Absolute risk integration using penalized logistic regression

Jesse Islam

2/27/2020

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.
  - Cox Regression

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.
    - Cox Regression
- When we want the absolute risk:

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.
    - Cox Regression
- When we want the absolute risk:
    - Breslow estimator

Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis.
  - Cox Regression
- When we want the absolute risk:
  - Breslow estimator
  - Parametric models

Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [2]

Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [2]
- They believe the stepwise nature is the reason, as it reduces interpretability. [2]

Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [2]
- They believe the stepwise nature is the reason, as it reduces interpretability. [2]
- Want to easily model non-proportional hazards. [2]

Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [2]
- They believe the stepwise nature is the reason, as it reduces interpretability. [2]
- Want to easily model non-proportional hazards. [2]
- A streamlined approach for reaching a **smooth absolute risk** curve. [2]

Dr. Cox's perspective

Reid: How do you feel about the cottage industry that's grown up around it [the Cox model]?
Cox: Don't know, really. In the light of some of the further results one knows since, I think I would normally want to tackle problems parametrically, so I would take the underlying hazard to be a Weibull or something. I'm not keen on nonparametric formulations usually.
Reid: So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right.
Cox: That's right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution [see, e.g., Cox and Oakes, Analysis of Survival Data, Chapter 8.5]. And if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically.

Index pt 2.

- SUPPORT study

Index pt 2.

- SUPPORT study
- Casebase sampling

Index pt 2.

- SUPPORT study
- Casebase sampling
- Penalized logistic regression on survival data

Index pt 2.

- SUPPORT study
- Casebase sampling
- Penalized logistic regression on survival data
- Maximum likelihood with regularization

Index pt 2.

- SUPPORT study
- Casebase sampling
- Penalized logistic regression on survival data
- Maximum likelihood with regularization
- Absolute risk comparison

Index pt 2.

- SUPPORT study
- Casebase sampling
- Penalized logistic regression on survival data
- Maximum likelihood with regularization
- Absolute risk comparison
- Future work

Index pt 2.

- SUPPORT study
- Casebase sampling
- Penalized logistic regression on survival data
- Maximum likelihood with regularization
- Absolute risk comparison
- Future work
- References

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**
- Design: Prospective cohort study.

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**
- Design: Prospective cohort study.
- Setting: 5 academic care centers in the United States.

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**
- Design: Prospective cohort study.
- Setting: 5 academic care centers in the United States.
- Participants: 9105 hospitalized.

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**
- Design: Prospective cohort study.
- Setting: 5 academic care centers in the United States.
- Participants: 9105 hospitalized.
- Follow-up-time: 5.56 years.

SUPPORT dataset (knaus)

- **Study to Understand Prognoses and Preferences for Outcomes and Risks Treatments**
- Design: Prospective cohort study.
- Setting: 5 academic care centers in the United States.
- Participants: 9105 hospitalized.
- Follow-up-time: 5.56 years.
- 68% incidence rate.

SUPPORT manual imputation (knaus)

- Notorious for missing data

Baseline Variable	Normal Fill-in Value
Bilirubin	1.01
BUN	6.51
Creatinine	1.01
PaO2/FiO2 ratio (pafi)	333.3
Serum albumin	3.5
Urine output	2502
White blood count	9 (thousands)

Table 1: Suggested imputation values. [3]

SUPPORT automated imputation

- Mice imputation package (R)

SUPPORT automated imputation

- Mice imputation package (R)

1. PMM (Predictive Mean Matching) – For numeric variables

SUPPORT automated imputation

- Mice imputation package (R)

1. PMM (Predictive Mean Matching) – For numeric variables
2. logreg(Logistic Regression) – For Binary Variables (with 2 levels)

SUPPORT automated imputation

- Mice imputation package (R)

1. PMM (Predictive Mean Matching) – For numeric variables
2. logreg(Logistic Regression) – For Binary Variables (with 2 levels)
3. polyreg(Bayesian polytomous regression) – For Factor Variables ($>=$ 2 levels)

Removed variables (knaus)

- Hospital Charges.

