

Fancy stuff

David L Miller

A word of warning



Jenny Bryan

@JennyBryan

All models are wrong, so why not start with one you actually understand?

Away from the exponential
family

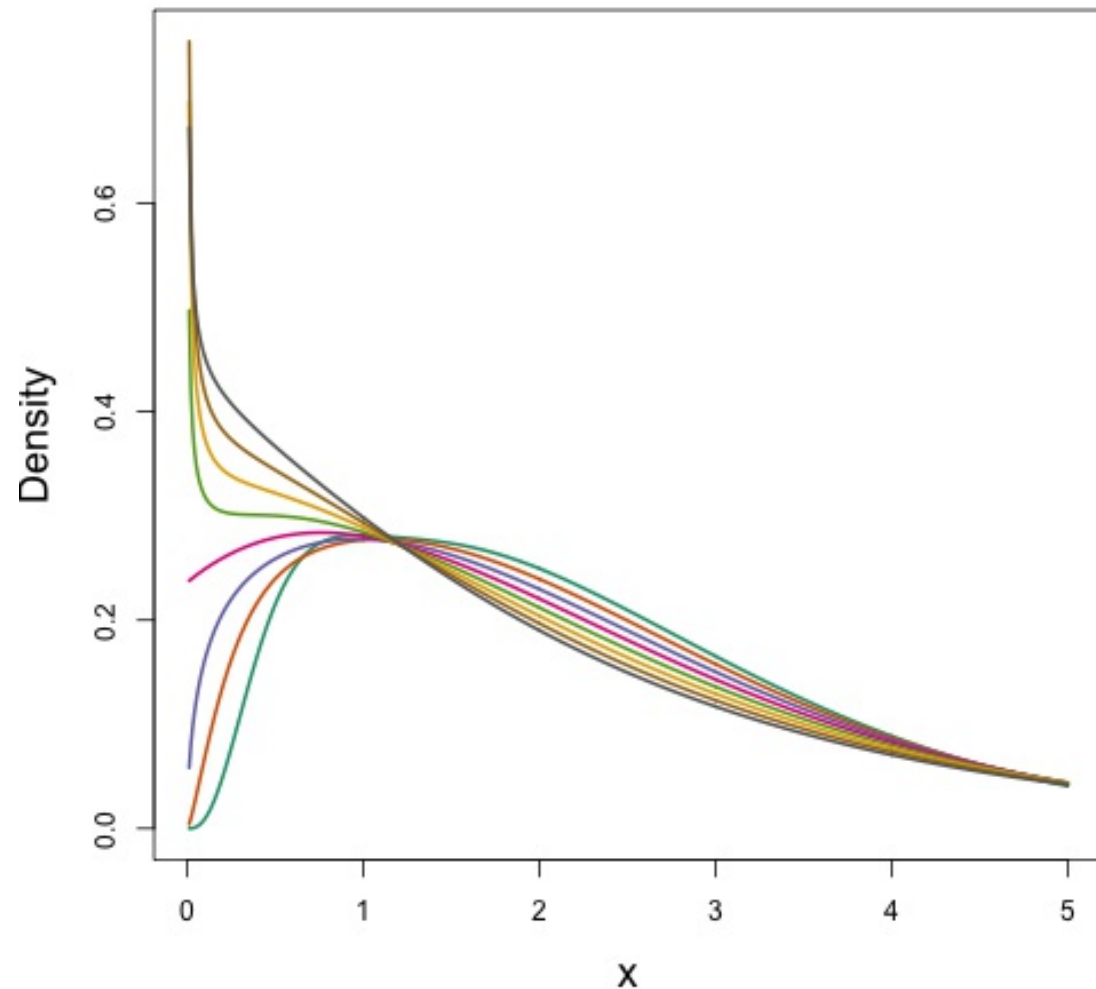
Modelling "counts"

Counts and count-like things

- Response is a count (not always integer)
- Often, it's mostly zero (that's complicated)
- Could also be catch per unit effort, biomass etc
- Flexible mean-variance relationship

