# Bayesian Cumulative Probability Models for Continuous Response Variables


Department of Biostatistics, Vanderbilt University

Nathan T. James, ScM   Frank E. Harrell, PhD   Bryan E Shepherd, PhD

Vanderbilt University Department of Biostatistics


Department of Biostatistics, Vanderbilt University

## Introduction

Cumulative Probability Models (CPM) are typically used for ordered categorical outcomes. Why use a Bayesian CPM with a **continuous** outcome?

1. Invariant to monotonic transformations
2. Models full conditional CDF; estimates of means and quantiles calculated from a single model
3. Handles ordered mix of discrete/continuous outcome values, e.g. lower limit of detection
4. Inference based on posterior probabilities

## Model

For observed ordered outcomes $y_1 \leq y_2 \leq \cdots \leq y_n$ the Cumulative Probability Model is:

$$G[P(Y \leq y_i|X)] = \alpha_i - \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}$$

where $\boldsymbol{X}$ is a matrix of covariates and $G(\cdot)$ a link function. The posterior distribution for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

$$\propto p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{i=1}^{n} [G^{-1}(\alpha_i - \beta x_i) - G^{-1}(\alpha_{i-1} - \beta x_i)]$$

Where $\alpha_i$ are ordered cutpoints which categorize the outcome ($\alpha_0 = -\infty$ and $\alpha_n = \infty$). For $Y$ with no ties, each category has one observation.

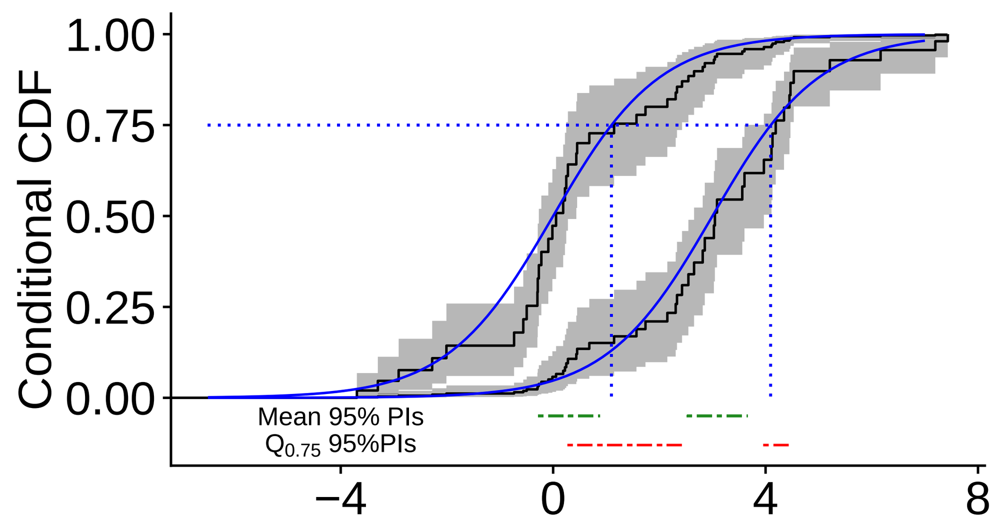
- `brms`: $\alpha_i \sim t(\nu = 3, \mu = 0, \sigma = 10)$ with ordering constraint
- `rstanarm`: let $\boldsymbol{\pi}$ be a simplex variable with $\pi_i = P(y = y_i|\bar{x})$ and Dirichlet pdf $\propto \prod_{i=1}^{n} \pi_i^{\gamma_i - 1}$ then $\alpha_i = G^{-1}\left(\sum_{j=1}^{i} \pi_j\right)$. By default, $\gamma_i = 1 \, \forall i$ (i.e., prior count of 1 in each bin)
- Flat (improper) priors used for $\beta$ with both

