Generalized BART For Non-Numeric Outcomes

Antonio R. Linero

LEARNING OBJECTIVES

1. **What models can be used beyond regression?**
2. **How does the workflow change when we try to use other models?**
3. **What are the main benefits of using BART, generally?**

The BART Model

Original Model:

$$Y_i = g(X_i; \mathcal{T}_1, \mathcal{M}_1) + \cdots + g(X_i; \mathcal{T}_m, \mathcal{M}_m) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

What if outcome is:

- Binary?
- Ordinal?
- Survival?
- Etc?

Why I Like BART

- Fast
- Fully-Bayes inference
- Gets interactions
- Gets non-linearities
- Easy to use
- **Model selection uncertainty!!!**
- **But difficult to interpret!!!**

Other Models

$$Y_i \sim \text{NegBin}(k, e^{r(X_i)})$$

Count Data

$$\lambda(t \mid X_i) = \lambda_0(t)\ \exp\{r(X_i)\}$$

Cox Proportional Hazards

$$Y_i \sim \text{Gamma}(\alpha, \alpha e^{-r(X_i)})$$

Non-Negative Outcomes

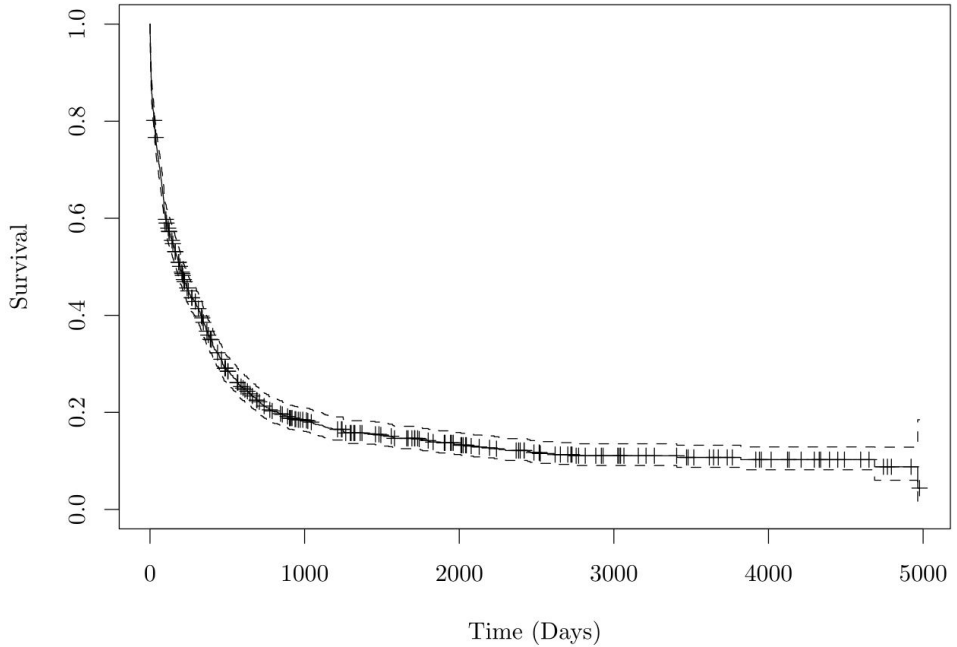
$$\Pr(Y_i = k \mid X_i) = \frac{e^{r_k(X_i)}}{\sum_j e^{r_j(X_i)}}$$

Multi-Category Outcomes

Survival Data: Leukemia

Marginal Distribution

Control for Covariates?



?LeukSurv in spBayesSurv

Two Seconds on Survival

Survival Function $\Pr(T > t \mid X = x) = S(t \mid x)$

Hazard Function $-\frac{d}{dt}\log S(t \mid x) = \lambda_0(t)\, e^{r(x)}$

Usually lots of right censoring!!!

$Y = \min(T, C)$ and $\delta = 1(T \leq C)$ model r with a BART!

