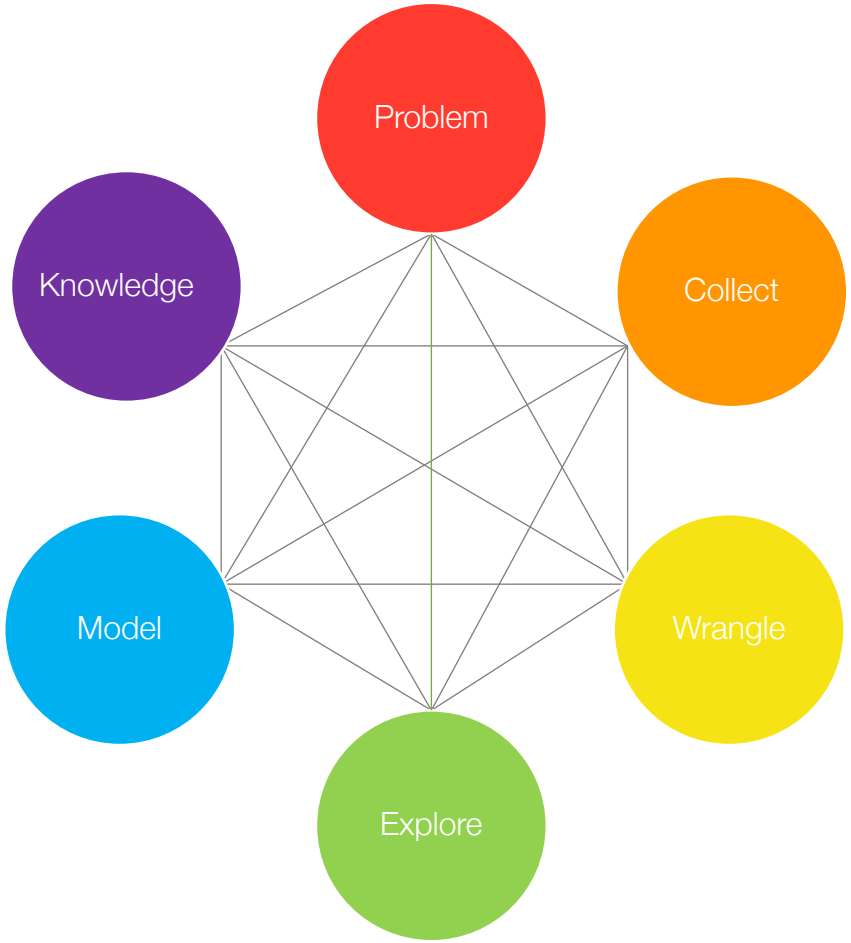
GEOG0125
ADVANCED TOPICS IN SOCIAL AND GEOGRAPHIC DATA SCIENCE

# INTRODUCTION TO BAYESIAN GENERALISED LINEAR MODELS (GLM)

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

# Contents

- What are Generalized Linear Models (GLMs)?
  - Link functions
- Selecting the appropriate type of statistical model
  - Linear regression model
  - Logistic regression model for Bernoulli OR Binomial
  - Poisson-based regression models (Normal, Negative Binomial & Zero-Inflated)
- What does each statistical model do?
  - Linear relationships
  - Log-odds and Odd Ratios (ORs)
  - Relative risk ratios (RRs)
- Interpretation of coefficients
- Model Specification from a Bayesian Framework



Remember in Term 1…
In Week 9's PSA lecture, we said…

# Multivariable Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Notes 1: Remember, in term 1 (week 9), we described what a linear regression model was before discussing at length what spatial lag and error models were etc.

