Journal: HV; Manuscript #: 43-2014HV0499

Estimate of influenza cases using generalized linear, additive and mixed models

Manuel Oviedo[2], Angela Domínguez[3,4], and M Pilar Muñoz[1,4,*]

[1]Department of Statistics and Op. Research; Technical University of Catalonia; Spain

[2]Department of Statistics and Op. Research; University of Santiago de Compostela; Spain

[3]Department of Public Health; University of Barcelona; Spain

[4]CIBER Epidemiología y Salud Pública; Carlos III Institute of Health; Madrid. Spain

Keywords: Influenza, mortality, incidence rate, hospitalization
Abbreviations: GLM, Generalized Lineal Model; GAM, Generalized Additive Model; GAMM, Generalized Additive Mixed Model; NegBin, Negative Binomial, adj,, adjusted; df, degrees of freedom; edf, estimated degrees of freedom Pr, Probability; sq, square.
*Correspondence to: M Pilar Muñoz; Email: pilar.munyoz@upc.edu

We investigated the relationship between reported cases of influenza in Catalonia (Spain). Covariates analyzed were: population, age, data of report of influenza, and health region during 2010–2014 using data obtained from the SISAP program (Institut Catala de la Salut - Generalitat of Catalonia). Reported cases were related with the study of covariates using a descriptive analysis. Generalized Linear Models, Generalized Additive Models and Generalized Additive Mixed Models were used to estimate the evolution of the transmission of influenza. Additive models can estimate data dependence such as serial correlation in the residuals of the model; and mixed models can measurement of the variability in factor variables using random effects. The incidence rate of influenza was calculated as the incidence per 100 000 people. The mean rate was 13.75 (range 0–27.5) in the winter months (December, January, February) and 3.38 (range 0–12.57) in the remaining months. Statistical analysis showed

that Generalized Additive Mixed Models were better adapted to the temporal evolution of influenza (serial correlation 0.59) than classical linear models.

Introduction

Influenza is an epidemic disease that is transmitted from person to person and causes mortality due to complications, mainly pulmonary and cardiovascular, in older age population. Seasonal flu is a serious public health problem that causes serious illness and death in high-risk populations. According to the World Health Organization,[1] seasonal influenza circulates worldwide and can affect anybody in any age group. Several causes make the flu virus very contagious. First of all, it is easily transmitted from person to person through the air we breathe, thus affecting anyone of any age group. A second problem of this disease is that this virus circulates around the world, with a very clear seasonal component. In this way, the virus causes annual epidemics in temperate areas of the world, during the winter. We must not forget the economic cost of this disease, due to the loss of labor productivity in the period related to the flu epidemic. Thus, influenza vaccination is recommended for preventing infection in high risk people; however, we must not forget that the flu virus is a mutant virus that changes throughout different influenza seasons, developing resistance to influenza antiviral medications.

Several authors monitor mortality as an indicator of influenza. The model proposed by Dominguez et al.,[2] based on general mortality, was useful for detecting epidemic activity of influenza. In that analysis, the indicator that best predicted large scale epidemic activity was reported morbidity, and mortality could be considered a complementary indicator. The main result was that, when the influence of one model on another was studied, it was seen that morbidity was influenced by mortality registered in the previous weeks, but the mortality series did not seem to be affected by previously

reported cases of influenza-like illness. This was a very surprising result for us, which allowed us to conclude that not only was mortality a good indicator of influenza activity in our milieu, but moreover it was independent of notified morbidity. In this way, Muñoz et al.,[3] studied the behavior of influenza with respect to morbidity and all-cause mortality in Catalonia, and their association with influenza vaccination coverage. Vaccination coverage was associated with a reduction in influenza associated morbidity but not with a reduction in all-cause mortality, concluding that all-cause mortality was a good indicator of influenza surveillance and vaccination coverage was associated with a reduction in influenza associated morbidity but not with all-cause mortality. This result is very close to the results obtained in Dominguez et al.,[2] thus demonstrating the consistency of these results.

The aim of this study was to investigate the relationship between reported cases of influenza in Catalonia (Spain) and develop an appropriate statistical model in order to understand and correctly predict flu epidemics.

