UNIVERSITY OF THE AEGEAN

DEPARTMENT OF MATHEMATICS

DIVISION OF STATISTICS AND ACTUARIAL FINANCIAL MATHEMATICS

STATISTICS AND DATA ANALYSIS

# GENERALIZED LINEAR MIXED MODELS WITH APPLICATIONS

MASTER THESIS

EMMANOUIL
NEKTARIOS
KALLIGERIS

SAMOS 2018

UNIVERSITY OF THE AEGEAN
DEPARTMENT OF MATHEMATICS
DIVISION OF STATISTICS AND ACTUARIAL FINANCIAL MATHEMATICS
STATISTICS AND DATA ANALYSIS

# Generalized Linear Mixed Models with Applications

Master Thesis

Emmanouil Nektarios
Kalligeris

Samos 2018

UNIVERSITY OF THE AEGEAN
DEPARTMENT OF MATHEMATICS
DIVISION OF STATISTICS AND ACTUARIAL FINANCIAL MATHEMATICS
STATISTICS AND DATA ANALYSIS

# Generalized Linear Mixed Models with Applications

Master Thesis
Emmanouil Nektarios
Kalligeris
20 February, Samos

**EVALUATION COMMITTEE**

Hatjispyros Spyridon
Assistant Professor UAegean

Karagrigoriou Alexandros
Professor UAegean (Supervisor)

Vasdekis Vasilios
Professor AUEB

**To my supervisor and to my family.**

# Contents

# Acknowledgements

I would like first to thank from the bottom of my heart, my supervisor Professor Alexandros Karagrigoriou of the University of the Aegean for his patience, motivation and support. His guidance not only through the writing of my thesis but also through my Masters' program and my life in general, was more than valuable to me and I am gratefully indebted to him.

I would also like to thank Dr. Christina Parpoula of the University of the Aegean, for providing me very valuable comments and her overall help was substantial to the completion of this thesis.

Moreover, I would like to thank the two readers of this thesis Assistant Professor Spyridon Hatjispyros of the University of the Aegean and Professor Vasilios Vasdekis of Athens University of Economics and Business.

In addition, I would like to thank the Hellenic Center for Disease Control and Prevention and the Hellenic National Meteorological Service who provided the needed data for the application presented in the last chapter of this thesis.

I must also thank Paschalini Trentou for her patience, understanding and support through my Masters' program.

My two very close friends Athanasios Spanoudis and Kimon Ntotsis must also be acknowledged for the support that they showed to me in my Master and in my life.

Finally, I would like to thank my family and especially my mother Aristea who I will always love no matter...

# Abstract

This thesis is conducted at the Department of Mathematics, Division of Statistics and Actuarial-Financial Mathematics of the University of the Aegean, in the context of the MSc program in *Statistics and Data Analysis*. Its purpose is to analyze the class of Generalized Linear Mixed Models (GLMMs) and their implementation in real life problems, through a thorough study on influenza-like illness (ILI) rate data. More specifically, we focus on a special class of GLMMs, the class of periodic regression mixed models for modeling the ILI time series data. For the trend, linear, quadratic, cubic and quartic trends are considered while for the seasonal component, the most widely used periodicities are implemented, i.e. 12, 6, and 3 months. The class extends further to include first and second order AR and MA parts while environmental covariates potentially affecting the output are also included.

The structure of the thesis consists of four Chapters. In Chapter 1, Generalized Linear (GLMs) and Generalized Linear Mixed Models are presented along with their properties. Some of the topics that will be discussed are, logistic regression model, maximum likelihood estimation and, test of hypotheses.

Chapter 2, constitutes an introduction to univariate time series analysis. Important terms such as autocorrelation and white noise are defined, as well as the backward shift operator. At the end, various time series models (i.e., AR, (S)ARIMA, and, Periodic) are presented.

Chapter 3, introduces four basic model selection criteria. Among them are, the Modified Divergence Information Criterion ($MDIC$) and $R^2_{GLMM(m)}$, which are being used for the selection of the "best overall" model of the application that follows.

Finally, in Chapter 4, an experimental study is applied on influenza-like illness (ILI) rate data, collected weekly through the sentinel surveillance system, provided by the Department of Epidemiological Surveillance and Intervention of the Hellenic Center for Disease Control and Prevention (H.C.D.C.P.). An exhaustive search process takes place, in order to provide guidelines for the selection of the optimal periodic regression mixed model for early and accurate outbreak detection in an epidemiological surveillance system, as well as for its proper use and implementation.

# Αβστραστ

Η παρούσα διπλωματική εκπονήθηκε στο Τμήμα Μαθηματικών, Κατεύθυνση Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου, στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών *Στατιστική και Ανάλυση Δεδομένων*. Σκοπός της, είναι η ανάλυση της κλάσης των Γενικευμένων Γραμμικών Μικτών Μοντέλων καθώς και η εφαρμογή τους σε προβλήματα της καθημερινότητας, μέσω μιας πλήρους μελέτης πάνω σε δεδομένα γρίπης (ILI). Συγκεκριμένα, επικεντρωνόμαστε σε μια ειδική κλάση των Γενικευμένων Γραμμικών Μικτών Μοντέλων, την κλάση των περιοδικών μικτών μοντέλων παλινδρόμησης για την μοντελοποίηση των ILI χρονολογικών δεδομένων. Για την τάση θεωρούνται οι γραμμικού, τετραγωνικού, κυβικού και η τετάρτου βαθμού τάσεις, ενώ για τον εποχικό παράγοντα εφαρμόζονται οι πιο ευρέως χρησιμοποιούμενες περιοδικότητες (12, 6 και 3 μηνών). Η κλάση επεκτείνεται περαιτέρω ώστε να συμπεριλάβει πρώτου και δευτέρου βαθμού AR και MA όρους ενώ περιλαμβάνονται και περιβαλλοντικές συμμεταβλητές που πιθανόν να επηρεάζουν το αποτέλεσμα.

Η δομή της δηπλωματικής αποτελείται από τέσσερα Κεφάλαια. Στο Κεφάλαιο 1, παρουσιάζονται τα Γενικευμένα Γραμμικά και τα Γενικευμένα Γραμμικά Μικτά Μοντέλα μαζί με τις ιδιότητές τους. Κάποια από τα θέματα που θα αναλυθούν είναι, το λογιστικό μοντέλο παλινδρόμησης, η εκτίμηση μέγιστης πιθανοφάνειας και ο έλεγχος υποθέσεων.

Στο Κεφάλαιο 2, γίνεται εισαγωγή στην μονομεταβλητή ανάλυση χρονοσειρών. Ορίζονται σημαντικές έννοιες, όπως αυτή της αυτοσυσχέτησης, του λευκού θορύβου όπως και του προς τα πίσω τελεστή. Τέλος, παρουσιάζονται διάφορα μοντέλα χρονοσειρών όπως τα AR, τα (S)ARIMA και τα Περιοδικά.

Στο Κεφάλαιο 3, γίνεται η παρουσίαση τεσσάρων βασικών κριτηρίων επιλογής μοντέλου. Μεταξύ αυτών είναι και τα Modified Divergence Information Criterion ($MDIC$) και $R^2_{GLMM(m)}$, τα οποία χρησιμοποιούνται στην εφαρμογή των δεδομένων γρίπης για την επιλογή του "καλύτερου" μοντέλου.

Τέλος, στο Κεφάλαιο 4, εφαρμόζεται μια πειραματική μελέτη σε δεδομένα γρίπης (ILI rates), που συλλέχθηκαν εβδομαδιαίως την περίοδο $2014-2016$ μέσω του συστήματος παρακολούθησης sentinel, τα οποία παραχωρήθηκαν από το Τμήμα Επιδημιολογικής Επιτήρησης και Παρέμβασης του Ελληνικού Κέντρου Ελέγχου και Πρόληψης Νοσημάτων (ΚΕ.ΕΛ.Π.ΝΟ). Πραγματοποιείται μια λεπτομερής έρευνα, με σκοπό τον προσδιορισμό του καταλληλότερου περιοδικού μικτού μοντέλου παλινδρόμησης για την έγκαιρη και έγκυρη ανίχνευση επιδημικών εξάρσεων σε ένα επιδημιολογικό σύστημα παρακολούθησης, καθώς και την κατάλληλη χρήση και εφαρμογή του.

CHAPTER 1

# Generalized Linear and Generalized Linear Mixed Models

In this chapter we will present and discuss the class of Generalized Linear Models (GLMs), as well as the extended class of Generalized Linear Mixed Models (GLMMs). GLMs is a covering algorithm allowing for the estimation of a number of otherwise distinct statistical regression models within a single and unified framework. First developed by John Nelder and R.W.M. Wedderburn in [75], the algorithm and the overall GLM methodology has proved to be of substantial value to statisticians in terms of the scope of models under its domain as well as the number of accompanying model statistics facilitating an analysis of fit. In the early days of statistical computing - from 1972 to 1990 - the GLM estimation algorithm also provided a substantial savings of computing memory and consequently, time compared to what was required using standard maximum likelihood techniques. Prior to Nelder and Wedderburn's efforts, GLM models were typically estimated using a Newton-Raphson type full maximum likelihood method, with the exception of the Gaussian model. Commonly known as normal or linear regression, the Gaussian model is usually estimated using a least squares algorithm. GLM, as we shall observe, is a generalization of ordinary least squares regression, employing a weighted least squares algorithm that iteratively solves for parameter estimates and standard errors.

In 1974, Nelder coordinated a project [76] to develop a specialized statistical application called GLIM, an acronym for Generalized Linear Interactive Modeling. Sponsored by the Royal Statistical Society and Rothamsted Experimental Station, which is one of the oldest agricultural research institutions in the world [86], GLIM provided the means for statisticians to easily estimate GLM models, as well as other more complicated models which could be constructed using the GLM framework. GLIM soon became one of the most used statistical applications worldwide, and was the first major statistical application to fully exploit the PC environment in 1981. However, it was discontinued in 1994. Presently, nearly all leading general purpose statistical packages offer GLM modeling capabilities; e.g., SAS, R, Stata, S-Plus, and SPSS [44].

An extension to the class of GLMs, as aforementioned, are GLMMs. GLMMs are being used in cases that we want to incorporate random factors in non-linear models. For example, suppose we wish to study factors affecting cost of hospitalization by taking a random sample of patient record of 15 teaching hospitals. The cost within a hospital must be regarded as correlated. They will be similar because of the general costs of running the hospital, billing practices, costs of nearby, and so on. In such cases, GLMMs encounter the complication of exploring the consequences of adding

random factors to non-linear models and thus they are considered useful.

## 1.1 Generalized Linear Models (GLMs)

Linear models have been used for the most part in situations where the observations are continuous. However, there are cases in practice where the observations are discrete or categorical. For example, consider the random variable $X = \{outcome\ of\ injured\ persons\}$, (see [54]). This is a discrete random variable since it can only take the values 0 (alive) or 1 (dead). McCullagh and Nelder in [66], proposed an extension of linear models, called Generalized Linear Models or GLMs. They noted that the key elements of a classical linear model, are $(i)$ the observations are independent, $(ii)$ the mean of the observations is a linear function of some covariates, and $(iii)$ the variance of the observation is a constant. The extension to GLM consists of modifications of $(ii)$ and $(iii)$ above; by $(ii)'$ the mean is associated with a linear function of some covariates through a link function; and $(iii)'$ the variance of the observation is a special function of the mean, that can be transformed and make the variance stable (econometrics) [47]. See [66] for details. In contrast to linear models, GLMs include a variety of models with Normal, Binomial, Poisson, and Multinomial as special cases. Thus, these models are applicable to cases where the observations may not be continuous.

### 1.1.1 Structure of a GLM

Constructing a generalized linear model involves three decisions:

**1.** The distribution of the data.

**2.** The function of the mean that will be modeled as linear in terms of the predictors (covariates).

**3.** The predictors.

#### 1.1.1.1 Distribution of $\boldsymbol{y}$

Let the vector $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ assumed to be consisted of independent measurements from a distribution with density belonging (or being similar to) the exponential family.

$$y_i \overset{indep.}{\sim} f_{Yi}(y_i), i = 1, ..., n \tag{1.1}$$

$$f_{Y_i}(y_i) = exp\left\{ \frac{y_i\gamma_i - b(\gamma_i)}{\tau^2 - c(y_i, \tau)} \right\}, \tag{1.2}$$

where the parameter $\gamma_i$ depends on the expected value of $y_i$, $\tau$ is a scale parameter, and $b$ and $c$ are arbitrary functions.
The quantity $f_{Y_i}(y_i)$ is written in what is called *canonical form.*

#### 1.1.1.2 Link function

Our intention, is to identify the connection between the parameters of the distribution and various predictors. This can be achieved by modeling a transformation of the mean, $\mu_i$, where $\mu_i$ will be some function of $\gamma_i$, as a linear model in the predictors

$$E[y_i] = \mu_i$$

$$g(\mu_i) = x_i'\beta, \tag{1.3}$$

where $g(\cdot)$ is a function known as *link function* (because it links together the mean of $y_i$ and the linear form of predictors), $\mathbf{x_i'}$ is the $\mathbf{i}^{th}$ row of the model matrix, and $\beta$ is the parameter vector in the linear predictor.

#### 1.1.1.3 Predictors

In practice, one must make decisions as to which predictors to be included in the right-hand side of (1.3) and in what functional form to be included them. For example, in [15] the suggested predictor of survival is log nicotine dose as opposed to nicotine itself.

It is important to notice that the modeling of the mean is the same for both GLMs and LMMs. For example, the similarity is obvious on issues related in the representation of predictors and interactions, on when and how to model non-linear relationships and in the incorporation of random factors (see 1.2).

#### 1.1.1.4 Example (Normal distribution - linear model)

We can write the normal distribution in the form (1.2) by defining

$$\gamma_i = \mu_i$$

$$b(\gamma_i) = \frac{1}{2}\mu_i^2$$

$$\tau^2 = \sigma^2 \tag{1.4}$$

$$c(y_i, \tau) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\frac{y_i^2}{\sigma^2}.$$

### 1.1.2 Maximum likelihood estimation

#### 1.1.2.1 Likelihood

By (1.2), the log likelihood is given by

$$l = \sum_{i=1}^{n} \left[ \frac{y_i \gamma_i - b(\gamma_i)}{\tau^2} \right] - \sum_{i=1}^{n} c(y_i, \tau). \tag{1.5}$$

#### 1.1.2.2 Useful identities

In this subsection it is useful to recall the following identities:

$$E[y_i] = \mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i}, \tag{1.6}$$

which, using (1.6) gives

$$var\left( \frac{y_i - \mu_i}{\tau^2} \right) = \frac{1}{\tau^2} \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2},$$

or

$$\begin{aligned} var(y_i) &= \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \\ &\equiv \tau^2 v(\mu_i), \end{aligned} \tag{1.7}$$

where $v(\mu_i) = \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$, is referred to as the *variance function* since it indicates the dependence between the variance and the mean of $y_i$.

The above results follow from the expression

$$E\left[ \frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i} \right] = 0, \tag{1.8}$$

and

$$var\left( \frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i} \right) = -E\left[ \frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \gamma_i^2} \right], \tag{1.9}$$

which require regularity conditions [22].
Using (1.2) and (1.8) we obtain

$$E\left[ \frac{\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\}}{\tau^2} \right] = 0. \tag{1.10}$$

Also, using (1.2) and (1.9) we obtain

$$var\left(\frac{\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\}}{\tau^2}\right) = -E\left[-\frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right] = 0. \tag{1.11}$$

The following two identities are also useful:

$$\left(\frac{\partial \gamma_i}{\partial \mu_i}\right) = \left(\frac{\partial \mu_i}{\partial \gamma_i}\right)^{-1} = \left(\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right)^{-1} = \frac{1}{v(\mu_i)}, \tag{1.12}$$

and using the chain rule and (1.3),

$$\left(\frac{\partial \mu_i}{\partial \beta}\right) = \frac{\partial \mu_i}{\partial g(\mu_i)}\frac{\partial g(\mu_i)}{\partial \mu_i} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)\frac{\partial x_i'\beta}{\partial \beta} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} x_i'. \tag{1.13}$$

#### 1.1.2.3 Likelihood equations

We can now derive the ML equations for the parameter the case of an one-dimensional parameter $\beta$. From (1.5) we have

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \sum\left[y_i\frac{\partial \gamma_i}{\partial \beta} - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \beta}\right] \\
&= \frac{1}{\tau^2}\sum(y_i - \mu_i)\frac{\partial \gamma_i}{\partial \beta} \qquad \text{,(by (1.6))} \\
&= \frac{1}{\tau^2}\sum(y_i - \mu_i)\frac{\partial \gamma_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta} \qquad \text{, (by (1.12) and (1.13))} \\
&= \frac{1}{\tau^2}\sum(y_i - \mu_i)w_i g_\mu(\mu_i)x_i',
\end{aligned} \tag{1.14}$$

where $w_i = [v(\mu_i)g_\mu^2(\mu_i)]^{-1}$. and $g_\mu$ as in (1.3)

Equation (1.14) can be written in matrix notation as follows

$$\frac{\partial l}{\partial \beta} = \frac{1}{\tau^2}\boldsymbol{X'W\Delta(y-\mu)}, \tag{1.15}$$

with $\boldsymbol{W} = \Big\{w_i\Big\}_{d\times d}$ and $\boldsymbol{\Delta} = \Big\{g_\mu(\mu_i)\Big\}_{d\times d}$.

With that said, the ML equations are given by

$$\boldsymbol{X'W\Delta y = X'W\Delta\mu}, \tag{1.16}$$

where $\boldsymbol{W, \Delta}$ and $\boldsymbol{\mu}$ involve the unknown parameter $\beta$. Due to the fact that the ML equations in (1.16) are typically non-linear functions of $\beta$, they can not be solved analytically.

In order to solve the ML equations and extract the large-sample variance of the estimator $\hat{\beta}$ of the parameter $\beta$, it is useful to have the expected value of the second derivative of the log likelihood:

$$\frac{\partial^2 l}{\partial\beta\partial\beta'} = -\frac{1}{\tau^2}\boldsymbol{X'W\Delta}\frac{\partial\mu}{\partial\beta'} + \frac{1}{\tau^2}\boldsymbol{X'}\frac{\partial\boldsymbol{W\Delta}}{\partial\beta}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{1.17}$$

so that

$$\begin{aligned} -E\left[\frac{\partial^2 l}{\partial\beta\partial\beta'}\right] &= -\frac{1}{\tau^2}\boldsymbol{X'W\Delta}\frac{\partial\mu}{\partial\beta'} + 0 \\ &= \frac{1}{\tau^2}\boldsymbol{X'W\Delta\Delta^{-1}X} \qquad \text{,(by (1.13))} \\ &= \frac{1}{\tau^2}\boldsymbol{X'WX}, \end{aligned} \tag{1.18}$$

where $\boldsymbol{W}$ as above.

#### 1.1.2.4 Large-sample variances

To extract the large-sample variance of the estimator $\hat{\beta}$ we first note that

$$\begin{aligned} -E\left[\frac{\partial^2 l}{\partial\beta\partial\tau^2}\right] &= -E\left[\frac{\partial}{\partial\tau^2}\frac{1}{\tau^2}\boldsymbol{X'W\Delta}(\boldsymbol{y}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\tau^4}\boldsymbol{X'W\Delta}E[\boldsymbol{y}-\boldsymbol{\mu}] \\ &= 0, \end{aligned} \tag{1.19}$$

so that estimation of $\tau^2$ does not affect the large-sample variance of $\hat{\beta}$.
The usual large-sample arguments along with (1.18) and (1.19), show that

$$var_\infty(\hat{\beta}) = \tau^2(\boldsymbol{X'WX})^{-1}, \tag{1.20}$$

where $var_\infty$ represents the limiting or asymptotic variance.

#### 1.1.2.5 Solving the ML equations

Solution of the ML equations given in (1.16), for $\beta$ is usually performed via a repeated weighted least squares method known as Fisher scoring. Fisher scoring is a repeated method for maximizing a likelihood. The form of the method is

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \boldsymbol{I}(\boldsymbol{\beta}^{(m)})^{-1}\frac{\partial}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{1.21}$$

where $(m)$ indicates the $\boldsymbol{m}^{th}$ iteration, $\boldsymbol{I(\beta)}$ is the Fisher's information matrix and $\boldsymbol{\theta}$ is the entire parameter vector.

With the use of (1.19), (1.18) and (1.15), the equation for $\beta$ is of the form

$$\beta^{(m+1)} = \beta^{(m)} + (\boldsymbol{X'WX})^{-1}\boldsymbol{X'W\Delta}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{1.22}$$

where it is easy to understand that $\boldsymbol{W}, \boldsymbol{\Delta}$ and $\boldsymbol{\mu}$ are evaluated at $\beta^{(m)}$.

## 1.1.3 Tests of hypotheses

### 1.1.3.1 Likelihood ratio tests

Likelihood ratio tests follow the typical procedure of comparing maximized values of the log likelihood both under $H_0$ and $H_1$. If the difference is large then $H_0$ is rejected.

When there are multiple parameters, our interest will most likely concern only in a subset of them. Thus, let the parameter vector $\boldsymbol{\beta}$ be partitioned into two components $\boldsymbol{\beta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and suppose interest focuses on $\boldsymbol{\theta}_1$ while $\boldsymbol{\theta}_2$ (often called as *nuisance parameter*) is left unspecified. Either or both of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ could be vector-valued and, with the condition that the entire parameter vector is of interest, $\boldsymbol{\theta}_2$ could be null.

Let us suppose that our hypothesis is of the form $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$, where $\boldsymbol{\theta}_{1,0}$ is a specified value of $\boldsymbol{\theta}_1$, and let $\hat{\boldsymbol{\theta}}_{2,0}$ be the MLE of $\boldsymbol{\theta}_2$ under the restriction that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$.
The likelihood ratio test statistic is given by

$$-2\log\Lambda = -2\Big[l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\Big], \tag{1.23}$$

where $\Lambda$ is the ratio of the two likelihood functions, and $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ and the large sample rejection region of the test is to reject $H_0$ in favor of $H_1$ when

$$-2\log\Lambda > \chi^2_{\nu,1-\alpha}, \tag{1.24}$$

where $\nu$ is the dimension of $\boldsymbol{\theta}_1$.

### 1.1.3.2 Wald tests

Another method of testing is to consider the large-sample normality of the ML estimator in order to form a test. From standard results

$$\hat{\boldsymbol{\theta}} \sim AN[\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})], \tag{1.25}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information for $\hat{\boldsymbol{\theta}}$.
Let us partition the Fisher information according to the dimensionality of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix}. \tag{1.26}$$

Then standard matrix algebra for partition matrices [90] and calculations for the multivariate normal show that large-sample variance of $\hat{\theta}_1$ is given by

$$var_\infty(\hat{\boldsymbol{\theta}}_1) = \left( \boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21} \right)^{-1}, \tag{1.27}$$

In order to test $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ we make use of the Wald statistic

$$W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0})'[var_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}), \tag{1.28}$$

which, under $H_0$, has the same large-sample $\chi^2$ distribution with degrees of freedom equal to the dimension of $\boldsymbol{\theta}_1$.
More precisely, we reject the $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ if

$$W > \chi^2_{\nu,1-\alpha}, \tag{1.29}$$

Likelihood ratio and Wald tests, although are available to test the same hypotheses and have the same limiting distribution, they have some differences. For large samples, and if the deviation from $H_0$ is not too extreme, they will give similar, but not identical, results [14]. On the other hand they may differ for small samples or for extreme deviations. Generally, various simulations ( [28]; [66]) have shown that use of the likelihood ratio test provides a more accurate approximation for small and moderate-sized samples than the use of the Wald test. Thus, the likelihood ratio test is to be preferred. However, the Wald test does have a computational advantage due to the fact that it does not require the calculation of $\hat{\boldsymbol{\theta}}_{2,0}$.

#### 1.1.3.3 Confidence intervals

Both likelihood ratio and Wald tests can be used to build large-sample confidence intervals for $\boldsymbol{\theta}_1$. For the likelihood ratio test we include in the confidence set all values $\boldsymbol{\theta}_1$ such that

$$-2\Big[l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\Big] \leq \chi^2_{\nu,1-\alpha}. \tag{1.30}$$

In (1.30) $\hat{\boldsymbol{\theta}}_{2,1}$ represents the MLE of $\boldsymbol{\theta}_2$ for each value of $\boldsymbol{\theta}_1$ checked for inclusion in the set.

