General Introduction to R^2 in survival

Janez Stare

IBMI, Faculty of Medicine, Ljubljana, Slovenia

Bordeaux, October 2013

Explained variation in linear regression - illustration 1

Explained variation in linear regression - illustration 2

Explained variation in linear regression - illustration 3

What do we mean by explained variation

Law of total variance

What do we mean by explained variation

Law of total variance

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

What do we mean by explained variation

Law of total variance

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

In linear regression under the usual assumptions

$$\sigma^2{}_Y = \sigma^2{}_{Res} + \sigma^2{}_{Reg}$$

What do we mean by explained variation

Law of total variance

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

In linear regression under the usual assumptions

$$\sigma^2{}_Y = \sigma^2{}_{Res} + \sigma^2{}_{Reg}$$

The population R^2 is then defined as

$$R^2 = 1 - \frac{\sigma^2{}_{Res}}{\sigma^2{}_Y}$$

Explained variation in linear regression

$$R^2 = 1 - \frac{\sigma^2{}_{Res}}{\sigma^2{}_{Y}}$$

Explained variation in linear regression

$$R^2 = 1 - \frac{{\sigma^2}_{Res}}{{\sigma^2}_Y}$$

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

Explained variation in linear regression

$$R^2 = 1 - \frac{{\sigma^2}_{Res}}{{\sigma^2}_Y}$$

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

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

Explained variation in linear regression

$$R^2 = 1 - \frac{{\sigma^2}_{Res}}{{\sigma^2}_{Y}}$$

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

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

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

Explained variation in linear regression

$$R^2 = 1 - \frac{{\sigma^2}_{Res}}{{\sigma^2}_{Y}}$$

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

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

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

$$\frac{SS_{tot}}{n-1} = \frac{SS_{res}}{n-1} + \frac{SS_{reg}}{n-1}$$

Explained variation in linear regression

- Most important: **decomposition** into explained and unexplained variation!!

Explained variation in linear regression

- Most important: **decomposition** into explained and unexplained variation!!
- If you lower the unexplained part, you increase the explained part by the **same amount**!

Explained variation in linear regression

- Most important: **decomposition** into explained and unexplained variation!!
- If you lower the unexplained part, you increase the explained part by the **same amount**!
- This ensures that we have a measure between 0 and 1

Illustration of wrong 'bad' properties of R^2

J. R. Statist. Soc. A (1984),
147, Part 1, pp. 100–103

242

2

The Box-Wetz Criterion Versus R^2

By NORMAN R. DRAPER

University of Wisconsin, USA

SUMMARY

The square of the multiple correlation coefficient in a regression fit, R^2, can be made small simply by increasing the number of repeat data points. It is argued here that R^2 is misleading in such cases, and that the Box-Wetz (1973) criterion, and a simpler but essentially equivalent criterion, are not.

Keywords: MULTIPLE CORRELATION COEFFICIENT; PURE ERROR; REGRESSION WITH LARGE DATA SETS

1. INTRODUCTION

Figs 1(a) and (b) indicate two sets of data. In Fig. 1(a) there are five X locations $r = 1, 2, \ldots, 5$ and one Y observation at each location. In Fig. 1(b), there are ten Y points at the five locations. Suppose we fit a straight line to each of these two sets and the underlying model is really a straight line. Because the additional data help to locate the true mean values at each location more precisely, the fit must be improved. However, the R^2 statistic will be lower than before indicating, it would seem, a worse fit. Why?

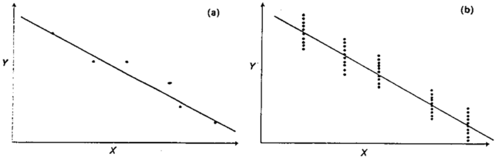


Fig. 1. Five data sites with (a) one run, (b) ten runs, per site.

2. ARGUMENT

The reason lies in the fact that it is impossible for a fitted model to explain pure error. Suppose we fit, by least squares, the model

$$Y = \eta + \varepsilon = X\beta + Z\psi + \varepsilon, \qquad (1)$$

where $X\beta$ is the part to be tested in a "test for regression" and $Z\psi$ represents effects such as the mean, block variables, time trends, and so on, that we wish to eliminate from the variation in

Illustration of wrong ‘bad’ properties of R^2

J. R. Statist. Soc. A (1984),
147, Part 1, pp. 100–103

242

2

The Box-Wetz Criterion Versus R^2

By NORMAN R. DRAPER

University of Wisconsin, USA

SUMMARY

The square of the multiple correlation coefficient in a regression fit, R^2, can be made small simply by increasing the number of repeat data points. It is argued here that R^2 is misleading in such cases, and that the Box-Wetz (1973) criterion, and a simpler but essentially equivalent criterion, are not.

Keywords: MULTIPLE CORRELATION COEFFICIENT; PURE ERROR; REGRESSION WITH LARGE DATA SETS

1. INTRODUCTION

Figs 1(a) and (b) indicate two sets of data. In Fig. 1(a) there are five X locations $r = 1, 2, \ldots, 5$ and one Y observation at each location. In Fig. 1(b), there are ten Y points at the five locations. Suppose we fit a straight line to each of these two sets and the underlying model is really a straight line. Because the additional data help to locate the true mean values at each location more precisely, the fit must be improved. However, the R^2 statistic will be lower than before indicating, it would seem, a worse fit. Why?

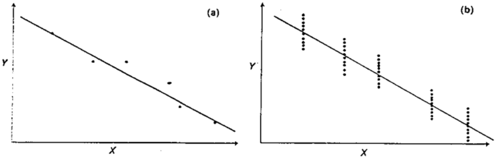


Fig. 1. Five data sites with (a) one run, (b) ten runs, per site.

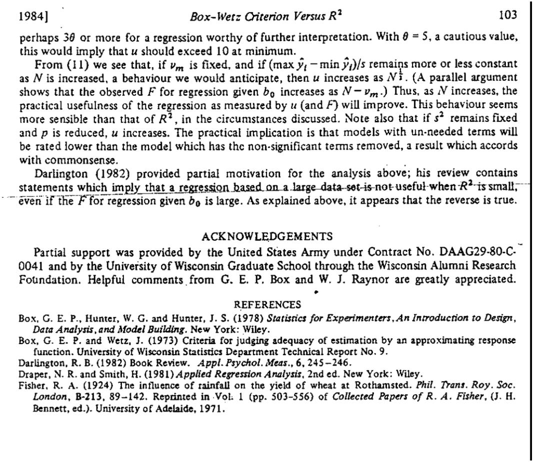
2. ARGUMENT

The reason lies in the fact that it is impossible for a fitted model to explain pure error. Suppose we fit, by least squares, the model

$$\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon}, \qquad (1)$$

where $\mathbf{X}\boldsymbol{\beta}$ is the part to be tested in a “test for regression” and $\mathbf{Z}\boldsymbol{\psi}$ represents effects such as the mean, block variables, time trends, and so on, that we wish to eliminate from the variation in

Illustration of wrong 'bad' properties of R^2



perhaps 3θ or more for a regression worthy of further interpretation. With $\theta = 5$, a cautious value, this would imply that u should exceed 10 at minimum.