Results

A total 9753 reported cases of influenza in Catalonia (Spain) were obtained during the years 2010–2014. These reported cases have been divided into three groups: 3202 cases under 14 y old; 4015 cases between 15 and 64 y old; and 2536 cases older than 64 y old.

Now we just present the two most popular models in this methodology: GAM with family Poisson and GAM with family Negative Binomial. Finally, a summary table comparing the different models related to this subject is presented.

Model 1: *GAM with family Poisson and log as a Link function*

Equation 1 fits the total number of cases as a function of Month (as a factor), Week day, FluSeason, Population obtained and the smoothed covariate s(day.year) which takes into account the day of the week for each year of study

Total cases ~+factor(Month)+Week.day+FluSeason+Population+s(day.year)+ ε (eq.1)

where ε is a random noise.

The statistical results for this model are in Table 1 showing us that the intercept and factor(Month)2, factor(Month)11 and factor(Month)12 are significant in relation to the intercept. This indicates that there is an increase in flu cases in months with winter temperature, as expected. The days of the week are all significant as well. The values presented by this table are the difference of these days with regard to Sunday, which is the day that is not on the table. FluSeason2011–2012, which picks up the number of cases of the flu for the season 2011–2012, is also significant in respect the intercept variable, showing us a decrease in the number of cases for this season, while FluSeason2012–2013 does not exhibit significant differences from the previous season. The remaining coefficients are also significant differences either.

In addition, this model also estimates, by a smoothing function, the smoothed parameter s(day.year) and the results are shown in **Table 2**.

The conclusion of this part is that smoothed term s(day.year) is significant and the statistics R-sq.(adj) has the value: 0.961, that in other words means that the 96.1% of the variability in the values of the number of cases of this model can be estimated by means of significant variables introduced in the model .

Model 2: *GAM with family Negative Binomial and log as a Link function*

Again, equation 2 fits the total number of cases as function of Month (as a factor), Week day, FluSeason, Population obtained and the smoothed covariate s(day.year) that takes into account the day of the week for each year of study

Total cases ~+factor(Month)+Week.day+FluSeason+Population+s(day.year)+ε (eq.2) where ε is a random noise.

The expression of eq.2 is the same as the eq.1; however the main difference with eq. 1 is very important. Now the probability family is not Poisson and it is negative binomial.

The results for this model are shown in **Table 3**.

Comparing these results with those of model 1 show that this model has worse results because they are only significant results related to the day of the week and the population and not with the monthly factor, as in the previous model in which the month factor detected the months were significant.

Model comparison

At this point, we will make two comparisons, one graphical and other numerical. In the following plot we will see graphically which models are better suited to the data obtained. **Figure 1** shows the value of the predicted cases vs the observed cases (+).

It is very clear that the GAM (Poisson) and GAMM (Poisson) models are the best performers in the sense that in the sense that fit better the evolution of the data, i.e., these two models predict very well the future daily values.

Conclusions

Table 4 shows the R^2-adj, mean square error (MSE) and relative MSE. The best models in terms on R^2-adj are GLM, link Poisson and with link NegBin, however in terms of MSE, the best model is the GAMM model (Generalized Additive Mixed Model) and in terms on Relative MSE the winner is the GAM model link Poisson. Those results give us an idea that is not easy to select the “best” model and we have to analyze the different properties of each model and to choose a model according to our needs.

Material and Methods

Three different approaches have been used to estimate the number of cases of influenza in Catalonia (Spain): Generalized linear model (GLM), the first model is the simplest one.

GAM (Generalized Additive Model) using a Poisson distribution and log as a link function. This model was originally introduced by Hastie and Tibshirani[4] and consists in replacing the coefficients associated with the covariates X_1, X_2,…, X_p, i.e., the linear form obtained in a multiple linear regression $\Sigma\ \beta_j\ X_j$ by a sum of smooth functions $\Sigma\ s_j\ (X_j)$ thereby obtaining a model in which the covariates exhibit a nonlinear behavior.

The main difference between the GAM and GAMM model is that GAM model only contains smooth functions to model the fixed covariate effects while GAMM[5] model allow more flexible functional dependence of the response variable on the covariates by adding fixed and random effects to the linear predictors.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was partially funded by CIBER Epidemiología y Salud Pública (CIBERESP), Spain and by AGAUR (expedient number 2014/SGR 1403).