Removed variables (knaus)

- Hospital Charges.
- Patient ratio of costs to charges.

Removed variables (knaus)

- Hospital Charges.
- Patient ratio of costs to charges.
- Patient Micro-costs.

Removed variables (knaus)

- Hospital Charges.
- Patient ratio of costs to charges.
- Patient Micro-costs.
- functional disability (ordinal).

Removed variables (knaus)

- Hospital Charges.
- Patient ratio of costs to charges.
- Patient Micro-costs.
- functional disability (ordinal).
- Income (ordinal).

Removed variables (knaus)

- Hospital Charges.
- Patient ratio of costs to charges.
- Patient Micro-costs.
- functional disability (ordinal).
- Income (ordinal).
- Surrogate activities of daily living.

Variable overview (knaus)

- **Response variables**

Variable overview (knaus)

- **Response variables**
    - follow-up time, death.

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)
  - Income, costs. (4)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)
  - Income, costs. (4)
  - Coma score, average Therapeutic Intervention Scoring System (2)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)
  - Income, costs. (4)
  - Coma score, average Therapeutic Intervention Scoring System (2)
  - Physiological variables. (11)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)
  - Income, costs. (4)
  - Coma score, average Therapeutic Intervention Scoring System (2)
  - Physiological variables. (11)
  - Activities of daily living. (2)

Variable overview (knaus)

- **Response variables**
  - follow-up time, death.
- **Covariates**
  - Age, sex, race, education (6)
  - Disease group/class, Number of comorbidities. (3)
  - Income, costs. (4)
  - Coma score, average Therapeutic Intervention Scoring System (2)
  - Physiological variables. (11)
  - Activities of daily living. (2)
  - Previous model findings. (8)

Original SUPPORT analysis (knaus)

- Determined SUPPORT prognostic model on phase I (4301 individuals).

Original SUPPORT analysis (knaus)

- Determined SUPPORT prognostic model on phase I (4301 individuals).
- Tested on Phase II (4028 individuals).

Original SUPPORT analysis (knaus)

- Determined SUPPORT prognostic model on phase I (4301 individuals).
- Tested on Phase II (4028 individuals).
- Both on the scale of 180 days.

Original SUPPORT analysis (knaus)

- Write out complicated model?????

SUPPORT question

- How does their model perform over 5.56 years?

SUPPORT question

- How does their model perform over 5.56 years?
- Absolute Risk comparison.

Analysis Process

1. Impute

Analysis Process

1. Impute
2. Compare SPS and APS over ∼5.56 years using absolute risk curves.

Analysis Process

1. Impute
2. Compare SPS and APS over ~5.56 years using absolute risk curves.
3. Compare to Kaplan-Meier curve

Analysis Process

1. Impute
2. Compare SPS and APS over ~5.56 years using absolute risk curves.
3. Compare to Kaplan-Meier curve
4. Compare to full model (excluding SPS and APS)

Analysis Process

1. Impute
2. Compare SPS and APS over ~5.56 years using absolute risk curves.
3. Compare to Kaplan-Meier curve
4. Compare to full model (excluding SPS and APS)

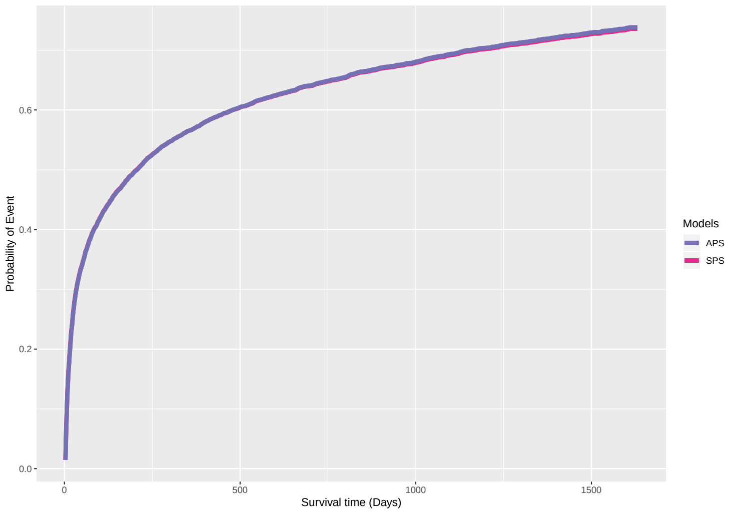
- All models is trained on 80% of the observations.

