International Astrostatistics Association

Bayesian Models for Astronomy

ADA8 Summer School

Rafael S. de Souza

rafael@caesar.elte.hu

May 24, 2016

Chapter 1

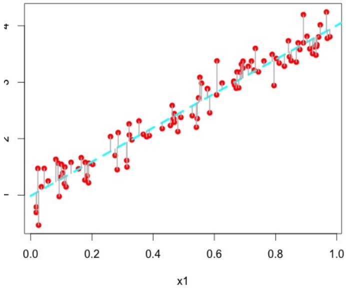
Gaussian Models

Linear Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim Norm(0, \sigma^2)$$

R script

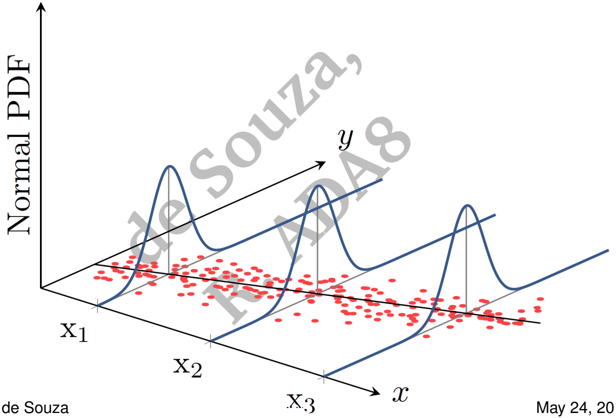
```
N = 100
sd=0.25
x1<-runif(nobs)
mu <- 1 + 3*x1
y<-rnorm(N,mu,sd)
# Fit model
lm(y ~ x1)
```



Linear Model

$Y_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$ Stochastic part

$\mu_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ Deterministic part



Normal Linear Models

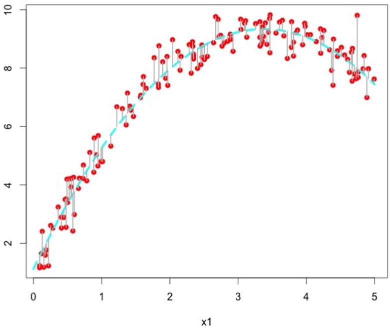
Gaussian

Linear Model

$$Y_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

R script

```
x1<-runif(nobs)
mu<-1+5*x1-0.75*x1^2
y<-rnorm(N,mu,0.5)
# Fit model
lm(y~x1+I(x^2))
```

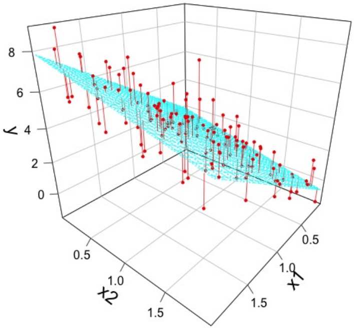


Linear Model

$$Y_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

R script

```
mu<-2+3*x1-1.5*x2
y<-rnorm(N,mu,sd)
# Fit model
lm(y~x1+x2)
```

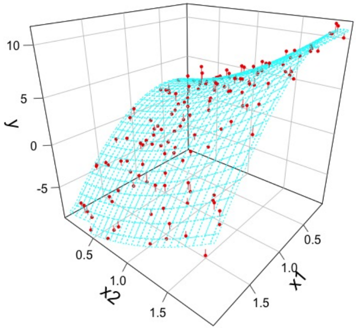


Linear Model

$Y_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$
$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +$
$\beta_3 x_1^2 + \beta_4 x_2^2$

R script

```
mu<-2+3*x1-4.5*x1^2
+ 1.5*x2+1.5*x2^2
y<-rnorm(N,mu,sd)
# Fit model
lm(y~x1+x2+
I(x1^2)+I(x2^2))
```



The Bayesian way

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

JAGS: Just Another Gibbs Sampler

http://mcmc-jags.sourceforge.net

A program for analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulation written with three aims in mind:

- To have a cross-platform engine for the BUGS (Bayesian inference Using Gibbs Sampling) language
- To be extensible, allowing users to write their own functions, distributions and samplers.
- To be a platform for experimentation with ideas in Bayesian modelling

