Chapter 5

Model Selection in Survival Analysis

Suppose we have a censored survival time that we want to model as a function of a (possibly large) 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*		*Tow*	Diff	*Length*	*Handling*	Total
	Time	*Censoring*	*Duration*	in	of Fish	Time	*log(catch)*
#	(min)	Indicator	(min.)	*Depth*	(cm)	(min.)	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
⋮							

Reading:

Collett, Section 3.6

Hosmer, Lemeshow, & May

Chapter 5: Model Development
Chapter 6: Assessment of Model Adequacy (sections 6.1-6.2)

5.1 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

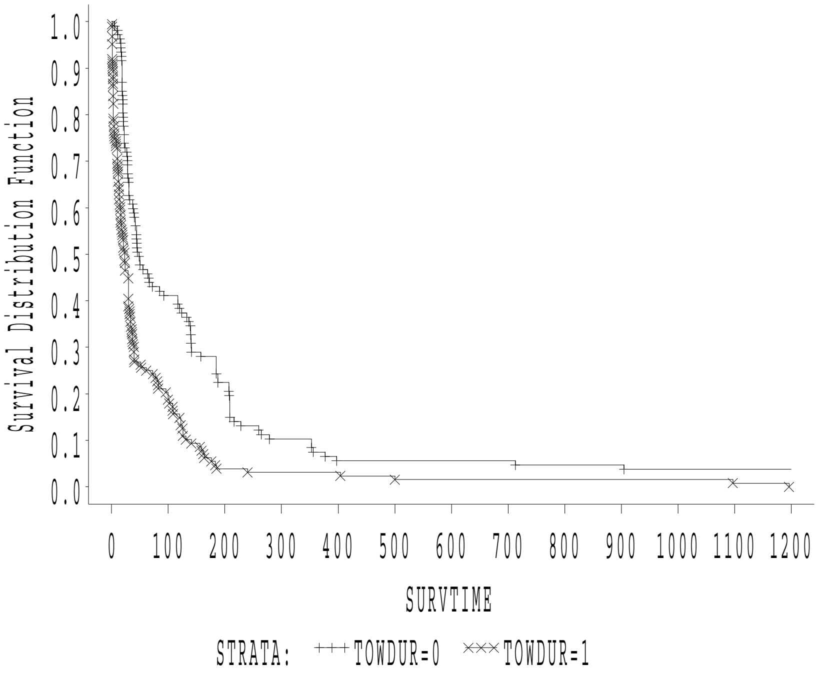
- knowledge of the science
- trial and error, common sense
- automatic variable selection procedures
  - forward selection
  - backward selection
  - stepwise selection

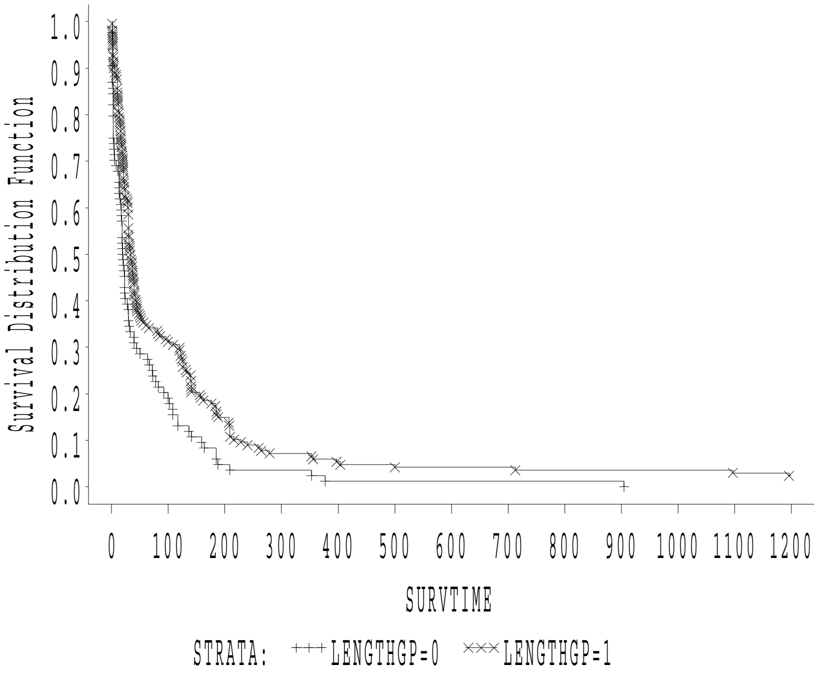
Many advocate the approach of first doing a univariate analysis to “screen” out potentially significant variables for consideration in the multivariate model (see Collett).

Let’s start with this approach!

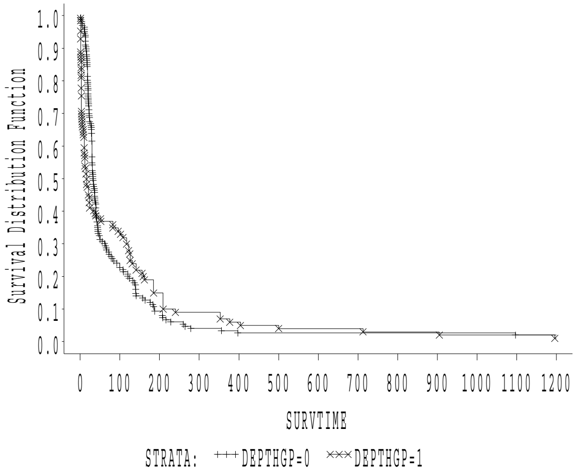
Univariate KM plots of Atlantic Halibut survival

(continuous variables have been dichotomized)

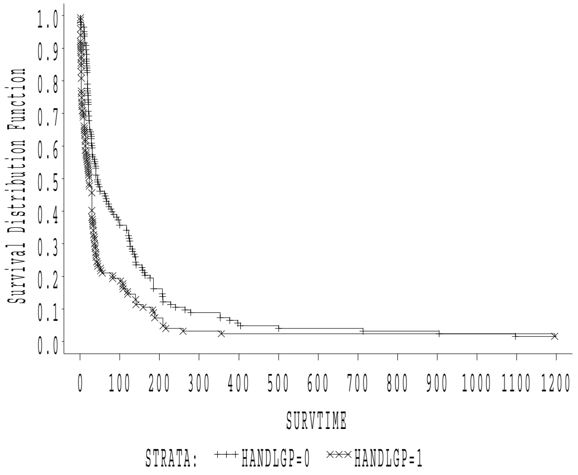

Towing Duration


Length of fish

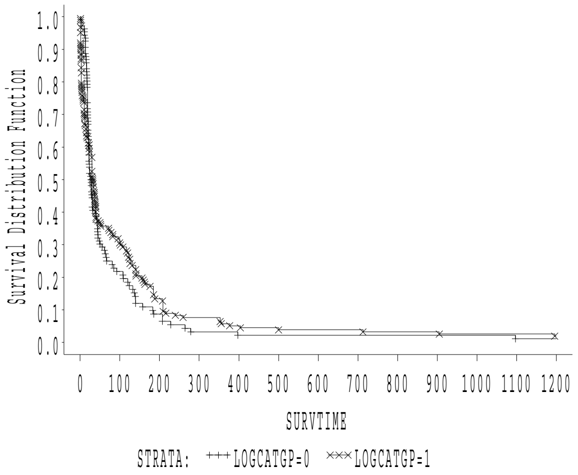
Difference in depth



Handling time



Log Total catch



Which covariates look like they might be important?

5.1.1 Automatic Variable Selection Procedures: Stata and SAS

Statistical Software:

- Stata: `stepwise` (or `sc`) 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)

(5) lasso (penalized regression)

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:

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

```
. use halibut

. stset survtime censor

. stepwise cox towdur depth length handling logcatch, 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.

