Intro to Generalized Linear Models (GLMs)

May 7, 2025

Prof. Gwendolyn Eadie

Some motivation

- Up to now we have been looking at the case where $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, or equivalently:

$$Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right)$$

or

$$E[Y|X] = \beta_0 + \beta_1 X$$

This assumes a continuous response variable. What if the response is binary? Integer?

- Generalized Linear Models (GLM) are an extension of classical linear model

Generalized Linear Models

- Another way to write

$$E[Y|X] = \beta_0 + \beta_1 X$$

is

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

- i.e., $\mathbf{E(Y)} = \boldsymbol{\mu}$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

- To *generalize* this, we rearrange this into three parts

1. *Random* component: $\mathbf{Y}$ have independent normal distr. with $\mathbf{E(Y)} = \boldsymbol{\mu}$ and constant variance
2. *Systematic* component: $\boldsymbol{\eta} = \Sigma_1^p \boldsymbol{x}_j \beta_j$
3. *Link* between the two above: $\boldsymbol{\mu} = \boldsymbol{\eta}$

McCullagh & Nelder, *Generalized Linear Models* 2nd ed., Monographs on Statistics and Applied Probability 37, Chapman & Hall/CRC, 1989.

Generalized Linear Models

1. *Random* component: $\mathbf{Y}$ have independent normal distr. with $\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and constant variance
2. *Systematic* component: $\boldsymbol{\eta} = \Sigma_1^p \boldsymbol{x}_j \beta_j$
3. *Link* between the two above: $\boldsymbol{\mu} = \boldsymbol{\eta}$

- Then we *generalize* the link:

$$\eta_i = g(\mu_i)$$

- The *link function* relates the linear predictor η to the expected value μ of a datum y

McCullagh & Nelder, *Generalized Linear Models* 2nd ed., Monographs on Statistics and Applied Probability 37, Chapman & Hall/CRC, 1989.

Common Link Functions

- The link function $g(\cdot)$:

$$\eta_i = g(\mu_i)$$

- The *link function* relates the linear predictor η to the expected value μ of a datum y
- So, in general, exponential families of distributions can have a GLM of the form

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

- Normal: $\eta = \mu$
- Poisson: $\eta = \ln \mu$ ← counts
- Binomial: $\eta = \ln\left(\frac{p}{1-p}\right)$ ← logistic regression
- Gamma: $\eta = 1/\mu$
- Inverse Gaussian: $\eta = 1/\mu^2$

McCullagh & Nelder, *Generalized Linear Models* 2nd ed., Monographs on Statistics and Applied Probability 37, Chapman & Hall/CRC, 1989.

Binomial Regression with the logit link

- Also known as *logistic regression*
- The log-likelihood using the logit link is

$$l(\beta; y) = \sum_{i=1}^{n} y_i \eta_i - n_i \log\left(1 + e_i^{\eta}\right) + \log \binom{n_i}{y_i}$$

- Do maximum likelihood estimation to get estimates of βs

- `glm` or `brms` packages in R

Example of a GLM: Logistic Regression

- Logistic regression can be useful when you have a *binary response* Y given *continuous covariates* $\mathbf{X}$
  - Toy astronomy example: we want to predict the probability that a star has a planet, given the star's mass
    - Covariate: stellar mass M_* of a bunch of stars (continuous)
    - Response: whether or not a the star has a planet (binary)

