Model Selection in Cox regression

Suppose we have a possibly censored survival outcome that we want to model as a function of a (possibly large) set of covariates. How do we decide which covariates to use?

An illustration example:

**Survival of Atlantic Halibut - Smith et al.**

<table>
<thead>
<tr>
<th>Obs #</th>
<th>Survival Time (min)</th>
<th>Censoring Indicator</th>
<th>Tow Duration (min.)</th>
<th>Diff in Depth (cm)</th>
<th>Length of Fish (min.)</th>
<th>Handling Time (min.)</th>
<th>Total log(catch) ln(weight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>353.0</td>
<td>1</td>
<td>30</td>
<td>15</td>
<td>39</td>
<td>5</td>
<td>5.685</td>
</tr>
<tr>
<td>109</td>
<td>111.0</td>
<td>1</td>
<td>100</td>
<td>5</td>
<td>44</td>
<td>29</td>
<td>8.690</td>
</tr>
<tr>
<td>113</td>
<td>64.0</td>
<td>0</td>
<td>100</td>
<td>10</td>
<td>53</td>
<td>4</td>
<td>5.323</td>
</tr>
<tr>
<td>116</td>
<td>500.0</td>
<td>1</td>
<td>100</td>
<td>10</td>
<td>44</td>
<td>4</td>
<td>5.323</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Process of Model Selection

There are various approaches to model selection. In practice, model selection proceeds through a combination of

- knowledge of the science
- trial and error, common sense
- automated (?) variable selection procedures
  - forward selection
  - backward selection
  - stepwise selection
- measures of explained variation
- information criteria, eg. AIC, BIC, etc.
- dimension reduction from high-d

Many advocate the approach of first doing a univariate analysis to “screen” out those variables that are likely noise.
(1) **Stepwise (including forward, backward) procedures** have been very widely used in practice, and active research is still being carried out on post selection inference, etc. They are available as automated procedures in Stata and SAS, but currently there does not appear to be a reasonable one in R (the one based on AIC gives strange results, and is not recommended).

We briefly describe the stepwise (back-and-forth) procedure there:

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.
The following is results of **Forward Selection** in Stata, using $p$-value < 0.05 as entry criterion.

```
begin with empty model

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>= 0.0000 &lt; 0.0500</td>
<td>adding handling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>= 0.0000 &lt; 0.0500</td>
<td>adding logcatch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>= 0.0010 &lt; 0.0500</td>
<td>adding towdur</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>= 0.0003 &lt; 0.0500</td>
<td>adding length</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 | .035341 | .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 |
```
The following is results of Backward Selection in Stata, using $p$-value $\geq 0.05$ as removal criterion.

begin with full model

$p = 0.1991 \geq 0.0500$ 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 |
The following is results of Stepwise Selection in Stata, using $p$-value < 0.05 as entry criterion, and $p$-value $\geq$ 0.10 as removal criterion.

begin with full model

$p = 0.1991 >= 0.1000$ removing depth

Cox Regression -- entry time 0

|                | 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|

Number of obs = 294
chi2(4) = 84.14
Prob > chi2 = 0.0000
Log Likelihood = -1257.6548

Pseudo R2 = 0.0324
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.

- Sometimes we want to force certain variables in the models during the whole selection process, even if they may not be significant (lockterm option in Stata and the include option in SAS).

- Depending on the software, different tests (Wald, score, or likelihood ratio) may be used to decide what variables to add and what variables to remove.
Interactions

It is always a good idea to check for interactions:

In this example, there are several important interactions. Here backward selection was used, while forcing all main effects to be included, and considering all pairwise interactions. Here are the results:

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOWDUR</td>
<td>1</td>
<td>-0.075452</td>
<td>0.01740</td>
<td>18.79679</td>
<td>0.0001</td>
<td>0.927</td>
</tr>
<tr>
<td>DEPTH</td>
<td>1</td>
<td>0.123293</td>
<td>0.06400</td>
<td>3.71107</td>
<td>0.0541</td>
<td>1.131</td>
</tr>
<tr>
<td>LENGTH</td>
<td>1</td>
<td>-0.077300</td>
<td>0.02551</td>
<td>9.18225</td>
<td>0.0024</td>
<td>0.926</td>
</tr>
<tr>
<td>HANDLING</td>
<td>1</td>
<td>0.004798</td>
<td>0.03221</td>
<td>0.02219</td>
<td>0.8816</td>
<td>1.005</td>
</tr>
<tr>
<td>LOGCATCH</td>
<td>1</td>
<td>-0.225158</td>
<td>0.07156</td>
<td>9.89924</td>
<td>0.0017</td>
<td>0.798</td>
</tr>
<tr>
<td>TOWDEPTH</td>
<td>1</td>
<td>0.002931</td>
<td>0.0004996</td>
<td>34.40781</td>
<td>0.0001</td>
<td>1.003</td>
</tr>
<tr>
<td>TOWLNGTH</td>
<td>1</td>
<td>0.001180</td>
<td>0.0003541</td>
<td>11.10036</td>
<td>0.0009</td>
<td>1.001</td>
</tr>
<tr>
<td>TOWHAND</td>
<td>1</td>
<td>0.001107</td>
<td>0.0003558</td>
<td>9.67706</td>
<td>0.0019</td>
<td>1.001</td>
</tr>
<tr>
<td>DEPLNGTH</td>
<td>1</td>
<td>-0.006034</td>
<td>0.00136</td>
<td>19.77360</td>
<td>0.0001</td>
<td>0.994</td>
</tr>
<tr>
<td>DEPHAND</td>
<td>1</td>
<td>-0.004104</td>
<td>0.00118</td>
<td>12.00517</td>
<td>0.0005</td>
<td>0.996</td>
</tr>
</tbody>
</table>

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, you might be able 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 - this is referred to as the ‘hierarchical principle’)

- 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 techniques we learned to check fit of model.
(2) $R^2$-type Measures

$R^2$-type measures have always been considered useful in practice, to quantify how much variation in the outcome is explained by the regressors or predictors.

It has also been used to quantify genetic heritability, and more recently, the ‘polygenic risk scores’.

For predictability, ‘out of sample’ $R^2$ has been used in machine learning approaches.

**Eg.** In a prognostic study in gastric cancer, we wanted to investigate the prognostic effects of blood-based acute phase reactant proteins (i.e. biomarkers) and stage on survival. Note that stage is only available after surgery. The types of questions we were interested in:

1. How much of the variability in survival is explained, by the biomarkers and/or stage?

2. How strong are the effects of certain prognostic variables once others have been accounted for?

3. How much predictability is lost, if at all, when replacing a continuously measured covariate by a binary coding?
4. In other disease areas with $Z(t)$, eg. CD4 counts in AIDS patients, how much is the effect on survival “captured” by such a surrogate?

As pointed out by Korn and Simon (1991), among others, the $R^2$ measure concerns explained variation, or predictive capability, but not the goodness-of-fit of a model (which is a common misunderstanding).
$R^2$ measure for linear regression

\[ y_i = \alpha + \beta x_i + \epsilon_i, \quad E(\epsilon_i) = 0 \]

\[ E_{\hat{\beta}}(Y|x_i) = \hat{\alpha} + \hat{\beta} x_i \]

\[ E_0(Y|x_i) = \bar{\alpha} = \bar{y} \]

- Residual

\[ r_i(\hat{\beta}) = y_i - E_{\hat{\beta}}(Y|x_i), \]

measures the discrepancy between observed response $y_i$ and its estimated expectation $\hat{y}_i$, under the model and given $X = x_i, \ i = 1, \ldots, n$,

- while

\[ r_i(0) = y_i - \bar{y}, \]

measures the discrepancy between observed response $y_i$ and its estimated expectation $\hat{y}_i$, without a model.

- Then, $\sum_{i=1}^{n} r_i^2(\hat{\beta})/n$ is average squared discrepancy between the observed $y_i$ and its estimated expected value under the model, and $\sum_{i=1}^{n} r_i^2(0)/n$ the average discrepancy without a model.

