University of Toronto

STA450 (Final Exam Review):

Topics in Statistics: Statistical Methods for Emerging Infectious Disease Management

Professor: Patrick Brown
Notes By: Jiaqi Bi

Winter 2022

Last Updated: April 24, 2022

Contents

1 Introduction to SIR Model

1.1 Classic SIR Compartmental Model

There are three states of a person within an infectious disease period:

- S_t is the proportion of susceptible individuals in the population at time t
- I_t is the proportion of infectious individuals in the population at time t
- R_t is the proportion of removed individuals in the population at time t

Note: $S_t + I_t + R_t = 1$ for all t

Definition 1.1 (SIR ODEs)**.** The assumption of three states lead us to a set of three ODEs for S_t, I_t, and R_t:

$$\frac{dS_t}{dt} = -\beta S_t I_t \tag{1.1}$$

$$\frac{dI_t}{dt} = \beta S_t I_t - \gamma I_t \tag{1.2}$$

$$\frac{dR_t}{dt} = \gamma I_t \tag{1.3}$$

$\beta \geq 0$ is the transmission rate, $\gamma \geq 0$ is the removal rate.

1.2 Density-Dependent SIR Model

Definition 1.2 (Density-Dependent SIR)**.** The SIR model is now based on the fixed population size (N), where

- S_t is the **number** of susceptible individuals in the population at time t
- I_t is the **number** of infectious individuals in the population at time t
- R_t is the **number** of removed individuals in the population at time t

Note that $S_t + I_t + R_t = N$ for all t

2 Non-Parametric Models

Non-Parametric models are typically time series models, with or without seasonal effects. However, in this course, we have only seen non-parametric models with seasonal effects. Thus, when constructing the model, we have to remember to define the covariates with different sinusoidal functions.

Example 2.1. According to the CO2 slides (co2cosines at Lec 7), the annual cycles looks sinusoidal:

$$s(t) = \rho \cos(2\pi t/365.25 + \phi) \tag{2.1}$$

while the professor defines 4 sinusoidal basis functions for covariates with proper R codes:

$$X_{i0} = 1 \tag{2.2}$$

$$X_{i1} = \cos(2\pi t_i/365.25) \tag{2.3}$$

$$X_{i2} = \sin(2\pi t_i/365.25) \tag{2.4}$$

$$X_{i3} = \cos(2\pi t_i/182.625) \tag{2.5}$$

$$X_{i4} = \sin(2\pi t_i/182.625) \tag{2.6}$$

3 Survival Analysis

3.1 Survival Analysis without Random Effects, with and without Censoring

This lecture introduces the survival analysis using Bayesian Inference with the example of Cricket data as shown in `survival lec8`. The survival analysis involves with **Density** and **Hazards**, and the distribution is set as **Weibull Distribution**.

Definition 3.1 (Density)**.** The density function illustrates what proportion of events happen at time t:

$$\pi(t) = \lim_{\delta \to 0} pr(t < Y < t + \delta)/\delta \tag{3.1}$$

Definition 3.2 (Hazards)**.** The hazards function estimates the probability of an event near time t given that we have made it to t:

$$h(t) = \lim_{\delta \to 0} pr(Y < t + \delta | Y > t)/\delta = \pi(t)/[1 - \int_0^t \pi(u)du] \tag{3.2}$$

Example 3.1 (Cricket Data)**.** While we are going to use the example of the slide in `survival lec8` of Cricket data. First thing we need to do is to justify our prior for α, and according to the graph of Weibulls with $E(Y) = 70$, and we know that Lifetimes are left-skewed, it seems $\alpha = 5$ is the most appropriate. Also, log-Normal$(\log(7.5), 2/3)$ prior seems reasonable. Them, we are able to construct our model with Bayesian inference:

$$Y_i \sim \text{Weibull}(\lambda_i, \alpha) \tag{3.3}$$

where λ_i is the scale parameter, and it has the equation:

$$\lambda_i = \exp(-\eta_i) \tag{3.4}$$

and

$$\eta_i = X_i\beta \tag{3.5}$$

Note that the default prior for β is $\beta \sim (0, 0.001)$. The prior for the setting in `inla` is when `"pc.prec"... param=c(a,b)`, $pr(\sigma > a) = b$. However, in the exam of STA442, we have seen the question is based on the survival analysis with censoring. Thus, we can also construct the censoring as a hierarchical model:

$$Z_i | Y_i, A_i = \min(Y_i, A_i) \tag{3.6}$$

where Z_i is the actual lifetime, and Y_i is the lifetime followed Weibull distribution, and A_i is the age for individual i despite the individual i's current state (either deceased or alive).

