Lattice Data

- **Disease Mapping**
- **Regression models**

Dra. ROSA ABELLANA

DPT. Fonaments Clínics (UB)

Count Data Models

- Variable coming from invidious account be studied using Poisson regression
- Number of cases ,$Y_{i,}$ has a Poisson distribution

$$Y_i \sim Poisson(\mu_i)$$

- Exemples:
  - Number of cancer cases in a region
  - Cells with chromosomal anomalies
  - Traffic accidents

Regression Poisson

$y_i \sim Poisson(\mu_i)$

$$log(\mu) = \alpha + \beta \cdot X$$

E(Y)=μ=exp(α+βX) Var(Y)=E(Y)

Exemple: Number of epileptic episodes in a month
Explanatory variable: Age, sex, treatment

R-package:
glm(num_epilectic~Age+sex+treatmet), family =poisson ,data = Dta.epi)

Model for a rates of incidence or mortality

If the observation units where the count has been carried out are not comparable because they correspond to:

- Different sample size
- Different duration of study period

$$Rate = \frac{number\ of\ cases}{People\ time\ at\ risk} = \frac{Y}{N \cdot t}$$

$$log\left(\frac{Y}{N \cdot t}\right) = \alpha + \beta \cdot X$$

$$log(Y) = \log(N \cdot t) + \alpha + \beta \cdot X$$

Log(N·t) is in the linear predictor but don't have regression coeffcient.

You must to defined as an offset

R-package:

glm(y~Age+sex **offset=lpob**, family =poisson , data = BBDD)

Interpretation of the coeficients:

exp(β) is a quocient of rates , risk relative=RR

$$RR = \frac{Rate\ incidence\ in\ exposed}{Rate\ incidence\ in\ not\ exposed}$$

RR>1. Risk factor: the probability to have the disease is superior if you have the risk factor

RR<1. Protective factor: the probability to have the disease is lower if you have the risk factor

RR=1. Disease and factor risk are not associated. The probability to have disease is the same if you have or not the risk factor.

R-package:

glm(y~Age+year, **offset=lpob**, family =poisson , data = BBDD)

Y =number of breast cancer in Tarragona

Age=age groups (<=50, >50 years) ⟶ Reference category <=50

Year=Year of diagnostic (2015,2016,2017)

Pob=population per year and age groups

- $\beta_{age>50} = 0{,}22 \longrightarrow \mathrm{RR}=\exp(0{,}22)=1{,}25$

  Woman older than 50 years has 25% more risk of cancer diagnostic

- $\beta_{year} = 0{,}024 \longrightarrow \mathrm{RR}_{\mathrm{year}}=\exp(0{,}024)=1{,}024$

  Percentage annual increase 2,4%

Application: Disease Mapping

Lattice data in the field of health sciences study the risk of suffering or dying of a particular area.

Main Objective: Analyze the geographical variability of rates of illness.

To eliminate possible differences among regions caused by variables, such as gender or age

↓

Indirect Standardized

- Internal
- External

Standardized mobility or mortality ratio

$$\hat{r}_i = \hat{SMR}_i = \frac{Y_i}{E_i}$$

Example of indirect standardization of rates: Rate of medical visits

REGION A

Age groups	Nº visits	Population	Rate
20-29	5	100	0,05
30-39	5	100	0,05
40-49	20	200	0,10
50-59	75	500	0,15
60-69	240	600	0,40
Total	345	1500	0,23

REGION B

Age groups	Nº visits	Population	Rate
20-29	20	500	0,04
30-39	24	400	0,06
40-49	20	200	0,10
50-59	30	300	0,10
60-69	32	100	0,32
Total	126	1500	0,084

Rate of medical visits all the country

Age groups	Nº visits	Population	Rates
20-29	25	600	0,04
30-39	29	500	0,06
40-49	40	400	0,10
50-59	105	800	0,13
60-69	272	700	0,39

Indirect Standardization (INTERNAL)