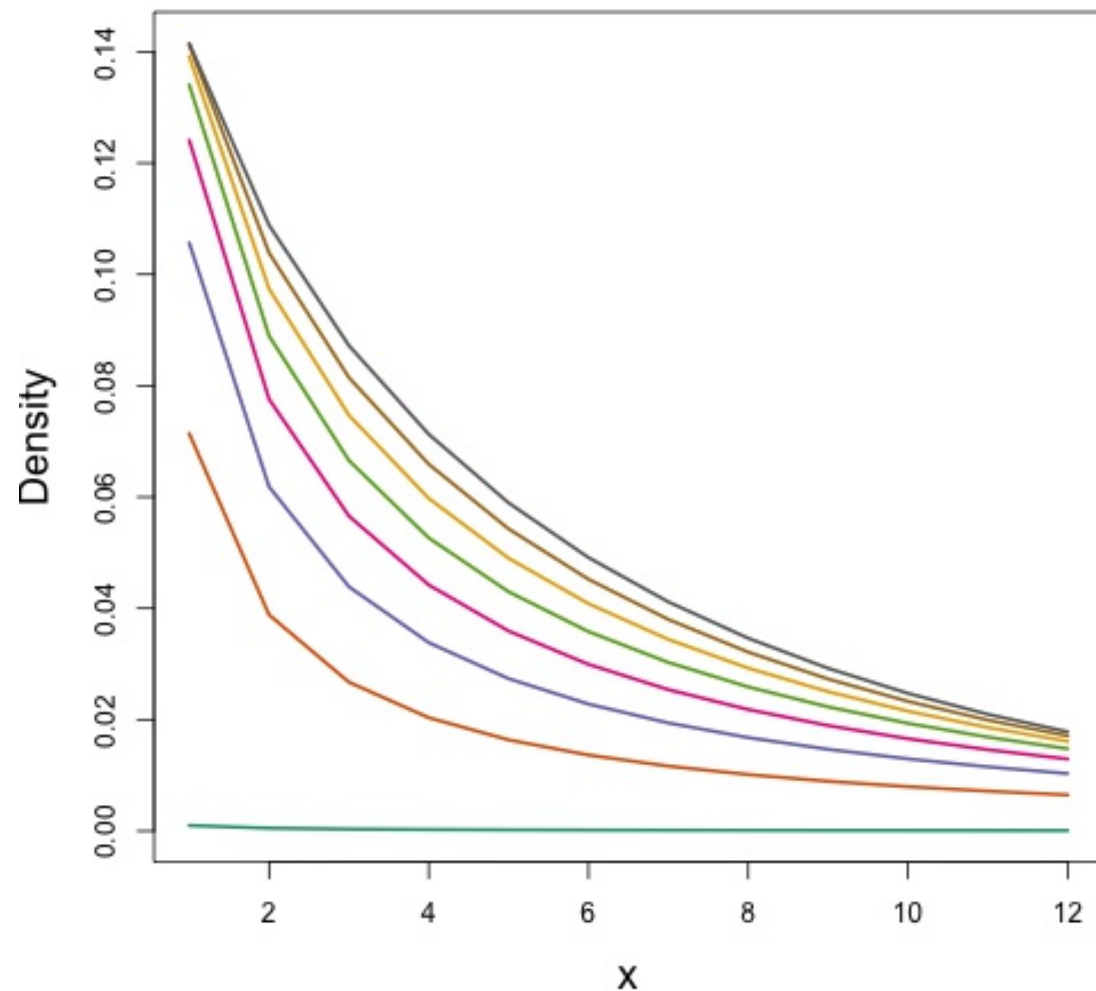


Tweedie distribution



- $\text{Var}(\text{count}) = \varphi(\text{count})^q$
- Common distributions are sub-cases:
 - $q = 1 \Rightarrow \text{Poisson}$
 - $q = 2 \Rightarrow \text{Gamma}$
 - $q = 3 \Rightarrow \text{Normal}$
- We are interested in $1 < q < 2$
- (here $q = 1.2, 1.3, \dots, 1.9$)
- `tw()`