- So define

\[ R^2(\hat{\beta}) = 1 - \frac{\sum_{i=1}^{n} r_i^2(\hat{\beta})}{\sum_{i=1}^{n} r_i^2(0)} = 1 - \frac{SS(\hat{\beta})}{SS(0)} \]
Cox regression

\[ \lambda(t|Z) = \lambda_0(t) \exp\{\beta Z\} \]

Here for simplicity we first assume \( Z \) of dimension one and time-independent.

- (Recall) Schoenfeld residual
  \[ r_i(\hat{\beta}) = Z_i - E_\hat{\beta}(Z|X_i), \]
  where
  \[ E_\beta(Z|t) = \frac{\sum_{j \in R(t)} Z_j e^{\beta Z_j}}{\sum_{j \in R(t)} e^{\beta Z_j}} \]

- Now \( E_0(Z|t) = \sum_{j \in R(t)} Z_j / |R(t)| \) is the empirical average of the \( Z \)’s in the risk set at time \( t \), corresponding to \( \beta = 0 \). Let
  \[ r_i(0) = Z_i - E_0(Z|X_i). \]

- In the absence of censoring, \( \sum_{i=1}^n r_i^2(\hat{\beta})/n \) is average discrepancy between the observed covariate and its expected value under the model, and \( \sum_{i=1}^n r_i^2(0)/n \) the average discrepancy without a model.
• Define (O’Quigley and Flandre 1994, *Proceedings of the National Academy of Science*)

\[ R^2(\hat{\beta}) = 1 - \frac{\sum_{i=1}^{n} \delta_i r_i^2(\hat{\beta})}{\sum_{i=1}^{n} \delta_i r_i^2(0)} = 1 - \frac{SS(\hat{\beta})}{SS(0)} \]
Interpretation (1)
Sum of squares decomposition

Let

\[ SS_{\text{res}} = SS(\hat{\beta}) = \sum_{i=1}^{n} \delta_i r_i^2(\hat{\beta}) \]

\[ SS_{\text{tot}} = SS(0) = \sum_{i=1}^{n} \delta_i r_i^2(0) \]

\[ SS_{\text{reg}} = \sum_{i=1}^{n} \delta_i \{ E_{\hat{\beta}}(Z|X_i) - E_0(Z|X_i) \}^2 \]

Then asymptotically we can show that

\[ SS_{\text{tot}} \overset{\text{asympt.}}{=} SS_{\text{res}} + SS_{\text{reg}} \quad \text{(why)} \]

[Recall in linear regression \( SS_{\text{tot}} = SS_{\text{res}} + SS_{\text{reg}} \).]

So that

\[ R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \approx \frac{SS_{\text{reg}}}{SS_{\text{tot}}} \]

— Just like in the linear regression!
Interpretation (2)
Predictability of failure rankings

Suppose we have $Z_i < Z_j$ and $\beta > 0$. What does this imply about the survival times of subject $i$ vs $j$?

- under ‘perfect prediction’, we would have $T_i > T_j$ for sure
- if $Z$ is independent of $T$ (no predictive power), $P(T_i > T_j)$ should be 0.5.

$P(T_i > T_j|Z_i < Z_j)$ (concordance) reflects the predictive capability of $Z$ under the model.

- Denote $\Omega^2(\beta)$ the population parameter that $R^2(\hat{\beta})$ is estimating.

- We can show that $\Omega^2(\beta)$ increases with $P(T_i > T_j|Z_i < Z_j)$ for $\beta > 0$, and similar results for $\beta < 0$.

- We can think of this as: in large samples, $R^2(\hat{\beta})$ increases with the probability of $T_i > T_j$ if $Z_i < Z_j$ and $\beta > 0$. (How do you state the result for $\beta < 0$?)
Interpretation (3)

Explain variation

We have approximately

$$R^2 \approx 1 - \frac{E\{\text{Var}(Z|T)\}}{\text{Var}(Z)} = \frac{\text{Var}\{E(Z|T)\}}{\text{Var}(Z)}.$$ 

This translates directly as the proportion of variation in $Z$ explained by $T$.

