REGRESSION ANALYSIS OF RECURRENT EVENT DATA

Jerry Lawless
University of Waterloo

Reference: The Statistical Analysis of Recurrent Events
by R.J. Cook and J.F. Lawless (Springer, 2007)

OUTLINE

1. Review of notation and types of problems

2. Introduction of examples

3. Regression methodology for mean and rate functions

4. Illustrations

5. Extensions and topics needing development

1. Review of notation and types of problems

- Repeated occurrences of some type of event

  - recurrent infections or disease episodes (e.g. bronchial infections, herpes simplex outbreaks)

  - epileptic seizures, asthma attacks

  - warranty claims for manufactured products

  - failures in software systems

- In general, consider multiple units or individuals $i = 1, 2, \ldots$ and a time scale t

$$N_i(t) = \text{number of events up to time } t \text{ for unit } i \qquad (t \geq 0)$$

- Objectives of analysis include

  - understanding and characterizing event occurrence (patterns over time, probabilities, dynamics)

  - explaining unit-to-unit variability (covariates, comparisons, treatments, excess variation)

  - assessing relationships with time-varying covariates or other processes

  - prediction

- Covariates (explanatory variables) x_i

  - fixed or time-varying

- Ways of looking at recurrent events

  - counts of cumulative numbers of events, or numbers in distinct time intervals

  - “gap” times between successive events

  - event intensities (probability of new event, given past events)

- Here we focus on counts: fundamental characteristics are then the means, variances and covariances, and distribution of counts for specified time intervals.

  mean (cumulative) function or MCF: $\mu_i(t) = E\left\{N_i(t)\right\}$

  rate of occurrence function: $\rho_i(t) = \mu_i'(t)$

2. Some examples

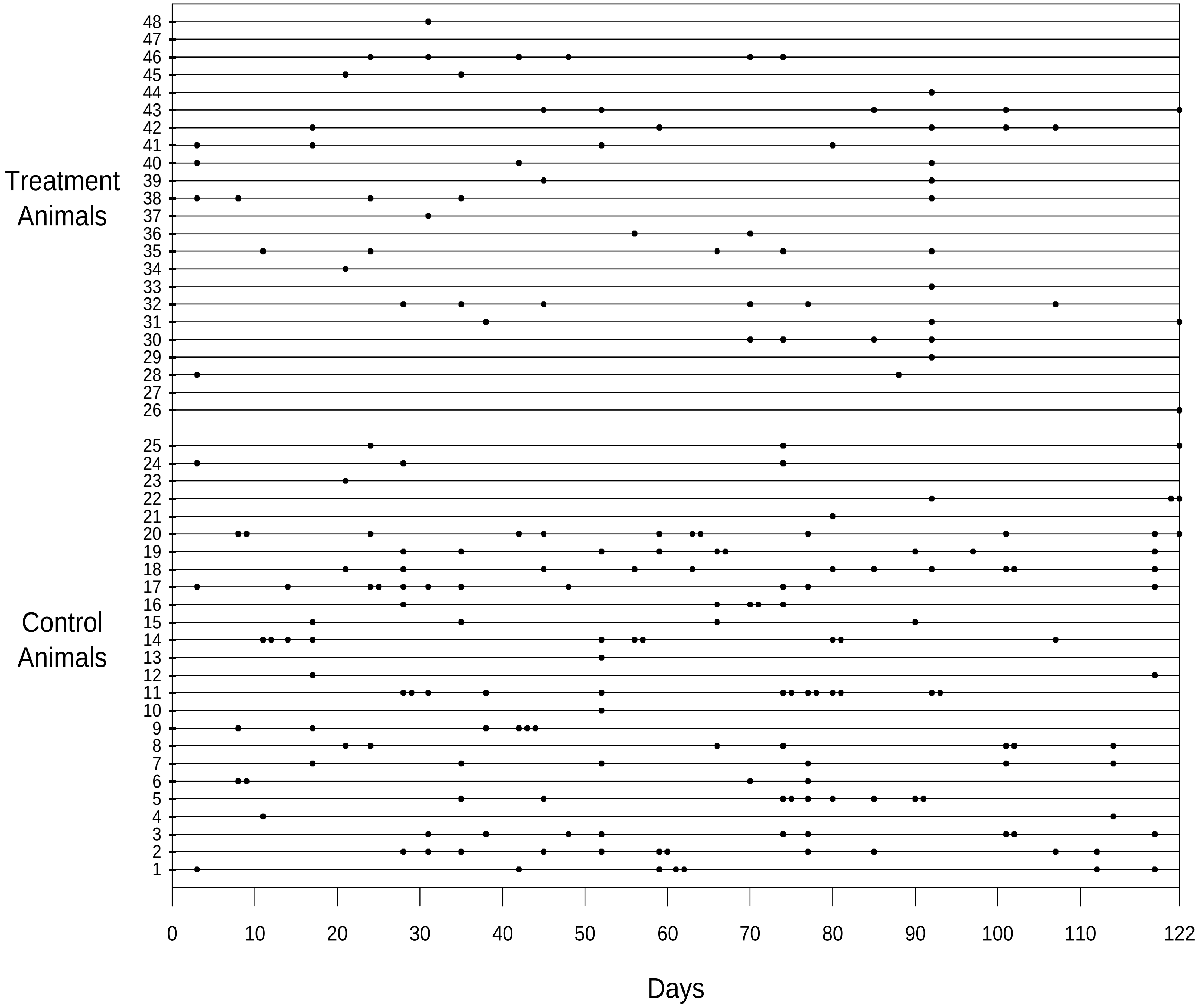
- Mammary tumors in a carcinogenicity study (Gail et al. 1980)

