Simon Says Stats
"People Are Strange"

Or how not everything is normal and homoscedastic

Simon White

Department of Psychiatry, University of Cambridge
MRC Biostatistics Unit, University of Cambridge

2020/Sep/15
Ed Group Meeting

Describe your outcome?

Percentage of correct answers scored on an assessment

- Is it continuous or discrete?
- Is the range open or closed?
- Is it a nominal, ordinal, interval, or ratio scale?

Describe your outcome?

Percentage of correct answers scored on an assessment

- Is it continuous or discrete?
- Is the range open or closed?
- Is it a nominal, ordinal, interval, or ratio scale?

- Discrete
  - There are a finite and countable set of percentages possible
  - If 8 questions, each worth 1 mark, there are nine values[1]
  - Ignoring issues around missing values — makes it hard
- Closed
  - Disregarding management-speak, cannot give 110% nor -10%
- Ordinal
  - Distance between values may be unequal

[1] 0.0%, 12.5%, 25.0%, 37.5%, 50.0%, 62.5%, 75.0%, 87.5%, 100.0%

Describe your outcome?

Percentage of correct answers scored on an assessment

- Is it continuous or discrete?
- Is the range open or closed?
- Is it a nominal, ordinal, interval, or ratio scale?

- Continuous
  - Conceptually we can get any percentage (and to as many decimal places)
  - We can imagine longer, more detailed, assessments with finer and finer detail
- Closed
  - Disregarding management-speak, cannot give 110% nor -10%
- Ratio (Interval)
  - Zero is well defined, ratios perhaps a little suspect?

Over use of the Central Limit Theorm

As $n \to \infty$, if x is well-behaved (*read ratio scale*), then mean of x will follow a normal distribution.

Over use of the Central Limit Theorm

As $n \to \infty$, if x is well-behaved (*read ratio scale*), then mean of x will follow a normal distribution.

If we assume an infinitly large population, we can readily apply the CLT.

Over use of the Central Limit Theorm

As $n \to \infty$, if x is well-behaved (*read ratio scale*), then mean of x will follow a normal distribution.

If we assume an infinitly large population, we can readily apply the CLT.

Not true if x is not "well-behaved".

The *z*-score

- *z*-scores have become common as a way to “standardise”
- They fundamentally assume
  - Truly continuous measures
  - open ranges
  - interval scales
- As with *p*-values (*which they are closely related too*), *z*-scores have become a single number summary of often complex data
- Also, in many cases they should really be called *t*-scores.

$$z = \frac{\bar{x}}{s} \qquad \text{and} \qquad t_{\mathrm{df}}^{-1}(z) = p \text{ or } 2p$$

What *s*?

GAMLSS

- Allow outcome to be non-normal (the "G" for generalised)
  - $X \sim \mathrm{Pois}()$: discrete, range is $[0, \infty)$
  - $X \sim \mathrm{Beta}()$: continuous, range is $(0, 1)$
  - $X \sim \mathrm{Binom}()$: discrete, range is $[0, n]$

GAMLSS

- Allow outcome to be non-normal (the "G" for generalised)
  - $X \sim \mathrm{Pois}()$: discrete, range is $[0, \infty)$
  - $X \sim \mathrm{Beta}()$: continuous, range is $(0, 1)$
  - $X \sim \mathrm{Binom}()$: discrete, range is $[0, n]$
- Enables – though not always sensible – to consider additive-smoothers (the "AM" for additve models)
  - Can incorporate data-driven estimation of non-straight-line smoothing (*read splines*, with penalty parameters)

GAMLSS

- Allow outcome to be non-normal (the "G" for generalised)
  - $X \sim \mathrm{Pois}()$: discrete, range is $[0, \infty)$
  - $X \sim \mathrm{Beta}()$: continuous, range is $(0, 1)$
  - $X \sim \mathrm{Binom}()$: discrete, range is $[0, n]$

fp

- Enables – though not always sensible – to consider additive-smoothers (the "AM" for additve models)
  - Can incorporate data-driven estimation of non-straight-line smoothing (*read splines*, with penalty parameters)
- Allows modelling multiple aspects of parametric distribution (the "LSS" for location, shape and scale)
  - $N(\mu, \sigma^2)$, traditional linear modelling defines $\boxed{\mu = f(\text{covariates})}$
  - GAMLSS allows functions for μ and σ

out ~ 1 + sex + age

Under GAMLSS *z*-score doesn't make sense

- GAMLSS defines a link function, relating the models for each paramter of the family (*read distribution*) choosen
- Link function relates the modelled parameters to a specific distribution
- All probability distributions can be mapped to $(0, 1)$ by definition (using the CDF (cumulative denisty function))
    - PMF (probability mass function) for discrete ←
    - PDF (probability denisty function) for continuous ←

Under GAMLSS *z*-score doesn't make sense

- GAMLSS defines a link function, relating the models for each paramter of the family (*read distribution*) choosen
- Link function relates the modelled parameters to a specific distribution
- All probability distributions can be mapped to $(0, 1)$ by definition (using the CDF (cumulative denisty function))
    - PMF (probability mass function) for discrete
    - PDF (probability denisty function) for continuous

Hence, using the inverse-link and link function we can define any observation (conditional on a model and covariates) as a given quantile between zero and one.