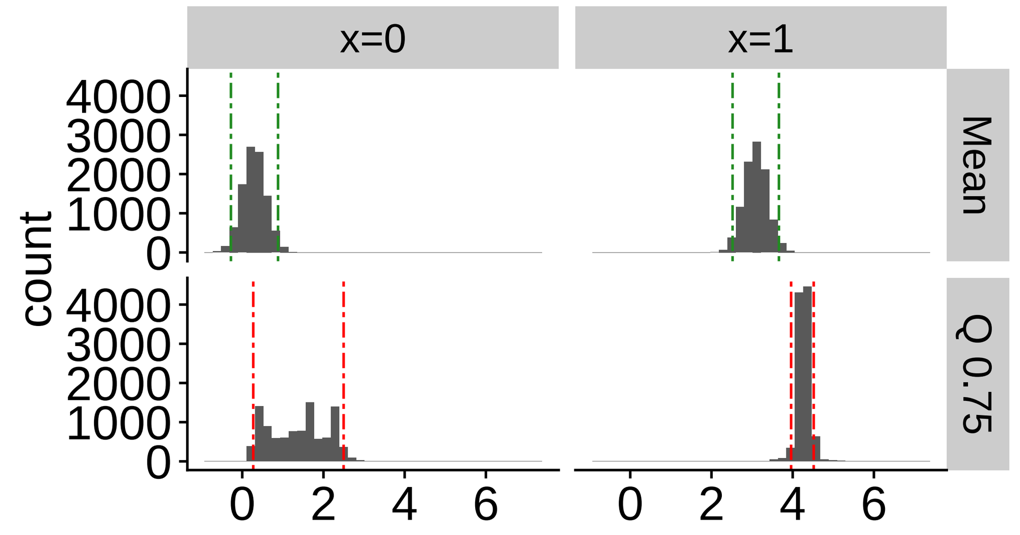
## Interpretation and Performance

- $\alpha_i$ estimate posterior CDF for $\boldsymbol{X} = 0$
- $\beta$ measure association between $\boldsymbol{X}$ and $\boldsymbol{Y}$; interpretation depends on link
  $\Rightarrow$ *ex.* if $G(\cdot)$ is logit link, $\beta$ are log-odds ratios
- Posterior mean and quantile estimates calculated from posterior conditional CDF

*Example*

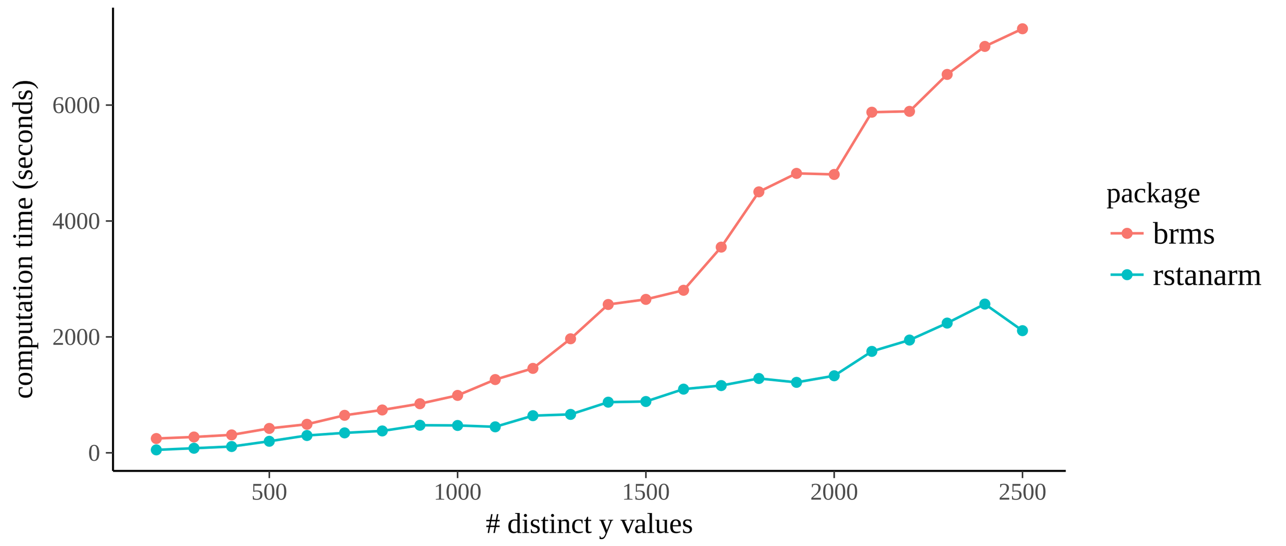
- Data for $n = 50$ observations generated from $Y = \beta X + \varepsilon$ with $\beta = 3$, $X \sim Bernoulli(0.5)$, and $\varepsilon \sim Logistic(0, 1)$
- 10,000 posterior MCMC draws produced using CPM with logit link $[G(p) = \log(\frac{p}{1-p})]$