```
. stepwise cox towdur depth length handling logcatch, 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 $pr(.) > pe(.)$

```
. stepwise cox towdur depth length handling logcatch, 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

SAS programming statements for model selection

```
data fish;
 infile ’fish.dat’;
 input ID SURVTIME CENSOR TOWDUR DEPTH LENGTH HANDLING LOGCATCH;
run;

title ’Survival of Atlantic Halibut’;
*** automatic variable selection procedures;
proc phreg data=fish;
  model survtime*censor(0)= towdur depth length handling logcatch
        /selection=forward slentry=0.1 details;
  title2 ’Forward selection’;
run;

proc phreg data=fish;
  model survtime*censor(0)= towdur depth length handling logcatch
        /selection=backward slstay=0.1 details;
  title2 ’Backward selection’;
run;

proc phreg data=fish;
  model survtime*censor(0)= towdur depth length handling logcatch
        /selection=stepwise slentry=0.1 slstay=0.1 details;
  title2 ’Stepwise selection’;
run;

proc phreg data=fish;
  model survtime*censor(0)= towdur depth length handling logcatch
        /selection=score;
  title2 ’Best subsets selection’;
run;
```

Final model for stepwise selection approach

```
                      Survival of Atlantic Halibut
                           Stepwise selection

                          The PHREG Procedure

                Analysis of Maximum Likelihood Estimates

                 Parameter    Standard      Wald         Pr >         Hazard
Variable  DF      Estimate      Error    Chi-Square  Chi-Square      Ratio

TOWDUR     1      0.007740     0.00202     14.68004      0.0001      1.008
LENGTH     1     -0.036650     0.01003     13.34660      0.0003      0.964
HANDLING   1      0.054899     0.00988     30.87336      0.0001      1.056
LOGCATCH   1     -0.184655     0.05101     13.10166      0.0003      0.831

                 Analysis of Variables Not in the Model

                                 Score          Pr >
                 Variable      Chi-Square    Chi-Square

                 DEPTH             1.6661        0.1968

           Residual Chi-square = 1.6661  with 1 DF (p=0.1968)

NOTE: No (additional) variables met the 0.1 level for entry into the
      model.
```

```
                    Summary of Stepwise Procedure

           Variable        Number    Score        Wald        Pr >
 Step  Entered    Removed       In  Chi-Square  Chi-Square  Chi-Square

    1  HANDLING                  1     47.1417           .      0.0001
    2  LOGCATCH                  2     18.4259           .      0.0001
    3  TOWDUR                    3     11.0191           .      0.0009
    4  LENGTH                    4     13.4222           .      0.0002
```

Although **DEPTH** was included in the list of potential covariates, it was not included in the final model from any approach.

Note that SAS uses the score test to decide what variables to add and the Wald test for what variables to remove.

However, once the final model is selected, it provides the separate Wald Chi-square tests for each individual parameter.

Using this model selection approach, one can obtain the SCORE Chi-square statistic for a particular variable adjusting for the others in the model.

Output from PROC SAS “score” option

```
  NUMBER OF        SCORE   VARIABLES INCLUDED
  VARIABLES        VALUE   IN MODEL

          1      47.1417   HANDLING
          1      29.9604   TOWDUR
          1      12.0058   LENGTH
          1       4.2185   DEPTH
          1       1.4795   LOGCATCH
---------------------------------
          2      65.6797   HANDLING LOGCATCH
          2      59.9515   TOWDUR HANDLING
          2      56.1825   LENGTH HANDLING
          2      51.6736   TOWDUR LENGTH
          2      47.2229   DEPTH HANDLING
          2      32.2509   TOWDUR LOGCATCH
          2      30.6815   TOWDUR DEPTH
          2      16.9342   DEPTH LENGTH
          2      14.4412   LENGTH LOGCATCH
          2       9.1575   DEPTH LOGCATCH
-------------------------------------
          3      76.8829   LENGTH HANDLING LOGCATCH
          3      76.3454   TOWDUR HANDLING LOGCATCH
          3      75.5291   TOWDUR LENGTH HANDLING
          3      69.0334   DEPTH HANDLING LOGCATCH
          3      60.0340   TOWDUR DEPTH HANDLING
          3      56.4207   DEPTH LENGTH HANDLING
          3      55.8374   TOWDUR LENGTH LOGCATCH
          3      52.4130   TOWDUR DEPTH LENGTH
          3      34.7563   TOWDUR DEPTH LOGCATCH
          3      24.2039   DEPTH LENGTH LOGCATCH
--------------------------------------------
```

SAS “Score” option, continued

```
          4      94.0062   TOWDUR LENGTH HANDLING LOGCATCH
          4      81.6045   DEPTH LENGTH HANDLING LOGCATCH
          4      77.8234   TOWDUR DEPTH HANDLING LOGCATCH
          4      75.5556   TOWDUR DEPTH LENGTH HANDLING
          4      59.1932   TOWDUR DEPTH LENGTH LOGCATCH
-------------------------------------------------
          5      96.1287   TOWDUR DEPTH LENGTH HANDLING LOGCATCH
------------------------------------------------------
```

When only 1 variable is included, `depth` is more informative than `logcatch`. However, once `handling` is included in the model, `depth` becomes less informative.

Best multivariate model for all 3 options

```
                      Survival of Atlantic Halibut
                         Best Multivariate Model

                          The PHREG Procedure

                         Summary of the Number of
                        Event and Censored Values
                                                    Percent
                 Total       Event     Censored    Censored

                   294         273           21        7.14

                 Testing Global Null Hypothesis: BETA=0

               Without       With
Criterion    Covariates    Covariates    Model Chi-Square

-2 LOG L       2599.449      2515.310      84.140 with 4 DF (p=0.0001)
Score              .             .         94.006 with 4 DF (p=0.0001)
Wald               .             .         90.247 with 4 DF (p=0.0001)

                Analysis of Maximum Likelihood Estimates

                 Parameter     Standard     Wald          Pr >     Hazard
Variable  DF      Estimate      Error    Chi-Square Chi-Square     Ratio

TOWDUR     1      0.007740     0.00202     14.68004     0.0001     1.008
LENGTH     1     -0.036650     0.01003     13.34660     0.0003     0.964
HANDLING   1      0.054899     0.00988     30.87336     0.0001     1.056
LOGCATCH   1     -0.184655     0.05101     13.10166     0.0003     0.831
```

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:

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

  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)

Questions:

- When might we want to force certain variables into the model?

  (1)

  (2)

  (3) ...

- Would it be possible to get different final models from SAS and Stata?

- Based on what we've seen in the behavior of Wald tests, would SAS or Stata be more likely to add a covariate to a model in a forward selection model?

- If we use the AIC criteria with $\alpha = 3$, how does that compare to the likelihood ratio test?

5.1.2 Use of AIC or Likelihood Ratio Tests to Compare Models

Say we have fit several models to our data, and wish to choose the best one. One approach is to compare the log-likelihoods (or somewhat equivalently, the AIC measures).

Consider the example from Collett (Example 3.4) on treatment of hypernephroma (malignant tumor in the kidney). This study includes survival times of 36 patients who were all treated with chemotherapy and immunotherapy, but some also had surgical remove of their kidney (nephrectomy). (see Lee & Wang, 2003)

The data include age (<60 yrs, 60-70 yrs, >70 yrs), nephrectomy status, survival time, and censoring indicator. The values of the $-2\log\hat{L}$ and AIC (from SAS) are shown below:

Model #	Type of model	Terms in model	Variables	$-2\log\hat{L}$	AIC
(1)	Null	(none)	(none)	177.667	177.667
(2)	Univariate	Age	`age6070 agegt70`	172.172	176.172
(3)		Nephrectomy	`nephrect`	170.247	172.247
(4)	Main effect	Age, Nephrectomy	`age6070 agegt70 nephrect`	165.508	171.508

(5)	Interaction	Age, Nephrectomy,
		Age x Nephrectomy
			`age6070 agegt70 nephrect`
			`age6070N agegt70N`
				162.479	172.479

Nephrectomy example:

(1) **Is there evidence of an interaction between age and nephrectomy status?**

Compare Model (4) to Model (5):

$$\chi^2_{LR} = 165.508 - 162.479 = 3.029 \qquad \text{with 2 df (p=0.22)}$$

.

(2) **Is survival related to age?**

There are multiple ways to answer this question. For example,

(a) Is age a significant predictor of survival? (not considering nephrectomy status)

(b) After adjusting for nephrectomy status, is age a significant predictor of survival?

To answer (a), compare Model (2) (only age effects) to Model (1) (no covariates):

$$\chi^2_{LR} = 177.667 - 172.172 = 5.495 \qquad \text{with 2 df (p=0.064)}$$

.

To answer (b), compare Model (3) (only nephrectomy) to Model (4) (nephrectomy and age):

$$\chi^2_{LR} = 170.247 - 165.508 = 4.739 \qquad \text{with 2 df (p=0.094)}$$

.

Nephrectomy example, continued:

(3) **Is nephrectomy significantly associated with survival?**

Again, we can consider this question both unadjusted and adjusted for other covariates:

(a) unadjusted: $\chi^2_{LR} = 177.667 - 170.247 = 7.42$ with 1df (p=0.006).

(b) adjusted for age: $\chi^2_{LR} = 172.172 - 165.508 = 6.664$ with 1df (p=0.010)

(4) **Which is the best model based on minimizing the AIC?**

For Model (4), the estimated model is:

```
                  Analysis of Maximum Likelihood Estimates

                  Parameter     Standard                               Hazard
Parameter   DF     Estimate        Error  Chi-Square  Pr > ChiSq        Ratio

age6070      1      0.01239      0.42460      0.0009      0.9767        1.012
agegt70      1      1.34156      0.59176      5.1396      0.0234        3.825
nephrect     1     -1.41176      0.51523      7.5080      0.0061        0.244
```

Note: this was not a randomized study, and it is possible that clinicians may have chosen patients for nephrectomy based on their age. Thus, it may be necessary to adjust for age to avoid bias in estimating the association of nephrectomy with survival. In this particular study, the unadjusted HR for nephrectomy was $\widehat{HR} = 0.228$, so there was little evidence of such bias.

Survival of Multiple Myeloma Patients:

The goal of the study was to identify risk factors for survival of patients diagnosed with multiple myeloma. 48 patients were included (see p.9 of Collett, Table 1.3), and the risk factors included age, sex, blood urea nitrogen (BUN), serum calcium (Ca) and hemoglobin (HGB), the percentage of plasma cells in the bone marrow (Pcells), and an indicator of whether a specific protein was present in the urine.

The final values of $-2\log\hat{L}$ are shown in the Table on next page:

- **Can you complete the AIC column, setting the penalty term $\alpha = 2$?**

- **What is the best single predictor of survival time?**

- **What is the best model following Collett's approach?**

- **What is the best model based on the AIC? Are they the same?**

Survival of Multiple Myeloma Patients: $-2\log\hat{L}$ and AIC Values

Model #	Type of model	Terms in model	$-2\log\hat{L}$	AIC
(1)	Null	(none)	215.940	
(2)	Univariate	Age	215.817	
(3)		Sex	215.906	
(4)		BUN	207.453	
(5)		Ca	215.494	
(6)		HGB	211.068	
(7)		Pcells	215.875	
(8)		Protein	213.890	
(9)	2-variables	HGB + BUN	202.938	
(10)		HGB + Protein	209.829	
(11)		BUN + Protein	203.641	
(12)	3-variables	BUN + HGB + Protein	200.503	
(13)		HGB + BUN + Age	202.669	
(14)		HGB + BUN + Sex	202.553	
(15)		HGB + BUN + Ca	202.937	
(16)		HGB + BUN + Pcells	202.773	

Note: Critical values for χ^2_1 are 3.84 for $p = 0.05$, 2.706 for $p = 0.10$, and 1.643 for $p = 0.20$.

Survival of Multiple Myeloma Patients:

Following Collett's approach (see slide 8):

(1) **Identify candidate predictors:** The variables `BUN`, `HGB`, and `Protein` all lead to reductions in the $-2\log\hat{L}$ of 1.643 or more (comparing models (4), (6), and (8) with model (1)), and thus have $p < 0.20$. So these are our initial candidate predictors.

(2) **Fit multivariate model and reduce:** Now we fit a multivariate model with these three candidate predictors. This is model (12), with $-2\log\hat{L} = 200.503$. We then use backwards elimination to remove unnecessary variables from this multivariate model. At this stage, we use a more stringent cutoff of p<0.10, and so the variable `Protein` is removed (based on comparing model (12) with model (9), $\chi^2_{LR} = 2.435$ with $p = 0.119$). Neither `BUN` nor `HGB` can be removed following this same strategy.

(3) **Consider other predictors not included previously:** Next we examine whether any of the variables which didn't originally get considered in step # 1 (Age, Sex, Ca, and Pcells) should be added back in, given `BUN` and `HGB` are already in the model. This involves comparing models (13)-(16) with model (9). None of these covariates result in a significant reduction in the $-2\log\hat{L}$.

(4) **Final pruning and consideration of interactions:** Since no additional variables were added, we don't need to do any additional pruning of the main effects model. However, it may be of interest to consider interactions (not done here, though).

So Model (9) was best based on Collett's approach, assuming we specifically set the significance level to 0.20 for covariates to be considered based on univariate models, and 0.10 to be retained in final multivariate model.

What is the best model based on minimizing the AIC?

In general, which approach – using Likelihood Ratio tests or minimzing AIC – is most likely to lead to inclusion of more covariates in the final model?

- For a single covariate to be added to a model, it must increase log-likelihood by 3.84 to be statistically significant
- However, to be considered for initial inclusion, it just needs to have $p < 0.20$, which corresponds to change in $-2logL$ of 1.643
- To be included in final model (or added to final model), it needs to have $p < 0.10$, which corresponds to change in $-2logL$ of 2.706.
- In contrast, the AIC would only need to be more than 2 points lower for each 1 df covariate added in order to result in a lower AIC.

5.1.3 LASSO for model selection (see Collett, Section 3.7)

LASSO=*least absolute shrinkage and selection operator*
This "shrinks" the coefficients of covariates toward zero, and sets some exactly equal to zero.

Covariates that are highly correlated with other covariates in the model, or have no association with the outcome, will have their coefficients set to zero.

Variable selection results from the set of covariates that end up with non-zero coefficients.

This shrinkage improves the predictive ability of the model.

The idea of the lasso is to estimate all of the β's in the model by maximizing the partial likelihood subject to constraining the sum of the absolute β's to be less than some value.

The partial likelihood is given by:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right\}^{\delta_i},$$

and the **constrained partial likelihood or penalized partial likelihood is given by**

$$L_\lambda(\beta) = L(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|.$$

The parameter, λ, is the *penalty* or *tuning* parameter. This is because a penalty is assigned to large β's.

Note: the variables should be standardized to be on the same scale.

This introduces some bias to the estimates of β. Why?

Estimates of the coefficients of all covariates are often obtained for a range of values of λ.

A plot of these estimates against λ (i.e., the *lasso trace*) displays the dependence of the estimates on the value of λ.

With small λ, there will not be much variable selection.
With large λ, there may be too much selection.

LASSO trace plot for multiple myeloma data:

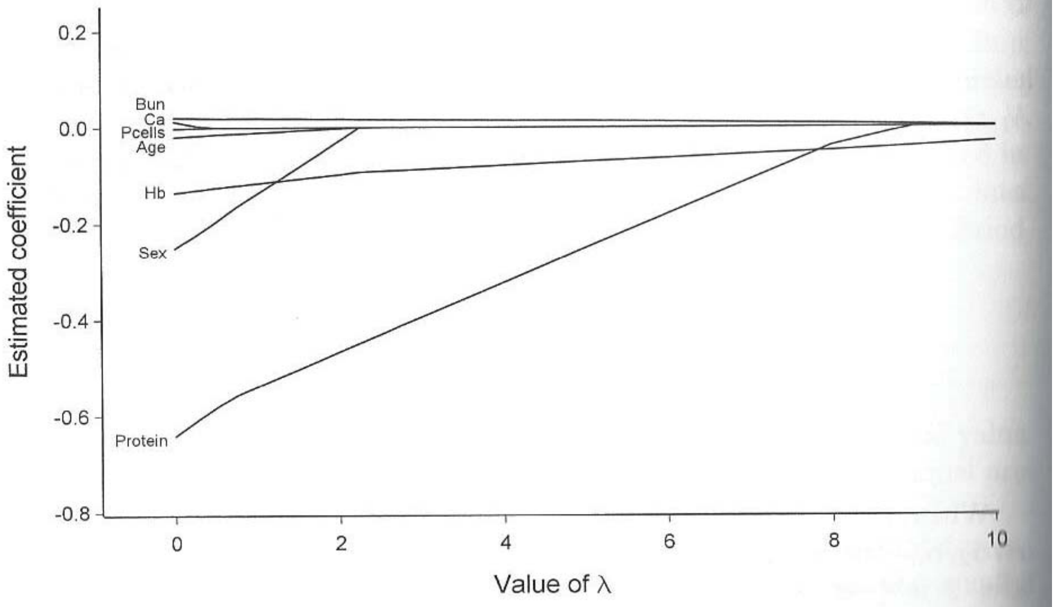


Figure 3.2 *Trace of the estimated coefficients of the explanatory variables as a function of the lasso parameter, λ.*

Cross-validation is usually used to find the optimal value of λ. The cross-validated partial log-likelihood is