REGION A

Age groups	Nº visits	Population	Expected
20-29	5	100	4
30-39	5	100	6
40-49	20	200	20
50-59	75	500	66
60-69	240	600	233
Total	345	1500	329

REGION B

Nº visits	Population	Expected
20	500	5
24	400	12
20	200	20
30	300	45
32	100	38
126	1500	120

$$SMR_i = \frac{Observed_i}{Expected_i}$$

Region A: $SMR_A = \frac{345}{329} = 1{,}05$

Region B: $SMR_B = \frac{126}{120} = 1{,}05$

Interpretation of SMR

SMR_i> 1 indicates that there is a higher risk in the region of study that in the population of reference

SMR_i <1 less risk

SMR_i = 1 equal risk to the population of reference

Model for SMR

$$Y_i \sim Poisson(\mu_i)$$

$$log\left(\frac{Y}{E}\right) = \alpha + \beta \cdot X$$

$$log(Y) = \log(E) + \alpha + \beta \cdot X$$

$$\mu = E \times \exp(\alpha + \beta X)$$

Generalized linear models

The distribution of Y_i is a member of an exponential family , such as the Gaussian (normal), binomial, poisson, gamma, or inverse-Gaussian families of distributions.

A linear predictor —that is a linear function of regressors

$$\eta i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

A smooth and invertible linearizing link function g(·) , which transforms the expectation of the response variable, $\mu_i \equiv E(Yi)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

The link function is invertible, we can also write

$$\mu_i = g^{-}(\eta_i) = g^{-}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

Some common link function g(·)

LINK	$g(\mu_i)$
Identity	μ_i
Log	$Log_e(\mu_i)$
Inverse	μ_i^{-1}
Inverse-square	μ_i^{-2}
Square-root	$\mu_i^{-1/2}$
Logit	$log\left(\frac{\mu_i}{1-\mu_i}\right)$
Probit	$\Phi^{-}(\mu_i)$
Log-log	$-log_e[-log_e(\mu_i)$
Complementary log-log	$log_e[-log_e(1-\mu_i)$

Problems with Poisson

- The equality of mean and variance of Poisson distribution places restriction on the applicability of this model in the real-world data.
- An issue of importance is when empirical variance in the data exceeds the nominal variance under presumed model.
- It is called overdispersion.
- Overdispersion lead to underestimated standard errors and overestimated significance of regression parameters

Causes of overdispersion

- variability of experimental material
  - individual level variability
- correlation between individual responses
  - e.g. litters of rats
- cluster sampling
  - e.g. areas; schools; classes; children
- aggregate level data
- omitted unobserved variables
- excess zero counts (structural and sampling zeros)
- Presence of spatial correlation

Deviance is a measure of how well the model fits the data

For a fitted Poisson regression the deviance is equal to

$$D = 2\sum_{i=1}^{n}\{Y_i \log(Y_i/\mu_i) - (Y_i - \mu_i)\}$$

Where $\widehat{\mu_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p)$

If the model fits well,
the observed values Y_i will be close to their predicted means μ_i,
causing both of the terms in D to be small
Deviance to be small.

Models for Over dispersed Count Data

If the Poisson model fits the data reasonably,
we would expect the **deviance** to be roughly equal to the **degrees of freedom (n-p)**.

That the residual deviance is so large suggests that the variation of the data **exceeds** the variation of a Poisson-distributed variable, for which the variance equals the mean. This common occurrence in the analysis of count data is termed over dispersion

Solutions of this problem

1. Fit a Poisson quasi likelihood. The quasi-poisson families differ from the Poisson families only in that the dispersion parameter is not fixed at one, so they can model over dispersion

2. Fit Negative Binomial GLM.

3. If the over dispersion is due to spatial correlation. Solution: Add random effects with a spatial correlation matrix. Extension of the generalized linear models are the generalized mixed models (GLMM)

4. Particular kind over dispersion arises from excess of zeroes
   Solution: Zero-inflated Poisson regression

1. Fit a Poisson quasi likelihood

To introduce a dispersion parameter into the Poisson model, so that the conditional variance of the response is

$$V(Y) = \phi\mu$$

If ϕ>1, therefore, the conditional variance of Y increases more rapidly than its mean.