Analysis Process

1. Impute
2. Compare SPS and APS over ~5.56 years using absolute risk curves.
3. Compare to Kaplan-Meier curve
4. Compare to full model (excluding SPS and APS)

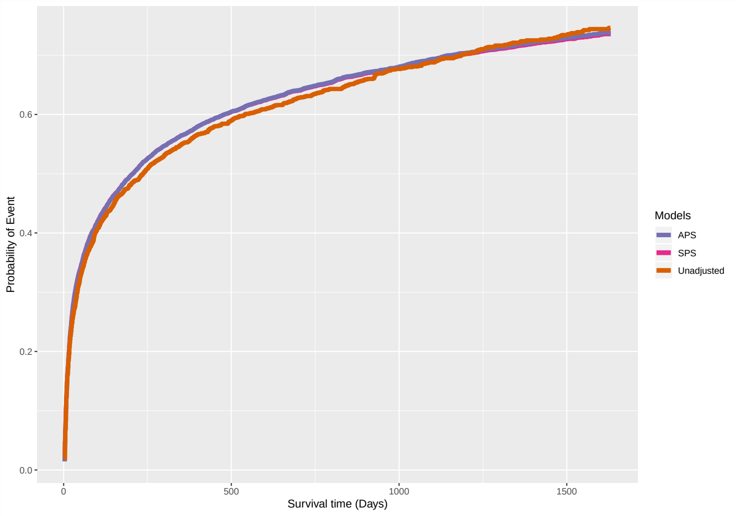
- All models is trained on 80% of the observations.
- Remaining observations are used to generate comparative absolute risk curves.

SPS vs APS

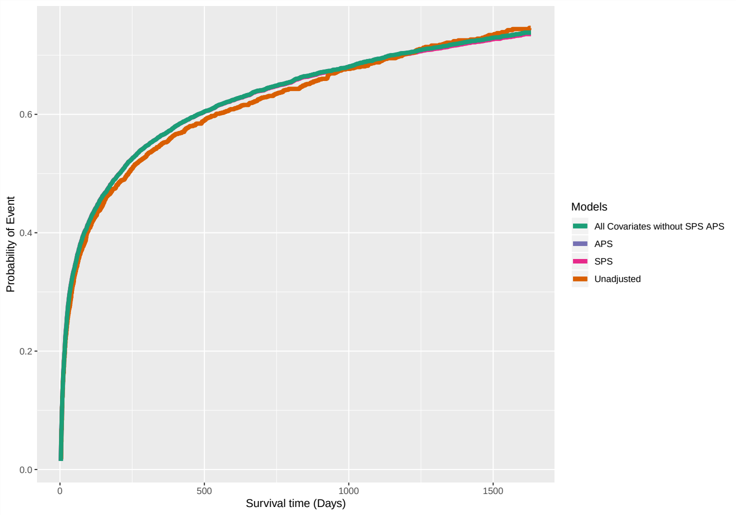


Probability of Event
0.0
0.2
0.4
0.6
0
500
1000
1500
Survival time (Days)
Models
APS
SPS

