Thoughts on validation

Terry Therneau

Oct 2024

Validation

- Software validation. This is discussed in the validation vignette.
- Internal validation: checks that further examine the fit of a model, using the original data set. These form an extension of the familiar trio of functional form, proportional hazards, and undue leverage which have been discussed elsewhere.
- External validation. Application of a model to new data set, one which was not used in the model's construction.

What does it mean to validate?

"If you don't know where you are going, you might end up someplace else.'' – Yogi Berra

Question: Is a chosen model M applicable outside of the data set on which it was developed?

- Applicable to what?
  - Korn and Simon (1990), Measures of explained variation for survival data
  - Altman and Royston(2000), What does it mean to validate a model
- Say $(t_i - \hat{t}_i)^2$ is the measure. Are the pair (.5, 1) and (5.5, 6) the same size error? Is (5.5, 6) an error at all?
- We want to use M to select subjects with expected survival of < 6 months for referral to supportive care
  - How well it predicts someone with 1 year vs 2 year is immaterial
  - We just need a yes/no wrt 6 months
- Use M for a proposed biologic therapy
  - No anticipated effect on deaths within 1 year
  - Separation of 2, 3, 4+ important for stratification

Dimensions

- There are at least 3 possible assessments
  1. The expected number of visits to state j, $E(N_j(t))$
  2. The probability in state j, $p_j(t)$
  3. The expected total sojourn time in state j, $s_j(t)$
- Each of these can be assessed at one or more chosen times τ.
- The validation data set is subject to censoring.
- Validation metrics

Censoring methods

1. Apply standard censored data methods to the validation data, and compare the results to the target model's predictions. I will sometimes call this the "yhat vs yhat' " approach.
2. Create uncensored data specific to a chosen assessment time τ, then use standard methods for uncensored data. Two approaches are
   - Use the redistribut-to-the-right (RTTR, IPC) algorithm to reassign the case weights of those observations censored prior to τ to others, resulting in a weighted subset who status at τ is known.
   - Replace the response with pseudovalues at time τ.
3. For assessment type 1, the total number of events, we can compare the observed events to the expected number given per-subject followup.
   - Essentially a standardized mortality ratio (SMR, SIR) approach
   - If the model is correct $N(t) - \Lambda(t)$ is a martingale, even if everyone has a different cut off time.
4. Ignore censoring.

Ideal order: 3,1,2,4; in the literature 2,4,1,3

Validation metrics

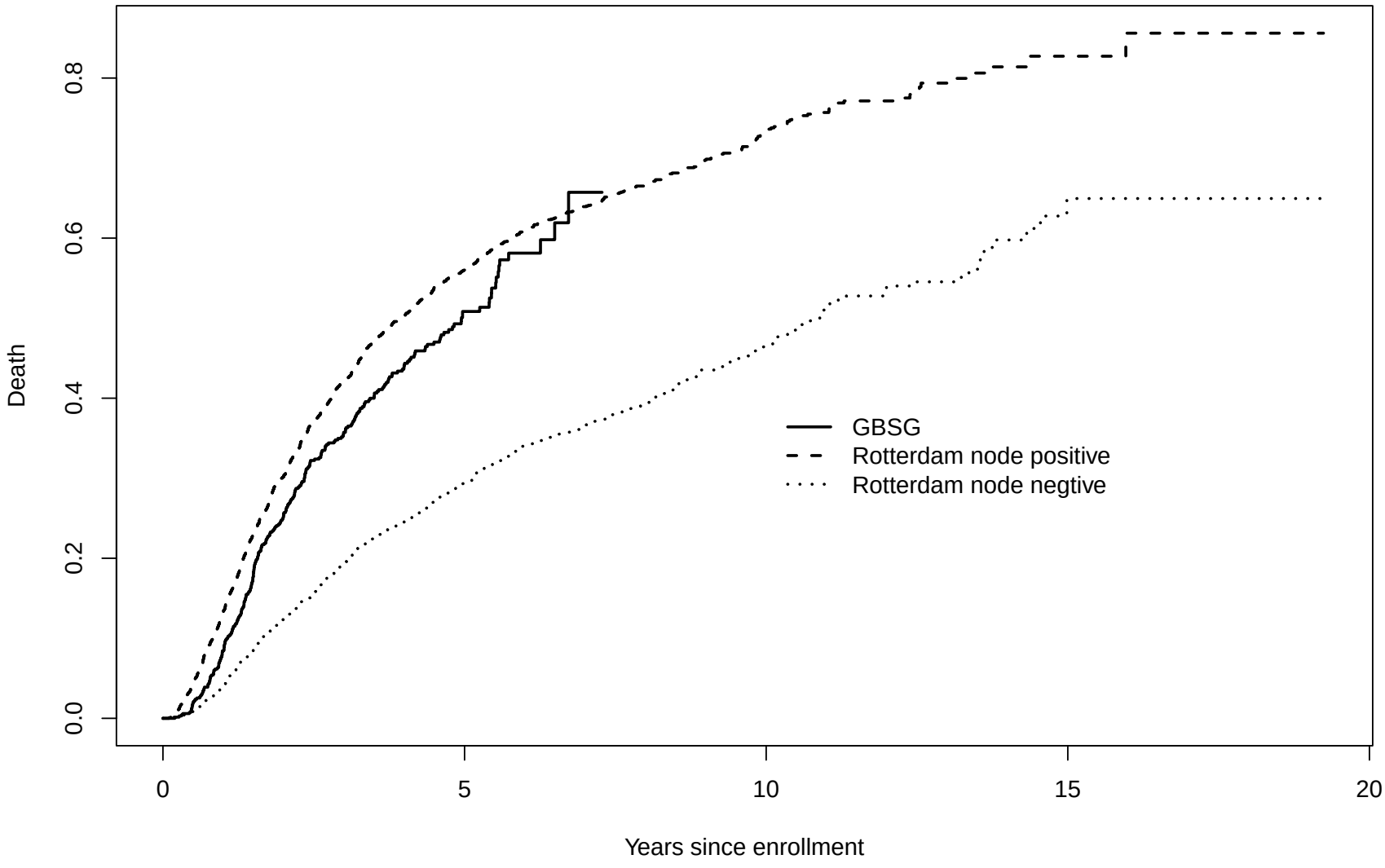
- Discrimination: are the predictions in the right order
- Calibration: are the predictions accurate, on an absolute scale
  - total is correct
  - linear rise of observed and predicted
  - pattern wrt prediction
  - responsive to individual covariates and combinations