But we are interested in predicting $T$ using $Z$. Why should we care about how much variation in $Z$ is explained by $T$?

- For bivariate normal, both are equivalent to the correlation coefficient squared.

- For the Cox model, predicting $Z$ given $T$ is equivalent to predicting the ranks of $T$ given $Z$.

- Also recall that the regression effect in the semiparametric Cox model only depends on the ranks of the failure times.
<table>
<thead>
<tr>
<th>$\beta$</th>
<th>Binary</th>
<th>Uniform</th>
<th>Normal</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>1.0</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.19</td>
</tr>
<tr>
<td>1.5</td>
<td>0.39</td>
<td>0.38</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>2.0</td>
<td>0.56</td>
<td>0.52</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>2.5</td>
<td>0.70</td>
<td>0.65</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>3.0</td>
<td>0.82</td>
<td>0.76</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td>3.5</td>
<td>0.91</td>
<td>0.84</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>4.0</td>
<td>0.98</td>
<td>0.91</td>
<td>0.98</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Population values of $R^2$ ($\Omega^2$) as a function of $\beta$** for different distributions of $Z$, all with the same variance (0.25).
Time-dependent covariates

It is immediate to generalize to time-dependent covariates, since the Schoenfeld residuals are well defined.

Define

\[ R^2(\hat{\beta}) = 1 - \frac{\sum_{i=1}^{n} \delta_i r_i^2(\hat{\beta})}{\sum_{i=1}^{n} \delta_i r_i^2(0)} \]

where

\[ r_i(\beta) = Z_i(X_i) - E_\beta(Z|X_i), \]

\[ E_\beta(Z|t) = \frac{\sum_{j \in R(t)} Z_j(t) e^{\beta Z_j(t)}}{\sum_{j \in R(t)} e^{\beta Z_j(t)}} \]

for \( \beta = \hat{\beta} \) and 0.
Properties of the $R^2$ measure

• have a value 0 when $\beta = 0$

• $\Omega^2(\beta)$ increases monotonically with $|\beta|$

• $\Omega^2(\beta)$ tend to 1 as $|\beta| \to \infty$

• invariant under linear transformations of $Z$ and monotonically increasing transformations of $T$

It is relatively easy to compute after fitting the Cox model.
Multiple covariates

• For multiple covariates, since the dependence of the survival time on the covariates is via the **prognostic index** $\beta'Z(t)$, we replace $r_i(\hat{\beta})$ in the above by
\[
\hat{\beta}' r_i(\hat{\beta}) = \hat{\beta}' Z_i(X_i) - E_{\hat{\beta}}(\hat{\beta}' Z|X_i),
\]
and replace $r_i(0)$ by
\[
\hat{\beta}' r_i(0) = \hat{\beta}' Z_i(X_i) - E_0(\hat{\beta}' Z|X_i),
\]

• The way to think of this is to imagine that every subject is ‘labelled’ by its $\beta'Z$. Two subjects with the same prognostic index should have the same survival distribution under the model. So in each risk set, choosing which subject to fail is equivalent to choosing its $\beta'Z$.

• When the dimension of $Z$ is one, the prognostic index is equivalent to $Z$ itself ($\beta \neq 0$).

• So define
\[
R^2(\hat{\beta}) = 1 - \frac{\sum_{i=1}^{n} \delta_i [\hat{\beta}' r_i(\hat{\beta})]^2}{\sum_{i=1}^{n} \delta_i [\hat{\beta}' r_i(0)]^2}.
\]
Examples

**Example 1.** For the Freireich data, where 42 leukemia patients were randomized to 6-MP ($Z = 0$) or placebo ($Z = 1$), we have $\hat{\beta} = 1.65$, and $R^2 = 0.42$, indicating a moderately strong degree of predictability.