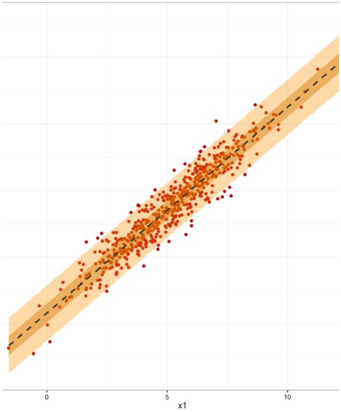
COIN
Cosmostatistics Initiative

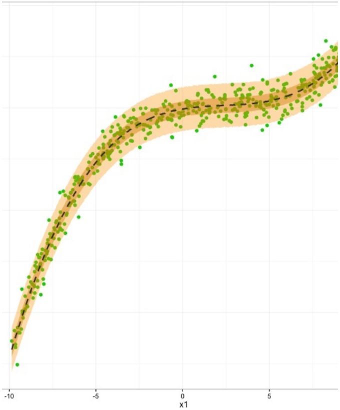
JAGS: Just Another Gibbs Sampler

An intuitive program for analysis of Bayesian hierarchical models using MCMC.

Normal Model	JAGS syntax
Likelihood	
$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$	$Y[i] \sim dnorm(mu[i], pow(sigma, -2))$
$\mu_i = \beta_1 + \beta_2 \times X_i$	$mu[i] < -inprod(beta[], X[i,])$
Priors	
$\beta_i \sim \mathcal{N}(0, 10^4)$	$beta[i] \sim dnorm(0, 0.001)$
$\sigma \sim \mathcal{U}(0, 100)$	$sigma \sim dunif(0, 100)$

Synthetic Normal model with JAGS





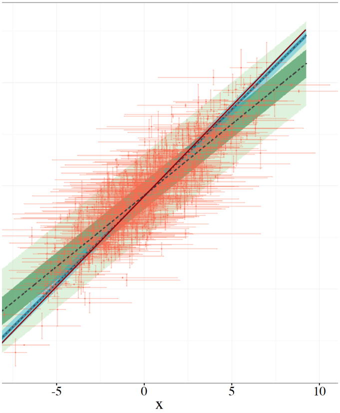
Normal Model

Likelihood

$$Y_{\text{obs;i}} \sim \mathcal{N}(Y_{\text{true;i}}, \epsilon_Y^2)$$
$$X_{\text{obs;i}} \sim \mathcal{N}(X_{\text{true;i}}, \epsilon_X^2)$$
$$Y_{\text{true;i}} \sim \mathcal{N}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_1 + \beta_2 \times X_{true;i}$$

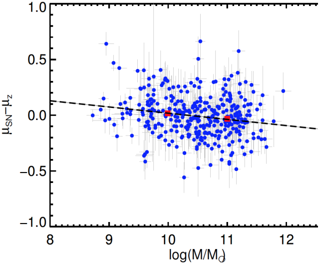
Priors

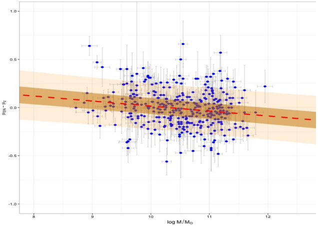
$$X_{true;i} \sim \mathcal{N}(0, 10^2)$$
$$\beta_1 \sim \mathcal{N}(0, 10^4)$$
$$\beta_2 \sim \mathcal{N}(0, 10^4)$$
$$\sigma \sim \mathcal{U}(0, 100)$$



Astronomical application of Normal model

Hubble residuals-data from Wolf, R. et al. arXiv:1602.02674



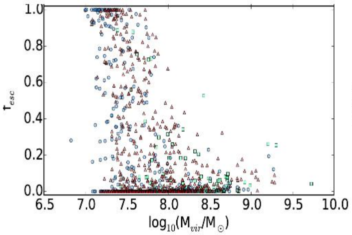


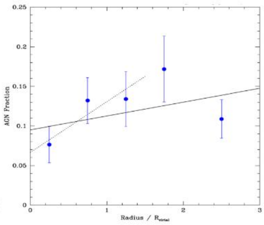
JAGS

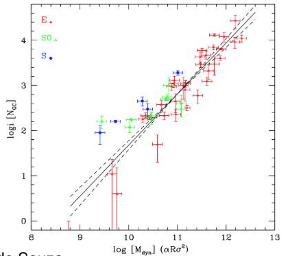
LINMIX

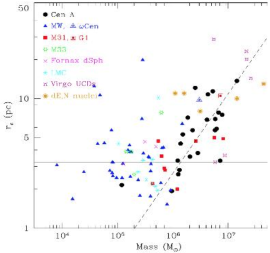
Astronomical examples of non-Gaussian data

Gaussian models are powerful, but fall short in many simple applications.