Negative binomial



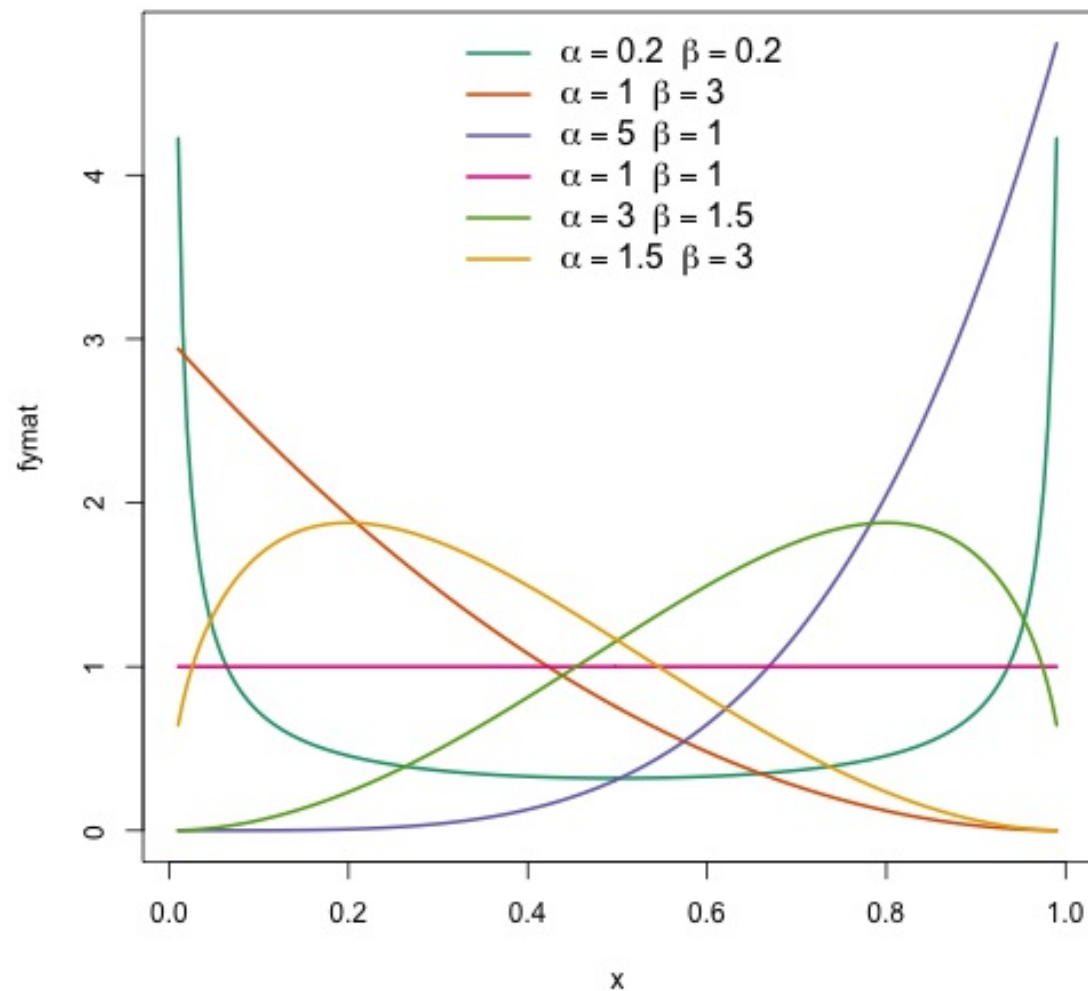
- $\text{Var}(\text{count}) = (\text{count}) + \kappa(\text{count})^2$
- Estimate κ
- Is quadratic relationship a “strong” assumption?
- Similar to Poisson:
 $\text{Var}(\text{count}) = (\text{count})$
- `nb()`

Zero-inflated distributions

- Models the probability of zeros separately from mean counts given that you've observed more than zero at a location.
- `ziP` and `zipLSS` (for location-scale models)
- zero inflation is assessed *conditional* on the model
 - is what you have zero inflation or just lots of zeros?
 - don't just jump straight to zero inflation

