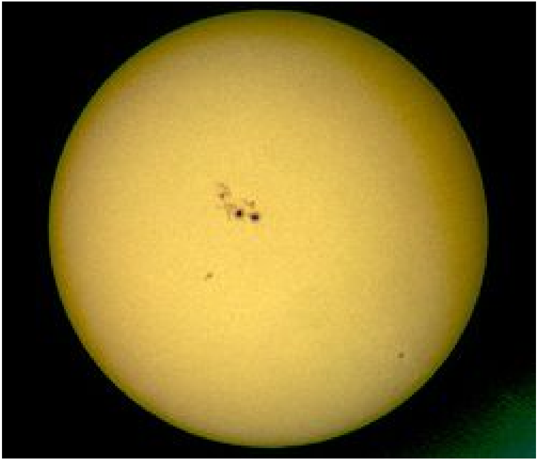
A Generalized Linear Mixed Model for Enumerated Sunspots

Jamie Riggs

Applied Statistics and Research Methods Colloquium
University of Northern Colorado

September 7, 2011

Solar Beauty Spots



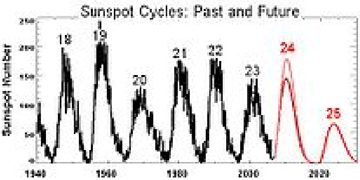
Presentation Outline

- Introduction
- Backgournd
- Wald Approach
- Statistical Models for Counts Data
- Future Development

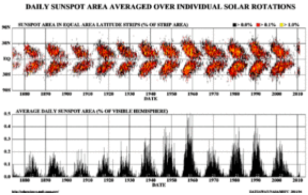
The Physics

- 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

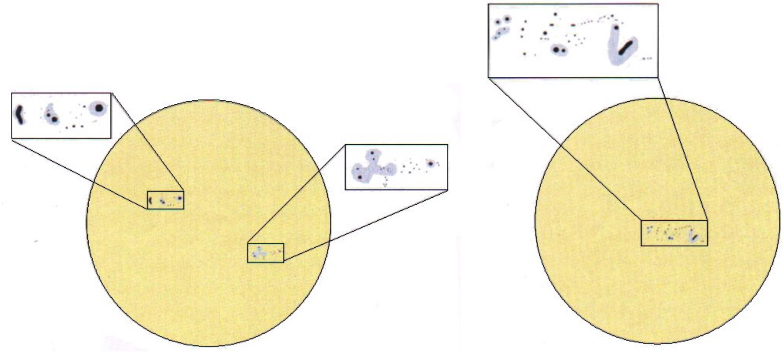


(b) Spörer's Law

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 recording the American Relative number in 1949
- AAVSO Solar Section experienced a serious loss of data
- 14 months of sunspot counts data collected since the loss
- Overall, weighted monthly count averages are assumed to be unbiased estimates of the true monthly sunspot numbers
- No sunspot number standard available so monthly counts are relative to the data provided
- As sunspot cycle in the last 3 months is increasing from a minimum, monthly corrections are anticipated

Sunspot Types



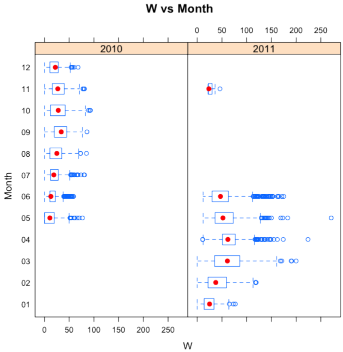
(c) Dual hemisphere

(d) Equatorial

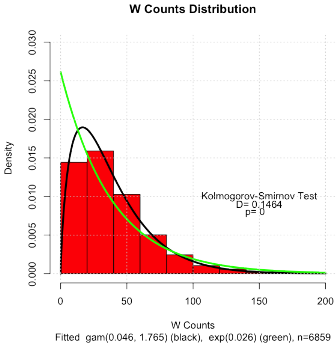
The Statistics

- Multiple observers ($\sim$ 80) worldwide
- Each may submit monthly (only $\sim$ 40 do so consistently)
- 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 predict sunspot numbers