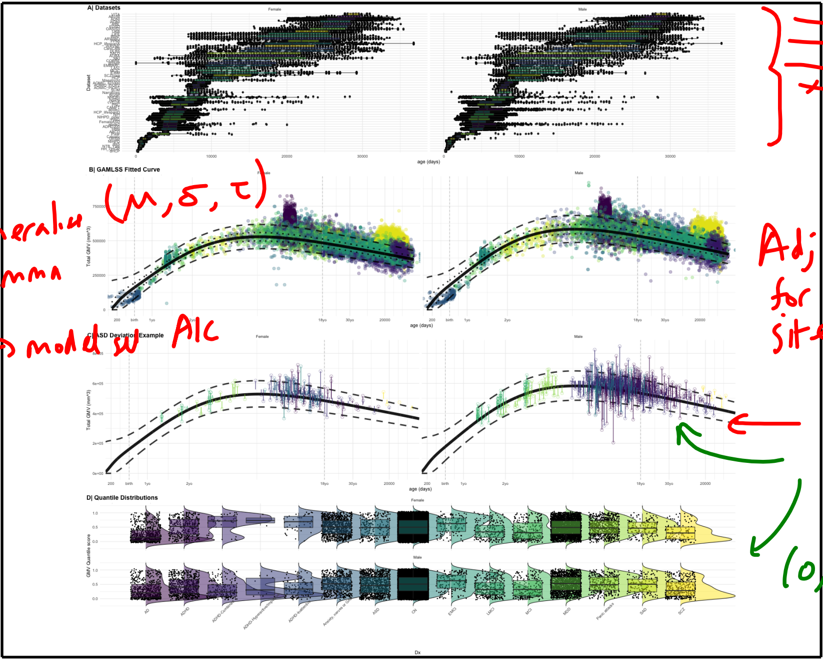
If the model is "good", then by definition the quantiles should be uniformly distributed. (*basically Q-Q plots from linear regression*)

Covariate-drive variance

$$\log \sigma = f(x_{\text{age}}, x_{\text{sex}}, \dots)$$

- GAMLSS models a change in **population** variance
- Without random-effects, this is not the same as increasing within-individual variance

People become more variable as they age is not the same as *as you age you become more variable*



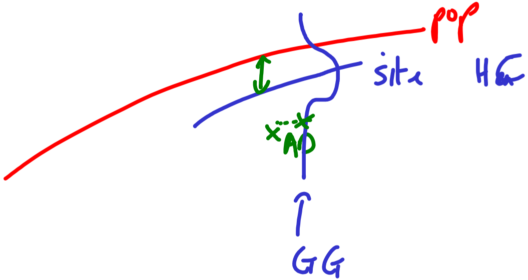
Generaliz Gamma

(μ, σ, τ)

→ model w AIC

Adj for site

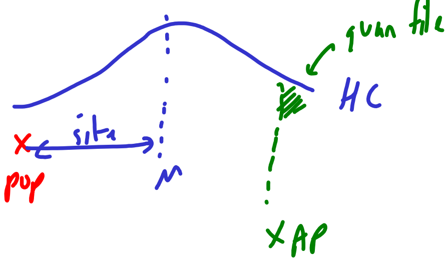
(0,1)



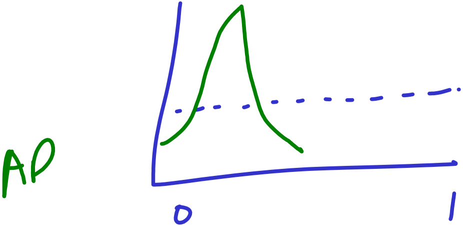
$\mu(age, re(site)$

$\sigma(age, re(site)$

$\tau(age, \ldots)$



HC



Summary

- Fundamental to analysis is understanding the characteristics of outcome
- Distinction between large-sample/population approximations and conceptual-truths
- GAMLSS more flexible. . . but requires bigger sample sizes!

Definitions are fun

Life-span or life-course

Life span refers to duration of life and characteristics that are closely related to age but that vary little across time and place. In contrast, the life course perspective elaborates the importance of time, context, process, and meaning on human development and family life (Bengtson and Allen 1993).