Notes 2: We mentioned that a linear regression model such as the above formula allows the user to quantify the relationship (or association) between an outcome (i.e. dependent variable) with one, or more predictors (i.e., independent variable(s))

These models are good causal and predictive inference

Variables

- $y$ is the dependent variable
- $x_1, x_2, x_3, \ldots, x_k$ are the independent variables

Parameters

- $\beta_0$ is the intercept
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the slopes (or coefficients) for the corresponding variables $x_1, x_2, x_3, \ldots, x_k$
- $\varepsilon$ is the error term

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Here is a board overview:

Notes 1: Recall that we described how linear regression models are best modelling outcomes that are only continuous measure, where by we made the assumption that such continuous measures are from a Gaussian/normal distribution. Before, deep diving into spatial lag and error regression models… because are from the family of linear models but a spatial component to it.

| Distribution of dependent variable | Suitable Model |
|---|---|
| Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc., | Linear regression |
| Binary measures (1 = "present" or 0 = "absent"): e.g., Person's voting for a candidate, Lung cancer risk, house infested with rodents etc., | Logistic Regression |
| Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc., | Logistic Regression |
| Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc., | Poisson Regression |
| Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc., | Survival Analysis with Cox regression |

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Notes 1: Recall that we only touched on the fact the outcomes can measures that are from a different distribution, but we never really touched on this matter and on these particular classes of regression models

Here is a board overview:

| Distribution of dependent variable | Suitable Model |
|---|---|
| Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc., | Linear regression |
| Binary measures (1 = "present" or 0 = "absent"): e.g., Person's voting for a candidate, Lung cancer risk, house infested with rodents etc., | Logistic Regression |
| Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc., | Logistic Regression |
| Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc., | Poisson Regression |
| Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc., | Survival Analysis with Cox regression |

# What are Generalized Linear Models?

# Definition:

Generalized linear model (GLMs) is a flexible generalization of ordinary linear regression model, which allows the user to link some outcome *y*, to a link function g($\eta$), when that outcome is characterised by distribution that is from one the exponential families of distribution.

$$g(\eta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Exponential family are set of parametric (i.e., discrete or continuous) probability distributions. There are many… but the most common examples are:

- Normal
- Binomial
- Poisson
- Multinomial
- Negative binomial

Notes 1: There are a tonne of them, but you really don't have to worry about any of them. You only need to concern yourself with how this link function works!

# What is a link function $g(\eta)$? [1]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

- The linear regression model does not support any other outcome whose distribution is not from a Gaussian/Normal distribution.
- However, by using some link function, allows the user to transform such outcome (i.e., that's considered binary, polychotomous, discrete etc.,) that's typical nonlinear

Notes 1: Tricking the model to thinking it is linear

$$g(\eta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

- The type of link function implemented dependents on the type of analysis you are going to perform

# What is a link function $g(\eta)$? [2]

Here are the most frequent examples which you will certainly encounter

| Distribution of dependent variable | Exponential Family (Distribution) | Link Function | Suitable Model |
|---|---|---|---|
| Continuous measures | Normal distribution | Identity (we've been using this all this while) | Linear regression |
| Binary measures (1 = "present" or 0 = "absent") | Bernoulli distribution | Logit | Logistic Regression |
| Binomial measure (or proportion) | Binomial distribution | Logit function on aggregated outcome for successful and failures | Logistic Regression |
| Counts or discrete measures | Poisson distribution | Log or ln | Poisson Regression |

# Logistic Regression [1]

- This model allows the user to model binary outcomes linearly with other independent variables
- Examples of such outcomes can be from **Bernoulli distribution** e.g., disease status: no disease = 0 or disease = 1; Victimisation status: not burgled = 0 or burgled = 1; etc.,
- Other examples can also be from a **Binomial distribution** where binary responses are aggregated: e.g. total number of individual surveyed in a village (*N*) and number people detected to be positive (*n*)
- Link function:

$$g(\eta) = \text{logit}(p), \text{ where p is a probability}$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \text{ is what we called the "log-odds"}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Notes 1: With binary outcomes, we are dealing with probabilities and not averages

# Logistic Regression [2]

$g(\eta) = \text{logit(p)}$, where p is a probability

$\text{logit(p)} = \ln\left(\frac{p}{1-p}\right)$ is what we called the "log-odds"

$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$

- When estimate our coefficients i.e., $\beta_i$, which shows the linear relationship between the binary or binomial response variable with independent variable $x_i$ - they are always on the log-odds scale.
- For interpretability: we always take the exponential of our coefficient i.e., exp($\beta_i$), to convert in onto the scale of **odds ratios (OR)**

This is the quantity i.e., **Odds Ratios (OR)**, we want to estimate and interpret from our logistic regression

# Interpretation of Odds Ratios (OR)

OR = 1 (null value), it means that independent variable has no effect on the outcome

OR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

OR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

Notes 1: We use p-values and 95% CIs to deem whether the odd ratios are statistically significant or not.

