On a Measure of Information Gain for Regression Models in Survival Analysis

Pascal Roy
Delphine Maucort-Boulch, Janez Stare

Equipe Biostatistique Santé, UMR CNRS 5558

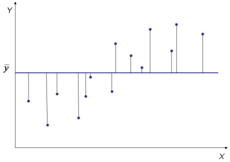
26th May 2010

Hôpitaux de Lyon

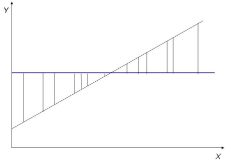
Outline

1. Information gain measure
   - Explained variation
   - Expected likelihood ratio
2. Simulations study
   - Method
   - Results
3. Conclusion

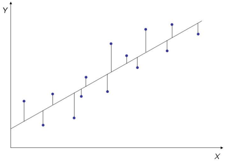
In linear regression



$=$



$+$



$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$SS_{tot} = SS_{reg} + SS_{res}$$

$$\rho^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_{reg}}{SS_{tot}}$$

In survival models

The most comprehensive (Kent *et al*)

$$\rho^2_{IG} = 1 - e^{-E(LR)}$$

Softwares output

$$\hat{\rho}^2_{\mathrm{n}} = 1 - e^{-\frac{2}{\mathrm{n}}[\mathrm{loglik}_{\mathrm{model}} - \mathrm{loglik}_{\mathrm{null}}]}$$

But no statistical justification

O'Quigley *et al* proposal

$$\hat{\rho}^2_{IG} = 1 - e^{-\frac{2}{\sharp\mathrm{events}}[\mathrm{loglik}_{\mathrm{model}} - \mathrm{loglik}_{\mathrm{null}}]}$$

Population value

We are interested in

$$E(LR) = 2\int_0^{\infty} log\left(\frac{f_M(t)}{f_0(t)}\right) dF_M(t)$$

Twice the Kullback Leibler information gain

The suggestion by O'Quigley *et al* was reported to be biased under censoring, although less than the ρ_n^2

- The jumps in the survival curve are not all equal to $1/k$
- The overall survival curve may not drop to 0

Last observed failure time τ

When the last observed failure occurs at time τ because of censoring or because we want to limit ourselves to observations less than a given time τ

$$E(LR) = 2\int_0^{\infty} log\frac{f_M(t)}{f_0(t)} dF_M(t|\tau) = 2\int_0^{\infty} log\frac{f_M(t)}{f_0(t)} \frac{dF_M(t)}{F_M(\tau)}$$

Two types of censoring

Before the last failure time τ

Attenuation of the sample size and, if random, only affect the variability of the estimator, so efficiency, not its expected value
$\rightarrow$ Weights, as jumps in the survival curve, should compensate for the missing information

$$\hat{E}_w(LR) = 2\sum_1^k \log(\hat{LR})\frac{\Delta\hat{F}_M(t)}{\hat{F}_M(\tau)}$$

After the last failure time τ

No information on precision after τ and since a measure of predictive accuracy is an overall measure, it will be affected by such censoring
$\rightarrow$ impute under the model

Data generation

Complete data

- Covariate
  - continuous $U[0, \sqrt{3}]$
  - binary
- $\beta \in \{1, 2, 5\}$
- Times generated under exponential model
- $n \in \{200, 500, 1000, 5000\}$
- Cox model fitted
- 100 iterations

Censoring

1. Complete data censored in two different ways
   1. random censoring
      τ highest failure time determined
   2. Type 1 censoring
      all times greater than τ are censored
2. Uniform censoring, percentages from 10% to 90%

Data generation

Complete data

- Covariate
  - continuous $U[0, \sqrt{3}]$
  - binary
- $\beta \in \{1, 2, 5\}$
- Times generated under exponential model
- $n \in \{200, 500, 1000, 5000\}$
- Cox model fitted
- 100 iterations

Censoring

1. Complete data censored in two different ways
   1. random censoring
      τ highest failure time determined
   2. Type 1 censoring
      all times greater than τ are censored
2. Uniform censoring, percentages from 10% to 90%

Measures

Different estimations

$$\rho_n^2 = 1 - \exp(\frac{2}{\mathrm{n}} \sum_1^k \log(\hat{LR}))$$

$$\rho_k^2 = 1 - \exp(\frac{2}{\mathrm{k}} \sum_1^k \log(\hat{LR}))$$

$$\rho_w^2 = 1 - \exp\left(2 \sum_1^k \log(\hat{LR}) \frac{\Delta \hat{F}_M(t)}{\hat{F}_M(\tau)}\right)$$

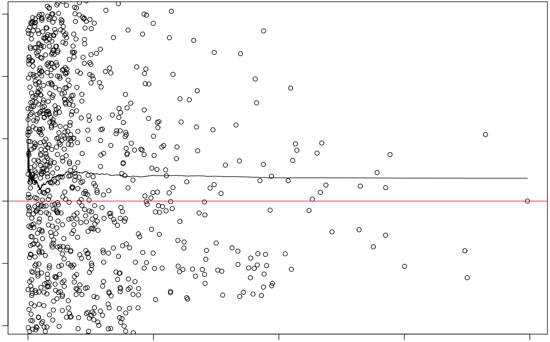
$$\rho_i^2 = 1 - \exp\left(2 \sum_1^{k'} \log(\hat{LR}) \Delta \hat{F}_M(t)\right)$$

- ρ_w^2: Kaplan-Meier estimates of $\hat{F}_M(t)$
- ρ_i^2: average on 10 imputations

Three data sets

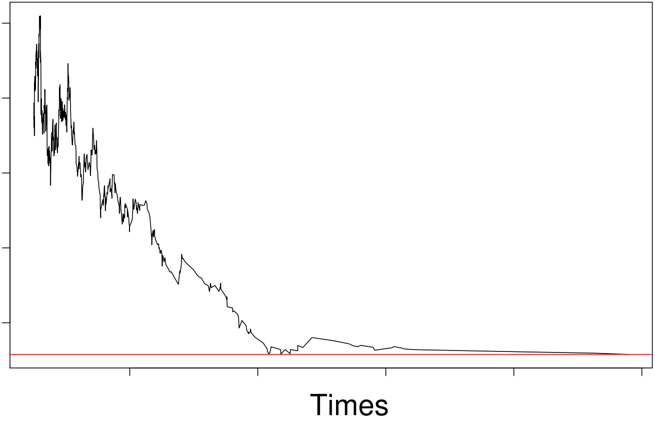
1. Randomly censored
2. Censored after τ
3. Complete

log likelihood ratio components over time



Times

ρ^2 over time



Randomly censored data

Population value 0.15 and 0.38 for continuous variable with $\beta \in \{1, 2\}$ respectively

β	Size	%	ρ_n	se	ρ_k	se	ρ_w	se	ρ_i	se
1	5000	80	0.04	0.01	0.19	0.03	0.19	0.04	0.17	0.03
1	5000	50	0.09	0.01	0.18	0.01	0.17	0.02	0.15	0.02
1	5000	20	0.13	0.01	0.16	0.01	0.15	0.01	0.14	0.01
1	200	80	0.04	0.03	0.20	0.12	0.21	0.15	0.16	0.11
1	200	50	0.10	0.03	0.19	0.06	0.18	0.07	0.15	0.06
1	200	20	0.12	0.04	0.15	0.05	0.14	0.05	0.14	0.05
2	5000	80	0.13	0.01	0.50	0.02	0.51	0.04	0.42	0.03
2	5000	50	0.29	0.01	0.50	0.01	0.48	0.02	0.39	0.02
2	5000	20	0.37	0.01	0.44	0.01	0.40	0.01	0.39	0.01
2	200	80	0.13	0.04	0.49	0.12	0.47	0.17	0.38	0.12
2	200	50	0.29	0.05	0.49	0.08	0.46	0.09	0.38	0.07
2	200	20	0.36	0.05	0.43	0.05	0.40	0.05	0.38	0.05

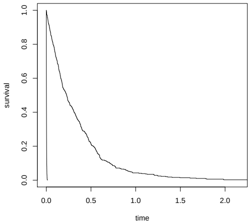
Uncensored before τ data

Population value 0.15 and 0.38 for continuous variable with $\beta \in \{1, 2\}$ respectively

β	Size	%	ρ_n	se	ρ_w	se	ρ_i	se
1	5000	80	0.07	0.01	0.19	0.02	0.17	0.02
1	5000	50	0.13	0.01	0.17	0.01	0.15	0.01
1	5000	20	0.14	0.01	0.15	0.01	0.15	0.01
1	200	80	0.07	0.03	0.21	0.09	0.18	0.08
1	200	50	0.13	0.04	0.18	0.05	0.16	0.05
1	200	20	0.14	0.04	0.15	0.05	0.14	0.04
2	5000	80	0.21	0.01	0.51	0.02	0.42	0.01
2	5000	50	0.37	0.01	0.48	0.01	0.39	0.01
2	5000	20	0.39	0.01	0.40	0.01	0.39	0.01
2	200	80	0.19	0.05	0.49	0.09	0.41	0.07
2	200	50	0.35	0.06	0.47	0.08	0.38	0.05
2	200	20	0.38	0.05	0.40	0.05	0.38	0.05

- Weights and imputation correct for the bias of O'Quigley *et al* proposal
- Gain in estimation comes with a price - a bigger variance

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

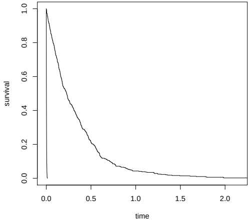
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model: $\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model: $\beta = 5.01$, $\rho^2_{IG} =$
Fit with Cox model: $\beta = 5.03$, $\rho^2_{IG} =$
Kent O'Quigley $\rho^2_{IG} =$

R output

```
Rsquare= 0.711
 (max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

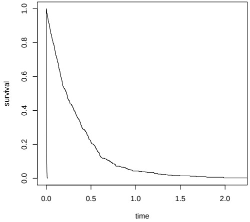
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model: $\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model: $\beta = 5.01$, $\rho^2_{IG} =$
Fit with Cox model: $\beta = 5.03$, $\rho^2_{IG} =$
Kent O'Quigley $\rho^2_{IG} =$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\to}$ 1

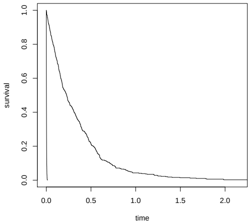
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model: $\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model: $\beta = 5.01$, $\rho^2_{IG} =$
Fit with Cox model: $\beta = 5.03$, $\rho^2_{IG} =$
Kent O'Quigley $\rho^2_{IG} =$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

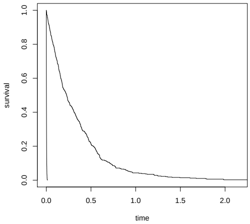
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\to}$ 1

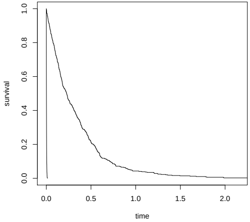
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

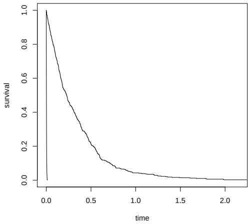
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.892$

R output

```
Rsquare= 0.711
 (max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

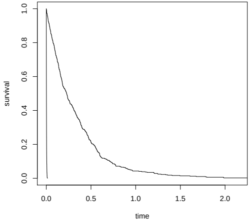
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$,	$\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$,	$\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$,	$\rho^2_{IG} = 0.711$
Kent O'Quigley		$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

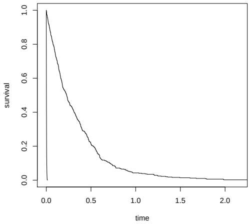
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} =$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta \to \infty}{\longrightarrow}$ 1

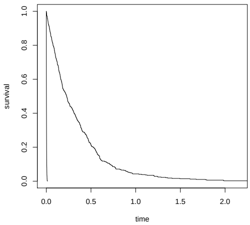
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

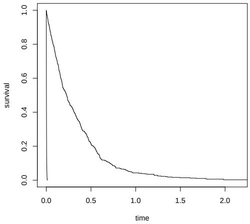
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\to}$ 1

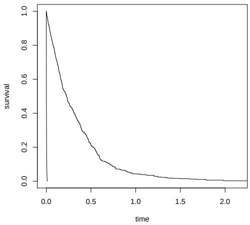
exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model: $\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model: $\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model: $\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley $\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

Two groups case



some measures $\overset{\beta\to\infty}{\longrightarrow}$ 1

exponential distribution
$\beta = 5$
sample size 1000

Fit with exponential model:	$\beta = 5.01$, $\rho^2_{IG} = 0.975$
Fit with Weibull model:	$\beta = 5.01$, $\rho^2_{IG} = 0.830$
Fit with Cox model:	$\beta = 5.03$, $\rho^2_{IG} = 0.711$
Kent O'Quigley	$\rho^2_{IG} = 0.832$

R output

```
Rsquare= 0.711
(max possible=1)
```

- Extensions to parametric models
  - Log likelihood

$$\sum_{\textit{uncensored}} \ln f(t|\beta) + \sum_{\textit{censored}} \ln S(t|\beta)$$

  - Influence of time values

References

- Kent JT, O'Quigley J. Measures of dependence for censored survival data. Biometrika 1988;75:525-534.
- Kullback S, Leibler RA. On information and sufficiency. Annals of Math. Stats.1951;22:7986.
- O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. Statist.Med. 2005;24:479489.