We use a method-of-moments estimator for the dispersion parameter.

$$\Phi = \frac{1}{n-k-1}\sum \frac{(Yi - \widehat{\mu_i})^2}{\widehat{\mu_i}}$$

2. A simple way to allow for a higher variance is to use **Negative Binomial** distribution instead of the **Poisson**

Under the Poisson the mean, μ, is assumed to be constant within classes.

But, if we define a specific distribution for μ, heterogeneity within classes can be used.

- Assume μ to be Gamma(α,β)
  $E(\mu) = \mu = \alpha/\beta$ then $\beta = \alpha/\mu$
  $Var(\mu) = \alpha / \beta^2 = \mu^2/\alpha$

The distribution of Y| μ to be the Poisson distribution with conditional mean E(Y | μ)= μ

The marginal distribution of Y_i, is **a binomial negative**

$$P(Y_i = y_i) = \int P(Y_i = y_i \mid \theta_i) f(\theta_i) d\theta_i$$

with

E(Y)=E[E[Y|μ]]=E[μ]= μ

Var(Y)=E[Var(Y|μ)] + Var(E[Y|μ])= E[μ]+ Var(μ)= =$\mu + \mu^2/\alpha$

By letting a = α^{-1} ,
Y follows a BN with E(Y)= μ , and Var(Y) = μ(1+a μ)
where **a** denotes the dispersion parameter.

Note: If **a=0,** there would be **no over dispersion.**

4. Overdispersion when there are more zeroes

A particular kind of over dispersion obtains when there are more zeroes in the data than is consistent with a Poisson (or negative-binomial) distribution, a situation that can arise when only certain members of the population are “at risk” of a nonzero count.

Model proposed for count data with an excess of zeroes, the zero-inflated Poisson regression (or ZIP) model, due to Lambert (1992).

ZIP model consists of two components:

1. Binary logistic regression model for membership in the latent class of individuals for whom the response variable is necessarily 0.
2. A Poisson-regression model for the latent class of individuals for whom the response may be 0 or a positive count.

Simple special case:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_0 \quad \text{(constant)}$$

And

$$log(\mu_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

The probability of observing a 0 count is:

$$P(Y_i = 0) = \pi_i + (1 - \pi_i)xe^{-\mu_i}$$

The probability of observing any particular nonzero count y_i is:

$$P(Y_i) = (1 - \pi_i)x\frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}$$

Expectation and variance for ZIP model

$$E(Y_i) = (1 - \pi_i) \cdot \mu_i$$

$$\text{Var}(Y_i) = (1 - \pi_i) \cdot \mu_i \cdot (1 + \pi_i \mu_i)$$

With Var>Expectation for π_i>0

MIXED models to incorporated random effects with spatial structure

Mixed model related to conditional mean to random effects.

$$E(y \mid b) = \mu^b$$

$$g\left(\mu^b\right) = \eta^b = X\beta + Zb$$

where X is the design matrix for the fixed effects

β Is the vector of the fixed parameters

Z is the design matrix for the random effects

b is the vector (dim= N) the random effects

$$b \sim NMV\left(0, \Sigma\left(\delta\right)\right)$$

Different structured for random effects

1. Heterogeneity model $b \sim NMV(0, \tau^2 I)$

Without spatial structure, assumes not spatial correlation

2. Spatial structured model

Introduced in this context by Clayton and Kaldor (1987)

Intrinsic CAR model $\rho=1$ $b \sim N\left(0, \tau_S^2 \left(D_w - W\right)^{-}\right)$