Logistic regression

- You have data with a binary response, so the model is

$$Y_i \sim \mathrm{Bin}(n_i, p_i)$$

- The link is the logit: $g(\mu) = \log(\frac{\mu}{1-\mu})$

$\eta = \log\left(\frac{\mu}{1-\mu}\right)$

$e^{\eta} = \mu/(1-\mu)$

$e^{\eta}(1-\mu) = \mu$

$e^{\eta} - e^{\eta}\mu = \mu$

$\mu = \frac{e^{\eta}}{1+e^{\eta}}$

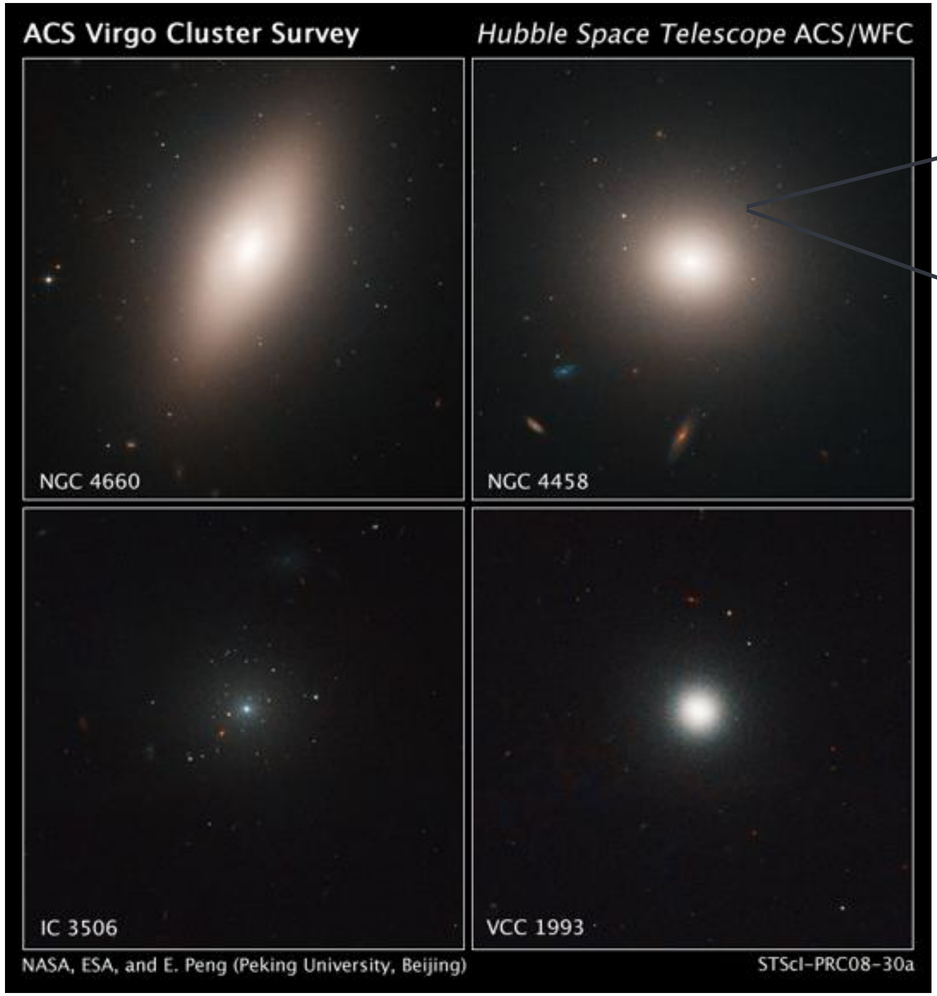
$\eta = X\beta$

- The probability p of "success" given some continuous covariate X is

$$P = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

- Can estimate βs via maximum likelihood

Example from astronomy



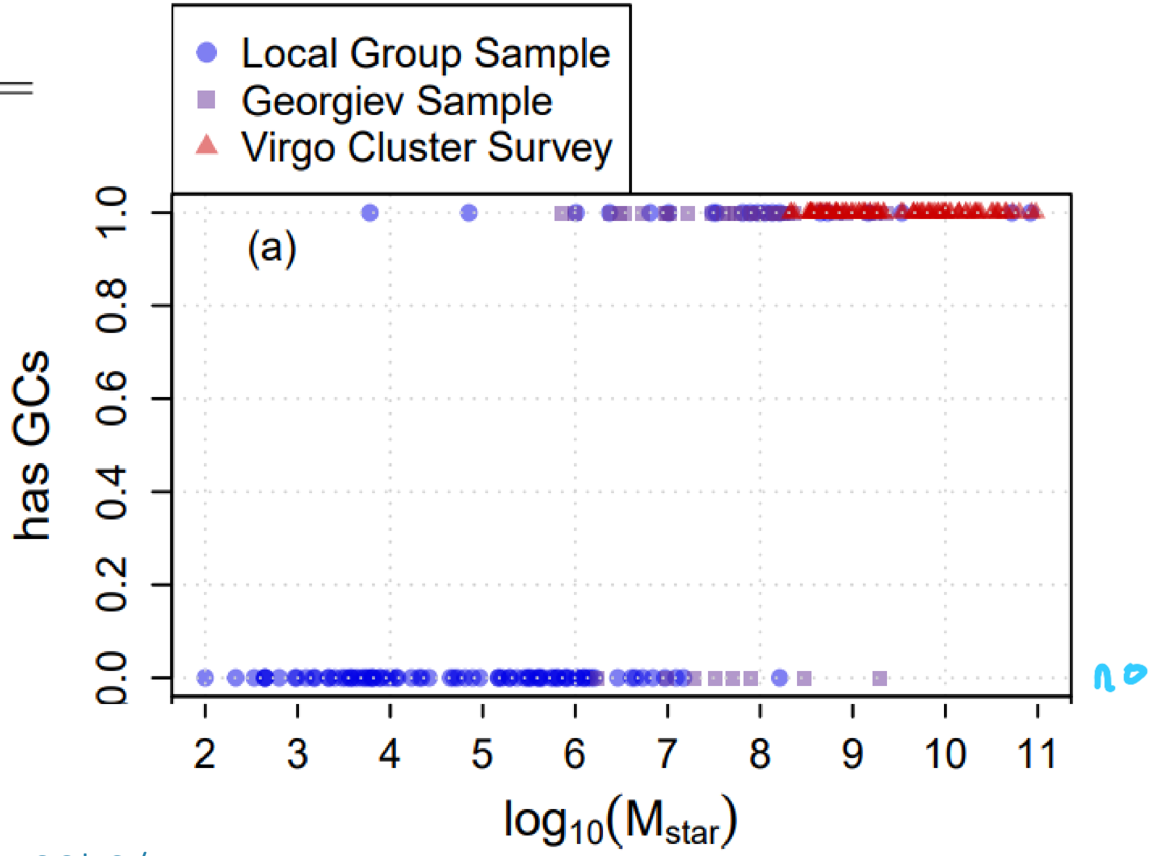
Almost every galaxy in the universe has a population of *Globular Clusters* (GCs, old star clusters), but lower mass galaxies do not always have a population of GCs

https://www.nasa.gov/mission_pages/hubble/science/virgo_cluster.html

Data: number of galaxies without or with GCs

	Has GCs? No (0)	Has GCs? Yes (1)	Total Galaxies
Local Group Sample	80	20	100
Georgiev Sample	8	31	39
Virgo Cluster Survey	0	93	93

Georgiev sample: Georgiev et al (2010), MNRAS, 406, 1967
Virgo sample: Cote et al 2004, ApJS,153, 223



Eadie, Harris, & Springford (2022), ApJ, 926:162
https://iopscience.iop.org/article/10.3847/1538-4357/ac33b0/meta

Example of Logistic Regression in Astronomy

- Covariate: $\log_{10}\frac{M_*}{M_\odot}$
- Binary Response: has GC populations (1) or does not (0)

$$p = (1 + e^{-(\beta_0 + \beta_1 \log M_\star)})^{-1}$$

Table 2
Logistic Regression Coefficients

	Dependent variable: has GCs	
	Local Group (1)	Entire Sample (2)
β_0	−10.31 (−14.59, −6.03)	−10.50 (−13.33, −7.67)
β_1	1.43 (0.79,2.07)	1.55 (1.15,1.94)
Observations	100	232

Note. Values in brackets are 95% confidence intervals.