References

<eref>1. World Health Organization. Influenza (seasonal). Fact sheet No.211, March 2014. Available from: http://www.who.int/topics/influenza/en/. Accessed 25 Juny 2014 http://www.who.int/topics/en/</eref>

<jrn>2. Domínguez A, Muñoz P, Martínez A, Orcau A. Monitoring mortality as an indicator of influenza in Catalonia, Spain. J Epidemiol Community Health 1996; 50:293-8; PMID:8935461; http://dx.doi.org/10.1136/jech.50.3.293</jrn>

<jrn>3. Muñoz MP, Soldevila N, Martínez A, Carmona G, Batalla J, Acosta LM, Domínguez A. Influenza vaccine coverage, influenza-associated morbidity and all-cause mortality in Catalonia (Spain). Vaccine 2011; 29:5047-52; PMID:21620915; http://dx.doi.org/10.1016/j.vaccine.2011.04.067</jrn>

<jrn>4. Hastie T, Tibshirani R. Generalized additive models. Stat Sci 1986; 1:297-318; http://dx.doi.org/10.1214/ss/1177013604</jrn>

<jrn>5. Schneipl F.Fitting generalized additive mixed models based on the mixed model algorithm. R-Package: amer. 2011</jrn>

Figure 1. Predicted number of cases as function of the predicted values of the estimated models

Table 1. Parametric coefficients in model 1

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	28.55	7.76	3.67	0.00236
factor(Month)2	15.51	0.04	3.82	0.000136
factor(Month)3	-0.08	0.07	-1.19	0.233926
factor(Month)10	0.19	0.46	0.42	0.673592
factor(Month)11	1.08	0.29	3.65	0.000263
factor(Month)12	0.98	0.10	9.32	9< 2e-16
Monday	1.89	0.05	35.28	< 2e-16
Tuesday	1.63	0.05	29.98	< 2e-16
Wednesday	1.54	0.05	28.10	< 2e-16
Saturday	0.21	0.07	3.05	0.002270
Friday	1.32	0.05	23.72	2e-16
FluSeason2011-2012	-46.68	7.48	-6.25	4.07e-10
FluSeason2012-2013	-12.30	10.57	-1.16	0.244442
Population	3.37e-06	4.49e-07	7.52	5.35e-14

Table 2. Approximate significance of smooth terms in model 1

	edf	**Ref.df**	**Chi.sq**	**p-value**
s(day.year)	8.89	8.99	1663	<2e-16

R-sq.(adj) = 0.961 Deviance explained = 96.6%

REML score = 1110.1 Scale est. = 1 n = 309

Table 3. Parametric coefficients in model 2

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.38	7.34	-0.18	0.8512
factor (Month)2	0.014	0.13	0.10	0.9188
factor (Month)3	-0.25	0.21	-1.21	0.2248
factor (Month)10	0.28	0.53	0.54	0.5915
factor (Month)11	0.33	0.35	0.95	0.3413
factor (Month)12	0.31	0.17	1.80	0.0714
Thusday	1.30	0.10	12.18	<2e-16
Monday	1.72	0.11	16.26	<2e-16
Tuesday	1.48	0.11	13.98	<2e-16
Wednesday	1.43	0.12	13.36	<2e-16
Saturday	0.26	0.11	2.22	0.0263
Friday	1.21	0.10	11.19	<2e-16
FlueSeason2011-2012	-2.18	6.53	-0.33	0.7378
FlueSeason2012-2013	5.48	10.06e+01	0.51	0.6070
Population	3.88e-06	7.199e-07	5.39	6.93e-08

Table 4. Comparison of different models

Model	**R^2-adj**	**MSE**	**Relative MSE**
GLM, link Poisson	**0.99**	**423.37**	**1.53**

GLM, link NegBin	**0.99**	**581.89**	**2.12**
GAM, link Poisson	**0.96**	**10.18**	**0.25**
GAM, link NegBin	**0.94**	**25.54**	**0.33**
GAMM, link Poisson	**0.95**	**13.82**	**0.30**
GAMM, link NegBin	**0.71**	**870.13**	**2.64**

GLM, Generalized Lineal Model; GAM, Generalized Additive Model; GAMM, Generalized Additive Mixed Model; NegBin, Negative Binomial