Notes 2: If you want to do a risk assessment and your study design is a either a cross-sectional or case-control study design with some binary outcome, use a logistic regression model.

**Table 2.** Results of logistic regression models of iAs exposure by quintile and BCC in the ASHRAM study population [OR (95% confidence interval)].

| Arsenic exposure index/quintile (range of exposure in controls) | Adjusted[a] | Additionally adjusted[b] | Trend test (*p*-value) |
|---|---|---|---|
| Lifetime average iAs concentration (μg/L) | | | 0.001 |
| 0.00–0.68 | 1.00 | 1.00 | |
| 0.68–0.98 | 1.27 (0.82, 1.97) | 1.39 (0.89, 2.19) | |
| 0.98–7.00 | 1.02 (0.67, 1.56) | 1.20 (0.77, 1.88) | |
| 7.10–19.43 | 1.63 (0.93, 2.85) | 1.73 (0.97, 3.11) | |
| 19.54–167.29 | 2.81 (1.62, 4.87) | 3.03 (1.70, 5.41) | |
| Peak daily iAs dose rate (μg/day) | | | 0.001 |
| 0.00–0.73 | 1.00 | 1.00 | |
| 0.73–1.48 | 0.93 (0.62, 1.39) | 0.91 (0.59, 1.39) | |
| 1.48–9.09 | 1.29 (0.86, 1.95) | 1.55 (1.00, 2.41) | |
| 9.09–32.23 | 1.78 (1.05, 3.02) | 1.76 (1.01, 3.07) | |
| 32.23–242.14 | 2.31 (1.32, 4.03) | 2.50 (1.39, 4.49) | |
| Cumulative iAs dose (g) | | | 0.001 |
| 0.00–0.01 | 1.00 | 1.00 | |
| 0.01–0.03 | 1.02 (0.68, 1.52) | 1.09 (0.72, 1.67) | |
| 0.03–0.13 | 1.19 (0.78, 1.81) | 1.46 (0.93, 2.27) | |
| 0.13–0.55 | 1.73 (1.02, 2.91) | 1.76 (1.02, 3.04) | |
| 0.55–4.46 | 2.45 (1.39, 4.32) | 2.63 (1.45, 4.78) | |

Notes 1: An example of logistic regression model, applied to health risk assessment study determining the impacts of arsenic exposure (biomarkers) and skin cancer risk in Eastern Europe.

Interpretation for independent variable that is categorical

Giovanni et al. 2012 [Source]

# Poisson Regression [1]

- This model allows the user to model count or discrete outcomes linearly with other independent variables
- Examples of such outcomes can be from **Poisson distribution** e.g., number of COVID cases in postcodes across London; Number of houses on street segments that were victims to burglary etc.
- With particular scenario we are dealing with aggregated units and its either **counts** or **rates**
- Link function:

$g(\eta) = \ln(\lambda_i)$ i.e., log-link function (log of some mean rate $\lambda_i$).

$\ln(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$

# Poisson Regression [2]

$g(\eta) = \ln(\lambda_i)$ i.e., log-link function (log of some mean rate $\lambda_i$).