**Example 2.** For the gastric cancer data, the following table is from univariate Cox model fits with one covariate at a time:

<table>
<thead>
<tr>
<th>covariate</th>
<th>$\hat{\beta}$</th>
<th>p-value</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>stage</td>
<td>1.78</td>
<td>$&lt;0.01$</td>
<td>0.48</td>
</tr>
<tr>
<td>ACT</td>
<td>2.26</td>
<td>$&lt;0.01$</td>
<td>0.29</td>
</tr>
<tr>
<td>log(CEA)</td>
<td>0.30</td>
<td>$&lt;0.01$</td>
<td>0.20</td>
</tr>
<tr>
<td>CRP</td>
<td>0.02</td>
<td>$&lt;0.01$</td>
<td>0.26</td>
</tr>
<tr>
<td>AGP</td>
<td>0.70</td>
<td>$&lt;0.01$</td>
<td>0.14</td>
</tr>
</tbody>
</table>

In multivariate regression, CRP and AGP are no longer of prognostic significance in the presence of other covariates. The value of $R^2$ for a model with ACT and log(CEA) is 0.37, and this increases to 0.54 when stage is also included. (This aimed to answer the scientific question if blood chemicals alone are sufficient to predict survival, since stage can only be obtained with surgery.)
Another measure

Another $R^2$ type measure (sometimes called ‘Generalized $R^2$’) is usually calculated using the likelihood ratio statistics. Here the partial likelihood ratio can be used.

- Kent (1983) defined the proportion of the explained randomness for i.i.d. data:
  Let $I(\beta) = E\{\log f(X; \beta)\} = E\{\log L(\beta)/n\}$, which corresponds to the Kullback-Leibler information.
  The randomness of $Z$ is defined by $D(Z) = \exp\{-2I(0)\}$, and the residual randomness of $Z$ given $T$ is $D(Z|T) = \exp\{-2I(\beta)\}$.
  Define
  $$\rho^2 = 1 - \frac{D(Z|T)}{D(Z)},$$

- Under the Cox model, let
  $$L(\beta) = \prod_{\delta_i=1} \pi_i(\beta) = \prod_{\delta_i=1} \frac{\exp\{\beta'Z_i(X_i)\}}{\sum_{R(X_i)} \exp\{\beta'Z_j(X_i)\}}$$
  be the partial likelihood.
Then a sample based $\rho^2$ measure can be defined
\[ \hat{\rho}^2 = 1 - e^{-\Gamma}, \]
where
\[ \Gamma = 2\{\log L(\hat{\beta}) - \log L(0)\}/k \]
\[ = \frac{\sum_{i=1}^k 2\log\{\pi_i(\hat{\beta})/\pi_i(0)\}}{k} \]
and $k$ is the number of events.

Note that with uncensored data, the above is divided by $n$ the sample size, here we need to use the number of events $k$.

\[ \Gamma \] estimates twice the Kullback-Leibler information gain, between the fitted model and the null model.

$\hat{\rho}^2$ and its population parameter enjoy the same properties that we listed for $R^2$, and is also simple to calculate.

**Example** For Freireich data, the partial likelihood ratio statistic is 15.2, and there were 30 events. So
\[ \hat{\rho}^2 = 1 - \exp(-15.2/30) = 0.40, \]
agreeing quite well with the $R^2 = 0.42$ that we had before.
Note that the measure computed in Allison’s SAS book was wrong, dividing the likelihood ratio statistics by $n$ instead of $k$. Here are some simulation results ($n=500$, 100 simulations):

<table>
<thead>
<tr>
<th>Hazard ratio</th>
<th>Percent censored</th>
<th>Covariate type</th>
<th>rho2_n</th>
<th>rho2_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>c</td>
<td>0.104</td>
<td>0.104</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>c</td>
<td>0.059</td>
<td>0.113</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>c</td>
<td>0.016</td>
<td>0.137</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>c</td>
<td>0.298</td>
<td>0.298</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>c</td>
<td>0.191</td>
<td>0.340</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>c</td>
<td>0.044</td>
<td>0.349</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>c</td>
<td>0.599</td>
<td>0.599</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>c</td>
<td>0.460</td>
<td>0.708</td>
</tr>
<tr>
<td>16</td>
<td>90</td>
<td>c</td>
<td>0.110</td>
<td>0.718</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>c</td>
<td>0.756</td>
<td>0.756</td>
</tr>
<tr>
<td>64</td>
<td>50</td>
<td>c</td>
<td>0.627</td>
<td>0.860</td>
</tr>
<tr>
<td>64</td>
<td>90</td>
<td>c</td>
<td>0.198</td>
<td>0.870</td>
</tr>
</tbody>
</table>