Beyond Gaussian

Generalized Linear Models

Beyond Normal Linear Regression

The tip of the iceberg

Non-exhaustive list of Regression Models:

- Linear/Gaussian models: $x \in \mathbb{R}$
- Generalized linear models
  - Gamma: $x \in \mathbb{R}_{>0}$
  - Log-normal: $x \in \mathbb{R}_{>0}$
  - Poisson: $x \in \mathbb{Z}_{\geq 0}$
  - Negative-binomial: $x \in \mathbb{Z}_{\geq 0}$
  - Beta: $\{0 < x < 1\}$
  - Beta-binomial: $\{0 < x < 1\}$
  - Bernoulli: $x = 0$ or $x = 1$

$\vdots$

Generalized Linear Models
Beyond Gaussian

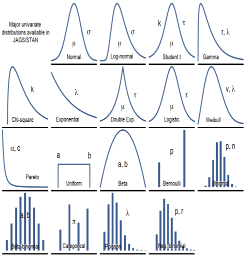
Generalized linear model is a natural extension of the Gaussian linear regression that accounts for a more general family of statistical distributions.

Linear Model

$$Y_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Generalized Linear Models

$$Y \sim f(\mu_i, a(\phi)V(\mu))$$
$$g(\mu) = \eta$$
$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$



Chapter II

Bernoulli Models

Bernoulli Model

Describes a random variable which takes the value 1 with success probability of p and 0 with failure probability of $1 - p$.

Logistic Model

$$Y_i \sim \mathrm{Bernoulli}(p) \equiv p^y(1-p)^{1-y}$$
$$\log \frac{p}{1-p} = \eta$$
$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Figure: Coin toss

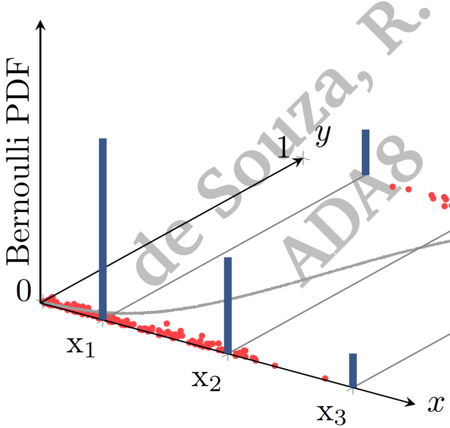


Figure: 3D visualization of a Bernoulli distributed data. Red points are the actual data and grey curve the fit.

Normal vs Bernoulli model

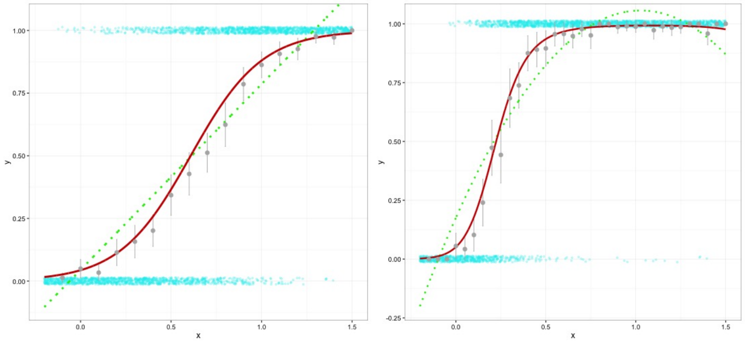
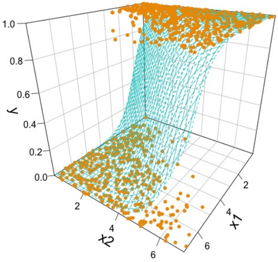
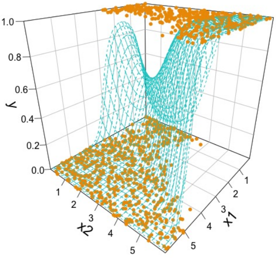


Figure: Linear regression predicts fractional (or probability) values outside the [0, 1] range.


1.0
0.8
0.6
0.4
0.2
0.0
y
2
4
6
x2
2
4
6
x1


1.0
0.8
0.6
0.4
0.2
0.0
y
1
2
3
4
5
x2
1
2
3
4
5
x1

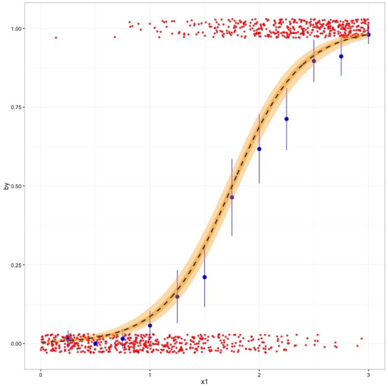
Bernoulli model in JAGS

Bernoulli Model

$$Y_i \sim \mathrm{Bern}(p_i)$$
$$log(\frac{p_i}{1-p_i}) = \eta_i$$
$$\eta_i = \beta_0 + \beta_1 X_i$$