From (11) we see that, if ν_m is fixed, and if $(\max \hat{y}_i - \min \hat{y}_i)/s$ remains more or less constant as N is increased, a behaviour we would anticipate, then u increases as $N^{\frac{1}{2}}$. (A parallel argument shows that the observed F for regression given b_0 increases as $N - \nu_m$.) Thus, as N increases, the practical usefulness of the regression as measured by u (and F) will improve. This behaviour seems more sensible than that of R^2, in the circumstances discussed. Note also that if s^2 remains fixed and p is reduced, u increases. The practical implication is that models with un-needed terms will be rated lower than the model which has the non-significant terms removed, a result which accords with commonsense.

Darlington (1982) provided partial motivation for the analysis above; his review contains statements which imply that a regression based on a large data set is not useful when R^2 is small, even if the F for regression given b_0 is large. As explained above, it appears that the reverse is true.

ACKNOWLEDGEMENTS

Partial support was provided by the United States Army under Contract No. DAAG29-80-C-0041 and by the University of Wisconsin Graduate School through the Wisconsin Alumni Research Foundation. Helpful comments from G. E. P. Box and W. J. Raynor are greatly appreciated.

REFERENCES

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building*. New York: Wiley.

Box, G. E. P. and Wetz, J. (1973) Criteria for judging adequacy of estimation by an approximating response function. University of Wisconsin Statistics Department Technical Report No. 9.

Darlington, R. B. (1982) Book Review. *Appl. Psychol. Meas.*, 6, 245–246.

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Fisher, R. A. (1924) The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. Roy. Soc. London*, B-213, 89–142. Reprinted in Vol. 1 (pp. 503–556) of *Collected Papers of R. A. Fisher*, (J. H. Bennett, ed.). University of Adelaide, 1971.

Illustration of wrong 'bad' properties of R^2

perhaps 3θ or more for a regression worthy of further interpretation. With $\theta = 5$, a cautious value, this would imply that u should exceed 10 at minimum.

From (11) we see that, if ν_m is fixed, and if $(\max \hat{y}_i - \min \hat{y}_i)/s$ remains more or less constant as N is increased, a behaviour we would anticipate, then u increases as $N^{\frac{1}{2}}$. (A parallel argument shows that the observed F for regression given b_0 increases as $N - \nu_m$.) Thus, as N increases, the practical usefulness of the regression as measured by u (and F) will improve. This behaviour seems more sensible than that of R^2, in the circumstances discussed. Note also that if s^2 remains fixed and p is reduced, u increases. The practical implication is that models with un-needed terms will be rated lower than the model which has the non-significant terms removed, a result which accords with commonsense.

Darlington (1982) provided partial motivation for the analysis above; his review contains statements which imply that a regression based on a large data set is not useful when R^2 is small, even if the F for regression given b_0 is large. As explained above, it appears that the reverse is true.

ACKNOWLEDGEMENTS

Partial support was provided by the United States Army under Contract No. DAAG29-80-C-0041 and by the University of Wisconsin Graduate School through the Wisconsin Alumni Research Foundation. Helpful comments from G. E. P. Box and W. J. Raynor are greatly appreciated.

REFERENCES

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building*. New York: Wiley.

Box, G. E. P. and Wetz, J. (1973) Criteria for judging adequacy of estimation by an approximating response function. University of Wisconsin Statistics Department Technical Report No. 9.

Darlington, R. B. (1982) Book Review. *Appl. Psychol. Meas.*, 6, 245–246.

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Fisher, R. A. (1924) The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. Roy. Soc. London*, B-213, 89–142. Reprinted in Vol. 1 (pp. 503–556) of *Collected Papers of R. A. Fisher*, (J. H. Bennett, ed.). University of Adelaide, 1971.

Illustration of wrong 'bad' properties of R^2

J. R. Statist. Soc. A (1985)
148, *Part* 4, *p.* 357

Corrections

The Box-Wetz Criterion Versus R^2

By NORMAN R. DRAPER

(*J. R. Statist. Soc.* A, 147, 1984, 100–103.)

Professor J. S. Cramer has kindly pointed out an error in this note. It lies immediately after equation (10) where my assumption that γ_{m} can be kept fixed while the n_i increase is untrue, and so subsequent remarks about R^2 are now wrong.

The correct conclusion from (10) is that, approximately, $R^2 \leqslant \sigma_\eta^2/(\sigma_\eta^2 + \sigma^2)$, the right hand side being the limit of R^2 as N increases, if the model is correct. For a useful regression, we would want σ_η/σ which measures the ratio of the spread of the η's to the spread of the errors to be large, and we see that R^2 provides good information on this since, for large N, $\sigma_\eta/\sigma \approx R/(1-R^2)^{1/2}$.

The behaviour of u and F is sensitive to N, however. From equation (11), if we assume the range of the $\hat{y}_i$ to be "about $\theta\sigma_\eta$" and with $s \approx \sigma$, we find $u \approx N^{1/2}\theta(\sigma_\eta/\sigma)/\nu_m^{1/2}$. Furthermore $F \approx (N - \nu_m - 1)(\sigma_\eta^2/\sigma^2)/\nu_m$. Thus, even if σ_η/σ is small, observed values of u and F can be large for large N, and may give a misleading impression of the explanatory power of the regression. Rather than compare F with the usual percentage point, we may perhaps wish to recalibrate the test, and this is what Box and Wetz (1973) investigated for *moderate* sample sizes.

Overall then, it appears that R^2 *is* a useful indicator for large data sets whereas F can be misleading if used in the ordinary way; thus Darlington (1982) *is* right.

Illustration of wrong 'bad' properties of R^2

J. R. Statist. Soc. A (1985)
148, *Part* 4, *p.* 357

Corrections

The Box–Wetz Criterion Versus R^2

By NORMAN R. DRAPER

(*J. R. Statist. Soc.* A, 147, 1984, 100–103.)