Other distributions

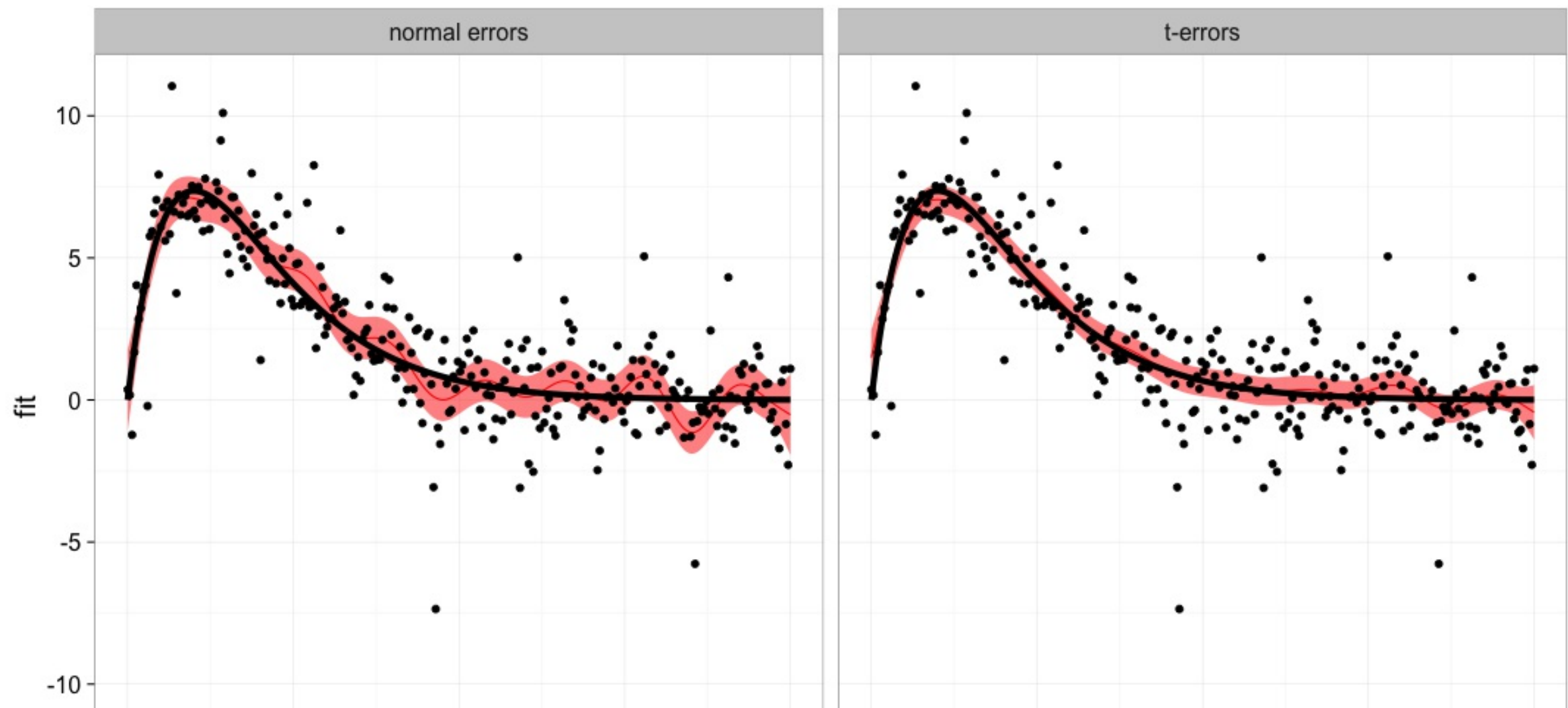
The Beta distribution



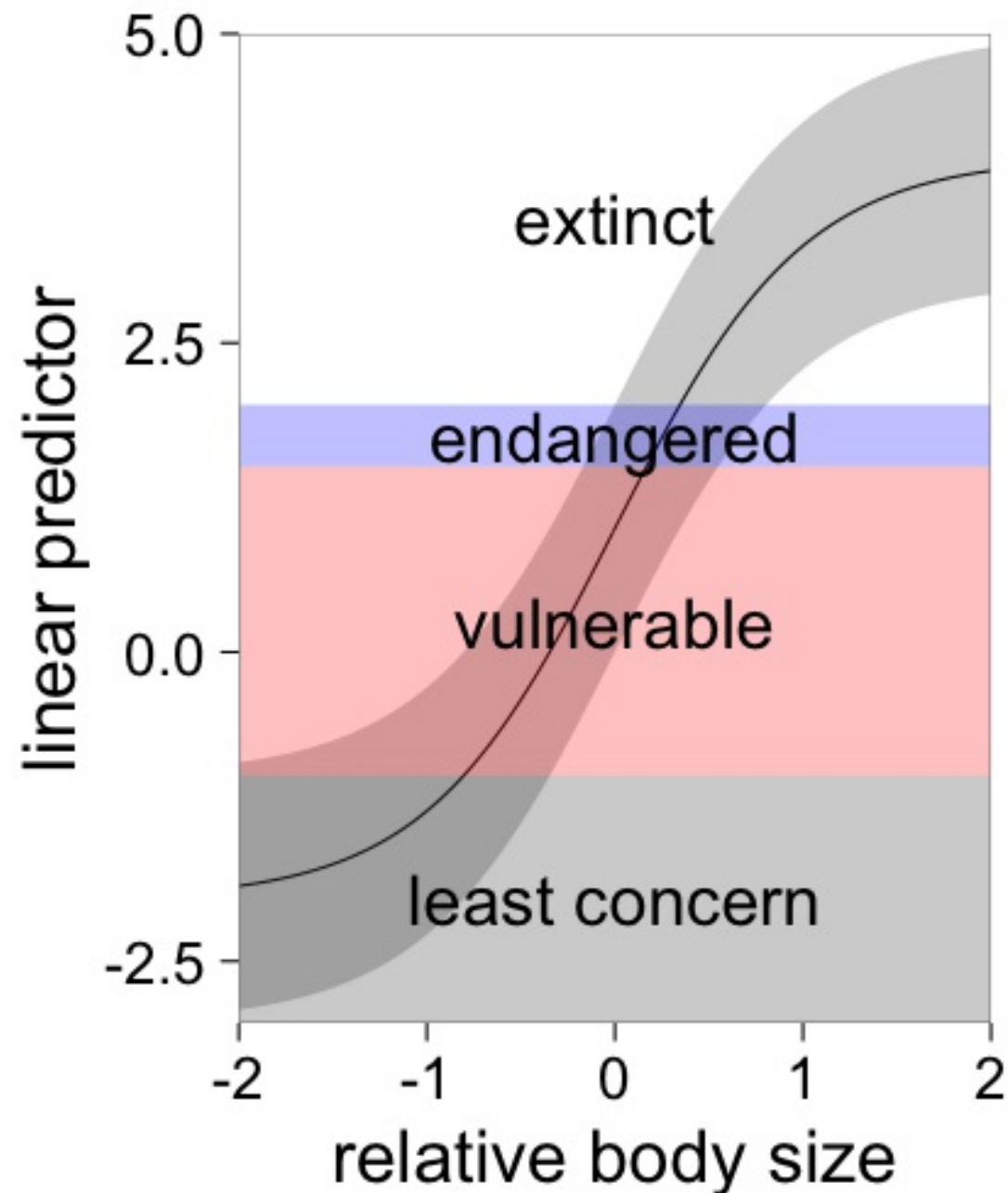
- Proportions; continuous, bounded at 0 & 1
- Beta distribution is convenient choice
- Two strictly positive shape parameters, α & β
- Has support on $x \in (0, 1)$
- Density at $x = 0$ & $x = 1$ is ∞ , fudge
- `betar()` family in **mgcv**