The conditional distribution of the random effects is a Normal

$$b_i \mid b_{j \in \delta_i} \sim N\left(\bar{b}_i, \sqrt{\frac{\sigma_s^2}{n_i}}\right)$$

Non- intrinsic CAR

Proposed by Besag, York and Mollié,1991, two variability components: over dispersion non-structured and another over dispersion spatial structured. The conditional distribution for random effects:

$$b \sim N\left(0, \tau_H^2 I_N + \tau_S^2 \left(D_w - \rho W\right)^{-}\right)$$

In this case

$$E\left[b_i \mid b_j \; i \neq j\right] = \bar{b}_i = \frac{\sum_{j \in \partial_i} b_i}{n_i}$$

$$Var\left(b_i \mid b_j \; i \neq j\right) = \sigma_H^2 + \frac{\sigma_S^2}{n_i}$$

Maxixum likelihood of the parameters

$$L\left(Y;\theta,\phi,\Sigma(\varpi)\right)=\int\prod_{i=1}^{N} f_Y(y;\varpi,\phi)\cdot f\left(b,\Sigma(\theta)\right)d(b)=$$

$$\int\prod_{i=1}^{N}\exp\left\{\frac{\left(y_i\cdot\theta_i-b(\theta_i)\right)}{a(\phi)}+c(y,\phi)\right\}\cdot f\left(b,\Sigma(\varpi)\right)d(b)$$

$$=\int\prod_{i=1}^{N}\exp\left\{\frac{\left(y_i\cdot\theta_i-b(\theta_i)\right)}{a(\phi)}+c(y,\phi)\right\}\cdot\frac{1}{\sqrt{(2\pi)^N\Sigma(\varpi)}}\exp\left\{-\left(1/2\right)\cdot b\cdot\Sigma^{-}(\varpi)\cdot b^T\right\}db$$

Since the random effects are not independent, the maximum likelihood function is not able to assess in a closed form.

Estimation of the parameters:

Perspective

- Bayesian → Fully Bayesian
- Frequentist → Penalized quasi-likelihood (PQL)

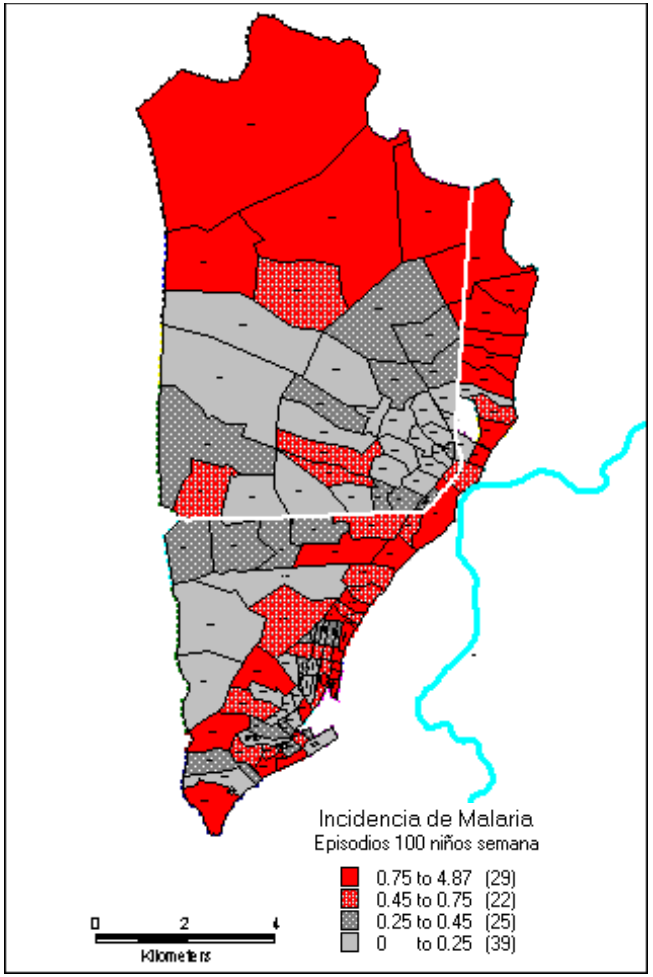
Besag and Clayton (1993)