$\ln(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$

- When we estimate our coefficients i.e., $\beta_i$, which shows the linear relationship between the counts or discrete response variable with independent variable $x_i$ - they are always on the log-scale.
- For interpretability: we always take the exponential of our coefficient i.e., exp($\beta_i$), to convert it onto the scale of **risk ratios (RR)**

This is the quantity i.e., **Risk Ratios (RR) (interchangeable with Relative Risk)**, we want to estimate and interpret from our Poisson regression

# Interpretation of Risk Ratios (RR)

RR= 1 (null value), it means that independent variable has no effect on the outcome

RR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

RR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

Notes 1: We use p-values and 95% CIs to deem whether the risk ratios are statistically significant or not.

Notes 2: If you want to do a risk assessment and you study design is a either an ecological study design where you have counts and denominators for rates, and also for cohort study when following groups of participants prospectively

**Table 2**

Using a negative binomial Poisson regression model to report multivariable associations between street-level exposures and residential burglaries in Kaduna, Nigeria.

| Street exposure variables | Residential Burglary | |
|---|---|---|
| | CRR (95% CI) | VIF |
| **Intercept** | 0.69 (0.50–0.95)* | – |
| **Length of street segment (m)** | 1.01 (0.99–1.02) | 1.83 |
| **Connectivity** | 1.05 (1.03–1.08)** | 2.62 |
| **Betweeness (normalised index) (quartiles)** | | 2.33 |
| 1st Quartile (lowest) | 1.00 (referent) | |
| 2nd Quartile | 1.29 (0.99–1.66) | |
| 3rd Quartile | 1.55 (1.20–2.02)* | |
| 4th Quartile (highest) | 1.64 (1.19–2.29)* | |
| **Closeness (normalised index) (quartiles)** | | 1.12 |
| 1st Quartile (lowest) | 1.00 (referent) | |
| 2nd Quartile | 0.62 (0.49–0.78)* | |
| 3rd Quartile | 0.78 (0.61–0.99)* | |
| 4th Quartile (highest) | 0.62 (0.48–0.81)* | |
| **Business activity index (z-scores) (quintiles)** | | 1.18 |
| 1st Quintile (lowest) | 1.00 (referent) | |
| 2nd Quintile | 0.96 (0.72–1.29) | |
| 3rd Quintile | 0.71 (0.57–1.03) | |
| 4th Quintile | 1.47 (1.15–1.86)* | |
| 5th Quintile (highest) | 1.31 (1.01–1.68)* | |
| **Socioeconomic status (z-scores) (quintiles)** | | 1.15 |
| 1st Quintile (lowest) | 1.00 (referent) | |
| 2nd Quintile | 1.28 (1.00–1.63) | |
| 3rd Quintile | 0.95 (0.74–1.21) | |
| 4th Quintile | 0.79 (0.61–1.02) | |
| 5th Quintile (highest) | 0.81 (0.63–1.05) | |

*Significant with p-value $< 0.05$; Crime Rate Ratios (CRR); 95% Confidence Intervals (95% CI); Variance Inflation Factor (VIF).

Notes 1: An example of Poisson-based regression model, applied to crime victimisation study to determine the impacts of various environmental and society society risk factor (quantified on a street-level) and burglary risk in Nigeria.

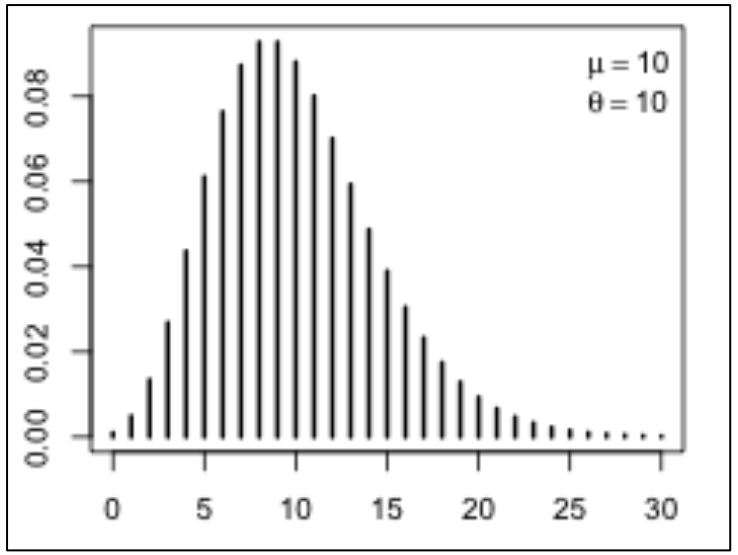
Interpretation for independent variable that are continuous as well as those that are categorical