Here $\rho_2_n$ is the measure from Allison’s book, and $\rho_2_k$ is what we defined. The censoring mechanism is $U(0, \tau)$. The changes in $\rho_2_k$ for different amount of censoring are mainly due to the upper bound $\tau$ so that we only have limited information on the survival distribution. See O’Quigley et al. (2005).
Remarks

• In *Handbook of Statistics in Clinical Oncology* (ed. Crowley), Chapter on ‘Explained variation in proportional hazards regression’, it is argued that in the definition of $R^2$, the squared Schoenfeld residuals should be weighted to offset the censoring effect, and the weights should be the increments of the marginal KM curve. This will make the measure closer to the population parameter $\Omega^2$. But for most practical purposes, both versions give similar results.

• Following the same reasoning as above, the $2 \log \left\{ \pi_i(\hat{\beta})/\pi_i(0) \right\}$’s in the definition of $\rho^2$, instead of summing up and dividing by $k$, strictly speaking should also be weighted proportionally to the increments of the marginal KM curve (Xu and O’Quigley, 1999). Again, in practice the two versions are similar, so you may not have to compute the more complicated one (O’Quigley et al., 2005).

• We saw some examples of $R^2$ and $\rho^2$ values. For certain data such as time-dependent CD4 counts, which often serves as a surrogate endpoint for survival in AIDS studies, the explained variation/randomness in survival by CD4 counts can be as high as over 80%.
(3) Information Criteria

Information criteria have been used for model selection.

**Akaike Information** (classical):

- Consider $g_\theta$ as a family of distributions indexed by parameter $\theta$, that is used to model the observed data $y$.

- Kullback-Leibler (KL) information $E\{\log g_\theta(y)\}$ gives a risk function, corresponding to the deviance loss function $l(y, \theta) = -2 \log g_\theta(y)$ (see for example, Lehmann ‘Point Estimation’).

- We want to choose the model (say between $g_\theta$ and $h_\phi$) that minimizes the risk, as well as given the model family, say $g_\theta$, to choose the $\theta$ value that minimizes the risk.

- Note that the KL information gain gives a kind of ‘distance’ from the true distribution $f$ that generates the data $y$, to $g_\theta$:

  $$KL(f, g_\theta) = E_f\{\log f(y) - \log g_\theta(y)\}.$$ 

- Minimum $KL$ is attained at $\theta_0$ such that $KL(f, g_{\theta_0}) = \min_{\theta} KL(f, g_\theta)$ or, equivalently,

  $$E\{\log g_{\theta_0}(y)\} = \max_{\theta} E\{\log g_\theta(y)\}.$$
• Then $g_{\theta_0}$ is the best approximation to $f$ within the family of models, and $\theta_0$ is sometimes called the ‘least-false’ parameter.

• When the model is correct, we have $f = g_{\theta_0}$. [ex]

• In practice $\theta_0$ is often estimated by the MLE $\hat{\theta}(y)$. (why)

• Then the risk $-2E\{\log g_{\theta_0}(y)\}$ is estimated by

$$-2E_{y^*}\{\log g(y^*|\hat{\theta}(y))\}.$$ 

Note that we use $y^*$ to denote the r.v. that the expectation is w.r.t., in order to distinguish from $y$ the observed data that’s used to estimate $\theta_0$.