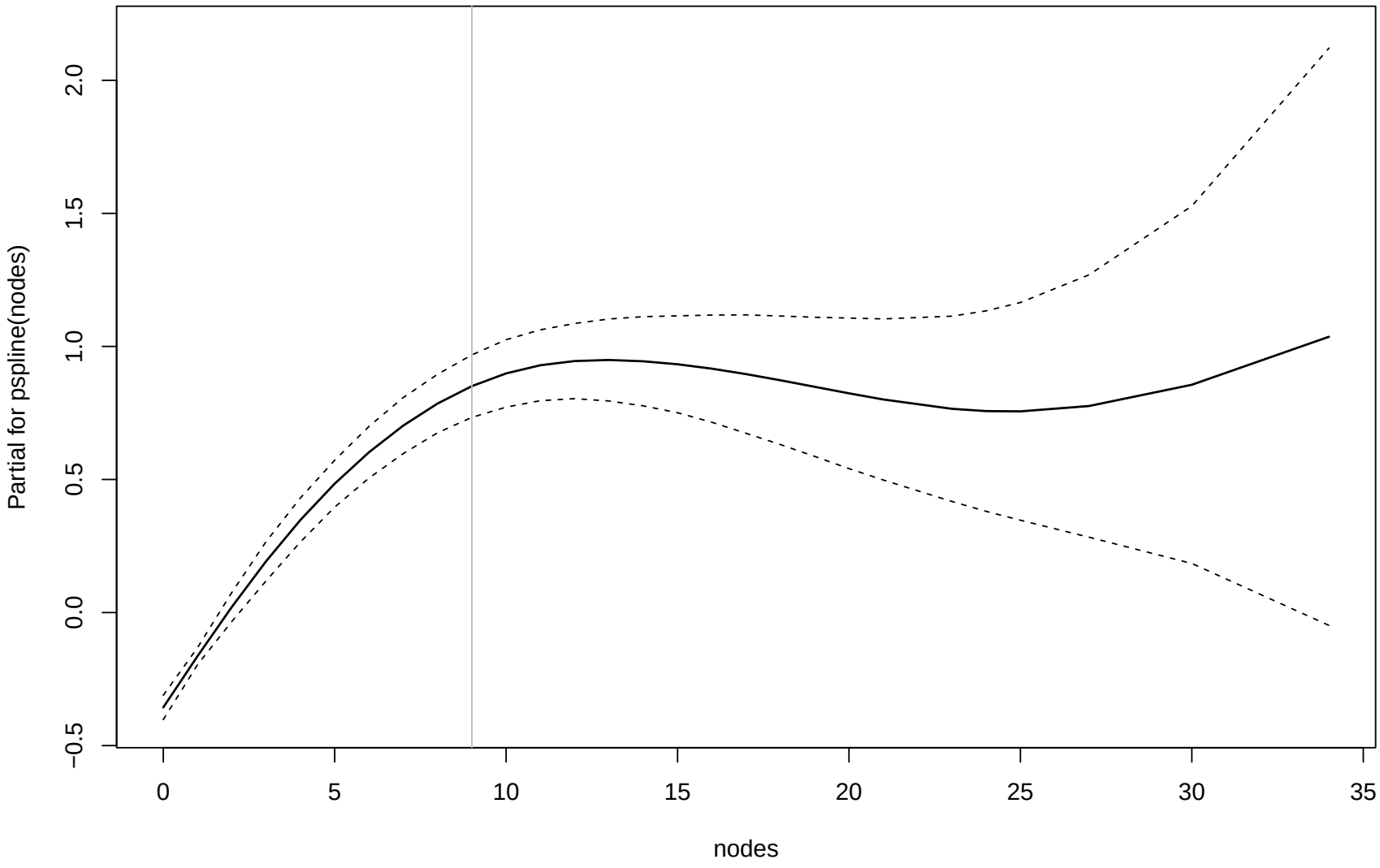
Complaint

- We know a lot about assessing binomial data (sensitivity, specificity, PPV, AUROC, . . .)
- Good literature on validating binomial data
- Literature on validating time to event data too often forces survival data onto a Procrustean bed.