$$\log \hat{L}_{CV}(\lambda) = \sum_{i=1}^{n} \left\{ \log L[\hat{\beta}_{(-i)}(\lambda)] - \log L_{(-i)}[\hat{\beta}_{(-i)}(\lambda)] \right\},$$

where $\hat{\beta}_{(-i)}(\lambda)$ is the estimated parameters for a given λ and excluding subject i and $\log L_{(-i)}(\beta)$ is the partial log-likelihood when the data for the ith subject are excluded.

Each term represents the independent contribution of individual i to the log likelihood.

The optimal value of λ is that which maximizes $\log \hat{L}_{CV}(\lambda)$.

Finally, the covariates selected for inclusion in the final model are those with coefficients that are non-zero for this optimal value of λ.

5.2 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 linearity assumptions for any continuous covariate included in your model

- Check proportionality assumption (this will be the topic of the next set of lectures)

- Check residuals!

  (a) generalized (Cox-Snell)

  (b) martingale

  (c) deviance

  (d) Schoenfeld

  (e) weighted Schoenfeld

- **Check other survival regression diagnostics**

5.2.1 Checking Linearity Assumptions

Whenever we put a covariate *X* in that is "continuous", we are making a strong assumption regarding the relationship with the log-hazard:

$$\log(HR) = \beta_1 X$$

This model assumes that there is the same change in $\log(HR)$ for each one unit change in X.

For example, if $X = BMI$ as for the NEJM Obesity article, then we are assuming there is the same change in the log HR for comparing someone with a BMI of 16 vs 15 as there is for comparing 25 vs 24.

Approaches for checking linearity:

- **Consider transformations: for some very highly skewed variables, or ones constrained to be positive (income), it may be appropriate to log-transform variables prior to modeling.**
- **Check linearity by adding higher order terms: Add quadratic (X^2) or both quadratic and cubic terms X^3) to the model and examine the changes in $-2\log\hat{L}$.**
- **Create A ordinal categories from the continuous measure: For the example of BMI, ordinal categories were created as $18.5-24.9$, 25-29.9, and 30 or higher.**
  - **Create four to five categories, depending on sample size and any external clinical relevance.**
  - **Each level should have approximately equal numbers of observations. For example, quartiles could be used for exposure measures.**
  - **Fit the Cox PH model with the categorical variable either using dummy variables to indicate the $(a-1)$ levels or using a `class` variable.**
  - **Descriptively evaluate the linearity of estimates (*on the log HR scale*!) across the categories; a plot may be useful.**
  - **A test for linearity can be conducted by comparing the model with continuous X to one with the categorical variables, based on comparing the $-2\log(\hat{L})$ or the AIC.**

Checking Linearity Assumptions: Multiple Myeloma Example

To evaluate whether the association between HGB and survival follows a linear relationship, we can categorize hemoglobin as:

(1) ≤ 7

(2) 7-10

(3) 10-13

(4) $>$ 13

Fitting the model (9) including BUN but substituting the categorical variable above for HGB, we obtain $-2\log(\hat{L}) = 200.417$.

Comparing to Model (4) with just BUN, the change is 7.036 (with 3df), so p=0.071. This suggests keeping the categorical HGB in the model (although not as strongly as for a continuous one).

If we instead fit the model with the ordinal predictor (1,2,3,4) reflecting the four groups above, we obtain $-2\log(\hat{L}) = 203.89$. So comparing the model with ordinal predictor to 4 separate categories yields a change of 203.891-200.417=3.474 (with 2 df), p=0.18.

We can conclude that there is not substantial evidence of non-linearity for the effect of `HGB` on survival in multiple myeloma patients.

Ordinal model is nested within categorical model; if it is not rejected then conclude linearity.

Consider covariate X, which takes on values $\{1, 2, 3\}$.

Let $I_1 = 1$ if $X = 1$, $I_1 = 0$ otherwise
Let $I_2 = 1$ if $X = 2$, $I_2 = 0$ otherwise
Let $I_3 = 1$ if $X = 3$, $I_3 = 0$ otherwise.

Note that $X = I_1 + 2I_2 + 3I_3$ and that

$$X^2 = I_1 + 4I_2 + 9I_3$$
$$X^3 = I_1 + 8I_2 + 27I_3.$$

The model with categorical X has terms: $\beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3$, which is equivalent to the model that has terms X, X^2 and X^3: $\alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3$, which equals

$$(\alpha_1 + \alpha_2 + \alpha_3)I_1 + (2\alpha_1 + 4\alpha_2 + 8\alpha_3)I_1 + (3\alpha_1 + 9\alpha_2 + 27\alpha_3)I_3.$$

Obesity and Heart Disease, NEJM 2002, portion of Table 3 (separately by sex)

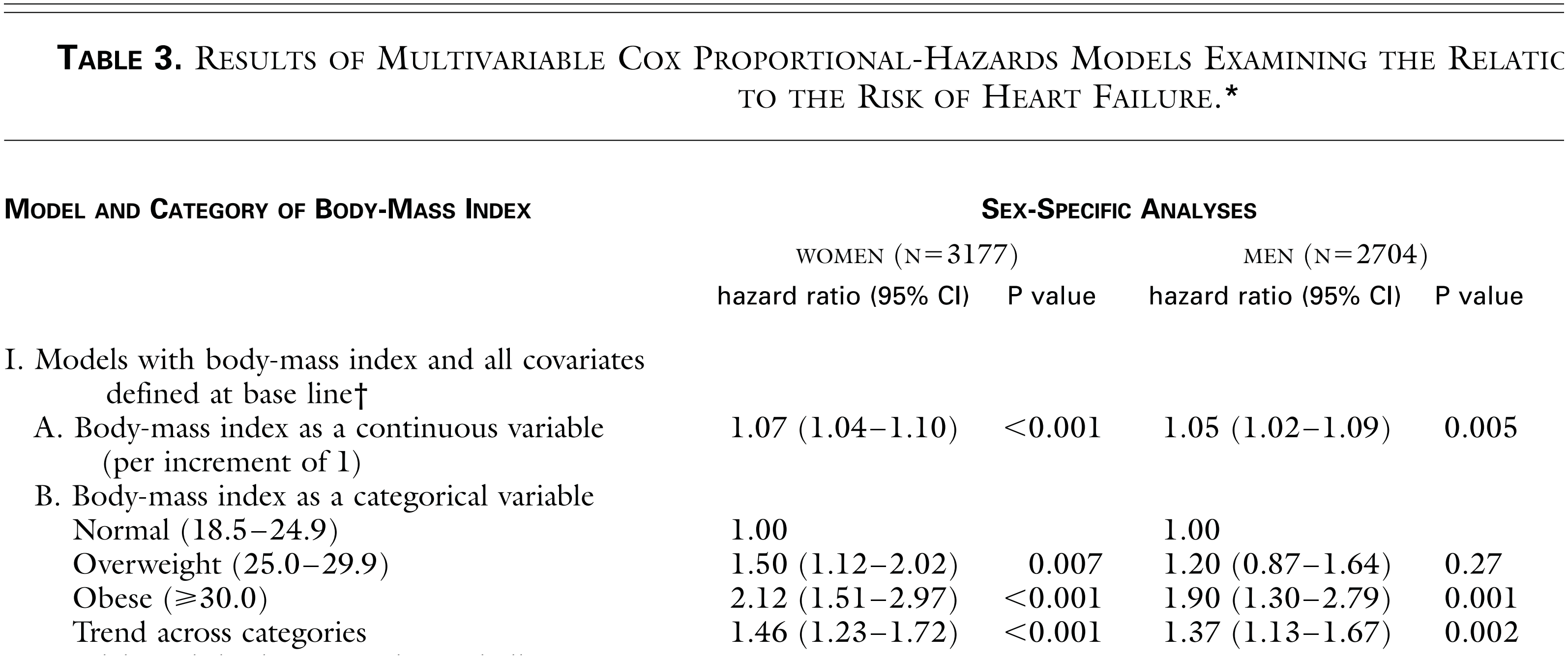


TABLE 3. RESULTS OF MULTIVARIABLE COX PROPORTIONAL-HAZARDS MODELS EXAMINING THE RELATIC TO THE RISK OF HEART FAILURE.*

MODEL AND CATEGORY OF BODY-MASS INDEX	SEX-SPECIFIC ANALYSES			
	WOMEN (N=3177)		MEN (N=2704)	
	hazard ratio (95% CI)	P value	hazard ratio (95% CI)	P value
I. Models with body-mass index and all covariates defined at base line†				
A. Body-mass index as a continuous variable (per increment of 1)	1.07 (1.04–1.10)	<0.001	1.05 (1.02–1.09)	0.005
B. Body-mass index as a categorical variable				
Normal (18.5–24.9)	1.00		1.00	
Overweight (25.0–29.9)	1.50 (1.12–2.02)	0.007	1.20 (0.87–1.64)	0.27
Obese (≥30.0)	2.12 (1.51–2.97)	<0.001	1.90 (1.30–2.79)	0.001
Trend across categories	1.46 (1.23–1.72)	<0.001	1.37 (1.13–1.67)	0.002

Does the assumption of linearity look to be satisfied for the Obesity and Heart Disease Cox models?

For females, $\beta_{OWT} = 0.406$ and $\beta_{OB} = 0.751$.

For males, $\beta_{OWT} = 0.182$ and $\beta_{OB} = 0.642$.

5.2.2 Residuals for Survival Data

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:

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

5.2.3 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 $\boldsymbol{Z}_i$, we can get

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

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

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

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.

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?**

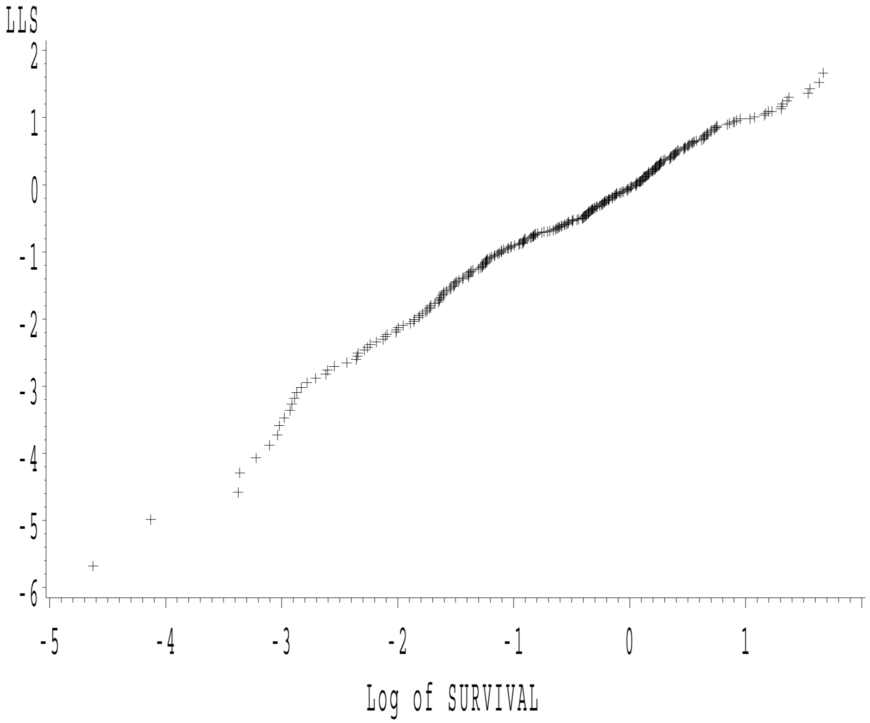
(Note: this does not necessarily mean that the underlying distribution of the original survival times is exponential!)

Obtaining the generalized residuals from Stata

- **Fit a Cox PH model with the `stcox` command, along with 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."

Obtaining the generalized residuals from SAS

The generalized residuals can be obtained from SAS after fitting a PH model using the output statement with the logsurv option.