Monthly Submissions and Histogram



(e) Monthly counts



(f) Histogram with fitted pdfs

Wolf, Wald, and Shapley

The Framers

- Wolf, R, 1848.
    - Developed the Wolf number (a 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 $X \sim f(Y)$
- Just for fun:
  - (1939). "A New Formula for the Index of Cost of Living". Econometrica
  - (1939). "Contributions to the Theory of Statistical Estimation and Testing Hypotheses". Annals of Mathematical Statistics
  - (June 1945). "Sequential Tests of Statistical Hypotheses". The Annals of Mathematical Statistics
  - (1947). Sequential Analysis. New York: John Wiley and Sons
  - (1950). Statistical Decision Functions. John Wiley and Sons, New York

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

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

Shapley via Wald

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

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

Shapley via Wald

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

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

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

The Derivation

$$Ey_{ij} = \beta_0 + \beta_1 x_{ij}$$

The Derivation

$$Ey_{ij} = \beta_0 + \beta_1 x_{ij}$$

$$\Rightarrow R_{sj} = k_i R_{ij}, \quad \text{per Shapley } \beta_0 = 0, \; \beta_1 = 1$$

The Derivation

$$Ey_{ij} = \beta_0 + \beta_1 x_{ij}$$

$$\Rightarrow R_{sj} = k_i R_{ij}, \quad \text{per Shapley } \beta_0 = 0, \ \beta_1 = 1$$

$$\frac{1}{N} \sum_{j=1}^{n_i} \log R_{sj} = \frac{1}{N} \sum_{j=1}^{n_i} \log k_i R_{ij}$$

The Derivation

$$Ey_{ij} = \beta_0 + \beta_1 x_{ij}$$

$$\Rightarrow R_{sj} = k_i R_{ij}, \quad \text{per Shapley } \beta_0 = 0, \; \beta_1 = 1$$

$$\frac{1}{N}\sum_{j=1}^{n_i} \log R_{sj} = \frac{1}{N}\sum_{j=1}^{n_i} \log k_i R_{ij}$$

$$\log k_i = \frac{1}{N}\left(\sum_{j=1}^{N} \log R_{sj} - \sum_{j=1}^{N} \log R_{ij}\right) \tag{4}$$

The Derivation

$$Ey_{ij} = \beta_0 + \beta_1 x_{ij}$$

$$\Rightarrow R_{sj} = k_i R_{ij}, \quad \text{per Shapley } \beta_0 = 0, \;\; \beta_1 = 1$$

$$\frac{1}{N}\sum_{j=1}^{n_i} \log R_{sj} = \frac{1}{N}\sum_{j=1}^{n_i} \log k_i R_{ij}$$

$$\log k_i = \frac{1}{N}\left(\sum_{j=1}^{N} \log R_{sj} - \sum_{j=1}^{N} \log R_{ij}\right) \tag{4}$$

$$w_i = \frac{N-1}{\sum_{j=1}^{N}\left(\log R_{sj} - \log R_{ij}\right)^2 - N \cdot a_i^2} \tag{5}$$

Poisson Models

Poisson Distribution

Poisson probability distribution function

$$f(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = e^{-\mu_i} \frac{1}{y_i!} e^{y_i \log(\mu_i)}, \quad i = 1, 2, \ldots, N \tag{6}$$

Poisson Distribution

Poisson probability distribution function

$$f(y_i;\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = e^{-\mu_i}\frac{1}{y_i!}e^{y_i \log(\mu_i)}, \quad i = 1, 2, \ldots, N \tag{6}$$

GLM canonical link to a monotone function of μ_i

$$\log(\mu_i) = \sum_{i,j} \beta_i x_{ij}, \quad i = 1, \ldots, N,\ j = 1, 2, \ldots, n_i \tag{7}$$

Poisson Distribution

Poisson probability distribution function

$$f(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = e^{-\mu_i} \frac{1}{y_i!} e^{y_i \log(\mu_i)}, \quad i = 1, 2, \ldots, N \tag{6}$$

GLM canonical link to a monotone function of μ_i

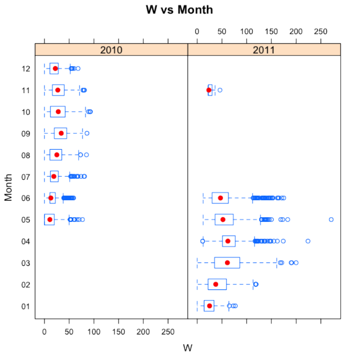
$$\log(\mu_i) = \sum_{i,j} \beta_i x_{ij}, \quad i = 1, \ldots, N, \ j = 1, 2, \ldots, n_i \tag{7}$$

The matrix form including observer, period, and seeing conditions

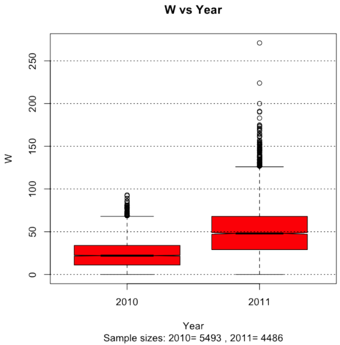
$$\log(\boldsymbol{\mu}_{\mathbf{f}}) = \mathbf{X}\boldsymbol{\beta}, \tag{8}$$

Determination of the k-Coefficients

Determination of the k-Coefficients

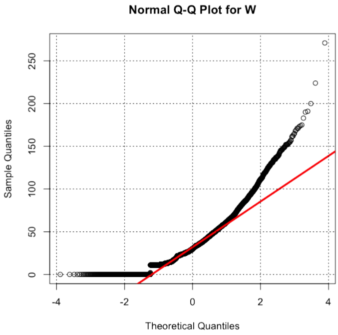


(g) Boxplots of Wolf numbers by Month

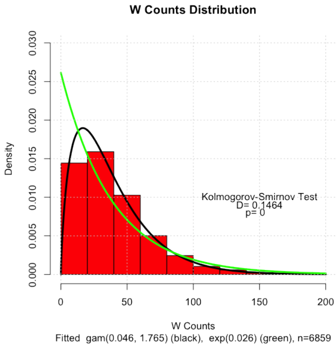


(h) Boxplots of Wolf numbers by Year

Determination of the k-Coefficients



(i) The normal Q-Q plot for the Wolf number.



(j) Wolf number distribution.

To Fix or Not to Fix...

- Should the factor "observer" be a fixed or a random effect?

To Fix or Not to Fix...

- Should the factor "observer" be a fixed or a random effect?
- Per Milliken, G.A. & Johnson, D.E., *Analysis of Messy Data, Volume 1: Designed Experiments*, 1998, Chapman & Hall, Boca Raton, Fl, p. 212.
  - A factor is *random* if its levels consist of a random sample of levels from a population of possible levels.

To Fix or Not to Fix...

- Should the factor "observer" be a fixed or a random effect?
- Per Milliken, G.A. & Johnson, D.E., *Analysis of Messy Data, Volume 1: Designed Experiments*, 1998, Chapman & Hall, Boca Raton, Fl, p. 212.
  - A factor is *random* if its levels consist of a random sample of levels from a population of possible levels.
  - A factor is *fixed* if its levels are selected by a nonrandom process or if its levels consist of the entire population of possible levels.

GLM with Poisson Error Structure

- Several models were fitted using the independent variables observer, seeing conditions, and time sequence
- Two error structures were used: Poisson and negative binomial
- The primary criterion for model selection is the ratio of the variance of the model-fitted values to the mean of the model-fitted values. Why?

GLM with Poisson Error Structure

- Several models were fitted using the independent variables observer, seeing conditions, and time sequence
- Two error structures were used: Poisson and negative binomial
- The primary criterion for model selection is the ratio of the variance of the model-fitted values to the mean of the model-fitted values. Why?
- The final model is

$$\log(y_{ij} = \beta_0 + \beta_1 x_{1ij} + beta_2 x_{1ij} + \eta_{ij},$$
$$x_{1ij} = j^{th} \text{ appearance of the } i^{th} \text{ observer}$$
$$x_{2ij} = j^{th} \text{ occurrence of the } i^{th}\text{observer's seeing condition}$$

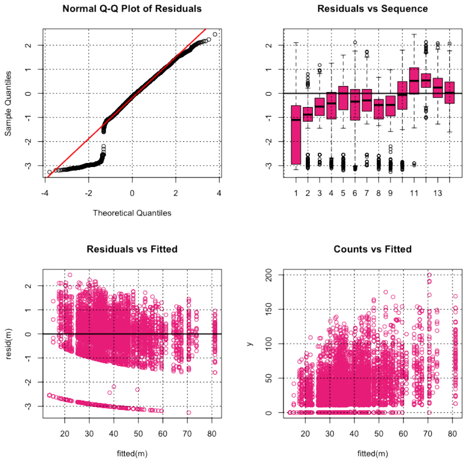
GLM with Poisson Error Structure

- The fitted values variance by mean ratio is 3.5173. Prefer < 2
- Residual deviance: 8439.5 on 6815 degrees of freedom (try for equality)

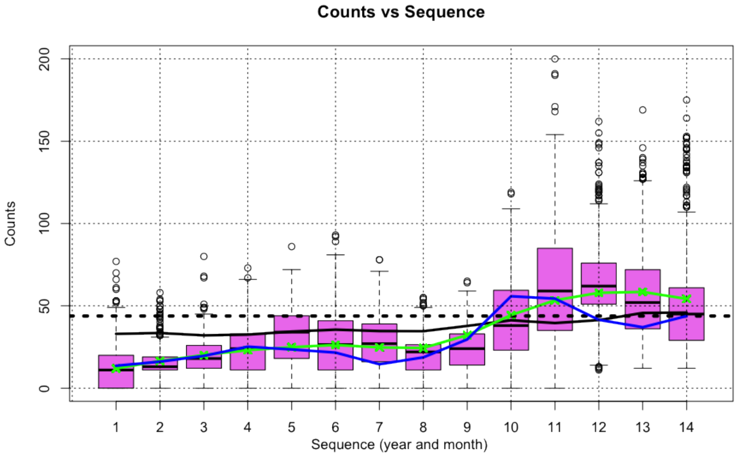
Table: ANOVA

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			6858	9200.01	
x1	40	665.97	6818	8534.04	0.0000
x4	3	94.55	6815	8439.48	0.0000

GLM with Poisson Error Structure Diagnostics



GLM with Poisson Error Structure Diagnostics



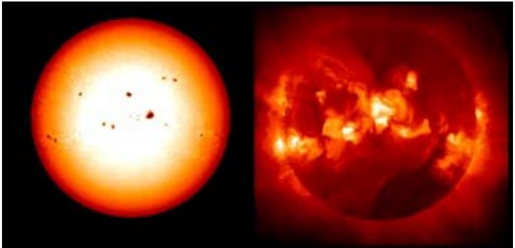
Green X's connected by the curve are a loess fit. Dotted blue curve is NOAA values.
Black dashed line is loglinear fit.

Future Development

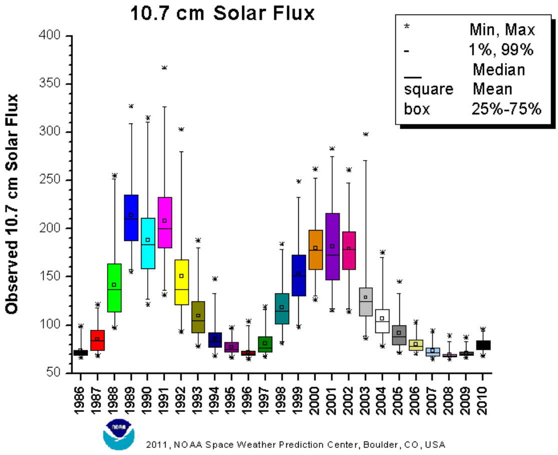
Future Development

- Introduce an observer's equipment factor (fixed)
- Test for the effect of the Solar hemisphere
- Braid in
    - Optical observations from Europe
    - X-ray observations from GOES-15 satellite
    - 10.7cm radio from Deep Space Exploration Society, Canada, and Australia

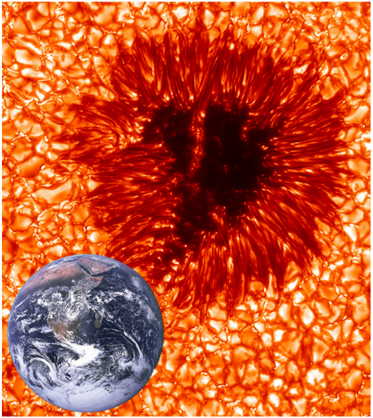
Soft X-rays


A visible light image (left) and an X ray image (right) of the sun.

10.7 cm (2800 MHz) Radio



It's a Matter of Scale



A Generalized Linear Mixed Model for Enumerated Sunspots

Jamie Riggs

Applied Statistics and Research Methods Colloquium
University of Northern Colorado

September 7, 2011