§New. Other Topics

PubH 7450

©Wei Pan

Email: weip@biostat.umn.edu

Http: www.biostat.umn.edu/∼weip

§New. Analysis of Recurrent Events

- Recurrent events: an event of interest can happen multiple times on each subject; e.g. multiple infections, cancer relapses...
- Closely related to correlated survival data analysis (i.e. assuming possible correlations among the multiple observations from the same subject), but more complex:
  need to prepare the data in a format matching one of the models you choose.
- A bladder cancer example: see SAS manual
  86 subjects;
  Recurrent event: recurrence of bladder cancer tumors after surgical removal;
  Covariates: Tx (0=placebo, 1=trt thiotepa), Num and Size (initial number and size of tumors);
  Subject 10 had the first two events at time points 12 and 16

months, then censored at 18 months;

- Two decisions:

  1. Marginal vs Conditional (vs Frailty?) model?
  2. What is the starting time after the previous event?

  Example: for Subj 10, what is the starting time for the second event, 12 or 0?

- Marginal 1: Counting Process model; assume the same type of recurrences, e.g. (non-severe) cold or flu or ear infections.

```
ID Enum Evt Start Stop Tx Num Size
10  1    1    0    12  0   1   1
10  2    1   12    16  0   1   1
10  3    0   16    18  0   1   1
11  ...
...
SAS code:
```

```
proc phreg covm covs(aggregate);
  model (Start, Stop)*Evt(0)=Tx Num Size;
  id ID;
```

- Remark: using the model-based covariance estimate `covm` (i.e. under the independence assumption) corresponds to *intensity model*, while using the sandwitch estimate `covs(aggregate)` (i.e. accounting for possible within-subject correlations) corresponds to *proportional means model* (Lin et al 2000).
- Marginal 2: Wei-Lin-Weissfeld (1989) WLW model; assume different types of recurrences, e.g. cancer relapses.

```
ID Enum Evt Start Stop Tx Num Size
10  1    1    0    12  0   1   1
10  2    1    0    16  0   1   1
10  3    0    0    18  0   1   1
10  4    0    0    18  0   1   1
...
```

```
proc phreg covs(aggregate);
  model Stop*Evt(0)=Tx Num Size;
  strata Enum;
  id ID;
```

- Remarks:
  1) the 4th obs for subj 10 is created since there are max 4 recurrences in the data;
  2) Throughout, one may first include interaction Tx*Enum, why?

- Conditional 1: Prentice-Williams-Peterson (1981) PWP total time model,

```
ID Enum Evt Start Stop  Tx Num Size
10  1    1    0    12   0   1   1
10  2    1   12    16   0   1   1
10  3    0   16    18   0   1   1
11  ...
```

```
...
proc phreg;
  model (Start, Stop)*Evt(0)=Tx Num Size;
  strata Enum;
```

- Conditional 2: PWP gap-time model, $Gaptime = Stop - Start$

```
ID Enum Evt Start Stop Gaptime Tx Num Size
10  1    1   0     12   12      0   1   1
10  2    1  12     16    4      0   1   1
10  3    0  16     18    2      0   1   1
11  ...
...
proc phreg;
  model Gaptime*Evt(0)=Tx Num Size;
  strata Enum;
```

§New. Penalized Semi-parametric PH Regression

- PHM: $h(x|Z) = h_0(x)\exp(Z'\beta)$.
- Inference: use the partial likelihood $L(\beta)$; e.g., MPLE

$$\hat{\beta} = \arg\max_{\beta} \log L(\beta),$$

- A problem: what happens if $p = dim(\beta)$ is close to or even larger than the sample size n in **high-dimensional data**?
  Example: in gene expression data, p 1,000s to 10,000s, n 10s to 100s.
- How to proceed?
  As usual, ...
- An alternative, simultaneous variable selection and parameter estimation via penalzied regression.
  Literature: most in linear regression.

- Penalized PH regression: MPPLE

$$\tilde{\beta} = \arg\max_{\beta} \log L(\beta) - g_\lambda(\beta),$$

  where λ is a tuning parameter to be decided.

- Examples:
  1) Ridge (Hoerl and Kennard 1970):

$$g_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2;$$

  2) Lasso (Tibshirani 1992):

$$g_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|;$$

- Typically, compared to MPLE $\hat{\beta}$, MPPLE $\tilde{\beta}$ is shrunken towards 0.

More importantly, if λ is large enough in Lasso (but not in Ridge), many $\tilde{\beta}_j = 0 \Longrightarrow$ variable selection!

- The MPPLE $\tilde{\beta}$ depends on the choice of λ:
  Use some model selection criteria (e.g. AIC or BIC), or cross-validation (CV).

- MPPLE has a Bayesian interpretation: $-g_\lambda(\beta)$ is log prior density; MPPLE is maximum *a posteriori* estimate (MAPE).
  Ridge: β_j iid $N(0, \sigma^2)$;
  Lasso: β_j iid Laplacian (i.e. double exponential) with mean 0 and scale σ;
  Both: $\sigma \sim \lambda$

- Performance in variable selection: any uniform winner?
  Compared to sequential (e.g. step-wise) var selection,
  Lasso performs better if the true model is ...
  Ridge?

- Performance in prediction:
  If $p \approx n$ or $p > n$, Lasso and ridge often perform better than MPLE (or MLE); why?
  Between Lasso and ridge:
  Combining Lasso and ridge: elastic net (Zou and Hastie 2005),

$$g_\lambda(\beta) = \lambda[\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|],$$

  where α (like λ) is another tuning parameter to be decided.

- Downsides:
  1) Biased parameter estimates!
  possible solutions: SCAD (Fan and Li 2002); Adaptive Lasso (Zou 2006); TLP (Shen, Pan & Zhu 2011),...
  2) More importantly, inference?

- An R example.

§New. Sample size calculations

- Reference: Shih (1995). Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials*, 16:395-407.
- SAS macro %size
- See also SAS Proc Power.

§New. More on model checking

- Goal: checking PHM with right censored data.
- Use Cumulative Sums of Martingale Residuals (Lin, Wei & Ying 1993)
  checking on a covariate:
  1) PH assumption: i) graphics (with simulated null realizations, like envelops or point-wise CIs); 2) p-value.
  2) Functional form: Pattern in a residual plot may suggest a possible transformation.
- SAS manual for Proc Phreg ASSESS statement.