t-distribution

- Models continuous data w/ longer tails than normal
- Far less sensitive to outliers
- Has one extra parameter: df.
- bigger df: t dist approaches normal



Ordered categorical data



- Data are categories, have order
- e.g.: conservation status: “least concern”, “vulnerable”, “endangered”, “extinct”
- fits a linear latent model using covariates, w/ threshold for each level
- see ?ocat
- for unordered categories, see ?multinom

Other distributions (quickly)

- Multivariate normal (`family = "mvn"`)
 - Multivariate response, each has different smooth, allow correlation
- Cox proportional hazards (`family = cox.ph`)
 - Censored data: time until an event occurs, or the study was stopped
- Gaussian location-scale models (`family = gauLss`)
 - mean (“location”) and variance (“scale”) as smooths

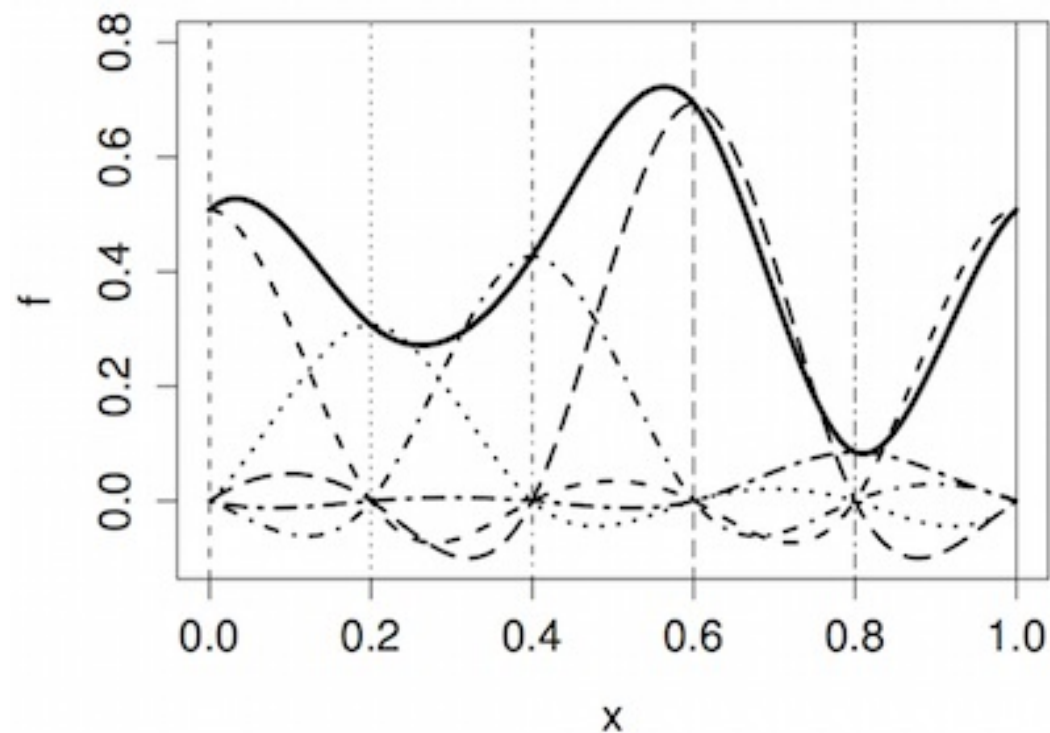
All of these distributions have quirks! Read the manual!