```
proc phreg data=fish;
  model survtime*censor(0) = towdur handling logcatch length;
  output out=phres logsurv=genres;

*** take negative log Pr(survival) at each persons survtime;
data phres;
  set phres;
  genres=-genres;

*** Now we treat the generalized residuals as the input dataset;
*** to evaluate whether the assumption of an exponential;
*** distribution is appropriate;
proc lifetest data=phres outsurv=survres;
  time genres*censor(0);

data survres;
  set survres;
  lls=log(-log(survival));
  loggenr=log(genres);

proc gplot data=survres;
  plot lls*loggenr;
run;
```

```
library(survival)

cph1 <- coxph(Surv(futime, fustat)~rx+age , data=ovarian)

residual <- residuals(cph1, type="")
```

5.2.4 Martingale Residuals

(see Fleming and Harrington, p.164)

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

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

estimated cumulative hazard for individua

Properties:

- **r_i's have mean 0**
- **range of r_i's is between $-\infty$ and 1**
- **approximately uncorrelated (in large 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., βZ_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
```

The martingale residuals can be obtained from SAS after fitting a PH model using the output statement with the resmart option.

Once you have them, you can

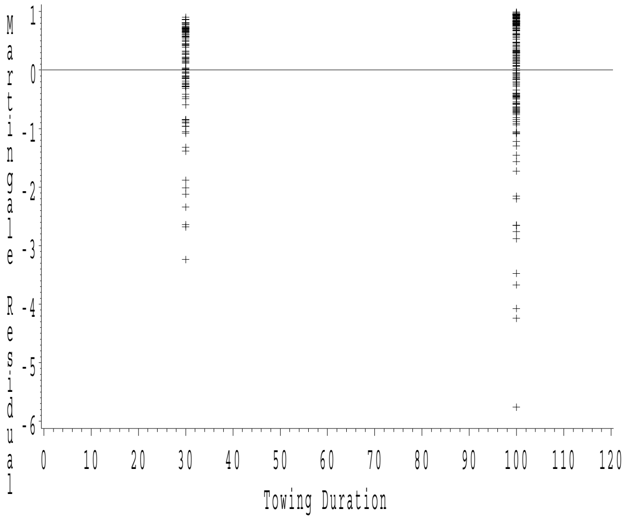
- **plot against predicted values**
- **plot against covariates**

```
proc phreg data=fish;
  model survtime*censor(0) = towdur handling logcatch length;
  output out=phres resmart=mres xbeta=xb;

proc gplot data=phres;
  plot mres*xb;              /* predicted values */
  plot mres*towdur;
  plot mres*handling;
  plot mres*logcatch;
  plot mres*length;
run;
```

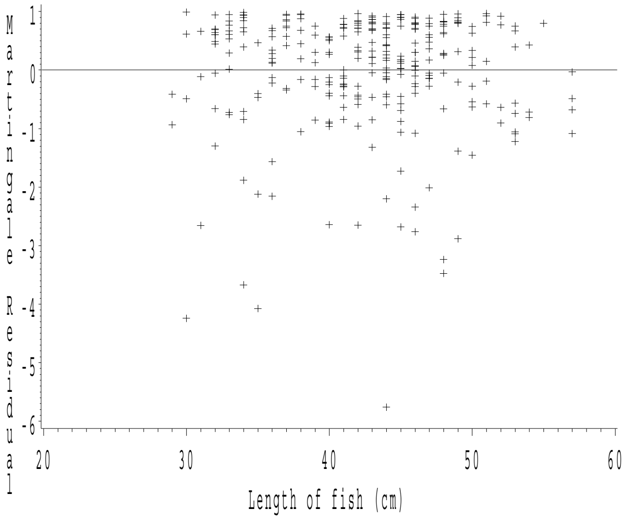
(Allison says “For most purposes, you can ignore the Cox-Snell and martingale residuals.”)

