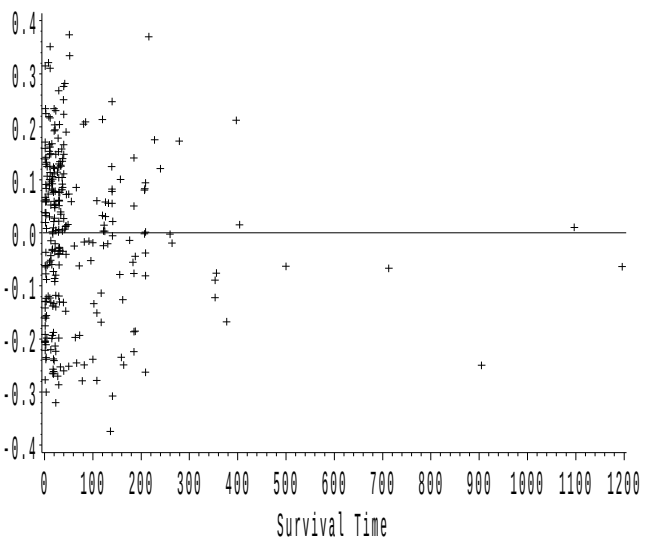
for up to k regressors in the model.

Weighted Schoenfeld Residuals

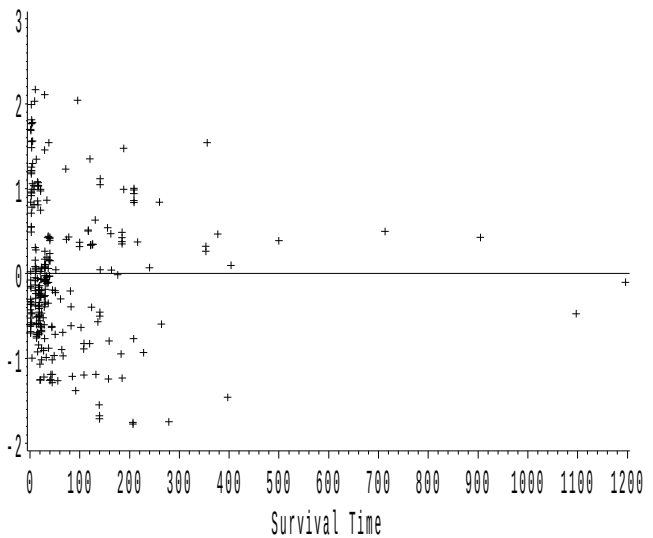
Weighted Schoenfeld resids for towing vs time



Schoenfeld resids for length vs time



Schoenfeld resids for log(catch) vs time



Schoenfeld resids for handling vs time