A Generalized Linear Mixed Model for Enumerated Sunspots

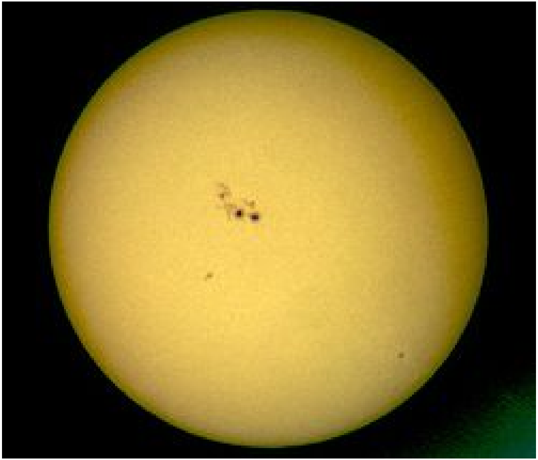
Jamie Riggs
Applied Statistics and Research Methods
Deep Space Exploration Society

Spring 2012 CO/WY Meeting

April 20, 2012

UNIVERSITY of NORTHERN COLORADO

Solar Beauty Spots



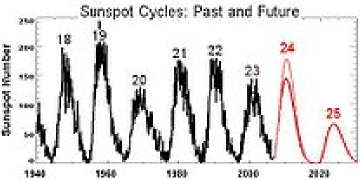
Presentation Outline

- Introduction
- Background
- Wald Approach
- General Linear Mixed Models
- Future Development
- Acknowledgments
  Rodney Howe, Solar Bulletin Editor, AAVSO
  Trent Lalonde, Applied Statistics, University of Northern Colorado

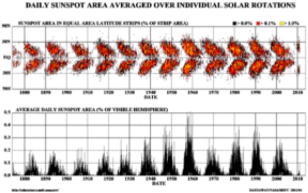
The Physics

- Sunspot generation a current research area
- Sunspots thought to be the visible counterparts of magnetic flux tubes in the Sun's convective zone
- Differential rotation (coriolis effect) stresses the tubes which puncture the Sun's surface
- Energy flux from the Sun's interior decreases and with it surface temperature
- Sunspot activity cycles about every eleven years
- Early in the cycle, sunspots appear in the higher latitudes and then move towards the equator as the cycle approaches maximum: this is called Spörer's law

Sunspot Cycle and Butterfly Plot



(a) Eleven-year cycle



(b) Spörer's Law

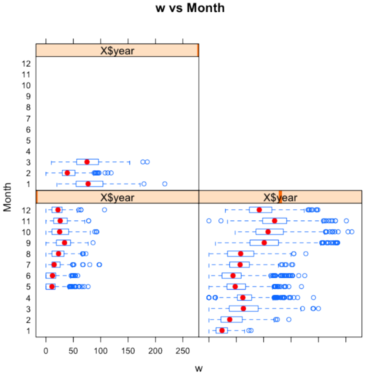
The Astronomy

- First noted sunspots in 364 BC by Chinese astronomer Gan De
- First telescopy in 1610 by English astronomer Thomas Harriot
- Rudolf Wolf established the Wolf Number in 1848
- AAVSO began the American Relative number in 1949
- Overall, weighted monthly count averages are assumed to be unbiased estimates of the true monthly sunspot numbers
- Sunspot cycle since 2010 is increasing from a minimum

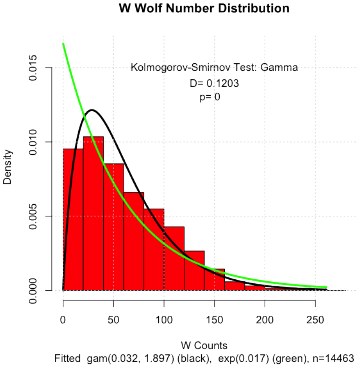
The Statistics

- Multiple observers (~ 60) worldwide
- Three random variables: sunspot counts, observers, and monthly sunspot numbers
- Sunspot numbers are known to follow an approximately 11-year sinusoidal cycle
- The statistical model needs to tie the average monthly sunspot numbers to the observer-reported counts
- The statistical model should tie historical numbers and predict future numbers

Monthly Submissions and Histogram



(c) Monthly counts



(d) Histogram with fitted pdfs

Wolf, Wald, and Shapley

The Framers

- Wolf, R, 1848.
  - Developed the Wolf number (an International sunspot number, relative sunspot number, or Zürich number)
  - A quantity measuring the number of sunspots and groups of sunspots on the Sun's surface
  - The relative sunspot number R is computed as

$$R = k(10g + s)$$

where

- s is the number of individual spots
- g is the number of sunspot groups
- k is a factor that varies with location and instrumentation

The Framers

- Wald, A., The Fitting of Straight Lines if Both Variables are Subject to Error, *Annals Mathematical Statistics*, 1940, Vol. 11, No. 3, pp. 284-300.
  - Response, Y, and predictor, X are random variables
  - Method of least squares (SLR) usually used
  - Fit parameters different for $Y \sim f(X)$ and $X \sim f(Y)$

The Framers

- Shapley, A.H., Reduction of Sunspot-Number Observations, *Publication of the Astronomical Society of the Pacific*, 1949, Vol. 61, No. 358, pp 13-21.
  - Adapted Wald's method to correct observations from many observers to the American Relative sunspot number
  - Correction factor accounts for variations in equipment and seeing conditions
  - A "statistical weight" per observer is also used

The American Relative Sunspot Number

Shapley via Wald

$$R_i = k_i(10g_i + s_i) \tag{1}$$

$$R_a = \frac{\sum_{i=1}^{N} w_i k_i R_i}{\sum_{i=1}^{N} w_i} \tag{2}$$

$$R_{sm} = \frac{1}{24}\left(R_{a,i-6} + R_{a,i+6} + 2\sum_{j=i-5}^{5} R_{a,j}\right) \tag{3}$$

Generalized Linear Mixed Models (GLMM)

GLM

The Poisson probability distribution function

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = e^{-\mu}\frac{1}{y!}e^{y\log(\mu)}, \quad y = 0, 1, 2, \ldots \qquad (4)$$

GLM

The Poisson probability distribution function

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = e^{-\mu}\frac{1}{y!}e^{y\log(\mu)}, \quad y = 0, 1, 2, \ldots \qquad (4)$$

Generalized Linear Models (GLM) use a 1-1 link to a monotone function of μ

$$\boldsymbol{\eta} = \mathbf{X}\beta = g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}) \qquad (5)$$

β is often estimated through iterative reweighted least squares

GLMM

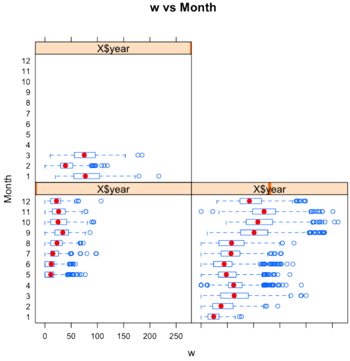
- In GLMM, $\boldsymbol{\eta}$ incorporates both fixed effects $\boldsymbol{\beta}$, and random effects $\mathbf{u}$ as

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{6}$$

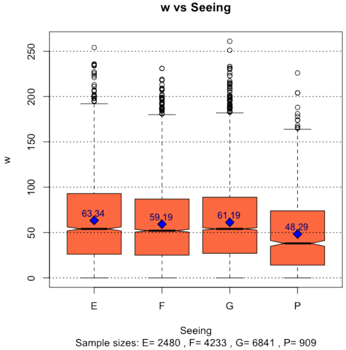
$$\begin{aligned} \boldsymbol{\mu} &= \text{ vector of mean sunspot numbers} \\ \mathbf{X} &= \text{ fixed effects matrix} \\ \boldsymbol{\beta} &= \text{ vector of fixed effects parameters} \\ \mathbf{Z} &= \text{ random effects matrix of observer identifiers} \\ \mathbf{u} &\sim \text{iid}\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),\text{ random effects parameter vector} \end{aligned} \tag{7}$$

Estimation of R_a

Estimation of R_a

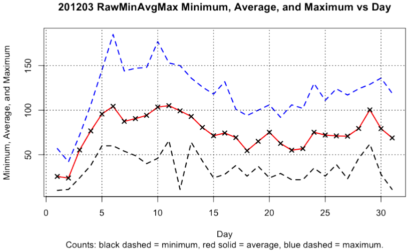


(e) Boxplots of Wolf numbers by Year and Month

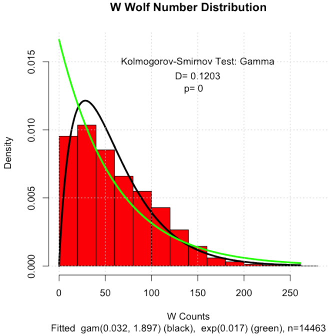


(f) Boxplots of Wolf numbers by seeing condition

Estimation of R_a



(g) Range of daily sunspot counts.

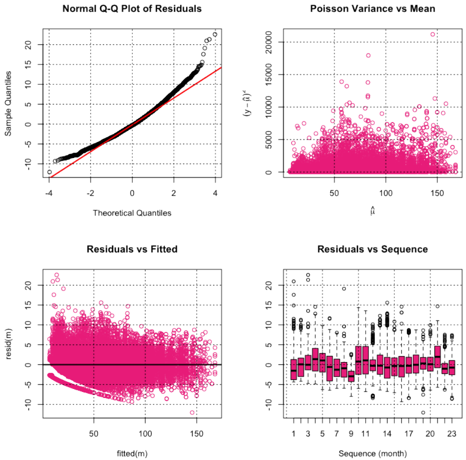


(h) Wolf number distribution.

Estimation of R_a

- Marginal likelihood estimation
  - Used on fixed effects model and Poisson/Normal model
  - Removes nuisance parameters by integrating them out
  - Time-consuming iterative integration
- Hierarchical likelihood estimation
  - Allow extra error components in the linear predictors of GLM
  - Distributions of these components not restricted to be normal
  - Uses Henderson's joint likelihood
  - Avoids integration as in marginal likelihood
  - Maximizing the h-likelihood gives
    - Fixed effect estimators asymptotically equivalent to marginal likelihood estimators
    - Obtain random effect estimates asymptotically BLUP

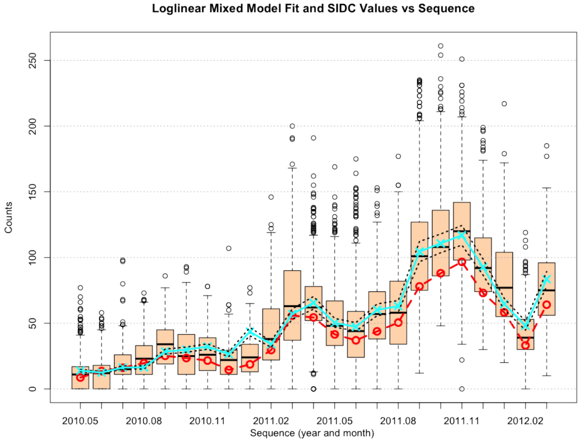
GLMM Diagnostics $\mathbf{y}|\mathbf{u} \sim Poi(\boldsymbol{\mu}),\ \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{u}}\mathbf{I})$



GLMM Diagnostics $\mathbf{y}|\mathbf{u} \sim Poi(\boldsymbol{\mu}),\ \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{u}}\mathbf{I})$

- $s^2/\bar{x} = 21.65875 >> 1$
- Concave up Normal Q-Q plot indicates right-skewed residuals
- Residuals vs. Fitted plot pattern indicates missing or misspecified predictors
- Preliminary use of Gamma error structure for observer random effect reduces the mean-variance ratio

GLMM Sunspot Number Estimates



Solid cyan curve connecting X'a is the loglinear (LL) model fit. Dashed red curve connecting O's SIDC values.
The dotted black curves are 99% lower and upper CIs for LL.

GLMM Overdispersion

Table: Improvements from Error Structure Changes

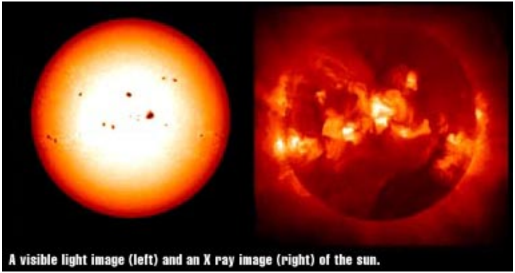
$\eta\|\mathbf{u}$ Dist	Link $g(\mu)$	$\mathbf{u}$ Dist	Link $v(\mathbf{u})$	Method	$s^2/\bar{x}$
Poisson	log	fixed	NA	GLS	22.87
Poisson	log	Normal	identity	log-likelihood	21.66
Poisson	log	Gamma	log	h-likelihood	18.49
Poisson	log	Poisson	identity	h-likelihood	?
Gamma	log	Gamma	identity	h-likelihood	?
Gamma	inverse	inverse Gamma	inversey	h-likelihood	?

Future Development

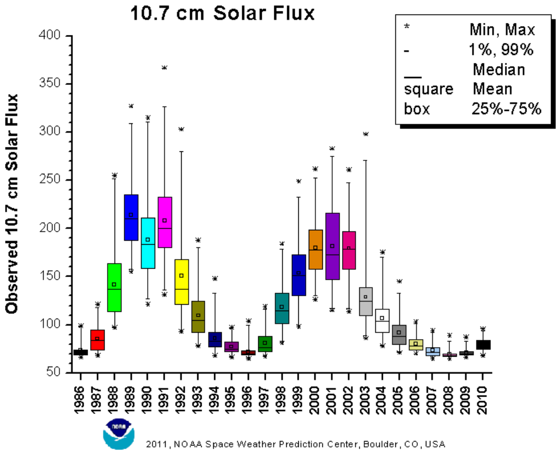
Future Development

- Introduce an observer's equipment factor (fixed)
- Test for the effect of the Solar hemisphere
- Test different error structures for counts and for observer random variables
- Braid in
  - Optical observations from Europe
  - X-ray
  - 10.7cm radio

Soft X-rays



10.7 cm (2800 MHz) Radio



A Generalized Linear Mixed Model for Enumerated Sunspots

Jamie Riggs
Applied Statistics and Research Methods
Deep Space Exploration Society

Spring 2012 CO/WY Meeting

April 20, 2012