Two level random coefficient longitudinal modelling of COVID-19 spreads in Italy

by

Youtian Hao

MBA in Management, Lanzhou University, 2016
Bachelor in Economics, Tianjin Foreign Studies University, 2010

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

In the Graduate Academic Unit of Mathematics and Statistics

Supervisor(s):	M. Tariqul Hasan, Ph.D, Statistics
	Renjun Ma, Ph.D, Statistics
Examining Board:	Nicholas Touikan, Ph.D, Mathematics, Chair
	M. Tariqul Hasan, Ph.D, Statistics
	Guohua Yan, Ph.D, Statistics
	Lin Wang, Ph.D, Mathematics

This report is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

November, 2020

© Youtian Hao, 2021

Abstract

The current statistical modeling of coronavirus (COVID-19) spread has mainly focused on spreading patterns and forecasting of COVID-19 development; these patterns have been found to vary among locations. As the survival time of coronaviruses on surfaces depends on temperature, some researchers have explored the association of daily confirmed cases with environmental factors. Furthermore, some researchers have studied the link between daily fatality rates with regional factors such as health resources, but found no significant factors. As the spreading patterns of COVID-19 development vary a lot among locations, fitting regression models of daily confirmed cases or fatality rates directly with regional factors might not reveal important relationships. In this study, we investigate the link between regional spreading patterns of COVID-19 development in Italy and regional factors in two steps. First, we characterize regional spreading patterns of COVID-19 daily confirmed cases by a special patterned Poisson regression model for longitudinal count; the varying growth and declining patterns as well as turning points among regions in Italy have been well captured by regional regression parameters. We then associate these regional regression parameters with regional factors. The effects of regional factors on spreading patterns of COVID-19 daily confirmed cases have been effectively evaluated.

Dedication

This report is dedicated to my wife Ting Shi, who has given me supports with her unconditional love, and my baby Carmela Muwen Hao, who has made me stronger and better than I could have ever imagined.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Dr. M. Tariqul Hasan for his constant support and inspiration throughout my academic career in UNB, and my supervisor Dr. Renjun Ma for providing the methodology and general instructions on this report, as well as his earnest help to me.

I also would like to thank Dr. Guohua Yan, for his great patience on providing detailed technical support and prudential instructions on this report.

I also would like to thank Dr. Jeffery D. Picka for his instruction and encouragement.

I would like to express my profound gratitude to Department of Mathematics and Statistics in UNB for providing me the excellent opportunity to learn and stay.

Finally, I would like to thank all my family members and friends for their love and support.

Table of Contents

Abstract ii

Dedication iii

Acknowledgments iv

Table of Contents v

List of Tables vii

List of Figures viii

1 Introduction 1

2 Literature Review 4

2.1 Introduction to Poisson model 4

2.2 Introduction to longitudinal study 5

2.2.1 Longitudinal data 5

2.2.2 Longitudinal analysis 6

2.3 Introduction to hierarchical linear model 7

2.3.1 Data structure 7

2.3.2 Model specification 8

2.3.3 Model assumptions 10

2.3.4 Estimation of effects 10

2.3.5 Hypothesis Testing 12

2.4 Introduction to PLEC function 13

3 Two level random coefficient Poisson longitudinal modelling 14

3.1 Data 14

3.2 Model formulation and assumptions 16

3.2.1 Level-one model with PLEC function 16

3.2.2 Level-two model 19

3.2.3 Combined model 20

3.3 Computational method 20

3.3.1 Matrix form of model formulation 20

3.3.2 Fixed effects estimation 22

4 Data analysis 23

4.1 Model selection 23

4.2 Data analysis results 27

4.3 Model performance 30

5 Discussion and Conclusion 33

5.1 Fixed effects interpretation in terms of PLEC parameters 33

5.2 Conclusion 35

Bibliography 43

Vita

List of Tables

3.1 Longitudinal data of daily confirmed case from 21 regions in Italy . . 15

3.2 Regional covariates in 21 regions of Italy 16

4.1 Estimates fixed effects with standard errors in Model 1 24

4.2 Estimates fixed effects with standard errors in Full Model 26

4.3 Estimates fixed effects with standard errors in Reduced Model 27

4.4 Estimates fixed effect grouped by Level-one model parameters in Final Model 28

4.5 Predicted random effects for various regions in Italy 30

List of Figures

2.1 Data structure for the three-level hierarchical model 8

3.1 Map of Italy Republic 17

3.2 Regional daily new cases curve with (Left) average temperature and (Right) population density legend. 17

4.1 Cumulative case by Jun.30 with legend of (Left) average temperature and (Right) GDP per capita. 25

4.2 Boxplots of random effects 29

4.3 Multivariate normal probability plots on random effects 29

4.4 Model performance at regional level using first 60 days as training data with 95% prediction interval. 31

4.5 Model performance comparison at national level using first 45 and 60 days as training data. 32

5.1 Transmission curves in Lombardia with 0.5 scaled unit increase on (Left) average temperature and (Right) population density. 35

5.2 Transmission curves with 0.25 and 0.5 scaled unit increase on average temperature in each region. 36

5.3 Transmission curves with 0.25 and 0.5 scaled unit increase on population density in each region. 37

5.4 Transmission curves with 0.25 and 0.5 scaled unit increase on (Left) average temperature and (Right) population density in Italy. 37

5.5 Transmission curves in Italy updated to end of October. 39

Chapter 1

Introduction

Much of the effort on statistical modeling of coronavirus (COVID-19) spread has contributed to spreading patterns and forecasting of COVID-19 development; these patterns have been found to vary among locations ([24, 6]). As the survival time of coronaviruses on surfaces has found to depend on temperature, some researchers have explored the association of daily or cumulative confirmed cases with environmental factors ([1, 13, 21]). To forecast COVID-19 related new cases and deaths, various authors have employed deep learning, artificial intelligence (AI) and time series approaches. For example, recurrent neural network (RNN) based deep learning techniques is proposed by ([16]) for forecasting COVID-19 related new cases. For predicting the dynamical behavior of COVID-19 cases various artificial intelligence (AI) based modeling techniques such as Bayesian regression neural network, cubist regression, k-nearest neighbors, quantile random forest, and support vector regression were employed in ([5]). Autoregressive integrated moving average (ARIMA) based time series forecasting techniques were also incorporated in the literature to envision COVID-19 related new cases as well as deaths ([15, 14, 18, 4, 23]). Furthermore, some researchers have studied the link between daily fatality rates with regional factors such as health resources, but found no significant factors [3]. As

the spreading patterns of COVID-19 development vary a lot among locations, fitting regression models of daily confirmed cases or fatality rates directly with regional factors might not always reveal important relationships. In this study, we investigate the link between regional spreading patterns of COVID-19 development in Italy and regional factors in two steps. First, we extend the method of Zhang et al.[24] to handle longitudinal count in order to characterize regional spreading patterns of COVID-19 daily confirmed cases; the varying growth and declining patterns as well as turning points among regions in Italy have been well captured by regional regression parameters of this model. We then associate these regional regression parameters with regional factors. The effects of regional factors on regional spreading patterns of COVID-19 daily confirmed cases have been effectively evaluated.

The coronavirus disease 2019 (COVID-2019) has so far spread to over 200 countries causing more than 20 million confirmed cases globally by August 15, 2020. In addition to countermeasures such as lockdowns and social distancing, it is also important to explore and reveal the relationship between virus transmission and potential environmental covariates such as temperature, to help us better understand the behavior of the virus, and further implement more efficient public interventions to contain the spread. Italy, as one of the countries first affected by the virus with a total number of more than 240,000 confirmed cases by the end of June, has experienced the transmission periods from exponential growth, turning point to flattened tail, which provides us an ideally mature data set for relevant regression analysis. Furthermore, potential covariates such as demographic and economic status vary from region to region in the country, and specially, as Italy has a geographically narrow and long shape from north to south, plus the mountainous terrain in midland, temperature conditions also vary at regional level. The variability of covariates makes regression analysis approach more feasible. In addition, the fact that Italy has carried out a unified countermeasure policy across the country helps minimize various effects

brought by human intervention at regional level.

Our proposed hierarchical linear model deals with the analysis of the daily confirmed cases from February 24 to June 30, 2020 from various regions in Italy. To accommodate the heterogeneity among various regions, we consider regional level random effects. To depict Infection-Time relationship, we adopt a power law with exponential cutoff (PLEC) function ([24] [11]), which has exponential increase and decay terms. By incorporating the PLEC function into the hierarchical linear model, we can achieve the two major goals in this study: First, we construct a new hierarchical model to fit and predict the COVID-19 transmission trend in Italy, and compare it with the PLEC model ([24] [11]). Second, we conduct regression analysis on the relationship between COVID-19 transmission in Italy and spatially-varied covariates, including population density, GDP per capita and temperature. Our main interest in this study is the relationship between transmission and temperature.

Construction of the manuscript is as follows. After the proposed methodology in Section 2, we discuss our analysis results in Section 3. Some further discussions and conclusions of the proposed research are in Sections 4 and 5, respectively.

Chapter 2

Literature Review

2.1 Introduction to Poisson model

Poisson distribution was named after French mathematician Simeon Denis Poisson (1781 - 1840). It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event (Haight 1967)[8]. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

A random variable Y is said to have a Poisson distribution with parameter λ if it takes integer values $y = 0, 1, 2, ...$ with probability

$$Pr\{Y = y\} = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

for $\lambda > 0$. The mean and variance of this distribution can be shown to be

$$E(Y) = \text{var(Y)} = \lambda$$

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data. The Poisson distribution has been used in different fields, such as biological studies, finance and insurance, astronomy studies, etc.

2.2 Introduction to longitudinal study

A longitudinal study is a research design to measure intra-individual (e.g., people or object) change over time, involving repeated observations of the same variables. A main objective of longitudinal study is to relate some change over time in individuals to their characteristics, or to an experimental condition. For example, HIV patients may be followed over time with monthly measures such as CD4 counts, or cities can be followed over time with annual measures of economic growth. Such repeated measures data are correlated within subjects (patients and cities) and thus require special statistical analytical techniques. Longitudinal studies can be retrospective (looking back in time, thus using existing data such as medical records or social science data) or prospective (requiring experimental design and collection of new data). Our data analysis approach uses retrospective data of Italian Covid-19 daily confirmed cases, and we tracked the daily case for a certain length of time, thus it is a typical longitudinal analysis procedure.