`?family` and `?family.mgcv`

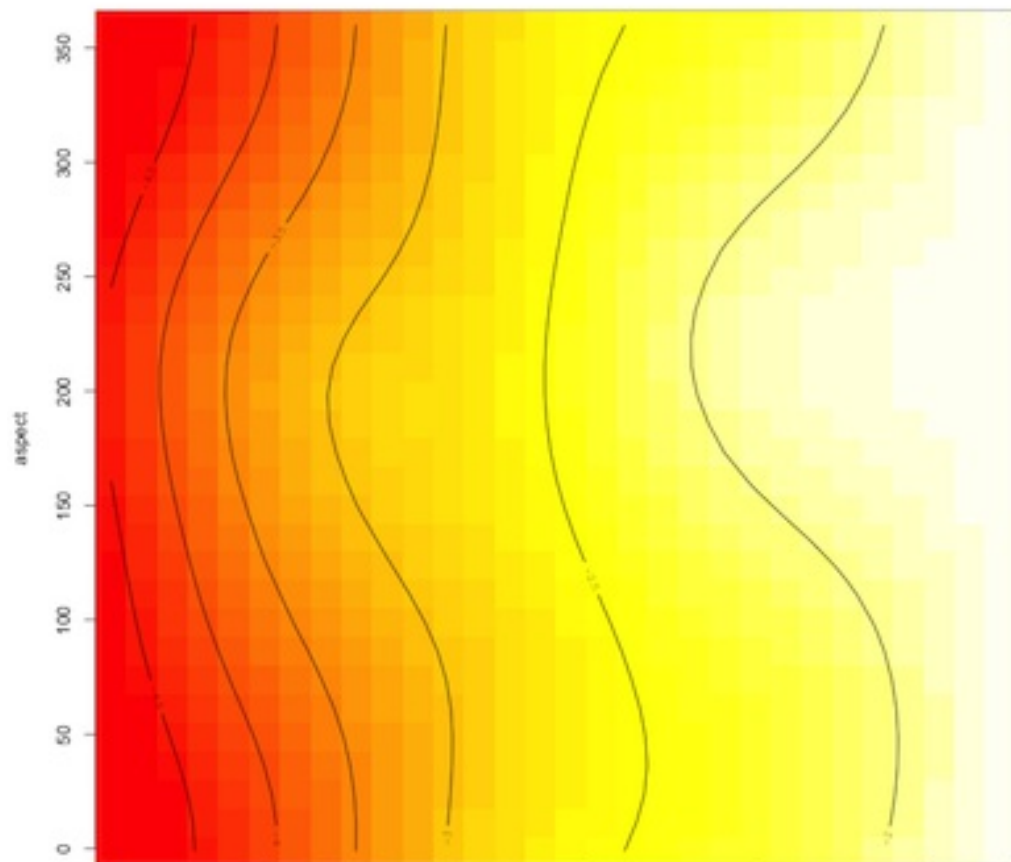
The end of the distribution zoo

Fancy smoothers

Cyclic smooths



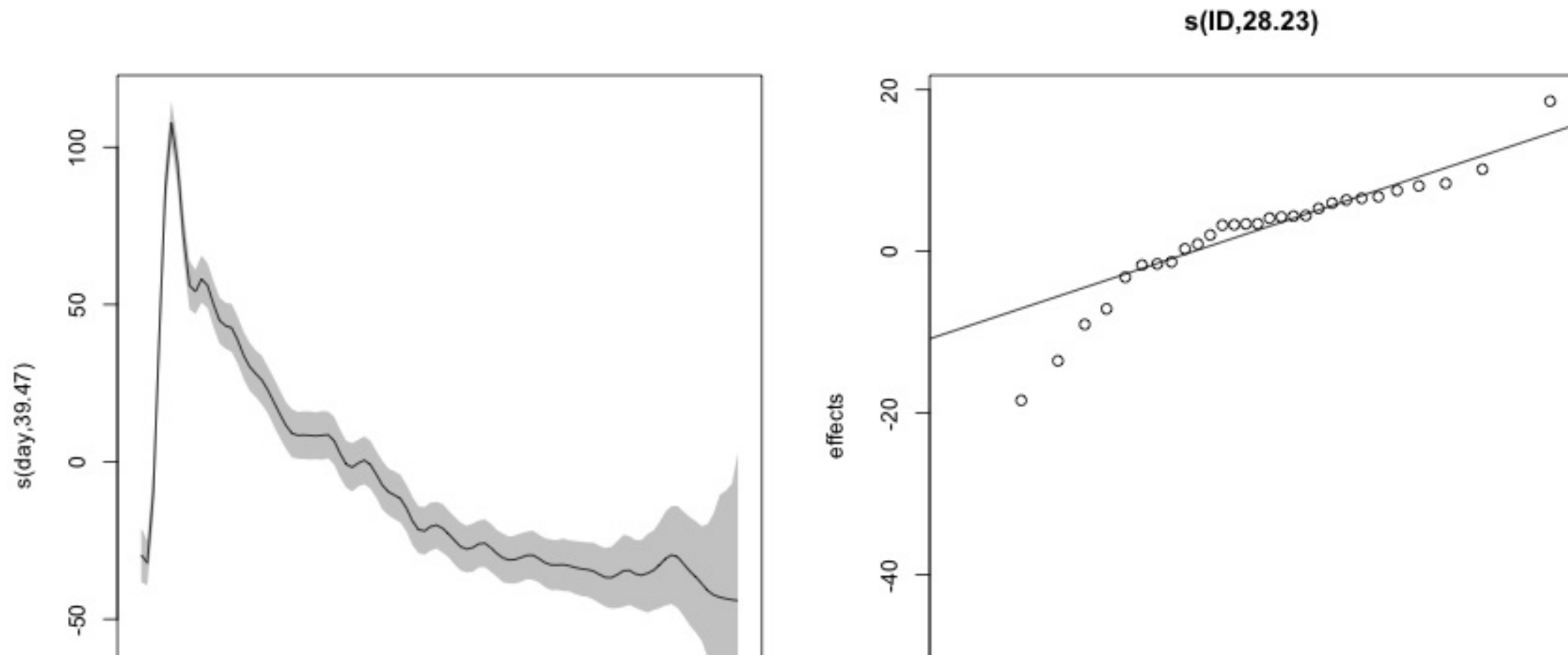
s(slope,aspect,6.037)



- cyclic smooths (`bs="cc"`, `bs`
 - what if smooths need to “mat
 - ensure up to 2nd derivs matc
 - need to be careful with end p
 - ?
- `smooth.construct.cc.s`

"Simple" random effects

- Earlier: "penalties can be thought of as variance components"
- We can think of random effects as splines too!
- in `mgcv` we can set `bs="re"`