Example

Data: Following of 2006 children between 1 and 10 years old

Cohort (2 years) visited once a week, considering a week at risk

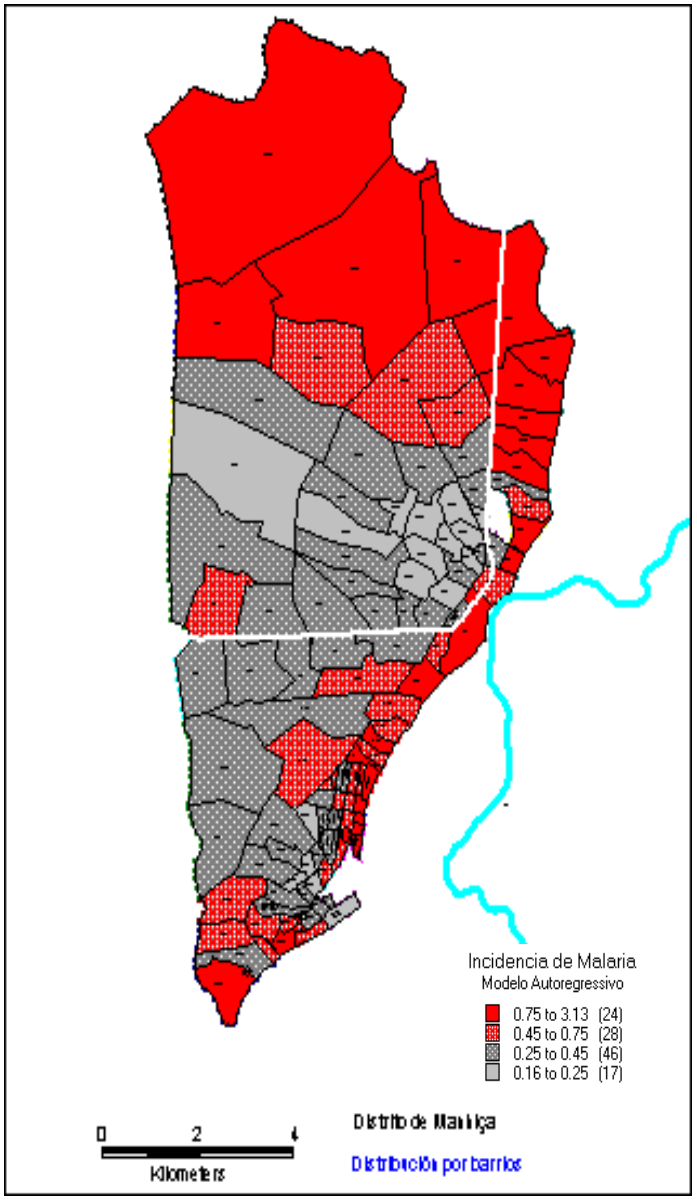
Incidence is expressed by number of episodes for 100 children week at risk