For the confidence interval of the Wald test we include in the confidence set all values of $\boldsymbol{\theta}_1$ such that

$$(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)'[var_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \leq \chi^2_{\nu,1-\alpha}. \tag{1.31}$$

The computational burden of the likelihood-based confidence interval is larger than that for the Wald-based interval. Though the small and moderate-sized sample

performance of the likelihood-based confidence region has typically been found to be better.

### 1.1.4 Logistic regression model

The relationships between $\pi(x)$ and $x$ are usually nonlinear. A fixed change in $x$ may have less impact when $\pi$ is near 0 or 1 than when $\pi$ is near the middle of its range.

In practice, $\pi(x)$ often either increases or decreases continuously as $x$ increases. The S-shaped curves displayed in Fig.(1.1.4) are often realistic shapes for the above relationship, with $\beta$ the shape parameter.

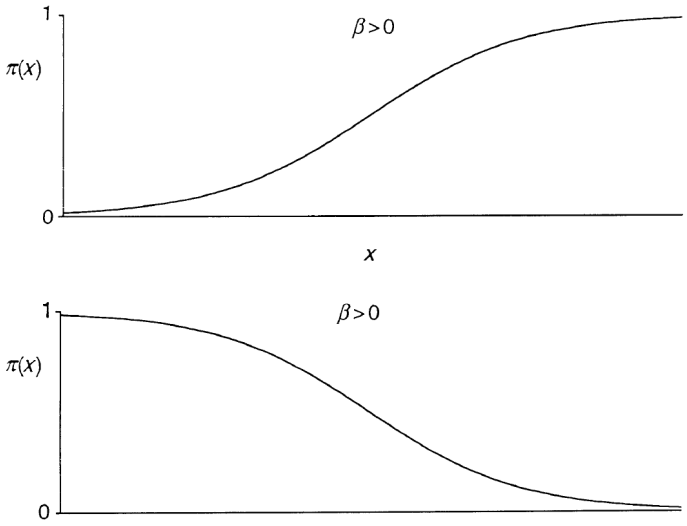


**Figure 1.1.4:** *Logistic regression functions*

The most important mathematical function of this shape is associated with the so called logistic regression which is given by

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

using the exponential function. The form of the corresponding logistic regression function is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \tag{1.32}$$

The logistic regression model (1.32) is a special case of a GLM. The random component for the outcomes (success, failure) has a binomial distribution. The link function which is called *the logit function* of $\pi$, and is denoted by "$logit(\pi)$", is given by $\log\left(\frac{\pi}{1-\pi}\right)$. The ratio $\frac{\pi}{1-\pi}$ is called *the odds ratio*, with the numerator defining the probability of success and the denominator the probability of failure. Logistic regression models are often called logit models. Since $\pi$ is a probability, it is restricted to the $0-1$ range and so the logit can be any real number. The real numbers are also the potential range for linear predictors (such as $\alpha+\beta x$) that form

the systematic component of a GLM, so this model does not have the structural problem that the linear probability models have.

As for the $\beta$ parameter in equation (1.32), it determines the rate of increase or decrease of the curve. When $\beta > 0$, $\pi(x)$ increases as $x$ increases, (Fig.(1.1.4(a))). Similarly when $\beta < 0$, $\pi(x)$ decreases as $x$ increases, (Fig.(1.1.4(b))). The magnitude of $\beta$ determines how fast the curve increases or decreases. As $|\beta|$ increases, the curve has a steeper rate of change. In the case that $\beta = 0$, the curve flattens to a horizontal straight line.

Remark: Another property of logistic regression relates to situations in which the explanatory variable $X$ rather than the response variable $Y$ is random. This occurs with retrospective sampling designs. Sometimes designs of this type are used because one of the response categories occurs rarely, and a prospective study might have too few cases to enable one to estimate effects of predictors well. For a given sample size, effect estimates have smaller standard errors when the number of outcomes of the two types are similar than when they are very different. Usually, retrospective designs are used with biomedical case-control studies. For samples of subjects having $Y = 1$ (cases) and having $Y = 0$ (controls), the value of $X$ is observed. Evidence exists of an association between $X$ and $Y$ if the distribution of $X$ values differs between cases and controls. For case control studies it is possible to estimate odds ratios but not other summary measures. Logistic regression parameters refer to odds and odds ratios. One can fit logistic regression models with data from case control studies and estimate effects of explanatory variables. The intercept term $\alpha$ in the model is not meaningful, because it relates to the relative numbers of outcomes of $Y = 1$ and $Y = 0$. We do not estimate this, because the sample frequencies for $Y = 1$ and $Y = 0$ are fixed by the nature of the case control study. With case control studies, it is not possible to estimate effects in binary models with link functions other than the logit. Unlike the odds ratio, the effect for the conditional distribution of $X$ given $Y$ does not equal that for $Y$ given $X$. This is the primary reason why in biosciences, logistic regression surpasses in popularity other models. Many control studies employ matching. Each case is matched with one or more control subjects (e.g. age). The model and subsequent analysis should take the matching into account. For a detailed illustration of logistic regression the reader is referred to [1].

## 1.2 Generalized Linear Mixed Models (GLMMs)

### 1.2.1 Introduction

Generalized Linear Mixed Models (GLMMs) are a powerful class of statistical models that combine the characteristics of GLMs (Section 1.1) and mixed models (models that include both fixed and random predictor variables). They handle a wide range of response distributions, and a wide range of scenarios where observations have been sampled in some kind of groups rather than completely independently. Even though they cannot perform satisfyingly under any setting, there are always situations where for greater flexibility one might choose a custom-built models. GLMMs are fast, powerful, can be extended to handle additional complexities such as zero-inflated responses [107], and can often be fitted with of-the-shelf software. However, there are a few real downside of GLMMs that come because of their

generality [37]: some of the standard recipes for model testing and inference may not apply. GLMMs are part of the statistical frontier, but not all of answers about their use and implementation are known.

## 1.2.2 Structure of the model

### 1.2.2.1 Conditional distribution of $\boldsymbol{y}$

To specify the model, we first have to define the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$. The response vector $\boldsymbol{y}$ is usually assumed to consist of conditional independent elements, with a distribution for each one with density from (or similar to) the exponential family:

$$y_i|\boldsymbol{u} \overset{indep.}{\sim} f_{Y_i|u}(y_i|\boldsymbol{u})$$

$$f_{Y_i|u}(y_i|\boldsymbol{u}) = exp\left\{\frac{y_i\gamma_i - b(\gamma_i)}{\tau^2 - c(y_i,\tau)}\right\}. \tag{1.33}$$

From (1.16) it is known that the conditional mean of $y_i$ is related to $\gamma_i$ in (1.33) via the identity $\mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i}$. It is a transformation of this mean that we wish to model as a linear model in both fixed and random factors:

$$E[y_i|\boldsymbol{u}] = \mu_i$$

$$g(\mu_i) = \boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}. \tag{1.34}$$

As in (1.1), $g(\cdot)$ is the *link function*, $\boldsymbol{x}_i'$ is the $\boldsymbol{i}^{th}$ row of the model matrix for the fixed effects, and $\beta$ is the fixed effects parameter vector. Furthermore, $z_i'$, is the $\boldsymbol{i}^{th}$ row of the model matrix for the random effects, and $\boldsymbol{u}$, the random effects vector. Note that in the present setting $\mu_i$ represents the conditional mean of $y_i$ given $\boldsymbol{u}$ and not the unconditional mean. To make the specification complete we assign a distribution to the random effects:

$$\boldsymbol{u} \sim f_{\boldsymbol{U}}(\boldsymbol{u}). \tag{1.35}$$

Taking into consideration the fact that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ is a notational extension of the generalized linear model presented in (1.1), most of the expressions and relationships derived there still hold. Consequently, as in (1.7), we denote the conditional variance of $y_i$ given $\boldsymbol{u}$ as $\tau^2 v(\mu_i)$ in order to exemplify its dependence on the conditional mean $\mu_i$.

## 1.2.3 Consequences of having random effects

### 1.2.3.1 Marginal versus conditional distribution

Based on the conditional distribution and characteristics in (1.33) and (1.34), we now extract aspects of the marginal distribution of $\boldsymbol{y}$ in order to fully understand the underlying mechanism of the observed data.

### 1.2.3.2 Mean of $\boldsymbol{y}$

The mean of $\boldsymbol{y}$ can be extracted by the standard device of iterated expectation:

$$E[y_i] = E[E[y_i|\boldsymbol{u}]]$$

$$= E[\mu_i]$$

$$= E[g^{-1}(\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u})]. \tag{1.36}$$

Generally, the above expression cannot be simplified due to the nonlinear function $g^{-1}(\cdot)$.

For illustrative purposes consider the log link namely $g(\mu) = \log \mu$ and $g^{-1}(x) = exp\{x\}$. Then we have

$$E[y_i] = E[exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}]$$

$$= exp\{E[y_i] = E[exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}]$$

$$= exp\{\boldsymbol{x}_i'\beta\}M_{\boldsymbol{u}}(\boldsymbol{z}_i), \tag{1.37}$$

where $M_{\boldsymbol{u}}(\boldsymbol{z}_i')$ is the moment generating function of $\boldsymbol{u}$ evaluated at $\boldsymbol{z}_i'$.

If we further assume that $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ and that each row of $\boldsymbol{Z}$ has a single entry equal to 1 with all the rest being zero, then

$$M_{\boldsymbol{u}}(\boldsymbol{z}_i) = exp\{\frac{\sigma_u^2}{2}\},$$

and

$$E[y_i] = exp\{\boldsymbol{x}_i'\beta\}exp\{\frac{\sigma_u^2}{2}\},$$

or

$$\log E[y_i] = \boldsymbol{x}_i'\beta + \frac{\sigma_u^2}{2}. \tag{1.38}$$

#### 1.2.3.3 Variances

To derive the marginal variance of $\boldsymbol{y}$ we make use of the following formula:

$$var(y) = var(E[y|u]) + E[var(y|u)]$$

Thus

$$var(y_i) = var(E[y_i|\boldsymbol{u}]) + E[var(y_i|\boldsymbol{u})]$$

$$var(y_i) = var(\mu_i) + E[\tau^2 v(\mu_i)]$$

$$= var(g^{-1}[\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}] + E[\tau^2 v(g^{-1}[\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}])]. \tag{1.39}$$

which is also not possible to be simplified substantially without making specific assumptions about the form $g(\cdot)$ and/or the conditional distribution of $\boldsymbol{y}$.

For illustrative purposes we consider again the log link and further assume that the elements of $\boldsymbol{y}$, given $\boldsymbol{u}$, are independent with a Poisson distribution. Hence the conditional variance of $y_i$ given $\boldsymbol{u}$ is $\tau^2 v(\mu_i) = \mu_i$. Using (1.39) we obtain

$$var(y_i) = var(\mu_i) + E[\mu_i]$$

$$= var(exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}) + E[exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}]$$

$$= E[(exp\{2(\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u})\})] + [E(exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\})]^2 + E[exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}] \tag{1.40}$$

$$= exp\{2\boldsymbol{x}_i'\beta\}\Big(M_{\boldsymbol{u}}(2\boldsymbol{z}_i) - [M_{\boldsymbol{u}}(\boldsymbol{z}_i)]^2 + exp\{-\boldsymbol{x}_i\beta\}M_{\boldsymbol{u}}(\boldsymbol{z}_i)\Big).$$

With the addition of the assumption that $u_i \sim N(0, \sigma_u^2)$ and that each row of $\boldsymbol{Z}$ has a single entry equal to 1 with all the rest being zero, then

$$var(y_i) = exp\{2\boldsymbol{x}_i'\beta\}\Big(exp\{2\sigma_u^2\} - exp\{\sigma_u^2\}\Big) + exp\{\boldsymbol{x}_i'\beta\}exp\{\frac{\sigma_u^2}{2}\}$$

$$exp\{\boldsymbol{x}_i'\beta + \frac{\sigma_u^2}{2}\}\Big(exp\{\boldsymbol{x}_i'\beta\}\Big[exp\{\frac{3\sigma_u^2}{2}\} - exp\{\frac{\sigma_u^2}{2}\}\Big] + 1\Big)$$

$$E[y_i]\Big(exp\{\boldsymbol{x}_i'\beta\}\Big[exp\{\frac{3\sigma_u^2}{2}\} - exp\{\frac{\sigma_u^2}{2}\}\Big] + 1\Big). \tag{1.41}$$

Notice that the term in parentheses in (1.41) is greater than 1 and so the variance is larger than the mean (overdispersion). Thus, although the conditional distribution of $y_i$ given $\boldsymbol{u}$ is Poisson, the marginal distribution cannot be. In fact, under these assumptions, it will always be overdispersed compared to the Poisson distribution. In this sense we can think of random effects as a way to model or attribute overdispersion to a particular source.

#### 1.2.3.4 Covariances and correlations

The use of random effects, introduces a correlation among observations which have any random effect in common. The same is true for generalized linear mixed models. Assuming conditional independence of the elements of $\boldsymbol{y}$ and with the use of the following formula for covariances

$$cov(y, w) = cov_u(E[y|u], E[w|u]) + E_u[cov(y, w|u)],$$

we have

$$cov(y_i, y_j) = cov(E[y_i|u], E[y_j|u]) + E[cov(y_i, y_j|u)]$$

$$= cov(\mu_i, \mu_j) + E[0]$$

$$cov(g^{-1}[\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}], g^{-1}[\boldsymbol{x}_j'\beta + \boldsymbol{z}_j'\boldsymbol{u}]). \tag{1.42}$$

In the case of a log link, this can be evaluated as

$$cov(y_i, y_j) = cov(exp\{\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u}\}, exp\{\boldsymbol{x}_j'\beta + \boldsymbol{z}_j'\boldsymbol{u}\}) + E[cov(y_i, y_j|u)]$$

$$= exp\{\boldsymbol{x}_i'\beta + \boldsymbol{x}_j'\beta\}cov(exp\{\boldsymbol{z}_i'\beta\}, exp\{\boldsymbol{z}_j'\beta\}) \tag{1.43}$$

$$= exp\{\boldsymbol{x}_i'\beta + \boldsymbol{x}_j'\beta\}[M_{\boldsymbol{u}}(\boldsymbol{z}_i + \boldsymbol{z}_j) - M_{\boldsymbol{u}}(\boldsymbol{z}_i)M_{\boldsymbol{u}}(\boldsymbol{z}_j)].$$

Again we make further the same assumptions, namely that $\boldsymbol{u} \sim N(0, \boldsymbol{I}\sigma_u^2)$ and that each row of $\boldsymbol{Z}$ has a single entry equal to 1 with all the rest being zero. Then

$$cov(y_i, y_j) = exp\{\boldsymbol{x}_i'\beta + \boldsymbol{x}_j'\beta\}\Big[exp\{\sigma_u^2\}(exp\{\boldsymbol{z}_i'\boldsymbol{z}_j\sigma_u^2\} - 1)\Big], \tag{1.44}$$

which if $\boldsymbol{z}_i'\boldsymbol{z}_j = 0$ (if the two observations do not share a random effect) or $\boldsymbol{z}_i'\boldsymbol{z}_j = 1$ then is equal to zero or positive respectively.

From (1.44) and (1.43), when $\boldsymbol{z}_i'\boldsymbol{z}_j = 1$, we can calculate the correlation (after canceling $exp\{\boldsymbol{x}_i'\beta + \boldsymbol{x}_j'\beta\}$ in the numerator and denominator) as:

$$corr(y_i, y_j) = \frac{e^{2\sigma_u^2} - e^{\sigma_u^2}}{\sqrt{e^{2\sigma_u^2} - e^{\sigma_u^2} + e^{-\boldsymbol{x}_i'\beta + \frac{\sigma_u^2}{2}}}}$$

$$= \frac{1}{\sqrt{(1 + \eta e^{-\boldsymbol{x}_i'\beta})(1 + \eta e^{-\boldsymbol{x}_j'\beta})}}, \tag{1.45}$$

where $\eta$ is given by $\frac{1}{e^{\frac{3\sigma_u^2}{2}} - e^{\frac{\sigma_u^2}{2}}}$.

## 1.2.4 Maximum likelihood estimation

### 1.2.4.1 Likelihood

From (1.33), (1.34), and (1.35) it is straightforward to write down the expression for the likelihood function:

$$L = \int \prod_i f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) f_U(\boldsymbol{u}) d\boldsymbol{u}, \tag{1.46}$$

where the integration is over the q-dimensional distribution of $\boldsymbol{u}$.

As an example consider the Poisson distribution for modeling data in correlated clusters thought to come from a Poisson distribution. An example of such a situation is described in [32], where the authors consider the analysis of the number of epileptic seizures in patients on a drug or placebo. In this context, the clusters would be repeated measurements taken on the same patients. We denote by $y_{ij}$ the $\boldsymbol{j}^{th}$ count taken in the $\boldsymbol{i}^{th}$ cluster. Hence we create a model as follows:

$$y_{ij}|\boldsymbol{u} \overset{indep.}{\sim} Poisson(\mu_{ij}); \quad i = 1, 2, ..., m; \; j = 1, 2, ..., n_i;$$

$$\log \mu_{ij} = \boldsymbol{x}_{ij}'\beta + u_i \tag{1.47}$$

$$u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2),$$

where a log link and a normal distribution for the random cluster (patient) effects are used. The normal distribution for the random effects is applicable since the log link carries the range of the parameter space for $\mu_{ij}$ into the entire real line. The random effects $u_i$ are shared among observations within the same cluster and hence those observations are being modeled as correlated.

The log likelihood can be simplified as follows

$$l = \log\left(\prod_{i=1}^{m}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}\frac{\mu_{ij}^{y_{ij}}e^{-\mu_{ij}}}{y_{ij}!}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}u_i^2}du_i\right)$$

$$= \boldsymbol{y}'\boldsymbol{X}\beta - \sum_{i,j}\log y_{ij}! + \sum_{i}\log\int_{-\infty}^{+\infty} exp\Big\{y_i u_i - \sum_{j} e^{\boldsymbol{x}'_{ij}\beta+u_i}\Big\}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}u_i^2}du_i. \quad (1.48)$$

Since (1.48), cannot be simplified further or evaluated in closed form, the same is true for maximizing values.

In the simplest cases, numerical integration for calculating the likelihood is straightforward and hence numerical maximization of the likelihood is carried out without difficulty.

The ML approach works relatively well in simple situations:

**1.** A single random effect.

**2.** Two or three nested random effects.

**3.** Random effects which come in cluster (e.g., longitudinal data with subjects having random intercepts and slopes).

For more complicated structures (e.g., crossed random factors) the approach fails.

#### 1.2.4.2 Likelihood equations

**-i. For the fixed effects parameters**

Although the likelihood equations are numerically complex, one can write them down in a simpler form. From (1.46)

$$l = \log\int f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})f_U(\boldsymbol{u})d\boldsymbol{u} = \log f_{\boldsymbol{Y}}(\boldsymbol{y}), \quad (1.49)$$

so that

$$\frac{\partial l}{\partial\beta} = \frac{\partial}{\partial\beta}\frac{\int f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})f_{\boldsymbol{U}(\boldsymbol{u})d\boldsymbol{u}}}{f_{\boldsymbol{Y}(\boldsymbol{y})}}$$

$$= \frac{\int\left[\frac{\partial}{\partial\beta}f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})\right]f_{\boldsymbol{U}(\boldsymbol{u})d\boldsymbol{u}}}{f_{\boldsymbol{Y}}(\boldsymbol{y})}, \quad (1.50)$$

since $f_U(u)$ does not involve $\beta$. Noting that

$$\frac{\partial}{\partial\beta}f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}) = \left(\frac{1}{f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})}\frac{\partial f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})}{\partial\beta}\right)$$

$$= \frac{\partial\log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})}{\partial\beta}f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}), \quad (1.51)$$

we can now rewrite (1.50) as

$$\frac{\partial l}{\partial \beta} = \frac{\int \frac{\partial \log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})}{\partial \beta} f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}) f_{U(\boldsymbol{u})d\boldsymbol{u}}}{f_{\boldsymbol{Y}(\boldsymbol{y})}}$$

$$\int \frac{\partial \log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u})}{\partial \beta} f_{\boldsymbol{U}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u}. \tag{1.52}$$

Combining (1.15) and (1.52) get

$$\frac{\partial l}{\partial \beta} = \int \boldsymbol{X}'\boldsymbol{W}^*(\boldsymbol{y} - \boldsymbol{\mu}) f_{\boldsymbol{U}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u}$$

$$= \boldsymbol{X}' E[\boldsymbol{W}^*|\boldsymbol{y}] - \boldsymbol{X}' E[\boldsymbol{W}^* \mu|\boldsymbol{y}], \tag{1.53}$$

where $\boldsymbol{W}^* = \left\{[\alpha(\phi) v(\mu_i g_\mu(\mu_i))]^{-1}\right\}_{d \times d}$

The likelihood equation for $\beta$ is therefore given by

$$\boldsymbol{X}' E[\boldsymbol{W}^*|\boldsymbol{y}] = \boldsymbol{X}' E[\boldsymbol{W}^* \mu|\boldsymbol{y}], \tag{1.54}$$

which is similar to (1.16), the difference being that $\boldsymbol{W}^*$ and $\boldsymbol{W}\mu$ are replaced by their conditional expected values given $\boldsymbol{y}$.

In cases like the Poisson example of (1.47), $\boldsymbol{W}^*$=$\boldsymbol{I}$ and the equations simplify to

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}' E[\boldsymbol{\mu}|\boldsymbol{y}]. \tag{1.55}$$

**-ii. For the random effects parameters**

We can easily derive a result similar to (1.52) for the ML equations for the parameters $\phi$ in the distribution $f_{\boldsymbol{U}(\boldsymbol{u})}$:

$$\frac{\partial l}{\partial \phi} = \int \frac{\partial \log f_{\boldsymbol{U}(\boldsymbol{u})}}{\partial \phi} f_{\boldsymbol{U}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u}$$

$$= E\left[\frac{\partial \log f_{\boldsymbol{U}(\boldsymbol{u})}}{\partial \phi}\middle| \boldsymbol{y}\right], \tag{1.56}$$

which though cannot be further simplified without providing a specific form of the random effects distribution.

### 1.2.5 Marginal versus conditional models

Suppose that we might directly hypothesize a model for the mean of $\boldsymbol{y}$ instead of starting from the conditional specification as in (1.33), (1.34), and (1.35).

As an example, suppose $y_{ij}$ is equal to 1 if the $\boldsymbol{j}^{th}$ child of woman $\boldsymbol{i}$ is born prematurely and *zero* otherwise. In addition, assume that we have a single predictor $x_{ij} =$ *number of drinks of alcohol per day.* If the marginal approach is chosen, then the marginal mean of $y_{ij}$ would be modeled directly by assuming, for instance, a logistic regression model:

$$logit(E[y_{ij}]) = logit(P\{y_{ij} = 1\}) = \alpha + \beta x_{ij}.$$

This means that the model would be for logit of the probability of premature birth, averaged over a population of women. In case the model was for correlated data, we would not be able to assume that the observations were independent.

On the other hand, the typical conditional approach corresponds to take into consideration the presence of a random factor for women and specifying a conditional model such as follows

$$logit(E[y_{ij}|\boldsymbol{u}]) = \alpha + \beta x_{ij} + u_i,$$

where $u_i$ represents the random woman effect. This corresponds to modeling the conditional probability of a premature birth for each woman [67].

From a probabilistic perspective, the marginal distribution of $\boldsymbol{y}$ can be calculated from the distribution of $\boldsymbol{u}$ and the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$. It is not possible to recover the marginal of $\boldsymbol{u}$ and the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$ from the marginal distribution of $\boldsymbol{y}$, which appears to favor the conditional specification of the model.

Note though that, in some cases, the marginal distribution (or perhaps only the marginal mean) may be adequate for answering questions of interest. For example, in the alcohol consumption example, a natural question of interest is what is the reduced rate of the incidence of premature birth if lowering, on average, women's alcohol consumption. In such cases, the potentially perplexed problem of specifying the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$ and the marginal distribution of $\boldsymbol{u}$ can be avoided. This can be considered as an advantage of marginal modeling and the basis of the generalized estimating equations. For more details the interested reader may refer to [67].

Distinguishing conditional from marginal models is straight forward probabilistically, but in practice it is frequently difficult. For example, a researcher might be interested in "*the influence of alcohol consumption on premature birth*", which would now specify which type of model to build. According to [67] researchers often thing about building models in a mechanistic way, which seems more compatible with the conditional approach. In the premature birth example, a researcher might think about the influence of alcohol consumption by trying to understand how alcohol influence individual women's physiology.