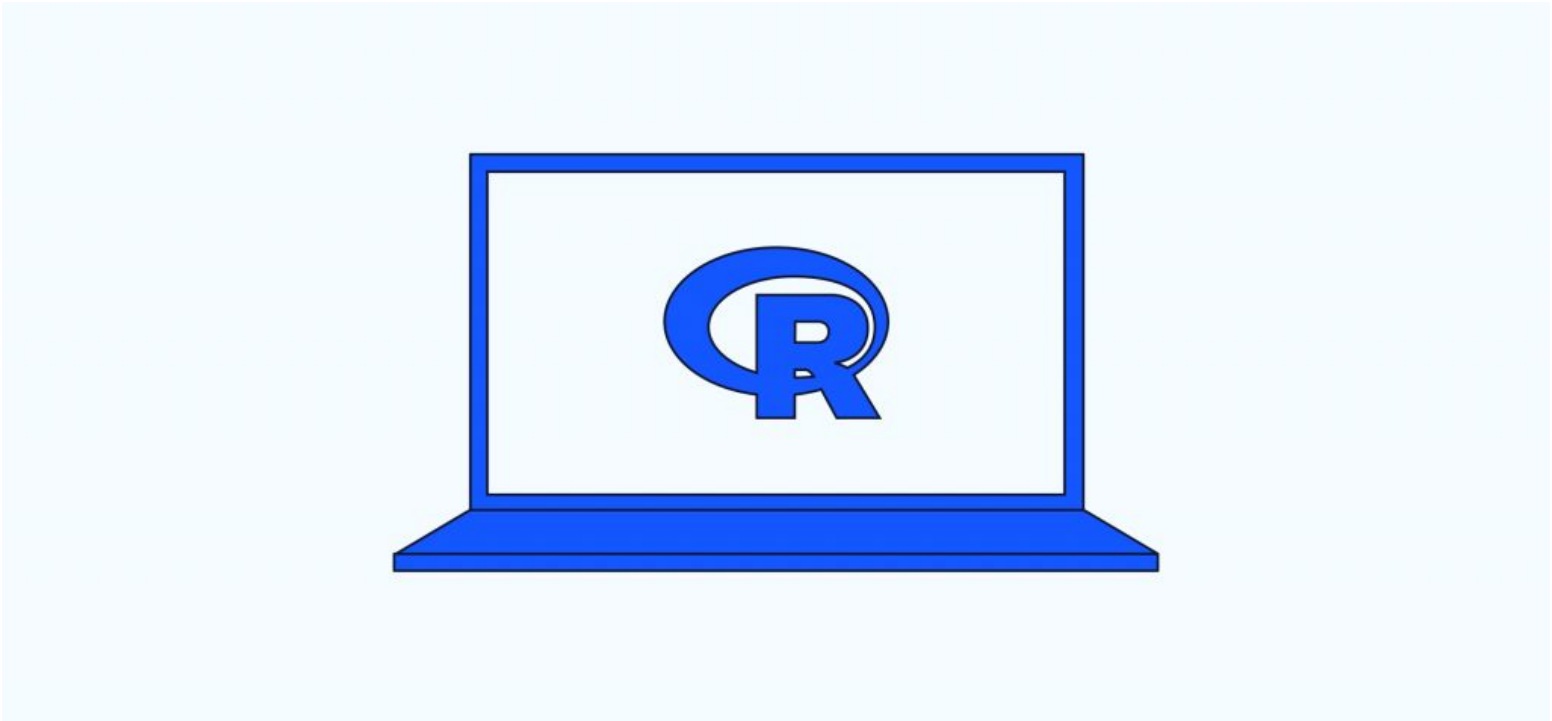
$$L = \prod_i \lambda_0(Y_i)^{\delta_i} \exp\left\{\delta_i r(X_i) - e^{r(X_i)} \int_0^{Y_i} \lambda_0(t)\, dt\right\}$$

Questions:

1. Is SES of living area associated with survival outcomes, and if so how?
    a. Measured in sample via the *Townsend Index*
2. More generally, what leads to better prognosis for individuals in the sample?
    a. We have *age*, *sex*, and *baseline white blood cell count* as additional prognostic factors
3. Do the relevance of any of the prognostic factors change as we look at longer time horizons?
    a. While generally quite lethal, the survival function estimates is suggestive of a *cured population*

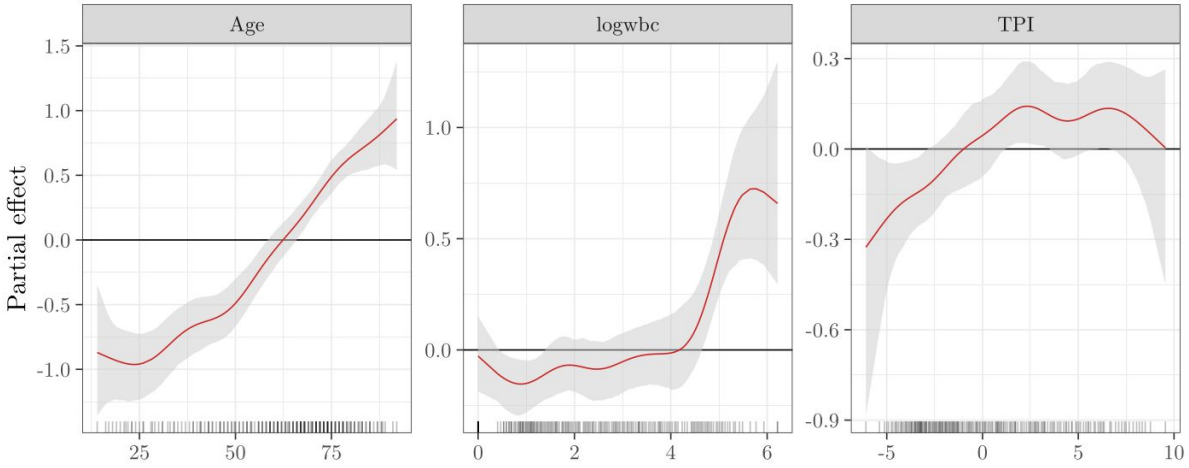
	time	cens	xcoord	ycoord	age	sex	wbc	tpi	district
1	1	1	0.205071665	0.497243660	61	0	13.3	-1.96	9
2	1	1	0.285556781	0.848952591	76	0	450.0	-3.39	7
3	1	1	0.176405733	0.736493936	74	0	154.0	-4.95	7
4	1	1	0.244762955	0.210584344	79	1	500.0	-1.40	24
5	1	1	0.327453142	0.907386990	83	1	160.0	-2.59	7
6	1	1	0.638368247	0.362734289	81	1	30.4	0.03	11
7	1	1	0.589856670	0.078280044	76	0	41.3	3.95	17
8	1	1	0.639470783	0.101433297	87	0	280.0	1.91	21
9	1	1	0.662624035	0.180815877	66	0	201.0	-3.50	18
10	1	1	0.152149945	0.873208379	78	1	3.9	0.38	7
11	1	1	0.589856670	0.127894157	57	0	0.0	6.70	17
12	1	1	0.235942668	0.549062845	87	1	1.4	-3.47	9
13	1	1	0.585446527	0.200661521	79	1	27.1	0.07	17
14	1	1	0.277839030	0.318632856	84	0	10.7	-2.06	4
15	1	1	0.117971334	0.831312018	77	0	291.0	3.17	7
16	1	1	0.656008820	0.309812569	69	1	181.0	4.87	19
17	1	1	0.625137817	0.026460860	64	0	36.6	1.12	21
18	1	1	0.552370452	0.109151047	67	1	149.0	3.19	23
19	1	1	0.267916207	0.482910695	60	0	0.0	4.44	9
20	1	1	0.320837927	0.218302095	53	0	1.3	4.90	24
21	1	1	0.302094818	0.244762955	55	0	159.0	-4.05	24
22	1	1	0.463065050	0.208379272	88	1	350.0	1.78	20