Risk ratio which has been operationalised and termed as Crime risk ratio (CRR)

Musah et al. 2020 [Source]

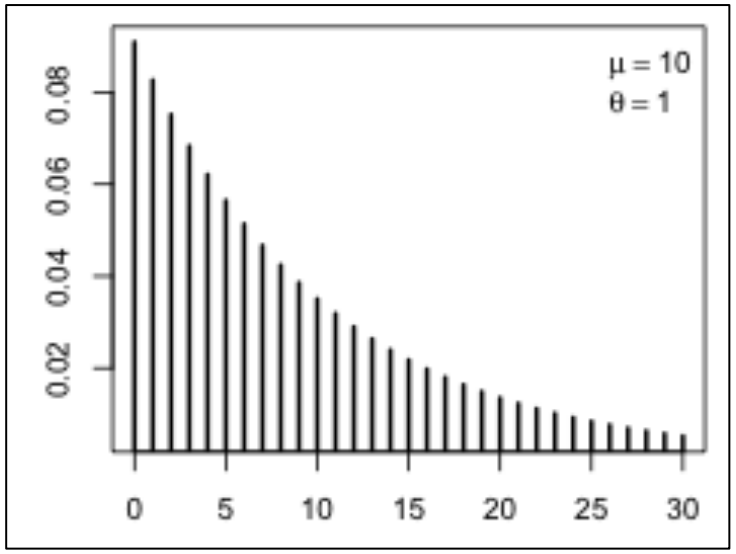
# Types of Poisson Regression

Examine the frequency distribution of the count response



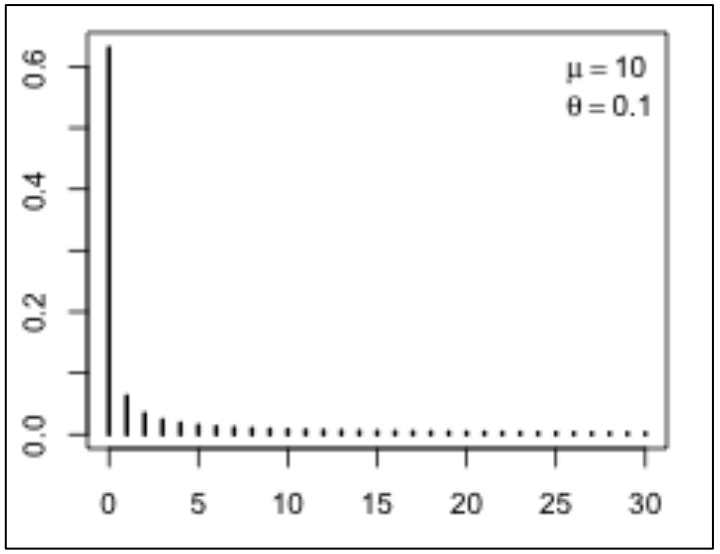

Scenario 1: Little to no dispersion

Use: regression Poisson model




Scenario 2: Over dispersed

Use: Negative Binomial Poisson model




Scenario 3: Strong over-dispersed response

Use: Zero-inflated Poisson model

Use as last restore if scenario is gives you unstable results.

# How do you code a Bayesian GLM in RStudio?

Specifications for model block:

- Logistic regression (Y : 1 or 0): **bernoulli_logit()**
- Logistic regression (Y : numerator & denominators): **binomial_logit()**
- Poisson regression (Y : counts or rates; normal): **poisson_log()**
- Poisson regression (Y : counts or rates; over-dispersed or zero-inflated): **neg_binomial_2_log()**

But we will do this practical together to demonstrate

Any questions?