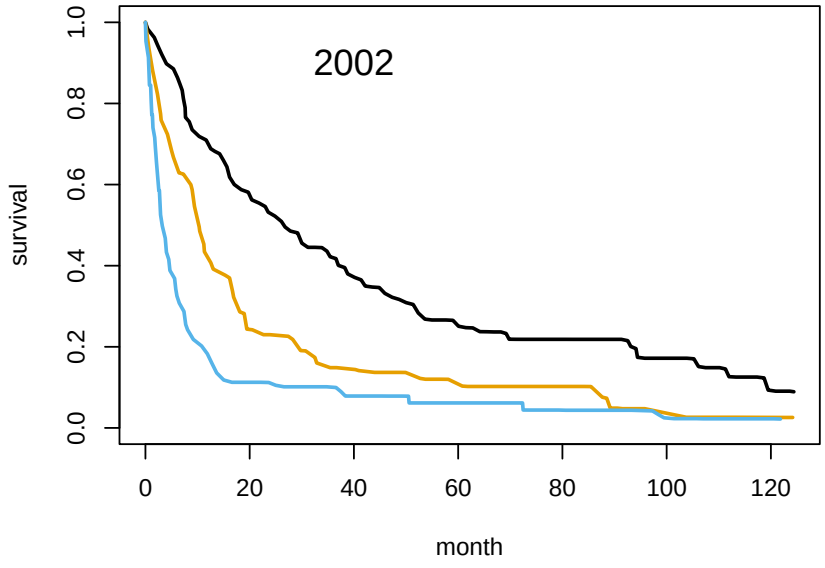
Data

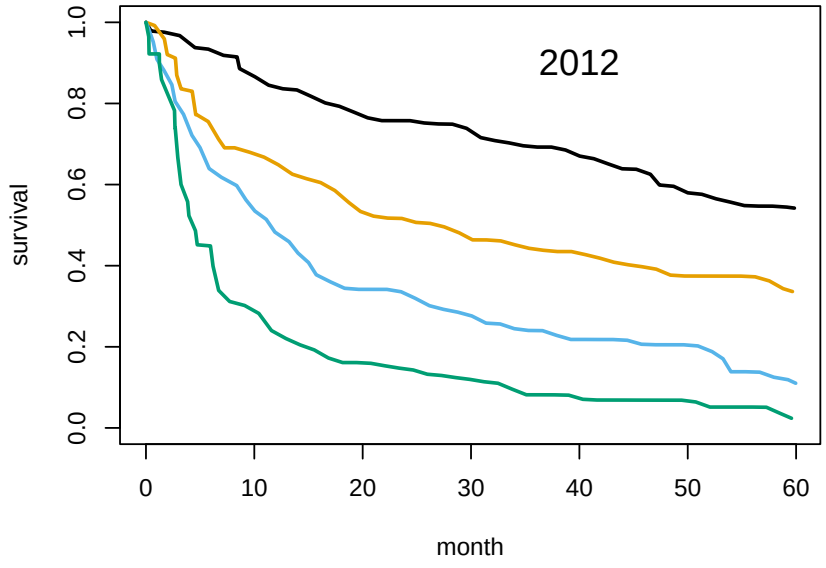
- Rotterdam: 2982 primary breast cancers recorded in the Rotterdam Tumor Bank
- GBSG: 686/720 subjects from a 1984-1989 trial conducted by the German Breast Cancer Study Group.
- Build a model on the Rotterdam data
  - ignoring GBSG
  - allow imperfection
- Validate it on GBSG

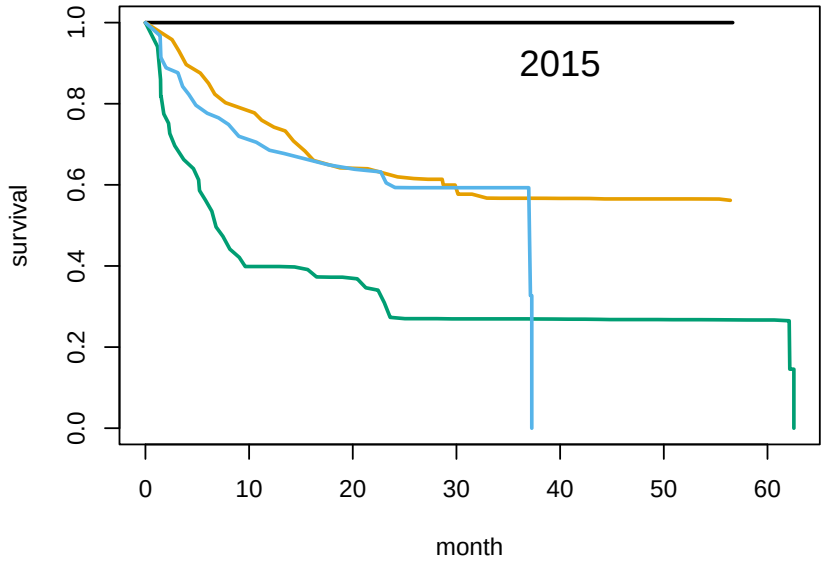

Death
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
Years since enrollment
GBSG
Rotterdam node positive
Rotterdam node negtive

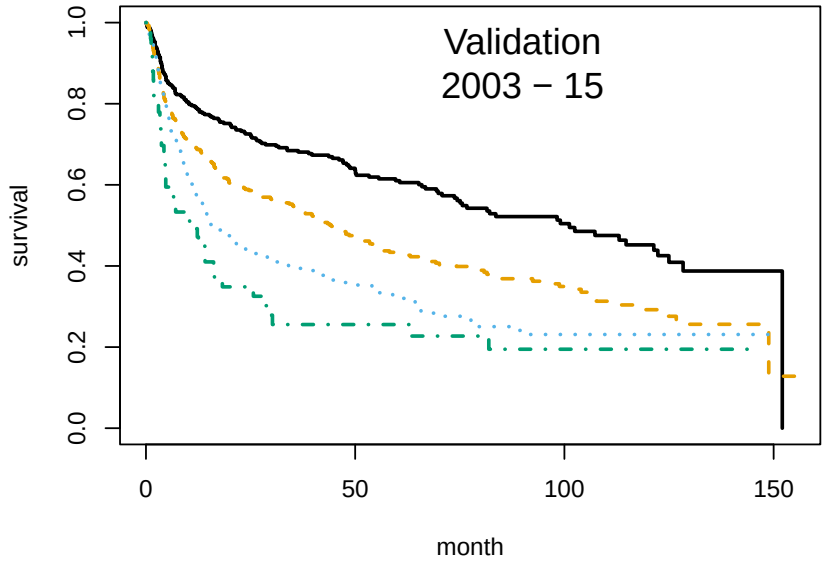

Partial for pspline(nodes)
−0.5
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
25
30
35
nodes