It is important to keep in mind the distinction between conditional and marginal models. For more details the reader may refer to [67]. Other advantages of the conditional approach are presented in the following examples:
If two studies are performed in different populations with different variances, then the marginal models will be different even though the conditional models are the same. Again for the alcohol consumption example, consider a small preliminary

study with a homogeneous study population (and hence a small variance for the subject random effect), a larger scale study with a heterogeneous study population. Even if the effect on every person in both studies is the same, the marginal models will differ due to different variances.

### 1.2.6 Generalized estimating equations

The generalized estimating equations (GEEs) approach begins by assuming in a marginal generalized linear model for the mean of $\boldsymbol{y}$ as a function of the predictors. Suppose for instance a logistic regression is hypothesized for the mean of binary data:

$$logit(E[\boldsymbol{y}]) = \boldsymbol{X}\beta.$$

Under the independence assumption as in [67] of all the elements of $\boldsymbol{y}$, the ML estimating equations for $\beta$ would be,

$$\boldsymbol{X'y} = \boldsymbol{X'}E[\boldsymbol{y}]. \tag{1.57}$$

Observe that the above are *unbiased estimating equations* meaning that the difference between the right and left hand side is zero, $(E(\boldsymbol{X'y} - \boldsymbol{X'}E[\boldsymbol{y}]) = 0)$. It should be pointed out that under regularity conditions, solutions to unbiased estimating equations give consistent estimators [43].

This estimator could be calculated by pretending that all the data were independent and conducting a standard logistic regression analysis.

### 1.2.7 Tests of hypotheses

#### 1.2.7.1 Likelihood ratio tests

As usual the likelihood ratio test for nested models can be performed by comparing $-2\log\Lambda$ to the appropriate percentile of a chi-square distribution. In the simple case where we are testing the null hypothesis that a single variance component is equal to zero, the large sample distribution is a 50/50 mixture of the constant 0 and a $\chi^2_1$ distribution. Hence the critical values are given by $\chi^2_{1,1-2\alpha}$ for an $\alpha$-level test.

The likelihood ratio test statistic cannot be evaluated analytically since the likelihood, in general, cannot, too. On the other hand, It can be calculated only numerically for a given data set.

#### 1.2.7.2 Asymptotic variances

It should be noted that the evaluation of even large-sample variances and standard errors can be a computational burden. In fact, we must rely on numerical methods to calculate even the observed Fisher information.

#### 1.2.7.3 Wald tests

For large samples, when construction of the observed or expected information is possible, Wald test can be formed by utilizing the large-sample normality of estim-

ators. Thus, we have for an individual parameter:

$$\frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\hat{var}_\infty(\hat{\beta}_i)}} \sim AN(0,1), \tag{1.58}$$

and for a set of linear combinations of the parameters

$$\boldsymbol{K}'\hat{\beta} - \boldsymbol{K}'\beta_0 \sim AN(0, \boldsymbol{K}'\boldsymbol{I}^{-1}\boldsymbol{K}), \tag{1.59}$$

where $\boldsymbol{I}$ represents the observed (or expected) information.

#### 1.2.7.4 Score tests

Various score tests have also been proposed( [25]; [46]; [58]; [24]) for testing the presence of a single random effect or multiple random effects. These tests have the advantage of not requiring the maximum likelihood estimators under the GLMM. On the other hand they usually have less power than the test based on random effects models.

CHAPTER 2

# Univariate Time Series Analysis

In this chapter univariate time series will be discussed. The main purpose of times series analysis, is to understand the past and to predict the future. A time series analysis, quantifies the main features in data and the random variations. Considering the improvement on computing power, time series methods have been widely applicable in government, industry, and commerce.

A *time series* is defined as a set of observation $y_t$, each one being recorded at a specific time $t$. In addition, if the set $T_0$ of times at which observations are made is a discrete set when observations are made at fixed time intervals, then it is called *a discrete time series*. On the other hand, *continuous time series* are obtained when observations are recorded continuously over some time interval (e.g., $T_0 = [0, 1]$).

Time series methods are used in everyday operational decisions. For example, gas suppliers in the United Kingdom have to place orders for gas from the offshore fields one day ahead of the supply [27]. Then, $\{y_t\}$ is the quantity ordered at time $t$. Thus, in the following Sections, some of the most common time series models such as, $AR(p)$, $MA(q)$, and $ARMA(p, q)$ are presented. At the end of this chapter, periodic regression models will be discussed, for which the implementation and performance will be fully examined in Chapter 4 through an experimental study.

## 2.1 Decomposition of series

### 2.1.1 Notation

We represent a times series consisting of $n$ values sampled at discrete times $1, 2, ..., n$ by $\{y_t \ : \ t = 1, ..., n\} = \{y_1, y_2, ..., y_n\}$. A time series model is a sequence of random variables and the observed time series is considered as a realization of the model. We will use the same notation for both, and rely on the context to make the distinction.
For the sample means of a series of length $n$ the following notation will be used:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{2.1}$$

The 'hat' notation ($\hat{y}$), will be used for the representation of a *prediction* or *forecast*. For the series $\{y_t \ : \ t = 1, ..., n\}$, a forecast at *lead time* $k$ is a predicted future value, and is denoted by $\hat{x}_{t+k|t}$.

### 2.1.2 Models

Many series are mainly dominated by *trend* and/or *seasonal effects*. A simple *additive decomposition* model is given by

$$y_t = m_t + s_t + \epsilon_t, \tag{2.2}$$

where, at the time $t$, $y_t$ is the observed series, $m_t$ is the trend, $s_t$ is the seasonal effect, and $\epsilon_t$ is an error term that generally is a sequence of correlated random variables with mean zero.

Suppose that the seasonal effect tends to increase as the trend increases, then a multiplicative model may be more appropriate:

$$y_t = m_t \cdot s_t \cdot \epsilon_t. \tag{2.3}$$

If the random variable is modeled by a multiplicative function and the variable is positive, the additive decomposition model given (2.2) can be used for $\log(y_t)$:

$$log(y_t) = m_t + s_t + \epsilon_t. \tag{2.4}$$

When it comes to the exponential function, some care is required when it is applied to the predicted mean of $\log(y_t)$ to obtain a prediction for the mean value $y_t$, as the effect is usually a biased prediction.
Let us suppose that

$$\epsilon_t \sim N(0, \sigma^2), \tag{2.5}$$

then the predicted mean value at time $t$ based in (2.4) is given by

$$\hat{x}_t = e^{m_t + s_t} e^{\frac{1}{2}\sigma^2}. \tag{2.6}$$

In the case of non-normally distributed and negatively skewed error series, as it is often the case after taking logarithms, the bias correction function will be an overcorrection and the implementation of an empirical adjustment is preferable.

### 2.1.3 Estimating trends and seasonal effects

There are various ways to estimate the trend $m_t$ at time $t$. The most relatively simple procedure, which does not assume any specific form, is to calculate a *moving average* centred on $y_t$. Cowpertwait and Metcalfe in [27] describe the moving average as an average of a specified number of time series values around each value in the time series, with the exception of the first few and last few terms. Thus, the length of the moving average is chosen to average out the seasonal effects so they can be estimated later. In the case of monthly series, assuming that the series begins at January ($t = 1$) and we average January up to December ($t = 12$), this average corresponds to a time $t = 6.5$, between June and July. For the estimation

of seasonal effects, we need a moving average at integer times. We can achieve so by averaging the average of January up to December and the average of February ($t = 2$) up to January ($t = 13$). This average of two moving averages corresponds to $t = 7$, and the process is called centring. Hence the trend at time $t$ can be estimated by the centred moving average

$$\hat{m}_t = \frac{\frac{1}{2}y_{t-6} + y_{t-5} + ... + y_{t-1} + y_t + y_{t+1} + ... + y_{t+5} + \frac{1}{2}y_{t+6}}{12}, \tag{2.7}$$

where $t = 7, ..., n - 6$. Our goal is to give equal weight to each month with the coefficients summing up to 1. Thus, the coefficients in (2.7) for each month are 1/12 (or sum to 1/12 in the case of the first and last coefficients). By using the seasonal frequency for the coefficients in the moving average, the procedure generalizes for any seasonal frequency (e.g., quarterly series), provided the condition that the coefficients sum to unity still holds.

The estimation of the monthly additive effect ($s_t$) at time $t$ can be obtained by subtracting $\hat{m}_t$:

$$\hat{s}_t = y_t - \hat{m}_t. \tag{2.8}$$

We can obtain a single estimate for each month by averaging these estimates of the monthly effect of each month.
Let us now consider the case that the time series is a whole number of years. Then the number of monthly effects averaged for each month is one less than the number of years of record. In this case, the average value of the twelve monthly additive components should be close, but not usually exactly equal to, zero. It is usual to adjust them by subtracting their mean so that they do average to zero. For the multiplicative monthly effect, the estimation is given by division; e.g., $\hat{s}_t = \frac{y_t}{\hat{m}_t}$. The adjustment to monthly multiplicative factors is done so that they average to unity. We can generalize the procedure, by using the same principle, to any seasonal frequency.
If the seasonal effect is additive, a seasonally adjusted series is given by $y_t - \bar{s}_t$. On the other hand, if the seasonal effect is multiplicative, an adjusted series is obtained from $\frac{y_t}{\bar{s}_t}$. The $\bar{s}_t$ term on both cases of additive and multiplicative seasonal effect, defines the seasonally adjusted mean for the month corresponding to time $t$.

### 2.1.4 Smoothing

*Smoothing* procedures can use points before and after the time at which the smoothed estimate is to be calculated. Thus the smoothed series will have some points missing at the beginning and at the end. An example of smoothing procedure is the centred moving average presented in Subsection (2.1.3). It is usual, instead of 'smoothing' to use the term *filtering*. Specifically, the term filtering is the process of obtaining the best estimate of some variable now, given the latest measurement of it and past measurements. The measurements are subject to random error and are described as being *corrupted by noise*. Filtering is an important part of control algorithms with a variety of applications.

## 2.2 Correlation

### 2.2.1 Expectation and variance

The mean function of a time series model is

$$\mu(t) = E(y_t), \tag{2.9}$$

and generally is a function of $t$.

If the mean remains constant over time, $\mu(t) \equiv \mu \ \forall t$, then the time series is considered to be stationary in the mean. The sample estimate of the population mean, $\mu$, is the sample mean, $\bar{y}$:

$$\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n}. \tag{2.10}$$

The above equation (2.10) relies on an assumption that a sufficiently long time series characterizes the hypothetical model. Models of this type are known as *ergodic*. For more details on ergodic series the reader may refer to [27].

The variance function of a time series model that is stationary in the mean is

$$\sigma^2(t) = E\Big[(y_t - \mu)^2\Big]. \tag{2.11}$$

which takes a different value at every time $t$. However, it is not possible to estimate a different variance at each time point from a single time series. Assuming that the model is stationary in the variance, the constant population variance, $\sigma^2(t) = \sigma^2$, can be estimated from the sample variance:

$$Var(y) = \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n-1}. \tag{2.12}$$

It is a common issue sequential observations in a time series analysis to be correlated. If the correlation is positive, then $Var(y)$ will tend to underestimate the population variance in a short time series since successive observations tend to be relatively similar. However, usually this does not present a problem since the biases decrease sharply as the length $n$ of the series increases.

### 2.2.2 Autocorrelation

In the study of statistical distributions, the mean and the variance play an important role since they summarize two key distributional properties, the central location and the spread. Similarly, in the study of time series models, a key role is played by *second-order properties*, which include the mean, variance, and autocorrelation.

Let us consider a time series model that "weakly stationary", namely is stationary in the mean and variance. In addition let us suppose that the variables are correlated. The model is *second-order stationary* if the correlation between variables depends only on the number of time steps separating them. The number of time

steps between the variables is known as *the lag*. A correlation of a variable with itself a different times is known as *autocorrelation* or *serial correlation*.
If a time series is second-order stationary, we can define an *autocovariance function (ACVF)*, $\gamma_k$, as a function of the lag $k$:

$$\gamma_k = E\Big[(y_t - \mu)(y_{t+k} - \mu)\Big]. \tag{2.13}$$

Notice that $ACVF$ does not depend on $t$ because the expectation, is constant, $\forall t$.
The *lag k autocorrelation function (ACF)*, $\rho_k$, is defined by

$$\rho_k = \frac{\gamma_k}{\gamma_0}. \tag{2.14}$$

By definition $\gamma_0 = \sigma^2$ and thus $\rho_0$ is 1.

The $ACVF$ and $ACF$ can be estimated from a time series by their sample equivalents. The sample autocovariance, $c_k$, is calculated as

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y}). \tag{2.15}$$

The autocovariance at lag 0, is the sample variance. In addition, a denominator $n$ is used when calculating $c_k$, although only $n - k$ terms are used for the estimation. Adopting this definition constrains all sample autocorrelations to lie between $-1$ and 1. The sample ACF is defined as

$$r_k = \frac{c_k}{c_0}. \tag{2.16}$$

## 2.3 The correlogram

Correlograms like the one in Figure (2.3) have the following features:

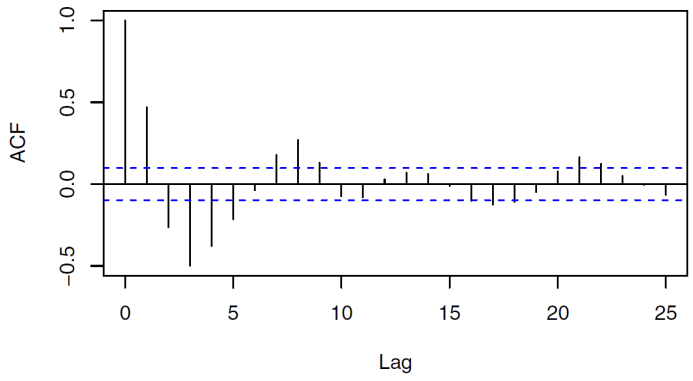


**Figure 2.3:** *Example of Correlogram*

- The $x$-axis gives the lag($k$) and the $y$-axis gives the sample autocorrelation ($r_k$) at each lag. The unit of lag is the sampling interval, 0.1 second. Correlation is dimensionless and thus there is no unit for the $y$-axis.

- If $\rho_k = 0$, the sampling distribution of $r_k$ is approximately normal, with a mean of $-\frac{1}{n}$ and a variance of $\frac{1}{n}$. The dotted lines at the correlogram are drawn at
$$-\frac{1}{n} \pm \frac{2}{\sqrt{n}},$$
with "2" being used as an approximation of $z_{0.05} = 1.96$.
In the case that $r_k$ falls outside these lines, we have evidence against the null hypothesis that $\rho_k = 0$ at the 5% level. However, we should be careful about interpreting multiple hypothesis test. Firstly, if $\rho_k = 0$ at all lags $k$, we expect 5% of the estimates, $r_k$, to fall outside the lines. Secondly, the $r_k$ are correlated, so if one falls outside the lines, the neighbouring ones are more likely to be statistically significant.

- The lag 0 autocorrelation is always 1 and is shown on the plot. It is included so that we can compare values of the other autocorrelations relative to the theoretical maximum of 1. Something like this is considered useful because, if we have a log time series, small values of $r_k$ with no practical consequence may be statistically significant. However, some understanding is needed to decide what constitutes a notworthy autocorrelation from a practical perspective. By squaring the autocorrelation we obtain the percentage of variability explained by a linear relationship between the variables. For example, a lag 1 autocorrelation of 0.1 implies that a linear dependence of $y_t$ on $y_{t-1}$ would only explain 1% of the variability of $y_t$. Usually we treat a statistically significant result as important when it has almost no practical consequence.

The correlogram provides graphically an idea about the stationarity of a time series. Specifically:

- If the spikes on the correlogram have a slow decrease from lag 1 then the time series is considered non-stationary.

- If the spikes on the correlogram have a fast decrease from lag 1 then the time series is considered stationary.

## 2.4 White noise

### 2.4.1 Definition

A time series $\{\epsilon_t : t = 1, 2, ..., n\}$ is discrete *white noise* if the variables $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are *independent* and *identically* distributed with mean equal to *zero*. This implies that all the variables have the same variance $\sigma^2$ and $Cor(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. In case of normality (i.e., $\epsilon_t \sim N(0, \sigma^2)$), the series is called *Gaussian white noise*, in which case $\epsilon_t \sim WN(0, \sigma^2)$.

## 2.5 Random walks

### 2.5.1 Definition

Let $\{y_t\}$ be a time series. Then $\{y_t\}$ is a random walk if

$$y_t = y_{t-1} + \epsilon_t, \tag{2.17}$$

where $\{\epsilon_t\}$ is a white noise series. By substituting $y_{t-1} = y_{t-2} + \epsilon_{t-1}$ in Eq.(2.17) and then for $y_{t-2}$, followed by $y_{t-3}$ and so on (a process known as "*back substitution*") we get:

$$y_t = \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} + \cdots . \tag{2.18}$$

In practice the series above will not be infinite but will start at some time $t = 1$. Hence,

$$y_t = \epsilon_1 + \epsilon_2 + \cdots + \epsilon_t. \tag{2.19}$$

### 2.5.2 The backward shift operator

We define the *backward shift operator* (also known as *lag operator*) as

$$\boldsymbol{B} y_t = y_{t-1}. \tag{2.20}$$

By repeatedly applying $\boldsymbol{B}$, it follows that

$$\boldsymbol{B}^n y_t = y_{t-n}. \tag{2.21}$$

Using $\boldsymbol{B}$, we can rewrite Eq.(2.17) as

$$y_t = \boldsymbol{B} y_t + \epsilon_t \Rightarrow (1 - \boldsymbol{B}) y_t = \epsilon_t \Rightarrow y_t = (1 - \boldsymbol{B})^{-1} \epsilon_t$$

$$\Rightarrow y_t = (1 + \boldsymbol{B} + \boldsymbol{B}^2 + \cdots)\epsilon_t \Rightarrow y_t = \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} + \cdots .$$

### 2.5.3 Random walk: Second-order properties

The second-order properties of a random walk follow as

$$\begin{cases} \mu = 0 \\ \\ \gamma_k(t) = Cov(y_t, y_{t+k}) = t\sigma^2 \end{cases} . \tag{2.22}$$

The first part of (2.22) is obvious since by (2.19) $\{y_t\}$ is a finite sum of white noise terms. As for the autocovariance in Eq.(2.22) it can be written as follows:

$$\gamma_k(t) = Cov\Big(y_t, y_{t+k}\Big) = Cov\Big(\sum_{i=1}^{t}\epsilon_i, \sum_{j=1}^{t+k}\epsilon_j\Big) = \sum_{i=j} Cov(\epsilon_i, \epsilon_j) = t\sigma^2.$$

Since the covariance is a function of time, the process is non-stationary. The variance is $t\sigma^2$ and thus it increases without limit as $t$ increases. It follows that a random walk is only suitable for short term predictions.

Using Eq.(2.22) we define the time-varying autocorrelation function for $k > 0$ as

$$\rho_k(t) = \frac{Cov(y_t, y_{t+k})}{\sqrt{Var(y_t)Var(y_{t+k})}} = \frac{t\sigma^2}{\sqrt{t\sigma^2(t+k)\sigma^2}} = \frac{1}{\sqrt{1+\frac{k}{t}}}. \tag{2.23}$$

Notice that for large $t$ with $k$ considerably less than $t$, $\rho_k$ is close to 1. Hence, the correlogram for a random walk is characterized by positive correlations that decay very slowly down from unity.

### 2.5.4 The difference operator

We can transform a non-stationary series to a stationary, by differencing its adjacent terms. For example, if the series $\{y_t\}$ is a random walk, it is non-stationary. However, from Eq.(2.17), the first-order differences of $\{y_t\}$ produce the stationary white noise series $\{\epsilon_t\}$ given by $\{y_t - y_{t-1} = \epsilon_t\}$. Thus, differencing has been proved to be a useful "filtering" procedure in the study of non-stationary time series. The difference operator $\nabla$ is defined by

$$\nabla y_t = y_t - y_{t-1} \equiv (1 - \boldsymbol{B})y_t, \tag{2.24}$$

Higher-order differencing can be expressed as (see [27])

$$\nabla^n = (1 - \boldsymbol{B})^n. \tag{2.25}$$

## 2.6 Autoregressive models (AR)

### 2.6.1 Definition

The series $\{y_t\}$ is an autoregressive process of order $p$, abbreviated by $AR(p)$ if

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \epsilon_t, \tag{2.26}$$

where $\epsilon_t$ is white noise and $\alpha_1, ..., \alpha_p$ are the model parameters, with $\alpha_p \neq 0$ for an order $p$ process. $AR(p)$ can be expressed as a polynomial of order $p$ in terms of the backward shift operator:

$$\theta_p(\boldsymbol{B})y_t = (1 - \alpha_1\boldsymbol{B} - \alpha_2\boldsymbol{B}^2 - \cdots - \alpha_p\boldsymbol{B}^p)y_t = \epsilon_t. \tag{2.27}$$

### 2.6.2 Stationary and non-stationary AR processes

The equation $\theta_p(\boldsymbol{B}) = 0$ where $\boldsymbol{B}$ is formally treated as a number (real or complex), is called *the characteristic equation*. The roots of the characteristic equation

must all exceed unity in absolute value for the process to be stationary.

Using $\boldsymbol{B}$, a stable $AR(1)$ process ($|\alpha| < 1$) can be written as

$$(1 - \alpha\boldsymbol{B})y_t = \epsilon_t$$

$$\Rightarrow y_t = (1 - \alpha\boldsymbol{B})^{-1}\epsilon_t$$

$$= \epsilon_t + \alpha\epsilon_{t-1} + \alpha^2_\epsilon t-2 + \cdots = \sum_{i=0}^{\infty} \alpha^i \epsilon_{t-i}. \tag{2.28}$$

Hence, the mean is given by

$$\mu = E(y_t) = E\left(\sum_{i=0}^{\infty} \alpha^i \epsilon_{t-i}\right) = \sum_{i=0}^{\infty} \alpha^i E(\epsilon_{t-i}) = 0.$$

As for the autocovariance, it is given by

$$\gamma_k = Cov(y_t, y_{t+k}) = Cov\left(\sum_{i=0}^{\infty} \alpha^i \epsilon_{t-i}, \sum_{j=0}^{\infty} \alpha^j \epsilon_{t+k-j}\right)$$

$$= \sum_{j=k+i} \alpha^i \alpha^j Cov(\epsilon_{t-i}, \epsilon_{t+k-j})$$

$$= \alpha^k \sigma^2 \sum_{i=0}^{\infty} \alpha^{2i} = \frac{\alpha^k \sigma^2}{(1 - \alpha^2)}.$$

By (2.14), the autocorrelation function is given by

$$\rho_k = \alpha^k \quad (k \geq 0), \tag{2.29}$$

where $|a| < 1$. Thus, the correlogram decays to zero more rapidly for small $a$.

The following two examples describe the procedure for determining whether an $AR$ process is stationary or non-stationary:

1. The $AR(1)$ model $y_t = \frac{1}{2}y_{t-1} + \epsilon_t$ is stationary because the root of $1 - \frac{1}{2}\boldsymbol{B} = 0$ is $\boldsymbol{B} = 2$, which is greater than 1.

2. The model $y_t = \frac{1}{2}y_{t-1} + \frac{1}{2}y_{t-2} + \epsilon_t$ is non-stationary because one of the roots is unity. Indeed, since $-\frac{1}{2}(\boldsymbol{B}^2 + \boldsymbol{B} - 2)y_t = -\frac{1}{2}(\boldsymbol{B} - 1)(\boldsymbol{B} + 2)y_t = \epsilon_t$, the polynomial $\theta(\boldsymbol{B}) = -\frac{1}{2}(\boldsymbol{B} - 1)(\boldsymbol{B} + 2)$ has roots $\boldsymbol{B} = 1, -2$. Since there is one unit root, the model is non-stationary.

### 2.6.3 Partial autocorrelation

Although the underlying model $y_t$ only depends on the previous value $y_{t-1}$, the autocorrelations are non-zero for all lags. The *partial autocorrelation* at lag $k$ is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. For example, suppose an $AR(1)$ process, then the partial autocorrelation of $AR(1)$ will be zero for all lags greater than 1. Generally, the partial autocorrelation at lag $k$ is the $\boldsymbol{k}^{th}$ coefficient of a fitted $AR(k)$ model. In case of an $AR(p)$ process, the coefficients $a_k$ will be zero for all $k > p$. Thus, an $AR(p)$ process has a correlogram of partial autocorrelations that is zero after lag $p$. As seen from above, a plot of the estimated partial autocorrelations can be useful when determining the order of a suitable $AR$ process for a time series.