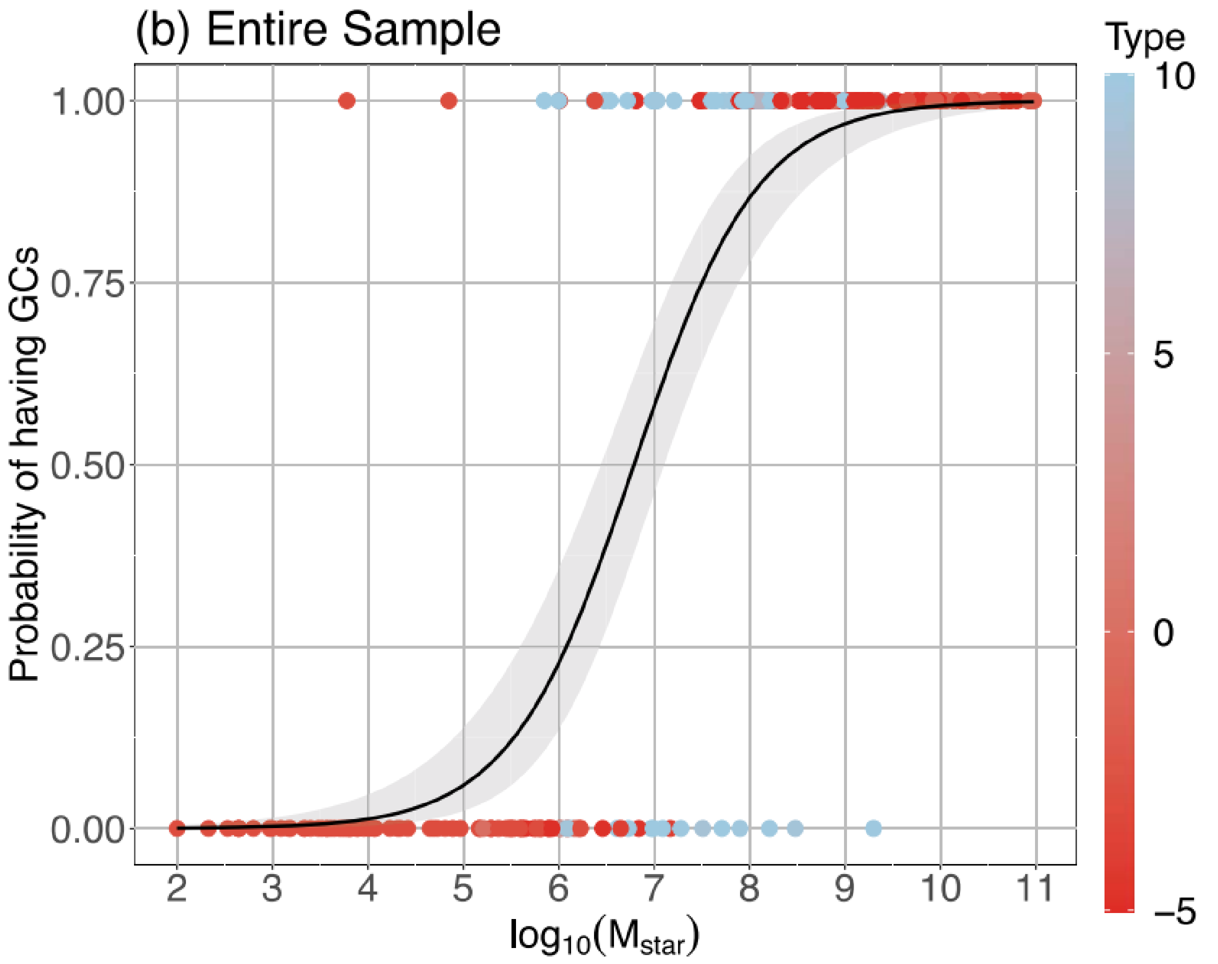


Figure 2. Logistic regression model assuming a single predictor (stellar mass) for (a) the LG, and (b) the entire sample. Galaxy morphological type from −5 (elliptical) to +10 (irregular) is color coded as shown by the color bar, but is not included in the analysis (see text). The logit regression curve obtained through maximum likelihood is shown as the solid black line, and the gray regions show the estimated 95% confidence intervals in probability for a given stellar mass.

Eadie *et al* 2022 *ApJ* **926** 162

Example of a GLM: Poisson Regression

- Poisson regression can be is useful when you have a *counts* as the response Y given *continuous covariates* $\mathbf{X}$
  - Toy astronomy example: we want to predict the number of GCs around a galaxy, given the galaxy's mass
    - Covariate: stellar mass M_* of a bunch of stars (continuous)
    - Response:
- The *link function* is the natural log:

$$\ln(\lambda) = X\beta$$

- Note: $\boldsymbol{\lambda}$ is a ***vector***, i.e.

$$E[Y_i] = \lambda_i$$

Example of Poisson Regression in Astronomy

- Covariate: $\log_{10}\frac{M_*}{M_\odot}$
- Count Response: Number GCs around galaxy

Likelihood:

$$L(\boldsymbol{\beta}; y) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

where

$$\ln(\lambda_i) = \beta_0 + \beta_1 X_i$$

Table 3
Poisson Regression Coefficients

	Dependent variable: N_{GC} Entire Sample
β_0	−7.47
	(−7.72, −7.21)
β_1	1.21
	(1.18, 1.24)
Observations	232

Note. Values in brackets are 95% confidence intervals.