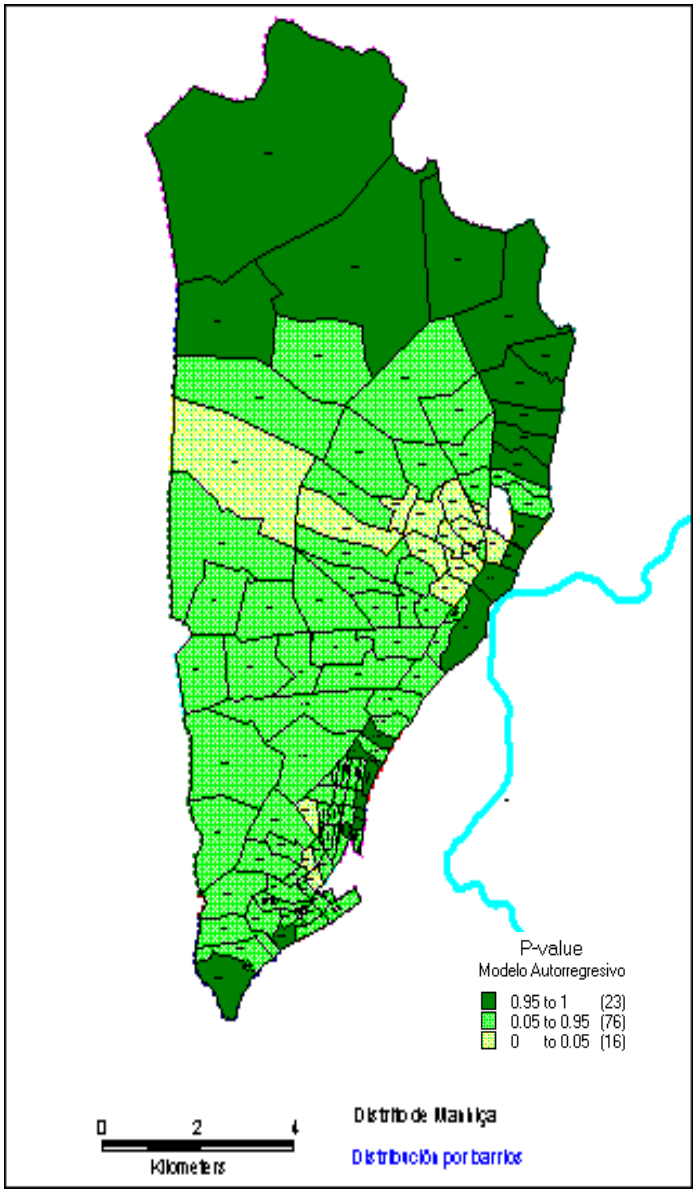


RESULTS

Model	Mean (1)	Desv. Het	Desv. CAR	DIC
Heterogeneity	0.43	0.748		194.97
CAR	0.43		0.731	176.11
Heterogeneity + CAR	0.42	0.327	0.664	178.65

100 children week at risk


Incidencia de Malaria
Modelo Autoregressivo
0.75 to 3.13 (24)
0.45 to 0.75 (28)
0.25 to 0.45 (46)
0.16 to 0.25 (17)
0 2 4
Kilometers
Distrito de Manlleu
Distribución por barrios


P-value
Modelo Autorregresivo
0.95 to 1 (23)
0.05 to 0.95 (76)
0 to 0.05 (16)
0 2 4
Kilometers
Distrito de Manlleu
Distribución por barrios

Bibliografia:

BOOKS

Cressie N.A. (**1993**). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. Canada. **Chapter 6**

Banerjee S Carlin BP, Gelfrand A.E. (2004) Hierarchical Modelling and Analysis for Spatial Data. Chapman & Hall /CRC. **Chapter 3**

Clayton D.and Bernardinelli L. (**1992**). *Bayesian methods for mapping disease risks*. In small Area Studies in Geographical and Environmental Epidemiology (P.Elliot, J. Cuzich, et al) Oxford University Press. 205-20.

Moreno R i Vayà E. (2000). Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial. Ediciones UB. 44 manuals.

PAPERS

Besag J. (**1974**). *Spatial interaction and the statistical analysis of lattice systems*. Journal of the Royal Statistical Society, Series B. **36**:192-236.

Besag J., York J. and Mollié A.(**1991**). Bayesian image restoration, with applications in spatial statistics. Annals of the Institute of Statistical Mathematics. **43**: 1-59.

Breslow N.E. and Clayton D.G.(**1993**). *Approximate Inference in Generalized Linear Mixed Models*, Journal of the American Statistical Association. **88**: 9-25.

Clayton D. and Kaldor J. (**1987)**. *Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping*. Biometrics, **43**: 671-681