ANALYSIS OF SPATIAL DATA IN EPIDEMIOLOGY

Prof. Dr. Maria A Barceló and Prof. Dr. Marc Saez

September 8, 10, 14 and 16, 2021

Research Group on Statistics, Econometrics and Health (GRECS), University of Girona
CIBER of Epidemiology and Public Health (CIBERESP)

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

COURSE INTRODUCTION

1. Course introduction
2. Introduction to epidemiology and spatial statistics
3. **Overview of mixed models**
4. Overview of mixed models - Practicals
5. Introduction to INLA and R INLA
6. R INLA - Practicals

Wednesday 8

Friday 10

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

7. Disease mapping. Standardisation of incidence and mortality rates
8. Disease mapping. Smoothing standardised incidence and mortality rates
9. Disease mapping – Practicals
10. Geographical association studies. Spatial ecological regression
11. Spatial ecological regression - Practicals

Tuesday 14

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

12. Clustering
13. Extensions: BYM2, point processes, leaflet, pc priors
14. Extensions – Practicals

Thursday 16

Marc Saez Maria Antònia Barceló

OVERVIEW OF MIXED MODELS

1. Linear regression model
2. Logistic regression model
3. Generalised linear model (GLM)
4. Mixed models

OVERVIEW OF MIXED MODELS

1. **Linear regression model**
2. Logistic regression model
3. Generalised linear model (GLM)
4. Mixed models

