Cannabis Data Science

Saturday Morning Statistics #16

March 19th, 2022

Question of the day.

- Can we apply **survival analysis** to study the question: what affects the likelihood of a cultivator, processor, or retailer being successful? Hypothesized effects may include:
  - Average THC concentration of products;
  - Rent in the zip code;
  - Your ideas?

Survival Analysis

First, define

- The <u>amount of time</u> individual i has been at risk at time j is t_{ij}.
- An indicator variable to denote if an individual i has <u>exited</u>

$$d_{ij} = \begin{cases} 1 \text{ if individual } i \text{ exited in period } j, \\ 0 \text{ otherwise.} \end{cases}$$

The chances of individual i surviving period j is given by the **survival function**

$$S(t_{ij}) = P(T > t_{ij}).$$

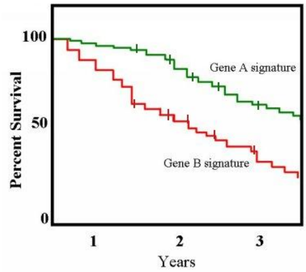
The instantaneous risk of failure for individual i in period j is the **hazard rate**

$$\lambda(t) = \frac{S'(t)}{S(t)}.$$

Kaplan–Meier Estimator

- Used to estimate survival functions.

- One of the most frequently used methods of survival analysis.



- The estimator is given by

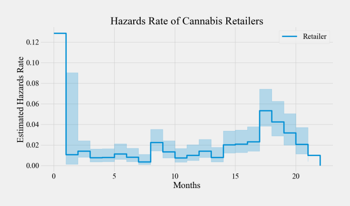
$$\hat{S}(t) = \sum_{t_i=0}^{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where

  - t_i is exposure time,
  - d_i is the number of events at time t_i,
  - n_i is the number of individuals surviving through time t_i.

Nelson–Aalen Estimator

An estimator of the <u>cumulative hazard rate function</u> given *censored data* or *incomplete data*.



- The estimator is given by

$$\hat{H}(t) = \sum_{t_i=0}^{t_i \leq t} \frac{d_i}{n_i}$$

where

- t_i is exposure time,
- $d_i$the number of events at time t_i,
- n_i is the number of individuals surviving through time t_i.

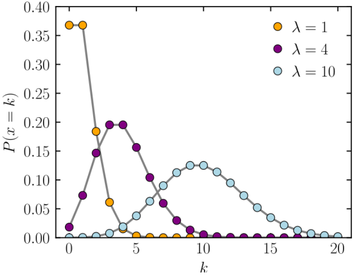
Poisson regressions to estimate hazards models

You can approximate d_{ij} with a **Poisson distribution**

$$d_{ij} \sim \text{Po}(\mu_{ij}) = \frac{\mu_{ij}^{d_{ij}} \exp(-\mu_{ij})}{d_{ij}!}$$

with means

$$\mu_{ij} = t_{ij}\lambda_{ij}.$$



The risk incurred by individual i in period j is

$$\lambda_{ij} = \lambda_j \exp(X_i\beta).$$

Cox's Proportional Hazards Model

Given covariates, X_i, and parameters, β, the hazard rate is modeled as

$$\lambda(t) = \lambda_0(t)\exp(X_i\beta),$$

where $\lambda_0(t)$ is the baseline hazard. Taking the log yields a Poisson log-linear model

$$\log\mu_{ij} = \log t_{ij} + \alpha_j + X_i'\beta.$$

A 1 unit increase in X_i is interpreted as a change in the average (and median) survival by a factor of $\exp(\beta)$.

Hazards Model with Time-varying Covariates

Adding time-varying covariates to the hazards model yields

$$\log\lambda_{ij} = \alpha_j + \beta X_{ij},$$

where X_{ij} are the values of the covariates of individual i in interval j.

Hazards Model with Time-dependent Effects

Adding time-dependent covariates to the hazards model yields

$$\log\lambda_{ij} = \alpha_j + \beta_j X_{ij},$$

where β_j is the effect of the hazard during interval j.

Note: Effects vary only at interval boundaries.

Bayesian Inference of Hazards Model

1. First, you specify your priors for the parameters

$$\begin{aligned} \beta &\sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \\ \sigma_\beta &\sim \mathcal{U}(a, b) \\ \lambda_j &\sim \Gamma(\alpha, \beta) \end{aligned}$$

   with hyperparameters μ_β, a, b, α, and β.

2. Second, you simulate draws from the posterior distributions.

3. Finally, you analyze and interpret your Bayesian estimates.

🙏 Thank you for coming.

Lessons of the Day

- Survival analysis can help identify factors that are and are not related to licensees exiting and their chances of surviving.