Session 5: Model selection

1. Univariate analyses of the halibut data set

Now we input the data set `halibut.dat`. This is done through the following SAS statements:

```
data halibut;
     infile 'halibut.dat';
     input id   survtime   censor   towdur   depth   length   handling   logcatch;
     label survtime='Length of survival'
           depth='Depth'
           handling='Handling time'
           towdur='Duration of towing'
           length='Length of fish'
           logcatch='Logarithm of the total catch';
run;
```

We would like first to produce univariate PH regressions. This is straightforward with PHREG, but, to generate a graph, we will discretize each continuous factor. One such straightforward discretization is to dichotomize the factor as below or above the median. In turn, to obtain the medians we use PROC UNIVARIATE as follows:

```
proc univariate data=halibut;
     var towdur length depth handling logcatch;
     title 'Descriptive statistics for main explanatory variables';
run;
```

From this we see that the median of `towdur` is 100. We can incorporate this directly into the statement of the PROC PHREG as follows:

```
proc phreg data=halibut noprint;
     model survtime*censor(0)=disctd;
     disctd=(towdur<100);
     id towdur;
     title 'Survival time by discretized tow duration (above vs. below median)';
     output out=outsurv survival=predsurv;
run;
```

The output from the previous statements is suppressed with the `noprint` statement (which allows us to produce the output data set `outsurv` without getting lengthy output.

The problem with this data set is that the variable towdur (since it is not part of the model) is not included in outsurv. To insert it there we use the id statement, i..e,

```
      id towdur;
```

Let's see what this data set looks like:

```
proc print data=outsurv;
     title 'Predicted survival with respect to tow duration';
```

```
run;
```

```
              Predicted survival with respect to tow duration
49

                      Obs    towdur    survtime    censor    disctd    predsurv

                        1        30       209.0         1         1     0.16346
                        2        30       209.0         1         1     0.16346
                        3        30       209.0         1         1     0.16346
                        4        30       209.0         1         1     0.16346
                        5        30        38.0         1         1     0.56354
                        6        30       209.0         1         1     0.16346
                        7        30       140.9         1         1     0.30191
                        8        30       140.9         1         1     0.30191
                        9        30       140.1         1         1     0.31248
                       10        30       208.0         1         1     0.19483
                       11        30       140.1         1         1     0.31248
                        .         .           .         .         .           .
                        .         .           .         .         .           .
                        .         .           .         .         .           .
```

Now to produce a graph we need to format the new variable `disctd`.

```
proc format;
     value bivarfmt 1='Below median' 0='Above median';
run;
```

We update the oursurv data set as follows:

```
proc format;
     value bivarfmt 1='Below median' 0='Above median';
run;

data outsurv;
      set outsurv;
      format disctd bivarfmt.;
run;
```

Now let's produce the graph. First we must define the symbols and lines used in the graph. We have two groups, so we define two symbols as follows:

```
symbol1 c=red   line=1 i=stepljs value=plus;
symbol2 c=green line=1 i=stepljs value=plus;
```

The option `c=red` and `c=green` determine that the first plot (for `disctd==0`) will be in red color and the second (for `disctd==1`) in green.

The next option `line=1` specifies that both lines will be solid. Broken lines of varying widths can be specified by increasing the number after the "=" sign. In both cases, the symbol itself will be a "+" (plus) sign, so `value=plus`.

Because we would like to generate a plot that will look like a Kaplan-Meier plot (even though the survival estimates were derived from a Cox model) we specify that the points between the lines must be interpolated as follows

```
i=stepljs
```

Now we must define the axes (the default axes in SAS are rather unattractive)

```
axis1 label=(angle=90 height=2.0 font='arial' 'Percent surviving' )
      value=(font='arial' height=1.5);
axis2 label=(height=2.0 font='arial' 'Length of survival')
      value=(font='arial' height=1.5) minor=NONE;
```

The label of the first axis (later to be defined as the *y* axis) has the text `'Percent surviving'` which is rotated by 90^{o} (`angle=90`) has size 2.0 (`height=2.0`) which you need to play around with since the SAS units of size are not obvious, and the PC font is arial (`font='arial'`). This means that this program code might not produce the expected results in a platform that does not have this font (e.g., in UNIX). You need to limit characteristics that make your code not portable as much as possible. We also can specify pretty much every aspect of the axis. Here we choose to make the markers a bit larger by specifying `value=(font='arial' height=1.5)`. Note the syntax that, for every attribute, has all the characteristics in a parenthesis that follows an equal sign.

The second axis, `axis2`, (later to be defined as the *x* axis) is similar, except that we have removed any minor tick marks by specifying `minor=NONE`.
The graph is generated by PROC GPLOT as follows:

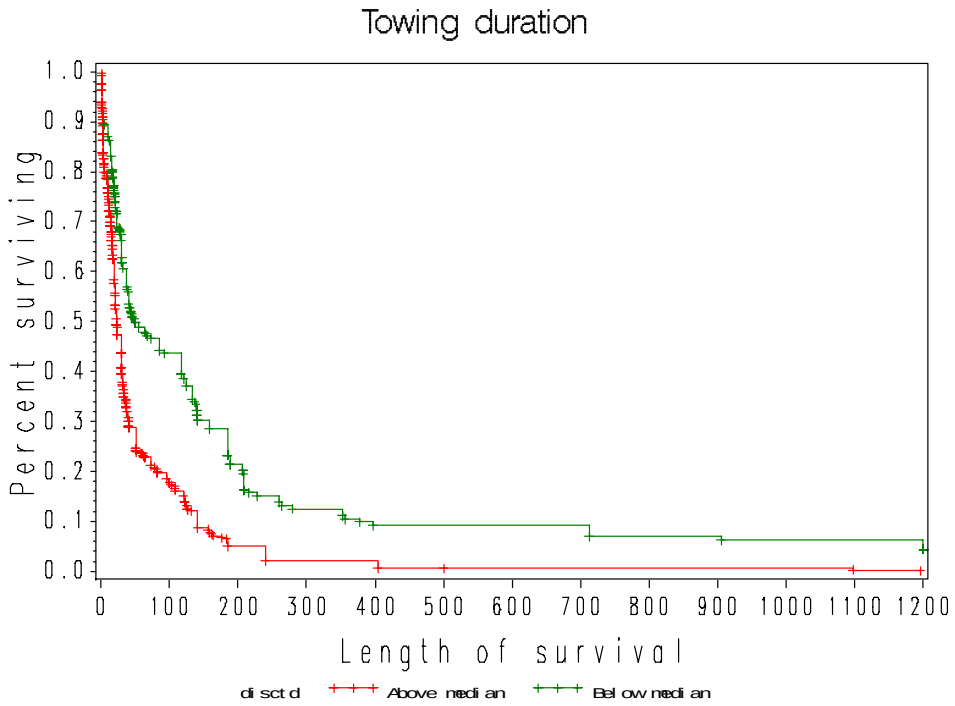
```
proc gplot data=outsurv;
     plot predsurv*survtime=disctd/overlay vaxis=axis1 haxis=axis2;
     title 'Towing duration';
run;
```

The syntax of the procedure itself is familiar. The plot is generated with the statement

```
plot predsurv*survtime=disctd/vaxis=axis1
                              haxis=axis2;
```

The syntax of the plot statement is `plot yvar*xvar` followed by options. You can also add a categorical variable that generates as many plots as there are categories in it. Here we have included variable `disctd` (the discretized tow duration variable). The only option we have made define which axis is the

vertical and which is the horizontal (`vaxis=axis1` and `haxis=axis2` repectively). The output is as follows:



This plot suggests that longer tow duration results in shorter survival times among the halibut fish. Various plots with respect to the other continuous variables, analyzed in a univariate manner are similarly obtained.

2. Model selection

Now we describe how model selection can be automatically accomplished with PROC PHREG. We will use the stepwise method as an example understanding that the forward and backgward selection methods are similar.

To carry out a model selection procedure in the halibut data set we proceed as follows:

```
proc phreg data=halibut;
     model survtime*censor(0)=towdur depth length handling
                              logcatch/selection=stepwise selentry=0.2
                                                          selstay=0.1;
     title 'Model selection of the halibut data set';
run;
```

This statement specifies that `selection=stepwise` and that the probability threshold for entering a variable is `slentry=0.2` and that of removing one is `slstay=0.1`.

The output is as follows:

Model selection of the halibut data set

The PHREG Procedure

Model Information

Data Set	WORK.HALIBUT	
Dependent Variable	survtime	Length of survival
Censoring Variable	censor	
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
294	273	21	7.14

Step 1. Variable handling is entered. The model contains the following explanatory variables:

handling

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2558.358
AIC	2599.449	2560.358
SBC	2599.449	2563.967

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.0914	1	<.0001
Score	47.1417	1	<.0001
Wald	46.0330	1	<.0001

Step 2. Variable logcatch is entered. The model contains the following explanatory variables:

handling logcatch

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2539.647
AIC	2599.449	2543.647
SBC	2599.449	2550.866

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.8023	2	<.0001
Score	65.6797	2	<.0001
Wald	63.0055	2	<.0001

Model selection of the halibut data set

The PHREG Procedure

Step 3. Variable towdur is entered. The model contains the following explanatory variables:

towdur handling logcatch

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2528.599
AIC	2599.449	2534.599
SBC	2599.449	2545.427

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	70.8507	3	<.0001
Score	76.3454	3	<.0001
Wald	72.7407	3	<.0001

Step 4. Variable length is entered. The model contains the following explanatory variables:

towdur length handling logcatch

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2515.310
AIC	2599.449	2523.310
SBC	2599.449	2537.748

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	84.1397	4	<.0001
Score	94.0062	4	<.0001
Wald	90.2476	4	<.0001

Step 5. Variable depth is entered. The model contains the following explanatory variables:

towdur depth length handling logcatch

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

The PHREG Procedure

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2513.769
AIC	2599.449	2523.769
SBC	2599.449	2541.817

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	85.6799	5	<.0001
Score	96.1287	5	<.0001
Wald	92.0770	5	<.0001

Step 6. Variable depth is removed. The model contains the following explanatory variables:

towdur length handling logcatch

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2599.449	2515.310
AIC	2599.449	2523.310
SBC	2599.449	2537.748

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	84.1397	4	<.0001
Score	94.0062	4	<.0001
Wald	90.2476	4	<.0001

NOTE: Model building terminates because the variable to be entered is the variable that was removed
in the last step.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
towdur	1	0.00774	0.00202	14.6800	0.0001	1.008	Duration of towing
length	1	-0.03665	0.01003	13.3466	0.0003	0.964	Length of fish
handling	1	0.05490	0.00988	30.8735	<.0001	1.056	Handling time
logcatch	1	-0.18466	0.05101	13.1017	0.0003	0.831	Logarithm of the otal catch

Summary of Stepwise Selection

Step	Variable Entered	Number Removed	In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	handling		1	47.1417	.	<.0001	Handling time
2	logcatch		2	18.4259	.	<.0001	Logarithm of the total catch
3	towdur		3	11.0191	.	0.0009	Duration of towing
4	length		4	13.4222	.	0.0002	Length of fish
5	depth		5	1.6661	.	0.1968	Depth
6		depth	4	.	1.6506	0.1989	Depth