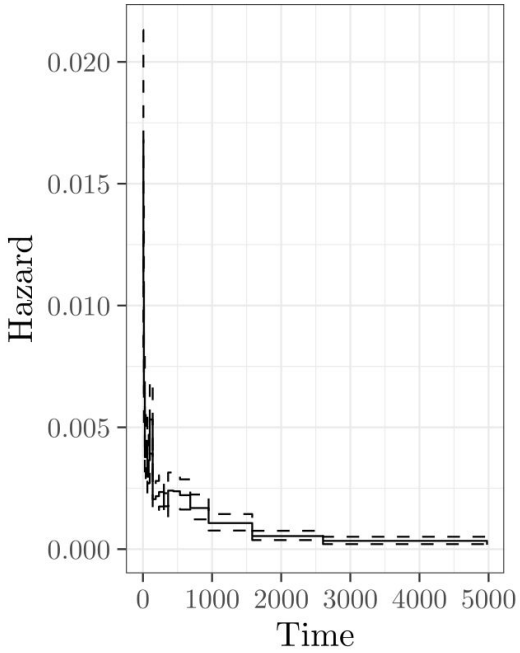
Go Over Code



Leukemia: Cox Proportional Hazards Model

Bigger means lower survival!





?LeukSurv in spBayesSurv

Models

Data Type	Model	Reference
Continuous	$\text{Normal}(\theta_x, \sigma^2)$	Chipman et al. (2010)
	$\text{Normal}(\theta_x, \tau_x^2\sigma^2)$	Pratola et al. (2020)
Quantile	$\text{ASL}_\tau(\theta_x, \sigma)$	Kindo et al. (2016)
Count	$\text{Poisson}(e^{\theta_x})$	Murray (2021)
	$\text{NegBin}(k, e^{\eta_x})$	Murray (2021)
Binomial	$\text{Binomial}\{n_i, \text{expit}(\theta_x)\}$	Murray (2021)
	$\text{Binomial}\{n_i, \Phi(\theta_x)\}$	Chipman et al. (2010)
Non-Negative	$\text{Gam}(\alpha, \alpha e^{-\theta_x})$	L. et al. (2020)
	Log $\text{Normal}(\theta_x, \sigma^2)$	Chipman et al. (2010)
Survival	Cox PH	L. et al. (2022)
	Fully Nonparametric	Sparapani et al. (2016)
	AFT	Sparapani et al. (2023)

Packages

- **BART** (or BART3 on GitHub)
- **Batman** (experimental, on GitHub) has *a lot* of models available
- Most BART packages can also do probit regression, but not much else…
  - dbarts
  - bartMachine
  - XBART (on GitHub)
  - *These packages are very fast!!!*

What If My Model Is Not On The List?

BAYESIAN ADDITIVE REGRESSION TREES FOR PROBABILISTIC PROGRAMMING

Miriana Quiroga IMASL-CONICET
Pablo G Garay IMASL-CONICET
Juan Martin Loyola IMASL-CONICET
Juan M. Alonso IMASL-CONICET
Osvaldo A Martin IMASL-CONICET-UNSL*
omarti@unsl.edu.ar

August 16, 2023

ABSTRACT

Bayesian additive regression trees (BART) is a non-parametric method to approximate functions. It is a black-box method based on the sum of many trees where priors are used to regularize inference, mainly by restricting trees' learning capacity so that no individual tree is able to explain the data, but rather the sum of trees. We discuss BART in the context of probabilistic programming languages (PPL), i.e., we present BART as a primitive that can be used as a component of a probabilistic model rather than as a standalone model. Specifically, we introduce the Python library PyMC-BART, which works by extending PyMC, a library for probabilistic programming. We showcase a few examples of models that can be built using PyMC-BART, discuss recommendations for the selection of hyperparameters, and finally, we close with limitations of our implementation and future directions for improvement.