Amyloidosis









Concordance

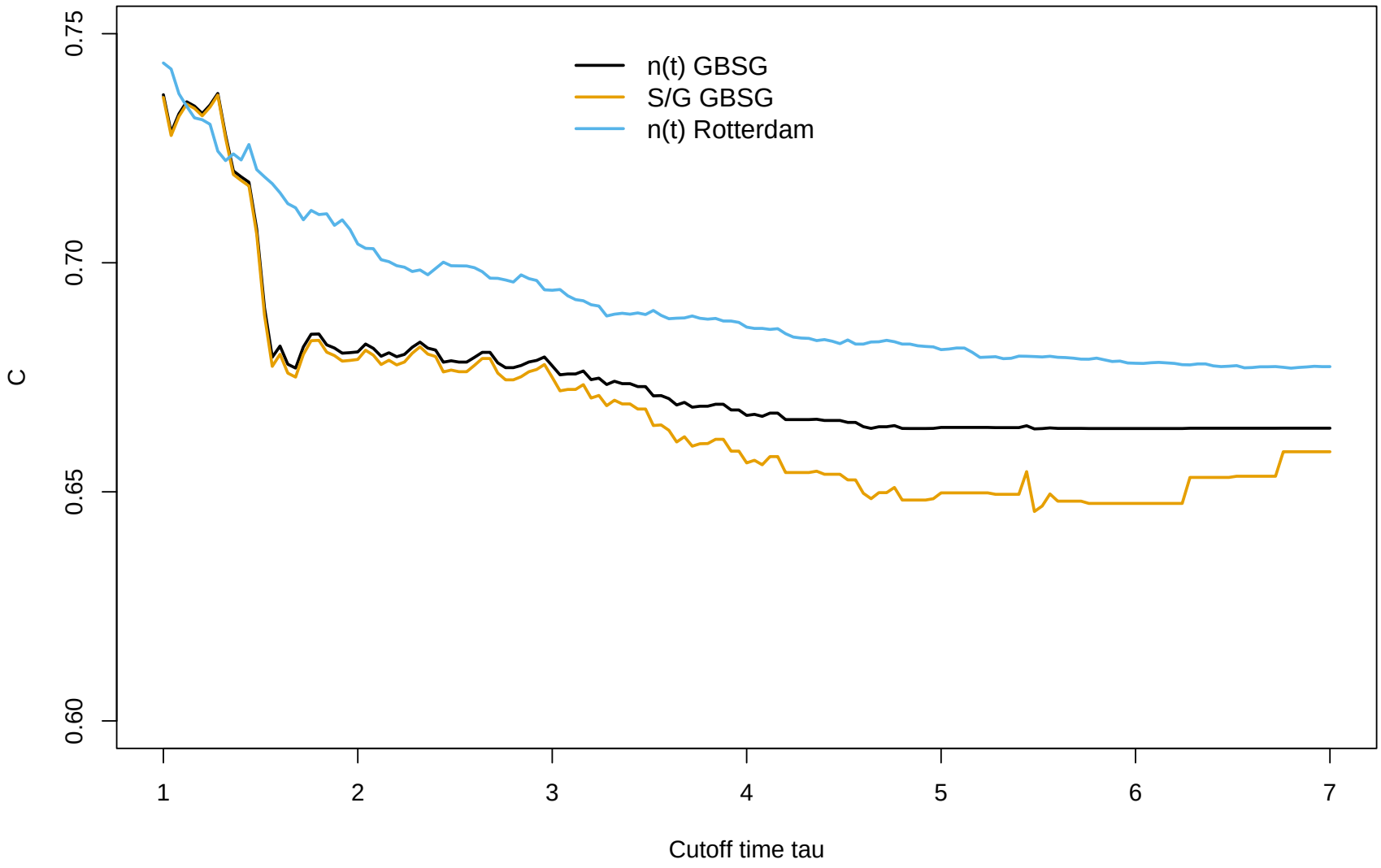
$$P(y_i > y_j | \hat{y}_i > \hat{y}_j) = P(\hat{y}_i > \hat{y}_j | y_i > y_j)$$

- $\hat{y}$ can be any of (deaths, prob, sojourn, $X\beta$) and get the same answer (for a Cox model)
- $X\beta$ does not need the intercept (baseline hazard)
- Math
  - τ_a, τ_b, γ, Somers' d are [-1, 1], differ in ties
  - C = (d+1)/2
  - If y is 0/1 then C = AUROC