## 2.7 Moving average models (MA)

### 2.7.1 MA(q) process: Definition and properties

A *moving average (MA)* process of order $q$ is a linear combination of the current white noise term and the $q$ most recent past white noise terms and is defined by

$$y_t = \epsilon_t + \beta_1 + \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q}, \tag{2.30}$$

where $\epsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma_w^2)$. We can rewrite Eq.(2.30) in terms of the backward shift operator $\boldsymbol{B}$ as

$$y_t = (1 + \beta_1 \boldsymbol{B} + \beta_2 \boldsymbol{B}^2 + \cdots + \beta_q \boldsymbol{B}^q)\epsilon_t = \phi_q(\boldsymbol{B})\epsilon_t, \tag{2.31}$$

where $\phi_q$ is a polynomial of order $q$. Because $MA$ processes consist of a finite sum of stationary white noise terms, they are stationary and hence have a time-variant mean and autocovariance.

For the derivation of the mean and variance of $\{y_t\}$, the mean is *zero*, since is is a sum of terms that all have a mean of *zero*, and the variance is $\sigma_w^2(1+\beta_1^2+\cdots+\beta_q^2)$ because each of the white noise terms has the same variance and the terms are mutually independent.
The autocorrelation function, for $k \geq 0$, is given by

$$\rho_k = \begin{cases} 1, & k = 0 \\ \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^{q} \beta_i^2}, & k = 1, ..., q \\ 0, & k > q \end{cases}, \tag{2.32}$$

where $\beta_0$ is unity.

An $MA$ process is *invertible* if it can be expressed as a stationary autoregressive process of infinite order without an error term. For example, the $MA$ process

$y_t = (1 - \beta \boldsymbol{B})\epsilon_t$ can be expressed as

$$\epsilon_t = (1 - \beta \boldsymbol{B})^{-1} y_t = y_t + \beta y_{t-1} + \beta^2 y_{t-2} + \cdots , \tag{2.33}$$

provided that $|\beta| < 1$, which is required for convergence.

## 2.8 Mixed models: The ARMA process

### 2.8.1 Definition

In Sec.(2.6) and Sec.(2.7) $AR$ and $MA$ processes, respectively, were discussed. A useful class of models are obtained when $AR$ and $MA$ terms are added together in a single expression.

A time series $\{y_t\}$ follows an autoregressive moving average ($ARMA$) process of order $(p, q)$, denoted ARMA$(p, q)$, when

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \cdots + \beta_q \epsilon_{t-q}, \tag{2.34}$$

where $\{\epsilon_t\}$ is white noise. Once again, we can express Eq.(2.34) in terms of the backward shift operator as

$$\theta_p(\boldsymbol{B}) y_t = \phi_q(\boldsymbol{B}) \epsilon_t. \tag{2.35}$$

After defining the $ARMA(p, q)$ process, the following points should be noted:

a. The process is stationary when the roots of $\theta$ all exceed unity in absolute value.

b. The process is invertible when the roots of $\phi$ all exceed unity in absolute value.

c. The $AR(p)$ model is the special case of $ARMA(p, q)$, with $q = 0$.

d. The $MA(q)$ model is the special case of $ARMA(p, q)$, with $p = 0$.

e. *Parameter parsimony.* When fitting to data, an $ARMA$ model will often be more parameter efficient (i.e., require fewer parameters) than a single $MA$ or $AR$ model.

f. *Parameter redundancy.* When $\theta$ and $\phi$ share a common factor, a stationary model can be simplified.

### 2.8.2 Second-order properties of an ARMA model

First we express the $\{y_t\}$ in terms of white noise components $\{\epsilon_t\}$ because white noise terms are independent. Below, we illustrate the procedure for the $ARMA(1, 1)$ model which is defined by

$$y_t = \alpha y_{t-1} + \epsilon_t + \beta \epsilon_{t-1}, \tag{2.36}$$

where $\epsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma_w^2)$. In order to express $\{y_t\}$ in terms of white noise components, we rearrange Eq.(2.36) as,

$$y_t = (1 - \alpha \boldsymbol{B})^{-1}(1 + \beta \boldsymbol{B})\epsilon_t.$$

Expanding the right-hand side,

$$y_t = (1 + \alpha \boldsymbol{B} + \alpha^2 \boldsymbol{B}^2 + \cdots)(1 + \beta \boldsymbol{B})\epsilon_t$$

$$= \left( \sum_{i=0}^{\infty} \alpha^i \boldsymbol{B}^i \right)(1 + \beta \boldsymbol{B})\epsilon_t$$

$$= \left( 1 + \sum_{i=0}^{\infty} \alpha^{i+1} \boldsymbol{B}^{i+1} + \sum_{i=0}^{\infty} \alpha^i \beta \boldsymbol{B}^{i+1} \right)\epsilon_t$$

$$= \epsilon_t + (\alpha + \beta) \sum_{i=1}^{\infty} \alpha^{i-1} \epsilon_{t-i}. \tag{2.37}$$

With the equation in the form above, the second-order properties follow:

a.

$$\mu = E(y_t) = 0$$

b.

$$Var(y_t) = \sigma_w^2 + \sigma_w^2(\alpha + \beta)^2(1 - \alpha^2)^{-1} \tag{2.38}$$

c.

$$\gamma_k = Cov(y_t, y_{t+k}) = (\alpha+\beta)\alpha^{k-1}\sigma_w^2 + (\alpha+\beta)^2\sigma_w^2\alpha^k(1-\alpha^2)^{-1}, \quad k > 0 \tag{2.39}$$

d.

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\alpha^{k-1}(\alpha + \beta)(1 + \alpha\beta)}{1 + \alpha\beta + \beta^2} \tag{2.40}$$

## 2.8.3 Non-seasonal ARIMA models

### 2.8.3.1 Integrated model

A series $\{y_t\}$ is *integrated* of order $d$, denoted as $I(d)$, if the $\boldsymbol{d}^{th}$ difference of $\{y_t\}$ is a white noise $\{\epsilon_t\}$; i.e., $\nabla^d y_t = \epsilon_t$. Since $\nabla^d \equiv (1-\boldsymbol{B})^d$, where $\boldsymbol{B}$ is the backward shift operator, a series $\{y_t\}$ is integrated of order $d$ if

$$(1-\boldsymbol{B})^d y_t = \epsilon_t. \tag{2.41}$$

Note that the random walk presented in Sec.(2.5), is a special case of $I(1)$.

### 2.8.3.2 Definition

A time series $\{y_t\}$ follows an $ARIMA(p,d,q)$ process if the $\boldsymbol{d}^{th}$ difference of the $\{y_t\}$ series is an $ARMA(p,q)$ process. If $y_t = (1-\boldsymbol{B})^d y_t$, then $\theta_p(\boldsymbol{B})y_t = \phi_q(\boldsymbol{B})\epsilon_t$. To obtain the more concise form for an $ARIMA(p,d,q)$ process, we can substitute for $y_t$ and thus

$$\theta_p(\boldsymbol{B})(1-\boldsymbol{B})^d y_t = \phi_q(\boldsymbol{B})\epsilon_t, \tag{2.42}$$

where $\theta_p$ and $\phi_q$ are polynomials of orders $p$ and $q$, respectively. Some examples of $ARIMA$ models are:

1. $\boldsymbol{ARIMA(1,1,1)}$ **:** $y_t = \alpha y_{t-1} + y_{t-1} + \epsilon_t + \beta\epsilon_{t-1}$, where $\alpha, \beta$ are model parameters.
2. $\boldsymbol{ARIMA(0,1,2)}$ **:** $y_t = y_{t-1} + \epsilon_t + \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2}$, where $\beta_1, \beta_2$ are model parameters.

For a more extensive presentation of $ARIMA$ processes, see [27].

## 2.8.4 Seasonal ARIMA models

### 2.8.4.1 Definition

A seasonal $ARIMA$ model, $((S)ARIMA)$, uses differencing at a lag $D$ equal to the number of seasons (s) to remove additive seasonal effects. The seasonal component of the $(S)ARIMA$ model consists of autoregressive and moving average terms at lag $m$. In terms of the backward shift operator, $(S)ARIMA(p,d,q)(P,D,Q)_m$ can be expressed as

$$\Theta_P(\boldsymbol{B^m})\theta_p(\boldsymbol{B})(1-\boldsymbol{B}^m)^D(1-\boldsymbol{B}^d)y_t = \Phi_Q(\boldsymbol{B}^m)\phi_q(\boldsymbol{B})\epsilon_t, \tag{2.43}$$

where $\Theta_P, \theta_p, \Phi_Q, \phi_q$ are polynomials of orders $P, p, Q$ and $q$, respectively. If $D = d = 0$ and the roots of the characteristic equation (left-hand side of Eq.(2.43)) all exceed unity in absolute value, the resulting model would be stationary. In general, except the case mentioned above, the model is non-stationary. Some examples of $(S)ARIMA$ models are:

1. $\boldsymbol{(S)ARIMA(0,0,0)(1,0,0)_{12}}$ **:** $y_t = \alpha y_{t-12} + \epsilon_t$, where $\alpha$ is model parameter. Such a model would be suitable for monthly data when only the value in the month of the previous year influences the current monthly value. The model is stationary when $|\alpha|^{-\frac{1}{12}} > 1$.

2. $\boldsymbol{(S)ARIMA(0,1,0)(0,1,1)_4}$ **:** $y_t = \alpha y_{t-1} + y_{t-4} + \epsilon_t + \beta\epsilon_{t-4}$, where $\alpha, \beta$ are model parameters. The model presented in this example, is appropriate if the seasonal terms contain a stochastic trend.

Differencing at lag $m$ will remove linear trend. Thus, there is a choice whether or not to include lag 1 differencing. If we include lag 1 differencing, when a linear trend is appropriate, it will introduce moving average terms into white noise series. For a more extensive presentation of $(S)ARIMA$ process, once again, we refer to [27].

## 2.9 Periodic Models

Consider a univariate time series $y_t$, which is observed quarterly for $N$ years, that is, $t = 1, 2, ..., n$. We assume, without loss of generality, that $n = 4N$.

### 2.9.1 Periodic autoregressive models (PAR)

A *periodic autoregressive model* of order $p$, $(PAR(p))$, is defined as

$$y_t = \mu_m + \alpha_{1(m)} y_{t-1} + \cdots + \alpha_{p(m)} y_{t-p} + \epsilon_t, \tag{2.44}$$

where $\mu_m$ is a seasonally-varying intercept term. The $\alpha_{1(m)}, ..., \alpha_{p(m)}$ are autoregressive parameters up to order $p(m)$ which may vary with the season $m$, where $m = 1, 2, 3, 4$. In addition, we assume that $\epsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma^2)$. This assumption may be relaxed by allowing $\epsilon_t$ to have seasonal variance $\sigma_m^2$ but this will shall not be discussed further in this work. Since some of $\alpha_{i(m)}$, $i = 1, ..., p$, can take zero values, the order $p$ in Eq.(2.44) is the maximum of all $p(m)$, where $p(m)$ denotes the $AR$ order per season $m$. Hence, we may also consider the so-called subset periodic autoregressions, which are investigated in [36].
For a more detailed discussion on $PAR$ models, see [27].

### 2.9.2 Periodic moving average models (PMA)

Periodic moving average processes are representatives of the class of periodic models suitable for the description of some seasonal time series and for the construction of multivariate moving average models.

A *periodic moving average model* of order $q$, $(PMA(q))$, for $y_t$ can be written as

$$y_t = \mu_m + \epsilon_t + \beta_{1(m)} \epsilon_{t-1} + \cdots + \beta_{q(m)} \epsilon_{t-q}, \tag{2.45}$$

where $\mu_m$ is a seasonally-varying intercept term and $\epsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma^2)$. The $\beta_{1(m)}, ..., \beta_{q(m)}$ are moving average parameters up to order $q(m)$ which may vary with the season $m$, where $m = 1, 2, 3, 4$. For a theoretical analysis of $PMA(q)$, see [23]

### 2.9.3 Periodic autoregressive moving average models (PARMA)

As discussed in [101], many seasonal time series cannot be filtered or standardized to achieve second-order stationarity. This happens because the correlation structure of the series depends on the season. For example, consider a river where high runoff periods occur in the spring and low flows coupled with irrigation diversions occur in the summer. The stream-flow correlations between spring months may be different from the correlations between summer months. In such situations, a useful class of models is that of *periodic autoregressive moving average* ($PARMA$) models ( [48]; [80]; [99]; [97]), which are extensions of commonly used $ARMA$ models to allow parameters that depend on season.

A *periodic autoregressive moving average model* of order $p, q$, denoted $PARMA(p, q)$, for $y_t$ can be written as

$$y_t = \mu_m + \alpha_{1(m)} y_{t-1} + \cdots + \alpha_{p(m)} y_{t-p} + \epsilon_t + \beta_{1(m)} \epsilon_{t-1} + \cdots + \beta_{q(m)} \epsilon_{t-q}, \qquad (2.46)$$

where $\mu_m$ is a seasonally-varying intercept term and $\epsilon_t \overset{i.i.d.}{\sim} WN(0, \sigma^2)$. The $\alpha_{1(m)}, ..., \alpha_{p(m)}$ and $\beta_{1(m)}, ..., \beta_{q(m)}$ are autoregressive and moving average parameters, respectively, up to order $p(m)$ and $q(m)$, which may vary with the season $m$, where $m = 1, 2, 3, 4$. For a theoretical analysis of $PARMA(p, q)$, see [101].

CHAPTER 3

# Model Selection Criteria

In this chapter, model selection criteria will be presented. Model selection criteria is an important part of any statistical analysis, and plays central role in the pursuit of the ideal underlying mechanism of the phenomenon under investigation. In 1951, Kullback and Leibler developed a measure to capture the information that is lost when approximating reality; that is, the Kullback and Leibler measure is a criterion for a good model that minimizes the loss of information. Twenty years later, Akaike established a relationship between Kullback-Leibler measure and maximum likelihood estimation method (AIC). Based on Kullback-Leibler information, several criteria have been developed. The criteria mentioned in this chapter are the most common ones when it comes to model selection. Kullback-Leibler Information (K-L), Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), Modified Divergence Information Criterion (MDIC), and Coefficient of Determination ($R^2$) for GLMMs, will be discussed in the following sections.

## 3.1 Kullback-Leibler Information

### 3.1.1 Definition and properties

Let $\boldsymbol{x}_n = \{x_1, x_2, ..., x_n\}$ be a set of $n$ independent observations from an unknown probability distribution function $G(x)$. Below we refer to the probability distribution function $G(x)$ that generates data as the true model or the true distribution. On the other hand, let $F(x)$ be an arbitrarily specified model. If the probability distribution functions $G(x)$ and $F(x)$ have density functions $g(x)$ and $f(x)$ respectively, then they are called *continuous models* (or *continuous distribution models*). Their expression as probabilities of events, given a set or a countably infinite set of discrete points $\{x_1, x_2, ..., x_k, ...\}$, is as follows

$$g_i = g(x_i) \equiv Pr(\{\omega;\ X(\omega) = x_i\}),$$

$$f_i = f(x_i) \equiv Pr(\{\omega;\ X(\omega) = x_i\}), \quad i = 1, 2, ..., \tag{3.1}$$

and thus these models are called *discrete models* (*discrete distribution models*).

We assume that the goodness of the model $f(x)$ is assessed in terms of the closeness as a probability distribution to the true distribution $g(x)$. Akaike in [5], proposed as a measure of this closeness the use of the following *Kullback-Leibler*

*information* (K-L):

$$I(g;f) = E_G\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right], \tag{3.2}$$

where $E_G$ represents the expectation with respect to the probability distribution $G$.

In case of continuous models with densities $g(x)$ and $f(x)$, the K-L information can be expressed as

$$I(g;f) = \int_{-\infty}^{+\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx. \tag{3.3}$$

Correspondingly, for discrete models with probabilities given by $\{g(x_i);\ i = 1, 2, ...\}$ and $\{f(x_i);\ i = 1, 2, ...\}$, the K-L information can be written as

$$I(g;f) = \sum_{i=1}^{\infty} g(x_i) \log\left\{\frac{g(x_i)}{f(x_i)}\right\}. \tag{3.4}$$

By unifying the continuous and discrete models, we can express the K-L information as follows:

$$I(g;f) = \int \log\left\{\frac{g(x)}{f(x)}\right\} dG(x)$$

$$= \begin{cases} \int_{-\infty}^{+\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx, & \text{for a continuous model.} \\ \\ \sum_{i=1}^{\infty} g(x_i) \log\left\{\frac{g(x_i)}{f(x_i)}\right\}, & \text{for a discrete model.} \end{cases} . \tag{3.5}$$

### 3.1.2 Properties of K-L information

The K-L information has the following properties:

**1.** $I(g;f) \geq 0$

**2.** $I(g;f) = 0, \Longleftrightarrow g(x) = f(x)$.

In view of these properties, we consider that the smaller the quantity of K-L information, the closer the model $f(x)$ is to $g(x)$.

For the proof of the above the interested reader may refer to [51].

### 3.1.3 Measures of similarity between distributions

The following quantities have been proposed in addition to K-L information:

$$\chi^2(g;f) = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i}, \quad (\chi^2) \text{ -statistics, ( [82])},$$

$$I_K(g;f) = \int \left\{\sqrt{f(x)} - \sqrt{g(x)}\right\}^2 dx, \quad \text{Hellinger distance, ( [42]; [65])},$$

$$I_\lambda(g;f)=\frac{1}{\lambda}\int\left\{\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}g(x)dx,\quad \lambda\neq 0,\ -1,\quad \text{Generalized information, ( [29])},$$

$$D(g;f)=\int u\left(\frac{g(x)}{f(x)}\right)g(x)dx,\quad u(x)=(1-\sqrt{u})^2,\quad \text{Divergence, ( [30]; [8])},$$

$$L_1(g;f)=\int|g(x)-f(x)|dx,\quad L^1\text{-norm},$$

$$L_1(g;f)=\int\{g(x)-f(x)\}^2dx,\quad L^2\text{-norm}.$$

In the case of $D(g;f)$, letting $u(x)=\log x$ produces K-L information $I(g;f)$. Similarly, letting $u(x)=\lambda^{-1}(x^\lambda-1)$ reduces to the generalized information $I_\lambda(g;f)$. Note that when $\lambda\longrightarrow 0$, $I_\lambda(g;f)\longrightarrow I(g;f)$.

In the following section, we will extend the above concept to parametric models involving a $p$-dimensional parameter $\boldsymbol{\theta}$.

## 3.2 Information Criterion AIC

### 3.2.1 Log-Likelihood and Expected Log-Likelihood

When we build a model using data, we assume that the data $\boldsymbol{x}_n=\{x_1,x_2,...,x_n\}$ are generated according to the true distribution $G(x)$ or $g(x)$. In order to capture the structure of the given phenomena, let us assume a parametric model $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta}\in\Theta\subset R^p\}$ having a $p$-dimensional parameter $\boldsymbol{\theta}$, to be estimated via maximum likelihood. Our intention is the construction of a statistical model $f(x|\hat{\boldsymbol{\theta}})$ by replacing the unknown parameter $\boldsymbol{\theta}$ by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. Our purpose is the evaluation of the goodness or badness of the statistical model $f(x|\hat{\boldsymbol{\theta}})$ that constructed. We now consider the evaluation of a model from the standpoint of making a prediction. Our task is to evaluate the expected goodness or badness of the estimated model $f(z|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $Z=z$ generated from the unknown true distribution $g(z)$. Below we describe the K-L information in order to measure the closeness of the two distributions:

$$I\{G(z;f(z|\hat{\boldsymbol{\theta}}))\}=E_G\left[\log\left\{\frac{g(Z)}{f(Z|\hat{\boldsymbol{\theta}})}\right\}\right]$$

$$=E_G\Big[\log g(Z)\Big]-E_G\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big],\tag{3.6}$$

where the expectation is taken with respect to the unknown probability distribution G(z) by fixing $\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$.

In view of the properties of the K-L information, the larger the expected log-

likelihood

$$E_G\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big] = \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z), \tag{3.7}$$

of the model is, the closer the model is to the true one. Therefore, it is crucial in order to define the information criterion to obtain a good estimator of the expected log-likelihood. One such estimator, which is unbiased, is

$$E_G\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big] = \int \log f(z|\hat{\boldsymbol{\theta}}) d\hat{G}(z)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}), \tag{3.8}$$

in which the unknown probability distribution $G$ contained in the expected log-likelihood is replaced with an empirical distribution function $\hat{G}$. This is the log-likelihood of the statistical model $f(z|\hat{\boldsymbol{\theta}})$ or the maximum log-likelihood

$$l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}). \tag{3.9}$$

At this point it is worth mentioning that the estimator of the expected log-likelihood $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ is $n^{-1}l(\hat{\boldsymbol{\theta}})$ and that the log-likelihood $l(\hat{\boldsymbol{\theta}})$ is an estimator of $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$.

### 3.2.2 Necessity of Bias Correction for the Log-Likelihood

Usually it is difficult to precisely capture the true structure of given phenomena from a limited number of observed data. Hence, we construct various candidate statistical models based on the observed data at hand and select the model that most closely approximates the mechanism of the occurrence of the phenomenon under consideration [51]. In this subsection we consider the situation in which multiple models $\{f_j(z|\boldsymbol{\theta}_j);\ j = 1, 2, ..., m\}$ exist, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_j$ has been obtained for the parameter $\boldsymbol{\theta}_j$ of the model.

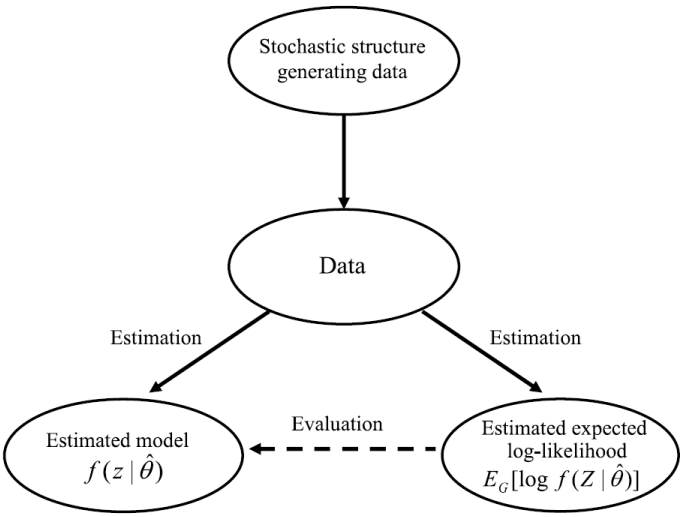


**Figure 3.2.2:** *Use of data in the estimations of the parameter of a model and of the expected log-likelihood.*

It appears that the goodness of the model specified by $\hat{\boldsymbol{\theta}}_j$, that is, the goodness of the maximum likelihood model $f_j(z|\hat{\boldsymbol{\theta}}_j)$, can be determined by comparing the magnitudes of the maximum log-likelihood $l_j(\hat{\boldsymbol{\theta}}_j)$. However with this approach the comparison of the models it is not fair, since the quantity $l_j(\hat{\boldsymbol{\theta}}_j)$, as an estimator of the expected log-likelihood $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$, contains a bias associated with the dimension of the parameter vector.

This is a result that may come in contrast to the fact that generally $l(\boldsymbol{\theta})$ is a good estimator of $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$. However, as is evident from the process by which the log-likelihood in (3.8) was derived, the log-likelihood was obtained by estimating the expected log-likelihood by reusing the data $\boldsymbol{x}_n$ that were initially used to estimate the model in place of the future data (Figure 3.2.2). If we make use of the same data twice for estimating the parameters and for estimating the evaluation measure (the expected log-likelihood) of the goodness of the fit estimated model, the bias will rise.

### 3.2.3 Relationship between log-likelihood and expected log-likelihood

The figure below (Figure 3.2.3) describes the relationship between the expected log-likelihood function and the log-likelihood function $f(x|\theta)$ with one-dimensional parameter $\theta$.

