Exploratory analysis

Dr Emanuele Giorgi
Lancaster University
e.giorgi@lancaster.ac.uk

March 23, 2022

Pre-requisites

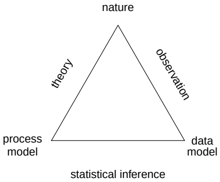
- Good knowledge of generalized linear models
- Basic knowledge of R
- Notions of probability calculus (e.g. conditional distribution and expectation).
- Basic mastering of mathematical equations

Learning outcomes of the workshop

You should be able:

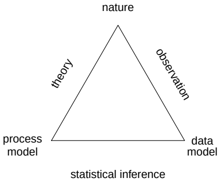
- to understand the limitations of generalized linear models;
- to test for the presence of spatial correlation using variogram-based techniques;
- to formulate a suitable geostatistical model for data-analysis;
- to understand and correctly interpret the results from a geostatistical analysis;
- to fit generalized linear geostatistical models and carry out spatial prediction using `PrevMap` in R.

Science and statistics



Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

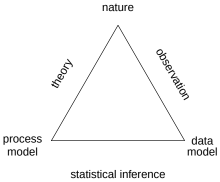
Science and statistics



Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

- $S =$ "process of nature"
- $Y =$ "data"

Science and statistics

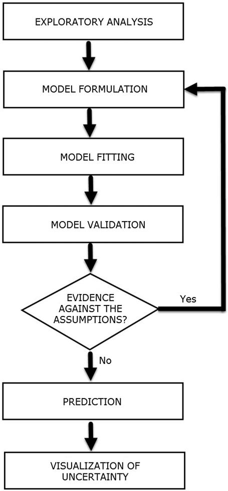


Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

- $S =$ "process of nature"
- $Y =$ "data"

$$[Y, S] = [S][Y|S]$$

Statistical analysis



A close look at generalized linear regression

Assumptions:

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)
- Our model for the data: $[S][Y|S]$.

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)
- Our model for the data: $[S][Y|S]$.
- Under the assumptions of classical GLMs, we can ignore $[S]$.

What is the purpose of statistical modelling?

- **Prediction:** developing a probabilistic model that can accurately predict future realizations of Y
- **Explanation:** developing a probabilistic model that can reliable explain and quantificy the association bewteen Y and a covariate d.

Example: River-blindness in Liberia

Example: River-blindness in Liberia

- An introduction to the disease:
  https://www.youtube.com/watch?v=PIJ8UYDAF3M

Example: River-blindness in Liberia

- An introduction to the disease: https://www.youtube.com/watch?v=PIJ8UYDAF3M
- Y_i="number of positively tested individuals for river-blindness out of n_i.

Example: River-blindness in Liberia

- An introduction to the disease:
https://www.youtube.com/watch?v=PIJ8UYDAF3M
- Y_i=“number of positively tested individuals for river-blindness out of n_i.
- d_i=“elevation of the i-th village”

Example: River-blindness in Liberia

- An introduction to the disease: https://www.youtube.com/watch?v=PIJ8UYDAF3M
- Y_i="number of positively tested individuals for river-blindness out of n_i.
- d_i="elevation of the i-th village"
- **Question:** How should we formulate and estimate a model to understand the association between d_i and the probability of being positive for river-blindness, p_i? (script1.R)

Over-dispersion

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim \text{Bernoulli}(p)$ and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim \text{Bernoulli}(p)$ and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim \text{Bernoulli}(p)$ and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

1. *Marginal models.* e.g. quasi-models, $E[Y_i] = m_i \mu_i$ and $V[Y_i] = \phi m_i V(\mu_i)$ where ϕ is the over-dispersion parameter.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim \text{Bernoulli}(p)$ and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

1. *Marginal models.* e.g. quasi-models, $E[Y_i] = m_i \mu_i$ and $V[Y_i] = \phi m_i V(\mu_i)$ where ϕ is the over-dispersion parameter.
2. *Random effects models.* $S = d^\top \beta + Z$, where Z is a stochastic process.

A class of generalized linear mixed models

Assumptions:

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i | Z_i \sim f(\cdot)$ belongs to the exponential family;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i | Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i | Z_i] = m_i \mu_i$ and $Var[Y_i | Z_i] = m_i V(\mu_i)$;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i | Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i | Z_i] = m_i \mu_i$ and $Var[Y_i | Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i | Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i | Z_i] = m_i \mu_i$ and $Var[Y_i | Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;
5. $Y_i | Z_i$ are mutually independent for $i = 1, \ldots, n$.

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i|Z_i] = m_i\mu_i$ and $Var[Y_i|Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;
5. $Y_i|Z_i$ are mutually independent for $i = 1, \ldots, n$.

Are the Y_i mutually independent?

- **Examples:** 1) $Y_i|Z_i \sim \text{Poisson}(e^{d_i^\top \beta + Z_i})$ and $Z_i \sim N(-\tau^2/2, \tau^2)$ i.i.d.; $E[Y_i] =$ .... and $Var[Y_i] =$ .... (Hint: Use the law of total expectation and variance)
  2) $Y_i|Z_i \sim \text{Poisson}(e^{d_i^\top \beta + Z_i})$, $e^{Z_i} \sim \text{Gamma}(k, k)$ i.i.d.; show that Y_i is a Negative Binomial distribution.

Parameter estimation

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \ldots, n$

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \ldots, n$
- $\theta = (\beta, \sigma^2)$ (vector of unknown parameters)

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \ldots, n$
- $\theta = (\beta, \sigma^2)$ (vector of unknown parameters)
- The likelihood function

$$L(\theta) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} [Z_i][Y_i|Z_i] \, dY_i$$

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \ldots, n$
- $\theta = (\beta, \sigma^2)$ (vector of unknown parameters)
- The likelihood function

$$L(\theta) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} [Z_i][Y_i|Z_i] \, dY_i$$

- Maximize the likelihood using the Laplace approximation (`glmer` in the `lme4` package).

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \ldots, n$
- $\theta = (\beta, \sigma^2)$ (vector of unknown parameters)
- The likelihood function

$$L(\theta) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} [Z_i][Y_i|Z_i] \, dY_i$$

- Maximize the likelihood using the Laplace approximation (`glmer` in the `lme4` package).
- Hypothesis testing on $\beta = \beta_0$ (H_0).

Parameter estimation

- $Z_i \sim N(0, \sigma^2)$ i.d.d. for $i = 1, \dots, n$
- $\theta = (\beta, \sigma^2)$ (vector of unknown parameters)
- The likelihood function

$$L(\theta) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} [Z_i][Y_i|Z_i]\, dY_i$$

- Maximize the likelihood using the Laplace approximation (`glmer` in the `lme4` package).
- Hypothesis testing on $\beta = \beta_0$ (H_0).
  1. Obtain $\hat{\theta}$ (MLE).
  2. Obtain $\hat{\theta}_0$, the MLE constrained by fixing p values of β to 0.
  3. Compute the log-likelihood ratio

$$D = 2(logL(\hat{\theta}) - logL(\hat{\theta}_0)) \sim \chi^2_p$$

  4. P-value: $P(D > D_{obs}|H_0)$

Example: River-blindness in Liberia (Revisited)

- Y_i=“number of positively tested individuals for river-blindness out of n_i.
- d_i=“elevation of the i-th village”
- **Question:** How should we account for overdispersion? (script2.R)