Keywords Bayesian inference · non-parametrics · PyMC · Python · binary trees · ensemble method

Generalized Bayesian Additive Regression Trees Models: Beyond Conditional Conjugacy

Antonio R. Linero*

Abstract

Bayesian additive regression trees have seen increased interest in recent years due to their ability to combine machine learning techniques with principled uncertainty quantification. The Bayesian backfitting algorithm used to fit BART models, however, limits their application to a small class of models for which conditional conjugacy exists. In this article, we greatly expand the domain of applicability of BART to arbitrary *generalized BART* models by introducing a very simple, tuning-parameter-free, reversible jump Markov chain Monte Carlo algorithm. Our algorithm requires only that the user be able to compute the likelihood and (optionally) its gradient and Fisher information. The potential applications are very broad; we consider examples in survival analysis, structured heteroskedastic regression, and gamma shape regression.

What Changes?

Prior Specification

Algorithms

Model Interpretation

What Changes?

Hyperparameter selection:

$$\sigma_\mu^2 = ???$$

Conjugate prior usually also changes:

$$\mu \sim \log \mathrm{Gam}(a, b)$$

$$\psi(a) = \log b$$

$$\psi'(a) = \sigma_\mu^2$$

Why this prior?

What I Usually Do

First: add an offset term

$$r(x) = o + \sum_{t=1}^{m} g(x; \mathcal{T}_t, \mathcal{M}_t)$$

Second: Use a hyperprior

$$\sigma_\mu \sim \text{Exp}(\text{mean} = 1/\sqrt{m})$$

$$o \sim \text{Flat} \qquad \text{or} \qquad \text{fixed}$$

What Changes?

Algorithm 1 One iteration of a generalized Bayesian backfitting algorithm for updating $(\mathcal{T}_t, \mathcal{M}_t)$

Input: $\{\mathcal{T}_t, \mathcal{M}_t : t = 1, \ldots, T\}$, $\boldsymbol{Y}, \boldsymbol{X}, \eta, q(\cdot \mid \cdot)$

1: **for** $t = 1, \ldots, T$ **do**
2: Compute $\lambda_i \leftarrow \sum_{k \neq t} g(X_i; \mathcal{T}_k, \mathcal{M}_k)$ for $i = 1, \ldots, N$.
3: Propose a new tree structure $\mathcal{T}' \sim q(\mathcal{T}' \mid \mathcal{T}_t)$.
4: Compute the integrated likelihoods $\Lambda(\mathcal{T}_t)$ and $\Lambda(\mathcal{T}')$ where

$$\Lambda(\mathcal{T}) = \prod_{\ell \in \mathcal{L}(\mathcal{T})} \int \pi_\mu(\mu) \prod_{i: X_i \overset{\mathcal{T}}{\rightsquigarrow} \ell} f_\eta(Y_i \mid \lambda_i + \mu) \, d\mu.$$

5: Compute the acceptance probability

$$A = \min \left\{ \frac{\Lambda(\mathcal{T}') \, \pi_{\mathcal{T}}(\mathcal{T}') \, q(\mathcal{T}_t \mid \mathcal{T}')}{\Lambda(\mathcal{T}_t) \, \pi_{\mathcal{T}}(\mathcal{T}_t) \, q(\mathcal{T}' \mid \mathcal{T}_t)}, 1 \right\}.$$

6: With probability A, set $\mathcal{T}_t \leftarrow \mathcal{T}'$; otherwise, leave $\mathcal{T}_t$ unchanged.
7: Sample $\mathcal{M}_t$ from its full conditional distribution.
8: **end for**