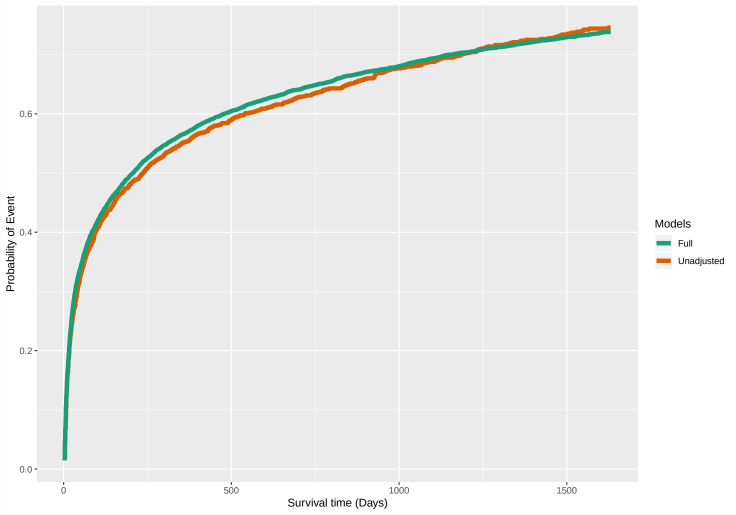
SPS vs. Kaplan-Meier



All covariates vs. physiology scores vs unadjusted



Chosen absolute risk comparisons



Chosen absolute risk comparisons: conclusion

- Linear associations without physiology scores perform similarly to physiology scores alone.

Casebase sampling overview

1. Clever sampling.

Casebase sampling overview

1. Clever sampling.
2. Implicitly deals with censoring.

Casebase sampling overview

1. Clever sampling.
2. Implicitly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

Casebase sampling overview

1. Clever sampling.
2. Implicitly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

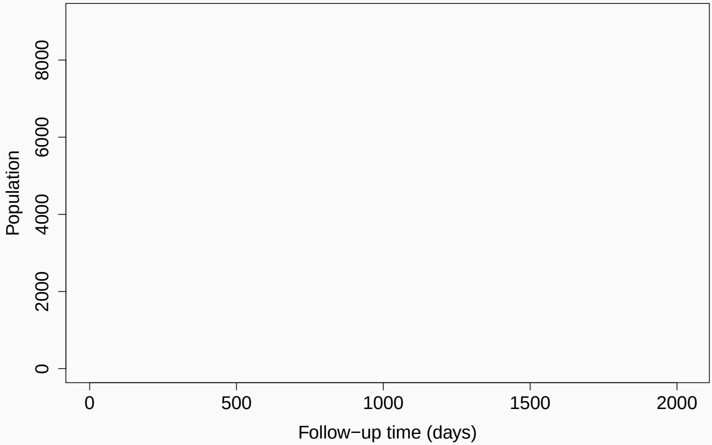
- Casebase is parametric, and allows different parametric fits by incorporation of the time component.

Casebase sampling overview

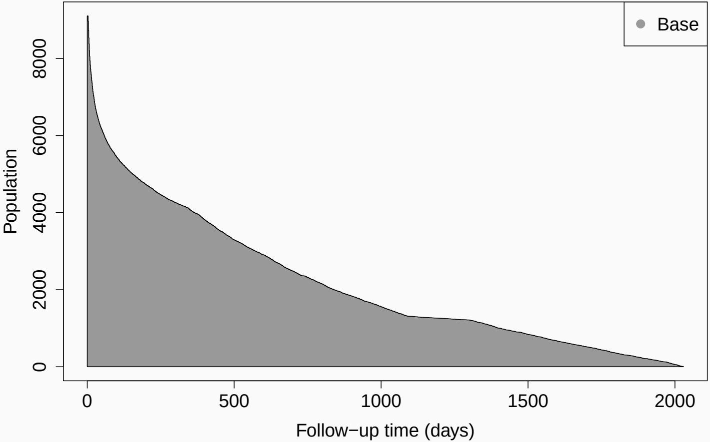
1. Clever sampling.
2. Implicitly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

- Casebase is parametric, and allows different parametric fits by incorporation of the time component.
- Package contains an implementation for generating *population-time* plots.