- these are **simple**, non-nested random effects



Complicated random effects

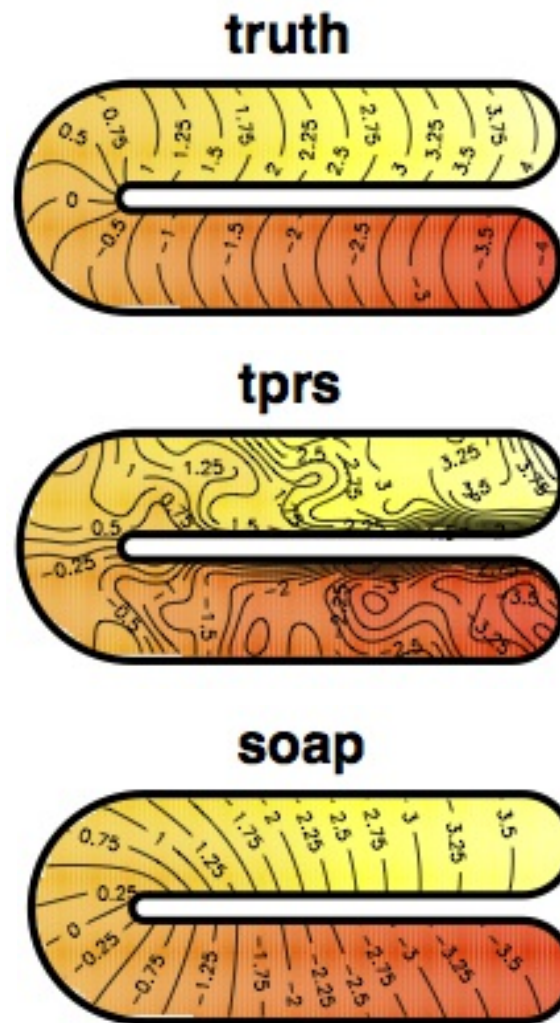
- `gamm` — uses spline-random effects equiv.
- cast splines as random effects, fit using `nlme`
- random effects are sparse, splines are dense
- often modelling problems with complex models
- `random= . . .` argument for nesting etc
- model has a `$gam` and `$lme` parts

Correlation structures

- again, need to use `gamm`
- `correlation=...` gives structure
- `corAR1`, `corARMA`, `corCAR1` etc
- tend to be hard to fit for SDMs