Count Data: Crabs

- What makes a female crab attractive to males?
- Via *Categorical Data Analysis* (Agresti, 2012)

	color	spine	width	satell	weight
1	medium	bad	28.3	8	3050
2	dark	bad	22.5	0	1550
3	light	good	26.0	9	2300
4	dark	bad	24.8	0	2100
5	dark	bad	26.0	4	2600
6	medium	bad	23.8	0	2100
7	light	good	26.5	0	2350
8	dark	middle	24.7	0	1900
9	medium	good	23.7	0	1950
10	dark	bad	25.6	0	2150
11	dark	bad	24.3	0	2150
12	medium	bad	25.8	0	2650
13	medium	bad	28.2	11	3050
14	darker	middle	21.0	0	1850

Showing 1 to 15 of 173 entries, 6 total columns

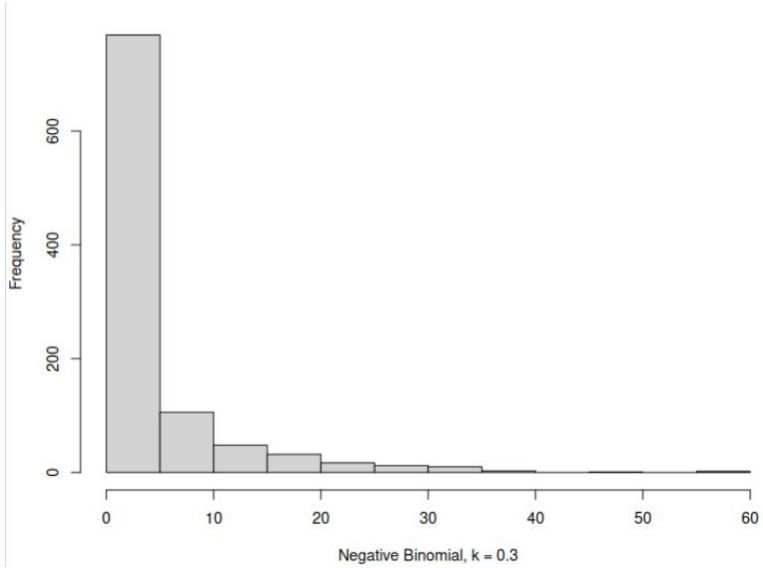
Console | Terminal × | Background Jobs ×

Negative Binomial Regression

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\xi_i) \\
\xi_i &\sim \text{Gamma}(k, k/\mu_i) \\
\mu_i &= \exp\{f(X_i)\} \\
f(X_i) &= \sum_t g(X_i; \mathcal{T}_t, \mathcal{M}_t)
\end{aligned}
$$