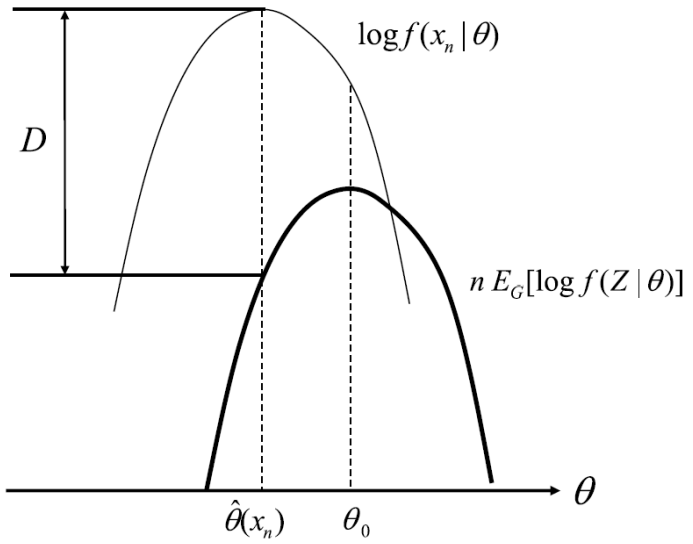


**Figure 3.2.3:** *Log-likelihood and expected log-likelihood.*

$$n\eta(\theta) = nE_G[\log f(Z|\theta)], \quad l(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta). \tag{3.10}$$

The value of $\theta$ that maximizes the expected log-likelihood is the true parameter $\theta_0$. On the other hand, the maximum likelihood estimator $\hat{\theta}(\boldsymbol{x}_n)$ is given as the maximizer of the log-likelihood function $l(\theta)$. The goodness of the model $f(z|\hat{\theta})$ defined by $\hat{\theta}(\boldsymbol{x}_n)$ should be evaluated in terms of the expected log-likelihood $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$. Therefore, the evaluation derives from using the log-likelihood $l(\hat{\theta})$ that can be calculated from the data. In such a case, as indicated in Figure 3.2.3, the true criterion should give $E_G[\log f(Z|\hat{\boldsymbol{\theta}})] \leq E_G[\log f(Z|\theta_0)]$ (see [51]). However, in the log-likelihood, the relationship $l(\hat{\theta}) \leq l(\theta_0)$ always holds.

The log-likelihood function fluctuates depending on data, and the geometry between the two functions also varies; however, the above two inequalities always hold. In case the two functions have the same form, the log-likelihood is actually interior to the extent that it appears to be better than the true model. The objective of the bias evaluation is to compensate for this phenomenon of reversal. Therefore, the prerequisite for a fair comparison of models in evaluations of and correction for the bias.

Let us assume that $n$ observations $\boldsymbol{x}_n$ generated from the true distribution $G(x)$ or $g(x)$ are realizations of the random variable $\boldsymbol{X}_n = (X_1, X_2, ..., X_n)^T$, and let

$$l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)) = \log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)), \tag{3.11}$$

represent the log-likelihood of the statistical model $f(z|\hat{\boldsymbol{\theta}})$ estimated by the maximum likelihood method. The bias of the log-likelihood as an estimator of the expected log-likelihood given in 3.7 is defined by

$$b(G) = E_{G(\boldsymbol{x}_n)}\Big[\log f(\boldsymbol{X_n}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\Big]\Big], \tag{3.12}$$

where the expectation $E_{G(\boldsymbol{x}_n)}$ is taken with respect to the joint distribution $\prod_{a=1}^{n} G(x_i) =$

$G(\boldsymbol{x}_n)$, of the sample $\boldsymbol{X}_n$, and $E_{G(z)}$ is the expectation on the true distribution $G(z)$. Hence, we can construct the general form of the information criterion (IC) by evaluating the bias and correcting for the bias of the log-likelihood as follows:

$$IC(\boldsymbol{X}_n; \hat{G}) = -2(\text{log-likelihood of the statistical model} - \text{bias estimator})$$

$$= -2\sum_{i=1}^{n} \log f(X_i|\hat{\boldsymbol{\theta}}) + 2\{\text{estimator for b(G)}\}. \tag{3.13}$$

Depending on the relationship between the true distribution generating the data and the specified model and on the method employed to construct a statistical model, the bias $b(G)$ can take various forms.

### 3.2.4 Derivation of Bias of the Log-Likelihood

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is given as the $p$-dimensional parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood function $l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta})$ or by solving the likelihood equation

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta}) = 0. \tag{3.14}$$

By taking the expectation, we obtain

$$E_{G(\boldsymbol{x}_n)}\Big[\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta})\Big] = nE_{G(z)}\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z|\boldsymbol{\theta})\Big]. \tag{3.15}$$

For a continuous model, if $\boldsymbol{\theta}_0$ is a solution of the equation

$$E_{G(z)}\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z|\boldsymbol{\theta})\Big] = \int g(z) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(z|\boldsymbol{\theta}) dz = 0, \tag{3.16}$$

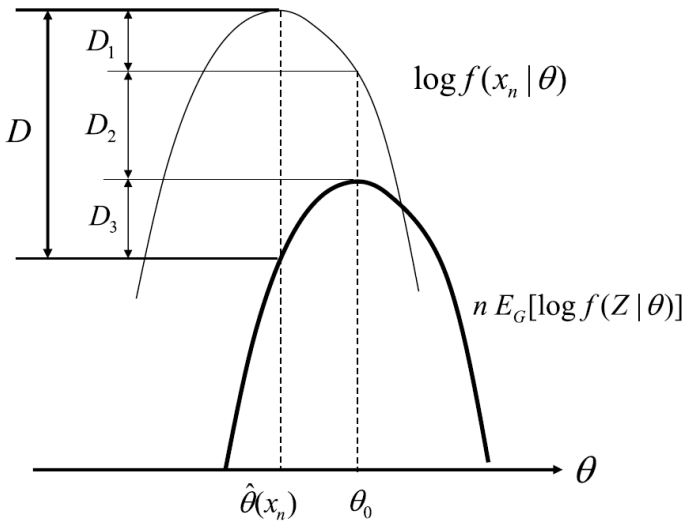


**Figure 3.2.4:** *Decomposition of the bias term.*

it can be shown that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ when $n \longrightarrow +\infty$.
For a discrete model see [51].

We can now evaluate the bias based on the above results:

$$b(G) = E_{G(\boldsymbol{x}_n)}\Bigg[\log f(\boldsymbol{X_n}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\Big]\Bigg], \tag{3.17}$$

when the expected log-likelihood is estimated using the log-likelihood of the statistical model. To this end, we first decompose the bias as follows (Figure 3.2.4):

$$b(G) = E_{G(\boldsymbol{x}_n)}\Bigg[\log f(\boldsymbol{X_n}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\Big]\Bigg]$$

$$= E_{G(\boldsymbol{x}_n)}\Bigg[\log f(\boldsymbol{X_n}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - \log f(\boldsymbol{X_n}|\boldsymbol{\theta}_0)\Bigg]$$

$$+E_{G(\boldsymbol{x}_n)}\Bigg[\log f(\boldsymbol{X_n}|\boldsymbol{\theta}_0) - nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big]\Bigg] \tag{3.18}$$

$$+E_{G(\boldsymbol{x}_n)}\Bigg[nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big] - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\Big]\Bigg]$$

$$= D_1 + D_2 + D_3.$$

We now calculate separately the three expectations $D_1$, $D_2$, and $D_3$.

### 1. Calculation of $\mathbf{D}_2$

The simplest case is the evaluation of $D_2$ because it does not contain an estimator. It can be seen that

$$D_2 = E_{G(\boldsymbol{x}_n)}\Bigg[\log f(\boldsymbol{X_n}|\boldsymbol{\theta}_0) - nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big]\Bigg]$$

$$D_2 = E_{G(\boldsymbol{x}_n)}\Bigg[\sum_{i=1}^{n}\log f(X_i|\boldsymbol{\theta}_0)\Bigg] - nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big]$$

$$= 0. \tag{3.19}$$

This means that in Fig.(3.2.4), although $D_2$ varies randomly depending on the data, its expectation becomes 0.

**2. Calculation of $D_3$**

First, we write

$$\eta(\hat{\boldsymbol{\theta}}) = E_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big]. \tag{3.20}$$

By performing a Taylor series expansion of $\eta(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ given a solution to (3.16), we obtain

$$\eta(\hat{\boldsymbol{\theta}}) = \eta(\boldsymbol{\theta}_0) + \sum_{i=1}^{p}(\hat{\theta}_i - \theta^(0)_i)\frac{\partial\eta(\boldsymbol{\theta}_0)}{\partial\theta_i} \tag{3.21}$$

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\theta}_i - \theta^(0)_i)(\hat{\theta}_j - \theta^(0)_j)\frac{\partial^2\eta(\boldsymbol{\theta}_0)}{\partial\theta_i\theta_j} + \cdots,$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta}_p)^T$ and $\boldsymbol{\theta}_0 = (\theta_1^{(0)}, \theta_2^{(0)}, ...\theta_p^{(0)})^T$. Due to the fact that $\hat{\boldsymbol{\theta}}$ is a solution of (3.16), it holds that

$$\frac{\partial\eta(\boldsymbol{\theta}_0)}{\partial\theta_i} = E_{G(z)}\left[\frac{\partial}{\partial\theta_i}\log f(Z|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = 0, \quad i = 1, 2, ..., p. \tag{3.22}$$

Hence, (3.21) can be approximated as

$$\eta(\hat{\boldsymbol{\theta}}) = \eta(\boldsymbol{\theta}_0) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{3.23}$$

where $J(\boldsymbol{\theta}_0)$ is the $p \times p$ matrix given by

$$J(\boldsymbol{\theta}_0) = -E_{G(z)}\left[\frac{\partial^2\log f(Z|\boldsymbol{\theta})}{\partial\theta\partial\theta^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -\int g(z)\frac{\partial^2\log f(Z|\boldsymbol{\theta})}{\partial\theta\partial\theta^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}dz, \tag{3.24}$$

such that the $\boldsymbol{a}^{th}$ and the $\boldsymbol{b}^{th}$ elements are given by

$$j_{ab} = -E_{G(z)}\left[\frac{\partial^2\log f(Z|\boldsymbol{\theta})}{\partial\theta_a\partial\theta_b}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -\int g(z)\frac{\partial^2\log f(Z|\boldsymbol{\theta})}{\partial\theta_a\partial\theta_b}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}dz. \tag{3.25}$$

Since $D_3$ is the expectation of $\eta(\boldsymbol{\theta}_0) - \eta(\hat{\boldsymbol{\theta}})$ with respect to $G(\boldsymbol{x}_n)$ we obtain approximately

$$D_3 = E_{G(\boldsymbol{x}_n)}\Big[nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big] - nE_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big]\Big]$$

$$= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big]$$

$$= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\Big[tr\Big\{J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\Big\}\Big] \tag{3.26}$$

$$= \frac{n}{2} tr\left\{ J(\boldsymbol{\theta}_0) E_{G(\boldsymbol{x}_n)}\left[ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \right] \right\}.$$

By substituting the (asymptotic) variance covariance matrix

$$E_{G(\boldsymbol{x}_n)}\left[ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \right] = \frac{1}{n} J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0) - 1, \tag{3.27}$$

of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ into (3.26), we have

$$D_3 = \frac{1}{2} tr\{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \}, \tag{3.28}$$

where $J(\boldsymbol{\theta}_0)$ is given in (3.24) and $I(\boldsymbol{\theta}_0)$ is the $p \times p$ matrix given by

$$I(\boldsymbol{\theta}_0) = E_{G(z)}\left[ \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right]$$

$$= \int g(z) \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} dz. \tag{3.29}$$

### 3. Calculation of $D_1$

If we write $l(\boldsymbol{\theta}) = \log f(\boldsymbol{X}_n|\boldsymbol{\theta})$ and apply a Taylor series expansion around the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we obtain

$$l(\boldsymbol{\theta}) = l(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots. \tag{3.30}$$

The quantity $\hat{\boldsymbol{\theta}}$ satisfies the equation $\frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = 0$ by advantage of the maximum likelihood estimator given as a solution of the likelihood equation $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$.

The quantity below converges in probability to $J(\boldsymbol{\theta}_0)$ in (3.24) when $n \longrightarrow +\infty$.

$$\frac{1}{n} \frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{1}{n} \frac{\partial^2 \log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{3.31}$$

This can be derived from the fact that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and from:

$$-\frac{1}{n} \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(x_i|\boldsymbol{\theta}) \bigg|_{\boldsymbol{\theta}_0} \longrightarrow J(\boldsymbol{\theta}_0),$$

where $\Big|_{\boldsymbol{\theta}_0}$ is the value of the derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

Using the results above we can obtain the following approximation for (3.30):

$$l(\boldsymbol{\theta}_0) - l(\hat{\boldsymbol{\theta}}) \backsimeq -\frac{n}{2} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0) (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}). \tag{3.32}$$

Based on (3.32) and the asymptotic variance covariance matrix (3.27) of the maximum likelihood estimator, $D_1$ can be calculated approximately as follows:

$$D_1 = E_{G(\boldsymbol{x}_n)}\Big[\log f(\boldsymbol{X_n}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - \log f(\boldsymbol{X_n}|\boldsymbol{\theta}_0)\Big]$$

$$= \frac{n}{2} E_{G(\boldsymbol{x}_n)}\Big[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\Big]$$

$$= \frac{n}{2} E_{G(\boldsymbol{x}_n)}\Big[tr\{J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T\}\Big] \tag{3.33}$$

$$= \frac{n}{2} tr\Bigg\{J(\boldsymbol{\theta}_0) E_{G(\boldsymbol{x}_n)}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\Big]\Bigg\}$$

$$= \frac{1}{2} tr\{I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}\}.$$

Thus, combining (3.19), (3.28), and (3.33), the bias resulting from the estimation of the expected log-lkelihood using the log-likelihood of the model is asymptotically obtained as

$$b(G) = D_1 + D_2 + D_3$$

$$= \frac{1}{2} tr\{I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}\} + 0 + \frac{1}{2} tr\{I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}\} \tag{3.34}$$

$$= \frac{1}{2} tr\{I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}\},$$

where $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are respectively given in (3.29) and (3.24).

### 3.2.5 Akaike Information Criterion (AIC)

The Akaike Information Criterion is one of the most useful criteria. It has played a significant role on solving problems in a wide variety of fields as a model selection criterion for analyzing actual data.
The AIC(p) is defined by

$$AIC(p) = -2(\text{maximum log-likelihood}) + 2(\text{p}), \tag{3.35}$$

where $p$ refers to the dimension of the parameter vector $\boldsymbol{\theta}$ contained in the specified model $f(x|\boldsymbol{\theta})$.

The AIC is an evaluation criterion for the badness of the model. We can estimate its parameters by the maximum likelihood method and it indicates that the bias of the log-likelihood approximately becomes the "number of free parameters contained in the model". The bias can be derived under the assumption that the true distribution $g(x)$ is included in the specified parametric model $\{f(x|\boldsymbol{\theta})\ \boldsymbol{\theta} \in \Theta \subset R^p\}$, there exists a $\boldsymbol{\theta}_0 \in \Theta$ such that the equality $g(x) = f(x|\boldsymbol{\theta}_0)$ holds.

At this point we also assume that the parametric model is $\{f(x|\boldsymbol{\theta})\ \boldsymbol{\theta} \in \Theta \subset R^p\}$ and that the true distribution $g(x)$ can be expressed as $g(x) = f(x|\boldsymbol{\theta}_0)$ for properly specified $\boldsymbol{\theta}_0 \in \Theta$. Hence the equality $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ holds for the $p \times p$ matrix $J(\boldsymbol{\theta}_0)$ and the $p \times p$ matrix $I(\boldsymbol{\theta}_0)$ given in 3.24 and 3.29 respectively. Therefore the bias of the log-likelihood is asymptotically given by

$$E_{G(X_n)} = \left[\sum_{i=1}^{n} \log f(X_i|\hat{\boldsymbol{\theta}}) - nE_{G(z)} \log f(Z|\hat{\boldsymbol{\theta}})\right]$$

$$= tr\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\} \equiv tr(I_p) = p, \tag{3.36}$$

where $I_p$ is the identity matrix of dimension $p$. Thus, the AIC is given by

$$\text{AIC(p)} = -2\sum_{i=1}^{n} \log f(X_i|\hat{\boldsymbol{\theta}}) + 2p, \tag{3.37}$$

which is the corrected asymptotic bias $p$ of the log-likelihood.

The AIC does not require any analytical derivation of the bias correction terms for individual problems and does not depend on the unknown probability distribution G, which removes variations due to the estimation of the bias. Moreover, Akaike in [5] states that if the true distribution that generated the data exists near the specified parametric model, the bias associated with the log-likelihood of the model based on the maximum likelihood method can be approximated by the number of parameters. These aspects make the AIC a very flexible technique from a practical point of view and explain is great popularity across scientific disciplines.

Findley and Wei in [34], provided a derivation of AIC and its asymptotic properties for the case of vector time series regression model (see also [35], [13]). For more details see [59], [87], [16], [50], [2], [68], [52] and [53].

# 3.3 Information Criterion BIC

## 3.3.1 Bayesian Model Evaluation Criterion

### 3.3.1.1 Definition of BIC

The Bayesian or Schwartz's information criterion proposed by Schwarz in [89], is an evaluation criterion for models in terms of their posterior probability (see [6]).

Let $M_1, M_2, ..., M_r$ be $r$ candidate models, and assume that each model $M_i$ is characterized by a parametric distribution $f_i(x|\boldsymbol{\theta}_i)$ $(\boldsymbol{\theta}_i \in \Theta_i \subset R^{k_i})$ and the prior distribution $\pi_i(\boldsymbol{\theta}_i)$ of the $k_i$-dimensional parameter vector $\boldsymbol{\theta}_i$. Suppose that $n$ observations $\boldsymbol{x}_n = \{x_1, x_2, ..., x_n\}$ are given. Then the marginal distribution or probability

of $\boldsymbol{x}_n$ for the $\boldsymbol{i}^{th}$ model $M_i$, is given by

$$p_i(\boldsymbol{x}_n) = \int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i, \tag{3.38}$$

which represents the *the marginal likelihood* of the data.

Based on Bayes' theorem, if $P(M_i)$ is the prior probability of the $\boldsymbol{i}^{th}$ model, then the corresponding posterior probability of the $\boldsymbol{i}^{th}$ model is given by

$$P(M_i|\boldsymbol{x}_n) = \frac{p_i(\boldsymbol{x}_n)P(M_i)}{\sum_{j=1}^{r} p_j(\boldsymbol{x}_n)P(M_j)}, \quad i = 1, 2, ..., r. \tag{3.39}$$

The above probability represents the probability of the data being generated from the $\boldsymbol{i}^{th}$ model when data are observed. Hence, if a model is to be selected from the class of $r$ models, that will be the one with the largest posterior probability. Since all models share the same denominator in (3.39), the model that maximizes the numerator $p_i(\boldsymbol{x}_n)P(M_i)$ must be selected.

In addition, if we assume that the prior probabilities $P(M_i)$ are equal in all models, it follows that the model that maximizes the marginal likelihood $p(\boldsymbol{x}_n)$ of the data must be selected.

The BIC is defined as

$$-2\log p_i(\boldsymbol{x}_n) = -2\log\left\{\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right\}$$

$$\simeq -2\log f_i(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_i) + k_i\log n, \tag{3.40}$$

where $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator of the $k_i$-dimensional parameter vector $\boldsymbol{\theta}_i$ of the model $f_i(x|\boldsymbol{\theta}_i)$.
Consequently, from the class of r models that are to be evaluated using the maximum likelihood method, the one that minimizes the value of BIC can be selected as the optimal for the data.

#### 3.3.1.2 Bayes factors

Consider for comparative purposes, models $M_1$ and $M_2$. When the data produce the posterior probabilities the posterior odds in favor of $M_1$ against $M_2$ are

$$\frac{P(M_1|\boldsymbol{x}_n)}{P(M_2|\boldsymbol{x}_n)} = \frac{p_1(\boldsymbol{x}_n)P(M_1)}{p_2(\boldsymbol{x}_n)P(M_2)}. \tag{3.41}$$

Then the ratio

$$\frac{p_1(\boldsymbol{x}_n)}{p_2(\boldsymbol{x}_n)} = \frac{\int f_1(\boldsymbol{x}_n|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int f_2(\boldsymbol{x}_n|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2}, \tag{3.42}$$

is defined as *the Bayes factor*.

Model comparisons based on the AIC are asymptotically equivalent to those based on Bayes factors, [4]. Moreover, Kass and Raftery in [49], commented that

from a Bayes point of view this is true only if the precision of the prior is comparable to that of the likelihood, but not in the more usual situation where prior information is limited relative to the information provided by the data.
For more details on Bayes factors the reader may refer to [49], [77], and [12].

### 3.3.2 Derivation of the BIC

Let us represent the marginal likelihood of (3.38) as

$$p(\boldsymbol{x}_n) = \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3.43}$$

where $\boldsymbol{\theta}$ is a $p$-dimensional parameter vector. Therefore, (3.43) can be rewritten as

$$p(\boldsymbol{x}_n) = \int exp\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \Rightarrow$$

$$p(\boldsymbol{x}_n) = \int exp\{l(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3.44}$$

where $l(\boldsymbol{\theta})$ is the log-likelihood function $l(\boldsymbol{\theta}) = \log f(\boldsymbol{x}_n|\boldsymbol{\theta})$.

### 3.3.3 Bayesian information criterion (BIC)

Let $f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ be a statistical model estimated by the maximum likelihood method. The Bayesian Information Criterion is given by

$$\text{BIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + p\log n. \tag{3.45}$$

The BIC is an evaluation criterion for models estimated by using the maximum likelihood method and that the criterion is obtained under the condition that the sample size $n$ is made large enough. The quantity in (3.45) was obtained by approximating the marginal likelihood associated with the posterior probability of the model by Laplace's method for integrals. For a more detailed discussion on Laplace's method for integrals see [51].

## 3.4 Information Criterion MDIC

The criteria discussed in the previous three sections are the most popular ones which are based on the log-likelihood function. An alternative class of criteria is based on measures of divergence or distance. Mantalos et. al in [62], proposed an improvement on the Divergence Information Criterion (DIC) [63] called the Modified Divergence Information Criterion (MDIC).

### 3.4.1 The development of the MDIC Criterion

Basu et al. in [11], proposed the Basu-Harris-Hjort-Jones (BHHJ) measure of divergence, which is indexed by a positive parameter $\alpha$, and defined as:

$$I^{\alpha}(g, f_{\theta}) = \int \Big\{ f_{\theta}^{1+\alpha}(z) - \Big(1 + \frac{1}{\alpha}\Big) g(z) f_{\theta}^{\alpha}(z) + \Big(\frac{1}{\alpha}\Big) g^{1+\alpha}(z) \Big\} dz. \tag{3.46}$$

This family of measures was proposed by Basu et al. in [11], for the development of a minimum divergence estimation method for robust parameter estimation.
The index $\alpha$ controls the trade-off between robustness and asymptotic efficiency of the parameter estimators that are the values of $\theta$ that minimize the measure over a parametric space $\mathbf{\Theta}$. The BHHJ family reduces to the Kullback-Leibler measure for $\alpha \downarrow 0$ (see [64]) and to the square of the standard $L_2$ distance between the candidate and the true model for $\alpha = 1$. Mattheou et al. in [63] developed a new criterion, the Divergence Information Criterion (DIC), by applying the same methodology mentioned above and using BHHJ measure in place of K-L information and log-likelihood.

Let us suppose a set of observations $\{x_1, x_2, ..., x_n\}$. Then the DIC is derived by

$$\mathrm{DIC(p)} = nQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}} (1+\alpha)^{2+\frac{p}{2}} p, \tag{3.47}$$

where $Q_{\hat{\theta}} = \int \Big\{ f_{\theta}^{1+\alpha}(z)dz - \Big(1 + \frac{1}{\alpha}\Big) \frac{1}{n} \sum_{i=1}^{n} f_{\hat{\theta}}^{\alpha}(x_i) \Big\}$ and $\hat{\theta}$ is a consistent and asymptotically normal estimator of $\theta$.

Although, the DIC criterion in preliminary simulations studies for regression models [64] showed a very good medium sample size performance for values of $\alpha$ close to zero, it is not computationally attractive for practitioners, primarily due to the calculation of the first term $Q_{\hat{\theta}}$, namely, the integral $\int f_{\hat{\theta}}^{1+\alpha}(z)dz$. Moreover, a simulation study shows that the difference in the calculation of the above integral for the different candidate models is negligible compared with the difference in the calculation for the entire quantity $Q_{\hat{\theta}}$. Hence, the integral term does not affect the selected model and therefore the criterion can be properly modified. Thus, Mantalos et al. in [62], proposed a modified criterion called the Modified Divergence Information Criterion (MDIC), which is given by

$$\mathrm{MDIC(p)} = n^{*} MQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}} (1+\alpha)^{2+\frac{p}{2}} p, \tag{3.48}$$

where $MQ_{\hat{\theta}} = -[\Big(1 + \frac{1}{\alpha}\Big) \frac{1}{n} \sum_{i=1}^{n} f_{\hat{\theta}}^{\alpha}(x_i)]$.

Note that a model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity that measures the distance between the true unknown model and a fitted approximating model.
If we choose the model with the smallest estimator of the expected overall discrepancy, we may end up with a selection with an unnecessarily large order. In that sense, MDIC is a criterion comparable to AIC since they are both provide unbiased estimates of the expected overall discrepancy relying on two different measures of divergence of information, namely BHHJ and K-L.

### 3.4.2 Optimal choice for the index $\alpha$

For practical purposes we have to decide the optimal choice of the positive index $\alpha$. Thus, Mantalos et al. in [62] simulated a 100 observation series for five differ-

ent models with $\alpha \in [0.01,\ 0.5]$. In the figure below the power of the selection is provided according to the specified value of the index $\alpha$.

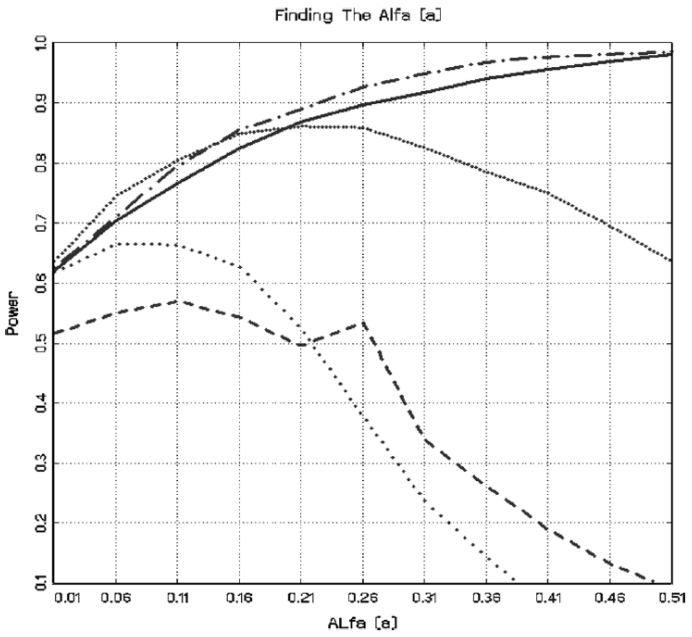


**Figure 3.4.2:** *Optimal choice of the index* $\alpha$. $AR(1) - \cdot -$, $AR(2)$—, $AR(3) \cdots$, $AR(4)$- - -, $AR(5) \cdot \ \cdot \ \cdot$.

- $AR(1):\ x_t = 1 + 0.65x_{t-1} + \epsilon_t$
- $AR(2):\ x_t = 1 + 1.5x_{t-1} - 0.5x_{t-2} + \epsilon_t$
- $AR(3):\ x_t = 1 + 0.2x_{t-1} + 0.5x_{t-2} - 0.35x_{t-3} + \epsilon_t$
- $AR(4):\ x_t = 1 + 0.2x_{t-1} + 0.5x_{t-2} - 0.35x_{t-3} - 0.22x_{t-4} + \epsilon_t$
- $AR(5):\ x_t = 1 + 0.23x_{t-1} - 0.22x_{t-3} - 0.45x_{t-5} + \epsilon_t$

Figure 3.4.2 shows that the power increases as the value of $\alpha$ increases for small lags (models $AR(1)$ and $AR(2)$). On the other hand, for lags $\geq 3$ the power increases up to a value of $\alpha$ and then decreases. Mantalos et al. in [62], concluded that and optimal index $\alpha$ value equals 0.25 since it appears to serve a fair balance between small and large lag models.

For more details on MDIC and the simulations made, see [62].

## 3.5 Coefficient of determination ($R^2$) for GLMMs

Using both linear (LMs) and generalized linear mixed-effects models (GLMMs) has become a common trend in social, medical, biological, ecology and evolution sciences. Information criteria, such as AIC and BIC, are usually employed as model comparison tools for mixed-effects models. On the other hand, the use of "variance explained" $R^2$ as a relevant summarizing statistic for mixed-effects models, is rare,

in contrast to LMs and GLMs. Coefficient of determination ($R^2$) is a powerful statistical tool, which in contrast to AIC, BIC, MDIC etc., provides an absolute value for the goodness of fit of a model.
One of the main reasons for the under-estimation of $R^2$ for mixed-effects models lies in the fact that $R^2$ can be defined in many ways. In addition, most definitions of $R^2$ for mixed-effects have serious theoretical problems (e.g., decreased or negative ($R^2$) values in larger models). Moreover, their use is constrained by practical difficulties (e.g. implementation).
Keeping in mind the problems mentioned above, we are going to present a method for calculating $R^2$ for GLMMs as proposed by Nakagawa et al. in [73].

### 3.5.1 Definitions of $R^2$

A standard (general) linear model (LM) can be written as:

$$y_i = \beta_0 + \sum_{h=1}^{p} \beta_h x_{hi} + \epsilon_i, \tag{3.49}$$

$$\epsilon_i \sim Gaussian(0, \sigma_\epsilon^2),$$

where $y_i$ is the $\boldsymbol{i}^{th}$ value for the $\boldsymbol{h}^{th}$ predictor, $\beta_0$ is the intercept, $\beta_h$ is the slope (regression coefficient) of the $\boldsymbol{h}^{th}$ predictor, $\epsilon_i$ is the $\boldsymbol{i}^{th}$ residual value and residual errors are normally (Gaussian) distributed with a variance of $\sigma_\epsilon^2$. Regression models of this type are fitted by ordinary least squares (OLS) methods that minimize the sum of squared distances between observed and fitted responses. The residual sum of squares appears in the formulation of the empirical definition for the coefficient of determination $R^2$ ( [57]; [33]).

$$R_O^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.50}$$

$$\hat{y} = \hat{\beta}_0 + \sum_{h=1}^{p} \hat{\beta}_h x_{hi},$$

where $n$ is the number of observations, $\bar{y}$ is the mean of the response, $\hat{y}_i$ is the $\boldsymbol{i}^{th}$ fitted response value, $\hat{\beta}_0$ and $\hat{\beta}_h$ are estimates of $\beta_0$ and $\beta_h$, respectively, and the subscript "$O$" in $R_O^2$, signifies OLS regression.
An interesting feature to note here, is that the definition of "variance explained" is rather indirectly composed of 1 minus the "variance unexplained". An equivalent quantity of $R_O^2$ is described as follows:

$$R_O^2 = 1 - \frac{var(y_i - \hat{y}_i)}{var(y_i)}, \tag{3.51}$$

$$R_O^2 = 1 - \frac{\sigma_\epsilon^2}{var(y_i)}, \tag{3.52}$$

where "var(·)" indicates the variance.
The quantity in (3.52) can also be expressed as the ratio between the residual variance of the model of interest and the residual variance of the null model (also referred as the empty model or the intercept model [73] described as:

$$R_O^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_{\epsilon 0}^2}, \tag{3.53}$$

where $\sigma_{\epsilon 0}^2$ is the residual variance of the null model.

Let us consider the case of generalizing the definition of $R_0^2$ to GLMMs. In such a case, there are two difficulties to deal with. Firstly, in the case of generalization to non-linear response variables (i.e. GLMs), it is not straightforward to get an appropriate estimate of the residual variance. Secondly, for the generalized class of mixed effects models that consist of error terms at different hierarchical levels, it is not obvious from the beginning which estimate should be used for the unexplained variance.
Thus, $R^2$ can be defined using the maximum likelihood of the full and null models [61].

$$R_g^2 = 1 - \left(\frac{L_0}{L_\beta}\right)^{\frac{2}{n}}, \tag{3.54}$$

where $L_\beta$ is the likelihood of the data given the fitted model of interest, and $L_0$ is the likelihood of the data given the null model, $n$ is the total sample size, the subscript "$g$" in $R_g^2$, signifies "general" (see [69]). The problem with $R_g^2$ is that it cannot become 1 even if the model of interest fits data perfectly. Nagelkerke in [72] proposed an adjustment to equation (3.54) to deal with this problem described as follows:

$$R_G^2 = \frac{\left[1 - \left(\frac{L_0}{L_\beta}\right)^{\frac{2}{n}}\right]}{\left[1 - (L_0)^{\frac{2}{n}}\right]}, \tag{3.55}$$

where the denominator can be interpreted as the maximum possible value of $R_g^2$ and the subscript "$G$" in $R_G^2$, signifies "General". A definition of $R^2$ comparable to this of $R_G^2$ is:

$$R_D^2 = 1 - \frac{-2\log(L_\beta)}{-2\log(L_0)}, \tag{3.56}$$

where "$D$" signifies "Deviance".
Note that $-2$ is left in both denominator and numerator, so that $R_D^2$ can be compared with (3.50). For a linear model, the term "$-2$ log-likelihood statistic" (sometimes referred to as deviance) is equal to the residual sum of squares based on OLS of this model [69]. For other likelihood-based definitions of $R^2$, see [21] and [69].

### 3.5.2 Common problems of generalizing $R^2$

Suppose we have an experimental design for which we repeatedly sample from the same set of individuals. In the case of generalizing the LM described in (3.49), we can fit a LMM with one random factor as:

$$y_{ij} = \beta_0 + \sum_{h=1}^{p} \beta_h x_{hij} + \alpha_j + \epsilon_{ij}, \tag{3.57}$$

$$\alpha_j \sim Gaussian(0, \sigma_\alpha^2),$$

$$\epsilon_{ij} \sim Gaussian(0, \sigma_\epsilon^2),$$

where $y_{ij}$ is the $\boldsymbol{i}^{th}$ response of the $\boldsymbol{j}^{th}$ individual, $x_{hij}$ is the $\boldsymbol{i}^{th}$ value of the $\boldsymbol{j}^{th}$ individual for the $\boldsymbol{h}^{th}$ predictor, $\beta_0$ is the intercept, $\beta_h$ is the slope (regression coefficient) of the $\boldsymbol{h}^{th}$ predictor, $\alpha_j$ is the individual-specific effect from a normal distribution of individual-specific effects with mean of the zero and variance of $\sigma_\alpha^2$ (between-individual variance) and $\epsilon_{ij}$ is the residual associated with the $\boldsymbol{i}^{th}$ value of the $\boldsymbol{j}^{th}$ individual from a normal distribution of residual with mean of zero and variance of $\sigma_\epsilon^2$ (within-individual variance). As is indicated the previous equation (3.57), LMMs have by default more than one variance component, ($\sigma_\alpha^2$ and $\sigma_\epsilon^2$) while LMs have only one (see 3.49).

A common definition of $R^2$ for mixed-effects models is based on the reduction of each variance component when including fixed-effect predictors separately, which means that there will be a separate $R^2$ for each random effect and the residual variance, respectively ( [85]; [20]). However, Snijders and Bosker in [94], pointed out that it is not uncommon that some predictors can reduce $\sigma_\epsilon^2$ while at the same time can increase $\sigma_\alpha^2$, and *vice versa*, even though the total sum of variance components ($\sigma_\epsilon^2 + \sigma_\alpha^2$) is usually reduced (see [94]). Variance components of this type oftenly can result in negative $R^2$. This happens due to the fact that $\sigma_\epsilon^2$ and $\sigma_\alpha^2$ can be larger than $\sigma_{\epsilon 0}^2$ and $\sigma_{\alpha 0}^2$, respectively.

In this respect, Snijders and Bosker in [94] proposed what they refer to as $R_1^2$ and $R_2^2$ for LMMs with one random factor (as in (3.57)) in order to avoid this problem. Thus, one $R^2$ value is calculated for each level of a LMM. $R_1^2$ can be expressed in two forms described as follows:

$$R_1^2 = 1 - \frac{var(y_{ij} - \hat{y}_{ij})}{var(y_{ij})}, \tag{3.58}$$

$$\hat{y}_{ij} = \hat{\beta}_0 + \sum_{h=1}^{p} \hat{\beta}_h x_{hij}, \tag{3.59}$$

$$R_1^2 = 1 - \frac{\sigma_\epsilon^2 + \sigma_\alpha^2}{\sigma_{\epsilon 0}^2 + \sigma_{\alpha 0}^2}, \tag{3.60}$$

where $R_1^2$ is the variance explained at the unit of analysis, $\hat{y}_{ij}$ is the $\boldsymbol{i}^{th}$ fitted value for $\boldsymbol{j}^{th}$ individual and the other notations are as above.
$R_2^2$ can be written in a similar manner as:

$$R_2^2 = 1 - \frac{var(\bar{y}_j - \hat{\bar{y}}_j)}{var(\bar{y}_j)}, \tag{3.61}$$

$$R_2^2 = 1 - \frac{\sigma_\epsilon^2 + \frac{\sigma_\alpha^2}{k}}{\sigma_{\epsilon 0}^2 + \frac{\sigma_{\alpha 0}^2}{k}}, \tag{3.62}$$

where $k = \frac{M}{\sum_{j=1}^{M} \frac{1}{m_j}}$, $R_2^2$ is the variance explained at the individual level, $\bar{y}_j$ is the mean observed value for the $\boldsymbol{j}^{th}$ individual, $\hat{\bar{y}}_j$ is the fitted value for $\boldsymbol{j}^{th}$ individual, $k$ is the harmonic mean of the number of replicates per individuals, $m_j$ is the number of replicates for the $\boldsymbol{i}^{th}$ individual, $M$ is the total number of individuals, and the other notations are as above.

The use of $R_1^2$ and $R_2^2$ has the advantage of calculating how much variance is explained at each level of the analysis. However, there are at least three problems to deal with in this approach:

**1.** $R_1^2$ and $R_2^2$ can decrease in larger models.

**2.** It is not clear how $R_1^2$ and $R_2^2$ can be extended to more than two levels.

**3.** It is not clear how $R_1^2$ and $R_2^2$ can be generalized to GLMMs.

The first problem occurs due to the fact that $\sigma_\epsilon^2 + \sigma_\alpha^2$ of a model with more predictors can be larger than that of a model of fewer predictors, and $R_1^2$ and $R_2^2$ could also take negative values [94]. In other words, the estimate of $(\sigma_\epsilon^2 + \sigma_\alpha^2)$ can be larger than that of $(\sigma_{\epsilon 0}^2 + \sigma_{\alpha 0}^2)$.
There are two explanations for the decreasing or/and negative values of $R^2$ in a larger model:

**1.** Random fluctuation that is most prominent when the sample size is small.

**2.** Misspecification of the model, when the new predictor is redundant in relation to one or more predictors.

It is suggested that $R_1^2$ and $R_2^2$ can be used as a diagnostic measure in model selection. However, such misspecification is not necessarily the cause of an increase in the quantity of $(\sigma_\epsilon^2 + \sigma_\alpha^2)$.

The second problem was addressed by Gelman and Pardoe in [38], which provided a solution of extending $R_1^2$ and $R_2^2$ to any arbitrary numbers of levels (or random factors) in a Bayesian framework. For more information, the interested reader may refer [38].

The third problem of generalizing $R_1^2$ and $R_2^2$ is considered as profound, since the residual variance $\sigma_\epsilon^2$, cannot be easily defined for non-Gaussian responses. For more details see [73].

### 3.5.3 General and simple $R^2$ for GLMMs

As aforementioned, the variance explained $R_O^2$ is defined using the variance unexplained by the model, and ($R_O^2$) can be redefined more directly in terms of variance explained as:

$$R_O^2 = \frac{\sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.63}$$

$$R_O^2 = \frac{var(\hat{y}_i)}{var(y_i)}, \tag{3.64}$$

for which the notations are identical with those in (3.50) and (3.52).

This direct formulation, will now be extended to LMMs and GLMMs.

- **LMMs**

  Let us assume a LMM with two random effects, namely $\gamma_k$ corresponding to "groups" (with individuals uniquely assigned to groups), $k = 1, ..., K$ and $a_{jk}$ corresponding to individuals within each group, $j = 1, ..., J$ (with multiple observations per individual) respectively. Hence, observations (denoted below by the index "$i$") are clustered in individuals (denoted below by the index "$j$") and individuals are nested within groups (denoted below by the index "$k$"), $i = 1, ..., n$, $j = 1, ..., J$, $k = 1, ..., K$, (see [88]).
  The model in such case can be written as:

$$y_{ijk} = \beta_0 + \sum_{h=1}^{p} \beta_h x_{hijk} + \gamma_k + \alpha_{jk} + \epsilon_{ijk}, \tag{3.65}$$

$$\gamma_k \sim Gaussian(0, \sigma_\gamma^2),$$

$$\alpha_{jk} \sim Gaussian(0, \sigma_\alpha^2),$$

$$\epsilon_{ijk} \sim Gaussian(0, \sigma_\epsilon^2),$$

  where $y_{ijk}$ is the $\boldsymbol{i}^{th}$ response (observation), $i = 1, ..., n$ of the $\boldsymbol{j}^{th}$ individual, $j = 1, .., J$, belonging to the $\boldsymbol{k}^{th}$ group, $k = 1, ..., K$, $x_{hijk}$ is the $\boldsymbol{i}^{th}$ value of the $\boldsymbol{j}^{th}$ individual in the $\boldsymbol{k}^{th}$ group for the $\boldsymbol{h}^{th}$ predictor, $h = 1, ..., p$, $\gamma_k$ is the group-specific effect from a normal distribution of group-specific effects with mean of zero and variance of $\sigma_\gamma^2$, $\alpha_{jk}$ is the individual-specific effect from a normal distribution of individual-specific effects with mean of zero and variance of $\sigma_\alpha^2$ and $\epsilon_{ijk}$ is the residual from a normal distribution of group-specific effects with mean of zero and variance of $\sigma_\epsilon^2$.
  An $R^2$ for LMM given by equation (3.65) can be defined as:

$$R^2_{LMM(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}, \tag{3.66}$$

where "$m$" stands for *marginal*, and

$$\sigma_f^2 = var\left(\sum_{h=1}^{p} \beta_h x_{hijk}\right), \tag{3.67}$$

where $\sigma_f^2$ is the variance calculated from the fixed effect components of the LMM [93].
The estimation of $\sigma_f^2$ can be derived by predicting fitted values based on the fixed effects alone followed by calculating the variance of these fitted values. Note that $\sigma_f^2$ should be estimated without degrees-of-freedom correction. An advantage of the above formulation (3.66) is that it will never be negative. However, it is possible that $R^2_{LMM(m)}$ can decrease by adding predictors, but this is unlikely since $\sigma_f^2$ should always increase when predictors are added to the model.

- **GLMMs**

  Regrading the case of extending the aforementioned direct formulation to GLMMs, it has been mentioned that for non-Gaussian responses, it is difficult to define the residual variance, $\sigma_\epsilon^2$. However, it is possible to define the residual variance on the latent (or link) scale, although the definition of the residual variance is specific to the error distribution and the link function used in the analysis.
  In GLMMs, $\sigma_\epsilon^2$ can be expressed with three components (see [74]) :

  **1.** Multiplicative dispersion ($\omega$).

  **2.** Additive dispersion ($\sigma_e^2$).

  **3.** Distribution-specific variance ($\sigma_d^2$).

  We can implement GLMMs in two distinct ways, either by multiplicative or additive dispersion. Dispersion is fitted to account for variance that exceeds or falls short of the distribution-specific variance (e.g., Binomial or Poisson distributions). In this thesis, (as in [73]) we consider only additive dispersion implementation of GLMMs. Nonetheless, the presented formulas, can be modified for the use of GLMMs that apply to multiplicative dispersion.

  When additive dispersion is used, $\sigma_\epsilon^2$ equals the sum of the additive dispersion component and the distribution-specific variance ($\sigma_e^2 + \sigma_d^2$), and thus, $R^2$ for GLMMs can be defined as:

$$R^2_{GLMM(mar)} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_e^2 + \sigma_d^2}, \tag{3.68}$$

  where "$mar$" stands for *marginal*, is the variance explained on the latent (or link) scale rather than original scale.

The quantity in (3.68) can be generalized to multiple levels as follows:

$$R^2_{GLMM(mar)} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}, \tag{3.69}$$

where $u$ is the number of random factors in GLMMs and $\sigma_l^2$ is the variance component of the $\boldsymbol{l}^{th}$ random factor. The equation on (3.69) can be modified so that it can express conditional $R^2$.

$$R^2_{GLMM(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}, \tag{3.70}$$

where "$c$" stands for *conditional*.
Notice that in (3.70), the conditional $R^2$, i.e. $R^2_{GLMM(c)}$, can be interpreted as the variance explained by the full model. Both marginal and conditional $R^2_{GLMM}$ convey unique and interesting information. Nakagawa et al. in [73], recommended the use of them for research purposes.

In the case of a Gaussian response and an identity link, the linked scale variance and the original scale variance are identical and the distribution-specific variance is zero. Thus, $\sigma_e^2 + \sigma_d^2$ reduces to $\sigma_\epsilon^2$ in (3.69) and (3.70). For other GLMMs, the link-scale variance is different from the original scale variance.

The equations in (3.69) and (3.70) can be applied to different GLMMs families, given the knowledge of a distribution specific variance $\sigma_d^2$, and a model that fits additive overdispersion (e.g., [41]). It is worth to be noted that when the denominators of (3.69) and (3.70) include $\sigma_d^2$, both types of $R^2_{GLMM}$ will never become 1 in contrast to traditional $R^2$.
The interested reader may refer to [73] for more details on $R^2$ for GLMMs and related issues.

CHAPTER 4

# An Application on Influenza-Like Illness Outbreaks

## 4.1 Materials and Methods

### 4.1.1 Sentinel Epidemiological Surveillance System

In Greece, since 1999, a sentinel system of epidemiological surveillance is operating which is based on voluntary participation of physicians, general practitioners and pediatricians of Primary Health Care (PHC) throughout Greece. The sentinel systems in PHC are the most important source of PHC epidemiological diseases data and through reporting processing, analysis, and results/conclusions export procedures, provide general guidelines for optimal decision making in health services. Through these systems, the evolution of the frequency of certain diseases is recorded by selected reporting sites and health professionals report cases of the disease or syndrome under surveillance, based on clinical diagnoses. The sentinel medical doctors send weekly epidemiological data regarding the number under surveillance, according to a specified clinical definition. Then, the Hellenic Center for Disease Control Prevention (HCDCP) estimates the weekly number of syndrome cases per 1000 visits using reporting forms.

During the period $2014-2015$ there was a reorganization of the Hellenic sentinel system and several changes in the operating parameters of the system took place. As a result, the national priorities were posed for the syndromes monitored through the sentinel system and the system was directed to main syndromes of interest, namely, influenza-like illness (ILI) and gastroenteritis. The study of the evolution of these two syndromes is of major public health concern for three key reasons. Firstly, they belong to the sentinel epidemiological surveillance priorities of the country, secondly they are monitored traditionally by sentinel systems in the European region, while they are high in terms of international interest, due to their potential for widespread transmission (or even potential pandemic risk). Thirdly, the surveillance of ILI and gastroenteritis through the sentinel system allows studying 1. the existence of seasonality, 2. the determination of signaled start and end weeks and the intensity of epidemic waves for ILI, 3. the determination of epidemic outbreaks for gastroenteritis nationwide [81].

### 4.1.2 Two Season Influenza Historical Data

As stated above, through the operation of the sentinel system, the evolution of the frequency of certain diseases is recorded by carefully sampled reporting

units, based on clinical diagnoses including clinical manifestations compatible with flu, i.e., ILI. Through the aforementioned reorganization of the sentinel system in Greece, the system was harmonized with the updated European and international standards-instructions. As a consequence, recent sentinel surveillance system data (week40/2014 to date) are not considered comparable to those of the previous years (past influenza seasons until week39/2014). In this work, we focus on the study of weekly ILI rate data between September 29, 2014 and October 2, 2016. This data was used for analysis purposes, in order to determine the past two seasonal influenza outbreaks (signaled standard end weeks) and establish optimal empirical epidemic thresholds. Thus, we conducted a retrospective analysis for the period from 2014 to 2016, based on a model fit to two−season historical data (week40/2014 to week39/2016). The main objectives of this work are:

**1.** The prediction of the time interval for which an influenza outbreak is expected;

**2.** The estimation of the duration of the epidemic waves;

**3.** The early detection of possible epidemics.

### 4.1.3 Research Methodology

There are two types of analysis for surveillance time series: retrospective and prospective analysis. The first one, locates and quantifies the impact of past epidemics, whereas the second one, is useful when it comes to real time detection of epidemics. This study focuses on retrospective analysis, epidemic detection and quantification from time series data. Four steps are necessary to be followed, in the case of detection of influenza epidemics in time series:

**1.** Determination of the training period;

**2.** Purge of the training period;

**3.** Estimation of the regression equation;

**4.** Epidemic alert notification.

#### 4.1.3.1 Determination of the Training Period

In general, not all data should be included in the training period even in cases that long times series are available [91]. Specifically, as discussed by Pelat et al. in [83], changes in case reporting or/and demographics will likely be present over long time periods and this fact may affect how well the baseline model fits the data. odeling of influenza morbidity or/and mortality typically makes use of the five proceeding years in baseline determination. In our study, we included all data in the training period, using the whole dataset in the model fitting for retrospective analysis, (as in [83], [92], [103]). Parpoula et al. in [81], pointed out that including more past seasons improves the seasonal components estimates, while limiting the quantity of the data allows capturing recent trends. Pelat et al. in [83] suggested, that a minimum of one year historical data is required to fit the models. If one wants to succeed more reliable predictions, then at least two or even three year historical data is required in order to calculate the baseline level.

#### 4.1.3.2 Purge of the Training Period

It is very important to fit the model on non-epidemic data, in order to model the non-epidemic baseline level. For seasonal diseases such as influenza, it is difficult to find long non-epidemic periods, since epidemics typically occur every year. There are two choices to make in order to deal with such a problem [83]. The first one is to identify epidemics, and then exclude the corresponding data from the series. The second and less common one, requires explicit modeling of the epidemic periods during the training data. In the second case, an epidemic indicator is required to be included as a covariate in the model. However, in such a case hidden Markov models may be appropriate, the consideration of which will be investigated in a future work (see e.g. [108]).

In the first choice, we must first identify epidemics. Several rules have been suggested in the literature in this respect, such as excluding the 25% higher values from the training period [103], removing all data above a given threshold [26] or excluding whole periods known to be epidemic prone [78]. For more details the reader may refer to [81] and [83]. In this work, we selected to exclude the 15% highest observations from the training period (the default value selected by Pelat et al. in [83]).

#### 4.1.3.3 Estimation of the Regression Equation

Several different formulations could be used for the regression equation, as discussed in [81] such as linear regression [26], linear regression on the log-transformed series [18], Poisson regression [96], and Poisson regression allowing for over-dispersion [102]. In this work, linear regression is applied on the available two season historical data series (weekly ILI rate data, i.e., week40/2014 to week39/2016). The weekly estimated ILI rate, is a time series with specific characteristic properties, such as trend and seasonality. In the regression equation the trend is usually modeled using a linear term or a polynomial (of 2nd or 3rd degree) [83], while seasonality is usually modeled using sine and cosine terms with period one year. However, refined models are found in literature, often with terms of period six months [26], 3 months [45], and smaller [60].

In this work, we follow an exhaustive search process, based on periodic mixed regression models discussed in Section 2.9, in order to identify the optimal fit of the baseline model. Thus, linear, quadratic, cubic and quartic (for comparison purposes) trends are considered, and regarding the seasonal component, the most widely used periodicities are implemented, i.e. 12, 6, and 3 months. The terms stated above, are not the only ones that can be included in the regression equation. Some authors included autoregressive terms in their models ( as done in [104], [79], [106]). Others incorporated additional variables into the regression equation, such as day of the week [71], sex and age [19], and environmental factors for example temperature and humidity [104]. Thus, in this work we make use of the following environmental factors:

- temperature,
- humidity,
- wind force, and
- wind direction,

because of the interest in the impact of extreme weather on human health that is

increasing in recent decades. Climate changes and how these affect public health ( [39]; [55]), showed that higher temperatures are likely to increase heat-related mortality worldwide. In addition, there is evidence that high temperatures are associated with mortality [9]. The relationship between temperature and mortality may be confounded by a range of measured or unmeasured confounders. Confounding factors are present when a covariate is strongly associated with both the outcome and exposure of interest, but it is not a result of the exposure and may distort the association being studied between two other variables. As pointed out by Touloumi et al. in [98] a fundamental consideration in epidemiological modeling, is to properly control for all potential confounders. Such confounders may include meteorological indicators, such as relative humidity, seasonality and long-term trends ( [10]; [17]; [31]; [40]; [95]). In addition, Tsangari et al. in [100], concluded that high temperatures during warm months can result in increased mortality rates. Since influenza causes an estimated $290,000 - 650,000$ deaths worldwide [105], it is considered reasonable to study the effect of environmental factors on influenza-like illness ( [70], [84]). In this word, additional environmental factors were incorporated into the regression model such as: minimum-maximum-median-mean temperature (temp), minimum-maximum-median-mean wind direction (wd), minimum-maximum-median-mean wind force (wf). Moreover, first-second order autoregressive terms and first-second order moving average terms, were also incorporated into the regression model. Meteorological data were provided by the Hellenic National Meteorological Service (HNMS).
Thus, the regression equation is defined as follows:

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) \\ + \gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right) \\ + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \lambda_2 \epsilon_{t-2} \\ + \zeta_1 \text{minwd} + \zeta_2 \text{maxwd} + \zeta_3 \text{medianwd} + \zeta_4 \text{meanwd} \\ + \theta_1 \text{minwf} + \theta_2 \text{maxwf} + \theta_3 \text{medianwf} + \theta_4 \text{meanwf} + \omega_1 \text{mintemp} \\ + \omega_2 \text{maxtemp} + \omega_3 \text{mediantemp} + \omega_4 \text{meantemp}. \end{aligned} \tag{4.1}$$

Then, a thorough comparison was made, among all candidate periodic mixed models (e.g., $PAR(1)$, $PAR(2)$, $PMA(1)$, $PMA(2)$, $PARMA(1,1)$, $PARMA(2,1)$, $PARMA(1,2)$, $PARMA(2,1)$), with respect to the significance of the environmental factors. Note that all regression equations for the observed value $y_t$ are special cases of equation (4.1).
The comparison is described step by step as follows: We start from the simplest model labeled as $PAR(1)$ by examining the significance of each of the environmental explanatory variables of the mixed model. If there is at least one significant, then we keep the model and go on to the next one (e.g. $PAR(2)$). The procedure goes on until the significance of the environmental factors for each model is examined. Finally, the models kept by the process, are being compared with respect to $MDIC$. The model with the lowest $MDIC$ is being selected. Thus, $PARMA(2,1)$ mixed

model with minimum temperature as a significant covariate ($p-value<.001$) was the model selected.
As a result, the time period and the minimum temperature are the explanatory variables, the observed time series values (weekly ILI rate) is the dependent variable and all regression equations for the observed value $y_t$ are special cases of the following $PARMA(2,1)$ mixed model:

$$\begin{aligned} y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \gamma_1 cos\left(\frac{2\pi t}{n}\right) + \delta_1 sin\left(\frac{2\pi t}{n}\right) \\ +\gamma_2 cos\left(\frac{4\pi t}{n}\right) + \delta_2 sin\left(\frac{4\pi t}{n}\right) + \gamma_3 cos\left(\frac{8\pi t}{n}\right) + \delta_3 sin\left(\frac{8\pi t}{n}\right) \\ +\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned} \tag{4.2}$$

where $\epsilon_t \sim WN(0,\sigma^2)$, $n$ denotes the sample size, and parameter coefficients are estimated by least squares regression. Selection of the best fitting model, relies on:
• 12 candidate models that combine 4 trends, namely, linear, quadratic, cubic and quartic, and 3 seasonal periodicities, namely, 12, 6, and 3 months.
• Analysis Of Variance ($ANOVA$) comparison (significance level $\alpha$ is chosen to be 5%), for nested models.
• Akaike Information Criterion ($AIC$) or Modified Divergence Information Criterion ($MDIC$), for non-nested models (see Sections 3.2 and 3.4).

The latter process is described step by step as follows: The process starts comparing, by $ANOVA$, the simplest model labeled as $M11$ with a linear trend and 12−month seasonal periodicity, and defined as

$$\begin{aligned} M11: \quad y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\Big(\frac{2\pi t}{n}\Big) + \delta_1 sin\Big(\frac{2\pi t}{n}\Big) + \phi_1 y_{t-1} \\ +\phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned} \tag{4.3}$$

with the two models within which it is nested, labeled as $M12$ (linear trend and 12− and 6−month seasonal periodicities) and $M21$ (quadratic trend and 12− month seasonal periodicity), which are defined as

$$\begin{aligned} M12: \quad y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\Big(\frac{2\pi t}{n}\Big) + \delta_1 sin\Big(\frac{2\pi t}{n}\Big) + \gamma_2 cos\Big(\frac{4\pi t}{n}\Big) + \delta_2 sin\Big(\frac{4\pi t}{n}\Big) \\ +\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}, \end{aligned} \tag{4.4}$$

and

$$\begin{aligned} M21: \quad y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 cos\Big(\frac{2\pi t}{n}\Big) + \delta_1 sin\Big(\frac{2\pi t}{n}\Big) + \gamma_2 cos\Big(\frac{4\pi t}{n}\Big) + \delta_2 sin\Big(\frac{4\pi t}{n}\Big) \\ +\gamma_3 cos\Big(\frac{8\pi t}{n}\Big) + \delta_3 sin\Big(\frac{8\pi t}{n}\Big) + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \end{aligned} \tag{4.5}$$

In the case that none of the alternative models ($M12$ and $M21$) is significantly

better than the initial one ($M11$), the process retains $M11$ and terminates. If one of the two alternative models is better than the initial one ($p - values < 0.05$), the algorithmic process keeps it and goes on. If both alternative models are better than the initial one, the algorithmic process keeps the one with the lowest $AIC$ or $MDIC$ and goes on. The procedure is repeated until finding the "best overall" model over the twelve considered models (combining the four choices for the trend and the three choices for the periodicity).

#### 4.1.3.4 Epidemic Alert Notification

As the model is fitted to the observations, one could make use of the standard deviation of the residuals $(y_t - \hat{y}_t)$ in order to estimate the variation around the model fit. In this way, if we assume that the baseline model holds in the future, it is possible to obtain forecast intervals for future observations [83]. The epidemic thresholds which signal an unexpected change are typically obtained by taking an upper percentile for the prediction distribution (assumed to me normal), usually the upper $95^{th}$ percentile [26], or upper $90^{th}$ percentile [92]. Increasing the value of the upper percentile, will lead to less observations outside the thresholds and more specific detection. On the other hand, decreasing this value will increase the sensitivity and timeliness of the alerts. In this study, we will obtain the epidemic threshold by taking the upper $95^{th}$ percentile of the prediction distribution. In addition, we then make use of the following rule "*a series of observations fall above the epidemic threshold*" in order to define when epidemic alerts are produced. The latter step is important since in this way we avoid making alerts for isolated data points. Thus, a minimum duration above the epidemic threshold is required. The final rule was set to be "*a series of observations fall above the epidemic threshold during 2 weeks*" (see [83], [103]). The beginning of the epidemic is signaled the first time the series exceeds the threshold, and the end the first time the series returns below the threshold.

## 4.2 Experimental Study

As in [81], we conducted a retrospective analysis; the whole time series, i.e., week40/2014 to week39/2016, was therefore included in the training period. Then, we chose to exclude the top 15% observations from the training period (89 kept values from the total of 105). Based on $ANOVA$ comparison, $AIC$ and $MDIC$ criteria, the model selected was $\boldsymbol{M11}$ with:
• linear trend,
• annual periodic term (one year harmonics),
• first and second order autoregressive terms,
• first order moving average term, and
• minimum temperature.
In addition, the forecast interval was set to be 95%, that is the upper limit of the prediction interval which is used as a threshold to detect epidemics. The alert rule, was chosen to be "*an epidemic is declared when 2 weekly successive observations are above the estimated threshold*".

The mathematical form of model $\boldsymbol{M11}$ is described as follows:

$$y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\Big(\frac{2\pi t}{n}\Big) + \delta_1 sin\Big(\frac{2\pi t}{n}\Big) + \phi_1 y_{t-1}$$

$$+\phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \tag{4.6}$$

Table 4.1 presents the estimated parameters, the standard errors (sd), the statistic values (t-value) and the associated p-values of the selected model. The twelve periodic regression mixed models are described in Table 4.2, in which the components included in each model are indicated by "*", along with $MDIC$ and $R^2_{GLLM(m)}$ values of each model. The model finally kept $M11$ is in bold italics. Figures 4.1 and 4.2 illustrate the model selection pathway, using $AIC$ and $MDIC$ criterion, respectively. In addition, Figure 4.3 illustrates the plots of the time series, the predicted baseline as well as the threshold. In Figure 4.3, the epidemics detected by the selected model ($\boldsymbol{M11}$) appear in light red. The predicted baseline and threshold values at each date in

**Table 4.1:** Selected Model M11 Output

| Parameter | Estimate | sd | t-value | p-value |
|---|---|---|---|---|
| $\alpha_0$ | 10.468 | 2.128 | 4.920 | $< .001$ |
| $\alpha_1$ | -0.075 | 0.027 | -2.789 | 0.007 |
| $\gamma_1$ | -10.315 | 1.444 | -7.144 | $< .001$ |
| $\delta_1$ | 12.726 | 2.144 | 5.937 | $< .001$ |
| $\phi_1$ | 0.811 | 0.114 | 7.098 | $< .001$ |
| $\phi_2$ | -0.234 | 0.090 | -2.613 | 0.011 |
| $\lambda_1$ | 0.261 | 0.135 | -2.613 | 0.058 |
| $\omega_1$ | 0.729 | 0.172 | 4.247 | $< .001$ |

the dataset are presented in Table 4.3. Also, in Table 4.3 the epidemics appear in bold. Finally, Table 4.4 presents the dates and the results of the retrospective evaluation of the excess influenza morbidity in Greece for $2014 - 2016$ along with excess percentages, using the $M11$ periodic regression mixed model. The excess morbidity is defined as the cumulative difference between observations and baseline over the entire epidemic period. Excess percentages were calculated as the observed size divided by the sum of expected values throughout each epidemic.

**Table 4.2:** Models Selected Through the Algorithm Pathway

| | T[a] | | | | P[b] | | | ARMA | | | LV[c] | IC[d] | | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M[e] | t | $t^2$ | $t^3$ | $t^4$ | 1 y[f] | 6 m[g] | 3 m | AR(1) | AR(2) | MA(1) | MT[h] | AIC | MDIC | $R^2_{GLMM(mar)}$ |
| ***M11*** | * | | | | * | | | * | * | * | * | 409.31 | 8.32 | 0.927 |
| M12 | * | | | | * | * | | * | * | * | * | 406.96 | 22.98 | 0.934 |
| M13 | * | | | | * | * | * | * | * | * | * | 406.56 | 39.08 | 0.928 |
| M21 | * | * | | | * | | | * | * | * | * | 410.72 | 13.68 | 0.938 |
| M22 | * | * | | | * | * | | * | * | * | * | 408.49 | 30.25 | 0.934 |
| M23 | * | * | | | * | * | * | * | * | * | * | 408.17 | 50.29 | 0.937 |
| M31 | * | * | * | | * | | | * | * | * | * | 403.82 | 21.75 | 0.938 |
| M32 | * | * | * | | * | * | | * | * | * | * | 402.83 | 34.30 | 0.941 |
| M33 | * | * | * | | * | * | * | * | * | * | * | 401.04 | 57.51 | 0.937 |
| M41 | * | * | * | * | * | | | * | * | * | * | 405.56 | 31.53 | 0.946 |
| M42 | * | * | * | * | * | * | | * | * | * | * | 403.90 | 46.56 | 0.942 |
| M43 | * | * | * | * | * | * | * | * | * | * | * | 402.48 | 73.05 | 0.946 |

[a] "T" denotes *trend*;
[b] "P" denotes *periodicity*;
[c] "LV" denotes *latent variable*;
[d] "IC" denotes *information criterion*;
[e] "M" denotes *model*;
[f] "y" denotes *year*;
[g] "m" denotes *months*;
[h] "MT" denotes *minimum temperature*.

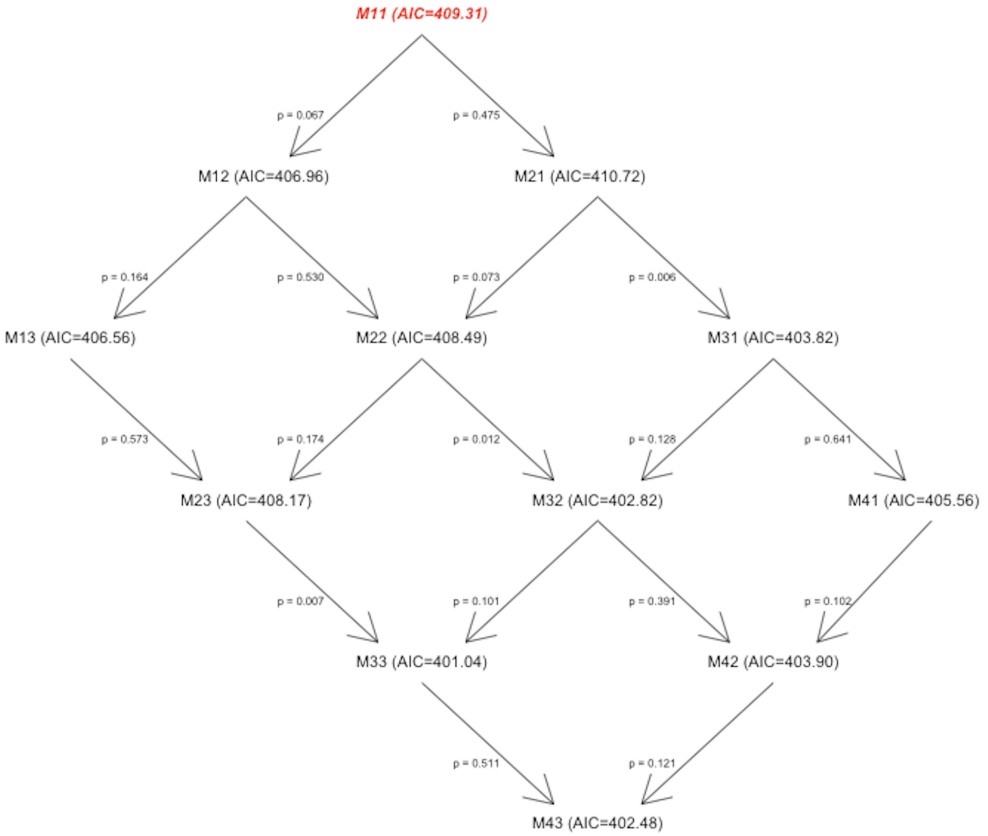


**Figure 4.1:** *Model selection pathway (ANOVA & AIC)*

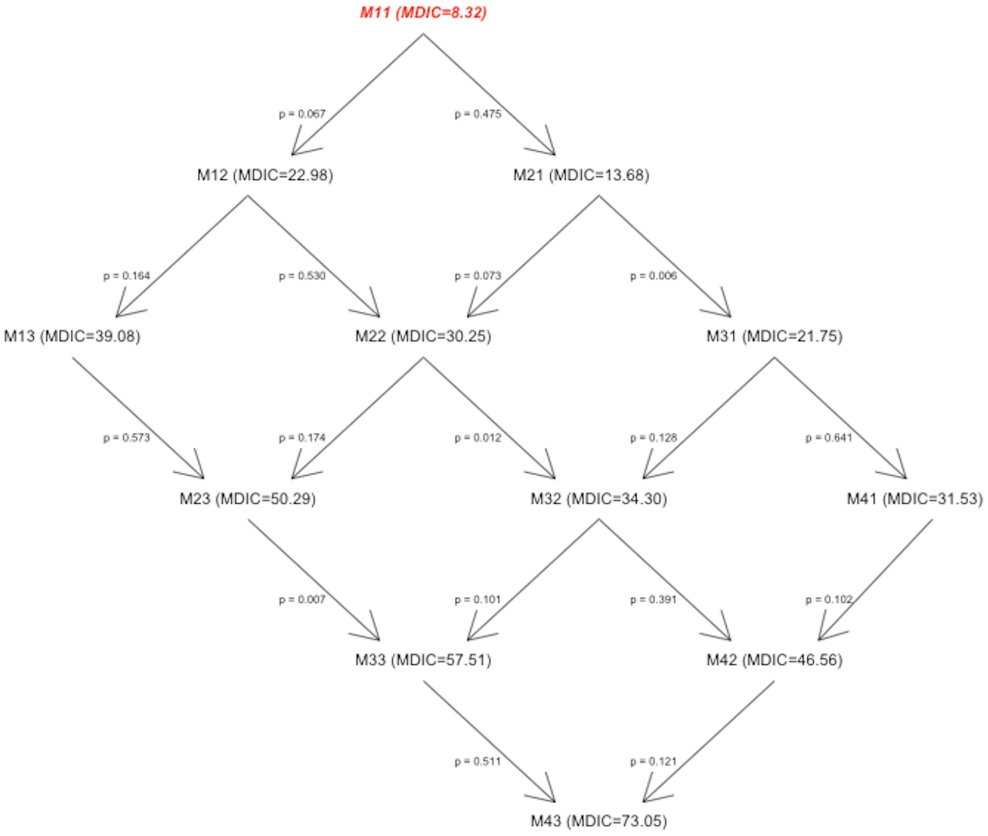


**Figure 4.2:** *Model selection pathway (ANOVA & MDIC)*

**Table 4.3:** Model Results: Predicted Baseline and Epidemic Threshold values

| Week | Date [a] | PB [b] | TH [c] | Week | Date [a] | PB [b] | TH [c] |
|---|---|---|---|---|---|---|---|
| 201440 | 2014-09-29 | 21.50 | 30.00 | 201441 | 2014-10-06 | 24.20 | 32.70 |
| 201442 | 2014-10-13 | 27.00 | 35.50 | 201443 | 2014-10-20 | 29.30 | 37.80 |
| 201444 | 2014-10-27 | 23.60 | 32.10 | 201445 | 2014-11-03 | 30.20 | 38.70 |
| 201446 | 2014-11-10 | 29.40 | 37.90 | 201447 | 2014-11-17 | 30.70 | 39.20 |
| 201448 | 2014-11-24 | 32.00 | 40.40 | 201449 | 2014-12-01 | 33.80 | 42.30 |
| 201450 | 2014-12-08 | 32.30 | 40.80 | 201451 | 2014-12-15 | 33.40 | 41.90 |
| 201452 | 2014-12-22 | 56.60 | 65.10 | **201501** | **2014-12-29** | **51.90** | **60.40** |
| **201502** | **2015-01-05** | **52.60** | **61.10** | **201503** | **2015-01-12** | **103.00** | **111.00** |
| **201504** | **2015-01-19** | **92.90** | **101.00** | **201505** | **2015-01-26** | **99.50** | **108.00** |
| **201506** | **2015-02-02** | **66.40** | **74.90** | **201507** | **2015-02-09** | **70.00** | **78.50** |
| **201508** | **2015-02-16** | **100.00** | **109.00** | **201509** | **2015-02-23** | **68.00** | **76.50** |
| **201510** | **2015-03-02** | **78.20** | **86.70** | **201511** | **2015-03-09** | **62.30** | **70.80** |
| **201512** | **2015-03-16** | **46.50** | **55.00** | 201513 | 2015-03-23 | 44.60 | 53.10 |
| 201514 | 2015-03-30 | 31.70 | 40.20 | 201515 | 2015-04-06 | 35.60 | 44.10 |
| 201516 | 2015-04-13 | 25.60 | 34.10 | 201517 | 2015-04-20 | 17.30 | 25.70 |
| 201518 | 2015-04-27 | 17.02 | 25.60 | 201519 | 2015-05-04 | 17.10 | 25.60 |
| 201520 | 2015-05-11 | 11.40 | 19.90 | 201521 | 2015-05-18 | 9.88 | 18.40 |
| 201522 | 2015-05-25 | 3.06 | 11.50 | 201523 | 2015-06-01 | 0.00 | 6.32 |
| 201524 | 2015-06-08 | 0.68 | 9.17 | 201525 | 2015-06-15 | 3.15 | 11.60 |
| 201526 | 2015-06-22 | 0.00 | 7.92 | 201527 | 2015-06-29 | 0.08 | 8.56 |
| 201528 | 2015-07-06 | 2.55 | 11.00 | 201529 | 2015-07-13 | 0.82 | 9.31 |
| 201530 | 2015-07-20 | 3.14 | 11.60 | 201531 | 2015-07-27 | 3.20 | 11.70 |
| 201532 | 2015-08-03 | 0.00 | 8.33 | 201533 | 2015-08-10 | 1.20 | 9.68 |
| 201534 | 2015-08-17 | 0.00 | 8.19 | 201535 | 2015-08-24 | 1.18 | 9.67 |
| 201536 | 2015-08-31 | 0.00 | 8.13 | 201537 | 2015-09-07 | 1.89 | 10.40 |
| 201538 | 2015-09-14 | 3.35 | 11.80 | 201539 | 2015-09-21 | 0.95 | 9.44 |
| 201540 | 2015-09-28 | 7.63 | 16.10 | 201541 | 2015-10-05 | 11.80 | 20.20 |
| 201542 | 2015-10-12 | 14.90 | 23.40 | 201543 | 2015-10-19 | 17.70 | 26.20 |
| 201544 | 2015-10-26 | 13.10 | 21.60 | 201545 | 2015-11-02 | 15.20 | 23.70 |
| 201546 | 2015-11-09 | 15.60 | 24.10 | 201547 | 2015-11-16 | 19.60 | 28.10 |
| 201548 | 2015-11-23 | 22.90 | 31.40 | 201549 | 2015-11-30 | 21.90 | 30.40 |
| 201550 | 2015-12-07 | 21.20 | 29.70 | 201551 | 2015-12-14 | 25.30 | 33.80 |
| 201552 | 2015-12-21 | 28.90 | 37.40 | 201553 | 2015-12-28 | 39.30 | 47.80 |
| 201601 | 2016-01-04 | 47.30 | 55.80 | 201602 | 2016-01-11 | 57.50 | 66.00 |
| 201603 | 2016-01-18 | 58.80 | 67.30 | 201604 | 2016-01-25 | 89.70 | 98.20 |
| **201605** | **2016-02-01** | **71.80** | **80.30** | **201606** | **2016-02-08** | **56.40** | **64.90** |
| **201607** | **2016-02-15** | **56.00** | **64.50** | **201608** | **2016-02-22** | **40.90** | **49.40** |
| 201609 | 2016-02-29 | 29.90 | 38.40 | 201610 | 2016-03-07 | 39.50 | 47.90 |
| 201611 | 2016-03-14 | 36.70 | 45.20 | 201612 | 2016-03-21 | 28.10 | 36.60 |
| 201613 | 2016-03-28 | 27.60 | 36.10 | 201614 | 2016-04-04 | 24.40 | 32.90 |
| 201615 | 2016-04-11 | 19.70 | 28.20 | 201616 | 2016-04-18 | 15.00 | 23.50 |
| 201617 | 2016-04-25 | 11.40 | 19.90 | 201618 | 2016-05-02 | 7.35 | 15.80 |

[a] Date denotes the week start date;
[b] PB denotes the predicted baseline;
[c] TH denotes the threshold.

**Table 4.4:** Retrospective Evaluation of the Excess Influenza Morbidity in Greece 2014-2016

| SW [a] | EW [a] | Excess cases | Expected cases | Cases | Excess percentage |
|---|---|---|---|---|---|
| 201501 | 201512 | 260 | 891 | 1151 | 29% |
| 201605 | 201608 | 91 | 225 | 316 | 40% |

[a] SW and EW denote the signaled start and end weeks for epidemics, respectively.

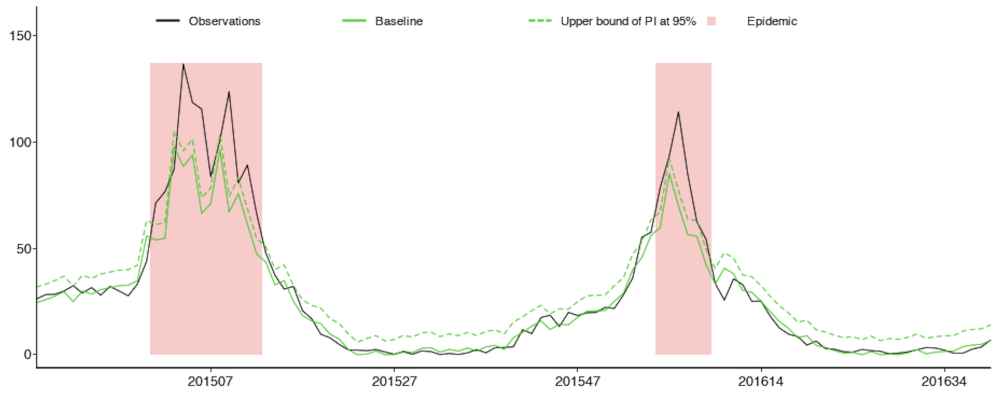


**Figure 4.3:** Detected epidemics in Greece 2014 – 2016

## 4.3 Conclusions

Based on Table 4.2 the selected periodic regression mixed model is **M11** which is described as follows:

$$y_t = \alpha_0 + \alpha_1 t + \gamma_1 cos\Big(\frac{2\pi t}{n}\Big) + \delta_1 sin\Big(\frac{2\pi t}{n}\Big) + \phi_1 y_{t-1}$$

$$+\phi_2 y_{t-2} + \epsilon_t + \lambda_1 \epsilon_{t-1} + \omega_1 \text{mintemp}. \tag{4.7}$$

However, in Fig.4.1 and Fig.4.2 the algorithm could have proceeded and stopped at M12, since the $p - value$ is close to $\alpha = 5\%$ (0.067). Thus, one could make a comparison between models M11 and M12, in order to select the "best overall" model.

The periodic mixed model M11, was selected based on the algorithm described in Subsection 4.1.3. It is worth to be noted that the exclusive use of AIC and $R^2_{GLMM(mar)}$ the "best overall" models would be M21 and M33, respectively. In Table 4.5 a comparison of the aforementioned models is presented, in order to evaluate the accuracy of models M11, M12, M21, M33, M43[1]. Comparing the models, we observe that model M21 is not considered acceptable, since its results are never the best for none of the accuracy measures examined. The results for models M11, M12, and M33, are very close, except the measures of MPE and MAPE for which M11 clearly outperforms. Thus, M11 is the "best overall"[2] model. It is also very important to notice the values of $MDIC$ and $AIC$ of the models which are in full support of the above results. Indeed, $MDIC$, is clearly in favor of model M11 (smallest $MDIC$ value). As for $AIC$ we observe that M11 is not the best, but the gain by choosing alternative models like M12, M33 or M43 is not

[1] M43 is the "full model", and thus it is considered useful to be compared with the rest.

[2] Even if model M33 was better, we would choose again M11, since it has the advantage of less explanatory variables, with a $R^2_{GLMM(mar)}$ very close to the one of M43.

**Table 4.5:** Results of Common Accuracy Measures

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| M11 | -2.359224e-16 | 4.298974 | 3.040053 | 0.2782516 | 48.19166 | 0.230213 |
| M12 | 2.179757e-16 | 4.10272 | 2.857186 | -23.29906 | 75.60382 | 0.2163651 |
| M21 | 4.937659e-17 | 4.280225 | 2.9915 | 7.396555 | 50.17105 | 0.2265362 |
| M33 | 7.6736e-17 | 3.703522 | 2.757 | -69.34156 | 127.622 | 0.2087783 |
| M43 | 9.953614e-17 | 3.68828 | 2.783278 | -69.73654 | 132.0232 | 0.2107683 |

significant enough to balance the complexity associated with such models. The values of $MDIC$, tend to get bigger as more explanatory variables are included to the model. In fact, the penalty given to the models by $MDIC$, is much bigger than the one given by $AIC$. Thus, the addition of explanatory variables, makes $MDIC$ stricter than $AIC$ and nearly the exact opposite of $R^2_{GLMM(mar)}$.

Conclusively, in this study, we conducted a retrospective analysis for the estimation of the influenza-like syndrome morbidity burden in Greece for the period $2014 - 2016$ (week40/2014 to week39/2016). Also, we made use of periodic regression mixed models in order to estimate the baseline level for the time series, associated with a prediction interval. Finally, as seen in Fig.(4.2) the model selected (e.g., M11), via an exhaustive search process (see Subsection 4.1.3), succeeded in detecting as epidemic the period between the two peaks of the epidemic wave for the period of $2014 - 2015$. The interconnection of statistical research with Health professional's structures is considered very useful, since it can serve, as pointed out by the present work, critical needs of Public Health. The very early and accurate detection of an outbreak or an epidemic activity, can help and benefit to the very early taking of the appropriate measures, in order to protect the population at risk.

# Bibliography


[1] Agresti, A.: An introduction to categorical data analysis, 2nd edn.. Wiley (2007)

[2] Akaike, H., Kitagawa, G.: The Practice of Time Series Analysis. Springer (1998)

[3] Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control, **AC−19**, 716−723 (1974)

[4] Akaike, H.: Information measures and model selection. In Proceedings 44th Session of the International Statistical Institute **1**, 277−291 (1983)

[5] Akaike, H.: Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory (Petrov, B.N., Csaki, F., eds.), Akademiai Kiado, Budapest, 267−281 (1973)

[6] Akaike, H.: On entropy maximization principle. Applications of Statistics, P. R. Krishnaiah, ed., North−Holland Publishing Company, 27−41 (1977)

[7] Akaike, H.: Statistical predictor identification. Ann. Inst. of Statist. Math., **22**, 203−217 (1970)

[8] Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. R. Statist. Soc. B, **28**, 131−142 (1966)

[9] Armstrong, B.: Models for the relationship between ambient temperature and daily mortality. Epidemiology, **17**, 624−631 (2006)

[10] Armstrong, B.G., Chalabi, Z., Fenn, B., Hajat, S., Kovats, S., Milojevic, A., Wilkinson, P.: Association of mortality with high temperatures in a temperate climate: England and Wales. J. Epidemiol. Commun. Health, **65**, 340−345 (2011)

[11] Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika, **85**, 549−559 (1998)

[12] Berger, J. Pericchi, L.: Objective Bayesian methods for model selection: introduction and comparison (with discussion). In Model Selection, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes − Monograph Series **38**, 135−207 (2001)

[13] Bhansali, R. J.: A derivation of the information criteria for selecting autoregressive models. Advances in Applied Probability, **18**, 360−387 (1986)

[14] Bishop, Y.M.M., Fienberg, S.E., Hollland P.W.: Discrete multivariate analysis: theory and practice. MIT (1975)

[15] Bliss, C.: The method of probits. Science, **79**, 38−39 (1934)

[16] Bozdogan, H.: Model selection and Akaikes information criterion (AIC): The general theory and its analytical extensions. Psychometrika, **52**, 345−370 (1987)

[17] Braga, L.F., Zanobetti, A., Schwartz, J.: The lag structure between particulate air pollution and respiratory and cardiovascular deaths in 10 US cities. J. Occup. Environ. Med., **43**, 927−933 (2001)

[18] Brillman, J.C., Burr, T., Forslund, D., Joyce, E., Picard, R., Umland, E.: Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. BMC Med. Inform. Decis. Mak. **5**(1), 4 (2005)

[19] Brinkhof, M.W., Spoerri, A., Birrer, A., Hagman, R., Koch, D., Zwahlen, M.: Influenza−attributable mortality among the elderly in Switzerland. Swiss Med Wkly, **136(19−20)**, 302−309 (2006)

[20] Bryk, A.S., Raudenbush, S.: Hierarchical Linear Models. Sage (1992)

[21] Cameron, A.C., Windmeijer, F.A.G.: An R−squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics, **77**, 329−342 (1997)

[22] Casella, G., Berger, R.L.: Statistical inference, 2nd edn.. Duxburry (2001)

[23] Cipra, T.: Periodic moving average processes. Aplikace Matematiky, **30**, 218−229 (1985)

[24] Commenges, D., Jacqmin−Gadda, H.: Generalized score test of homogeneity based on correlated random effects models. J. R. Statist. Soc. B, **59**, 157−171 (1997)

[25] Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T., Dartigues, J.−F.: Test of homogeneity of binary data with explanatory variables. Biometrics, **50**, 613−620 (1994)

[26] Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J., Valleron, A.−J.: A routine tool for detection and assessment of epidemics of influenza−like syndromes in France. American Journal of Public Health **81**, 97−99 (1991)

[27] Cowpertwait, P.S.P., Metcalfe, A.V.: Introductory time series with R. Springer (2009)

[28] Cox, D.R., Hinkley D.V.: Theoritical statistics. Chapman and Hall (1995)

[29] Cressie, N., Read, T.R.C.: Multinomial goodness−of−fit tests. J. R. Statist. Soc. B, **46**, 440−464 (1984)

[30] Csiszar, I.: Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten. Publ. Math. Inst. Hungarian Academy Sci. **8**, 84−108 (1963)

[31] Curriero, F.C., Heiner, K.S., Samet, J.M., Zeger, S.L., Strug, L., Patz, J.A.: Temperature and mortality in 11 cities of the Eastern United States. Am. J. Epidemiol., **155**, 80−87 (2002)

[32] Diggle, P., Liang K.−Y., Zeger, S.L.: Longitudinal data analysis. Oxford University (1994)

[33] Draper, N.R., Smith, H.: Applied Regression Analysis, 3rd edn.. Wiley (1998)

[34] Findley, D. F., Wei, C. Z.: AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. Journal of Multivariate Analysis, **83**, 415−450 (2002)

[35] Findley, D. F.: On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. Journal of Time Series Analysis, **6**, 229−252 (1985)

[36] Fox, G.A., Negrete−Yankelevich, S., Sosa, V.J.: Ecological statistics: Contemporary theory and application. Oxford University (2015)

[37] Franses, P. H.: Periodically intergrated subset for Dutch industrial production money stock. Journal of Forcasting, **12**, 601−613 (1993)

[38] Gelman, A., Pardoe, L.: Bayesian measures of explained variance and pooling in-multilevel (hierarchical)models. Technometrics, **48**, 241−251 (2006)

[39] Gosling, S.N., Lowe, J.A., McGregor, G.R., Pelling, M., Malamud, B.D.: Associations between elevated atmospheric temperature and human mortality: a critical review of the literature. Clim. Change, **92**, 299−341 (2009)

[40] Guo, Y., Barnett, A.G., Pan, X., Yu, W., Tong, S.: The impact of temperature on mortality in Tianjin, China: a case-crossover design with a distributed lag nonlinear model. Environ. Health Perspect., **119**, 1719−1725 (2011)

[41] Hadfield, J.D. : MCMC methods for multi−response Generalised Linear Mixed Models: the MCMCglmm R package. Journal of Statistical Software, **33**, 1−22 (2010)

[42] Hellinger, E.: Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik, **136**, 210−271 (1909)

[43] Heyde, C.C.: Quasi−likelihood and its application: A general approach to optimal parameter application. Springer (1997)

[44] Hilbe, J. M.: Generalized Linear Models. StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies (2010)

[45] Housworth, J., Langmuir, A.D.: Excess mortality from epidemic influenza, 1957−1966. Am. J. Epidemiol., **100**, 40−48 (1974)

[46] Jacqmin−Gadda, H., Commenges, D.: Tests of homogeneity for generalized linear models. J. A. S. A., **90**, 1237−1246 (1995)

[47] Jiang, J.: Linear and generalized linear mixed models and their applications. Springer (2007)

[48] Jones, R.H., Brelsford, W.M.: Time series with periodic structure. Biometrika, **54**, 403−407 (1967)

[49] Kass, R.E., Raftery, A.E. . Bayesian factors. J. A. S. A., **90**, 773−795 (1995)

[50] Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series. Lecture Notes in Statistics, **116**, Springer (1996)

[51] Konishi, S., Kitagawa, G.: Information criteria and statistical modeling. Springer (2008)

[52] Konishi, S.: Statistical model evaluation and information criteria. In Multivariate Analysis, Design of Experiments and Survey Sampling, S. Ghosh ed., 369−399 (1999)

[53] Konishi, S.: Theory for statistical modeling and information criteria −functional approach. American Mathematical Society, **15−1**, 89−106 (2002)

[54] Koukouvinos, C., Parpoula, C.: Variable selection and computation of the prior probability model via ROC curves methodology. J. D. S., **10**, 653−672, (2012)

[55] Kovats, R.S., Hajat, S.: Heat stress and public health: a critical review. Ann. Rev. Public Health, **29**, 41−55 (2008)

[56] Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics, **22**, 79−86 (1951)

[57] Kvålseth, T.O.: Cautionary note about $R^2$. American Statistician, **39**, 279−285 (1985)

[58] Lin, X.: Variance component testing in generalized linear models with random effects. Biometrika, **84**, 309−326 (1997)

[59] Linhart, H., Zuccini, W.: Model Selection. Wiley, (1986)

[60] Lui, K.J., Kendal, A.P.: Impact of influenza epidemics on mortality in the United States from October 1972 to May 1985. Am. J. Public Health, **77**, 712−716 (1987)

[61] Maddala, G.S.: Limited−Dependent and Qualitative Variables in Econometrics. Cambridge University Press (1983)

[62] Mantalos, P., Mattheou, K., Karagrigoriou, A.: An improved divergence information criterion for the determination of the order of an AR process. Communications in Statistics − Simulation and Computation, **39**, 865−879 (2010)

[63] Mattheou, K., Lee, S., Karagrigoriou, A.: A model selection criterion based on the BHHJ measure of divergence. Journal of Statistical Planning and Inference, **139**, 128−135 (2009)

[64] Mattheou, K.: On New Developments in Statistical Inference for Measures of Divergence, Ph.D. thesis, Department of Mathematics and Statistics, University of Cyprus (2007)

[65] Matusita, K.: On the notion of affinity of several distributions and some of its applications. Ann. Inst. Statist. Math., **19**, 181−192 (1967)

[66] McCullagh, P., Nelder, J.A.: Generalized linear models, 2nd edn.. Chapman and Hall (1989)

[67] McCulloch, C.E., Searle, S.R., Neuhaus, J.M.: Generalized, linear, and mixed models, 2nd edn.. Wiley (2008)

[68] McQuarrie, A.D.R., Tsai, C.−L.: Regression and Time Series Model Selection. World Scientific (1998)

[69] Menard, S.: Coefficients of determination for multiple logistic regression analysis. American Statistician, **54**, 17−24 (2000)

[70] Minh An, D.T., Bich Ngoc, N.T., Nilsson M.: Influenza−like illness in a Vietnamese province: epidemiology in correlation with weather factors and determinants from the surveillance system. Global Health Action, **7**, 10 (2014)

[71] Mostashari, F., Fine, A., Das, D., Adams, J., Layton, M.: Use of ambulance dispatch data as an early warning system for community wide influenza−like illness, New York City. J. Urban Health, **80(2 Suppl 1)**, i43−49 (2003)

[72] Nagelkerke, N.J.D.: A note on a general definition of the coefficient of determination. Biometrika, **78**, 691−692 (1991)

[73] Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining $R^2$ from generalized linear mixed−effects models. Methods in Ecology and Evolution, **4**, 133−142 (2013)

[74] Nakagawa, S., Schielzeth, H.: Repeatability for Gaussian and non−Gaussian data: a practical guide for biologists. Biological Reviews, **85**, 935−956 (2010)

[75] Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Statist. Soc. A, **135**, 370−384 (1972)

[76] Nelder, J.A.: Announcement by the working party on statistical computing: GLIM (Generalized Linear Interactive Modelling Program). J. R. Statist. Soc. A, **24**, 259−261 (1975)

[77] O'Hagan, A.: Fractional Bayes factors for model comparison (with discussion). J. R. Statist. Soc. B, **57**, 99−138 (1995)

[78] Olson, D.R., Simonsen, L., Edelson, P.J., Morse, S.S.: Epidemiological evidence of an early wave of the 1918 influenza pandemic in New York City. Proc. Natl. Acad. Sci. U.S.A, **31**, 11059−11063 (2005)

[79] Ozonoff, A., Forsberg, L., Bonetti, M., Pagano, M.: Bivariate method for spatio−temporal syndromic surveillance. MMWR Morb Mortal Wkly Rep, **53 Suppl**, 59−66 (2004)

[80] Pagano, M.: On periodic and multiple autoregressions. Annals of Statistics, **6**, 1310−1317 (1978)

[81] Parpoula, C., Karagrigoriou, A., Lambrou, A.: Epidemic intelligence statistical modelling for biosurveillance. J. Blömer et al. (Eds.): MACIS 2017, LNCS 10693, 1−15 (2017)

[82] Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 157−175 (1900)

[83] Pelat, C., Bolle, P.−Y., Cowling, B.J., Carrat, F., Flahault, A., Ansart, S., Valleron, A.−J.: Online detection and quantification of epidemics. BMC Medical Informatics and Decision Making **5**, 29 (2007)

[84] Pica, N., Bouvier, N.M.: Environmental Factors Affecting the Transmission of Respiratory Viruses. Current Opinion in Virology, **2(1)**, 90−95 (2012)

[85] Raudenbush, S., Bryk, A.S.: A hierarchical model for studying school effects. Sociology of Education, **59**, 1−17 (1986)

[86] Rothamsted Research $< https://www.rothamsted.ac.uk >$ (2017)

[87] Sakamoto, Y., Ishiguro, M., Kitagawa, G.: Akaike Information Criterion Statistics. D. Reidel Publishing Company (1986)

[88] Schielzeth, H., Nakagawa, S.: Nested by design: model fitting and interpretation in a mixed model era. Methods in Ecology and Evolution (2012)

[89] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics, **6**, 461−464 (1978)

[90] Searle, S.R.: Matrix algebra useful for statistics. Wiley (1982)

[91] Sebastiani, P., Mandl, K.: Biosurveillance and Outbreak Detection. Data Mining: Next Generation Challenges and Future Directions, 185−198 (2004)

[92] Simonsen, L., Reichert, T.A., Viboud, C., Blackwelder, W.C., Taylor, R.J., Miller, M.A.: Impact of influenza vaccination on seasonal mortality in the US elderly population. Arch. Intern. Med., **165**, 265−272 (2005)

[93] Snijders, T., Bosker, R.: Multilevel Analysis: An Introduction to basic and advanced multilevel modeling. Sage (1999)

[94] Snijders, T.A., Bosker, R.J.: Modeled variance in two−level models. Sociological Methods and Research, **22**, 342−363 (1994)

[95] Soebiyanto, R.P., Clara, W., Jara, J., et al.: The Role of Temperature and Humidity on Seasonal Influenza in Tropical Areas: Guatemala, El Salvador and Panama, 2008−2013. Colijn C, ed. PLoS ONE, **9(6)**, e100659 (2014)

[96] Thompson, W.W., Shay, D.K., Weintraub, E., Brammer, L., Cox, N., Anderson, L.J., Fukuda, K.: Mortality associated with influenza and respiratory syncytial virus in the United States. Jama, **289**, 179−186 (2003)

[97] Tiao, G.C., Grupe, M.R.: Hidden periodic autoregressive−moving average models in time series data. Biometrika, **67**, 365−373 (1980)

[98] Touloumi, G., Atkinson, R., Tertre, L., Samoli, A.E., Schwartz, J., Schindler, C., Vonk, J.M., Rossi, G., Saez, M., Rabszenko, D., Katsouyanni, K.: Analysis of health outcome time series data in epidemiological studies. Environmetrics, **15**, 101−117 (2004)

[99] Troutman, B.M.: Some results in periodic autoregression. Biometrika, **66**, 219−228 (1979)

[100] Tsangari, H., Paschalidou, A., Vardoulakis, S., Heaviside, C., Konsoula, Z., Christou, S., Georgiou, K.E., Ioannou, K., Mesimeris, T., Kleanthous, S., Pashiardis, S., Pavlou, P., Kassomenos, P., Yamasaki, E.N.: Human mortality in Cyprus: the role of temperature and particulate air pollution. Reg . Environ. Change, **16**, 1905−1913 (2016)

[101] Vecchia, A.V.: Maximum likelihood estimation for periodic autoregressive moving average models. Technometrics, **27**, 375−384 (1985)

[102] Vergu, E., Grais, R.F., Sarter, H., Fagot, J.P., Lambert, B., Valleron, A.J., Flahault, A.: Medication sales and syndromic surveillance, France. Emerg. Infect. Dis. **12**, 416−421 (2006)

[103] Viboud, C., Boelle, P.Y., Pakdaman, K., Carrat, F., Valleron, A.J., Flahault, A.: Influenza epidemics in the United States, France, and Australia, 1972−1997. Emerg. Infect. Dis., **10**, 32−39 (2004)

[104] Wong, C.M., Yang, L, Chan, K.P., Leung, G.M., Chan KH, Guan, Y., Lam, T.H., Hedley, A.J., Peiris, J.S.: Influenza−associated hospitalization in a subtropical city. PLoS Med, **3(4)**, e121 (2006)

[105] World Health Organization $< http://www.who.int/mediacentre/factsheets/fs211/en/ >$ (2017)

[106] World Health Organization $< http://www.who.int/topics/public\_health\_surveillance/en/ >$ (2017)

[107] Zero−inflated model $< https://en.wikipedia.org/wiki/Zero-inflated\_model >$, as published in 25 November (2017)

[108] Zucchini, W., MacDonald, I.: Hidden Markov models for time series: An introduction using R. Chapman & Hall/CRC (2009)