2.2.1 Longitudinal data

Consider a sample of N randomly selected units with n_i measurements of response on each unit, $i = 1,...,N$,

$$\underset{n_i \times 1}{\boldsymbol{Y_i}} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

where the $\boldsymbol{Y_i}$ are independent vectors and n_i is a time-related index, and n_i may or may not be the same for all units i. Associated with the jth measurement on the ith unit is a vector with the following form:

$$\underset{p\times 1}{\boldsymbol{X_{ij}}} = \begin{pmatrix} X_{ij1} \\ \vdots \\ X_{ijp} \end{pmatrix}$$

and can be further writen as:

$$\underset{n_i\times p}{\boldsymbol{X_i}} = \begin{pmatrix} \boldsymbol{X_{i1}}^\intercal \\ \vdots \\ \boldsymbol{X_{ip}}^\intercal \end{pmatrix}$$

$\boldsymbol{X_i}$ is the design matrix, and its rows correspond to the covariates associated with different times of measurement, and the columns correspond to different variables. Thus each subject has a vector of outcomes, $\boldsymbol{Y_i}$, and a matrix of covariates, $\boldsymbol{X_i}$.

2.2.2 Longitudinal analysis

Given the assumptions that the the sample consists of N independently selected units, we can write the conditional mean of the jth outcome of unit i as a linear function of X_{ij}:

$$\begin{aligned} E(Y_{ij}|X_i) &= \boldsymbol{X_{ij}}^\intercal \boldsymbol{\beta} \\ &= \beta_1 X_{ij1} + ... + \beta_p X_{ijp} \end{aligned}$$

so that proper regression models for the marginal mean of Y_{ij} can be used to explore and analyze our interest, such as to estimate $\boldsymbol{\beta}$. We further have

$$\text{cov}(\boldsymbol{Y_i}) = \boldsymbol{\Sigma_i}$$

where $\boldsymbol{\Sigma_i}$ is some known function of the variance-covariance matrix, and mixed effects model and corresponding variance-covariance structures are usually used to handle the correlation of multiple measurements of variables, and it will be discussed in Section 2.3.

2.3 Introduction to hierarchical linear model

Hierarchical linear models (also known as multilevel models, mixed models) are statistical models with parameters varying at more than one level. This type of model is useful to explore relationships in hierarchical data structures, such as patients within hospitals, students within schools, and cities within counties. In our data analysis procedure, we chose 21 regions in Italy for longitudinal study, and each region has repeated measurements, thus it is a two-level hierarchical model.

2.3.1 Data structure

In hierarchical linear models, data is collected on random samples in lower-level cluster nested within higher-level unit. This hierarchical structure can be more than two levels. One example is students nested in classes, and classes nested in schools. In this example, students in the same classroom could share variance due to the effect from their teacher, and their classmates. Variances between classrooms could differ as different classes have different teachers and students, and the same explanation on variance parameter can be applied to a higher level in data structure. Figure 2.1 demonstrates this three-level hierarchical data structure idea. In this example, researchers may be interested in what factors affect students' average grades in each of the level. At the highest level, these factors could be location, regional budget, and at the class-level, teachers' experience, teaching methodology could have influence on the average grades in that class, and naturally, there could be factors like students'

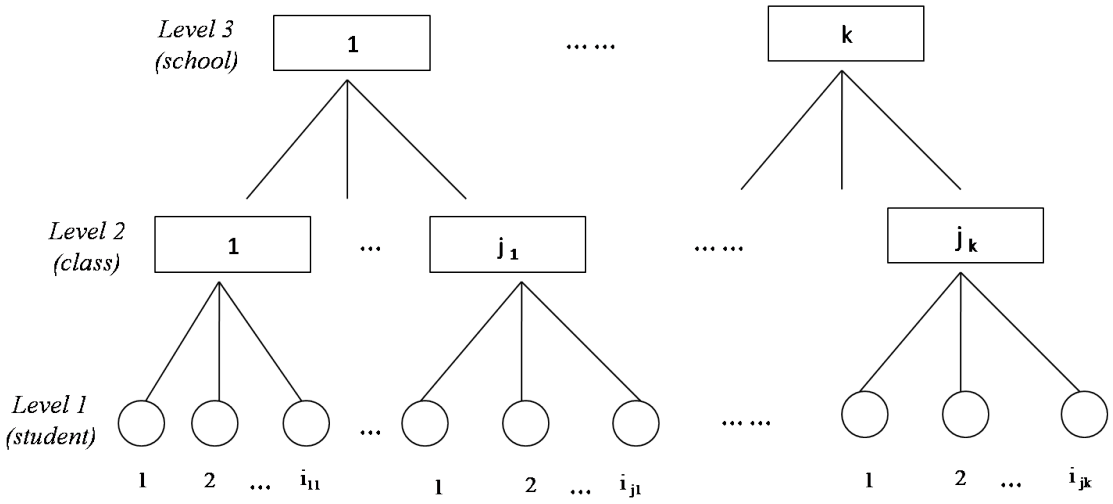

Figure 2.1: Data structure for the three-level hierarchical model

individual intelligence quotient (IQ) and behavior in studying at student-level. At the same time, variables at level-one are nested within level-two groups and share in common impact of level-two variables. For example, a teacher's teaching style could influence students' studying behavior, and this also applies to higher level structure, as teachers' experience could be affected by a school's budget.

2.3.2 Model specification

We will use a two-level hierarchical linear model as an example for discussion purpose in this section. Consider a continuous dependent variable Y (for example, students' grades), and a single, continuous level-one covariate X (for example, students' IQ), then level-one model has the following form:

$$Y_{ij} = \gamma_{0j} + \gamma_{1j} X_{ij} + \epsilon_{ij} \tag{2.1}$$

where Y_{ij} is the response variable measured on the ith level-one unit (student's grade), nested within the jth level-two unit (class), γ_{0j} is the intercept for the jth level 2 unit, X_{ij} is the level-one predictor (student's IQ), and γ_{1j} is the regression

coefficient associated with level-one predictor X for the jth level-two unit, and ϵ_{ij} is the random error associated with the ith level-one unit nested within the jth level-two unit. In the level-one model, we have J models as shown in (2.1), and each model could have different intercept and slope coefficients(γ_{0j} and γ_{1j}) as explained in previous subsection. Thus, in level-two models, we consider γ_{0j} and γ_{1j} as dependent variables and are related to appropriate level-two covariates. In our example, we set the level-two covariate (teacher's experience) as a continuous covariate, W, and we can then write the level-two model in the following expression:

$$\begin{cases} \gamma_{0j} &= \beta_{00} + \beta_{01} W_j + e_{0j} \\ \gamma_{1j} &= \beta_{10} + \beta_{11} W_j + e_{1j} \end{cases} \tag{2.2}$$

where γ_{0j} and γ_{1j} are intercept and slope for the jth level-two unit(class). β_{00} and β_{01} are the regression coefficients (intercept and slope) associated with level-one intercept term (γ_{0j}), and β_{10} and β_{11} are the regression coefficients (intercept and slope) associated with level-one slope term (γ_{1j}), and e_{0j} and e_{1j} are the random effects of the jth level-two unit (class) on intercept and slope, respectively. By substituting (2.2) into (2.1) we have the combined model:

$$Y_{ij} = (\beta_{00} + e_{oj}) + (\beta_{10} + e_{1j}) X_{ij} + \beta_{01} W_j + \beta_{11} W_j X_{ij} + \epsilon_{ij} \tag{2.3}$$

Now it is clear that the combined model is equivalent to a mixed effects model with random intercept and random slope, with both level-one and level-two covariates and an interaction term involved. We have four fixed-effect parameters β_{00} (intercept), β_{11}, β_{01} and β_{10}, and random effects are e_{0j} on intercept, and e_{1j} on X_{ij} predictor.

2.3.3 Model assumptions

Based on [17] Lisa M. Sullivan's work, we assume that errors term in level-one model are normal random variables with mean zero and common variance σ^2, and also assume that level-one coefficients γ_{0j}, and γ_{1j} follow a bivariate normal distribution with means α_0 and α_1 respectively, and variances τ_{00} and τ_{11} respectively. The covariances of the coefficients are denoted by τ_{01} and τ_{11}:

$$\begin{gathered} E(\epsilon_{ij}) = 0 \quad \text{var}(\epsilon_{ij}) = \sigma^2 \\ E(e_{0j}) = 0 \quad E(e_{1j}) = 0 \\ E(\gamma_{0j}) = \alpha_0 \quad E(\gamma_{1j}) = \alpha_1 \\ \text{var}(\gamma_{0j}) = \text{var}(e_{0j}) = \tau_{00} \quad \text{var}(\gamma_{1j}) = \text{var}(e_{1j}) = \tau_{11} \\ \text{cov}(\gamma_{0j}, \gamma_{1j}) = \text{cov}(e_{0j}, e_{1j}) = \tau_{01} \\ \text{cov}(e_{0j}, \epsilon_{ij}) = \text{cov}(e_{1j}, \epsilon_{ij}) = 0 \end{gathered} \tag{2.4}$$

2.3.4 Estimation of effects

Hierarchical models involve the estimation of two types of regression parameters. The first type of the parameter is fixed effects, which do not vary across groups. In our example in Section 2.3.2, these fixed effects are β_{00}, β_{01} and β_{10} and β_{11} in 2.3. Ordinary Least Squares (OLS) approach, which is widely used in linear regression, is no longer appropriate to estimate level-two fixed effects, as the assumption of homoscedasticity from linear regression is not met in hierarchical models, due to the fact that variance of level-one parameters are possible to vary across groups. Generalized Least Squared (GLS) estimate can be used to estimate fixed effects, as it generates a weighted level-two regression which ensures that groups (e.g., class) with more accurate estimates of the outcome variable are allocated more weight in the level-two regression equation. Math derivation to use GLS in matrix form will be discussed in next chapter.

The second type of the parameter is the random effects, e_{0j} and e_{1j} in our example, which are allowed to vary across groups (class). R package "lme4"[2] provides Best Linear Unbiased Predictors (BLUPs), and it is interacted with the method to estimate parameters of variance-covariance matrix.

To estimate variance-covariance matrix parameter, we write:

$$\mathbf{e}_j \sim N(0, \boldsymbol{G}), \quad \boldsymbol{G} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}. \tag{2.5}$$

where matrix $\boldsymbol{G}$ is multivariate normal with mean zero and a 2×2 variance-covariance matrix, and we can further write it into parameterized form:

$$\boldsymbol{G_\theta} = \sigma^2 \boldsymbol{\Sigma_\theta} \boldsymbol{\Sigma_\theta}^\intercal \tag{2.6}$$

where $\boldsymbol{\Sigma}$ is positive semi-definite, and now we express the variance-covariance matrix in terms of a relative covariance factor $\boldsymbol{\Sigma_\theta}$, which is 2 by 2 matrix, depending on the variance-component parater $\boldsymbol{\theta}$. R pakage "lme4" provides restricted maximum likelihood (REML) criterion to estimate parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and σ^2 by minimizing the following log-likelihood expression[12]:

$$-2L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_{obs}) = log\frac{|\boldsymbol{L_\theta}|^2}{|\boldsymbol{W}|} + n\log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta})}{\sigma^2} + \frac{||\boldsymbol{R}_X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})||}{\sigma^2} \tag{2.7}$$

where $r^2(\boldsymbol{\theta})$ is the penalized weighted residual sum-of-squares in terms of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ^2. $\boldsymbol{W}$ is a diagonal matrix of known prior weights, $|\boldsymbol{L_\theta}|^2$ is the Jacobian determinant of the transformation from $\boldsymbol{u}$ to $\boldsymbol{v}$, and $|\boldsymbol{R}_X|$ is the Jacobian determinant of the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{v}$, and

$$\boldsymbol{v} = \boldsymbol{R}_X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) \tag{2.8}$$

The REML estimate of σ^2 is[12]

$$\hat{\sigma_\theta^2} = \frac{r^2(\boldsymbol{\theta})}{n-p} \tag{2.9}$$

and the final REML criterion has the following form[12]:

$$-2L_R(\boldsymbol{\theta}|\boldsymbol{y}_{obs}) = log\frac{|\boldsymbol{L_\theta}|^2|\boldsymbol{R_X}|^2}{|\boldsymbol{W}|} + (n-p)[1+\log(\frac{2\pi r^2(\boldsymbol{\theta})}{n-p})] \tag{2.10}$$

2.3.5 Hypothesis Testing

We will focus on discussing the hypothesis test for fixed effects, which is our main interest in data analysis chapter. We know that the test statistic can be computed by taking the ratio of the ML (or REML) estimate to its estimated standard error as follows:

$$t = \frac{\hat{\beta_k}}{\sqrt{\hat{\text{var}}(\hat{\beta_k})}} \tag{2.11}$$

This hypothesis test is versus H_0: $\beta_k = 0$. the right part of formula(2.11) follows a t-distribution. In "lmer" function under "lme4" package, a t-as-z approach is used to achieve hypothesis testing results[10]. This approach simply uses the z distribution to obtain p-values from the Wald t-values provided. The logic behind this is that to use t distribution, we need to know the degrees of freedom, but in linear mixed model, it is usually not clear what the appropriate degrees of freedom are. We also know that t distribution begins to approximate the z distribution as degrees of freedom increase, and at infinite degrees of freedom they are identical. Most data sets analyzed using mixed models contain at least hundreds of data point. Out data used in next chapter has more than 2000 data points, so this t-as-z approach from "lme4" package is reasonable.

2.4 Introduction to PLEC function

PLEC (power law with exponential cutoff) model has the following general form:

$$I = cT^w \exp(dT) \tag{2.12}$$

Ihe model was was initially used for modeling the classic species-area relationship, and in Ma[11]'s work, it was also used for more general diversity-area relationship. $exp(dT)$ in (2.5) is the exponential decay term that acts as taper-off parameter to eventually overwhelm the power law behavior at large value of T. The parameters in (2.5) can be used for hierarchical regression, and by interpreting regression results directly on these parameters, we can explore the relationship between level-two predictors and the response' growth trend. This will be further discussed in next chapter.

Chapter 3

Two level random coefficient Poisson longitudinal modelling

3.1 Data

We first briefly discuss the dataset used in the manuscript before introducing the model. We collected daily confirmed cases data in Italy from Github [7] website from February 24, 2020 to June 30, 2020, from each of the 21 regions in Italy (including autonomous areas). There are situations when daily new cases are negative in certain regions due to data correction by the Italian authority, and we treat these observations as 0. Therefore, we have longitudinal observations for 128 consecutive days in 21 regions. Data structure is demonstrated in Table (3.1).

Italian demographic and economic data were collected from Wikipedia [20], including each region's population density and Gross Domestic Product (GDP) per capita. We also collected each region's historical daily average temperature from January 1 to June 30, 2020 on Wunderground website [22], and then an average temperature in Fahrenheit during this period was calculated for each region. Regional covariates data is demonstrated in Table (3.2).

Table 3.1: Longitudinal data of daily confirmed case from 21 regions in Italy

Region No.	Region	Day Serial	Date	Daily Case
1	Abruzzo	1	02/24	0
1	Abruzzo	2	02/25	0
1	Abruzzo	3	02/26	0
...	...	...	...	...
1	Abruzzo	48	04/11	106
1	Abruzzo	49	04/12	40
1	Abruzzo	50	04/13	53
...	...	...	...	...
1	Abruzzo	126	06/28	1
1	Abruzzo	127	06/29	1
1	Abruzzo	128	06/30	0
...	...	...	...	...
...	...	...	...	...
21	Veneto	1	02/24	33
21	Veneto	2	02/24	10
21	Veneto	3	02/24	28
...	...	...	...	...
21	Veneto	48	04/11	347
21	Veneto	49	04/12	309
21	Veneto	50	04/13	174
...	...	...	...	...
21	Veneto	126	06/28	11
21	Veneto	127	06/29	3
21	Veneto	128	06/30	8

To better understand the data structure as well as regional covariates' effects, we present the map of Italy with various regions presented in Figure 3.1. We consider regional factors such as population density, GDP per capita, and average temperature on the transmission of COVID-19 virus in Italy. To accommodate the different longitudinal patterns among the regions, we introduce three region-specific random effects, corresponding the three parameters in the PLEC function. Figure 3.2 demonstrates how fixed and random effects affect the transmission of the virus in different regions. Curves in different regions generally follow patterns that can be described by some PLEC curves, experiencing exponential growth and decay process, but the starting and turning point, as well as the growth rate of each curve differ across the

Table 3.2: Regional covariates in 21 regions of Italy

No.	Region	Pop.Density	Ave.Temperature	GDPPerCap.
1	Abruzzo	121	52.46	25000
2	Basilicata	56	60.61	21400
3	P.A. Bolzano	71.78	51.65	42900
4	Calabria	128	59.25	17400
5	Campania	424	56.09	18500
6	Emilia-Romagna	199	51.99	35800
7	Friuli Venezia Giulia	153	51.79	31400
8	Lazio	341	55.99	33200
9	Liguria	286	57.18	32000
10	Lombardia	422	52.06	38500
11	Marche	162	52.30	27200
12	Molise	69	60.61	20100
13	Piemonte	172	51.01	30700
14	Puglia	206	54.69	18700
15	Sardegna	68	57.64	20900
16	Sicilia	194	60.23	17700
17	Toscana	162	54.00	30900
18	P.A. Trento	87.18	51.65	36600
19	Umbria	104	50.89	24800
20	Valle d'Aosta	39	48.47	35700
21	Veneto	267	51.27	33500

regions in Italy, due to the influence from both fixed and random effects. By adding legends on average temperature and population density, some fixed effects from these two covariates can be visualized: Regions with higher average temperature appear to have curves at lower position, while curves of higher population density regions are at higher positions in the graph.

3.2 Model formulation and assumptions

3.2.1 Level-one model with PLEC function

In this section we present the PLEC function which will be used in the level-one model. It is noteworthy to point out that the PLEC function has a satisfactory



Figure 3.1: Map of Italy Republic

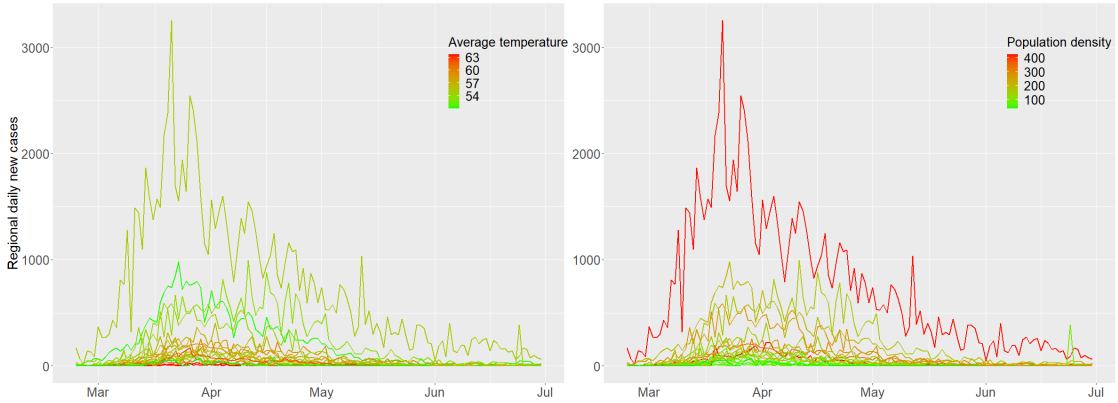

Figure 3.2: Regional daily new cases curve with (Left) average temperature and (Right) population density legend.

performance on depicting COVID-19 transmission curves in different countries [11, 19]. Following Ma [11],Wei and Zhang [19]'s results, the PLEC function has the following form

$$I = cT^w \exp(-dT) \tag{3.1}$$

where T is the time(in days) since the transmission begins, I is the number of daily new infections, c, w, and d are three positive parameters to be estimated from the data. Each of the parameters has its biological interpretation: cT^w term is a power function with w dominating the transmission growth with a multiplier c, and the exponential term $\exp(-dT)$ characterizes the declining trend of the transmission curve. Following [24], we can get the maximum value of infection by taking the derivative of $df(I)/dT$ and setting it to zero as,

$$I_{\max} = c\left(\frac{w}{d}\right)^w \exp(-w), \tag{3.2}$$

which occurs at the turning point of

$$T_{\max(\text{peak})} = \frac{w}{d}. \tag{3.3}$$

In order to incorporate the PLEC function with spatially-varied covariates in Italy, we assume that the daily confirmed case, which is a count response, follows a Poisson distribution as,

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}),$$

where

$$\mu_{ij} = e^{c_j - d_j T_i} T_i^{w_j}, \tag{3.4}$$

is the PLEC function in an alternative form, and e^{c_j} corresponds to c in Equation (3.1). In (3.4), μ_{ij}, the mean value of daily new cases, on the ith level-one unit (T), representing 128 longitudinal daily observations from February 24 to June 30,

nested within jth level-two (regional level) unit, where j = 1,2,...,21, representing 21 regions in Italy. The maximum value of infection and its turning point in jth region can be derived as

$$\mu_{\max,j} = e^{c_j - w_j} \left(\frac{w_j}{d_j}\right)^{w_j}, \quad T_{\max,j} = \frac{w_j}{d_j}. \tag{3.5}$$

If we further apply log-transformation, Equation (3.4) can be expressed as

$$\log(\mu_{ij}) = c_j + w_j \log(T_i) + (-d_j T_i). \tag{3.6}$$

In (3.6), T, and its log form $\log(T)$ are level-one (day) predictors, which take identical day serial values across regions, with $-d_j$ and w_j as coefficients, and c_j can be treated as the intercept term.

3.2.2 Level-two model

In level-two model, we consider the regression coefficients c_j, w_j, and $-d_j$ from level-one models as dependent variables incorporated with corresponding level-two covariates.

$$\begin{cases} c_j &= \alpha_0 + \mathbf{x}_j^\intercal \boldsymbol{\beta}_0 + e_{0j} \\ w_j &= \alpha_1 + \mathbf{x}_j^\intercal \boldsymbol{\beta}_1 + e_{1j} \\ -d_j &= \alpha_2 + \mathbf{x}_j^\intercal \boldsymbol{\beta}_2 + e_{2j} \end{cases} \tag{3.7}$$

In our case, $\mathbf{x}_j$ is the vector of spatially-varied covariates at regional level, incorporated with level-one regression coefficients, and it may contain multiple covariates of interest. In addition, the $\mathbf{x}_j$ expression can be different for the three equations in (2.7). $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are coefficients of level-two covariates and α_0, α_1, α_2 are corresponding intercepts. The residual terms e_{0j}, e_{1j}, and e_{2j} are the random effects of the jth level-two unit (region) on the level-one parameters c_j, w_j, and $-d_j$ respectively. Following the derivation of Sullivan et al.[17], We assume that the error terms $\mathbf{e}_j = (e_{0j}, e_{1j}, e_{2j})^\intercal$ follow a tri-variate normal with mean (0,0,0) and an unknown

covariance matrix G.

3.2.3 Combined model

We can derive our combined model by substituting (3.7) into (3.6):

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log(\mu_{ij}) &= (\alpha_0 + e_{0j}) + (\alpha_1 + e_{1j})\log(T_i) + (\alpha_2 + e_{2j})T_i \\ &+\mathbf{x}_j^\intercal \boldsymbol{\beta}_0 + \mathbf{x}_j^\intercal \boldsymbol{\beta}_1 \log(T_i) + \mathbf{x}_j^\intercal \boldsymbol{\beta}_2 T_i \end{aligned} \tag{3.8}$$

Now both level-one and level-two covariates (time and spatially-varied covariates), interaction terms, and error terms are included in one model. It is then clear that the hierarchical formulation is equivalent to the random-intercept-and-random-slope model.

3.3 Computational method

3.3.1 Matrix form of model formulation

We derive our computational method in matrix form. First, we rewrite Equation (3.6) into the following expression,

$$\log(\boldsymbol{\mu}_j) = \mathbf{y}_j = D_j \boldsymbol{\gamma}_j, \tag{3.9}$$

where $\log(\boldsymbol{\mu}_j) = (\log(\mu_{1j}), \log(\mu_{2j}), \ldots, \log(\mu_{128,j}))^\intercal$, is the log form of mean response in 128 days in jth region, and $\boldsymbol{\gamma}_j = (c_j, w_j, -d_j)^\intercal$, is the parameter vector at level-

one. D_j is a known design matrix with following form identical across regions,

$$D_j = \begin{bmatrix} 1 & \log(1) & 1 \\ 1 & \log(2) & 2 \\ . & . & . \\ . & . & . \\ 1 & \log(128) & 128 \end{bmatrix}. \tag{3.10}$$

The first column in D_j is a designed column for intercept term c_j, and third and second columns take day serial values and their log forms. We further rewrite level-two equation in the following matrix form

$$\boldsymbol{\gamma}_j = W_j\boldsymbol{\beta} + \mathbf{e}_j, \tag{3.11}$$

where $\boldsymbol{\beta} = (\alpha_0, \boldsymbol{\beta}_0^\intercal, \alpha_1, \boldsymbol{\beta}_1^\intercal, \alpha_2, \boldsymbol{\beta}_2^\intercal)^\intercal$ is a vector of level-two coefficients, including both intercept and slope terms, which need to be estimated in the regression, and $\mathbf{e}_j = (e_{1j}, e_{2j}, e_{3j})^\intercal$ is a residual vector for random effects associated with three covariates from level-one model. As the known designed matrix of $\boldsymbol{\beta}$, W_j takes the following form as

$$W_j = \begin{bmatrix} 1 & \mathbf{x}_j^\intercal & 0 & \mathbf{0}^\intercal & 0 & \mathbf{0}^\intercal \\ 0 & \mathbf{0}^\intercal & 1 & \mathbf{x}_j^\intercal & 0 & \mathbf{0}^\intercal \\ 0 & \mathbf{0}^\intercal & 0 & \mathbf{0}^\intercal & 1 & \mathbf{x}_j^\intercal \end{bmatrix}, \tag{3.12}$$

where $\boldsymbol{\beta}$, $\mathbf{x}_j$ and $\mathbf{0}$ have the same length, depending on the number of level-two covariates we select in $\mathbf{x}_j$. For a combined model, it has the following matrix form,

$$\begin{aligned} log(\boldsymbol{\mu}_j) &= D_j\boldsymbol{\gamma}_j = D_j(W_j\boldsymbol{\beta} + \mathbf{e}_j) \\ &= D_jW_j\boldsymbol{\beta} + D_j\mathbf{e}_j, \end{aligned} \tag{3.13}$$

where

$$\mathbf{e}_j \sim N(0, G), \quad G = \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{bmatrix}. \tag{3.14}$$

3.3.2 Fixed effects estimation

We define $A_j = D_j W_j$, and it has a matrix expression as following:

$$A_j = \begin{bmatrix} 1 & \mathbf{x}_j^\intercal & \log(1) & \log(1)\mathbf{x}_j^\intercal & 1 & 1\mathbf{x}_j^\intercal \\ 1 & \mathbf{x}_j^\intercal & \log(2) & \log(2)\mathbf{x}_j^\intercal & 2 & 2\mathbf{x}_j^\intercal \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & \mathbf{x}_j^\intercal & \log(128) & \log(128)\mathbf{x}_j^\intercal & 128 & 128\mathbf{x}_j^\intercal \end{bmatrix} \tag{3.15}$$

Then we can use weighted least squares method or generalized least squares method to estimate $\boldsymbol{\beta}$ by applying following formula:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (A^\intercal \hat{V}^{-1} A)^{-1} A^\intercal \hat{V}^{-1} Y \\ V &= \mathrm{var}(Y) = DGD^\intercal \end{aligned} \tag{3.16}$$

A, D are design matrices across regions, and $\hat{V}$ can be achieved with G's maximum likelihood estimates. Variances of $\hat{\boldsymbol{\beta}}$ can be estimated by:

$$\widehat{\mathrm{var}(\boldsymbol{\beta})} = (A^\intercal \hat{V}^{-1} A)^{-1} \tag{3.17}$$

As our data is balanced with identical day counts in each region, G can be estimated by using closed-form maximum likelihood formulae. Random effects can be predicted by using best linear unbiased prediction (BLUP) method.

Chapter 4

Data analysis

4.1 Model selection

In this section we present our analysis results. We first include all potential spatially-varied covariates of interest into the level-two model, including regional average temperature (AveTemp), GDP per capita (GDPPerCap) and population density (PopDens). By inserting these level-two covariates into (3.7) and (3.8), we have the expressions of the first model as follows:

$$\begin{cases} c_j &= \alpha_0 + \beta_{00}\text{AveTemp}_j + \beta_{01}\text{PopDens}_j + \beta_{02}\text{GDPPerCap}_j + e_{0j} \\ w_j &= \alpha_1 + \beta_{10}\text{AveTemp}_j + \beta_{11}\text{PopDens}_j + \beta_{12}\text{GDPPerCap}_j + e_{1j} \\ -d_j &= \alpha_2 + \beta_{20}\text{AveTemp}_j + \beta_{21}\text{PopDens}_j + \beta_{22}\text{GDPPerCap}_j + e_{2j} \end{cases} \tag{4.1}$$

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= (\alpha_0 + e_{0j}) + (\alpha_1 + e_{1j})\log(T_i) + (\alpha_2 + e_{2j})T_i + \\
&\beta_{00}\text{AveTemp} + \beta_{01}\text{PopDens} + \beta_{02}\text{GDPPerCap} + \\
&\beta_{10}\text{AveTemp} * \log(T_i) + \beta_{11}\text{PopDens} * \log(T_i) + \beta_{12}\text{GDPPerCap} * \log(T_i) \\
&+\beta_{20}\text{AveTemp} * T_i + \beta_{21}\text{PopDens} * T_i + \beta_{22}\text{GDPPerCap} * T_i
\end{aligned}
\tag{4.2}
$$

We use R function "glmer" in "lme4" package [2] to estimate the coefficients and predict random effects by specifying a nested structure at regional level on all three level-one covariates, which are the intercept term, day and log(day). Before applying "glmer" function, all level-two covariates were scaled so that the algorithm can converge more easily. We present the results of Model 1 in Table 4.1. Based

Table 4.1: Estimates fixed effects with standard errors in Model 1

Coefficients	Fixed effects	Estimate	Std.Error	z value	P-value
α_0	Intercept	-7.3347	0.5248	-13.977	$< 2e-16$
α_1	log(day)	4.6780	0.1918	24.395	$< 2e-16$
α_2	day	-0.1330	0.0062	-21.486	$< 2e-16$
β_{00}	AveTemp	0.0885	0.7353	0.120	0.9042
β_{01}	PopDens	2.3096	0.5264	4.388	1.15e-05
β_{02}	GDP per capita	1.8953	0.7234	2.620	0.0088
β_{10}	log(day)*AveTemp	-0.3293	0.2689	-1.225	0.2206
β_{11}	log(day)*PopDens	-0.6027	0.1923	-3.134	0.0017
β_{12}	log(day)*GDPPerCap	-0.6909	0.2644	-2.613	0.0090
β_{20}	day*AveTemp	0.0160	0.0086	1.857	0.0633
β_{21}	day*PopDens	0.0185	0.0062	2.980	0.0029
β_{22}	day*GDPPerCap	0.0225	0.0086	2.628	0.0086

on the p-values, the three spatially-varied covariates or their interaction terms have significant effects at 0.1 significance level. However, we need to take the geographic factor into consideration when studying virus transmission, and certain geographic transmission trend could produce confounding variables. When COVID-19 broke out in Italy, it was believed that the virus hit the north first, and then it started

to spread to the south. Therefore, if level-two covariates have a corresponding geographic trend, they could be confounded. Figure 4.1 helps visualize the confounding

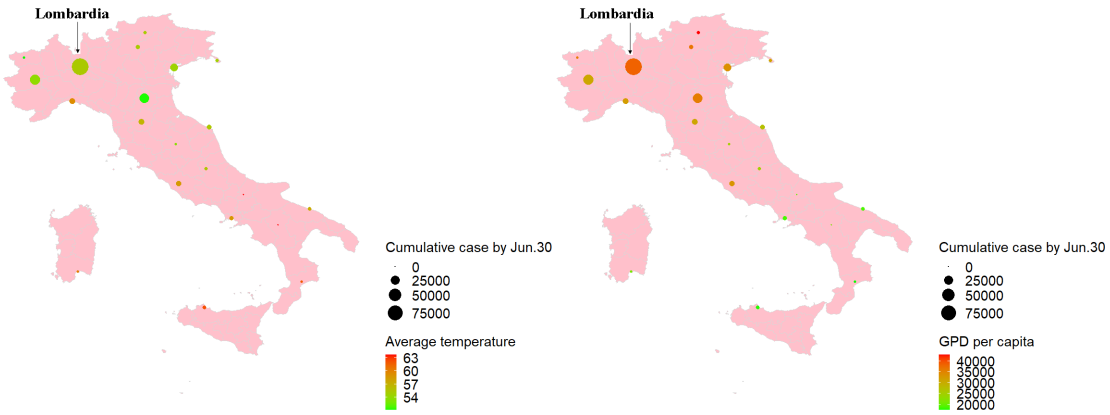

Figure 4.1: Cumulative case by Jun.30 with legend of (Left) average temperature and (Right) GDP per capita.

effects caused by the geographic pattern. The size of the circles, with gradient color, represent number of cumulative cases, and the circles' growth trend can be observed by adding animation effects. Figure 3.2 implies that transmission associated with both average temperature and GDP per capita also follows a certain geographic pattern from north to south. To incorporate this potential confounding geographic effect, we introduce a new covariate "Distance" into the level-two model. The coordinates are used to calculate the geographic distance between each region and region "Lombardia", where people believe the virus first started to spread. The level-two covariate matrix in this "full model" becomes:

$$\begin{aligned} \boldsymbol{\beta}_0 &= (\beta_{00}, \beta_{01}, \beta_{02}, \beta_{03})^\intercal \\ \boldsymbol{\beta}_1 &= (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13})^\intercal \\ \boldsymbol{\beta}_2 &= (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23})^\intercal \\ \mathbf{x}_j &= (\text{AveTemp}_j, \text{PopDens}_j, \text{GDPPerCap}_j, \text{Distance}_j)^\intercal \end{aligned} \tag{4.3}$$

Results from the full model are displayed in Table 4.2. After bringing "Distance" covariate into the model, all GDP per capita related terms become insignificant at

any conventional level of significance, so the GDP per capita covariate could be confounded with distance, and thus should be dropped.

Table 4.2: Estimates fixed effects with standard errors in Full Model

Coefficients	Fixed effects	Estimate	Std.Error	z value	P-value
α_0	Intercept	-7.3336	0.5166	-14.197	$< 2e-16$
α_1	log(day)	4.6775	0.1897	24.659	$< 2e-16$
α_2	day	-0.1330	0.0061	-21.862	$< 2e-16$
β_{00}	AveTemp	0.2644	0.7534	0.351	0.7256
β_{01}	PopDens	2.0910	0.5314	4.157	3.22e-05
β_{02}	GDP per capita	1.1906	1.1067	1.0760	0.2820
β_{03}	Distance	-0.9478	1.1468	-0.827	0.4085
β_{10}	log(day)*AveTemp	-0.3821	0.2767	-1.381	0.1673
β_{11}	log(day)*PopDens	-0.5721	0.1951	-2.932	0.0034
β_{12}	log(day)*GDPPerCap	-0.4770	0.4071	-1.172	0.2413
β_{13}	log(day)*Distance	0.2873	0.4213	0.682	0.4953
β_{20}	day*AveTemp	0.0181	0.0088	2.055	0.0399
β_{21}	day*PopDens	0.0173	0.0063	2.761	0.0058
β_{22}	day*GDPPerCap	0.0134	0.0131	1.062	0.2881
β_{23}	day*Distance	-0.0115	0.0136	-0.848	0.3966

After dropping the level-two covariate of GDP per capita, we have the following result in Table (4.3):

The only interaction term that is not statistically significant is log(day)*AveTemp (P-value = 0.254), and we use backward stepwise method to drop AverageTemp*log(T) term, thus we have our final level-two model which can be expressed in the following form:

$$\begin{cases} c_j &= \alpha_0 + \beta_{00}\text{AveTemp}_j + \beta_{01}\ \text{PopDens}_j + \beta_{02}\text{Distance}_j + e_{0j} \\ w_j &= \alpha_1 + \beta_{11}\text{PopDens}_j + \beta_{12}\text{Distance}_j + e_{1j} \\ -d_j &= \alpha_2 + \beta_{20}\text{AveTemp}_j + \beta_{21}\text{PopDens}_j + \beta_{22}\text{Distance}_j + e_{2j} \end{cases} \tag{4.4}$$

Table 4.3: Estimates fixed effects with standard errors in Reduced Model

Fixed effects	Estimate	Std.Error	z value	P-value
Intercept	-7.3339	0.5311	-13.808	$< 2e-16$
log(day)	4.6777	0.1961	23.858	$< 2e-16$
day	-0.1330	0.0062	21.307	$< 2e-16$
AveTemp	0.1087	0.7590	0.143	0.8861
PopDens	2.1443	0.5437	3.944	8.01e-05
Distance	-1.9013	0.7572	-2.511	0.01204
log(day)*AveTemp	-0.3194	0.2800	-1.141	0.2540
log(day)*PopDens	-0.5462	0.2006	-2.722	0.0065
log(day)*Distance	-0.0226	0.0089	-2.538	0.0111
day*AveTemp	0.016298	0.0089	1.837	0.0663
day*PopDens	0.0165	0.0064	2.588	0.0097
day*Distance	0.6691	0.2793	2.395	0.0166

By substituting (4.4) into (3.8), we have the combined final model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
log(\mu_{ij}) &= (\alpha_0 + e_{0j}) + (\alpha_1 + e_{1j}) \log(T_i) + (\alpha_2 + e_{2j}) T_i + \\
&\beta_{00}\text{AveTemp} + \beta_{01}\text{PopDens} + \beta_{02}\text{Distance} + \\
&\beta_{11}\text{PopDens} * \log(T_i) + \beta_{12}\text{Distance} * \log(T_i) + \\
&\beta_{20}\text{AveTemp} * T_i + \beta_{21}\text{PopDens} * T_i + \beta_{22}\text{Distance} * T_i
\end{aligned}
\tag{4.5}
$$

The final model is also equivalent to a random-intercept-and-random-slope model.

4.2 Data analysis results

The values of estimated fixed effects using the above mentioned final model are given in Table 4.4. It turns out that all remaining terms are significantly associated with COVID-19 transmission at 0.1 significance level, including the one of our main interests, temperature. Comparing the final model with results in Table 4.1, we find that bringing the geographic covariate "Distance" makes all GDP per capita related terms insignificant, but it does not change much on coefficients of population density

related terms. As of our main interest, the sign and magnitude of coefficients related to average temperature effect have changed, compared with results in Table 4.1. We

Table 4.4: Estimates fixed effect grouped by Level-one model parameters in Final Model

L1	L2 Coef.	Fixed effects	Estimate	Std.Error	z value	P-value
c	α_0	(Intercept)	-7.3458	0.5431	-13.526	$< 2e-16$
c	β_{00}	AveTemp	-0.7244	0.2132	-3.398	0.0007
c	β_{01}	PopDens	2.2358	0.5509	4.0580	4.94e-05
c	β_{02}	Distance	-1.3180	0.5741	-2.2960	0.0217
w	α_1	log(day)	4.6820	0.2010	23.3060	$< 2e-16$
w	β_{11}	log(day)*PopDens	-0.5811	0.2036	-2.855	0.0043
w	β_{12}	log(day)*Distance	0.4453	0.2047	2.175	0.0296
-d	α_2	day	-0.1331	0.0064	-20.862	$< 2e-16$
-d	β_{20}	day*AveTemp	0.0070	0.0036	1.9310	0.0534
-d	β_{21}	day*PopDens	0.0175	0.0065	2.7040	0.0069
-d	β_{22}	day*Distance	-0.0161	0.0070	-2.2960	0.0217

also present predicted random effects results from “lme4” package [2] using BLUP method in Table 4.5. Random effects have a range of -3.98 to 4.22 on intercept, a range from -0.075 to 0.063 on ”day” covariate, and a range from -2 to 1.63 on ”logday” covariate. We can visualize the ranges of random effects on each of the level-one covariates in Figure 4.2 , and further detect possible outliers of Region Molise, P.A. Trento and Umbria, as indicated in the figure. To check whether our predicted random effects satisfy the assumptions of the multivariate normal distribution, we draw normal probability plots and present results in Figure 4.3. Our graphs in Figure 4.3 show that the predicted random effects satisfy the multivariate normality assumption, and possible outliers from normality plots are consistent with the results in box plots.

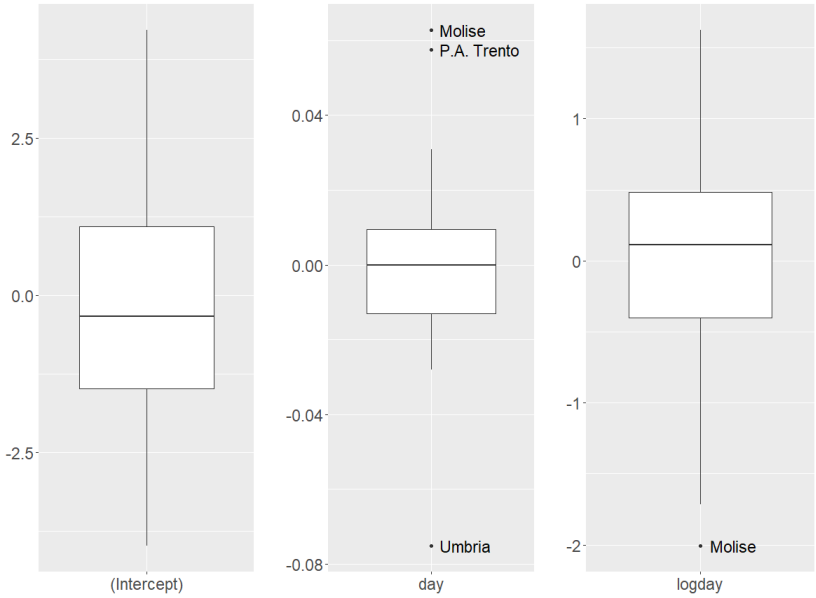

Figure 4.2: Boxplots of random effects

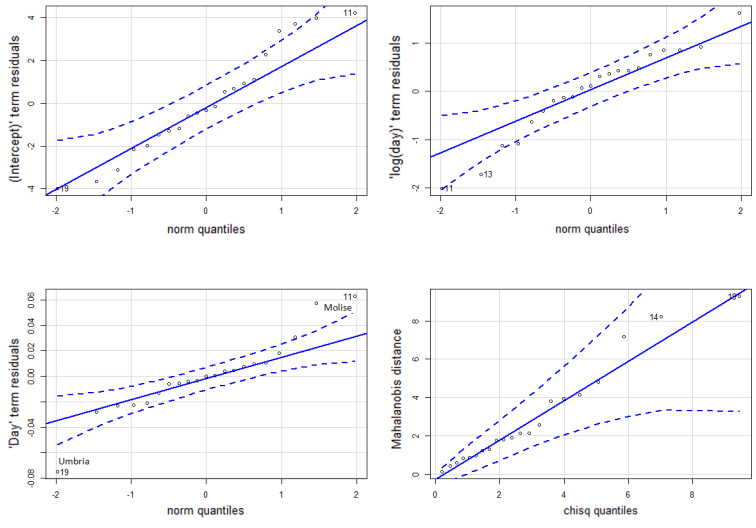

Figure 4.3: Multivariate normal probability plots on random effects

Table 4.5: Predicted random effects for various regions in Italy

Region	Intercept (e_0)	logday (e_1)	day (e_2)
Abruzzo	-1.4778	0.3643	4.8068e-03
Basilicata	-0.1659	0.0777	-5.8952e-03
Calabria	1.0968	-0.1965	-1.8788e-05
Campania	-3.1115	0.7663	-2.2389e-02
Emilia-Romagna	3.3658	-1.0811	3.0806e-02
Friuli Venezia Giulia	0.6879	-0.4002	1.0400e-02
Lazio	-0.4633	-0.1086	9.4084e-03
Liguria	-1.1703	0.3127	-5.6459e-03
Lombardia	2.2558	-0.6270	7.4296e-03
Marche	3.7051	-1.1192	1.7932e-02
Molise	4.2204	-2.0073	6.2852e-02
P.A. Bolzano	-2.1480	0.8520	-2.8011e-02
P.A. Trento	3.9894	-1.7144	5.7475e-02
Piemonte	-1.9840	0.8564	6.7807e-04
Puglia	-1.3040	0.4321	3.9145e-03
Sardegna	-0.5926	0.4834	-2.1216e-02
Sicilia	0.5414	0.1143	-4.0148e-03
Toscana	-0.3367	0.4356	-1.3019e-02
Umbria	-3.9813	1.6267	-7.5136e-02
Valle d'Aosta	-3.6609	0.9154	-2.3201e-02
Veneto	0.9086	-0.1221	-3.6045e-03

4.3 Model performance

To check the performance of the proposed model we use first days of observations as the training data to construct the model to predict the daily new cases in the rest of the days, and then check the deviation. With the variance-covariance matrix of both fixed effects and random effects given in "lme4" [2] results of our final model, We can further calculate a prediction interval for our proposed model by using R package "merTools" [9]. Figure 4.4 demonstrates the results using the first 60 days of data to construct the model and then to predict the rest of the days in each region.

Once we sum the fitted and predicted values from every region, we can compare our proposed model with models only using the PLEC function at national level. Based on the result from Figure 4.5, the proposed approach has less deviation in expected

prediction than only using the PLEC function.

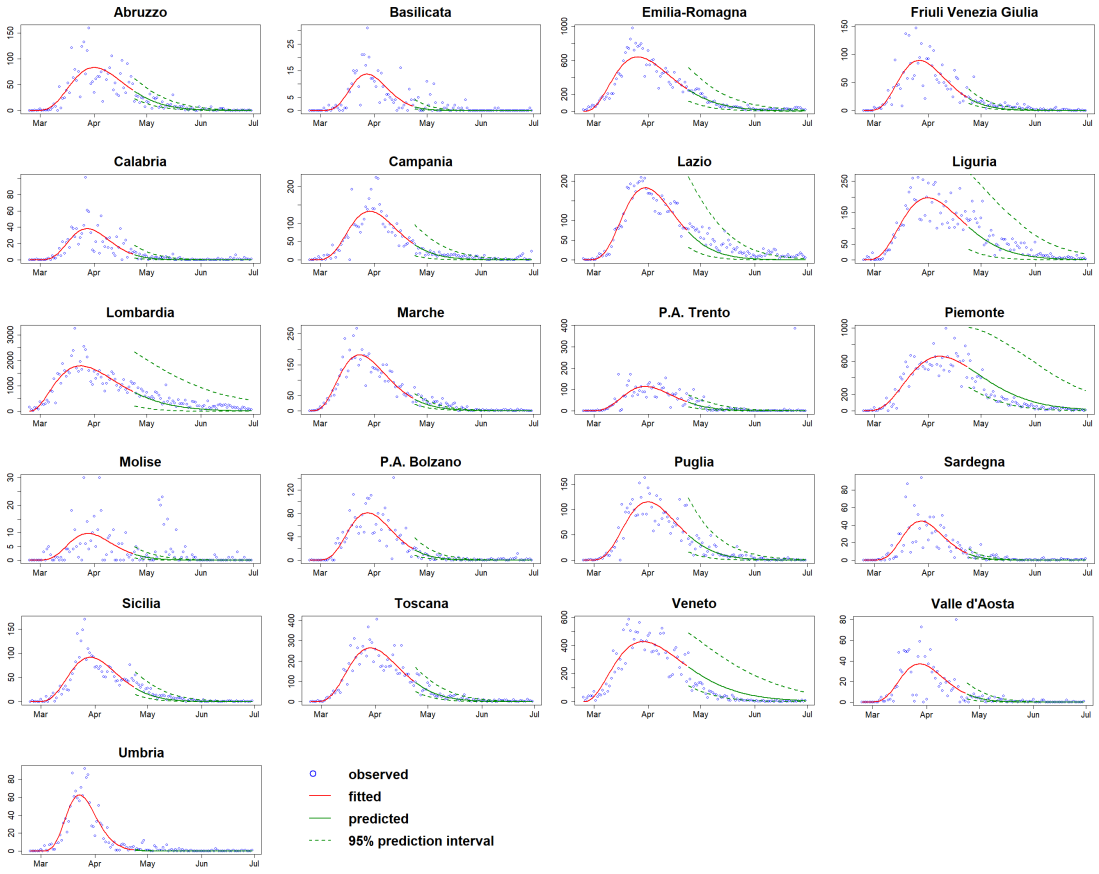

Figure 4.4: Model performance at regional level using first 60 days as training data with 95% prediction interval.

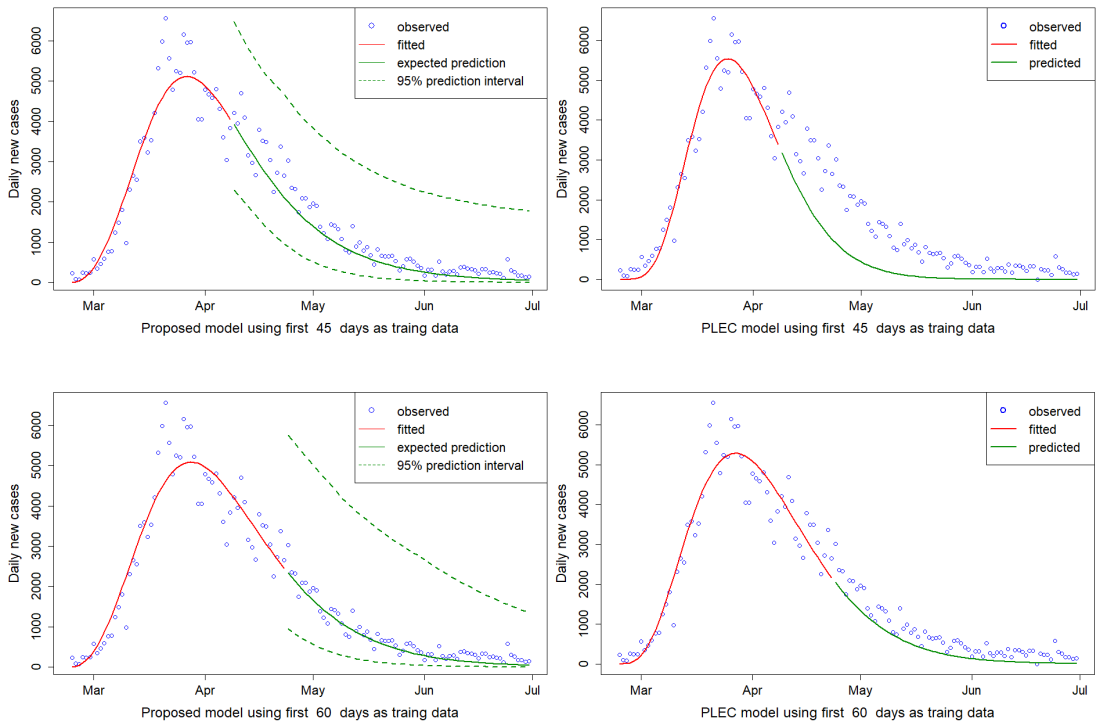

Figure 4.5: Model performance comparison at national level using first 45 and 60 days as training data.

Chapter 5

Discussion and Conclusion

5.1 Fixed effects interpretation in terms of PLEC parameters

Results in Table 4.4 show that all three estimated effects associated with population density are significant at 0.01 significance level, the effect of "AverageTemperature" is significant at 0.01 significance level, and that associated with the interaction term "day*AverageTemperature" is significant at 0.1 significance level. So we have decided to keep "day*AverageTemperature" term in the model as it is an interaction term of our main interest and its p-value is close to 0.05. In order to link results from combined model to PLEC parameters, we group coefficients in Table 4.4 by their corresponding level-one parameters. Thus, we can directly interpret these covariates' fixed effects on parameters c, w and $-d$ respectively.

Our results indicates that average temperature is negatively associated with parameter c. Holding other variables constant, the increase of every one scaled unit average temperature will lead to a 0.7244 decrease on parameter c, which downsizes the scale of the gamma curve. Also, due to the fact that average temperature is positively associated with parameter $-d$, the same increase on average temperature

will cause $-d$ to increase 0.0069. As a result, with time parameter T as a multiplier, $d * T$ will counter off the shrinkage effect from parameter c. Thus, the effect from change of average temperature on mean value of daily case is also related to time predictor, and we can derive its pattern by following approach. Based on (3.4), we set $\mu_0 = e^{c_0 - d_0 T} T^{w_0}$ and $\mu_1 = e^{c_1 - d_1 T} T^{w_1}$, where μ_1 is the new function by only increasing x scaled units average temperature ($x > 0$) while other variables are all held constant, thus, $c_1 - c_0 = -0.7244x$, $(-d_1) - (-d_0) = d_0 - d_1 = 0.00699x$, and $w_0 = w_1$. Further, we take $\mu_1/\mu_0 = e^{(c_1 - c_0) + T(d_0 - d_1)} = e^{-0.7244x + 0.00699xT}$. It is clear that $\mu_1 > \mu_0$ when $-0.7244 + 0.00699T > 0$, and it is free from x, which means at day 104, the effect from parameter $-d$ will exceed that from parameter c, and before this, it is always true that $\mu_1 < \mu_0$. This pattern is applicable to all regions. Similarly, we can also derive the effect of temperature change on the maximum value of infection and its turning point based on the result in (3.5). We set $\mu_{\max,0} = e^{c_0 - w_0}(w_0/d_0)^{w_0}$, $T_{\max,0} = w_0/d_0$, and $\mu_{\max,1} = e^{c_1 - w_1}(w_1/d_1)^{w_1}$, $T_{\max,1} = w_1/d_1$. Thus, $T_{\max,1}/T_{\max,0} = d_0/d_1 = d_0/(d_0 - 0.00699x) > 1$, which means higher average temperature leads to a delayed turning point, and $\mu_{\max,1}/\mu_{\max,0} = e^{c_1 - c_0}(T_{\max,1}/T_{\max,0}) = e^{-0.7244x}(d_0/(d_0 - 0.00699x))^{w_0}$, suggesting that whether or not the ratio is larger than 1 depends on the value of parameter $-d$ and w in each region. However, from the previous derived property, we know that as long as $T_{\max,0}$ occurs before the 104th day, $\mu_{\max,1} < \mu_{\max,0}$, which means in every region in Italy, increase of average temperature will reduce maximum infections and postpone the turning point. Similarly, we can derive the properties of effects from population density change. From Table 4.4, we know that population density is positively associated with parameter c and $-d$, and negatively associated with parameter w. Therefore, the final effect from population density on mean response depends on a comprehensive results from all three parameters. We assume an increase of x scaled units ($x > 0$) on population density while all other variables are held con-

stant. Thus $c_1 - c_0 = 2.2358x$, $w_1 - w_0 = -0.5811x$, and $d_0 - d_1 = 0.0175x$. We then derive $\mu_1/\mu_0 = e^{(c_1-c_0)+(d_0-d_1)T}T^{w_1-w_0} = e^{2.2358x+0.0175xT}T^{-0.5811x}$. If we set $d(\mu_1/\mu_0)/dT = 0$ and solve for T, it can be shown that μ_1/μ_0 takes its minimal value when $T = 33.206$. Thus, $(\mu_1/\mu_0)_{\min} = e^{2.8169x}33.2057^{-0.5811x}$. We can further derive $(d(\mu_1/\mu_0)/dx)_{\min} = 0.7815e^{2.8169x}33.2057^{-0.5811x} > 0$ (strictly increasing) at x's domain, and $(\mu_1/\mu_0)_{\min} = 1$ when $x = 0$, so it's proved that $\mu_1 > \mu_0$ when $x > 0$, which suggests the number of mean infections will always become larger when population density is increased.

Figure 5.1 demonstrates the pattern illustrated above in Region Lombardia. The pattern is consistent across regions in the country, as shown in Figures 5.2, 5.3 and 5.4.

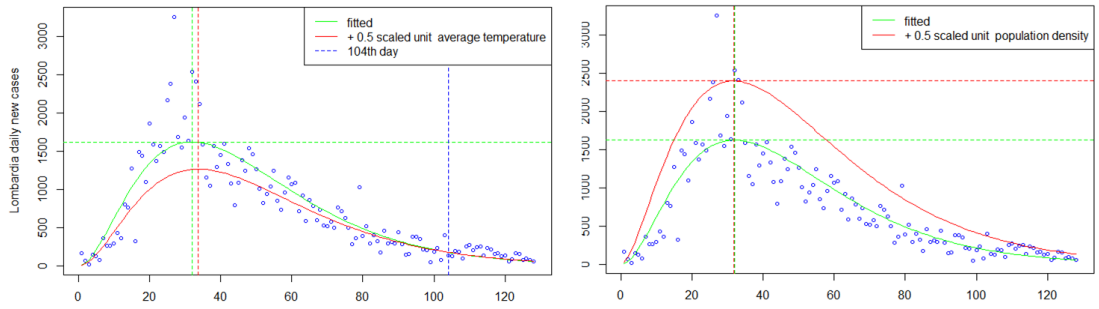

Figure 5.1: Transmission curves in Lombardia with 0.5 scaled unit increase on (Left) average temperature and (Right) population density.

5.2 Conclusion

In this study, we have examined the association between regional spreading patterns of COVID-19 development in Italy and regional factors in two steps. In the first step, we have symbolized the regional spreading patterns of the daily confirmed cases of COVID-19 in Italy by a special patterned Poisson regression model for repeated daily

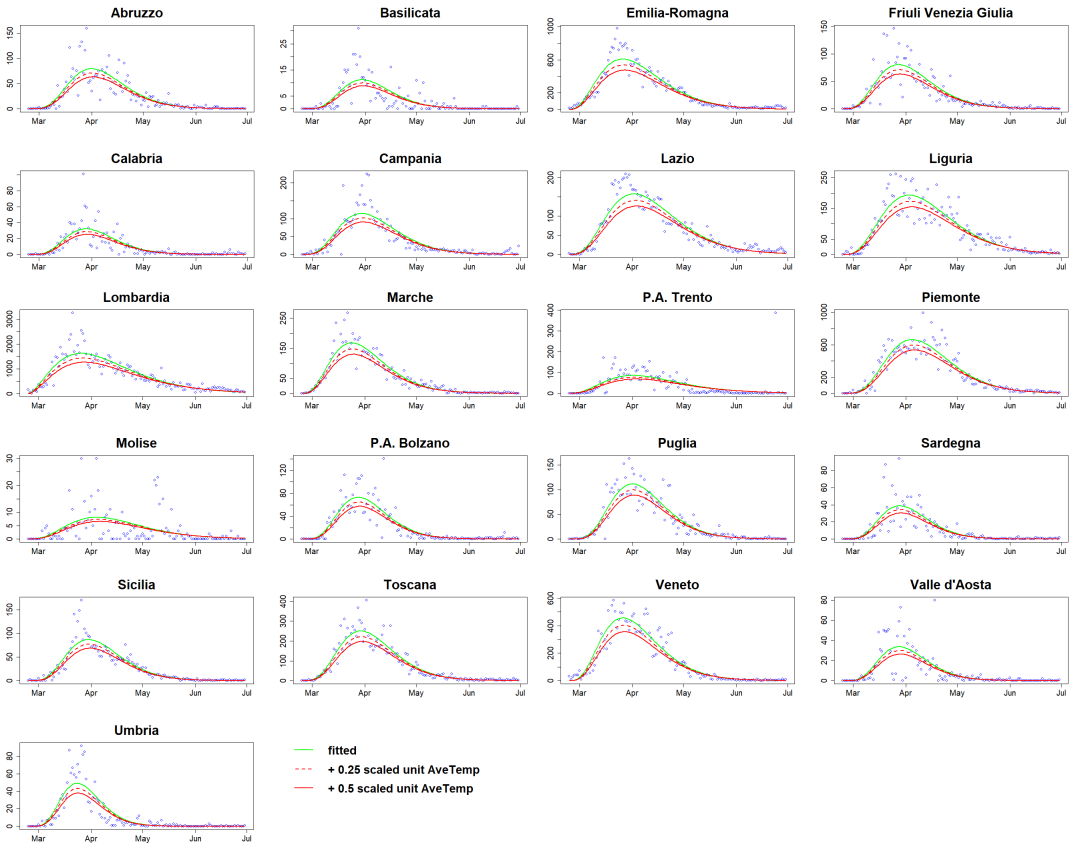

Figure 5.2: Transmission curves with 0.25 and 0.5 scaled unit increase on average temperature in each region.

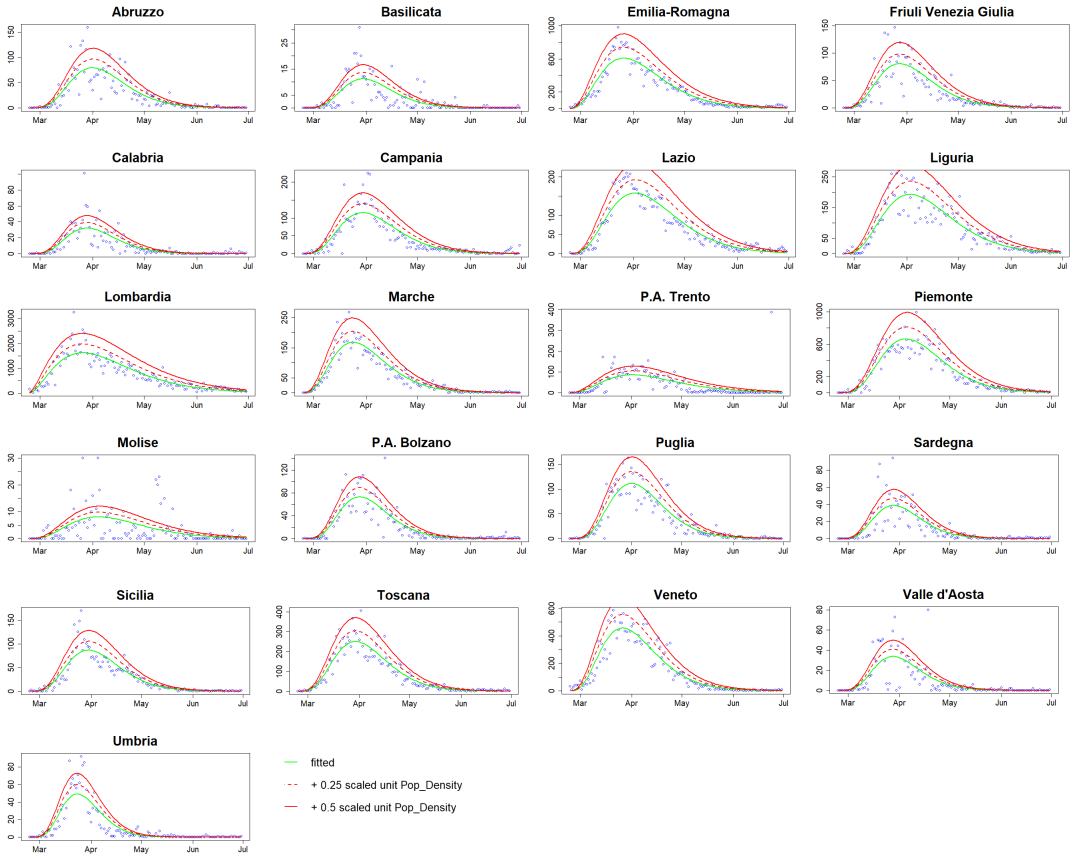

Figure 5.3: Transmission curves with 0.25 and 0.5 scaled unit increase on population density in each region.

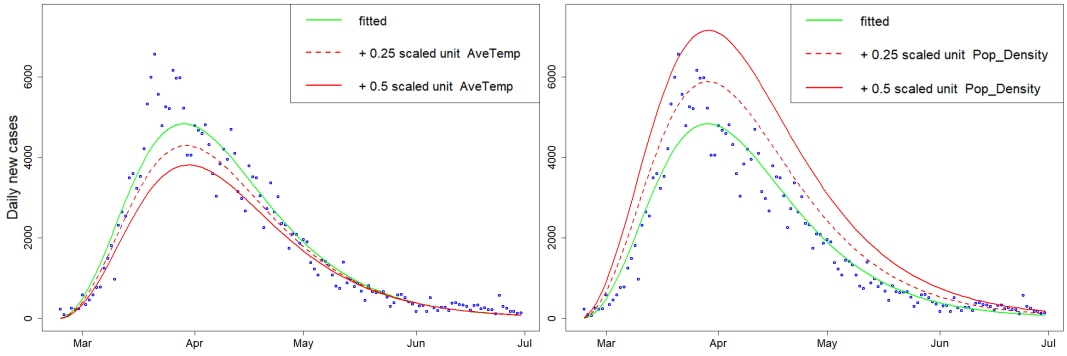

Figure 5.4: Transmission curves with 0.25 and 0.5 scaled unit increase on (Left) average temperature and (Right) population density in Italy.

counts. Our proposed regression model can capture the varying growth and declining patterns as well as turning points among regions in Italy by regional regression parameters. We then incorporate these regional regression parameters with regional factors as a second step to effectively evaluate the affect of regional factors on regional spreading patterns of COVID-19 daily confirmed cases.

Our two level longitudinal model performs better on prediction than models only using the PLEC function. This is because of the fact that two level model can incorporate regional covariates into the modelling approach, increasing the goodness of fit and predictive power. It also gives some mathematical properties linking to the level-one PLEC parameters, inferring that regional covariates average temperature and population density are associated with COVID-19 infection. This methodology could be useful in future epidemiological study when transmission pattern is potentially associated with social or environmental covariates. Neither our modelling approach nor PLEC function alone works well if we use data before the turning point to construct the model, due to the fact that lockdowns policy plays a crucial role affecting the curve. To incorporate this effect, we can use segmented models suggested by Zhang et al [24] and Wei and Zhang [19].

Our analysis results indicate that geographical patterns perform an important role as it could confound other variables in epidemiological studies. During the model selection process, we dropped GDP related effects as they were confounded with geographic factor "Distance". This is highly likely true as when the virus originally breaks out, it indeed starts from one certain region, or several regions, and then starts to spread, so geographic factor could play a more important role than other economic or environmental covariates at the first stage, and this was explained by the fact that "Distance" effect always had very high significance and made other terms insignificant or less significant in model selection process.

However, the main modelling and data analysis part were finished before September,

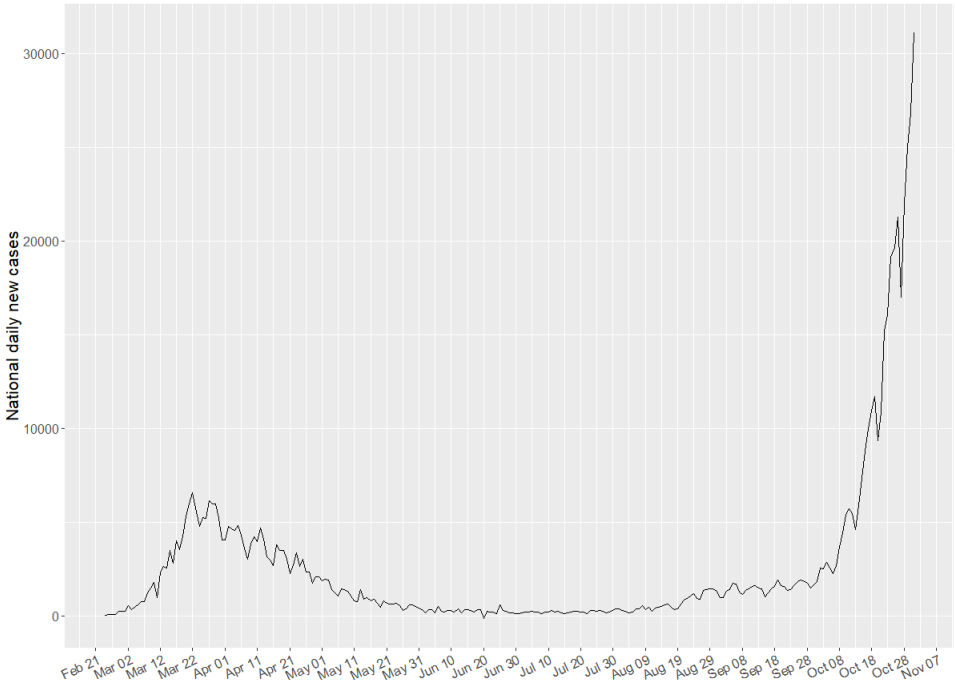

Figure 5.5: Transmission curves in Italy updated to end of October.

when there was no appearance of the second wave of the Covid-19 virus, as shown in Figure (5.5). Intuitively, we know that economic activity could play a very important role in the spread of the virus, due to the fact that economic activity could be correlated with human interaction, and this fact is especially important when all regions have the same start in the second wave of spreading (geographic factor could be less important). GDP Per Capita's effect should be considered in the next stage of analysis.

Considering the fact that the curve now has two peaks, or even more in future, ordinary PLEC function won't be applicable if we use data that has multiple peaks. Possible solutions are using piece-wise data to make sure each period of data only has one peak, and then apply the methodology from this paper to conduct regression analysis, or finding more appropriate level-one function to incorporate the overall multi-peak trend. Besides, more complicated models specifying geographic correlation could be further explored.

A final point of caution is that this modelling procedure is not a designed experiment, which means that it can only explore associations between potential covariates and response, but can not reveal causal relationship, and there might be more confounding or lurking variables to be found.

Bibliography

[1] M.F. Bashir, B. Ma, Bilal, and M. Bashir, *Correlation between climate indicators and covid-19 pandemic in new york, usa*, Science of The Total Environment **728** (1 August 2020), 10.1016/j.scitotenv.2020.138835.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker, *Fitting linear mixed-effects models using lme4*, Journal of Statistical Software **67** (2015), 1–48.

[3] K. Cai and W. He, *Covid-19 fatality: A cross-sectional study using adaptive lasso penalized sliced inverse regression*, Journal of Data Science **18 (3)** (July 2020), 483–494.

[4] Zeynep Ceylan, *Estimation of covid-19 prevalence in italy, spain, and france*, Science of The Total Environment **729** (2020), https://doi.org/10.1016/j.scitotenv.2020.138817.

[5] R. G. da Silva, M. H. Dal Molin Ribeiro, V. C. Mariani, and L. dos Santos Coelho, *Forecasting brazilian and american covid-19 cases based on artificial intelligence coupled with climatic exogenous variables*, Chaos, Solitons & Fractals **139** (2020), https://doi.org/10.1016/j.chaos.2020.110027.

[6] A. Feoli, A.L. Iannella, and E. Benedetto, *Three pictures of covid-19 behavior in italy: similar growth and different degrowth*, medRxiv (2000), https://doi.org/10.1101/2020.05.09.20096149.

[7] Github, https://github.com/pcm–dpc/COVID–19/tree/master/dati–regioni.

[8] Frank Avery Haight, *Handbook of the poisson distribution*, (1967).

[9] Jared E. Knowles and Carl Frederick, *mertools: Tools for analyzing mixed effect regression models*, Journal of Statistical Software (2020), https://CRAN.R–project.org/package=merTools.

[10] Luke and Steven G, *Evaluating significance in linear mixed-effects models in r*, Behavior research methods **49** (2017), no. 4, 1494–1502.

[11] Z.S. Ma, *A simple mathematical model for estimating the inflection points of covid-19 outbreaks*, medRxiv **Cold Spring Harbor Laboratory Press** (2020), DOI: 10.1101/2020.03.25.20043893.

[12] D. Bates. M. Mächler and M. Bolker, *Fitting linear mixed-effects models using lme4*, cran.r (2015), DOI 10.18637/jss.v067.i01.

[13] D.N. Prata, W. Rodrigues, and P.H. Bermejo, *Temperature significantly changes covid-19 transmission in (sub) tropical cities of brazil*, Science of The Total Environment **729** (10 August 2020), //doi.org/10.1016/j.scitotenv.2020.138862.

[14] M. H. Dal Molin Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, *Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil*, Chaos, Solitons & Fractals **135** (2020), https://doi.org/10.1016/j.chaos.2020.109853.

[15] A. K. Sahai, N. Rath, V. Sood, and M. P. Singh, *Arima modelling & forecasting of covid-19 in top five affected countries*, Diabetes & Metabolic Syndrome: Clinical Research & Reviews **14** (2020), 1419–1427.

[16] S. Shastri, K. Singh, S. Kumar, and P. Kour, *Time series forecasting of covid-19 using deep learning models: India-usa comparative case study*, Chaos, Solitons & Fractals **140** (2020), https://doi.org/10.1016/j.chaos.2020.110227.

[17] L.M. Sullivan, K.A. Dukes, and E. Losina, *Tutorial in biostatistics: An introduction to hierarchical linear modelling*, Statistics in medicine **18** (1999), 855–888.

[18] A. Vena, D. R. Giacobbe, A. Di Biagio, and M. Mikulska, *Clinical characteristics, management and in-hospital mortality of patients with coronavirus disease 2019 in genoa, italy*, Clinical Microbiology and Infection (2020), https://doi.org/10.1016/j.cmi.2020.07.049.

[19] W. Wei and X. Zhang, *An updated analysis of turning point, duration and attack rate of covid-19 outbreaks in major western countries with data of daily new cases*, Data Brief (2020), doi: 10.1016/j.dib.2020.105830.

[20] Wikipedia, https://en.wikipedia.org/.

[21] Y. Wu, W. Jing, J. Liu, and Q. Ma, *Effects of temperature and humidity on the daily new cases and new deaths of covid-19 in 166 countries*, Science of The Total Environment **729** (10 August 2020), //doi.org/10.1016/j.scitotenv.2020.139051.

[22] Wunderground, https://www.wunderground.com/history.

[23] Q. Yang, J. Wang, H. Ma, and X. Wang, *Research on covid-19 based on arima model$^{\delta}$—taking hubei, china as an example to see the epidemic in italy*, Journal of Infection and Public Health **13** (Issue 10, October 2020,), 1415–1418.

[24] X.L. Zhang, R.J. Ma, and L. Wang, *Predicting turning point, duration and attack rate of covid-19 outbreaks in major western countries*, Chaos, Solitons & Fractals **10.1101** (2020), 109829.

Vita

Candidate's full name: Youtian Hao

University attended : Master of Science, December 2020, University of New Brunswick, Fredericton, NB, Canada

Master of Business Administration, 2016, Lanzhou University, Lanzhou, Gansu, China

Bachelor of Economics, 2010, Tianjin Foreign Studies University, Tianjin, China