Absolute Risk integration using penalized logistic regression

Jesse Islam

2/16/2020

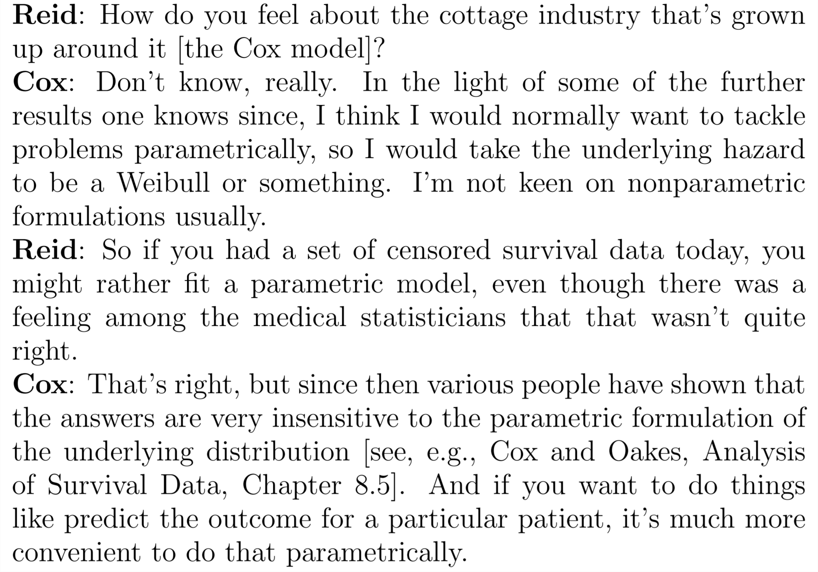
Popular methods in time-to-event analysis

- 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. [1]
- They believe the stepwise nature is the reason, as it reduces interpretability. [1]
- Want to easily model non-proportional hazards. [1]
- A streamlined approach for reaching a **smooth absolute risk** curve. [1]

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

- SUPPORT study
- Casebase sampling
- Logistic regression on survival data
- Maximum likelihood with regularization
- Comparing hazard models in SUPPORT study
- Absolute risk comparison
- Future work
- References

SUPPORT dataset

- **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

- 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. [Support site reference]

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)
4. Proportional odds model (ordered, $>=$ 2 levels)

Removed variables

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

Variable overview

- Age, sex, race, education, follow-up time, death. (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. (3)
- Previous model findings. (8)

Original SUPPORT analysis

- Determined SUPPORT prognostic model on phase I (4301 individuals).
- Tested on Phase II (4028 individuals).
- Both on the scale of 180 days.
- Write out complicated model?????
- image of SPS vs APS ???????

SUPPORT question

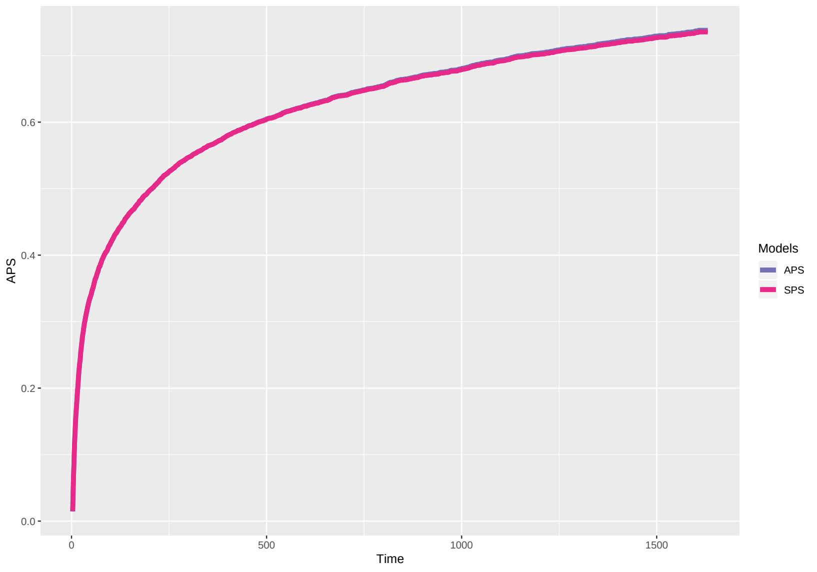
- How does their model perform over 5.56 years?
- Absolute Risk comparison.