$$\text{Var}(Y_i) = \mu_i + \mu_i^2/k$$

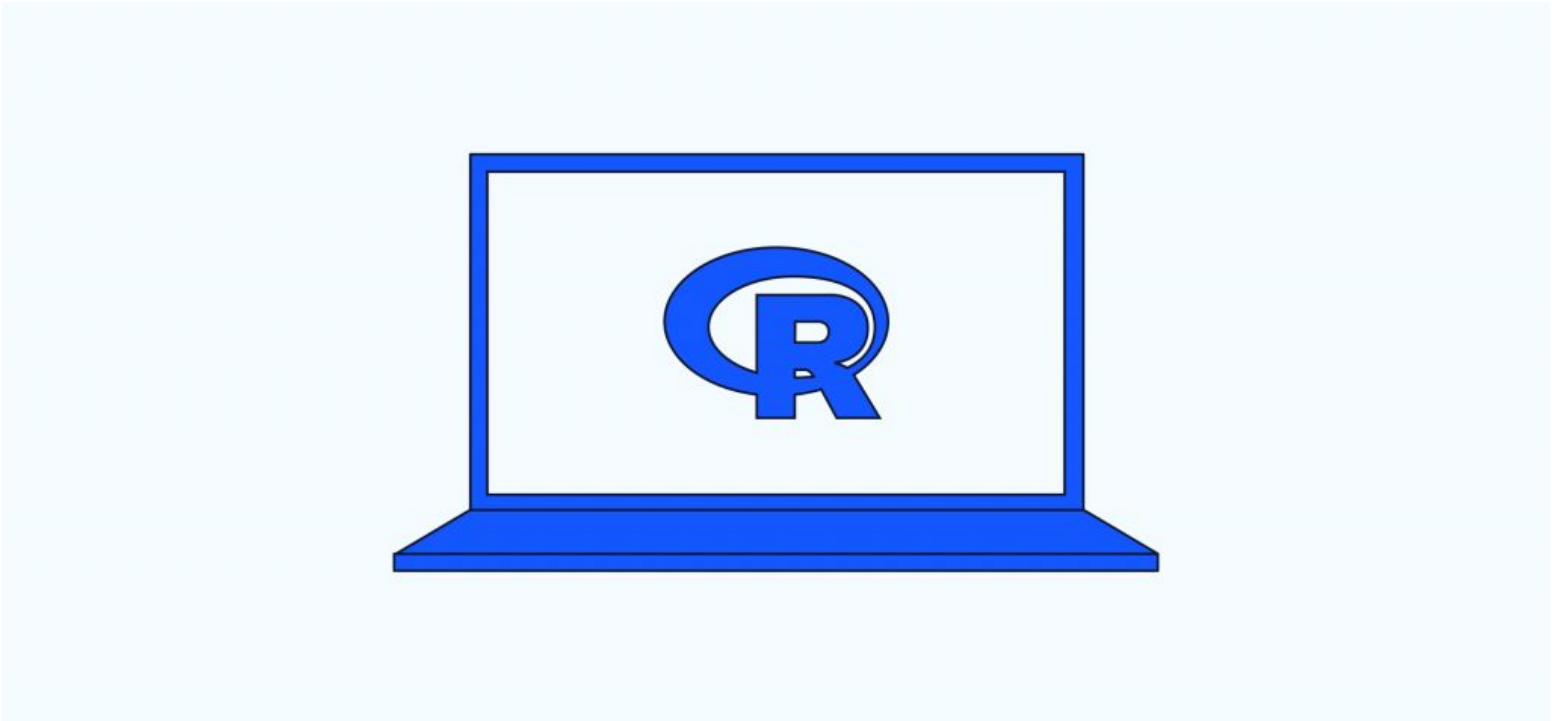
$$\frac{1}{\sqrt{k}} \sim \text{Exp}(1)$$



Questions:

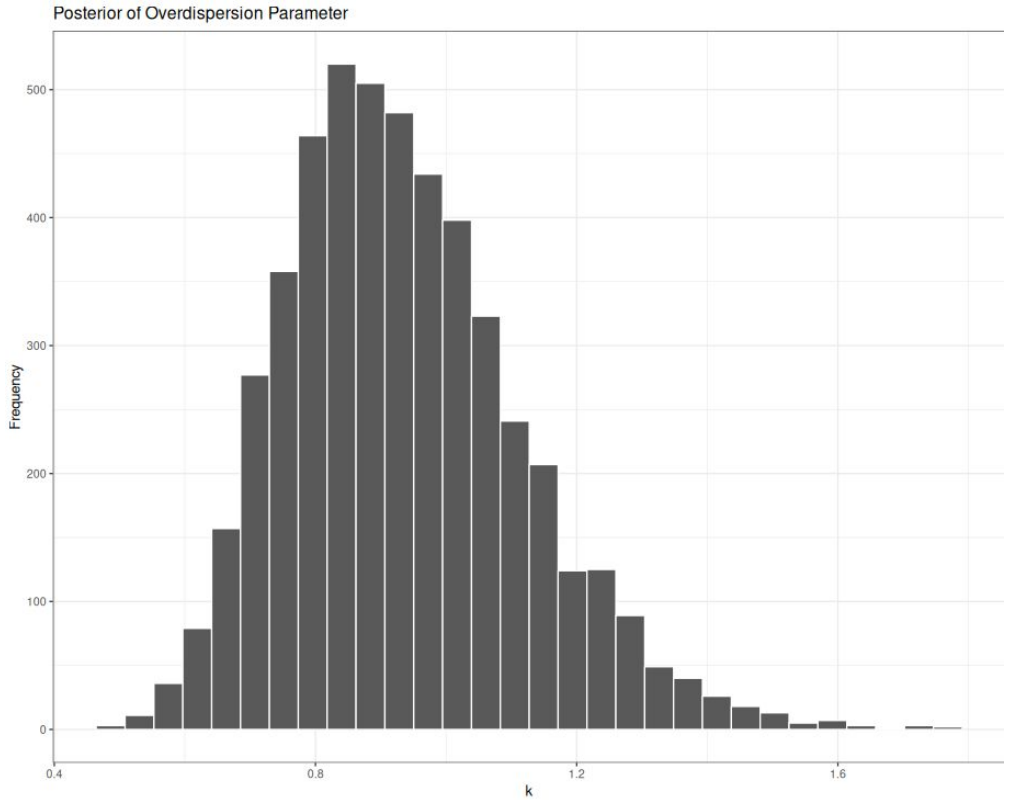
1. Are there likely unmeasured sources of variability we should be controlling for?
    a. *If not, we should expect Poisson outcomes*
2. Which features are most important in determining a crab's popularity?
3. What color is preferred, all else being equal?