$$E_i | Y_i, A_i = I(Y_i < A_i) \tag{3.7}$$

and (3.7) is the event indicator. Then, we have the distribution:

$$Y_i \sim \text{Weibull}(g(\eta_i, \kappa), \kappa) \tag{3.8}$$

and

$$\eta_i = X_i\beta \tag{3.9}$$

According to the example 3.1 we can apply the likelihood as well:

$$L(\mathbf{Z}, \mathbf{E}; \beta, \kappa) = \prod_{i; E_i = 1} f(Z_i) \prod_{i; E_i = 0} \int_{Z_i}^{\infty} f(u)du \tag{3.10}$$

that when $E_i = 1$, then $Y_i = Z_i$ and it means this individual is deceased; when $E_i = 0$ then $Z_i < Y_i < \infty$ and it indicates the individual is alive.

3.2 Survival Analysis with Random Effects

Definition 3.3 (Hierarchical Survival Models with a Random Effect)**.** Some survival analysis has random effects such as historical factors... Therefore, we have the hierarchical survival models with random effects as following

$$Z_{ij}|Y_{ij}, A_{ij}, U_i = \min(Y_{ij}, A_{ij}) \tag{3.11}$$
$$E_{ij}|Y_{ij}, A_{ij}, U_i = I(Y_{ij} < A_{ij}) \tag{3.12}$$
$$Y_{ij}|U_i \sim \text{Weibull}(\lambda_{ij}, \alpha) \tag{3.13}$$
$$-\log(\lambda_{ij}) = \eta_{ij} = X_{ij}\beta + U_i \tag{3.14}$$
$$U_i \sim \text{N}(0, \sigma^2) \tag{3.15}$$

We have observations j nested within subjects i, and each subject has a random effect U_i. In survival models, random effects are usually called **frailties**. The random effect U_i is usually normal distributed. However, Gamma distributed frailties are more commonly used in the real life (but not INLA). While doing this, we need priors for β, α and σ.

Example 3.2 (Fiji Data)**.** This example can be found at slide in `survival lec8`. The lecture uses interval censoring to fit the model. The case studies about the marriage age of people in Fiji:

$$Z_{i2}|Y_i, B_i = \max(Y_i, B_i) \tag{3.16}$$
$$Z_{i1}|Y_i, A_i = \min(Y_i, A_i) \tag{3.17}$$

where Z_{i2} and Z_{i1} are the interval censorings with the marriage age interval within $(A_i, B_i]$.

$$E_i|Y_i, A_i, B_i = I(Y_i < A_i) \tag{3.18}$$

Similar to the example 3.1, E_i still plays the role of indicator. As long as it is a survival analysis, we have

$$Y_i \sim \text{Weibull}(g(\eta_i, \kappa), \kappa) \tag{3.19}$$

and

$$\eta_i = X_i\beta \tag{3.20}$$

Moreover, the software inla will produce the likelihood as following:

$$L(\mathbf{Z}, \mathbf{E}; \beta, \kappa) = \prod_{i;E_i=1} f(Z_i) \prod_{i;E_i=0} \int_{Z_{i1}}^{Z_{i2}} f(u)du \tag{3.21}$$

Example 3.3 (Smoking Data)**.** We want to study on the age at first smoking a cigar. Firstly, we are using one random effect with school level. Note that the hazard function is

$$h(x; \rho, \kappa) = \rho\kappa x^{\kappa-1} \tag{3.22}$$

We have

$$f(x; \rho, \kappa) = \rho\kappa x^{\kappa-1} \exp(-\rho x^{\kappa}) \tag{3.23}$$