Analysis Process

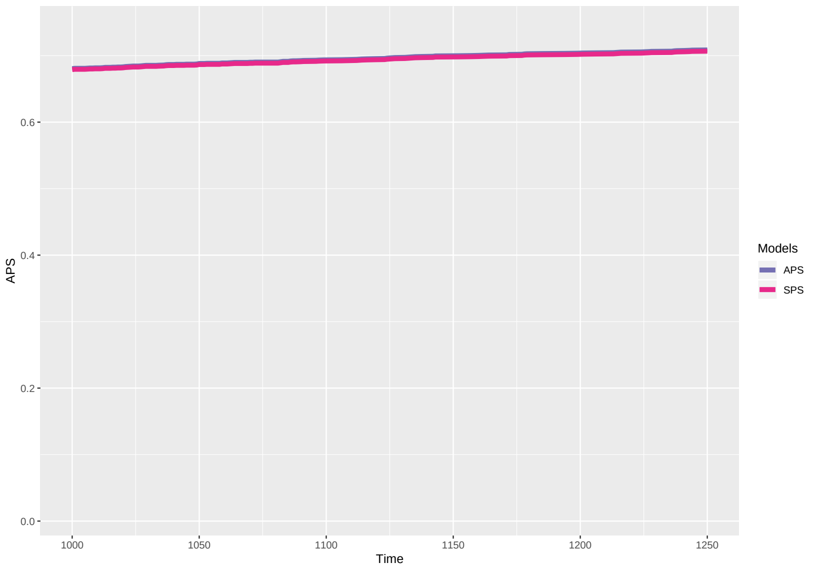
1. Impute
2. Compare SPS and APS over ~5.56 years using absolute risk curves.
3. Compare to Kaplan-Meier curve
4. Compare to full model (excluding SPS and APS)

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

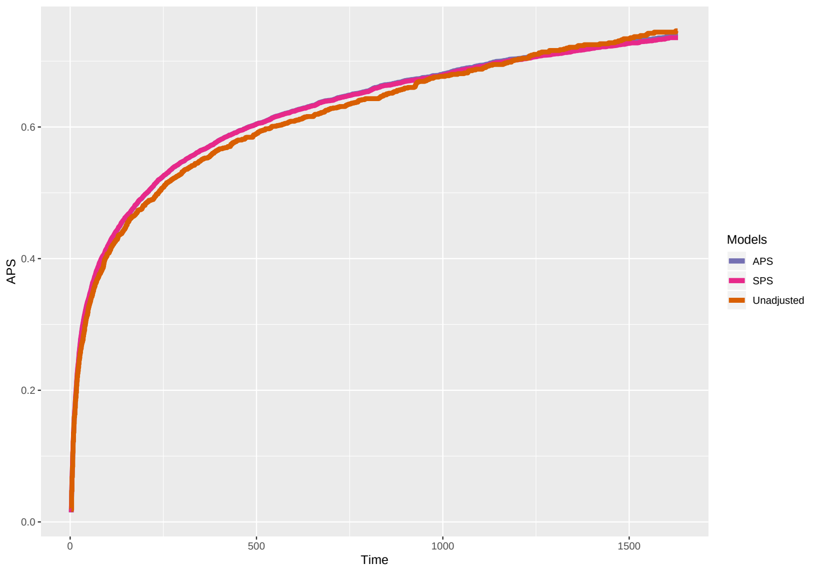
SPS vs APS



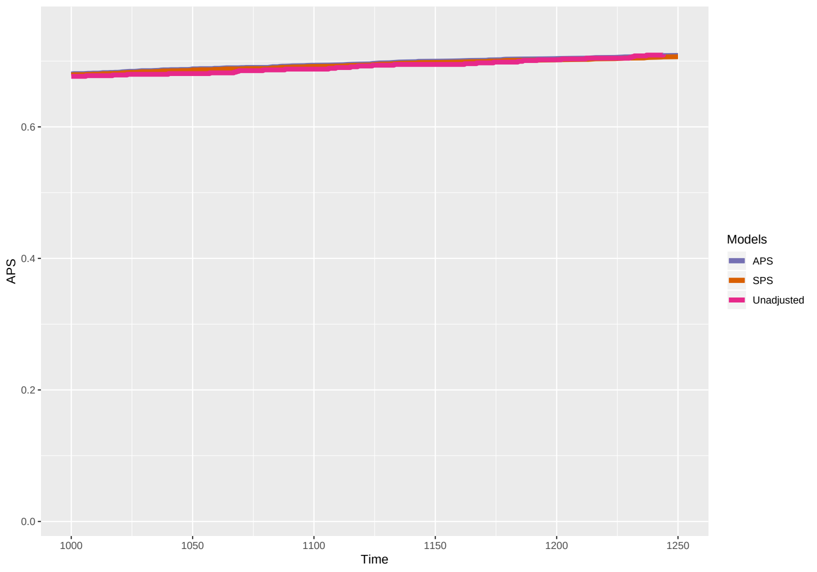
SPS vs APS


APS
0.6
0.4
0.2
0.0
1000
1050
1100
1150
1200
1250
Time
Models
APS
SPS

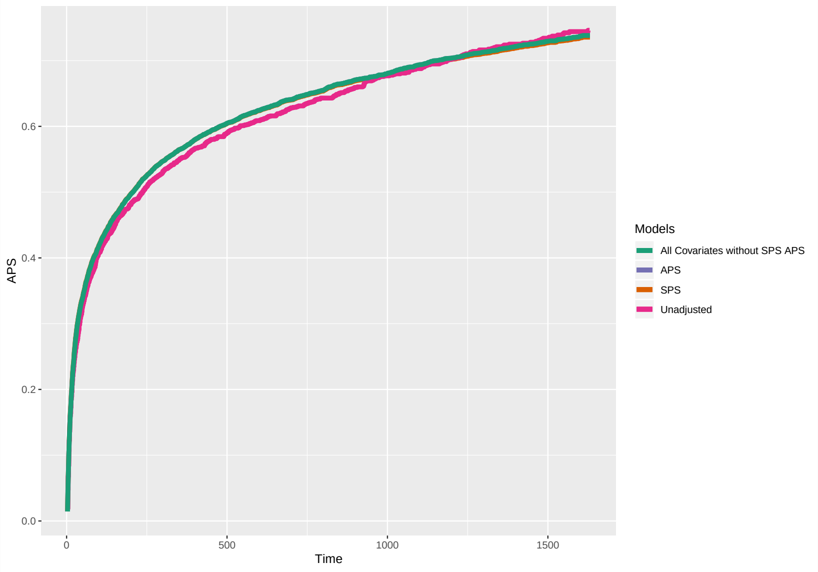
SPS vs. Kaplan-Meier



SPS vs. Kaplan-Meier



