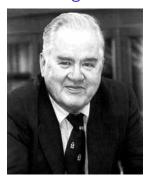


Cannabis Data Science

Saturday Morning Statistics #15

March 12th, 2022

A Brief Background



John Tukey (1915 - 2000) Professor of Statistics at Princeton University

Cannabis Data Science

John Tukey

- Created the box plot.
- Coined the term "bit".
- First published use of the word **software**.
- Creator of the <u>median-median line</u> (an alternative to the linear regression).
- Creator of the <u>trimean</u> measure of central tendency

$$TM = \frac{Q_1 + 2Q_2 + Q3}{4}$$

- Exploratory data analysis vs. confirmatory data analysis.
- The data should determine the methodology used.

Survival Analysis

Question of the day.

• Survival analysis has largely been pioneered by medical researchers to study <u>lifetimes</u>. Can we apply survival analysis to study the question: what is the natural <u>lifetime</u> of a retailer or producer in in the cannabis industry?

Survival function: S(t) = P(T > t).

Hazard function: $\lambda(t) = -\frac{S'(t)}{S(t)}$.

Poisson regressions have historically been used to approximate **proportional hazards models**.

- Calculation is quicker.
- Originally important when computers were <u>slower</u>.
- Also helpful with large data sets or complex models.

"we do not assume [the Poisson model] is true, but simply use it as a device for deriving the likelihood."

- Laird and Olivier (1981)

Kaplan–Meier Estimator

- Used to estimate survival functions.
- One of the most frequently used methods of survival analysis.
- The estimator is given by

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

where

Cannabis Data Science

- \triangleright t_i is exposure time.
- d_i is the number of events at time t_i.
- n_i is the number of individuals known to have survived up to 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

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

where

- t_i is exposure time,
- d_ithe number of events at time t_i,
- n_i is the number of individuals known to have survived up to time t_i .

Saturday Morning Statistics #15

Cox's Proportional Hazards Model

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

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

where $\lambda_0(t)$ is the baseline hazard.

A couple of important assumptions:

- The baseline hazard, $\lambda_0(t)$, is assumed to be independent of the covariate, x.
- The matrix of covariate, x, should not include a constant term.

Saturday Morning Statistics #15

Poisson regressions to approximate proportional hazards models

If you treat the event indicators, d_{ij} , as if they were independent <u>Poisson-distributed</u> observations with means

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

where t_{ij} is the exposure time and λ_{ij} is the hazard for individual i in interval j, then taking the log yields a Poisson log-linear model

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

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.

Assumptions:

Effects vary only at interval boundaries.

Bayesian Inference of Hazards Model

First, you specify your priors for the parameters

$$eta \sim \mathcal{N}(\mu_eta, \sigma_eta^2) \ \sigma_eta \sim \mathcal{U}(a,b) \ \lambda_j \sim \Gamma(lpha,eta)$$

with hyperparameters μ_{β} , a, b, α , and β .

- Second, you simulate draws from the posterior distributions.
- 3 Finally, you analyze and interpret your Bayesian estimates.



Lessons of the Day

- Borrowing from other fields is fruitful.
- Having the right tools (models) for the data at hand is critical.
- **Survive, then thrive** if you are a cannabis producer or retailer.