Martingale Residuals

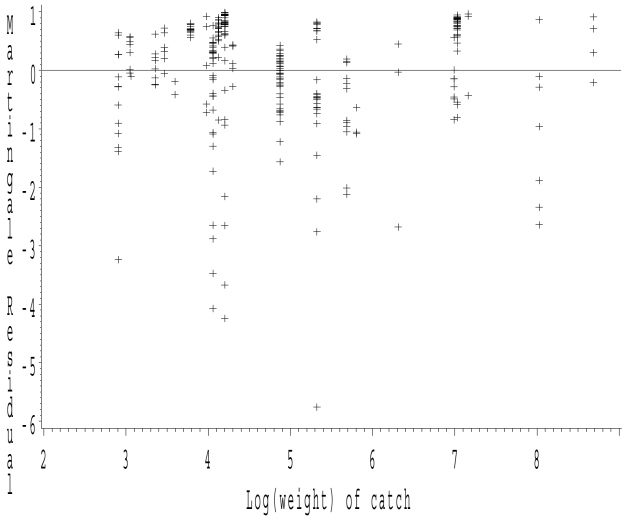
Martingale residuals vs towing duration



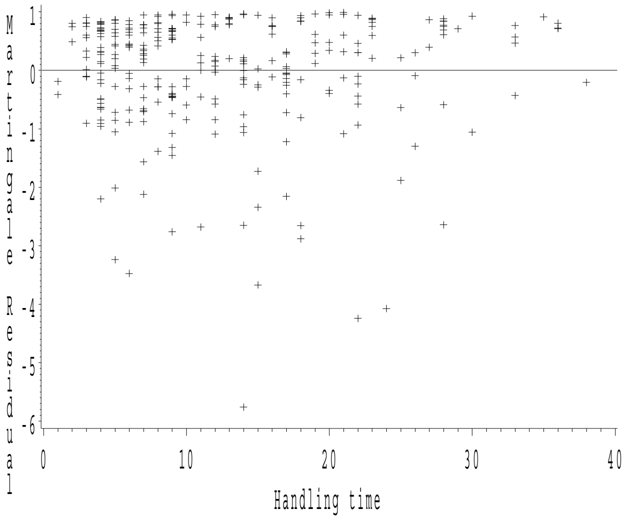
Martingale residuals vs length of fish



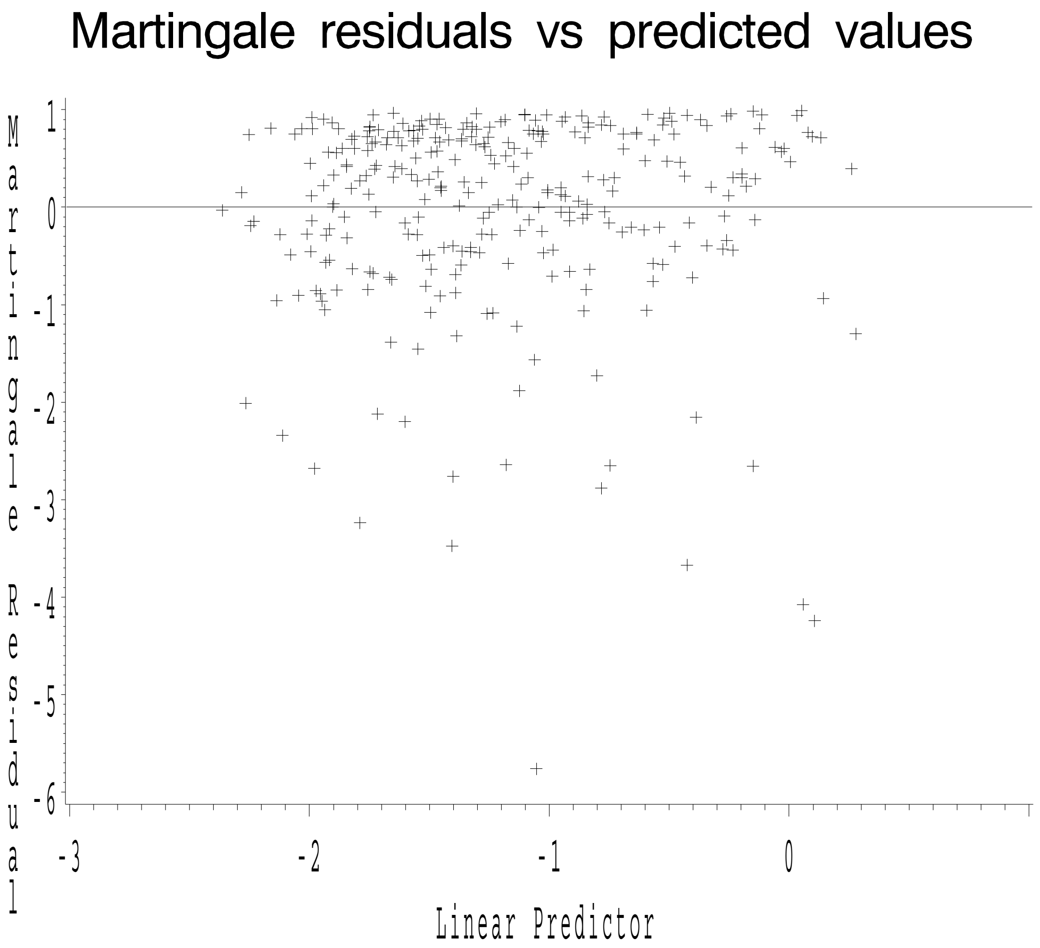
Martingale residuals vs log(catch)



Martingale residuals vs handling



Martingale Residuals versus predicted values



5.2.5 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):

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

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

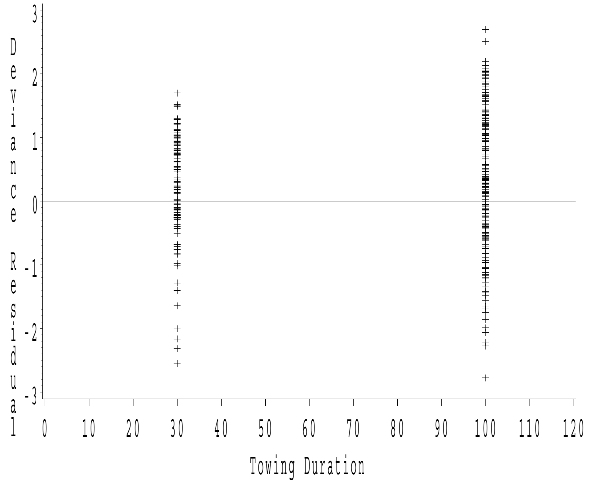
```
. stcox towdur handling length logcatch, mgale(mg)

. predict devres, deviance
```

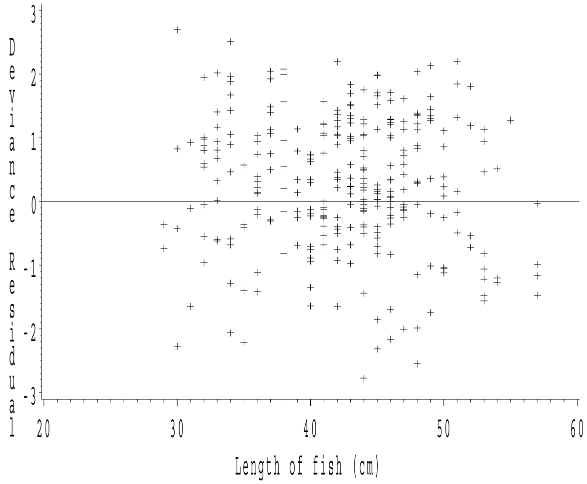
and then they can be plotted versus the predicted log(HR) or the individual covariates, as shown for the Martingale residuals.

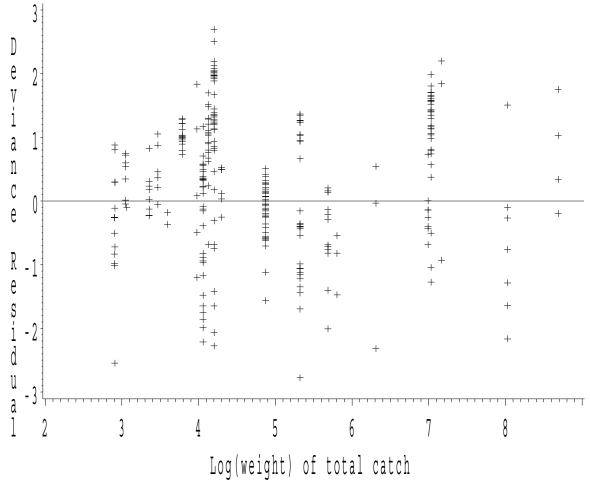
In SAS, just use resdev option instead of resmart.

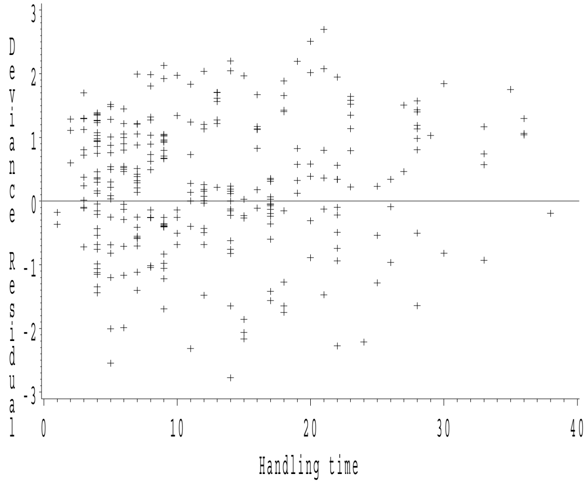
Deviance residuals behave much like residuals from OLS regression (i.e., mean=0, s.d.=1). They are negative for observations with survival times that are smaller than expected.