The numerator of C can be written as

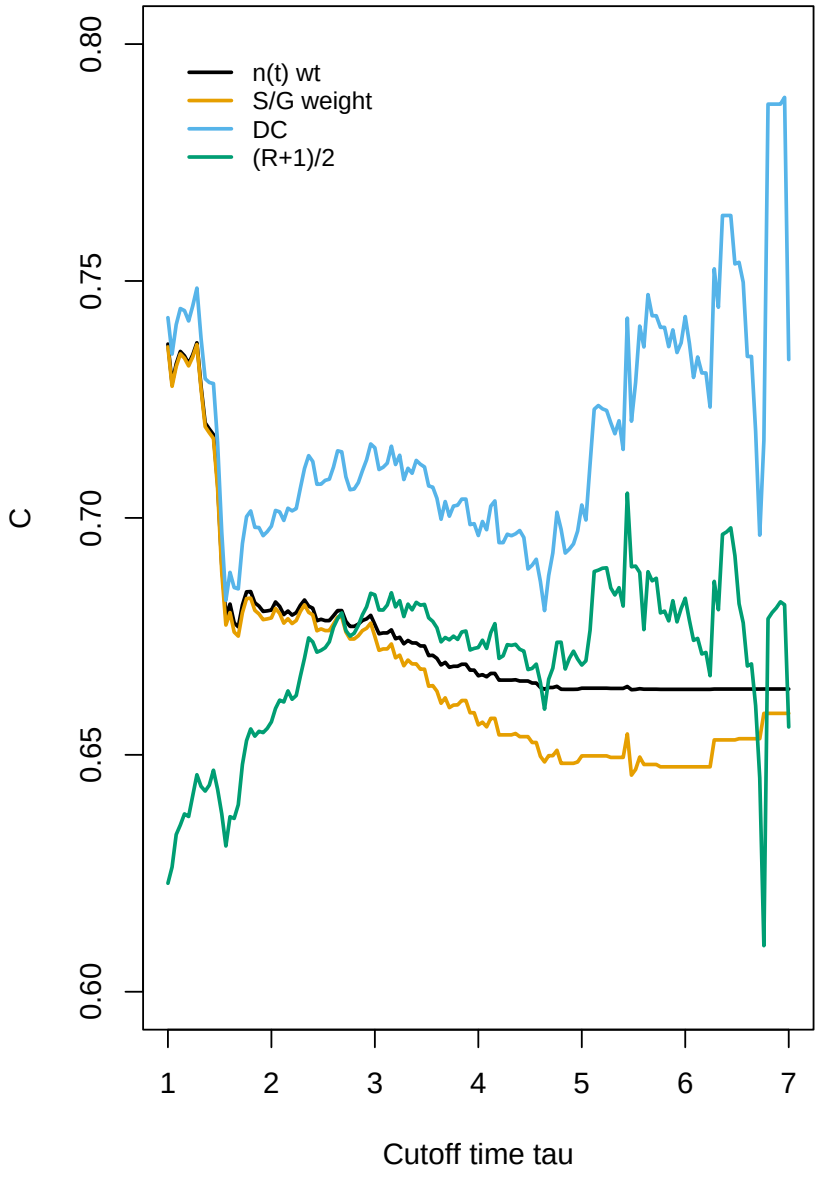
$$\sum_i \delta_i w(t)(r_i(t) - \overline{r}(t))$$

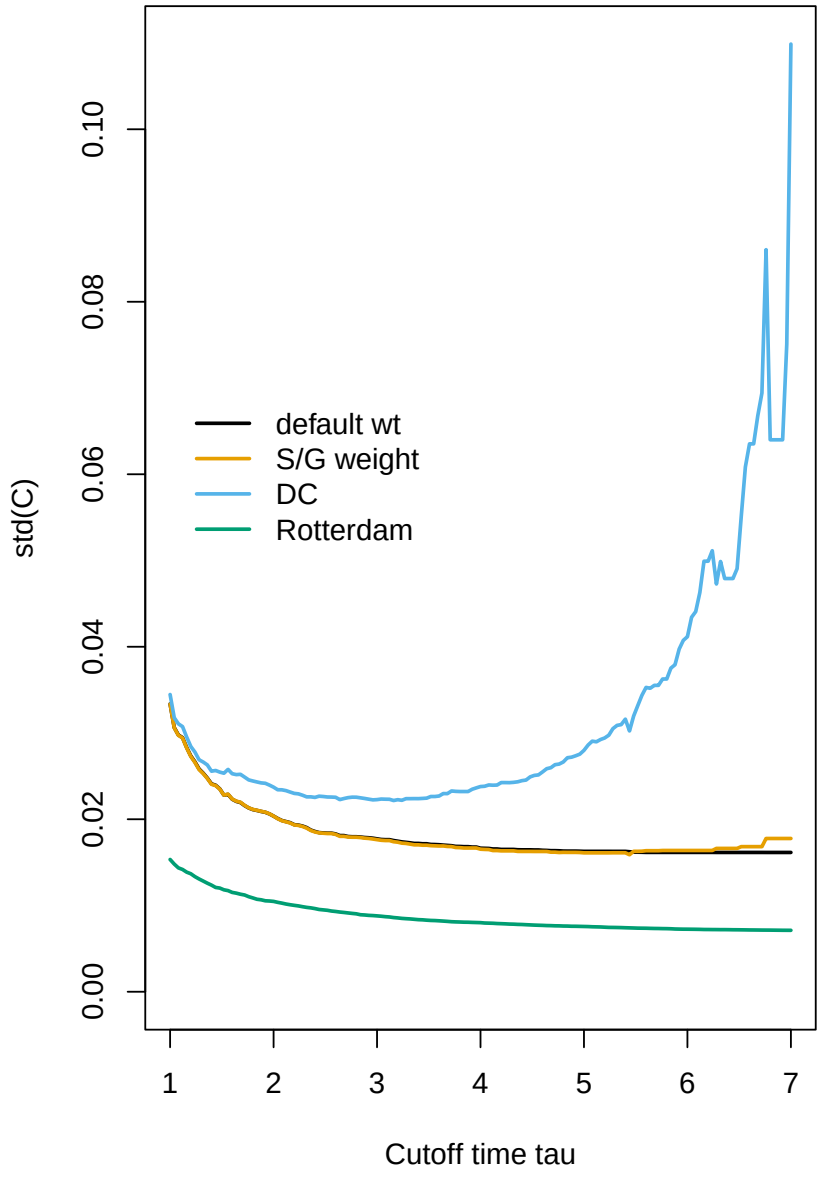
- w(t) = 1: log-rank test
- w(t)= n(t): Gehan-Wilcoxon test, Harrell C
- w(t)= S(t): Peto-Wilcoxon test
- w(t)= S(t)/G(t): Schemper test, Uno C
- ...


n(t) GBSG
S/G GBSG
n(t) Rotterdam
C
0.75
0.70
0.65
0.60
1
2
3
4
5
6
7
Cutoff time tau

Dichotomized concordance

- Use $I(t_i \leq \tau)$ rather then t_i as the response
- Does not estimate the same quantity
- Need to replace censored before τ using RTTR


n(t) wt
S/G weight
DC
(R+1)/2
C
0.60
0.65
0.70
0.75
0.80
1
2
3
4
5
6
7
Cutoff time tau

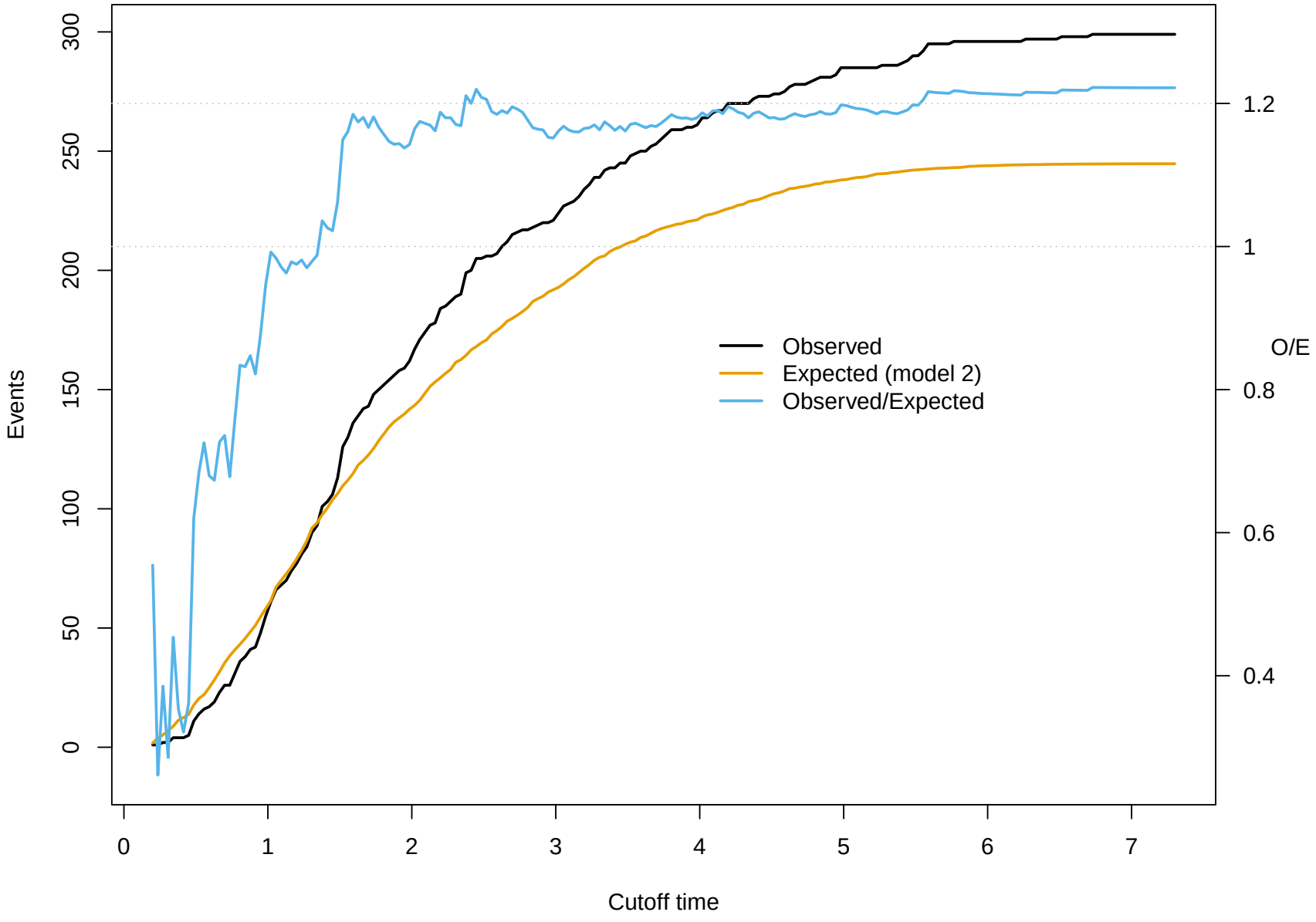

default wt
S/G weight
DC
Rotterdam
std(C)
0.00
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
7
Cutoff time tau

SMR

- R code
- data2$expect <- predict(coxfit1, type="expected", newdata=data2)
- glm(status ~ offset(log(expect)), poisson, data=data2)
- exp(intercept) = SMR
- valid estimates, std, CI
- add eta to the predictors for regression calibration

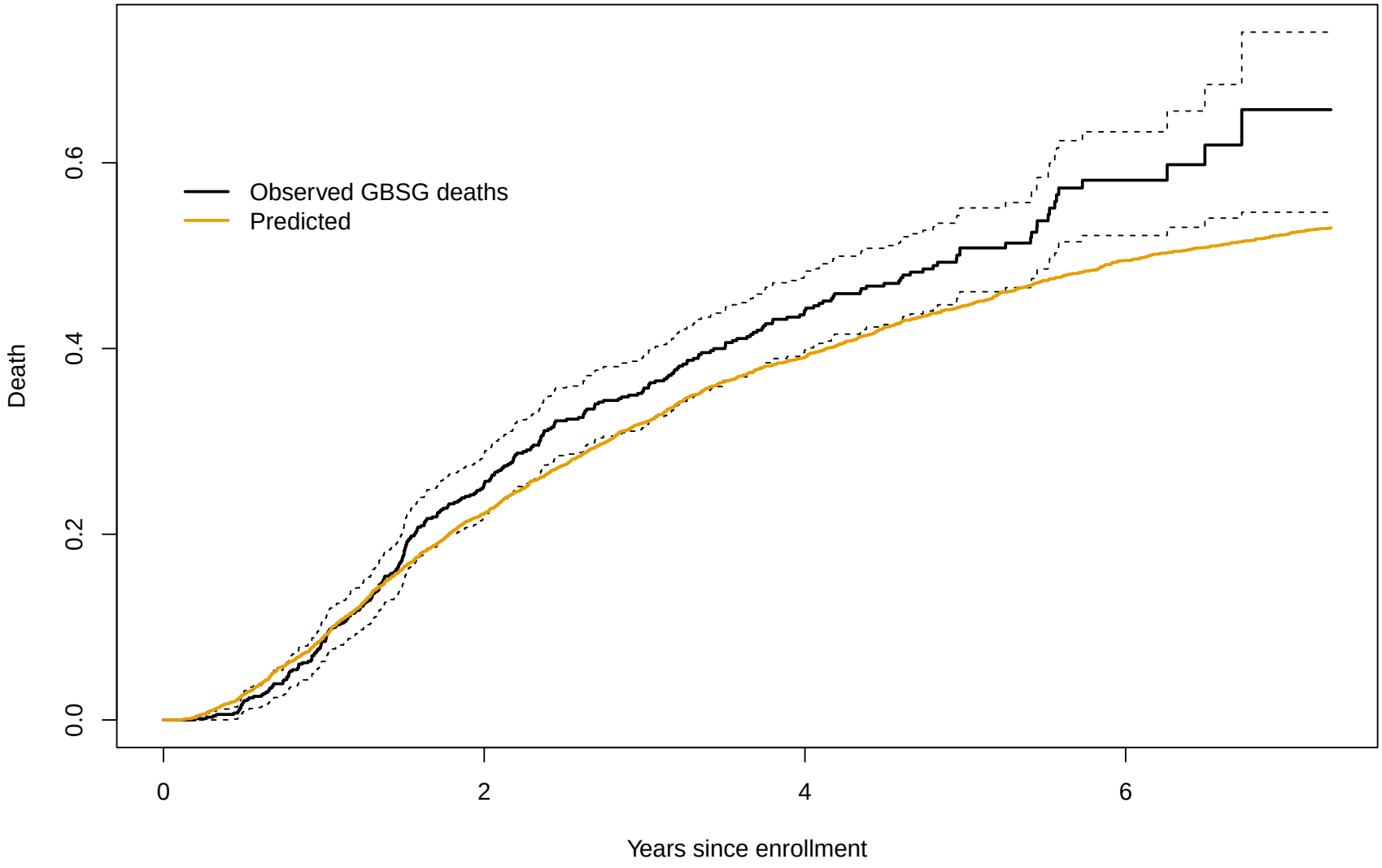
	Model 1	Model 2	Model 3
Observed	299.00	299.00	299.00
Expected	269.46	244.70	246.22
O/E	1.11	1.22	1.21

	Model 1	Model 2	Model 3
SMR	1.11	1.22	1.09
lower CI	0.99	1.09	0.97
upper CI	1.24	1.37	1.23


Events
300
250
200
150
100
50
0
O/E
1.2
1
0.8
0.6
0.4
Observed
Expected (model 2)
Observed/Expected
0
1
2
3
4
5
6
7
Cutoff time

Survival models

- Get predicted survival curve, per subject, from the fitted model
- Overall prediction vs KM (all time, all subjects)
- Per subject predictions at a given τ, vs new per-subject predictions

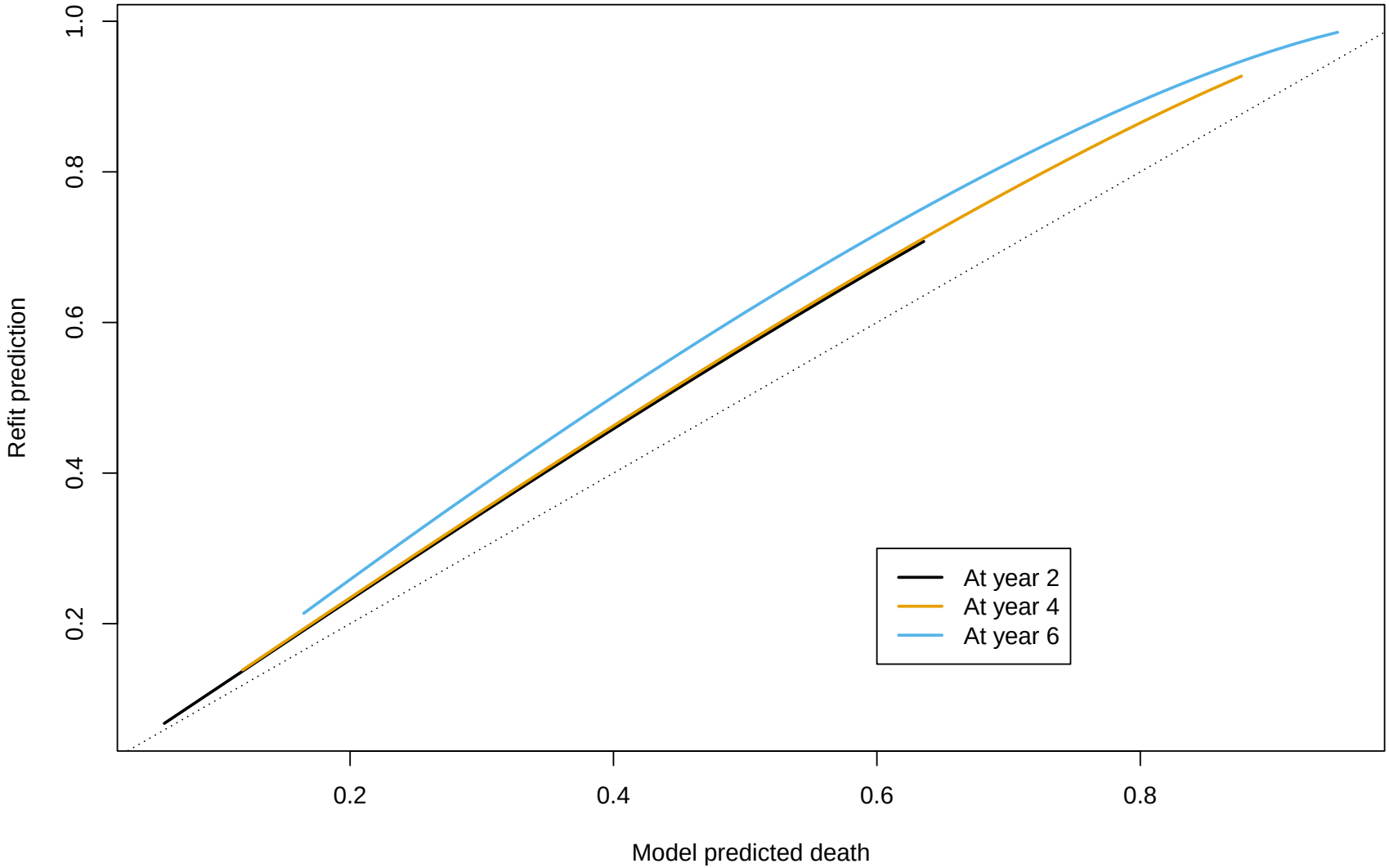

Observed GBSG deaths
Predicted
Death
0.0
0.2
0.4
0.6
0
2
4
6
Years since enrollment

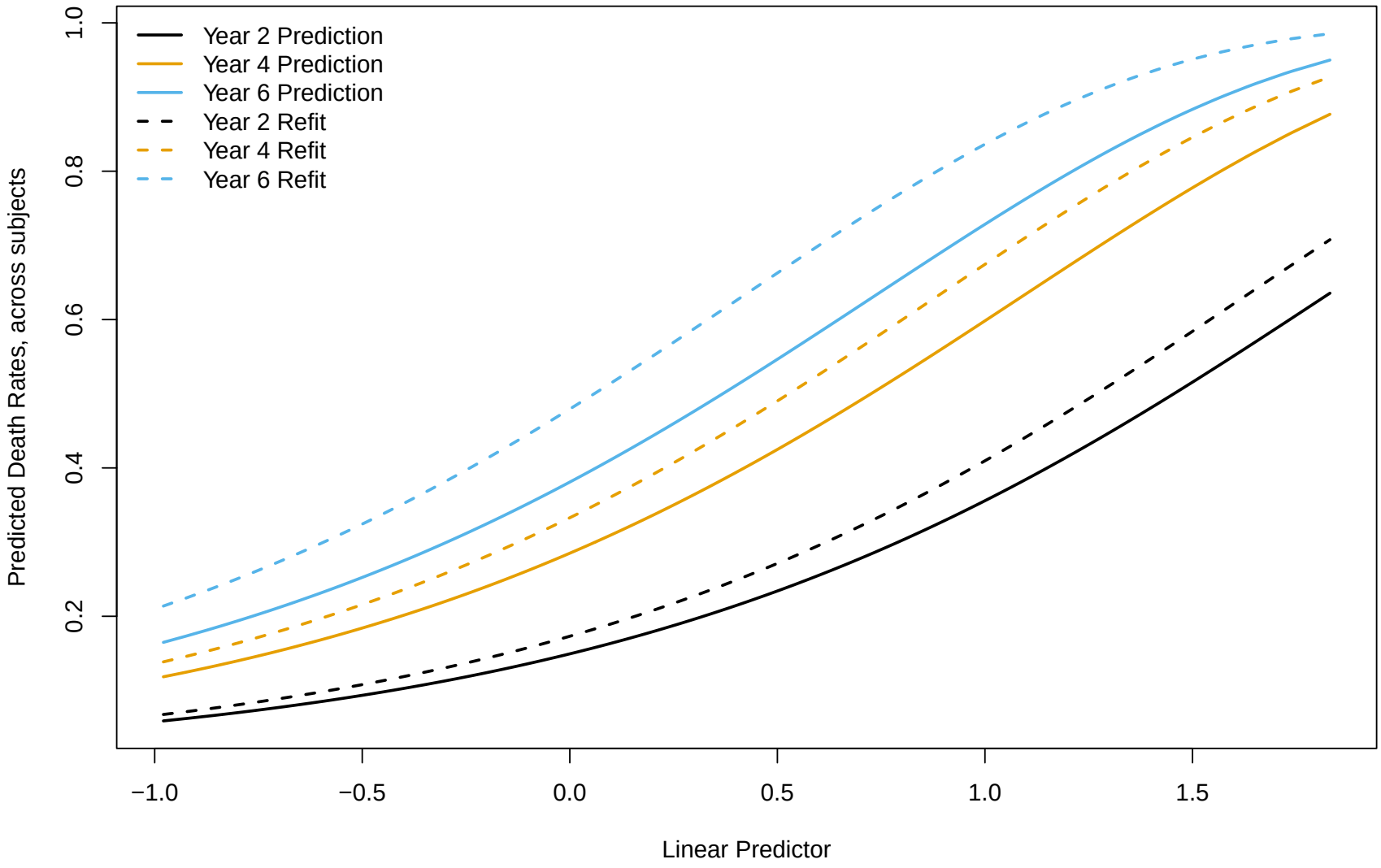
```
cfit1 <- coxph(Surv(ryear, rfs) ~ eta2, gbsg2)
cfit1

Call:
coxph(formula = Surv(ryear, rfs) ~ eta2, data = gbsg2)

        coef exp(coef) se(coef)     z      p
eta2 1.01981   2.77268  0.09966 10.23 <2e-16

Likelihood ratio test=102.8  on 1 df, p=< 2.2e-16
n= 686, number of events= 299
```


Refit prediction
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
Model predicted death
At year 2
At year 4
At year 6


Year 2 Prediction
Year 4 Prediction
Year 6 Prediction
Year 2 Refit
Year 4 Refit
Year 6 Refit
Predicted Death Rates, across subjects
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
Linear Predictor

Binomial models

- Pick a time τ
- Dichotomize the data
- $\hat{p}_i(\tau) =$ predictions from orignal model
- Direct (weighted)
- sensitivity, specificity, PPV, NPC
- AUROC
- Fit logistic regression models
- regression slope
- extra variables
- For GBSG
  - same song, third verse
  - larger variance

Multistate models

to be continued. . .