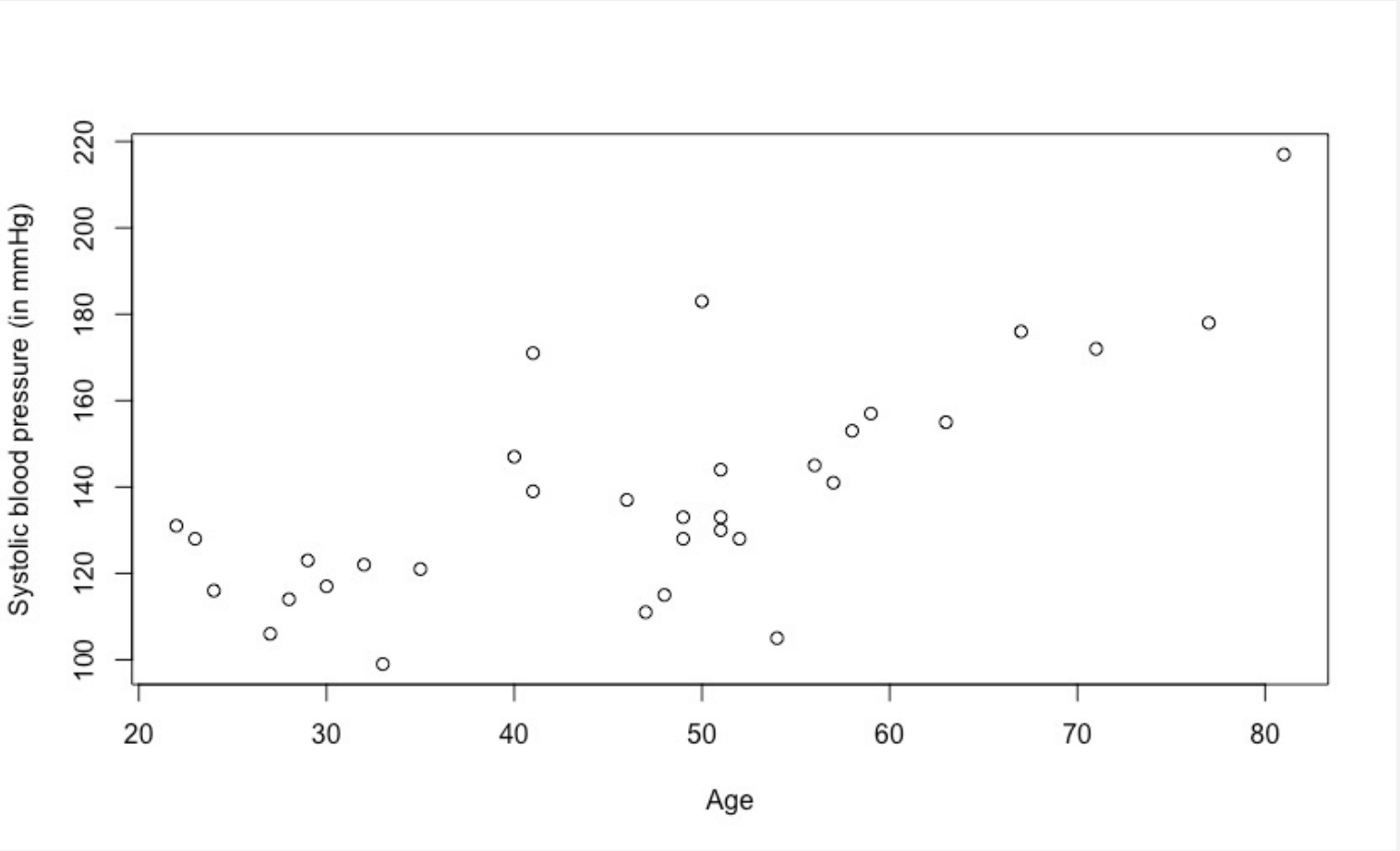
LINEAR REGRESSION MODEL

Table 1 Age and systolic blood pressure (SBP) among 33 adult women

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

Adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

LINEAR REGRESSION MODEL

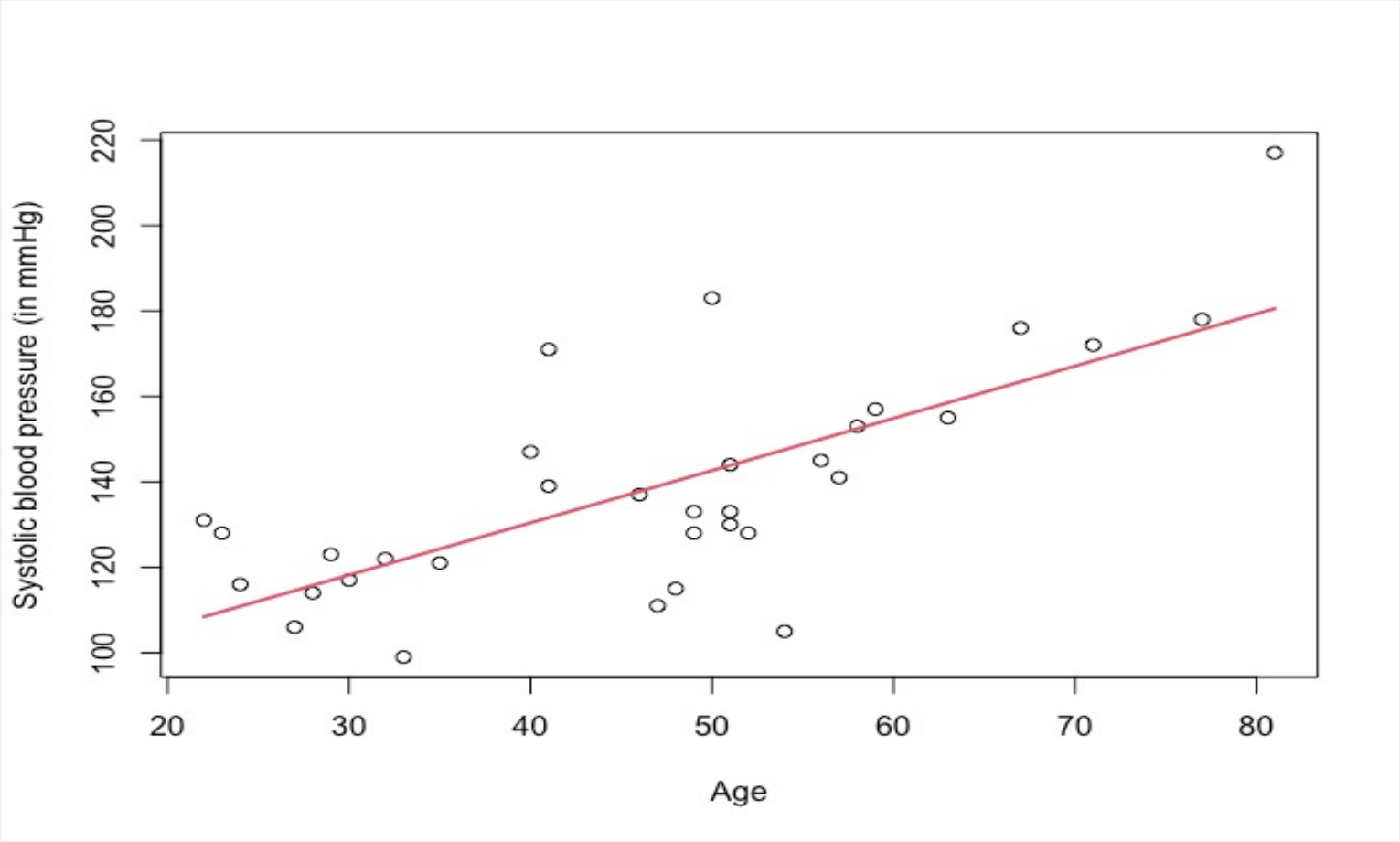


LINEAR REGRESSION MODEL

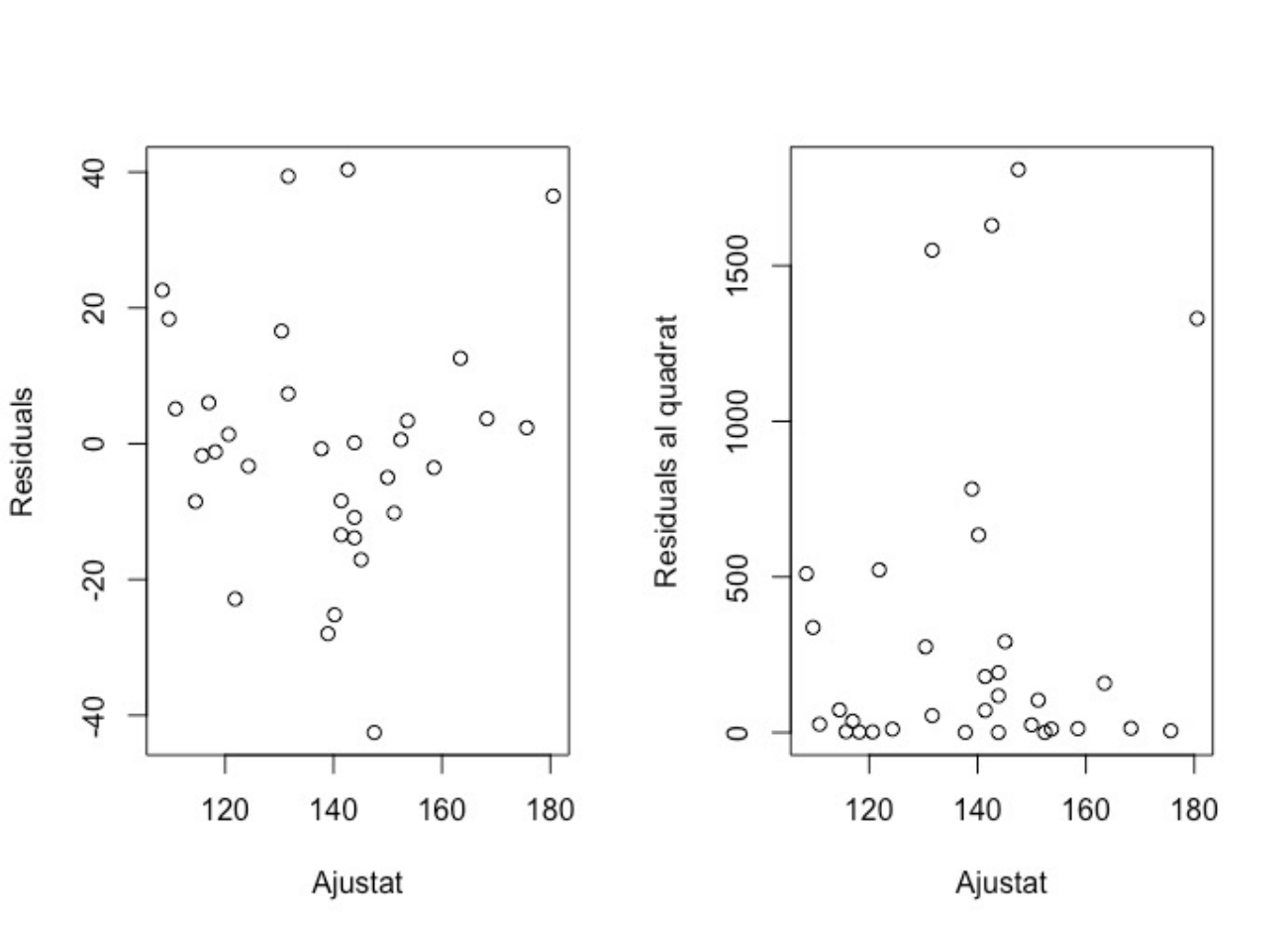
$$SBP_i = \beta_0 + \beta_1 age_i + u_i$$

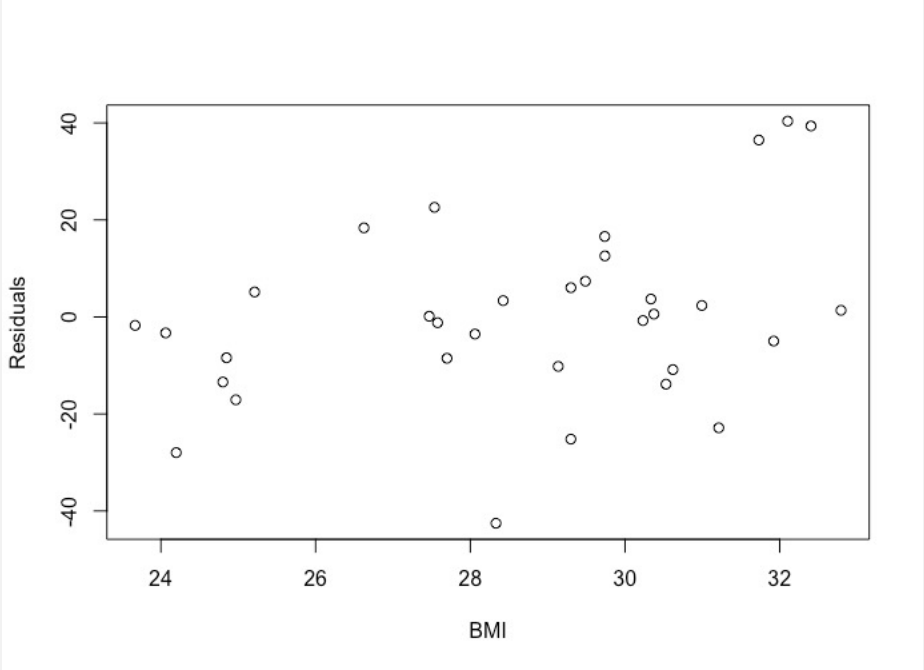
$$\widehat{SBP}_i = \hat{\beta}_0 + \hat{\beta}_1 age_i$$