All covariates vs. physiology scores vs unadjusted



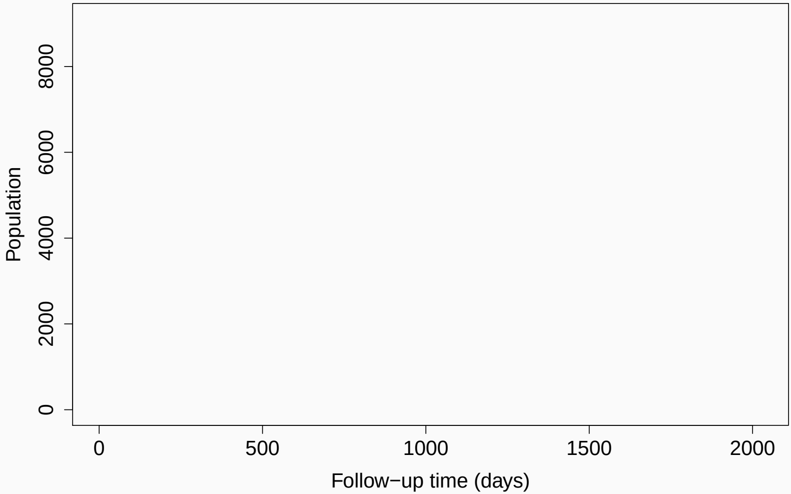
- Conclusion: linear associations without physiology scores

Casebase Overview

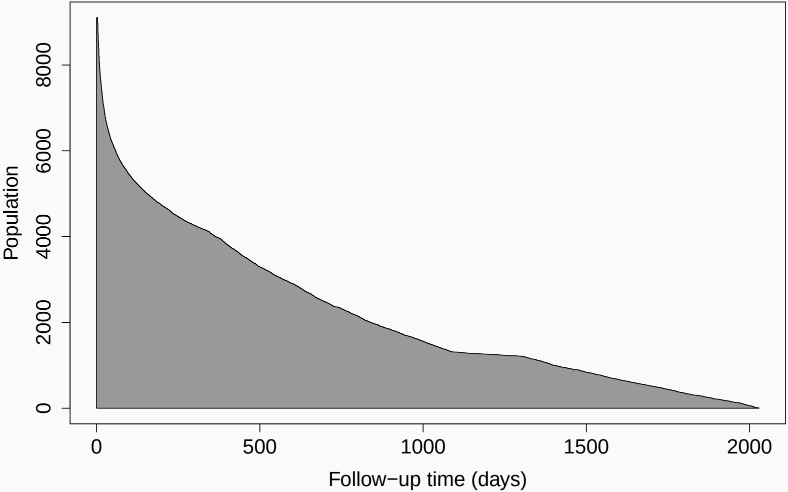
1. Clever sampling.
2. Implicitly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