  - Treatment ($n = 23$) and control ($n = 25$) groups of female rats, each exposed to a carcinogen

  - Animals followed for 122 days and times of occurrence of new tumors were recorded (see figure).

$$\overline{N}_T(122) = 2.65 \qquad \overline{N}_c(122) = 6.04$$

  - Objective is to compare treatment groups with respect to the frequency of tumor occurrence


Treatment Animals
Control Animals
48
47
46
45
44
43
42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80
90
100
110
122
Days

- RCT for treatment of Herpes Simplex Virus (HSV) infections

  - 48 week multi-center crossover trial for persons with HSV infections (Romanowski et al. 2003)

  - Patients randomized to two treatment groups

    A: suppressive treatment for weeks 1 - 24, then episodic treatment for weeks 25 - 48

    B: episodic treatment for weeks 1 - 24; suppressive for weeks 25 - 48

  - Other explanatory variables include age, sex, race, virus type, history re previous occurrences

- Automobile warranty claims

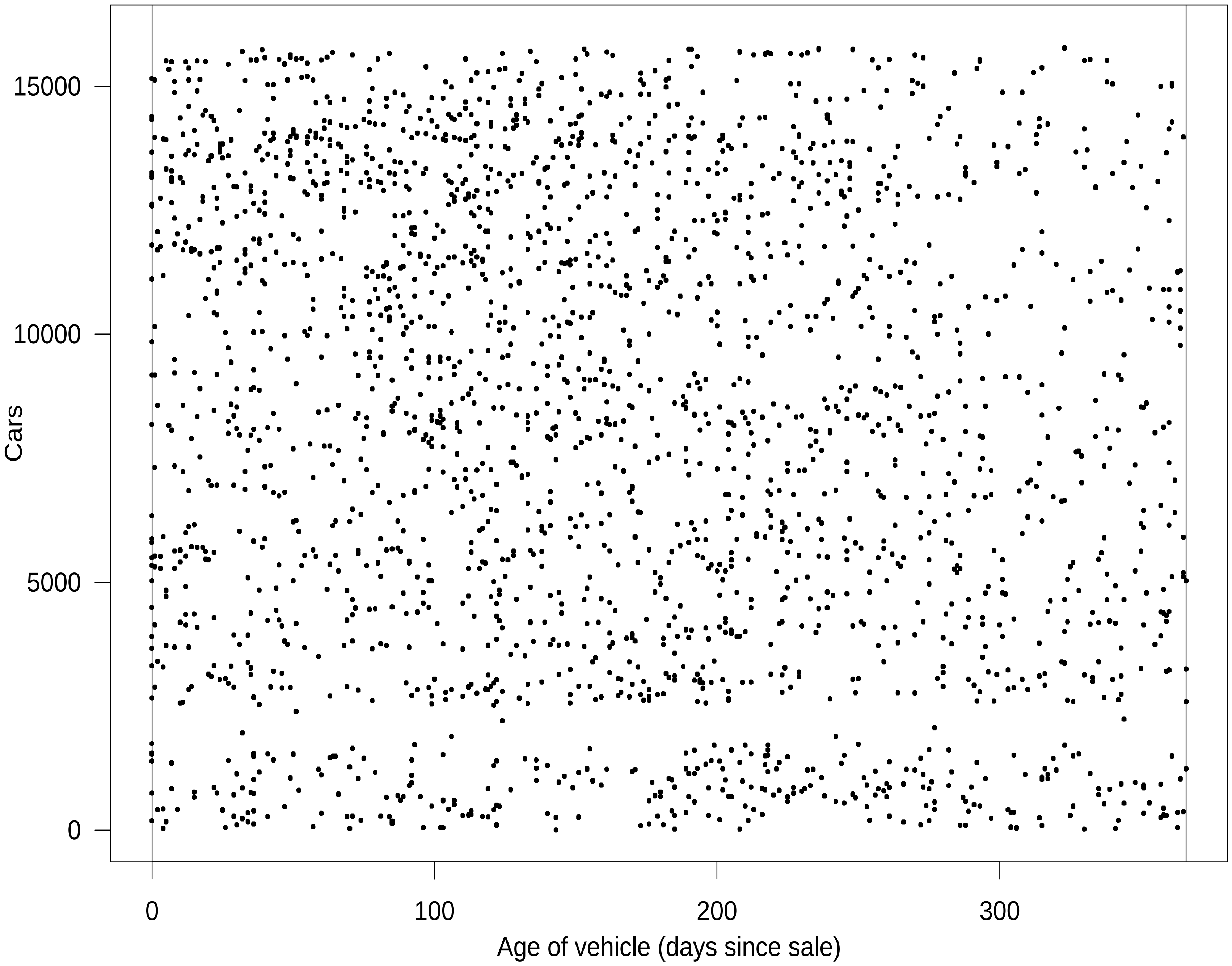
  - data on warranty claims (under a 1-year, 12,000-mile warranty) for 38,401 cars of one model type

  - “time” scale could be age (days since sale of vehicle) or mileage