Eadie *et al* 2022 *ApJ* **926** 162

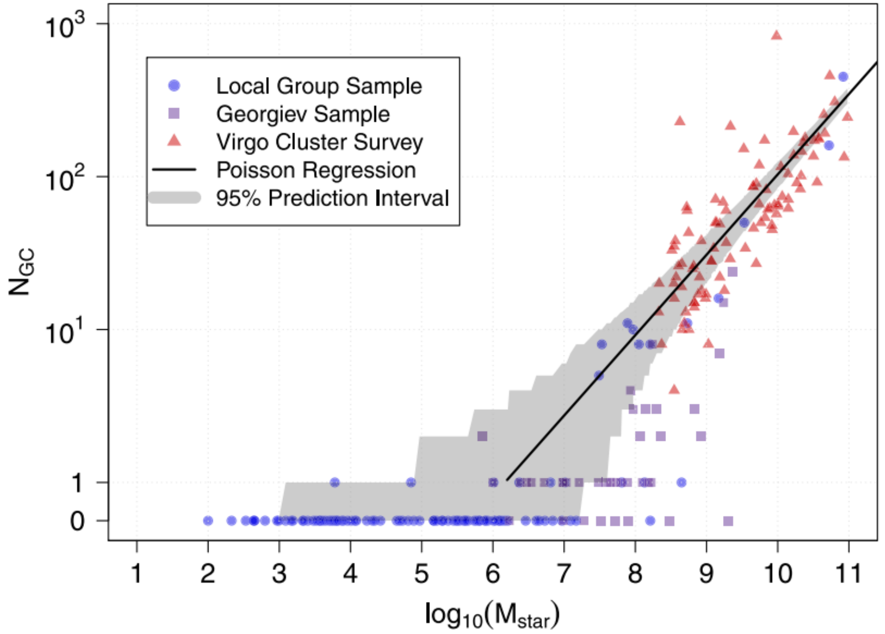
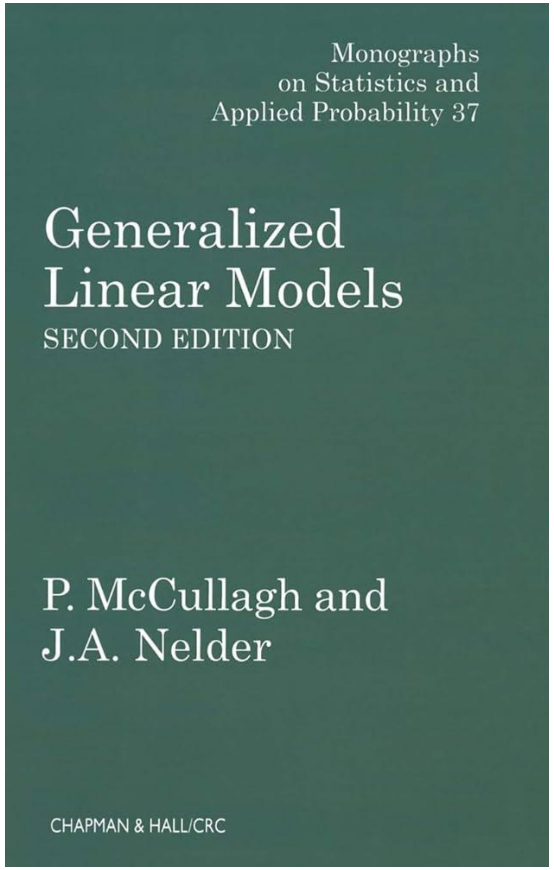
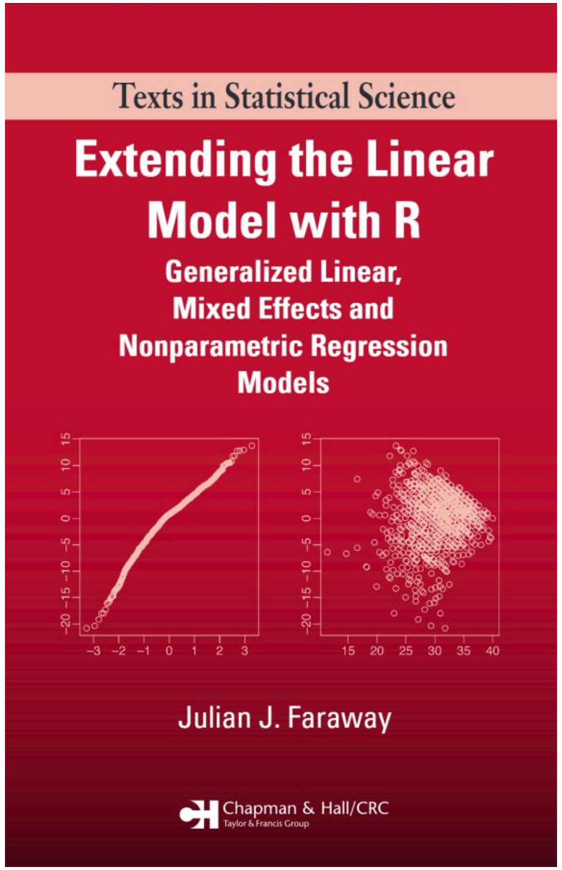


Figure 4. Top panel: the number of GCs in each galaxy as a function of galaxy stellar mass (points) and the Poisson regression fit (solid line). The line is not

Intro to Generalized Linear Models (GLMs)

- Introduced by Nelder and Wedderburn (1972)
- McCullagh and Nelder (1989)
- Concise introduction (ch.6) in Faraway (2005)
- GLMs *generalize* the linear model through a *link function*





Ch 6, *Extending the Linear Model with R (ELM),* Faraway

Some common GLMs

- Multinomial logit
  - Same as logistic regression, but for multiple categories
  - Categories may be ordered (e.g., strongly disagree, disagree, agree, …) or unordered (e.g., vanilla, strawberry, chocolate, …)
- Poisson
  - For responses that are counts
- Negative binomial (overdispersed Poisson)
- Zero-inflated models
- Logistic-binomial
- Hurdle

(Extra materials for reference)
Generalized Linear Models (GLMs)

GLM Definition

- Distribution of the response Y should be a member of the exponential family of distributions
- Exponential family of distributions:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Where θ is the *canonical parameter* (gives location! E.g., μ in Gaussian)

Where ϕ is the *dispersion parameter* (gives scale!) E.g., σ in Gaussian)

We write the above with functions $a(\phi)$, $b(\theta)$, and $c(y,\phi)$ because this makes the equation generalize to the entire exponential family.

Exponential family examples

- General form for exponential family of distributions

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

- Normal:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(y^2-2y\mu+\mu^2)}{2\sigma^2}\right] = \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

$\boxed{\mu=\theta}$

$\boxed{\phi=\sigma^2}$

$a(\phi)=1$

$b(\theta)=\frac{\theta^2}{2}$

$c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi^2} + \log(2\pi\phi)\right)$

GLM Definition

- Exponential family of distributions:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - \boxed{b(\theta)}}{a(\phi)} + c(y,\phi)\right]$$

- Poisson

$$f(y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = e^{-\mu} e^{y\log\mu} e^{-\log y!}$$

$$= \exp\left[\boxed{-\mu} + y\log\mu - \log y!\right]$$

$$\boxed{\begin{array}{l}\theta = \log\mu \\ \phi = 1\end{array}}$$

$a(\phi) = 1$

$b(\theta) = \exp(\theta)$

$c(y,\phi) = -\log y!$ (NOTE: doesn't depend on ϕ)

Exponential family distribution properties

- Expected value of Y

$$E[Y] = \mu = b'(\theta)$$

- Variance of Y

$$Var[Y] = b''(\theta)a(\phi)$$

(function of both position and scale parameters)

Link function – examples

- Example: in *Poisson Regression* (a GLM) use

$$g(\mu) = \log(\mu)$$

- Example: in *Logistic Regression* (a GLM) use

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

- Simplest example: in regular *(normal) Regression* (a LM)

