PoliSci 4782 Political Analysis II

Duration Outcome Models

Seamus Wagner

The Ohio State University

Duration Model/Survival Analysis/Event History

- Originated from epidemiology and public health studies
- Models duration (with an expectation that the chance of getting a particular outcome, often failure/death, is increasing over time)
- Examples of survival models in political science:
  - "survival" of political leaders
  - duration of military conflicts
- Consists of a group of models, most of which belong to generalized linear models (e.g. exponential, Weibull)
- Has its own lingo given its root in epidemiology, but essentially there is not so much new

Different Strategies

- The Kaplan-Meier estimator focuses on the survival function $Pr(T > t)$.
- Exponential and Weibull models focus on duration (or "failure time"/time-to-event).
- Some other models (e.g. the Cox model) focus on the hazard function $h(t)$, which captures the chance that a particular event occurs right at a give moment.

The Kaplan-Meier Estimator

The Kaplan-Meier Estimator

$$\hat{S}(t) = \prod_{i=1}^{t}(1 - \frac{d_i}{n_i})$$

- i is the index of time variable t.
- d is the number of failures/deaths that happened at time t.
- n is the cases known to have survived (not yet failed) up to time t.

In application, we use this estimator to do between-group analysis.

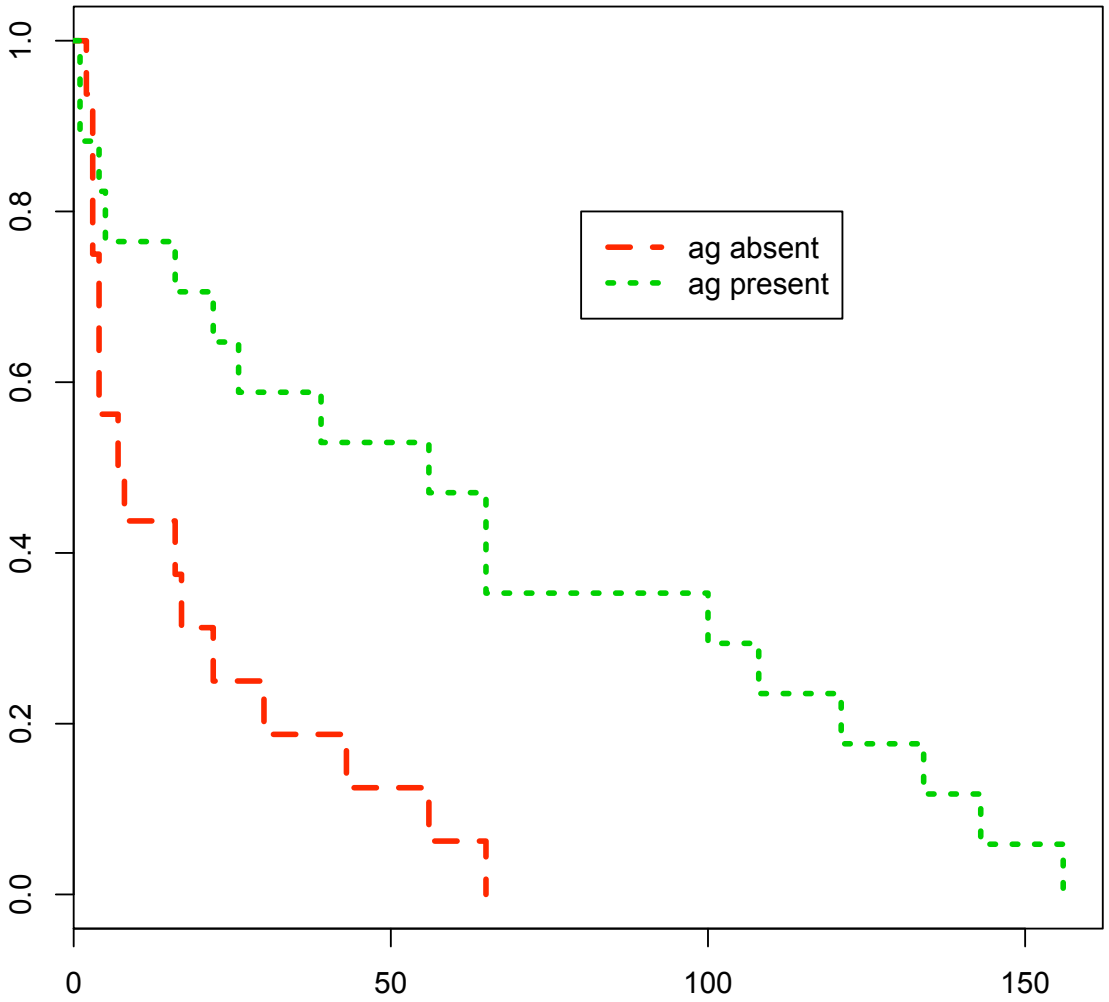
Application of Kaplan-Meier Estimator

Data of leukemia patients in an medical study:

- Unit of analysis is one patient.
- We know their survival times.
- We also know whether a patient was diagnosed as "present" or "absent" in terms of a morphologic characteristic of white blood cells ("A/G presence").

Question: Does "A/G presence" serve as an important indicator for patient survival?

Kaplan-Meier Survivor Analysis



The Exponential Model and the Weibull Model (GLMs on Duration)

Building GLM for Duration

- Again, we need to introduce a proper link function to equalize both sides of the equation.
- The range of the systematic component is always $(-\infty, +\infty)$ in theory.
- The mean of the outcome is always positive.

$$g(T) = \beta X$$

$$g(.) = ?$$

The Exponential and Weibull Models

- The exponential:

$$T = \exp^{(\beta X)}$$

or

$$log(T) = \beta X$$

- The Weibull:

$$\lambda T^{\lambda - 1} = \exp^{(\beta X)}$$

or

$$log(\lambda T^{\lambda - 1}) = \beta X$$

X in both models affect the duration T in a multiplicative way, because of the exponential component.