  - variable lengths of followup, according to date of sale of vehicle and date of analysis

  - plot shows ages of claims for 15,775 cars which have 1 year of followup

- Objectives include comparison of claims for vehicles manufactured in different time periods or locations; predictive modeling; early detection of problems


Cars
15000
10000
5000
0
0
100
200
300
Age of vehicle (days since sale)

3. Regression methodology for mean and rate functions

- Individuals $i = 1, \ldots, m$ with covariate vectors x_i or (if time-varying) $x_i(t)$, $t \geq 0$
- Denote times of events for individual i as $t_{i1}, t_{i2}, \ldots$
- Conditional on covariates, let

$$\mu_i(t) = E\left\{N_i(t)\right\} \qquad \rho_i(t) = \mu_i'(t)$$

- Common model:

$$\rho_i(t) = \rho_0(t) \exp\left(x_i(t)'\beta\right) \tag{1}$$

$$\mu_i(t) = \int_0^t \rho_i(u) du$$

- If $x_i(t) = x_i$ then $\mu_i(t) = \mu_0(t) \exp(x_i'\beta)$

Approaches to Modelling and Analysis

- Process intensity functions: let $H_i(t)$ be the history of events and covariates up to time t. Then

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{\Pr\left\{N_i(t+\Delta t) - N_i(t) = 1 | H_i(t)\right\}}{\Delta t}$$

  is called the intensity function.

- In continuous time, assume two events cannot occur simultaneously.

- If individual i is observed over a specified time interval $(0, \tau_i)$ then the probability density for the outcome "n_i events, at times $t_{i1} < t_{i2} < \ldots t_{in_i}$" $(n_i \geq 0)$ is

$$\left\{ \prod_{j=1}^{n_i} \lambda_i(t_{ij}) \right\} \exp\left\{ -\int_0^{\tau_i} \lambda_i(u) du \right\}. \qquad (2)$$

- Approach 1: Specify a model for $\lambda_i(t)$ and use (2) to get the likelihood function and MLEs.

  e.g. Poisson process: $\lambda_i(t) = \rho_i(t) =$ rate function
  e.g. Negative binomial process: "includes" Poisson process

$$\lambda_i(t) = \left\{ \frac{1 + \phi N_i(t-)}{1 + \phi \mu_i(t-)} \right\} \rho_i(t)$$

- Approach 2: To reduce model misspecification problems, just model the mean and rate functions e.g.

$$\rho_i(t) = \rho_0(t) \exp(x_i' \beta) \tag{3}$$

  without assuming the process is Poisson or any other specific process.

- Approach 2 is sometimes called "robust"

Robustness of Poisson process estimators

- If $\lambda_i(t) = \rho_i(t;\theta)$ then from (2) the log likelihood estimating equations from m independent individuals $i = 1, \ldots, m$ are

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n_i} \frac{\partial \log \rho_i(t_{ij};\theta)}{\partial \theta} - \int_0^{\tau_i} \frac{\partial \rho_i(t;\theta)}{\partial \theta} dt \right\}$$

- Trick: re-write $U(\theta)$ as

$$U(\theta) = \sum_{i=1}^{m} \int_0^{\tau_i} \frac{\partial \log \rho(t;\theta)}{\partial \theta} \{dN_i(t) - \rho_i(t;\theta)dt\} \qquad (4)$$

NOTE: $E\{U(\theta)\} = 0$ even if process is not Poisson.

- Approach 2 is similar to generalized estimating equation (GEE) methods where only means (and sometimes variances) are modelled. (Reference: Lawless and Nadeau, Technometrics 1995)

- This methodology is available for models (1) and (3) in R,S-PLUS and SAS, for the case where $\rho_0(t)$ is not specified parametrically.
  - When the process is a Poisson process this is called the Andersen-Gill (AG) model.
  - It is similar to the Cox model in survival analysis, and software for the Cox model (e.g. coxph, phreg) has been extended to cover both the AG model and robust estimation for (1) and (3).

4. Illustrations

- Mammary tumors in rats
  $x_i = 0$ if animal in control group and $= 1$ if in treatment group.

$$\rho_i(t) = \rho_0(t)\exp(\beta x_i) \qquad \mu_i(t) = \mu_0(t)\exp(\beta x_i)$$

$$\text{Note: } \exp(\beta) = \frac{\text{treatment } \rho(t)}{\text{control } \rho(t)} = \frac{\text{treatment } \mu(t)}{\text{control } \mu(t)}$$

- Following slide shows S-PLUS/R data frame and code

$$\hat{\beta} = -0.82, \ \text{Poisson s.e.} = 0.15, \text{robust s.e.} = 0.21.$$

$$\exp(\hat{\beta}) = .44$$

- Model checking can be carried out

The data for the first three rats in the treated group
are displayed below in the so-called "counting process" format.

```
>  rats[1:5, ]
    id start stop status   enum trt
  1  1     0  122      1      1   1
  2  2     0  122      0      1   1
  3  3     0    3      1      1   1
  4  3     3   88      1      2   1
  5  3    88  122      0      3   1
```

Robust Semiparametric Analysis

```
coxph(Surv(start,stop,status) ~ trt + cluster(id),
              data=rats, method="breslow")
   n= 254
            coef exp(coef) se(coef) robust se        z          p
   trt -0.815774  0.442297 0.151836   0.19809 -4.11819 3.8186e-05

       exp(coef) exp(-coef) lower .95 upper .95
   trt  0.442297    2.26092  0.299985  0.652122

   Likelihood ratio test= 31.69 on 1 df, p=1.81146e-08
   Wald test            = 16.96 on 1 df, p=3.8186e-05
   Score (logrank) test = 30.54 on 1 df, p=3.26554e-08, Robust = 11.2
p=0.000816617

   (Note: the likelihood ratio and score tests assume independence of
    observations within a cluster, the Wald and robust score tests do not).
```

RCT for Treatment of Herpes Simplex Virus Infections

- 48-week crossover trial (Romanowski et al. 2003)

  Group A: Suppressive then episodic therapy (24 weeks each)

  Group B: Episodic then suppressive

Episodic therapy - 1000 gms/day of valcyclovir only during an outbreak

Suppressive therapy - 500 gms/day every day; switch to 1000 gms/day during an outbreak

- Plot of estimated MCF's for Groups A and B, ignoring covariates (figure)

- Regression analysis:

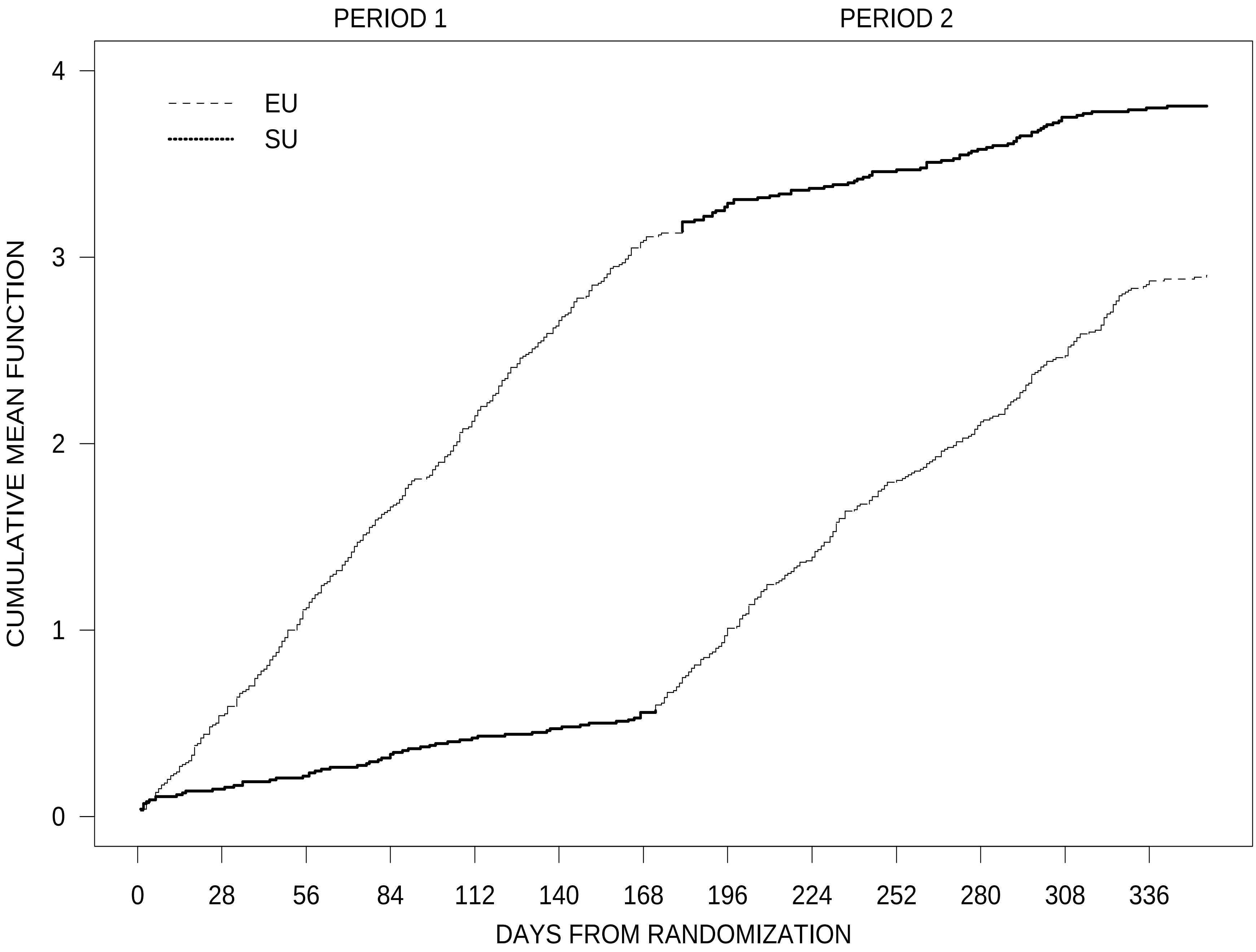
  $x_1(t) = I$ (on suppressive regime)