We have the individual i at school j with the Weibull response:

$$Y_{ij} \sim f(\rho_{ij}, \kappa) \tag{3.24}$$

and

$$\rho_{ij} = \exp(\eta_i) \tag{3.25}$$

and

$$\eta_{ij} = X_{ij}\beta + U_i \tag{3.26}$$

and the school-level random effect follows:

$$U_i \sim \mathrm{N}(0, \sigma^2) \tag{3.27}$$

We can use the survival model with more than one random effects as well, with state i, school j and individual k:

$$Y_{ijk} \sim \text{Weibull}(\rho_{ijk}, \kappa) \tag{3.28}$$
$$\rho_{ijk} = \exp(-\eta_{ijk}) \tag{3.29}$$
$$\eta_{ijk} = X_{ijk}\beta + U_i + V_{ij} \tag{3.30}$$
$$U_i \sim \mathrm{N}(0, \sigma_U^2) \tag{3.31}$$
$$V_{ij} \sim \mathrm{N}(0, \sigma_V^2) \tag{3.32}$$

4 Generalized Linear Models

4.1 Ordinary Least Squares

How to write Ordinary Least Squares:

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2) \tag{4.1}$$
$$\mu_i = X_i^\top \beta \tag{4.2}$$

Y_i usually indicates the Y for individual i of the given graph. With a specific R code, usually `glm(y~x)`, we write

$$Y_i \sim \mathrm{N}(\beta_0 + X_i\beta_1, \sigma^2) \tag{4.3}$$

4.2 Generailized Linear Models

Definition 4.1 (GLM)**.**

$$Y_i \sim G(\mu_i, \theta) \tag{4.4}$$
$$h(\mu_i) = X_i^\top \beta \tag{4.5}$$

- G is the distribution of the response variable
- μ_i is the location parameter for observation i
- θ are the additional parameters for the density of G
- h is the link function
- X_i are covariates for observation i
- β is a vector of regression coefficients

Example 4.1. Let's use the example of equations (4.1) and (4.2): G is a Normal distribution, θ is the variance parameter, denoted σ^2, and h is the identity function.

4.3 Binomial (Logistic) Regression

Definition 4.2 (Binomial)**.**

$$Y_i \sim \text{Binomial}(N_i, \mu_i) \tag{4.6}$$

$$\log(\frac{\mu_i}{1-\mu_i}) = X_i\beta \tag{4.7}$$

Note that μ_i is between 0 and 1. G is a Binomial distribution, or say it's a Bernoulli if $N_i = 1$, h is the logit link.

4.4 The Use of θ

Recall

$$Y_i \sim G(\mu_i, \theta) \tag{4.8}$$

$$h(\mu_i) = X_i^\top \beta \tag{4.9}$$

When G is Poisson or Binomial, θ is not used. When G is exponential family, θ factors out, such as σ^2 for Gaussian models.

4.5 Interpretation of Logistic Models

$$Y_i \sim \text{Binomial}(N_i, \mu_i) \tag{4.10}$$

$$\log(\frac{\mu_i}{1-\mu_i}) = \sum_{p=1}^{P} X_{ip}\beta_p \tag{4.11}$$

$$(\frac{\mu_i}{1-\mu_i}) = \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}} \tag{4.12}$$

where μ_i is the probability of the event, $\log(\mu_i/(1-\mu_i)$ is the log-odds, $\mu_i/1-\mu_i$ is odds. If $\mu_i \approx 0$ then $\mu_i \approx \mu_i/(1-\mu_i)$.

5 Spatial Statistics

5.1 Geostatistical Model

Definition 5.1 (Geostatistical Model)**.** We write s_i as the location that observation Y_i was made at, and we have:

$$Y_i \sim \text{N}(\lambda(s_i), \tau^2) \tag{5.1}$$

$$\lambda(s) = \mu + X(s)\beta + U(s) \tag{5.2}$$

$$\text{cov}[U(s+h), U(s)] = \sigma^2\rho(h/\phi; v) \tag{5.3}$$

- $U(s)$ is the **Spatial Random Effect** or we say **Residual Spatial Variation** or **Signal Process** or **Latent Spatial Process**.
- $X(s)$ are **covariates** at s or say **explanatory variables**, or **independent variables**, or **confounders**, or **fixed effects**.

- τ^2 is the **observation variance** or **error variance**, **nugget effect**, **noise variance** or more correctly the **independent observation-level random term**.