$$g(\mu) = \mu$$

Canonical Link

- The *canonical* link has

$$\eta = g(\mu) = \theta$$

So $g(b'(\theta)) = \theta$

eg. Poisson

$b(\theta) = e^{\theta}$

$b'(\theta) = e^{\theta}$

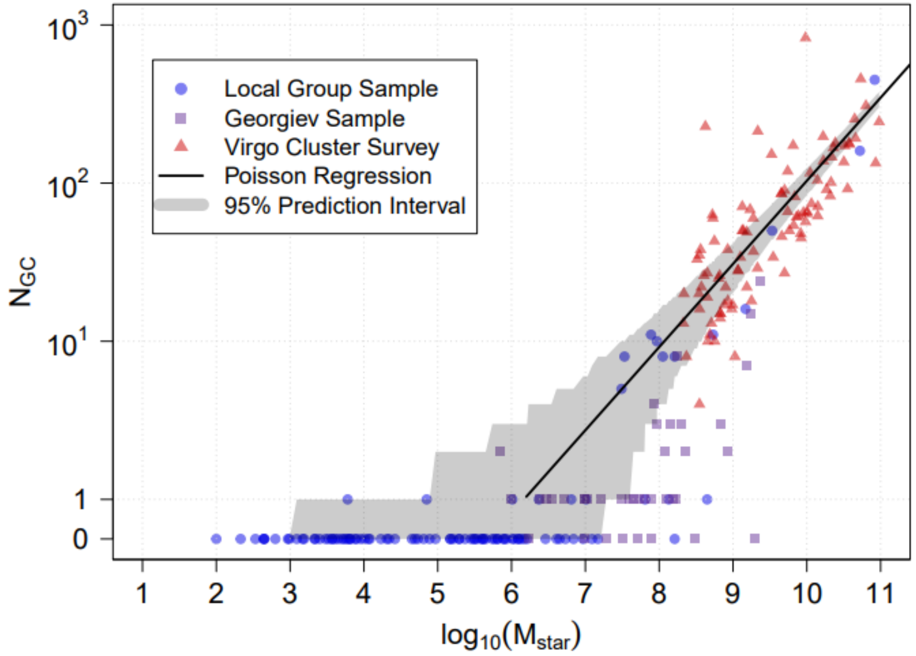
$g(\mu) = \theta = \log \mu$

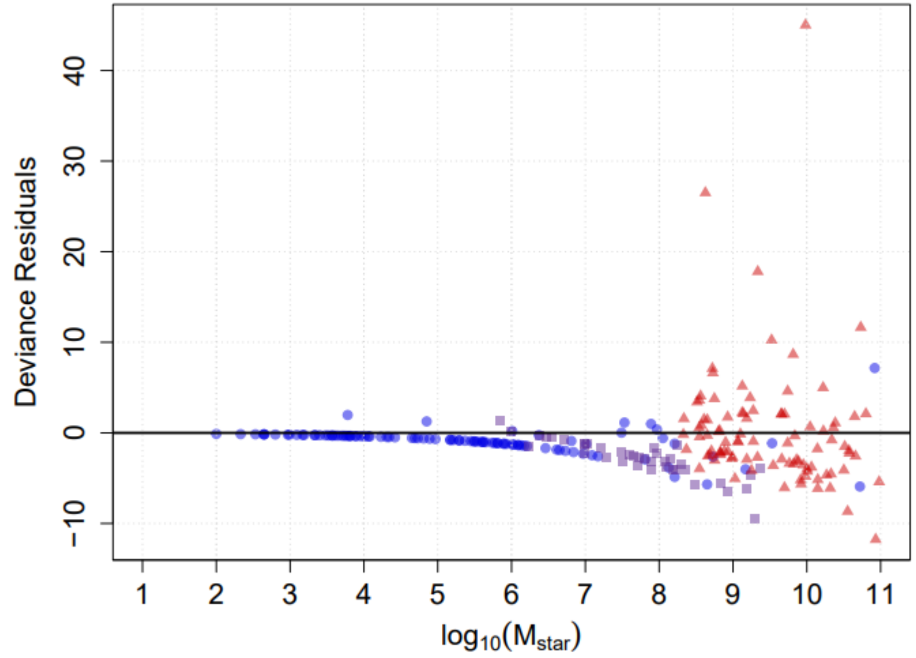
$g(b'(\theta)) = \log(e^{\theta}) = \theta$

- Common GLMs and their canonical links (Table 6.1 in ELM):

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = \log \mu$	μ
Binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse Gaussian	$\eta = \mu^{-2}$	μ^3

Ch 6, ELM





Results using Poisson Regression

$$\ln E[Y_i] = \beta_0 + \beta_1 \log M_{\star,i},$$

This is NOT a good model for GC counts!

$$\widehat{\beta_0} = -7.74 \;\; (- -7.72, -7.21)$$
$$\widehat{\beta_1} = 1.21 \; (1.18, 1.24)$$

(brackets are 95% confidence intervals)

Poisson Regression does not describe the data well:

- Counts are *overdispersed* (null hypothesis that data are equidispersed is rejected)
- Galaxies $10^6 - 10^9 M_{solar}$ have fewer GCs than expected by the model

Eadie, Harris, and Springford (2022), ApJ 926 162