  $x_2(t)$ measures carryover effect of Suppressive Therapy in period 2 for Group A subjects

Covariate	$\hat{\beta}$	Robust SE	Z
Suppressive regime	- 1.58	0.11	- 14.6
Suppressive carryover	- 0.28	0.21	- 1.34
Age (years)	0.0007	0.0007	1.00
Sex ($M = 1$)	- 0.14	0.12	- 1.13
Race 1 (Hisp. vs Wh.)	- 1.13	0.71	- 1.60
Race 2 (Asian vs Wh.)	- 1.33	1.19	- 1.11
Virus type	0.20	0.11	1.79
Occurrences in prev. year	0.071	0.025	2.84

H

ALL PATIENTS



5. Extensions and Needed Development

- Robust estimation methods require the end-of-followup times τ_i to be independent of the recurrent events.
  Reason: robust method uses estimating equations (4),
  $$U(\theta) = \sum_{i=1}^{m} \int_0^{\infty} Y_i(t) \frac{\partial \log \rho_i(t)}{\partial \theta} \{dN_i(t) - \rho_i(t;\theta)dt\} = 0$$
  where $Y_i(t) = I(t \leq \tau_i)$.
  - If $\{Y_i(t), t \geq 0\}$ is independent of $\{N_i(t), t \geq 0\} = 0$ then
  $$E\{U(\theta)\} = 0 \text{ since } E\{dN_i(t)\} = \rho_i(t;\theta)dt.$$
  - Not true that $E\{U(\theta)\} = 0$ more generally.

- New approach: Use "inverse probability of censoring" weights

$$\pi_i(t) = \Pr\{Y_i(t) = 1 | H_i(t)\}$$

$$U_{iw}(\theta) = \int_0^\infty \frac{Y_i(t)}{\pi_i(t)} \frac{\partial \log \rho_i(t)}{\partial \theta} \{dN_i(t) - \rho_i(t;\theta)dt\}$$

- Note $E\{U_{iw}(\theta)\} = 0$ by taking $E_{H_i} E_{Y_i|H_i}$
  Cook and Lawless (2007)

- Intermittent observation: individuals observed at discrete time points so exact event times are not known.
  - parametric models are OK, but semiparametric model considered here is more difficult.

- Prediction of future events or costs

  e.g. Warranty claims, medical costs, software testing and debugging

  - Robust methods can produce only “point” predictions and estimates of rate or mean functions at future times

  - To get prediction intervals we require a probability model, which takes us beyond the present discussion

- Probability models

  - Poisson models with unit-level random effects, e.g.

    $$\mu_i(t|u_i, x_i) = \mu_0(t) u_i \exp\,(\beta' x_i)$$

    coxph with frailty(unit) option
    Quite robust; mimics robust analysis in many cases

  - Event intensity models: condition on past event history $H_i(t)$, e.g.

    $$\lambda_i\,(t|x_i, H_i(t)) = \lambda_0(t) \exp\,\left(\beta' x_i + \gamma N_i(t-)\right)$$

Final Remarks

- Robust methods discussed here are very useful when we wish to assess baseline covariates or treatments in randomized experiments

- Also valuable in observational studies for describing effects of fixed or external time-varying covariates

- More generally, intensity modelling is used to examine the dynamics of a process (effect of past event history on subsequent event occurrence)

e.g. $\lambda_i(t) = \lambda_0(t) \exp\left(x_i'\beta + \gamma N_i(t-)\right)$

Cook and Lawless (2007): wide range of methods