$$\widehat{SBP}_i = 81{,}5161 + 1{,}224\ age_i$$



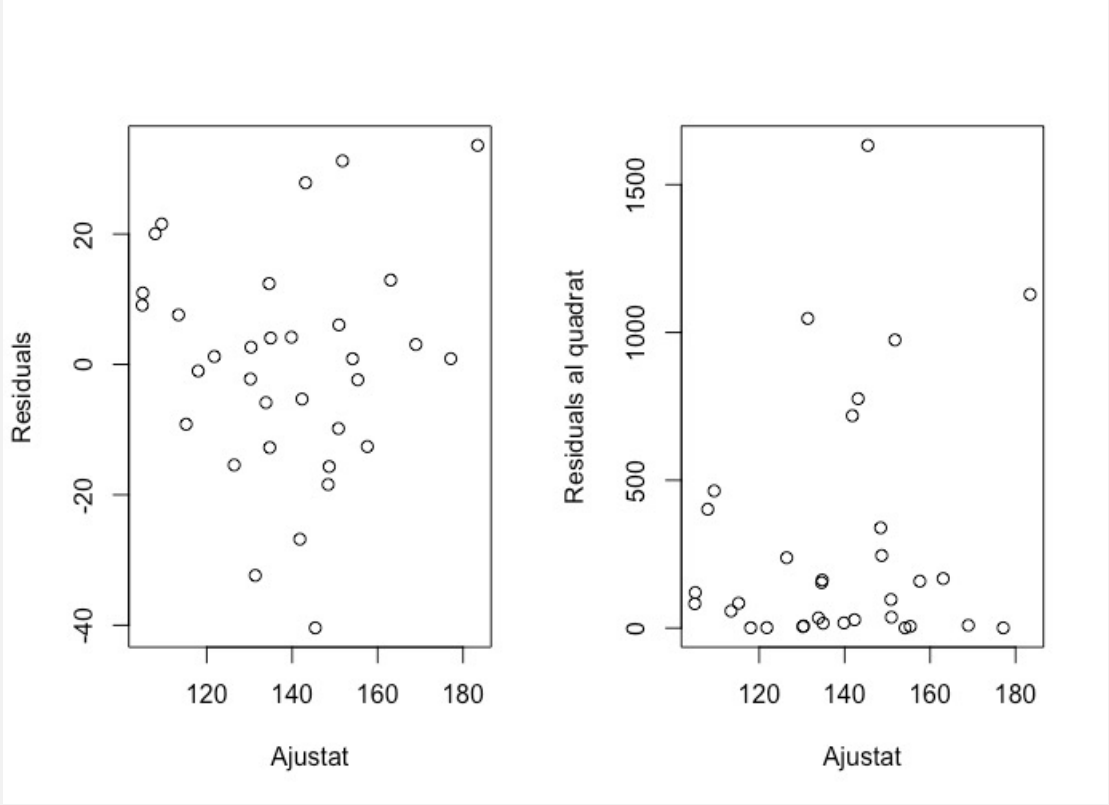
LINEAR REGRESSION MODEL





LINEAR REGRESSION MODEL

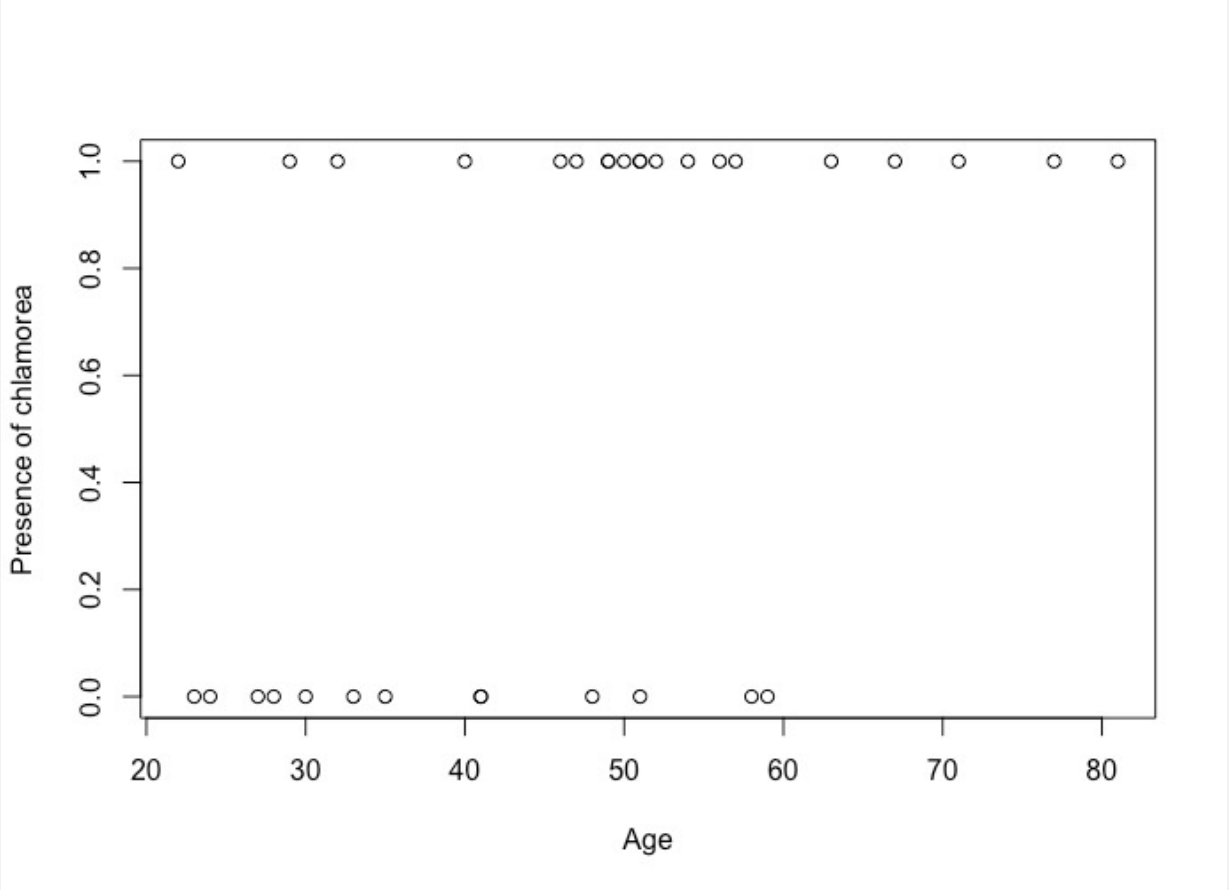
$$\widehat{SBP}_i = 8{,}9631 + 1{,}0538\, age_i + 2{,}8075\, BMI_i$$



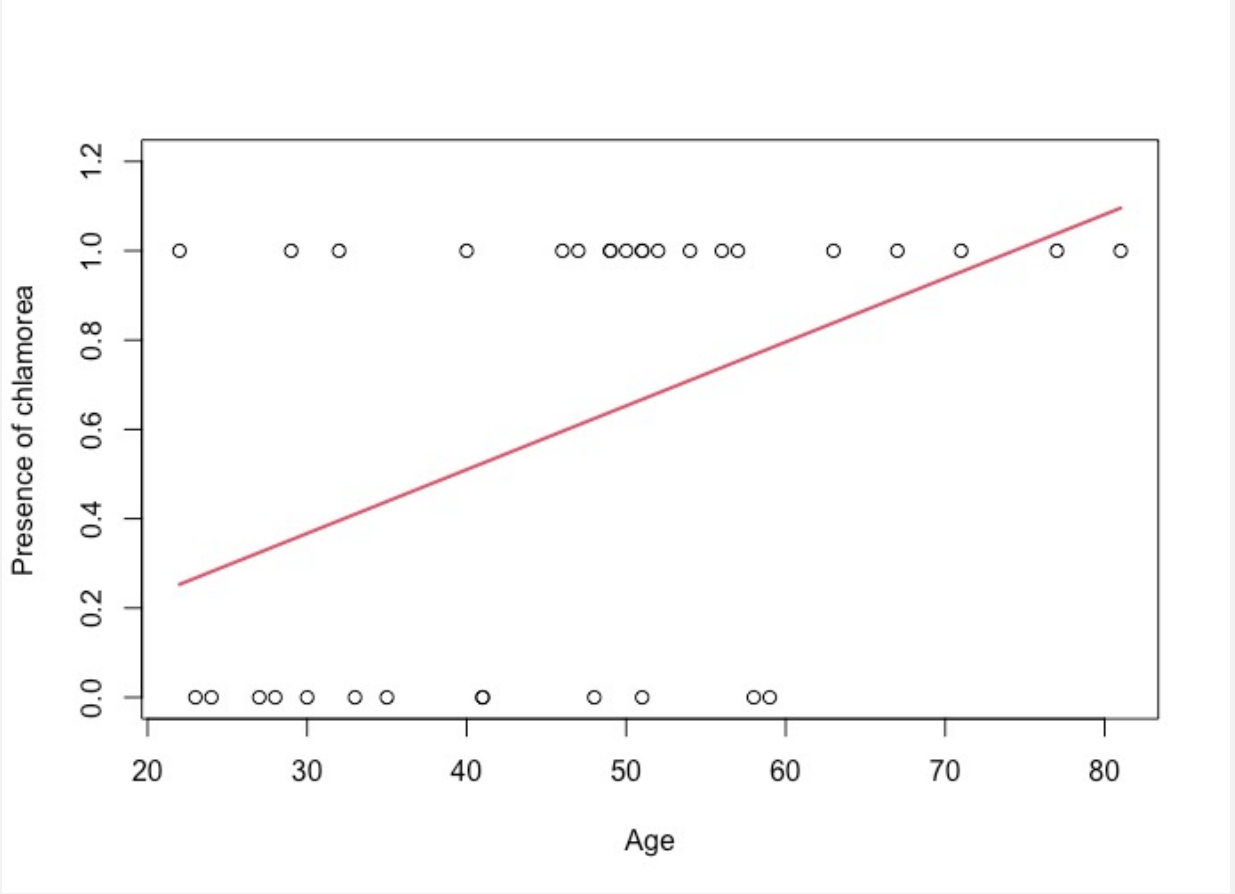
OVERVIEW OF MIXED MODELS

1. Linear regression model
2. **Logistic regression model**
3. Generalised linear model (GLM)
4. Mixed models

LOGISTIC REGRESSION MODEL



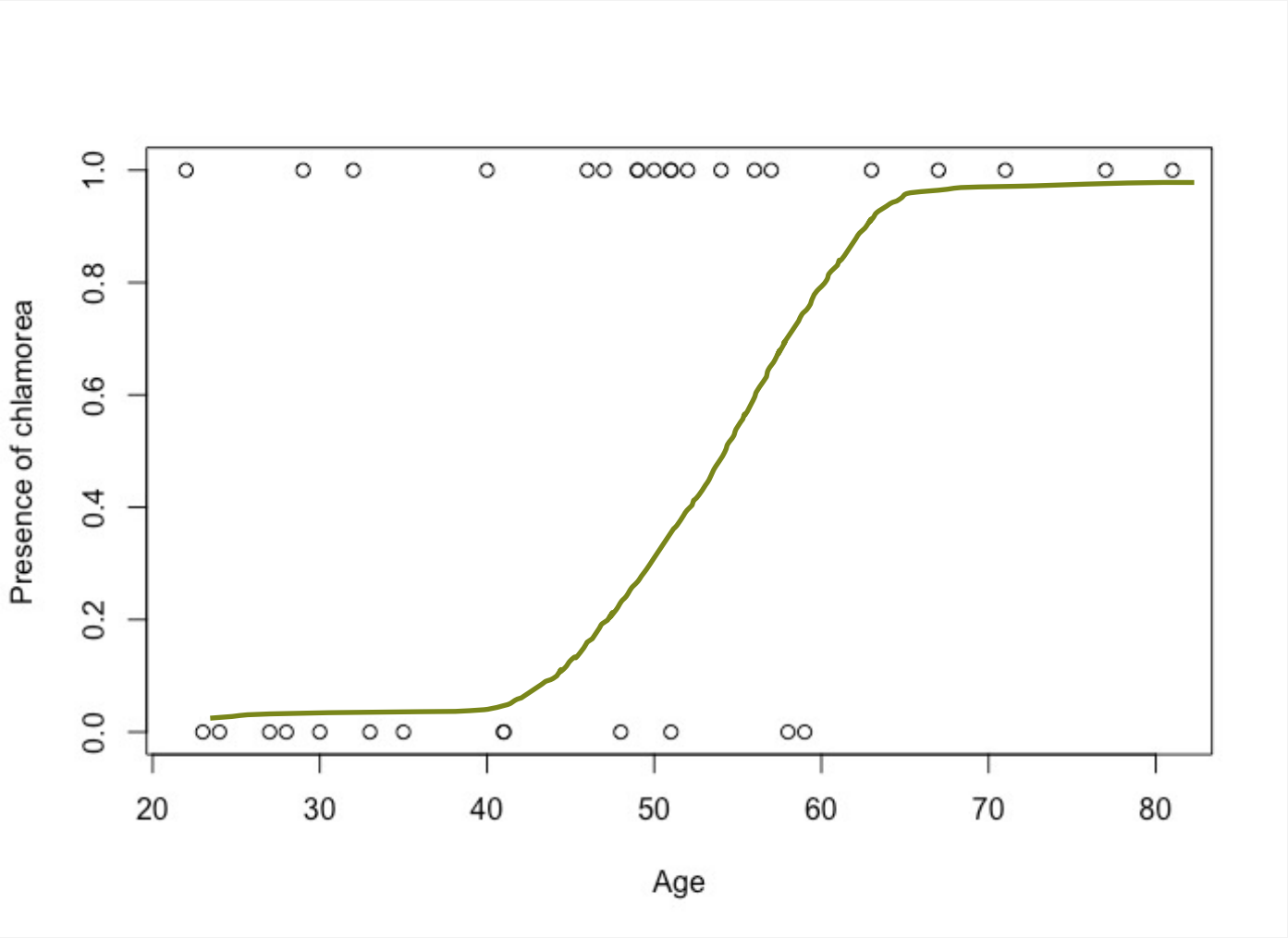
LOGISTIC REGRESSION MODEL



LOGISTIC REGRESSION MODEL

$$Prob(chlamorea = 1) = \frac{exp^{\beta_0+\beta_1 age_i}}{1 + exp^{\beta_0+\beta_1 age_i}}$$

$$ln\left(\frac{Prob(chlamorea_i = 1}{1 - Prob(chlamorea_i = 1)}\right) = \beta_0 + \beta_1 age_i$$



OVERVIEW OF MIXED MODELS

1. Linear regression model
2. Logistic regression model
3. **Generalised linear model (GLM)**
4. Mixed models

GENERALISED LINEAR MODEL (GLM)

- As its name suggests, the **generalised linear model, GLM** (McCullagh and Nelder, 1989), generalises the linear regression model for non-normal dependent variables, such as for example count variables (Poisson distribution) or dichotomous variables (binomial distribution), thus allowing us to use the same type of modelling, specification, estimation and diagnosis.
- The generalised linear model is an extension of the linear model when the observations are independent, but the assumptions about the normality of the disturbances are not met.