JAGS

$$Y[i] \sim \mathrm{Bern}(p[i])$$
$$logit(p[i]) = \eta[i]$$
$$\eta[i] = \beta[1] + \beta[2] * X[i]$$



Astronomical application of Bernoulli model

Star formation activity in early Universe, data from Biffi and Maio, arXiv:1309.2283

Bernoulli Model

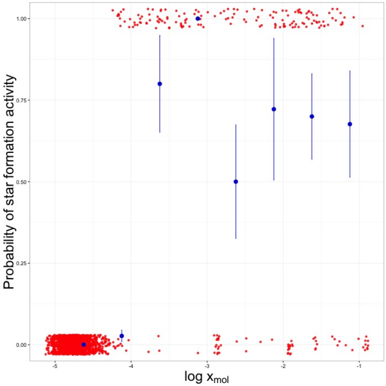
$SFR \sim \mathrm{Bern}(p_i)$

$log(\frac{p_i}{1-p_i}) = \eta_i$

$\eta_i = \beta_0 + \beta_1 \log x_{\mathrm{mol}} +$

$\beta_2(\log x_{\mathrm{mol}})^2 +$

$\beta_3(\log x_{\mathrm{mol}})^3$



Astronomical application of Bernoulli model

Star formation activity in early Universe, data from Biffi and Maio, arXiv:1309.2283, see also de Souza et al. arxiv.org/abs/1409.7696

Bernoulli Model

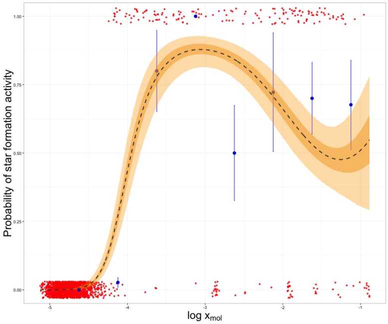
$$SFR \sim \mathrm{Bern}(p_i)$$

$$log(\frac{p_i}{1-p_i}) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 \log x_{\mathrm{mol}} +$$

$$\beta_2 (\log x_{\mathrm{mol}})^2 +$$

$$\beta_3 (\log x_{\mathrm{mol}})^3$$



Chapter III

Count Models

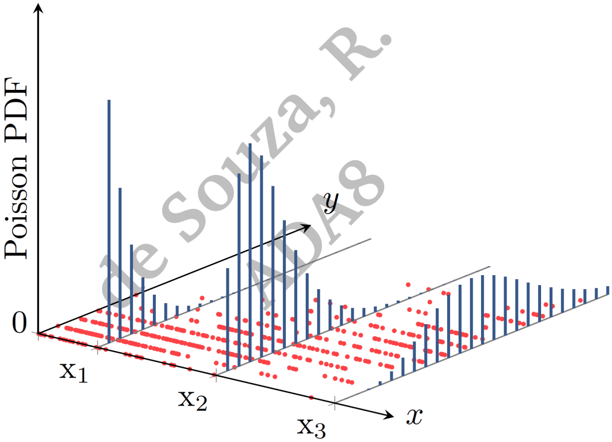


Figure: 3D visualization of a Poisson distributed data.

Poisson Models

Number of events

The Poisson distribution model the number of times an event occurs in an interval of time or space.

Linear Models

$Y \sim \mathrm{Normal}(\mu, \sigma^2)$

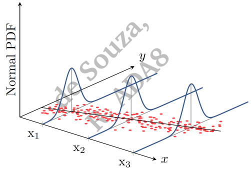
$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

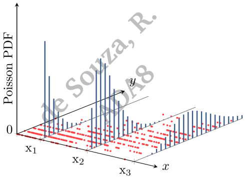
Poisson Model

$Y \sim \mathrm{Poisson}(\mu) \equiv \frac{\mu^{\mathrm{y}} \mathrm{e}^{-\mu}}{\mathrm{y}!}$

$\log(\mu) = \eta$

$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$





Poisson model
Mean equals variance

Although traditional, it is rarely useful in real situations.