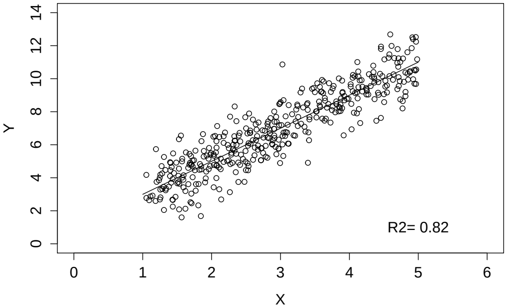
Professor J. S. Cramer has kindly pointed out an error in this note. It lies immediately after equation (10) where my assumption that γ_m can be kept fixed while the n_i increase is untrue, and so subsequent remarks about R^2 are now wrong.

The correct conclusion from (10) is that, approximately, $R^2 \leqslant \sigma_\eta^2/(\sigma_\eta^2 + \sigma^2)$, the right hand side being the limit of R^2 as N increases, if the model is correct. For a useful regression, we would want σ_η/σ which measures the ratio of the spread of the η's to the spread of the errors to be large, and we see that R^2 provides good information on this since, for large N, $\sigma_\eta/\sigma \approx R/(1-R^2)^{1/2}$.

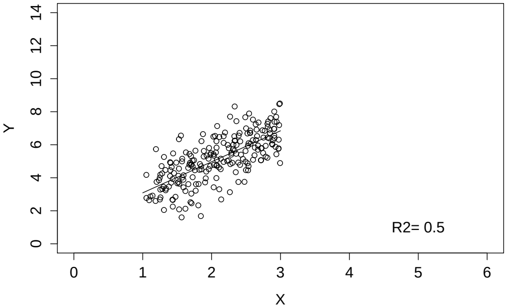
The behaviour of u and F is sensitive to N, however. From equation (11), if we assume the range of the $\hat{y}_i$ to be "about $\theta\sigma_\eta$" and with $s \approx \sigma$, we find $u \approx N^{1/2}\theta(\sigma_\eta/\sigma)/\nu_m^{1/2}$. Furthermore $F \approx (N - \nu_m - 1)(\sigma_\eta^2/\sigma^2)/\nu_m$. Thus, even if σ_η/σ is small, observed values of u and F can be large for large N, and may give a misleading impression of the explanatory power of the regression. Rather than compare F with the usual percentage point, we may perhaps wish to recalibrate the test, and this is what Box and Wetz (1973) investigated for *moderate* sample sizes.

Overall then, it appears that R^2 *is* a useful indicator for large data sets whereas F can be misleading if used in the ordinary way; thus Darlington (1982) *is* right.

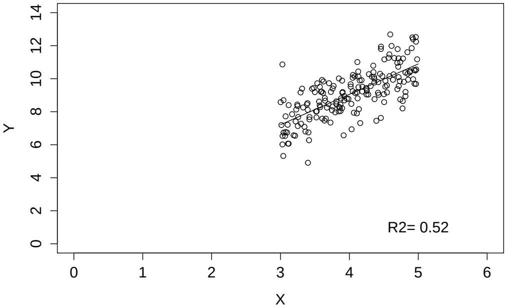
‘Bad’ properties - dependence on distribution of X



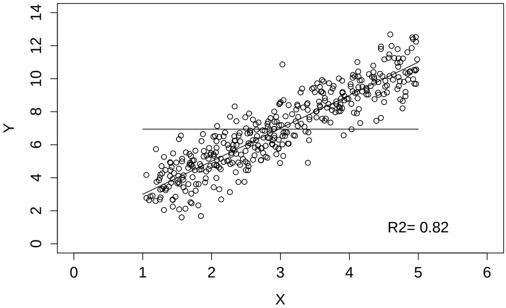
'Bad' properties - dependence on distribution of X



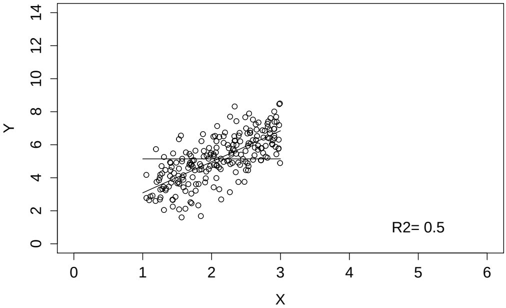
'Bad' properties - dependence on distribution of X



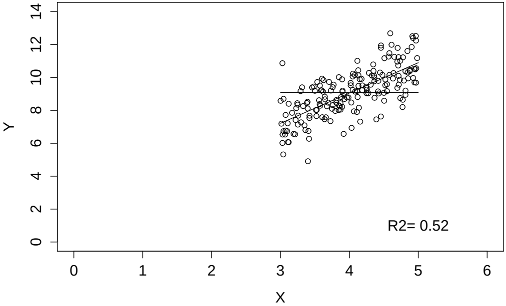
‘Bad’ properties - dependence on distribution of X



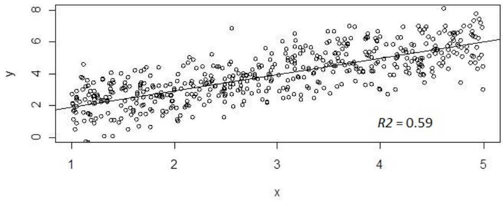
‘Bad’ properties - dependence on distribution of X

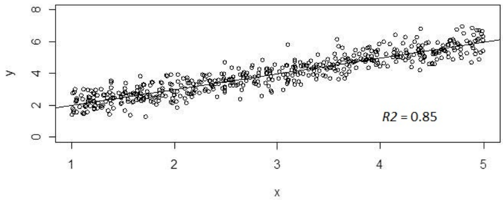


‘Bad’ properties - dependence on distribution of X

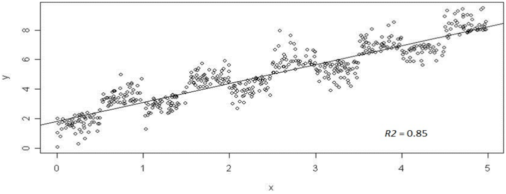


Misunderstanding - R2 is NOT a measure of fit





Misunderstanding - R2 is NOT a measure of fit



Explained variation in survival analysis

Many measures proposed under different names

- explained variation
- explained randomness
- prognostic value
- correlation
- concordance
- ...

A recent paper in SiM recommends the following measures:

A recent paper in SiM recommends the following measures:

- Kent and O'Quigley's

$$R^2 = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z) + \frac{\pi^2}{6}}$$

(This is really just an approximation to Kent & O'Quigley, but ok)

A recent paper in SiM recommends the following measures:

- Kent and O'Quigley's

$$R^2 = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z) + \frac{\pi^2}{6}}$$

(This is really just an approximation to Kent & O'Quigley, but ok)

- Royston and Sauerbrei's

$$R^2 = \frac{D^2/\sqrt{(}8/\pi)}{D^2/\sqrt{(}8/\pi) + \sigma_\epsilon^2}$$

A recent paper in SiM recommends the following measures:

- Kent and O'Quigley's

$$R^2 = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z) + \frac{\pi^2}{6}}$$

(This is really just an approximation to Kent & O'Quigley, but ok)

- Royston and Sauerbrei's

$$R^2 = \frac{D^2/\sqrt(8/\pi)}{D^2/\sqrt(8/\pi) + \sigma_\epsilon^2}$$

The others are said to perform poorly.

The recommended measures have the following properties (among others)

- They assume that the model holds everywhere
- They do not allow for covariates or effects to change in time
- They cannot be used with repeated events

The authors' recommendation is based on one criterium

The authors' recommendation is based on one criterium

Bias under censoring

The authors' recommendation is based on one criterium

Bias under censoring

(Sounds reasonable)

The usual simulations setup to study bias under censoring

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.
6. see if the value of the measure on censored data is close to the value on uncensored (on average).

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.
6. see if the value of the measure on censored data is close to the value on uncensored (on average).
7. if not - conclude **bias**.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.
6. see if the value of the measure on censored data is close to the value on uncensored (on average).
7. if not - conclude **bias**.
8. if comparing measures - choose the one with the smallest (or no) bias.

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.
6. see if the value of the measure on censored data is close to the value on uncensored (on average).
7. if not - conclude **bias**.
8. if comparing measures - choose the one with the smallest (or no) bias.

7 is wrong, so 8 may be wrong!

The usual simulations setup to study bias under censoring

1. generate covariates (decide on number - often 1, distribution - often uniform, range).
2. generate survival times, conditional on covariates under some model, often exponential.
3. fit a model, calculate a measure (or more) on this uncensored data.
4. generate censoring times, given a required proportion of censoring. Censoring distribution often **uniform**.
5. censor survival times, fit the model, calculate the measure again.
6. see if the value of the measure on censored data is close to the value on uncensored (on average).
7. if not - conclude **bias**.
8. if comparing measures - choose the one with the smallest (or no) bias.

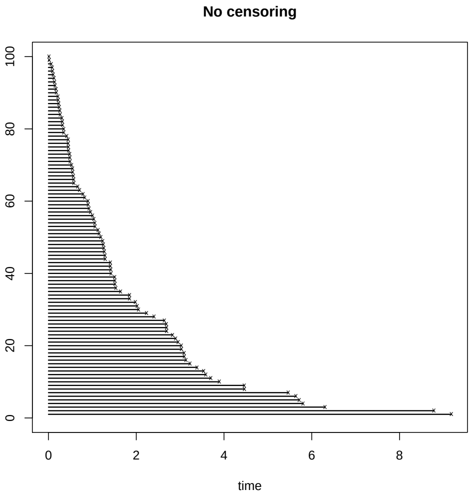
I'd better say: **7 doesn't make much sense!**

Usual reporting of simulation results in papers

Table 42 : Effect of censoring

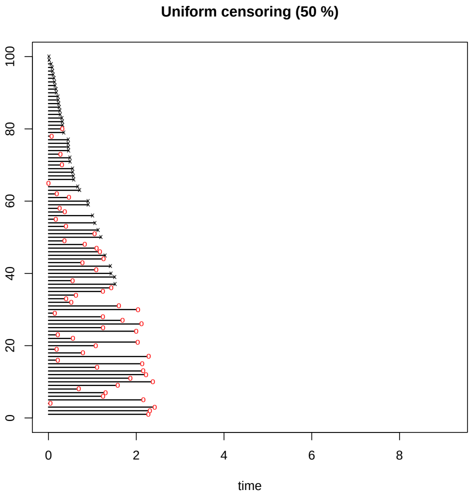
		censoring	
paper		0%	50%
1996	some measure	0.276	0.187
2011	some measure	0.139	41.8 % increase

What is going on?



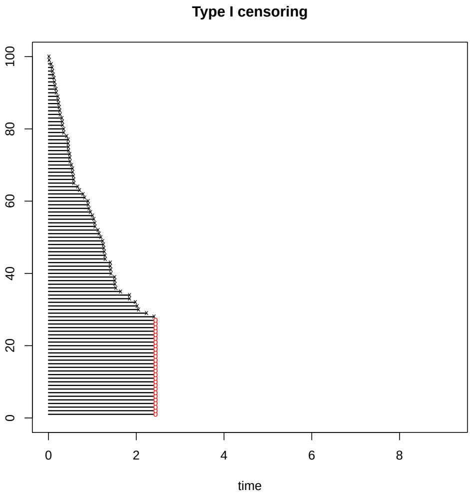
We estimate a measure on such data.

What is going on?



Why do we think that we should get the same value on such data?

What is going on?



One can study bias up to τ, nothing else!

Why 7 doesn't make sense?

$$\int_0^\infty S(t)dt$$

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \qquad \int_0^{\tau} S(t)dt$$

Why 7 doesn't make sense?

$$\int_0^\infty S(t)dt \neq \int_0^{\tau} S(t)dt$$

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \neq \int_0^{\tau} S(t)dt$$

Estimator A B

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \qquad \neq \qquad \int_0^{\tau} S(t)dt$$

$\uparrow$ unbiased

Estimator A B

Why 7 doesn’t make sense?

$$\int_0^{\infty} S(t)dt \qquad \neq \qquad \int_0^{\tau} S(t)dt$$

unbiased	unbiased

Estimator A B

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \neq \int_0^{\tau} S(t)dt$$

biased

unbiased

unbiased

Estimator A B

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \neq \int_0^{\tau} S(t)dt$$

unbiased

biased

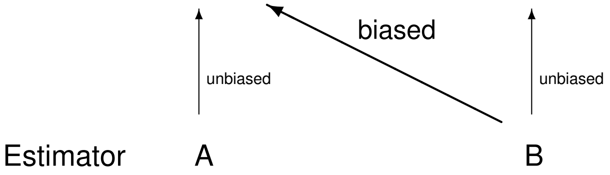
unbiased

Estimator A B

If we have a model for $S(t)$ and ASSUME that it holds everywhere, we can correct B.

Why 7 doesn't make sense?

$$\int_0^{\infty} S(t)dt \neq \int_0^{\tau} S(t)dt$$



If we have a model for $S(t)$ and ASSUME that it holds everywhere, we can correct B.
Some measures inherently assume that the model holds everywhere, so they SEEM better than the others.

Example - information gain measure

$$\rho^2 = 1 - e^{-E(LR)}, \tag{1}$$

where $E(LR)$ is

$$E(LR) = 2 \int_{D(x)} \int_0^\infty \log \frac{f_M(t|x)}{f_0(t)} dF_M(t|x) dG(x) \tag{2}$$

Example - information gain measure

$$\rho^2 = 1 - e^{-E(LR)}, \tag{1}$$

where $E(LR)$ is

$$E(LR) = 2 \int_{D(x)} \int_0^\infty \log \frac{f_M(t|x)}{f_0(t)} dF_M(t|x) dG(x) \tag{2}$$

If there is a time τ beyond which all the observations are censored, we CANNOT estimate (2).

Example - information gain measure

$$\rho^2 = 1 - e^{-E(LR)}, \tag{1}$$

where $E(LR)$ is

$$E(LR) = 2 \int_{D(x)} \int_0^\infty \log \frac{f_M(t|x)}{f_0(t)} dF_M(t|x) dG(x) \tag{2}$$

If there is a time τ beyond which all the observations are censored, we CANNOT estimate (2).

For parametric models, we replace the densities with values of $S(\tau|x)$, for the Cox model we drop them from the likelihood.

Example - information gain measure

$$\rho^2 = 1 - e^{-E(LR)}, \tag{1}$$

where $E(LR)$ is

$$E(LR) = 2 \int_{D(x)} \int_0^\infty \log \frac{f_M(t|x)}{f_0(t)} dF_M(t|x) dG(x) \tag{2}$$

If there is a time τ beyond which all the observations are censored, we CANNOT estimate (2).

For parametric models, we replace the densities with values of $S(\tau|x)$, for the Cox model we drop them from the likelihood.

Neither helps (of course)!

Example - information gain measure

$$\rho^2 = 1 - e^{-E(LR)}, \tag{1}$$

where $E(LR)$ is

$$E(LR) = 2 \int_{D(x)} \int_0^\infty \log \frac{f_M(t|x)}{f_0(t)} dF_M(t|x) dG(x) \tag{2}$$

If there is a time τ beyond which all the observations are censored, we CANNOT estimate (2).

For parametric models, we replace the densities with values of $S(\tau|x)$, for the Cox model we drop them from the likelihood.

Neither helps (of course)!

So, again, why do we think we should estimate (2) with observations limited to τ?

Example continued - estimation of information gain measure

Example continued - estimation of information gain measure

For the Cox model we used the following estimator:

$$\hat{E}(LR) = \frac{2}{k} \sum_{i=1}^{k} (L\hat{R})_i.$$

Example continued - estimation of information gain measure

For the Cox model we used the following estimator:

$$\hat{E}(LR) = \frac{2}{k}\sum_{i=1}^{k}(L\hat{R})_i.$$

So k (number of events), **NOT** n (number of cases)!

Example continued - estimation of information gain measure

For the Cox model we used the following estimator:

$$\hat{E}(LR) = \frac{2}{k}\sum_{i=1}^{k}(L\hat{R})_i.$$

So k (number of events), **NOT** n (number of cases)!
We imputed values for censored observations after t_{max}.

Example continued - estimation of information gain measure

For the Cox model we used the following estimator:

$$\hat{E}(LR) = \frac{2}{k} \sum_{i=1}^{k} (\hat{LR})_i .$$

So k (number of events), **NOT** n (number of cases)!
We imputed values for censored observations after t_{max}.

For parametric models the story is different.

Some results on Information gain measure

Table 43 : Cox model

% censored	ρ_i^2	se	ρ^2	se
90	0.43	0.12	0.44	0.04
80	0.43	0.08	0.44	0.04
70	0.44	0.05	0.44	0.04
60	0.43	0.05	0.43	0.04
50	0.45	0.04	0.45	0.04
40	0.43	0.04	0.43	0.04
30	0.43	0.04	0.43	0.04
20	0.44	0.04	0.44	0.04
10	0.44	0.04	0.44	0.04

Things to consider

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.
2. If we have data which are censored after certain t_{max}, then we can only estimate our measures until t_{max}.

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.
2. If we have data which are censored after certain t_{max}, then we can only estimate our measures until t_{max}.
3. Our estimates should be unbiased (or at least consistent) until t_{max}.

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.
2. If we have data which are censored after certain t_{max}, then we can only estimate our measures until t_{max}.
3. Our estimates should be unbiased (or at least consistent) until t_{max}.
4. When we censor with a uniform distribution, we will always have a maximum observable event time τ, everything beyond that is censored. And, BTW, using exponential will not solve the problem.

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.
2. If we have data which are censored after certain t_{max}, then we can only estimate our measures until t_{max}.
3. Our estimates should be unbiased (or at least consistent) until t_{max}.
4. When we censor with a uniform distribution, we will always have a maximum observable event time τ, everything beyond that is censored. And, BTW, using exponential will not solve the problem.
5. If some measure pretends to be (or looks) unbiased on the whole support of T, then we should be suspicious.

Things to consider

1. We need to distinguish between censoring BEFORE t_{max} and after t_{max} where t_{max} is the last observed event time.
2. If we have data which are censored after certain t_{max}, then we can only estimate our measures until t_{max}.
3. Our estimates should be unbiased (or at least consistent) until t_{max}.
4. When we censor with a uniform distribution, we will always have a maximum observable event time τ, everything beyond that is censored. And, BTW, using exponential will not solve the problem.
5. If some measure pretends to be (or looks) unbiased on the whole support of T, then we should be suspicious.
6. All measures (?) can be made to look unbiased on the whole support of T.

Criteria for choosing R^2

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?
- Time dependent covariates and effects.

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?
- Time dependent covariates and effects.
- Comparability across models (includes applicability for different models). Null model?

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?
- Time dependent covariates and effects.
- Comparability across models (includes applicability for different models). Null model?
- What about repeated events, competed risks, multi-state models?

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?
- Time dependent covariates and effects.
- Comparability across models (includes applicability for different models). Null model?
- What about repeated events, competed risks, multi-state models?
- Interpretation (explained variation?).

Criteria for choosing R^2

- Consistency (unbiasedness) up to τ. Most (all?) measures can be adjusted to meet this criterium.
- Expected value. And add variance of the estimator as a bonus.
- Scale? Approaching 1?
- Time dependent covariates and effects.
- Comparability across models (includes applicability for different models). Null model?
- What about repeated events, competed risks, multi-state models?
- Interpretation (explained variation?).

Once we satisfy these criteria, not many measures will be left ...

... but at least one will

BIOMETRICS 67, 750–759
September 2011

DOI: 10.1111/j.1541-0420.2010.01526.x

A Measure of Explained Variation for Event History Data

Janez Stare,[1,*] Maja Pohar Perme,[1] and Robin Henderson[2]

[1]Department of Biostatistics and Medical Informatics, University of Ljubljana, Vrazov trg 2, SI-1000 Ljubljana, Slovenia
[2]Mathematics & Statistics, Newcastle University, U.K.
[*]*email:* janez.stare@mf.uni-lj.si

Explained risk ranking

Rank of event under model $r^{(M)}(t)$

Explained risk ranking

Rank of event under model	$r^{(M)}(t)$
Rank of event under perfect knowledge	$r^{(P)}(t)$

Explained risk ranking

Rank of event under model	$r^{(M)}(t)$
Rank of event under perfect knowledge	$r^{(P)}(t)$
Rank of event without model	$r^{(0)}(t)$

Explained risk ranking

Rank of event under model	$r^{(M)}(t)$
Rank of event under perfect knowledge	$r^{(P)}(t)$
Rank of event without model	$r^{(0)}(t)$

Difference $r^{(0)}(t) - r^{(P)}(t)$ needs to be explained

Explained risk ranking

Rank of event under model	$r^{(M)}(t)$
Rank of event under perfect knowledge	$r^{(P)}(t)$
Rank of event without model	$r^{(0)}(t)$

Difference $r^{(0)}(t) - r^{(P)}(t)$ needs to be explained
Difference $r^{(0)}(t) - r^{(M)}(t)$ *is* explained

Explained risk ranking

Rank of event under model	$r^{(M)}(t)$
Rank of event under perfect knowledge	$r^{(P)}(t)$
Rank of event without model	$r^{(0)}(t)$

Difference $r^{(0)}(t) - r^{(P)}(t)$ needs to be explained
Difference $r^{(0)}(t) - r^{(M)}(t)$ *is* explained

$$R_E \simeq \frac{\sum_t (r^{(0)}(t) - r^{(M)}(t))}{\sum_t (r^{(0)}(t) - r^{(P)}(t))}$$

Our proposal

$$R_E = \frac{\sum_i \int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}{\sum_i \int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}$$

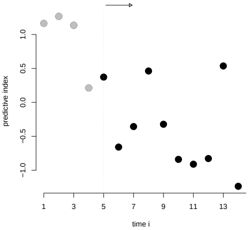
Our proposal

$$R_E = \frac{}{r_i^{(0)}(t) - r_i^{(P)}(t)}$$

perfect

continuous variable

$\sharp$ at risk	10
$\text{rank}_{\text{null}}$	5.5
$\text{rank}_{\text{perfect}}$	1



Our proposal

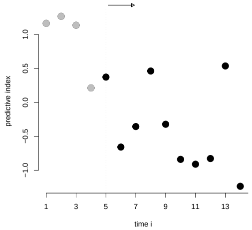
observed

$$R_E = \frac{r_i^{(0)}(t) - r_i^{(M)}(t)}{r_i^{(0)}(t) - r_i^{(P)}(t)}$$

perfect

continuous variable

♯ at risk	10
$\text{rank}_{\text{null}}$	5.5
$\text{rank}_{\text{perfect}}$	1
$\text{rank}_{\text{model}}(t_5)$	3

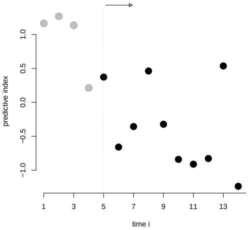


Our proposal

$$R_E = \frac{\int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right) \right\} dN_i(t)}{\int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right) \right\} dN_i(t)}$$

continuous variable

$\sharp$ at risk	10
$\text{rank}_{\text{null}}$	5.5
$\text{rank}_{\text{perfect}}$	1
$\text{rank}_{\text{model}}(t_5)$	3

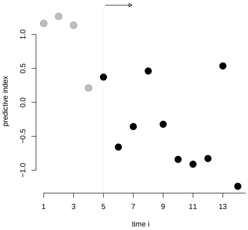


Our proposal

$$R_E = \frac{\int_0^{\tau} \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}{\int_0^{\tau} \left\{ \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}$$

continuous variable

♯ at risk	10
$\mathrm{rank}_{\mathrm{null}}$	5.5
$\mathrm{rank}_{\mathrm{perfect}}$	1
$\mathrm{rank}_{\mathrm{model}}(t_5)$	3

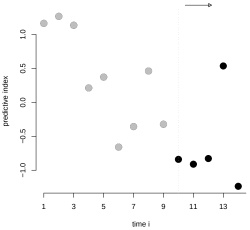


Our proposal

$$R_E = \frac{\sum_i \int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}{\sum_i \int_0^\tau \left\{ \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right) / \hat{G}_{it} \right\} dN_i(t)}$$

continuous variable

$\sharp$ at risk	5
$\text{rank}_{\text{null}}$	3
$\text{rank}_{\text{perfect}}$	1
$\text{rank}_{\text{model}}(t_{10})$	3



Variance of R_E - in short...

$$\hat{\mathrm{Var}}(R_E) =$$

$$\frac{1}{n^2 Q_2(\tau)^2} \int_0^\tau \sum_i \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right)^2 / \hat{G}_{it}^2 \right\} \hat{\lambda}_i(t|\mathscr{F}_{t-}) dt$$

$$- \frac{2 Q_1(\tau)}{n^2 Q_2(\tau)^3} \int_0^t \sum_i \left\{ \left(r_i^{(0)}(t) - r_i^{(M)}(t) \right) \right.$$

$$\left. \times \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right) / \hat{G}_{it}^2 \right\} \hat{\lambda}_i(t|\mathscr{F}_{t-}) dt$$

$$+ \frac{Q_1(\tau)^2}{n^2 Q_2(\tau)^4} \int_0^\tau \sum_i \left\{ \left(r_i^{(0)}(t) - r_i^{(P)}(t) \right)^2 / \hat{G}_{it}^2 \right\} \hat{\lambda}_i(t|\mathscr{F}_{t-}) dt$$

Population value

$$\mathcal{R}_E = 2\frac{E\left\{\int_0^\tau S(t)H_t(u)\alpha(t|\mathscr{F}_{t-},u)dt\right\}}{E\left\{\int_0^\tau S(t)\alpha(t|\mathscr{F}_{t-},u)dt\right\}} - 1,$$

Special cases, uncensored Cox survival, univariate X

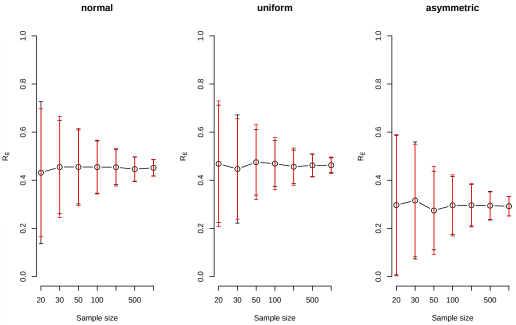
$$\mathcal{R}_E = 1 - \frac{4}{\beta}\int_0^1 \log\left(\frac{2e^{\beta x}}{1+e^{\beta x}}\right)dx \qquad X \sim U(0,1)$$

$$\mathcal{R}_E = \frac{1}{2}\left(\frac{e^\beta - 1}{e^\beta + 1}\right) \qquad X \sim Ber(0.5)$$

Some simulations

⊢⊣ variance of the R_E estimates
⊢⊣ average of variance estimates

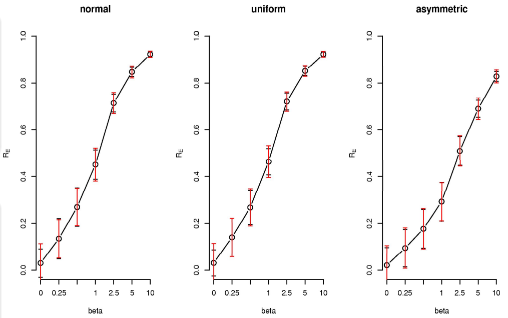
sample size: $20 \rightarrow 1000$
$\beta = 1$
no censoring
100 simulation runs



Some simulations

$\vdash\!\dashv$ variance of the R_E estimates
$\color{red}{\vdash\!\dashv}$ average of variance estimates

β : $0 \rightarrow 10$
sample size 250
no censoring
100 simulation runs



Properties

1. the null model IS always the same

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects
3. explained variation interpretation?

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects
3. explained variation interpretation?
4. scale? Is the difference between 0.5 and 0.45 the same as between 0.7 and 0.65?

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects
3. explained variation interpretation?
4. scale? Is the difference between 0.5 and 0.45 the same as between 0.7 and 0.65?
5. approaching 1?

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects
3. explained variation interpretation?
4. scale? Is the difference between 0.5 and 0.45 the same as between 0.7 and 0.65?
5. approaching 1?
6. dependency on censoring?

Properties

1. the null model IS always the same
2. R_E CAN be used with time dependent covariates or effects
3. explained variation interpretation?
4. scale? Is the difference between 0.5 and 0.45 the same as between 0.7 and 0.65?
5. approaching 1?
6. dependency on censoring?
7. CAN be used with any model

Some more properties

- between 0 and 1

Some more properties

- between 0 and 1
- simple interpretation

Some more properties

- $\smile$ between 0 and 1
- $\smile$ simple interpretation
- $\smile$ computationally simple

Some more properties

- between 0 and 1
- simple interpretation
- computationally simple
- partial R_E can be easily computed

Some more properties

- between 0 and 1
- simple interpretation
- computationally simple
- partial R_E can be easily computed
- only ranks are important

Some more properties

- between 0 and 1
- simple interpretation
- computationally simple
- partial R_E can be easily computed
- only ranks are important

Some more properties

- between 0 and 1
- simple interpretation
- computationally simple
- partial R_E can be easily computed
- only ranks are important
- For single episodes with time-constant effects and covariates

$$R_E = 2(c - 1/2)$$

Real data examples - Cox model

Real data examples - Cox model

Freireich data - remission times of leukaemia patients

42 patients, 12 events
covariate: treatment (binary: 21/21)

$$R_E = 0.374$$

Real data examples - Cox model

Freireich data - remission times of leukaemia patients

42 patients, 12 events
covariate: treatment (binary: 21/21)

$$R_E = 0.374$$

PBC data - Mayo clinic primary biliary cirrhosis data

312 patients, 125 events
covariates: age, log bilirubin, log albumin, presence of edema, log blood clotting time

$$R_E = 0.580$$

Real data examples - Cox model

Freireich data - remission times of leukaemia patients

42 patients, 12 events
covariate: treatment (binary: 21/21)

$R_E = 0.374$

PBC data - Mayo clinic primary biliary cirrhosis data

312 patients, 125 events
covariates: age, log bilirubin, log albumin, presence of edema, log blood clotting time

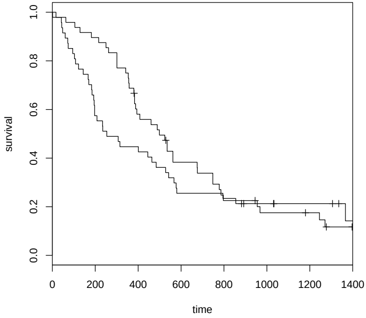
$R_E = 0.580$

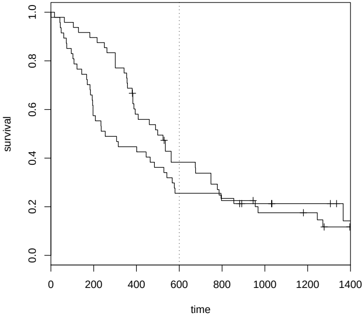
Stablein data

95 patients, 78 events
covariate: rx (binary: 48/47)

$R_E = 0.127$

Stablein data - improving the fit





$\beta = 0.22$

$\beta = 0.49 \quad \beta = -1.20$

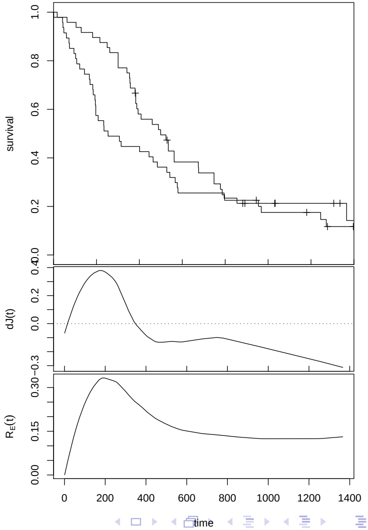
$R_E = 0.127$

$R_E = 0.173$

Measure in time - Stablein data

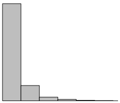
95 patients, 78 events
covariate: rx
binary: 48/47

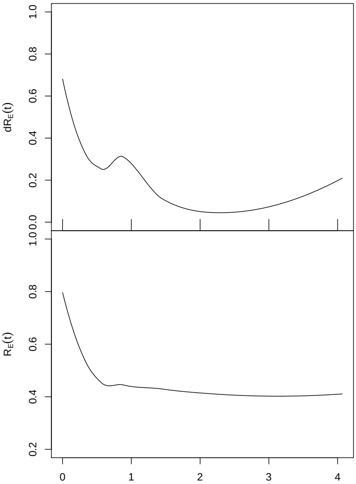
$$R_E = 0.127$$



Measure in time - asymmetric Z

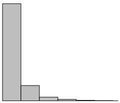
$Z \sim \Gamma(\frac{1}{2}, 1)$
$\beta = 1$
$R_E = 0.42$

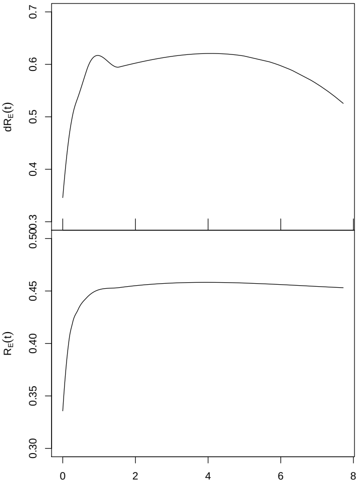




Measure in time - asymmetric Z

$Z \sim \Gamma(\frac{1}{2}, 1)$
$\beta = -1$
$R_E = 0.45$





Instead of conclusions

Instead of conclusions

- Can we do better than R_E?

Instead of conclusions

- Can we do better than R_E?
- Hopefully. (ask me in a year)

Explained variation interpretation

What is variation?

Explained variation interpretation

What is variation?

Any measure of the extent to which a distribution is not degenerate (Nagelkerke).

Explained variation interpretation

What is variation?

Any measure of the extent to which a distribution is not degenerate (Nagelkerke).

$$\left(r_i^{(0)}(t) - r_i^{(P)}(t)\right) = \left(r_i^{(0)}(t) - r_i^{(M)}(t)\right) + \left(r_i^{(M)}(t) - r_i^{(P)}(t)\right).$$

Scale

$$R_E := \frac{\sum_{i=1}^{n} [\mathrm{r}_{\mathrm{null}}(\mathrm{t_i}) - \mathrm{r}_{\mathrm{model}}(\mathrm{t_i})]}{\sum_{i=1}^{n} [\mathrm{r}_{\mathrm{null}}(\mathrm{t_i}) - 1]} = \frac{\sum_{i=1}^{n} \mathrm{r}_{\mathrm{null}}(\mathrm{t_i})}{\sum_{i=1}^{n} [\mathrm{r}_{\mathrm{null}}(\mathrm{t_i}) - 1]} - \frac{\sum_{i=1}^{n} \mathrm{r}_{\mathrm{model}}(\mathrm{t_i})}{\sum_{i=1}^{n} [\mathrm{r}_{\mathrm{null}}(\mathrm{t_i}) - 1]}$$

$$= a + b \sum_{i=1}^{n} \mathrm{r}_{\mathrm{model}}(\mathrm{t_i})$$

R_E is simply a linear function of the sum of predicted conditional ranks of failures!

Example - Cox model

Binary covariate

$$E(R_E) = 2p(1-p)\frac{e^{\beta x_1} - e^{\beta x_2}}{e^{\beta x_1} + e^{\beta x_2}} \stackrel{\beta\to\infty}{\to} 2p(1-p)$$

$\beta = 5.03$ $\quad$ $R_E = 0.493, \; E(R_E) = 0.493$

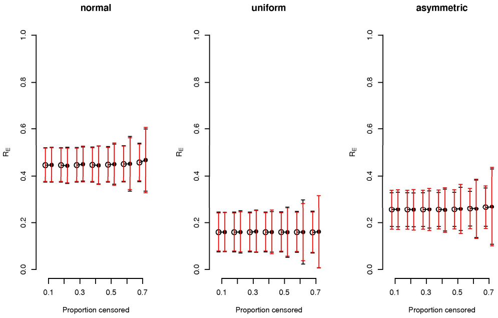
Example - Cox model

Uniform covariate

$\hat{\beta}$	R_E
0.8	0.141
1.9	0.284
3.0	0.435
3.9	0.502
5.0	0.575
6.9	0.667
10.1	0.762
13.8	0.810
18.8	0.863
25.1	0.890

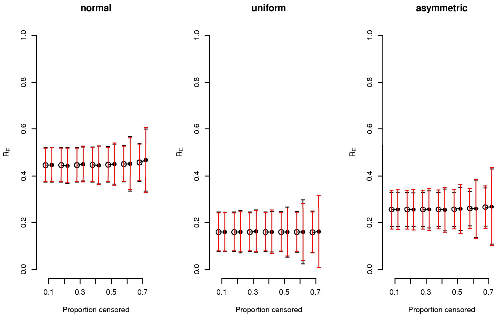
The effect of censoring before the last observed failure

$\beta = 1$
sample size 250
100 runs $Z \sim \mathcal{N}(0, 1)$
$Z \sim Unif[0, 1]$
$Z \sim asymmetric$
$\circ$ = censored data
$\bullet$ = uncensored data



The effect of censoring before the last observed failure

$\beta = 1$
sample size 250
100 runs $Z \sim \mathcal{N}(0, 1)$
$Z \sim Unif[0, 1]$
$Z \sim asymmetric$
$\circ$ = censored data
$\bullet$ = uncensored data

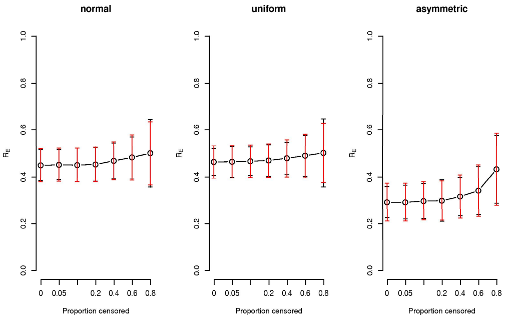


Conclusion

R_E does not depend on censoring before the last observed failure

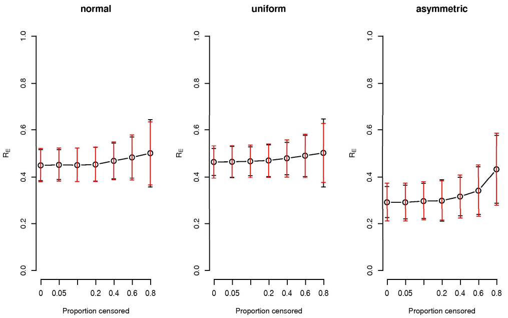
The effect of censoring at a given time

$\beta = 1$
sample size 250
100 runs $Z \sim \mathcal{N}(0, 1)$
$Z \sim Unif[0, 1]$
$Z \sim asymmetric$



The effect of censoring at a given time

$\beta = 1$
sample size 250
100 runs $Z \sim \mathcal{N}(0, 1)$
$Z \sim Unif[0, 1]$
$Z \sim asymmetric$



Conclusion

The effect of censoring at a given time depends on the covariate distribution. Possible solution: imputation.