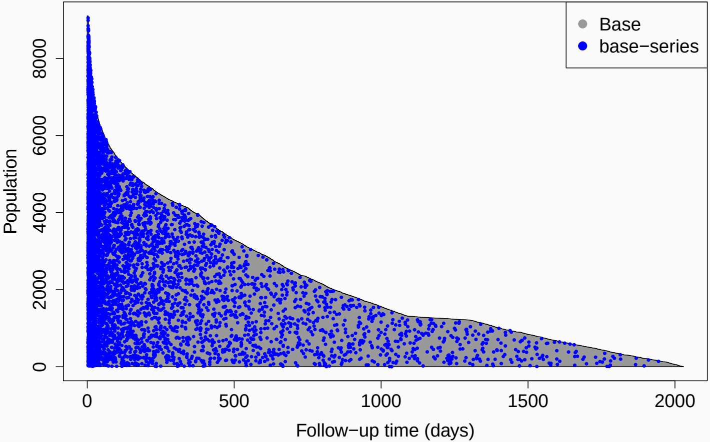
Casebase: Sampling



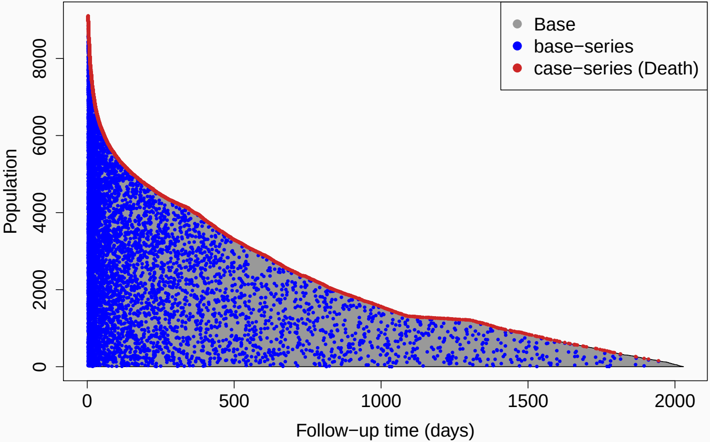
Casebase: Sampling



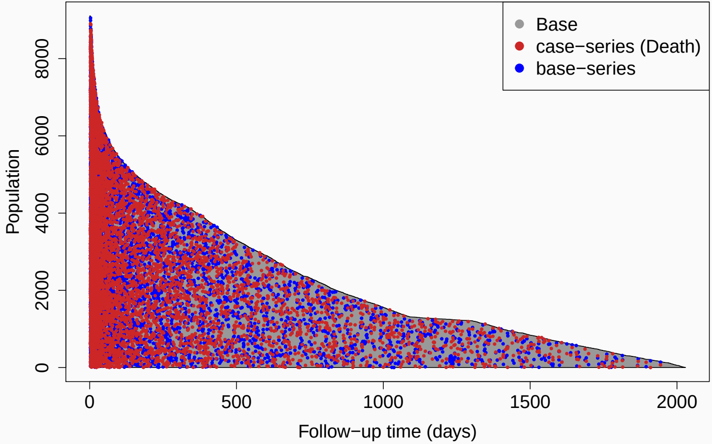
Casebase: Sampling



Casebase: Sampling



Casebase: Sampling



Casebase: Sampling

$$e^L = \frac{Pr(Y = 1|x, t)}{Pr(Y = 0|x, t)} = \frac{h(x, t) * B(x, t)}{b[B(x, t)/B]} = \frac{h(x, t) * B}{b}$$

- $L = \beta X$

Casebase: Sampling

$$e^L = \frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)} = \frac{h(x,t)*B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t)*B}{b}$$

- $L = \beta X$
- $b =$ base-series.

Casebase: Sampling

$$e^L = \frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)} = \frac{h(x,t) * B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t) * B}{b}$$

- $L = \beta X$
- $b =$ base-series.
- $B =$ Base.

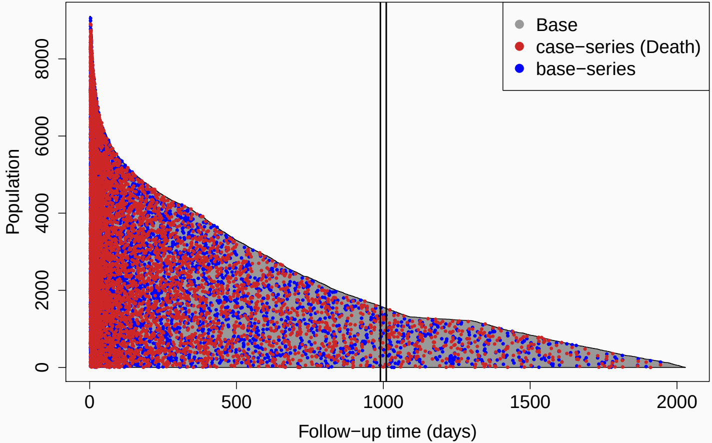
Casebase: Sampling

$$e^L = \frac{Pr(Y = 1|x,t)}{Pr(Y = 0|x,t)} = \frac{h(x,t) * B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t) * B}{b}$$

- $L = \beta X$
- b = base-series.
- B = Base.
- *B(x,t)* = Risk-set for survival time t.

Casebase: Sampling

$$e^L = \frac{Pr(Y{=}1|x,t)}{Pr(Y{=}0|x,t)} = \frac{h(x,t) * B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t) * B}{b}$$



log-odds = log hazard

$$e^L = \frac{\hat{h}(x,t) * B}{b}$$

$$\hat{h}(x,t) = \frac{b * e^L}{B}$$

$$log(\hat{h}(x,t)) = L + log(\frac{b}{B})$$

Maximum log-likelihood [1]

$$log(l(\beta)) = \sum_{i=1}^{N}(\sum_{k=0}^{K} x_{ik}\beta_k) - n_i log(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k})$$

Maximum log-likelihood, with offset

$$log(l(\beta)) = \sum_{i=1}^{N}(\sum_{k=0}^{K} x_{ik}\beta_k + log(\frac{b}{B})) - n_i log(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k + log(\frac{b}{B})})$$

Maximum log-likelihood, with offset and lasso

$$log(l(\beta)) = \sum_{i=1}^{N} (\sum_{k=0}^{K} x_{ik}\beta_k + log(\frac{b}{B})) - n_i log(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k + log(\frac{b}{B})}) - \lambda ||\beta||$$

Casebase: Parametric families

- We can now fit models of the form:

$$log(h(t;\alpha,\beta)) = g(t;\alpha) + \beta X$$

Casebase: Parametric families

- We can now fit models of the form:

$$log(h(t; \alpha, \beta)) = g(t; \alpha) + \beta X$$

- By changing the function $g(t; \alpha)$, we can model different parametric families easily:

Casebase: Parametric models

Exponential: $g(t; \alpha)$ is equal to a constant

```r
casebase::fitSmoothHazard(status ~ X1 + X2)
```

Gompertz: $g(t; \alpha) = \alpha t$

```r
casebase::fitSmoothHazard(status ~ time + X1 + X2)
```

Weibull: $g(t; \alpha) = \alpha log(t)$

```r
casebase::fitSmoothHazard(status ~ log(time) + X1 + X2)
```

Absolute Risk

- We have a bunch of different parametric hazard models now.

Absolute Risk

- We have a bunch of different parametric hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x, t) = 1 - e^{-\int_0^t h(x,u)du}$$

Absolute Risk

- We have a bunch of different parametric hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)

Absolute Risk

- We have a bunch of different parametric hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)
- *h(x,u)*= Hazard function

Absolute Risk

- We have a bunch of different parametric hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- *CI(x,t)*= Cumulative Incidence (Absolute Risk)
- *h(x,u)*= Hazard function
- Lets use the weibull hazard

models to be compared

- Recall: Case study to demonstrate regularization using our method.