GENERALISED LINEAR MODEL (GLM)

- The distribution of the disturbances can be chosen inside the set of distributions that belong to the family of **exponential distributions** of a parameter, which includes the Binomial, Poisson, Negative Binomial, Gamma, Beta, and Inverse Gaussian distributions, and within which the Normal Distribution is a special case.

GENERALISED LINEAR MODEL (GLM)

- The aim of the GLM is to describe the dependence of the response variable (or the dependent variable) y, with respect to the explanatory variables, $E(y \mid x) = \mu$.
- The GLM have two functions: the 'link' function, $g()$, and the 'variance' function, $v()$. There is also sometimes a scale parameter, ϕ.

GENERALISED LINEAR MODEL (GLM)

- Thus, for example, when the **dependent variable is continuous**, the GLM is specified as (equivalent to a linear regression):

$$\mu = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

$$Var(y \mid x) = \phi$$

The link is linear and the variance is constant.

GENERALISED LINEAR MODEL (GLM)

- When the **dependent variable is dichotomous**, the GLM (equivalent to a logistic regression), would be specified:

$$ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

$$Var(y|x) = \phi\mu(1-\mu)$$

In this case, the link is a logit. The parameter ϕ is called overdispersion (if ϕ > 1) or underdispersion (if ϕ < 1).

GENERALISED LINEAR MODEL (GLM)

- When the **dependent variable is discrete (or a count variable)**, the GLM (equivalent to a Poisson regression), would be specified:

$$ln(\mu) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

$$Var(y|x) = \phi\mu$$

The link is logarithmic. The parameter ϕ is called overdispersion (if $\phi > 1$) or underdispersion (if $\phi < 1$).

OVERVIEW OF MIXED MODELS

1. Linear regression model
2. Logistic regression model
3. Generalised linear model (GLM)
4. **Mixed models**

MIXED MODELS

Mixed designs are characterised by simultaneously considering one or more dimensions of analysis.

Mixed designs include:

- **Multilevel designs** or multiple level designs (also called hierarchical)
- **Longitudinal designs** or repeated measures designs.

MIXED MODELS – MULTILEVEL DESIGNS

- They have a hierarchical structure, with data groupings in groups or clusters.
- In the literature, these hierarchical levels are called level 1, level 2, etc.; stage 1, stage 2, etc.; individual level and population level; individual level and cluster level.
- For example, when we have countries and regions, we have two hierarchical levels.

MIXED MODELS – MULTILEVEL DESIGNS

Example: Patients in health centres

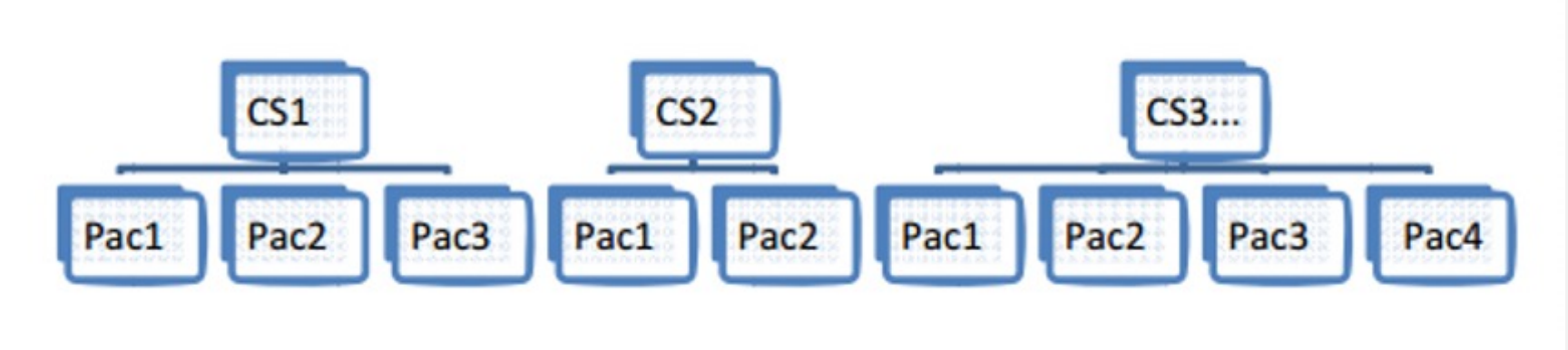


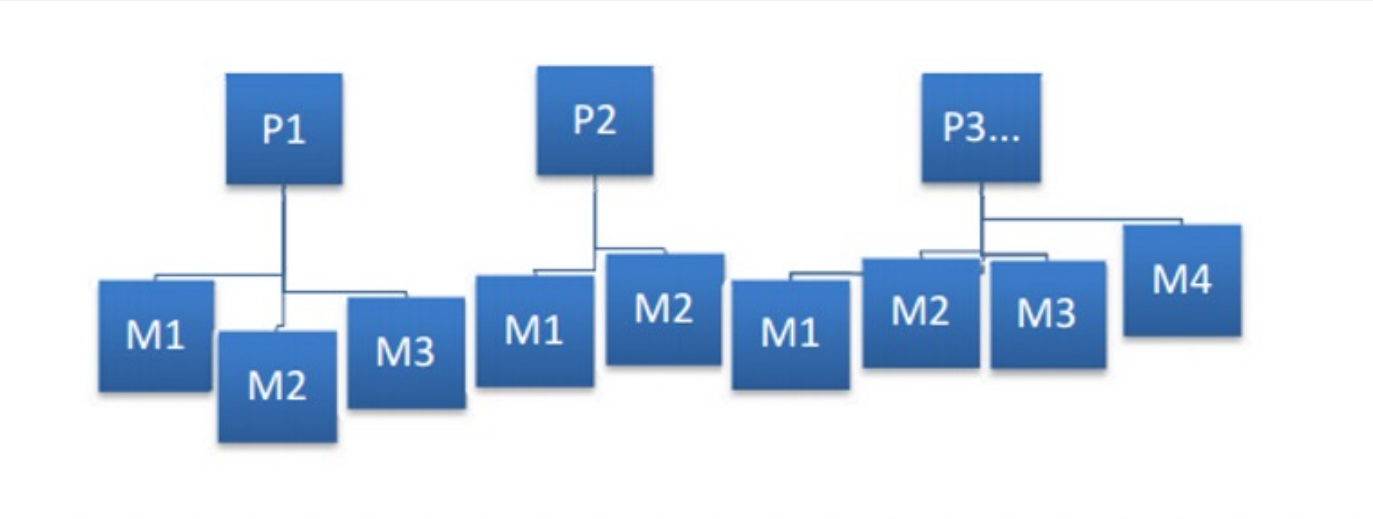
Diagram of 2 nested levels

MIXED MODELS – LONGITUDINAL DESIGNS

- The most peculiar characteristic of **longitudinal studies** is the repeated measurement over time of each individual or object of study.
- In econometrics, the most frequently used longitudinal designs are **panel data models,** also known as **data panels**.
- An **economic panel data model** is one that includes a sample of economic agents or agents of interest for a determined period, combining both types of data (temporal and structural dimensions).

MIXED MODELS – LONGITUDINALS DESIGNS

Example: Repeated measures, panel data



Categorisation and diagram of the units of a repeated measures design, nested measures in patients

MIXED MODELS – LONGITUDINAL DESIGNS

Types of panel

- Panels with a large number of individual units (transversal observations), are called **micropanels** (for example, many companies).
- Contrarily, if they have a large number of temporal observations they are called **macropanels** (for example, the evolution of a variable over time).
- When both situations occur, both a large cross-sectional dimension and a large temporal dimension, it is called a **random field**.

MIXED MODELS – LONGITUDINAL DESIGNS

Unlike the longitudinal designs used in other disciplines, in economics the data are usually observed in regular intervals.

The design of **panel data** can be:

- **Balanced**, when the observations are repeated the same number of times in all the individual units.
- **Unbalanced**, when there is at least one individual unit in which the observations are not repeated the same number of times as the rest.

MIXED MODELS – LONGITUDINAL DESIGNS

General advantages of longitudinal studies

Panel data designs compared with con other unidimensional designs:

- **They have a higher number of observations** and, therefore, of degrees of freedom. This means greater efficiency (or precision) in the estimates.
- **They enable more variability to be captured in the variables.** This will imply a lower probability of committing a type II error. In short, they have a higher statistical power.