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

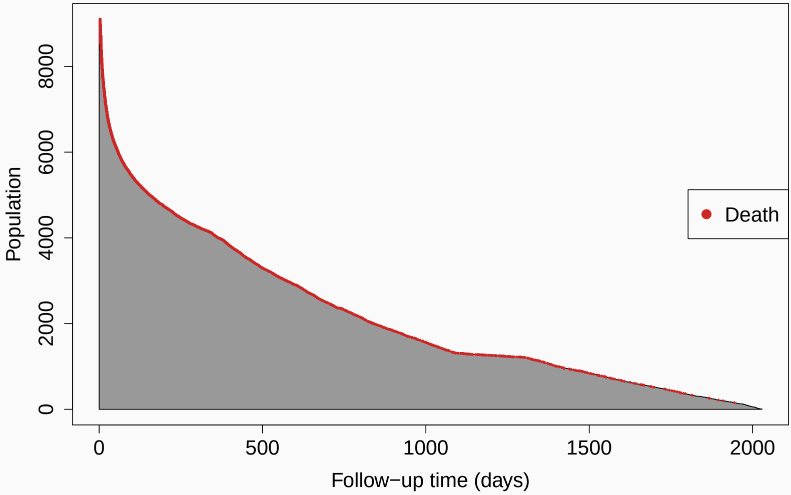
Casebase: Sampling



Casebase: Sampling

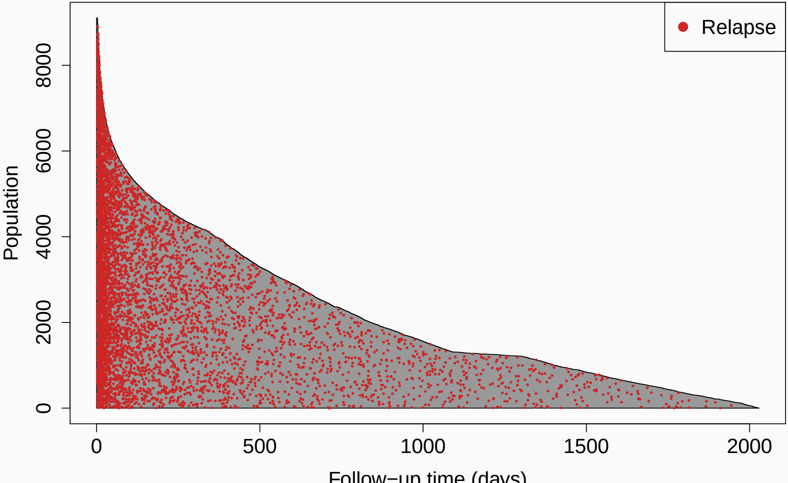


Casebase: Sampling



Casebase: Sampling

```
casebase::popTime(Data,Event,Time)
```



log-odds = log hazard

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

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

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

- $\hat{L} = \beta X$
- $b =$ base-series.
- $B =$ Base.
- *B(x,t)* = Base at time t.

Wolfe’s variance for case-to-base ratio

$$(\frac{1}{c} + \frac{1}{b})^{-1}$$

$$(\frac{1}{c} + \frac{1}{100c})^{-1}$$

maximum likelihood, with regularization

- should I show it?

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

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

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

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

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

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

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:

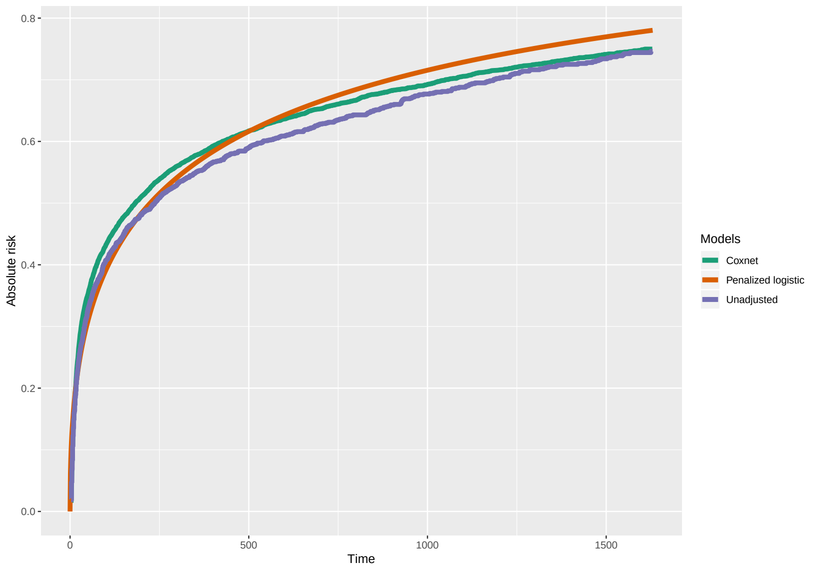
$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

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

models to be compared

- casebase surv weibull-> LASSO
- cox surv
- cox surve -> LASSO
- Kaplan-meier

Survival comparison



Covariate comparison plot

- there will be a lollipop plot but I wanted to sleep

IPA score

- Brier score equation
- Calibration and discrimination
- IPA score equation
- In progress

Future work

- survival GWAS

Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
    - Population-Time plots
    - Hazard functions
    - Absolute Risk
    - Casebase can deal with competing risks.

References 1

1.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)*.

2.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.

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

References 2

4.Schroder FH, et al., for the ERSPC Investigators.Screening and Prostate-Cancer Mortality in a Randomized European Study. *N Engl J Med* 2009;360:1320-8.

5.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.

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

Tutorial and Slides

Tutorial:

http://sahirbhatnagar.com/casebase/

Slides:

https://github.com/Jesse-Islam/UseR–CaseBase-Presentation

Questions?