Deviance Residuals

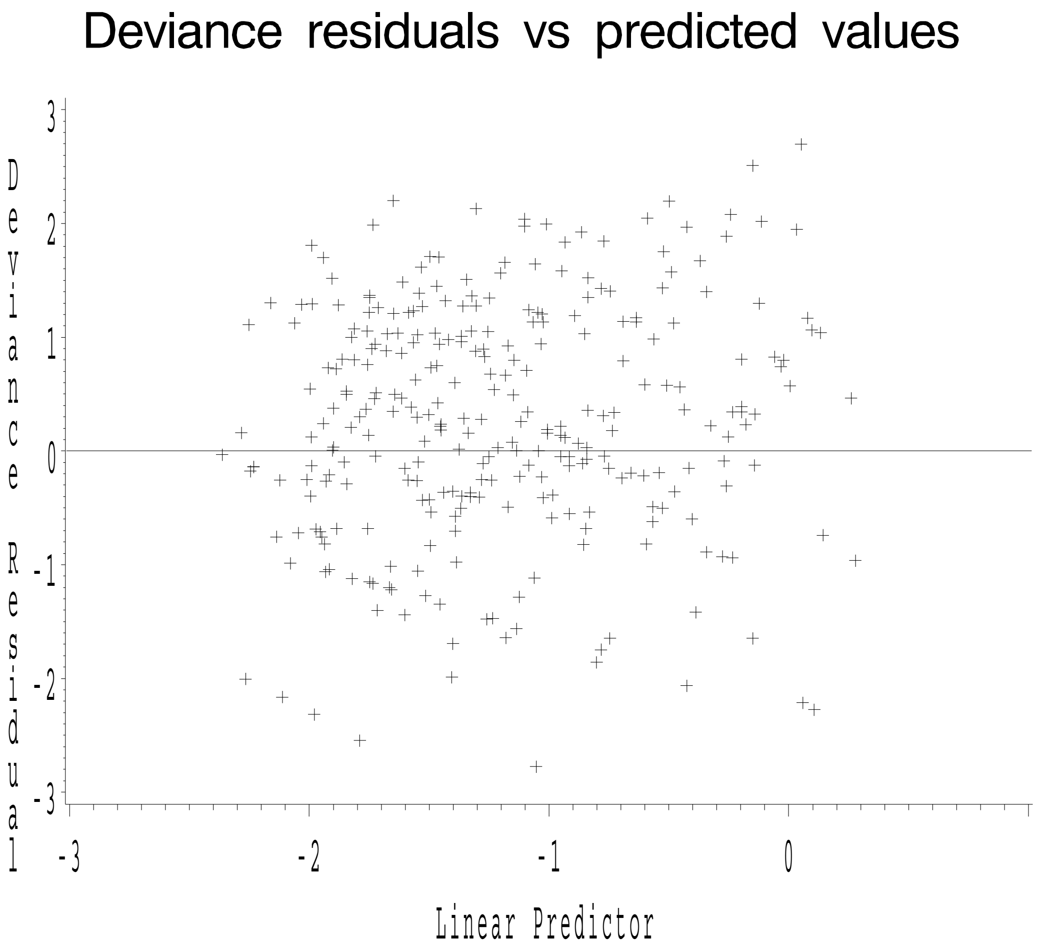

Deviance residuals vs towing duration


Deviance residuals vs length of fish


Deviance residuals vs log(catch)


Deviance residuals vs handling

Deviance Residuals vs predicted values



5.2.6 Schoenfeld Residuals

These are defined at each observed failure time as:

$$r^s_{ij} = Z_{ij}(t_i) - \bar{Z}_j(t_i)$$

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 large samples)

Software for Schoenfeld residuals

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
```

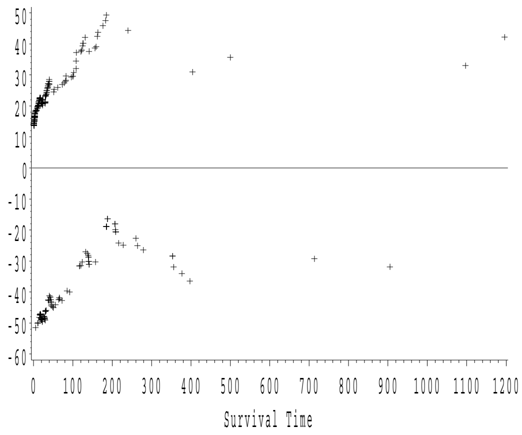
In SAS, add to the output line

```
RESSCH=name1 name2 ... namek
```

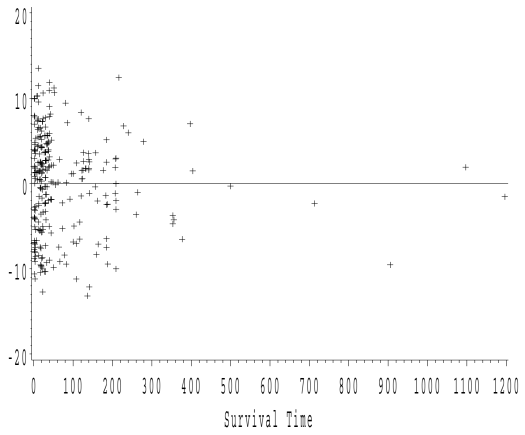
for up to k regressors in the model.

Schoenfeld Residuals

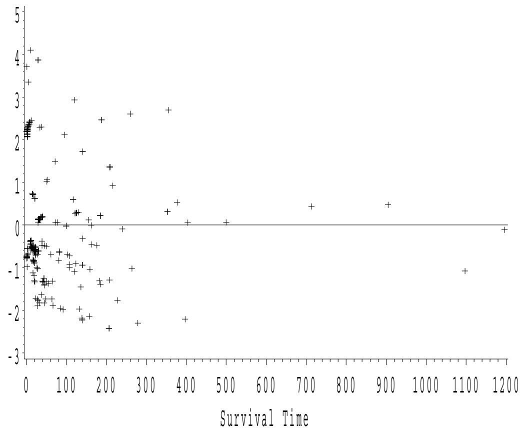
Schoenfeld resids for towing vs survival time



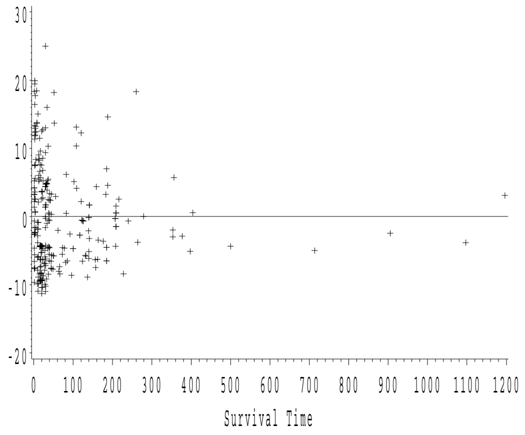
Schoenfeld resids for length vs survival time



Schoenfeld resids for log(catch) vs survival time



Schoenfeld resids for handling vs survival time



5.2.7 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:

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

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.

Software for Weighted Schoenfeld residuals

In Stata, use the command:

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

. graph logres2 survtime
```

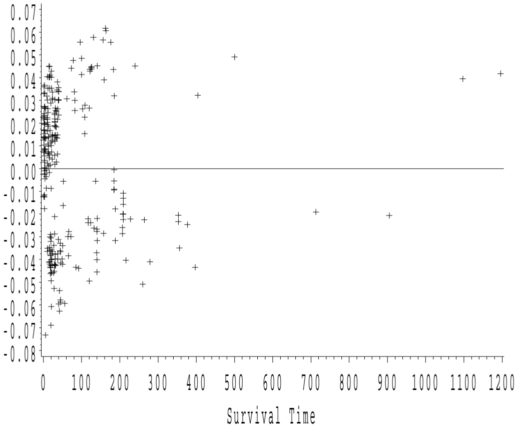
In SAS, add to the output line

```
WTRESSCH=name1 name2 ... namek
```

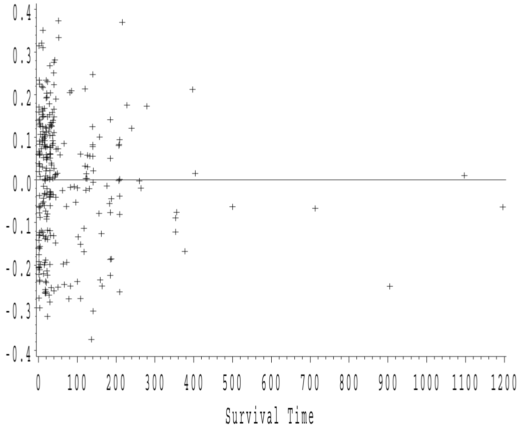
for up to k regressors in the model.