MIXED MODELS – LONGITUDINAL DESIGNS

Specific advantages of longitudinal studies

Panel data designs compared with temporal series designs:

- **They make use of the transversal variability**. If the variables included do not present a large temporal variability but they do have a wide transversal variability, the panel data design will allow for more precise (efficient) estimates of the parameters.
- Transversal designs **make use of the temporal variability**. Some variables can present greater temporal variability but not transversal variability. Therefore, their effect can only be captured by means of a temporal dimension.

MIXED MODELS – LONGITUDINAL DESIGNS

Specific advantages of longitudinal studies

Last, and probably most importantly, the **panel data designs** enable **non-observable factors** for which we do not have any information to be included. These factors, called **heterogeneity**, can be of two types:

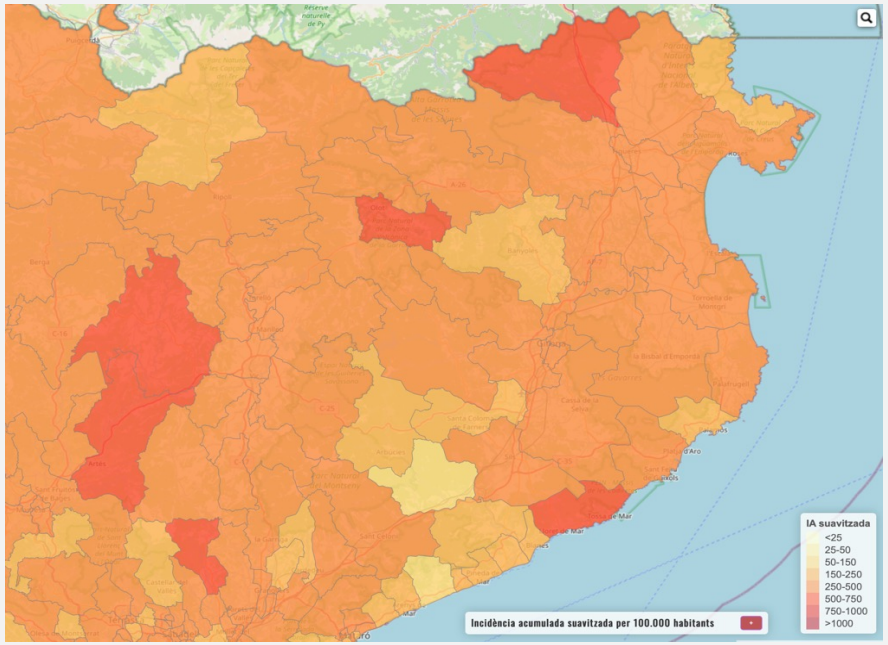
- **Individual heterogeneity**. Specific to each individual or unit of study and constant over time. Corresponds to the transversal dimension.
- **Temporal heterogeneity**. Common to all individuals or study units and time-varying, but constant for cross-sectional observations. Corresponds to the temporal dimension.

PRACTICAL EXAMPLE

In order to better understand the two types of heterogeneity, we will use, as an example, information on the COVID-19 pandemic.

First, we will show, the representation of the accumulated incidence per 100,000 inhabitants in the last fortnight on a map of municipalities with a focus on the Girona Health Region (which practically coincides with the province of Girona except for La Cerdanya, which belongs to the Alt Pirineu and Aran Health Region) in the period corresponding to the third wave (December 2020-February 2021).

PRACTICAL EXAMPLE



Source: COVIDCAT (https://ubidi.shinyapps.io/covidcat/)

We observe than during the third wave most of the municipalities in the Region had **an incidence of between 250 and 500 cases per 100,000 inhabitantes** in the last 14 days, but we also observe **some with between 150 and 250 cases** (Cadaqués, Llançà, Palamós, Santa Coloma de Farners, Arbúcies, Sant Hilari de Sacalm, Banyoles, Ribes de Fresser and Campdevànol) and **a few with between 500 and 750 cases** (Lloret de Mar, Tossa de Mar, Olot and a cluster of municipalities around Agullana in Alt Empordà).

PRACTICAL EXAMPLE

How can we explain, for example, that Banyoles, Besalú and Olot, being neighboring municipalities, have such different incidences? Banyoles is in the range 150-250, Besalú in the range 250-500, and Olot in the range 500-750.

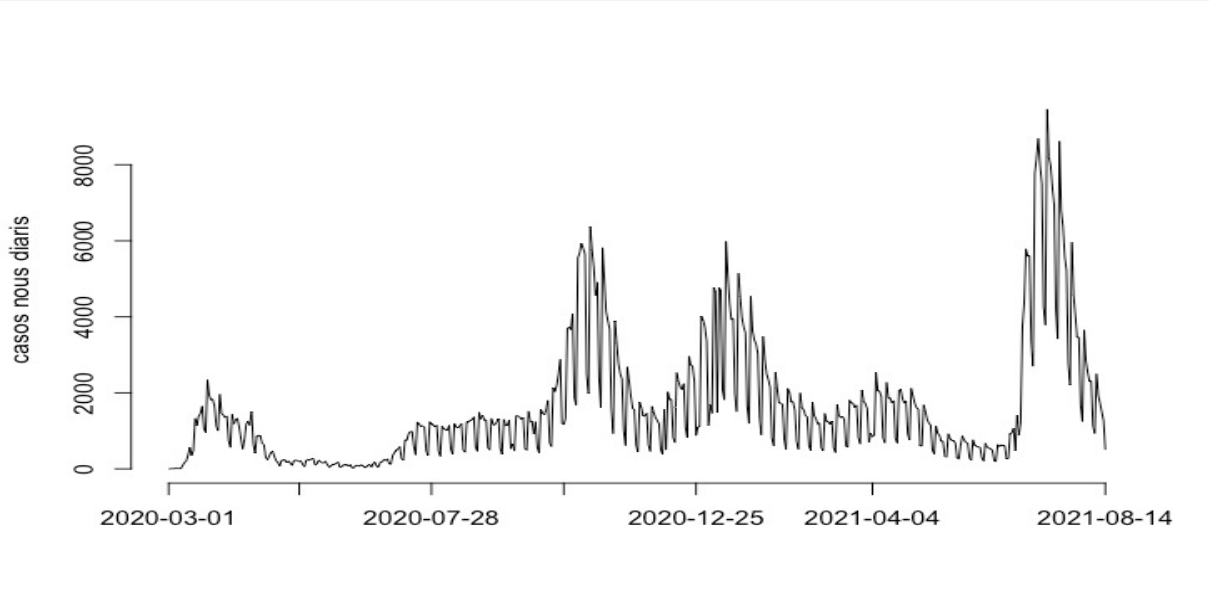
PRACTICAL EXAMPLE

How can we explain, for example, that Banyoles, Besalú and Olot, being neighboring municipalities, have such different incidences? Banyoles is in the range 150-250, Besalú in the range 250-500, and Olot in the range 500-750.

These differences and, in general, the non-homogeneous distribution of the incidence in the territory, could be due to the existence of non-observed variables specific to each cross-sectional unit (municipality in this case) and which may not vary over time. This is what is called **individual heterogeneity**.

PRACTICAL EXAMPLE

We will now show the temporal evolution of the new daily cases in Catalonia between 1 March 2020 and 14 August 2021.



Data source: Register of COVID-19 cases carried out in Catalonia. Segregation by sex and área Basic Health Area (ABS) (https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/xuwf-dxjd)

PRACTICAL EXAMPLE

The five waves are clearly observable. As can be seen, both the magnitude (height of the peaks) and the duration vary among the different waves. Why is this?

PRACTICAL EXAMPLE

The five waves are clearly observable. As can be seen, both the magnitude (height of the peaks) and the duration vary among the different waves. Why is this?

These differences could be due to non-observed variables, specific to each time unit (the waves, in this case) and, presumably, common to all cross-sectional units (Basic Health Areas or ABS in this case). These differences are known as **temporal heterogeneity**.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or longitudinal mixed model)

Starting with a linear (or normal) response, a panel model can be specified in the following way:

$$y_{it} = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3it} + \cdots + \beta_k x_{kit} + u_{it}$$

where the subindex $i = 1, \ldots, N$ denotes the individual units (N individual units) and the subindex$t = 1, \ldots, t_i$ ($T = max(t_i)$, number of periods) denotes the different observations repeated for each individual unit.

For example, i=each patient and t=three observations of the patients in time.

MIXED MODELS – LONGITUDINAL DESIGNS

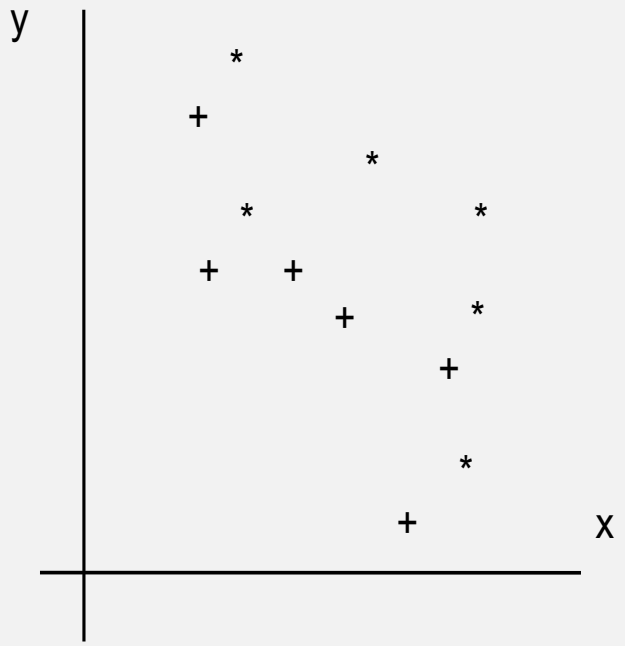
Specification of the panel data model (or longitudinal mixed model)

Example: In a simplified way, let's supose that we want to estimate the effect of age (x) on the variation in reading comprehension (y) of six children aged under 6 years of age, $i = 1, 2, 3, 4, 5, 6$ $(N = 6)$. The children are observed in two moments in time, two periods $t = 1\,(+), 2(*)$ $(T = 2)$ (balanced model). We say that the model is balanced because measures of reading ability are available for all children and for all periods. In this case, the explanatory variable is time-dependent, since the age of the children is different in the two moments of time in which they are observed..