models to be compared

- Recall: Case study to demonstrate regularization using our method.
- **unadjusted**: (death, time) ~ 1

models to be compared

- Recall: Case study to demonstrate regularization using our method.
- **unadjusted**: (death, time) ~ 1
- **Cox**: (death, time) $\sim \beta X$

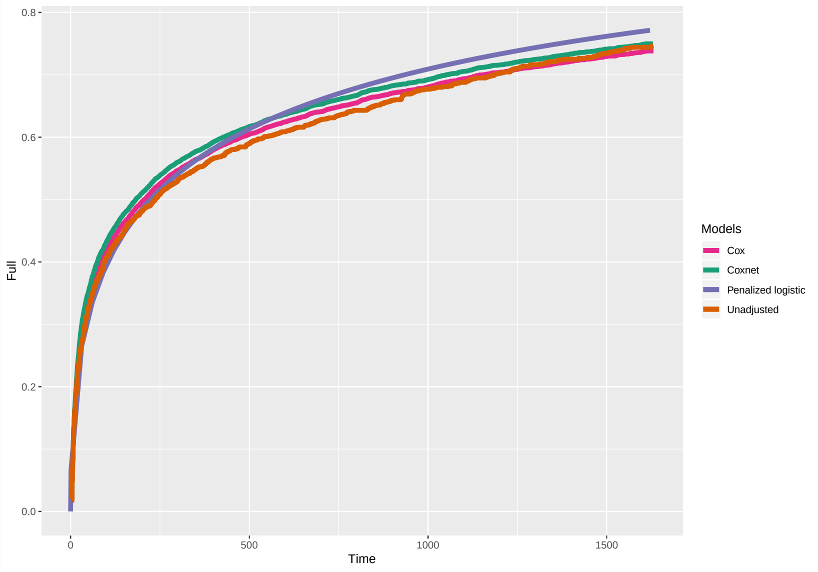
models to be compared

- Recall: Case study to demonstrate regularization using our method.
- **unadjusted**: (death, time) ~ 1
- **Cox**: (death, time) $\sim \beta X$
- **Coxnet**: (death, time) $\sim \beta X$ <–Lasso

models to be compared

- Recall: Case study to demonstrate regularization using our method.
- **unadjusted**: (death, time) ~ 1
- **Cox**: (death, time) ~ βX
- **Coxnet**: (death, time) ~ $\beta X <$–Lasso
- **Penalized logistic**: death ~ log(time) + $\beta X <$–Lasso

Survival comparison



Future work

- IPA score

Future work

- IPA score
- Survival GWAS

References 1

1.Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep. net/stat/mlelr. pdf, 1825252548-1564645290. ????

2.Hanley, James A, and Olli S Miettinen. 2009. "Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression." *The International Journal of Biostatistics 5 (1)*.

3.Harrell, F. (2020). SupportDesc < Main < Vanderbilt Biostatistics Wiki. [online] Biostat.mc.vanderbilt.edu. Available at: http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc [Accessed 25 Feb. 2020].

References 2

4.Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., . . . & Layde, P. (1995). The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. Annals of internal medicine, 122(3), 191-203.

5.Saarela, Olli, and Elja Arjas. 2015. "Non-Parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment." Scandinavian Journal of Statistics 42 (2). Wiley Online Library: 609–26.

6.Saarela, Olli. 2015. "A Case-Base Sampling Method for Estimating Recurrent Event Intensities." *Lifetime Data Analysis*. Springer, 1–17

References 3

7.Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplant*. 2007 Aug;40(4):381-7. doi: 10.1038/sj.bmt.1705727.

8.Turgeon, M. (2017, June 10). Retrieved May 05, 2019, from https://www.maxturgeon.ca/slides/MTurgeon-2017-Student-Conference.pdf

Tutorial and Slides

\begin{center} Tutorial:

http://sahirbhatnagar.com/casebase/

Slides:

https://github.com/Jesse-Islam/ATGC_survival_presentation_Feb.27.2020

Questions? \end{center}