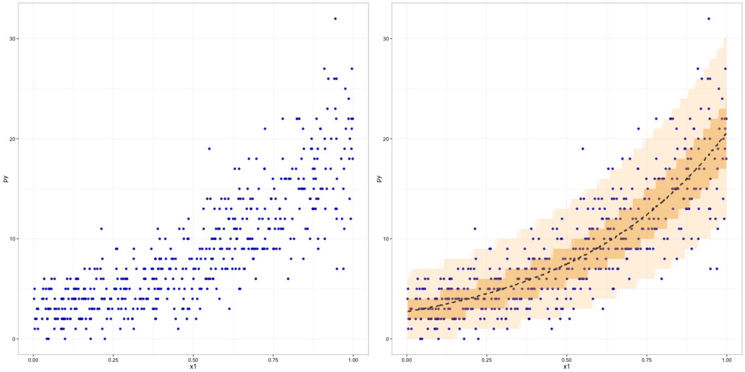


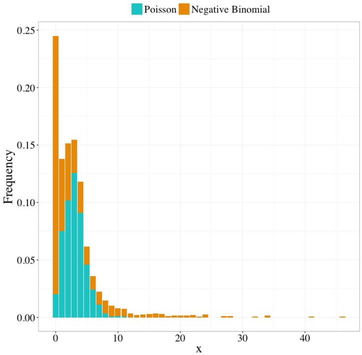
Figure: Poisson distributed data

The NB distribution can be seen as a Poisson distribution, where the mean is itself a random variable, distributed as a gamma distribution.

Negative Binomial distribution

$$\frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(1-\frac{\theta}{\mu+\theta}\right)^{y}$$

$$\text{Mean} = \mu \qquad \text{Var} = \mu + \frac{\mu^2}{\theta}$$



Accounts for Poisson over-dispersion

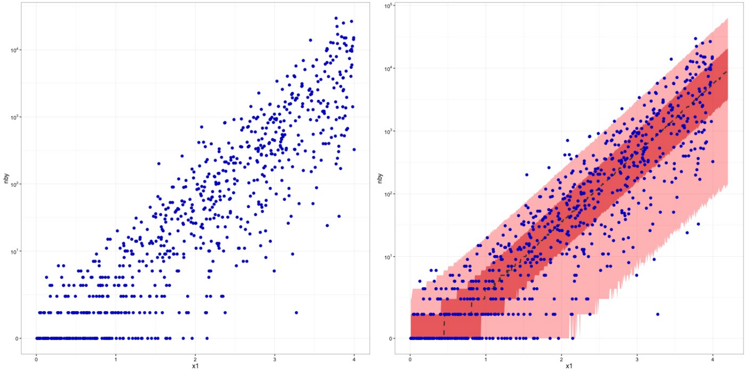


Figure: Negative binomial distributed data

Astronomical application of COUNT models

Globular Cluster Population, see R. S. de Souza, et al. MNRAS, 453, 2, 1928, 2015

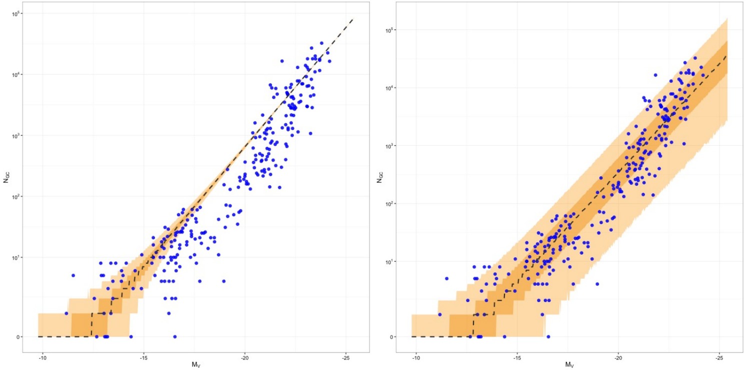


Figure: Globular Cluster population as a function of the galaxy brightness

Final Remarks

- Generalized linear models are a cornerstone of modern statistics, but nearly *Terra incognita* in astronomical investigations.
- Recent examples of GLM applications in astronomy are photo-z (Gamma-distributed; Elliott et al., 2015), GCs population (NB-distributed; de Souza at al. 2015), AGN activity (Bernoulli-distributed; Souza et al. 2016).

COIN
Cosmostatistics Initiative

References

- **R. S. de Souza**, M. L. Dantas, et al. arXiv:1603.06256
- **R. S. de Souza**, E. Cameron, et al. A&C, 12, 21, 2015
- J. Elliott, **R. S. de Souza**, et al. A&C, 10, 6, 2015
- **R. S. de Souza**, J. M. Hilbe, et al. MNRAS, 453, 2, 1928, 2015
- J. Hilbe, **R. S. de Souza** and E. E. O. Ishida, Bayesian Models for Astrophysical data, Cambridge University Press, in prep
- J. Hilbe, Practical Guide to Logistic Regression, Chapman and Hall/CRC, 2015
- S. Andreon & B. Weaver, Bayesian Methods for the Physical Sciences, Springer, 2015