MIXED MODELS – LONGITUDINAL DESIGNS

Example: In this case, the specification of our model will be:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$



MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model)

Depending on the proposed objective, panel data can be estimated marginally or conditionally.

MIXED MODELS – LONGITUDINAL DESIGNS

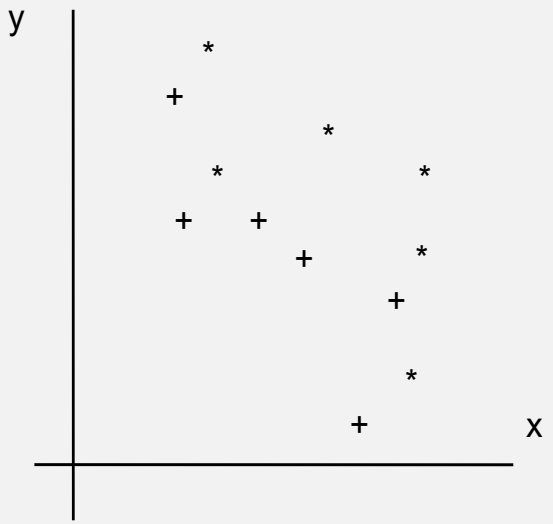
Specification of the panel data model (or the mixed longitudinal model). Marginal approach

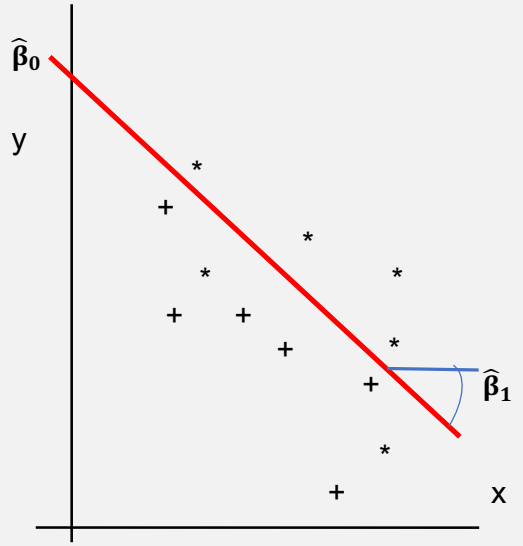
We use **the marginal approach** when we want to make **"population" inferences**, that is, if we want to explain the relationship between the dependent variable and the explanatory variables independently of intraindividual variability.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Marginal approach

In our example:





MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Marginal approach

The ordinate at the origin $\hat{\beta}_0$ and the coefficient associated with the explanatory variable $\hat{\beta}_1$ are common to all individuals. There is no individual heterogeneity. In other words, all the effects (of the explanatory variables, including the ordinate at the origin) are fixed.

We observe that, on average, the older the child, the less variation in reading comprehension.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Marginal approach

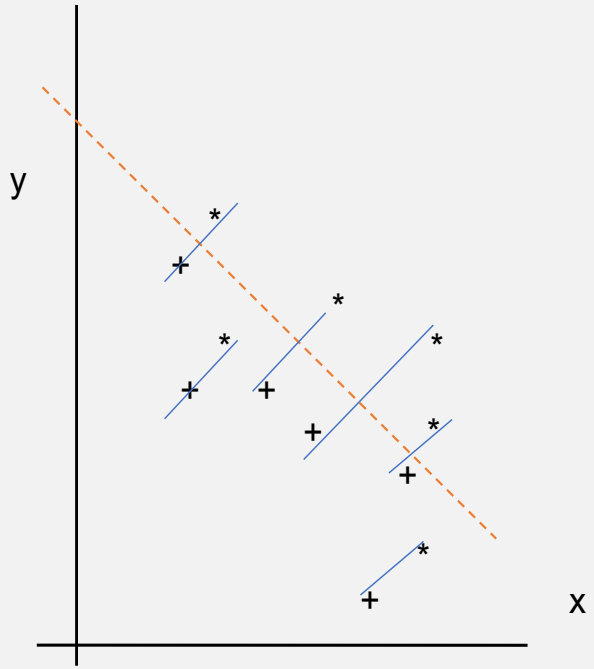
In the **marginal approach**, the aim is to estimate the parameters corresponding to the mean, β. Very rarely are the covariance parameters of interest. In fact, they are usually treated as a nuisance in the marginal approach, controlled, but not estimated.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Conditional approach

In the conditional approach, we want to make “individual” inferences.

In our example:



MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Conditional approach

In this case, the coefficient associated with the explanatory variable $\hat{\beta}_1$ is approximately equal for all individuals. But, the ordinate at the origin $\hat{\beta}_0$ is different for each of them. There is individual heterogeneity.

Regardless of the average behavior, for each individual child, the older the age, the more variation in reading comprehension.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model). Conditional approach

In the **conditional approach**, the mean of the dependent variable (inter-individual variability) and the structure of covariances or correlations (intra-individual variability) are modeled simultaneously. In this approach, the parameters defining the correlation have the same interest as those corresponding to the mean, sometimes even more.

MIXED MODELS – LONGITUDINAL DESIGNS

Specification of the panel data model (or the mixed longitudinal model)

Note that in the marginal approach, the average (or population inference) is of interest.

In the conditional approach, each child is of interest (or individual inference).

MIXED MODELS – LONGITUDINAL DESIGNS

Let's suppose that we specify the following models:

$$y_{it} = \beta x_{it} + a_i + \tau_t + u_{it}$$

where a_i captures **the non-observable individual heterogeneity** (factors specific to each individual unit, which do not vary in time); and τ_t captures **the temporal heterogeneity**, common to all the individual units.

MIXED MODELS – LONGITUDINAL DESIGNS

Pooled model

The simplest specification of the model, called the **pooled model**, consists in ignoring the heterogenity ($a_i = a$ and $\tau_t = 0$):

$$y_{it} = a + \beta x_{it} + u_{it}$$

and estimating the model by Ordinary Least Squares (OLS). This method of estimation is known as **Pooled Least Squares** (PLS).

MIXED MODELS – LONGITUDINAL DESIGNS

Pooled model. Problems of ignoring the heterogeneity in a mixed design

That is, even when we have a mixed design, we omit relevant variables (heterogeneity).

This means that a specification error is made. Therefore, **the PLS estimation of the parameters will be biased and their variances will be miscalculated**.

Possible solution:

Estimate the model using alternative estimation methods either by the marginal approximation or by the conditional approximation. We will use the conditional approach.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

The **random effects model** is the most well-known among the conditional models. It assumes that the dependence that may exist between the observations of the dependent variable (in other words, between the repeated responses of the same individual) is due to the fact that the regression coefficients (of the mean) are not the same for all individuals. In fact, the simplest interpretation is that which assumes that individual heterogeneity is due to non-observable factors (or omitted variables) fixed over time, but variable between individuals. These factors are what cause the dependence between the different observations of the dependent variable.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Therefore, this model **assumes that the coefficients of the regression are not the same for all the individuals**:

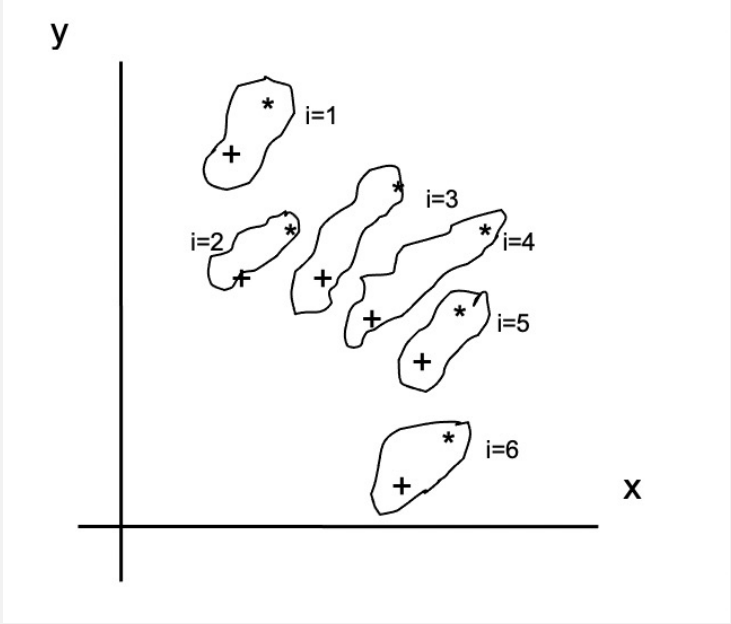
$$y_{it} = \beta_{it} x_{it} + a_i + \tau_t + u_{it}$$