More about the Weibull

- Compared to the exponential model, the Weibull has one more parameter (λ), making its shape more flexible to the actual variation of data (similar to Poisson versus negative binomial).

- The Weibull model is a member of the **accelerated failure time (AFT)** models, in the sense that the effect of X is to accelerate or decelerate some baseline duration determined by λ.

Application

Data of leukemia patients in an medical study:

- Unit of analysis is one patient.
- We know their survival times.
- We also know whether a patient was diagnosed as "A/G present".
- One more covariate: white blood count of each patient (`wbc`).

The Exponential Model

```
Call:
survreg(formula = Surv(time) ~ ag + log(wbc), data = leuk, dist = "exponential")
             Value   Std. Error  z       p
(Intercept)  5.815       1.263  4.60 4.15e-06
agpresent    1.018       0.364  2.80 5.14e-03
log(wbc)    -0.304       0.124 -2.45 1.44e-02

Scale fixed at 1

Exponential distribution
Loglik(model)= -146.5   Loglik(intercept only)= -155.5
Chisq= 17.82 on 2 degrees of freedom, p= 0.00014
Number of Newton-Raphson Iterations: 5
n= 33
```

- A/G presence increases survival time or decelerate the speed of death by a factor of exp(1.018) (176.8% increase)
- One unit rise in white blood count on the log scale decreases survival time by a factor of exp(-0.304) (a multiplier of 0.738, so 26.2% decrease)

The Weibull Model

```
Call:
survreg(formula = Surv(time) ~ ag + log(wbc), data = leuk)
             Value    Std. Error  z        p
(Intercept)  5.8524       1.323  4.425 9.66e-06
agpresent    1.0206       0.378  2.699 6.95e-03
log(wbc)    -0.3103       0.131 -2.363 1.81e-02
Log(scale)   0.0399       0.139  0.287 7.74e-01

Scale= 1.04

Weibull distribution
Loglik(model)= -146.5   Loglik(intercept only)= -153.6
Chisq= 14.18 on 2 degrees of freedom, p= 0.00084
Number of Newton-Raphson Iterations: 6
n= 33
```

- Similar coefficient estimates, the same interpretation method (speak to survival time).
- The scale parameter is estimated than assumed this time, which is close to 1, so no big difference from the exponential model.

The Cox (proportional hazard) Model

The Cox Model

- Cox (1972) introduced a less parametric approach to model proportional hazards.
- The model assumes a baseline hazard function $h_0(t)$ that is modified multiplicatively by X.
- So the actual hazard $h(t)$ is

$$h(t) = h_0(t) \exp^{\beta X}$$

- The effect of X on the actual hazard is proportional to the baseline hazard (an important feature as well as the underlying assumption of the Cox model).
- Because we do not specify the formula of $h_0(t)$ but instead let data decide it, this model is a semi-parametric one.

Interpreting the Cox Model

$$h(t) = h_0(t) \exp^{\beta X}$$

can be rewritten as

$$\frac{h(t)}{h_0(t)} = \exp^{\beta X}$$

- We call $\frac{h(t)}{h_0(t)}$ a hazard ratio (HR), which is affected by X.
- If $\exp^{\beta} > 1$, the effect is an increased HR (more dangerous!).
- If $\exp^{\beta} < 1$, the effect is an decreased HR (less dangerous).
- Be careful: if X has a **positive** coefficient in the Weibull (on duration), it will almost certainly have a **negative** coefficient in the Cox (on HR)

The Cox Model in the Leukemia Example

```
Call:
coxph(formula = Surv(time) ~ ag + log(wbc), data = leuk)

n= 33
             coef   exp(coef) se(coef)  z     p
agpresent -1.069     0.343    0.429 -2.49 0.0130
log(wbc)   0.368     1.444    0.136  2.70 0.0069

             exp(coef) exp(-coef) lower .95 upper .95
agpresent        0.343      2.913     0.148     0.796
log(wbc)         1.444      0.692     1.106     1.886

Rsquare= 0.377   (max possible= 0.994 )
Likelihood ratio test= 15.6  on 2 df,   p=0.000401
Wald test            = 15.1  on 2 df,   p=0.000537
Score (logrank) test = 16.5  on 2 df,   p=0.000263
```

Use the column of `exp(coef)` for interpretation: A/G presence decreases the hazard ratio by a factor 0.343 (65.7% decrease)

Censoring

- The major distinguishing feature of survival analysis (from other GLMs that we learned) is *censoring*.
- Consider an individual case that is not observed throughout its whole life course. We do not know its true life time.
- In other words, censoring happens when part of the duration in our observations is truncated by our study.
- We must treat censored observations differently, not simply removing them or assuming their observed duration as their true "lifetime."

Dealing with Right Censoring in Theory

- Right censoring: it it the situation when the observed duration (t) does not really indicate the actual "lifetime" (l) because of the termination of research (oppositely, left censoring is "truncation on the left")

- With right censoring, we need to include an indicator for the event outcome in question (e.g. death/failure) and treat those censored observations separately:

$$L = \prod_{\delta_i=0} Pr(T > t) \prod_{\delta_i=1} S(t_i)$$

- For those uncensored ($\delta_i = 1$), we stick with the duration model in likelihood estimation; for those censored, we calculate their probability of living pass a certain time point instead.

Coming Up

- Lab on categorical and duration outcome models in R.
- With GLM in hand, we will start to consider how to improve their performance and better hit our research targets.