Go Over Code



Overdispersion Posterior

Outcome is evidently overdispersed, consistent with a *geometric* rather than *Poisson* random variable

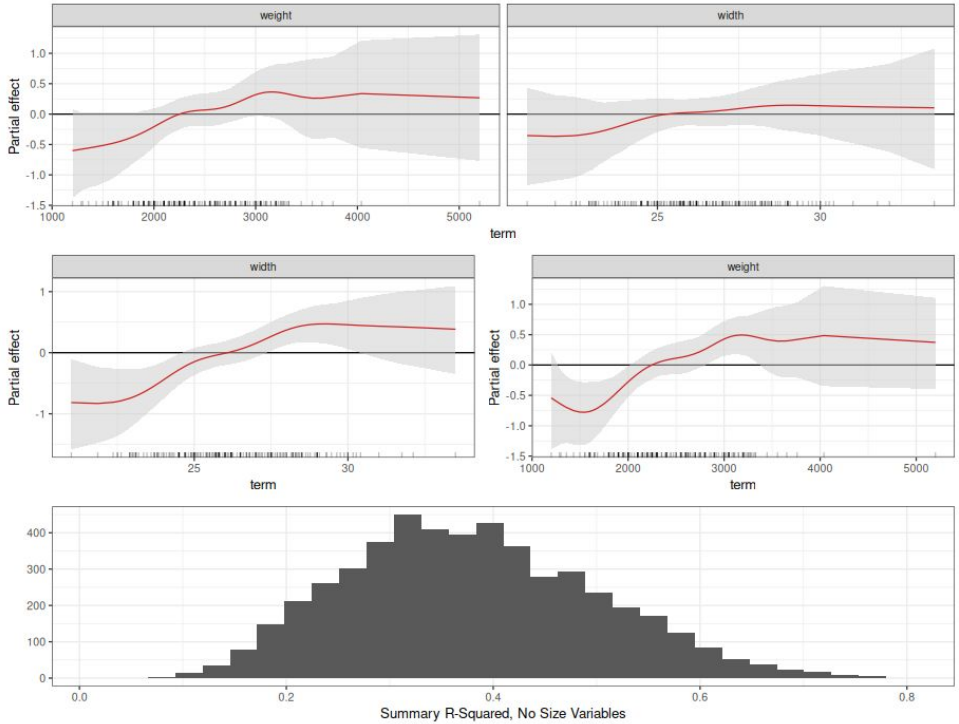


Assessing Size

Include Both

Drop 1

Drop Both



Color Contrasts

$$c_k = \frac{1}{N} \sum_i r(\text{color}_k, \text{spine}_i, \text{width}_i, \text{weight}_i)$$

$$d_k = c_k - \bar{c}$$