They can vary among individuals

MIXED MODELS – LONGITUDINAL DESIGNS

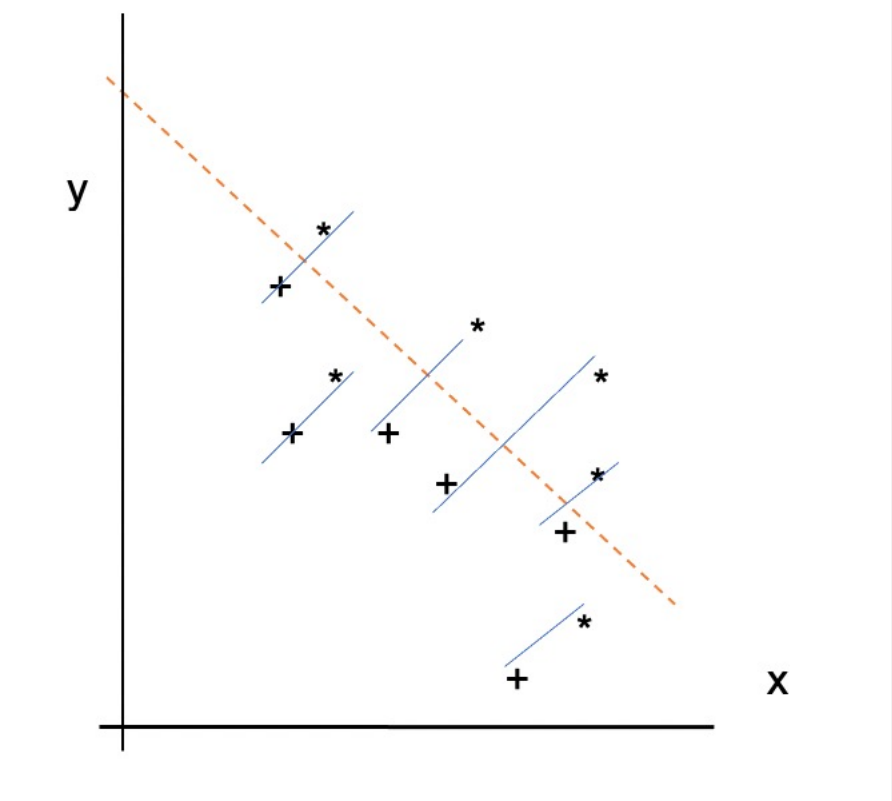
Conditional approach. Random effects model

Let's assume the example we have seen before, in which we consider 6 children and two periods (we represent the first period by the symbol + and the second by the symbol *):



MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model



In this case, the coefficient associated with the explanatory variable is approximately the same for all children. That is, reading comprehension varies approximately the same way for all the children.

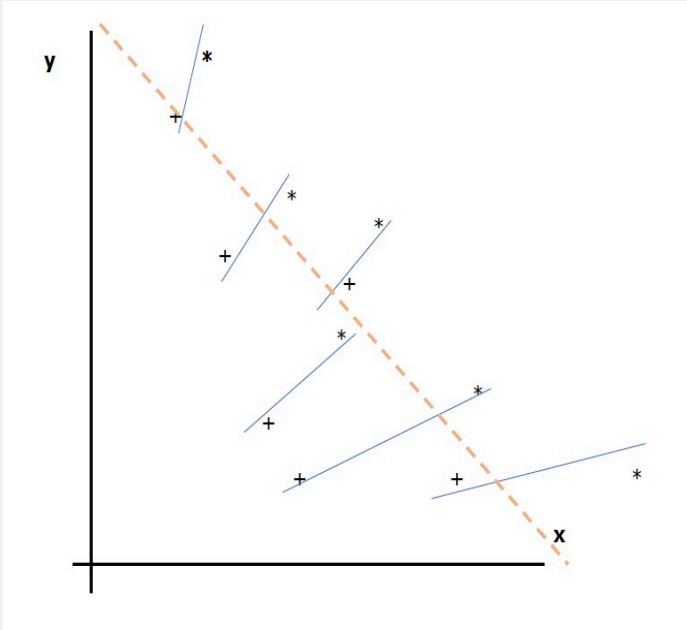
On the contrary, the ordinate at the origin is different for each one of them. The basal reading comprehension is different.

There is individual heterogeneity!!!

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

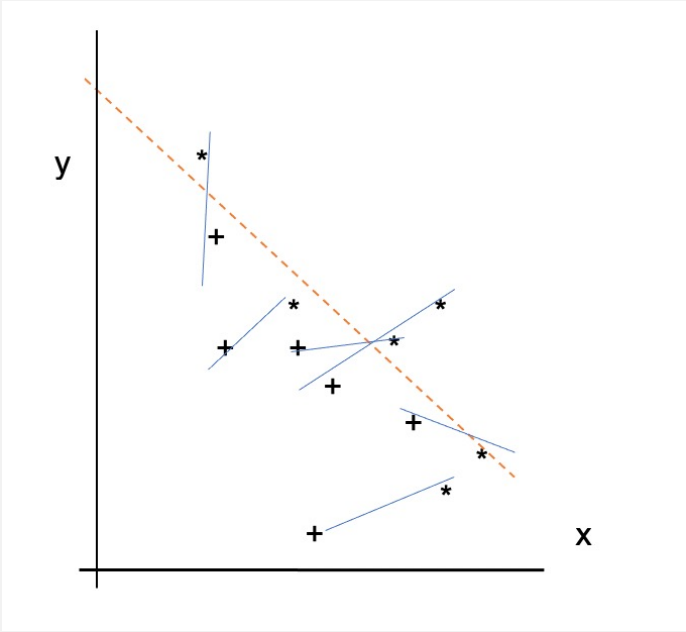
In the following example, the parameter associated with the explanatory variable, and not the one associated with the ordinate at the origin, is the one that is the random effect.



MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Both situations could also occur. That is, both the ordinate at the origin and the slope are random effects.



MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

In econometrics, when we talk about random effects models, we refer precisely to the first model. That is, there is only a random effect corresponding to the ordinate at the origin (i.e., non-observed individual heterogeneity).

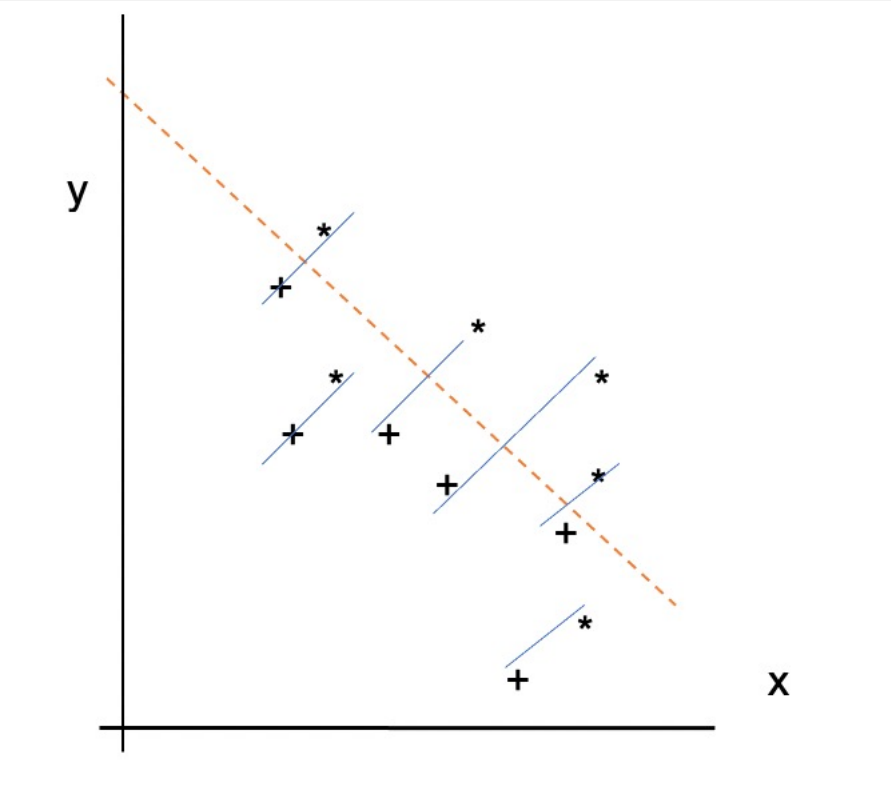
$$y_{it} = \beta_{0i} + \beta_1 (x_{it} - x_{it-1}) + \beta_p x_{it} + u_{it}$$

- On the other hand, in the second model, we would be talking about a random coeficient model. In this case, the random effect would be associated with an explanatory variable.

$$y_{it} = \beta_0 + \beta_{1i} (x_{it} - x_{it-1}) + \beta_p x_{it} + u_{it}$$

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model



MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

In summary and continuing with our example, in which the random effect is in the ordinate at the origin but not in the slope:

$$
\begin{aligned}
&Per\ a\ \ i = 1 && y_{1t} = \beta_{01} + \beta_1 x_{1t} + u_{1t} \\
&Per\ a\ \ i = 2 && y_{2t} = \beta_{02} + \beta_1 x_{2t} + u_{2t} \\
&Per\ a\ \ i = 3 && y_{3t} = \beta_{03} + \beta_1 x_{3t} + u_{3t} \\
&Per\ a\ \ i = 4 && y_{4t} = \beta_{04} + \beta_1 x_{4t} + u_{4t} \\
&Per\ a\ \ i = 5 && y_{5t} = \beta_{05} + \beta_1 x_{5t} + u_{5t} \\
&Per\ a\ \ i = 6 && y_{6t} = \beta_{06} + \beta_1 x_{6t} + u_{6t}
\end{aligned}
$$

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

The problem lies in the fact that each individual usually has too few observations to estimate the β based only on (y_{it}, x_{it}). For example, in our case we have 7 unknown parameters $(6\ \beta_{oi}, i = 1, \dots, 6\ and\ 1\ \beta_1)$ with 12 observations (6 children x 2 periods). **INEFFICIENT ESTIMATES!!!**

We must estimate the model using a **generalised linear mixed model (GLMM)**.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

We should assume that the β are nothing more than independent realizations of some probability distribution (the normal one in linear models) with mean β:

$$y_{it} = \beta_{0i} + \beta_1 x_{it} + u_{it}$$
$$\beta_{0i} = \beta_0 + \eta_i$$

Summarizing, we have specified a mixed model with a fixed effect (β_1) and a random effect (β_{0i}), meaning that β_{0i} can vary among individuals (i = 1, 2, ..., 6).