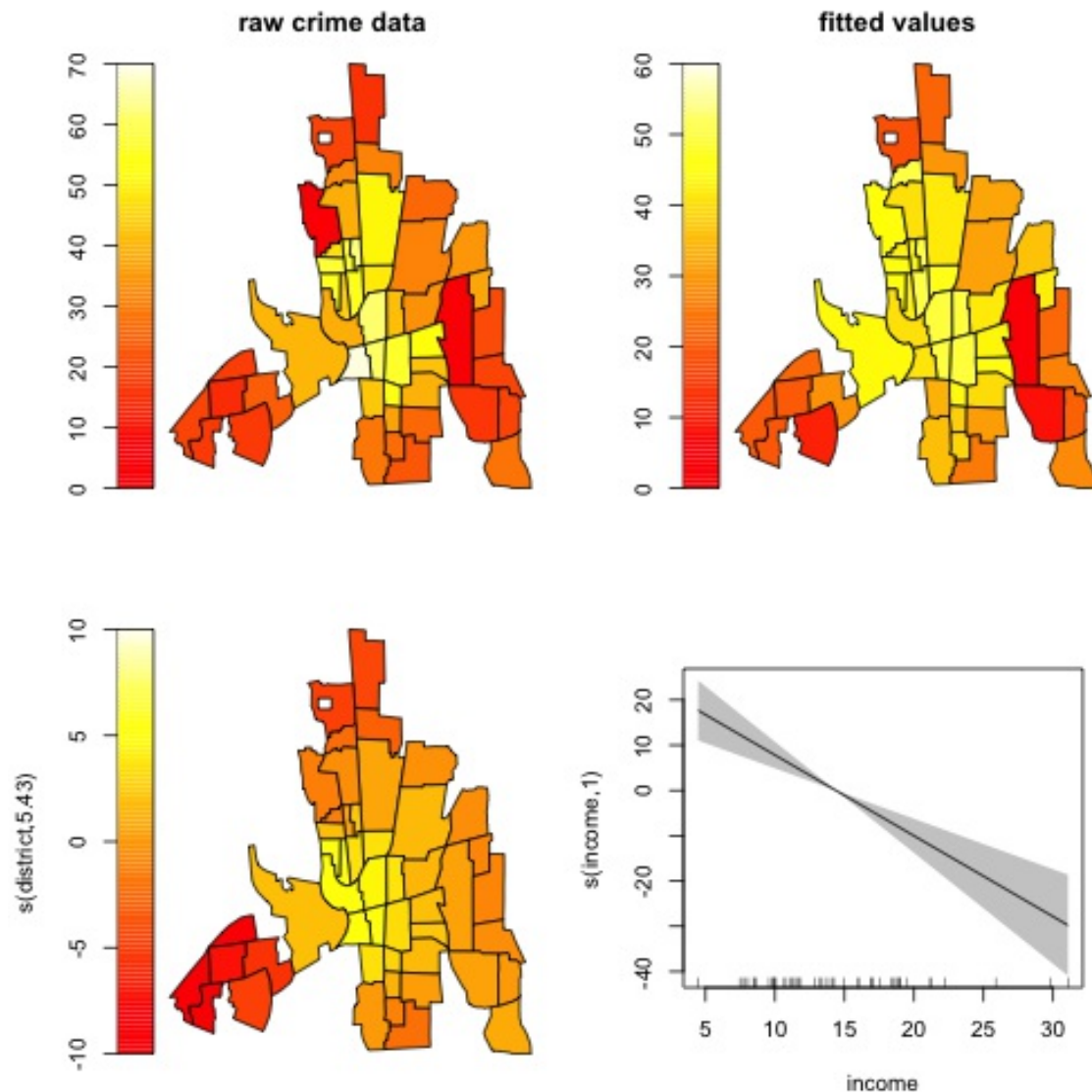
Fancy 2D smoothing

Funny-shaped regions



- Soap film smoother (bs="so")
- Model takes boundary into account by construction
- Need to specify a boundary and internal knots
- see ?soap

Spatial models using areas



- Markov random fields (bs="mrf")
- Need to specify polygons or adjacency matrix
- Not necessarily that useful for marine work?
- see ?mrf

Very general modelling

mgcv can fit *anything* you can write as (on the link scale):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \quad \text{s.t.} \quad \sum_j \boldsymbol{\beta} \mathbf{S}_j \boldsymbol{\beta}$$

if you can write your likelihood in a quadratic form, it can be part of a model in mgcv

?paraPen

Models for large datasets

- bam for big additive models
- can handle simple correlation structures
- parallel (block QR decompositions)
- fast! (still experimental)
- Wood, Goude, Shaw (2015)

Fancy summary

- You can do *a lot* of things in mgcv
- Start small, work up to complex models
- Sometimes convergence is against you
- There is *a lot* of information in the manual



Okay, that's enough

converged.yt/mgcv-workshop