## Computational Efficiency

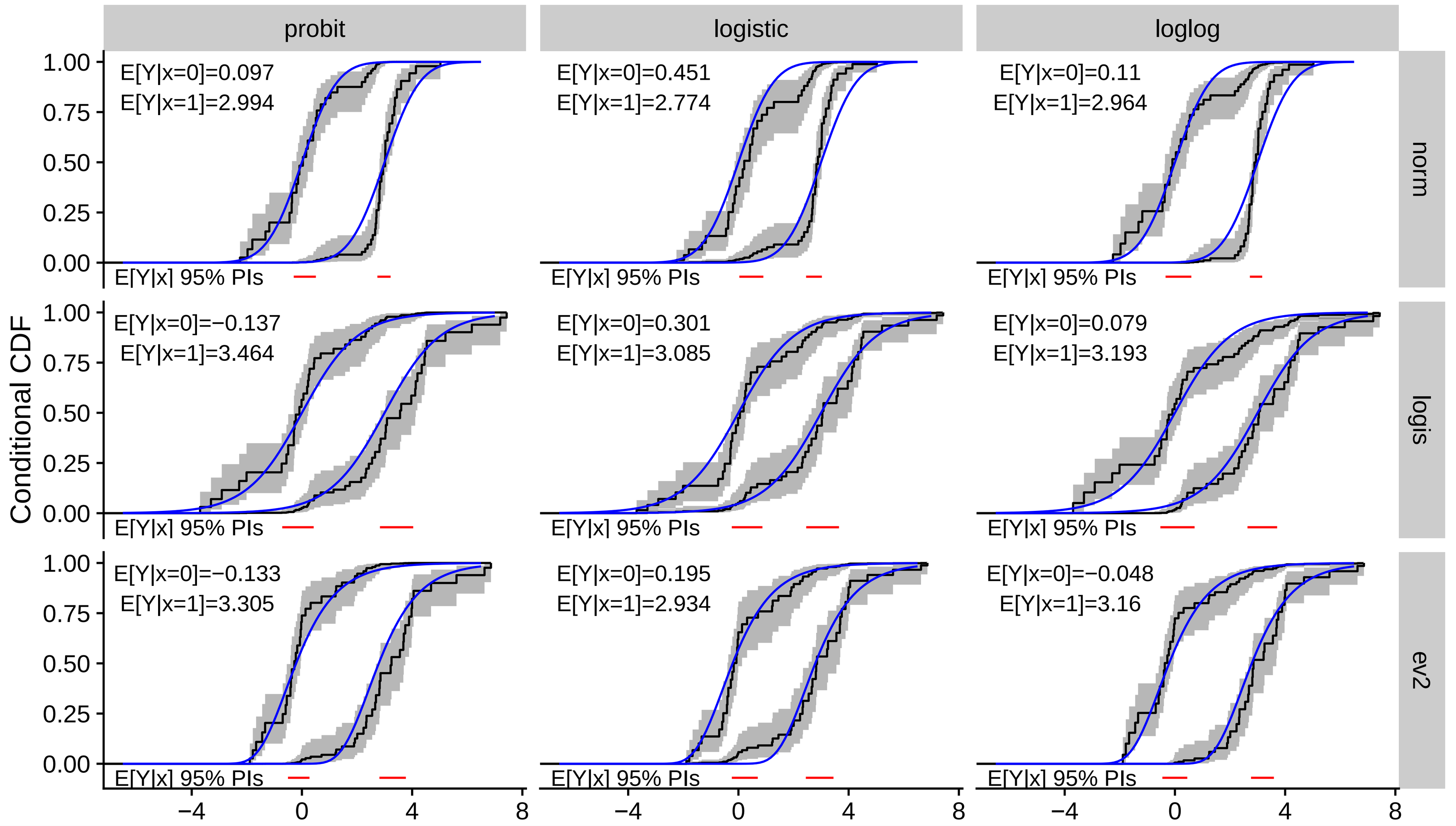
- `brms` needs to compile `C++` code, `rstanarm` uses pre-compiled code
- Convergence and speed depends on package and link; symmetric link functions faster than non-symmetric
- For moderate datasets time is approximately linear; in larger datasets compute time increases at a faster rate for `brms`




## Model Misspecification

*Example (cont.)*
$\varepsilon$ either *Normal*, *Logistic*, or *Extreme Value (type 2)* and CPM fit using probit, logit, or loglog link

- With moderate sample size, reasonably robust to link function misspecification
- Can account for uncertainty in link function using a mixture of links




## Conclusions

- Bayesian CPMs for continuous outcomes work best for small to moderate datasets
- Using default priors, `rstanarm` able to fit more link functions and more closely matches true generative model

## References


- Liu, Qi, et al. "Modeling continuous response variables using ordinal regression." *Stat. in Med.* 36.27 (2017): 4316-4335.
- Albert, James H., and Siddhartha Chib. "Bayesian analysis of binary and polychotomous response data." *JASA* 88.422 (1993): 669-679.