• Twice the expected risk in this case is the Akaike Information:

$$AI = -2E_yE_{y^*}\{\log g(y^*|\hat{\theta}(y))\}. \tag{1}$$

It is also referred to as the predictive log-likelihood, or the expected KL. Note that $y^*$ is an independent replicate of $y$, i.e. from the same distribution as $y$.

• The model should be chosen to minimize the AI, which itself needs to be estimated.
• Q: how would you estimate AI?

• It is well-known that the ‘apparent’ estimate $-2 \log g(y|\hat{\theta}(y))$ under-estimates AI. (why)

• Instead Akaike (1973) showed using asymptotic 2nd-order Taylor expansion that

$$AIC = -2 \log g(y|\hat{\theta}(y)) + 2p$$  \hspace{1cm} (2)

is an approximately unbiased estimator of AI, where $p$ is the dimension of $\theta$ in the classical parametric case.

• Therefore the model is chosen to minimize the AIC.
AIC for the Cox Model
(or Semiparametric Models in General)

Since under the Cox model, \( \lambda_0(t) \) is unspecified, model selection focuses on the \( \beta \) part.

- It is perhaps intuitive that the partial likelihood should be used in forming AIC.

- This turns out to be the case (Xu et al., 2009), because the partial likelihood possesses the properties of a classical likelihood, especially in terms of the 2nd order Taylor expansion that was used in the proof of classic AIC (recall partial likelihood as a profile likelihood).

- So

\[
AIC = -2 \, pl(y|\hat{\beta}(y)) + 2p,
\]

where \( pl() \) is the log partial likelihood, and \( p \) is the dimension of \( \beta \).
Interpretation

- In terms of the underlying KL information, since the partial likelihood does not exactly correspond to a density function, it has to be viewed via the profile likelihood.

- In this case, denote $\theta = (\beta, \lambda_0)$, where $\beta$ is the parameter of interest, and $\lambda_0$ is the nuisance parameter.

- Since $\lambda_0$ is left unspecified, the relevant ‘distance’ is that between the true distribution $f$ and the subfamily of models $\{g_{\beta,\lambda_0} : \lambda_0 \in \Lambda\}$, which is: $\min_{\lambda_0 \in \Lambda} KL(f, g_{\beta,\lambda_0})$.

- This is equivalent to

  $$\max_{\lambda_0} E \{ \log g_{\beta,\lambda_0}(y) \},$$

  which in the empirical version corresponds to the profile likelihood

  $$pl(y|\beta) = \max_{\lambda_0} \log g(y|\beta, \lambda_0).$$

- You can read more in Xu, Vaida and Harrington (2009) about profile Akaike information.
**BIC for the Cox Model**

For classical parametric models, the Bayesian information criterion is

\[ BIC = -2 \log g(y|\hat{\theta}(y)) + p \cdot \log(n), \]  

(3)

where \( n \) is the sample size.

For censored survival data, the number of events \( d \), instead of the number of subjects \( n \), is the ‘effective’ sample size.

Note that \( d \) is the number of terms in the partial likelihood, as well as the number of terms in the Fisher information.

Volinsky and Raftery (2000) proposed BIC for the Cox model:

\[ BIC = -2 \log l(y|\hat{\beta}(y)) + p \cdot \log(d), \]

where \( d \) is the number of events.
(4) Penalized log (partial) likelihood

Almost all the methods for model selection we discuss here can be written as choosing $\beta$ to maximize a penalized log (partial) likelihood:

$$pl(y|\beta) - P_\lambda(\beta),$$

where $\lambda \geq 0$ is the penalty parameter, and often we can use the penalty $P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^m$.

1. $m = 0$, $L_0$ penalty: best subset (AIC), stepwise (might require orthonormal design under linear regression), adjusted $R^2$, generalized cross-validation (GCV).

2. $m = 1$, $L_1$ penalty: least absolute shrinkage and selection operator (LASSO; Tibshirani, 1997).

3. $m = 2$, $L_2$ penalty: ridge regression (Huang and Harrington, 2002).

4. Other penalties: smoothly clipped absolute deviation (SCAD; Fan and Li, 2002), non-concave (Bradic et al. 2011).