Example 5.1 (Interpretation Example)**.** We have the following model:

$$Y_i \sim \mathrm{N}[\lambda(s_i), \tau^2] \tag{5.4}$$
$$\lambda(s) = \mu + X(s)\beta + U(s) \tag{5.5}$$
$$\mathrm{cov}[U(s+h), U(s)] = \sigma^2 \rho(h/\phi; v) \tag{5.6}$$

- Y_i: The mercury concentration measured at location s_i
- $X(s)$: A vector of altitude, nighttime light, and categorical variables of land use type at s
- τ: The error associated with mercury measurement, or very localized factors influencing mercury
- $U(s)$: The residual spatial variation, the difference between actual mercury concentration and what the covariates predict. Dependes on
  - σ Variability in residual variation
  - ϕ Range parameter
  - v shape parameter

5.2 The Generalized Linear Geostatistical Model

Definition 5.2 (The Generalized Linear Geostatistical Model)**.**

$$Y_i | U(s_i) \sim [\lambda(s_i), \theta] \tag{5.7}$$
$$g[\lambda(s_i)] = \beta X(s_i) + U(s_i) \tag{5.8}$$
$$\mathrm{cov}[U(s_i), U(s_j)] = \sigma^2 \rho[(s_i - s_j)/\phi, v] \tag{5.9}$$

- $U(s)$ is the Spatial Random Effect
- $X(s)$ are covariates
- Distribution of the observed Y_i conditional on the mean $\lambda(s_i)$
- g is the link function
- θ are additional parameters for π

Example 5.2 (Interpretation)**.** How to interpret?

- Y_i: Coffee shop type (1=latte...) at location s_i
- π: Bernoulli distribution
- $\lambda(s)$: Probability of a cafe at s serves lattes
- $X(s)$: A vector of covariates at s (proportion of Liberal voters, incomes...)
- g: logit transformation

- $U(s)$: The residual spatial variation, the difference between actual probability and what the covariates predict. Depends on
  - σ variability in residual variation
  - ϕ range parameter
  - v shape parameter

5.3 BYM Model

The BYM model is the most commonly seen in the real life. This is also something that can be highly on the test or exam.

Definition 5.3 (BYM Model)**.** The BYM model can be written as follows

$$Y_i \sim \text{Poisson}(E_i\lambda_i) \tag{5.10}$$
$$\log(\lambda_i) = \mu + X_i\beta + U_i \tag{5.11}$$
$$U_i \sim \text{BYM}(\sigma^2, \tau^2) \tag{5.12}$$
$$\theta_1 = \sqrt{\sigma^2 + \tau^2} \tag{5.13}$$
$$\theta_2 = \sigma/\sqrt{\sigma^2 + \tau^2} \tag{5.14}$$

Note that we need to identify the priors for θ_1 and θ_2:
usually we will see the code as `...sd=c(a,b), propSpatiao=c(c,d)...` then we will write

$$pr(\theta_1 > a) = b \tag{5.15}$$
$$pr(\theta_2 < c) = d \tag{5.16}$$

as the priors.

Example 5.3 (Voting in Georgia in 2016)**.** The model can be written as

$$Y_i \sim \text{Binomial}(N_i, \rho_i) \tag{5.17}$$
$$\log[\rho_i/(1-\rho_i)] = \mu + X_i\beta + U_i \tag{5.18}$$
$$U_i = \text{BYM}(\sigma^2, \phi) \tag{5.19}$$

where the N_i is the number of voters in the Republican primary in county i, Y_i is the number of Trump voters.

6 Compartmental Models of Infectious Disease Epidemics

Definition 6.1 (SIR)**.** The system of SIR ODEs:

$$S' = -\beta SI/N \tag{6.1}$$
$$I' = \beta SI/N - \gamma I \tag{6.2}$$
$$R' = \gamma I \tag{6.3}$$

where $N = S + I + R$.

Besides, we need to know the Basic Reproduction Number as well:

Definition 6.2 (Basic Reproduction Number)**.**

$$\mathcal{R}_0 = \frac{\beta}{\gamma} \tag{6.4}$$

and the final size is:

$$\ln \frac{S(0)}{S(\infty)} = \mathcal{R}_0(1 - \frac{S(\infty)}{N}) \tag{6.5}$$