Weighted Schoenfeld Residuals

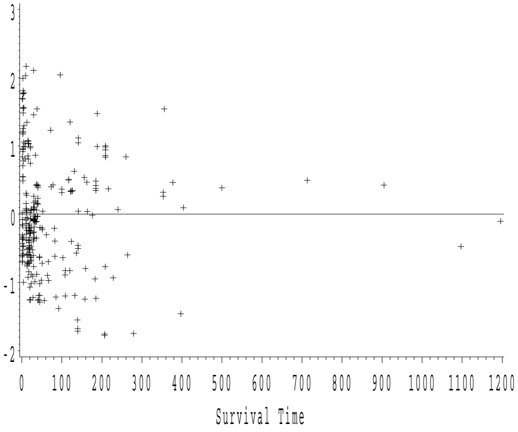
Weighted Schoenfeld resids for towing vs time



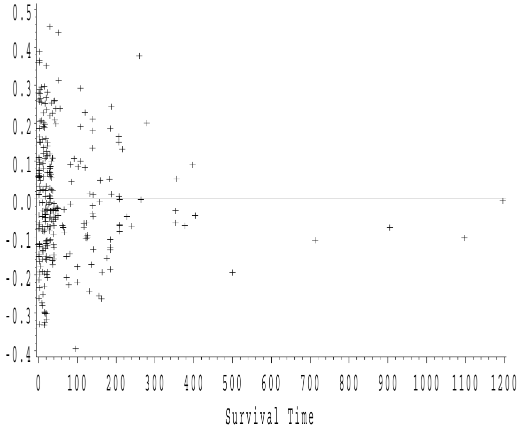
Schoenfeld resids for length vs time



Schoenfeld resids for log(catch) vs time



Schoenfeld resids for handling vs time



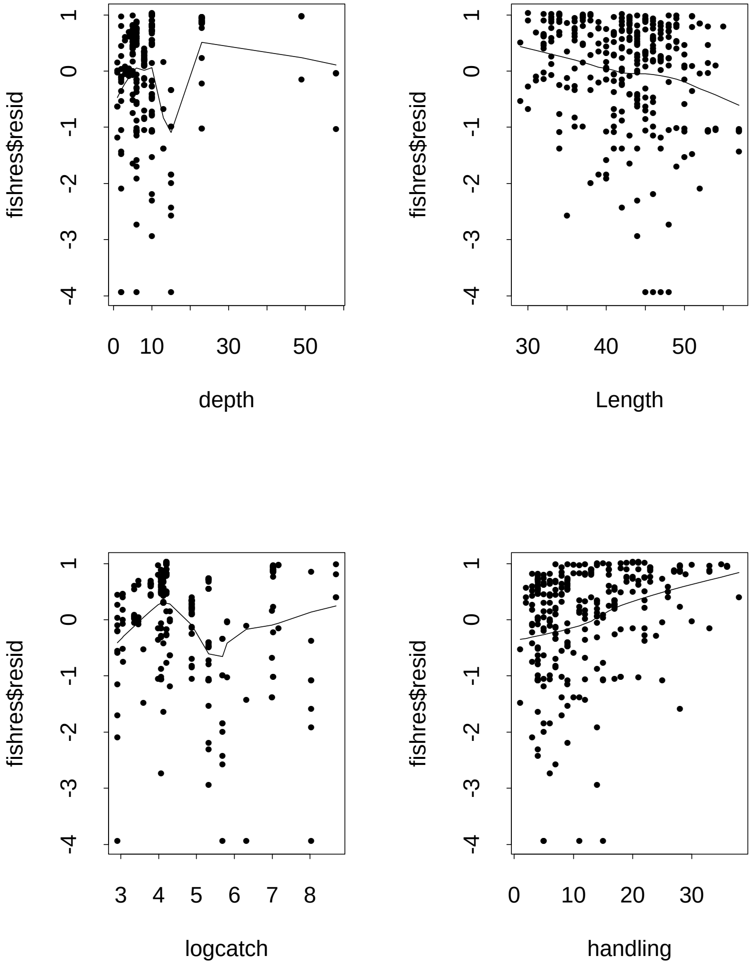
Using Residual plots to explore relationships

If you calculate martingale or deviance residuals without any covariates in the model and then plot against covariates, you obtain a graphical impression of the relationship between the covariate and the hazard.

In Splus, it is easy to do this (also possible in stata using the "estimate" option)

```
** read in the dataset and fit a cox PH model
fish_read.table('fish.data',header=T)
x_fish$towdur
fishres_coxreg(fish$time, fish$censor, x, resid="martingale",iter.max=0)

** the 2 commands below set up the postscript file, with 4 graphs
postscript("fishres.plt",horizontal=F,height=10,width=7)
par(mfrow=c(2,2),oma=c(0,0,2,0))

** plot the martingale residuals vs each of the other covariates
** and add a lowess smoothed fit to the plot
plot(fish$depth, fishres$resid, xlab="depth")
lines(lowess(fish$depth,fishres$resid,iter=0))

plot(fish$length,  fishres$resid, xlab="length")
lines(lowess(fish$length,fishres$resid,iter=0))

plot(fish$handling,  fishres$resid,  xlab="handling")
lines(lowess(fish$handling,fishres$resid,iter=0))

plot(fish$logcatch,  fishres$resid,  xlab="logcatch")
lines(lowess(fish$logcatch,fishres$resid,iter=0))
```

Splus Plots of Martingale Residuals for Cox Model containing only towing duration as a predictor, vs other covariates



5.2.8 Deletion Diagnostics

Deletion diagnostics are defined generally as:

$$\delta_i = \hat{\beta} - \hat{\beta}_{(i)}$$

In other words, they are the difference between the estimated regression coefficient using all observations and that without the i-th individual. This can be useful for assessing the influence of an individual.

In SAS PROC PHREG, we use the dfbeta option:
(Note that there is a separate dfbeta calculated for each of the predictors.)

```
proc phreg data=fish;
  model survtime*censor(0)=towdur handling logcatch length;
  id id;
  output out=phinfl dfbeta=dtow dhand dlogc dlength ld=lrchange;

proc univariate data=phinfl;
  var dtow dhand dlogc dlength lrchange;
  id id;
run;
```

The proc univariate procedure will supply the 5 smallest values and the 5 largest values. The "`id`" statement means that these will be labeled with the value of id from the dataset.

5.2.9 Other Influence Diagnostics

The LD option is another method for checking influence. It calculates how much the log-likelihood (x2) would change if the i-th person was removed from the sample.

$$LD_i \;=\; 2\left[logL(\widehat{\boldsymbol{\beta}}) - logL(\widehat{\boldsymbol{\beta}}_{-i})\right]$$

$\widehat{\beta}$ = MLE for all parameters with everyone included
$\widehat{\beta}_{-i}$ = MLE with i-th subject omitted

Again, the proc univariate procedure in SAS will identify the observations with the largest and smallest values of the `lrchange` diagnostic measure.

Can we improve the model?

The plots appear to have some structure, which indicate that we could be leaving something out. It is always a good idea to check for interactions:

In this case, there are several important interactions. I used a backward selection model forcing all main effects to be included, and considering all pairwise interactions. Here are the results:

```
                Parameter    Standard      Wald         Pr >       Hazard
Variable  DF     Estimate       Error   Chi-Square   Chi-Square    Ratio

TOWDUR     1    -0.075452     0.01740    18.79679       0.0001     0.927
DEPTH      1     0.123293     0.06400     3.71107       0.0541     1.131
LENGTH     1    -0.077300     0.02551     9.18225       0.0024     0.926
HANDLING   1     0.004798     0.03221     0.02219       0.8816     1.005
LOGCATCH   1    -0.225158     0.07156     9.89924       0.0017     0.798

TOWDEPTH   1     0.002931   0.0004996    34.40781       0.0001     1.003
TOWLNGTH   1     0.001180   0.0003541    11.10036       0.0009     1.001
TOWHAND    1     0.001107   0.0003558     9.67706       0.0019     1.001
DEPLNGTH   1    -0.006034     0.00136    19.77360       0.0001     0.994
DEPHAND    1    -0.004104     0.00118    12.00517       0.0005     0.996
```

Interpretation:

Handling alone doesn’t seem to affect survival, unless it is combined with a longer towing duration or shallower trawling depths.

An alternative modeling strategy when we have fewer covariates

With a dataset with only 5 main effects, it would make sense to consider interactions from the start. How many would there be?

- **Fit model with all main effects and pairwise interactions**
- **Then use backward selection to eliminate non-significant pairwise interactions (remember to force the main effects into the model at this stage)**
- **Once non-significant pairwise interactions have been eliminated, you could consider backwards selection to eliminate any non-significant main effects that are not involved in remaining interaction terms**
- **After obtaining final model, use residuals to check fit of model.**

Contents