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Note that, in the random effects model, within the conditional approach, we will also estimate the population parameters, β_0 and β_1, corresponding to the marginal approach.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Now, in addition to assuming:

$$E(u_{it}) = 0$$
$$E(u_{it}^2) = \sigma_u^2 \; constante$$
$$E(u_{it} \times u_{js}) = 0 \; \forall \; i \neq j \; i \; \forall \; t \neq s$$

We must also assume that:

$$E(\eta_i) = 0$$
$$E(\eta_i^2) = \sigma_\eta^2 \; constante$$
$$E(\eta_i, \eta_j) = 0 \; \forall \; i \neq j$$
$$E(u_{it}, \eta_j) = 0 \; \forall \; i, t, j$$

The last two hypotheses are particularly important.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

- First, $\boldsymbol{E(\eta_i, \eta_j) = 0 \; \forall \; i \neq j}$, means that the higher levels must be independent among themselves. Failure to comply with this hypothesis (called crossover effects) requires the use of much more complex estimation methods than those explained here (for example, Bayesian methods).

- Second, $\boldsymbol{E(u_{it}, \eta_j) = 0 \; \forall \; i, t, j}$, requires that the random effects be independent of the fixed effects. Otherwise, the effects must be effectively completely random.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

We will now illustrate the last two hypotheses in our example.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

First, $\boldsymbol{E(\eta_i, \eta_j) = 0 \; \forall \; i \neq j}$, means that the higher levels must be independent among themselves

Continuing with our example, children are the upper level, while time constitutes the lower level, since each child "contains" (in fact, is observed) two time periods. Therefore, the variation in a child's reading comprehension does not depend on the variation in reading comprehension of any other child. That is, the children are independent among themselves, so this hypothesis is satisfied.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

However, it could happen that the children belong to the same family (for example, let us suppose that children 2 and 5 are siblings). In this case, the variation in reading comprehension of child 2 and child 5 would not be independent, and the hypothesis would not be fulfilled.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

But, for the hypothesis to be met, it is enough to re-specify the model by adding an additional level:

$$y_{kit} = \beta_{ok} + \beta_1 x_{kit} + u_{kit}$$
$$\beta_{ok} = \beta_0 + \eta_k$$

where k $(k = 1\ children\ 2\ and\ 5; k = 2\ children\ 3; k = 3\ children\ 4; k = 4\ children\ 6)$ denotes the family, i $(i = 1, 2, \dots, 6)$ denotes the child and t $(t = +,*)$ denotes the period. Now, the higher level units (families) are indeed independent, thus fulfilling the hypothesis.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Second, $\boldsymbol{E(u_{it}, \eta_j) = 0 \; \forall \; i, t, j}$, requires that the random effects be independent of the fixed effects.

In our example, suppose that in the model we are using, the differences in the ordinates at the origin between children are explained by some variable, for example, the sex of the child.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

That is:

$$y_{it} = \beta_{oi} + \beta_1 x_{it} + u_{it}$$
$$\beta_{oi} = \beta_0 + \gamma sexe_i + \eta_i$$

Thus, isolating:

$$y_{it} = \beta_o + \gamma sexe_i + \eta_i + \beta_1 x_{it} + u_{it}$$
$$y_{it} = \beta_o + \beta_1 x_{it} + u_{it} + \gamma sexe_i + \eta_i$$

And, finally, rearranging terms:

$$y_{it} = \beta_o + \beta_1 x_{it} + \nu_{it}$$

where $\nu_{it} = u_{it} + \gamma sexe_i + \eta_i$.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model

Evidently, v_{it} will be co-related with u_{it}, not supporting the hypothesis that **the random effects are independent of the fixed effects,** and the estimators will be inconsistent.

So that the hypothesis is once again met, all the explanatory variables must be included in the main equation (the mean) and the effect left completely random. That is to say:

$$y_{it} = \beta_{oi} + \beta_1 x_{it} + \gamma sexe_i + u_{it}$$
$$\beta_{oi} = \beta_0 + \eta_i$$

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model. Model estimation

If the two hypotheses are satisfied, the models **are estimated by methods based on maximum likelihood**. The problem is that a complete likelihood function only exists when there is no correlation, whether or not there are random effects. But, quasi-likelihood can be defined, which has exactly the same form as the partial derivative of the complete likelihood.

MIXED MODELS – LONGITUDINAL DESIGNS

Conditional approach. Random effects model. Model estimation

Quasi-likelihood maximization is very complex. On the other hand, the Bayesian methods, to which we can turn, involve the full evaluation of the quasi-likelihood, explaining why they tend to be very intense computationally.

Alternatively, pseudo-likelihood or penalized pseudo-likelihood, PQL, restricted likelihood (REML) or Laplace's method can be used.

MIXED MODELS – LONGITUDINAL DESIGNS

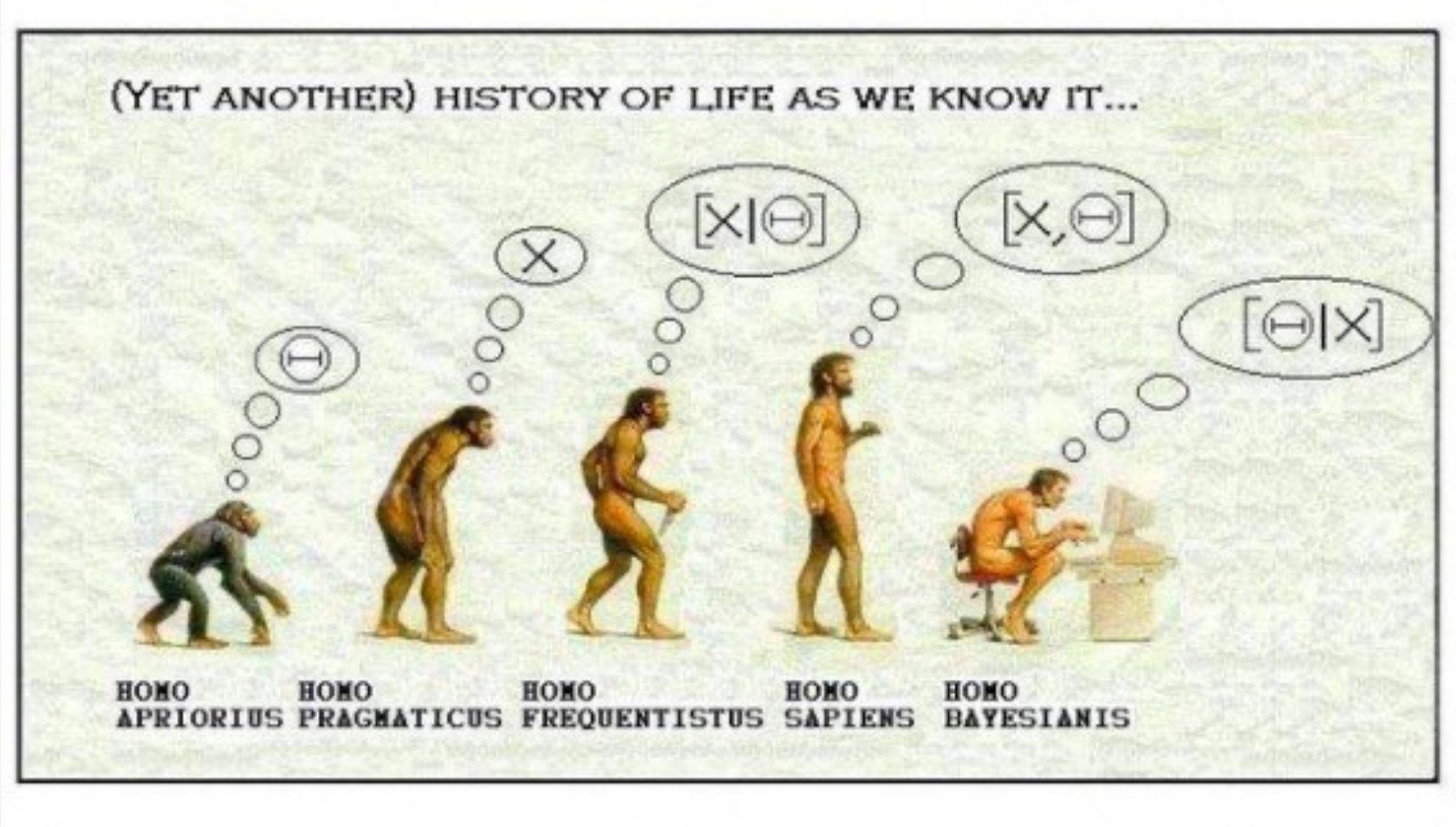
Resumiendo:

- The fixed effects estimator (marginal approximation) allows the model to be estimated under less restrictive assumptions than the random effects estimator.
- Nonetheless, if the regularity conditions are met, the random effects estimator is more efficient than the fixed effects estimator. But it may be inconsistent if there is correlation between the explanatory variables and individual heterogeneity.
- By Bayesian methods, random effects estimators are always consistent.

MIXED MODELS – LONGITUDINAL DESIGNS

Remember that the regularity conditions referred to in the second point are as follows:

$$E(u_{it}) = 0$$
$$Var(u_{it}) = E(u_{it}^2) = \sigma_u^2 \; constante$$
$$Cov(u_{it}, u_{js}) = E(u_{it} \times u_{js}) = 0 \; \forall \; i \neq j \; y \; \forall \; t \neq s$$


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
[X|Θ]
[X,Θ]
X
Θ
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS