National Institute of Economic and Social Research

TIME SERIES MODELS FOR EPIDEMICS: LEADING INDICATORS, CONTROL GROUPS AND POLICY ASSESSMENT

Andrew Harvey

Faculty of Economics, Cambridge University and NIESR

NIESR Discussion Paper No. 517
19 October 2020

About the National Institute of Economic and Social Research

The National Institute of Economic and Social Research is Britain's longest established independent research institute, founded in 1938. The vision of our founders was to carry out research to improve understanding of the economic and social forces that affect people's lives, and the ways in which policy can bring about change. Over eighty years later, this remains central to NIESR's ethos. We continue to apply our expertise in both quantitative and qualitative methods and our understanding of economic and social issues to current debates and to influence policy. The Institute is independent of all party political interests.

National Institute of Economic and Social Research
2 Dean Trench St
London SW1P 3HE
T: +44 (0)20 7222 7665
E: enquiries@niesr.ac.uk
www.niesr.ac.uk
Registered charity no. 306083

This paper was first published in October 2020

TIME SERIES MODELS FOR EPIDEMICS: LEADING INDICATORS, CONTROL GROUPS AND POLICY ASSESSMENT

Andrew Harvey

Abstract

This article shows how new time series models can used to track the progress of an epidemic, forecast key variables and evaluate the effects of policies. A class of univariate time series models was developed by Harvey and Kattuman (2020). Here the framework is extended to modelling the relationship between two or more series. The role of common trends is discussed, and it is shown that when there is balanced growth in the logarithms of the growth rates of the cumulated series, simple regression models can be used to forecast using leading indicators. Data on daily deaths from Covid-19 in Italy and the UK provides an example. When growth is not balanced, the model can be extended by including a stochastic trend: the viability of this model is investigated by examining the relationship between new cases and deaths in the Florida second wave of summer 2020. The balanced growth framework is then used as the basis for policy evaluation by showing how some variables can serve as control groups for a target variable. This approach is used to investigate the consequences of Sweden's soft lockdown coronavirus policy.

Keywords: Balanced growth; Co-integration; Covid-19; Gompertz curve; Kalman filter; Stochastic trend.

JEL Classifications: C22, C32

Contact details

Professor Andrew Harvey (ach34@cam.ac.uk), Corpus Christi College, Trumpington Street, Cambridge, CB1 2HD

1 Introduction

The aim of this article is to show how time series models can be used to track the progress of an epidemic, forecast key variables and evaluate the effects of policies. The new methods draw much of their inspiration from techniques in econometrics. However, the characteristics of time series for epidemics are different from those of most time series in economics and these differences need to be taken into account.

Harvey and Kattuman (2020a) - hereafter HK - developed a class of univariate time series models for predicting future values of a variable which when cumulated is subject to an unknown saturation level. In these models, the logarithm of the growth rate of the cumulated series depends on a time trend. Allowing this trend to be time-varying introduces flexibility which, in the context of an epidemic, enables the effects of changes in policy and population behaviour to be tracked. Nowcasts and forecasts of the variables of interest, such as the daily number of cases, its growth rate and the instantaneous reproduction number, R_t, can be made. Estimation of the models is by maximum likelihood and goodness of fit can be assessed by standard statistical test procedures.

Time series models can also be used to address other questions by exploring relationships between different series. One application concerns how the time path of an epidemic in a country which suffers an outbreak before another can be used as a leading indicator. The rationale for modelling the logarithm of the growth rate (of the cumulated series) comes from the properties of a Gompertz growth curve and when two such curves follow the same time path, but one lags the other, the trends in the series on the logarithms of the growth rate will be a constant distance apart. This suggests that when the trends are stochastic, the same will be true. This situation, known as balanced growth, arises in macroeconomics and is a special case of what econometricians call co-integration[1]; see, for example, Stock and Watson (1988). Balanced growth leads to a leading indicator regression model in which the logarithm of the growth rate in one series depends on lags in the logarithm of the growth rate of another series. The model is illustrated by showing how deaths in the UK in the first few months of the coronavirus epidemic can be predicted by deaths in Italy two weeks earlier.

[1] Maddala and Kim (1998) give a review of co-integration.

The requirement that two series exhibit balanced growth, while highly desirable, is not necessary for one to be a good leading indicator of the other. One way of dealing with this more general situation is by adding a stochastic trend to the regression model. The need for the additional flexibility is explored with data from the 'second wave' in Florida where it is shown how daily new cases of coronavirus can be used to predict deaths. The modelling framework is then extended to show how the data on new cases can be combined with daily deaths to better estimate the path of an epidemic and the associated values of R_t.

Time series modelling of an intervention can be used to assess the impact of a policy. This was done in HK in connection with the UK lockdown of March 2020. Here an attempt is made to answer the question 'What if lockdown had been imposed a week earlier?' The impact of lockdown is then explored further by developing the ideas on the logarithms of growth rates following a common trend to try to estimate the number of coronavirus deaths in Sweden had a more stringent lockdown been imposed. The methodology draws on the study of control groups in time series by Harvey and Thiele (2020). It is argued that the fact that death rates in Sweden were roughly ten times those in neighbouring countries could be misleading; the growth paths of the UK and Italy provide more relevant information.

2 Growth curves and time series models

This section sets out the basic model in which the logarithm of the growth rate of the cumulated series consists of a stochastic trend plus an irregular term. It is then shown how the framework may be extended to model the relationship between two series.

2.1 Dynamic trend models

The observational model uses data on the time series of the cumulated total of confirmed cases or deaths, Y_t, and the daily change, $y_t = \Delta Y_t = Y_t - Y_{t-1}$. HK show how the theory of generalized logistic growth curves suggests models for $\ln y_t$ and $\ln g_t$, where $g_t = y_t/Y_{t-1}$ or $\Delta \ln Y_t$. For the special case of the Gompertz growth curve, the models simplify to

$$\ln y_t = \ln Y_{t-1} + \delta - \gamma t + \varepsilon_t, \quad \gamma > 0, \quad t = 2, ..., T, \tag{1}$$

and

$$\ln g_t = \delta - \gamma t + \varepsilon_t, \quad t = 2, ..., T, \tag{2}$$

where ε_t is a random disturbance term.

A stochastic, or time-varying, trend may be introduced into (2), to give the dynamic trend model

$$\ln g_t = \delta_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \qquad t = 2, ..., T, \tag{3}$$

where

$$\begin{array}{rcll} \delta_t & = & \delta_{t-1} - \gamma_{t-1} + \eta_t, & \eta_t \sim NID(0, \sigma_\eta^2), \\ \gamma_t & = & \gamma_{t-1} + \zeta_t, & \zeta_t \sim NID(0, \sigma_\zeta^2), \end{array} \tag{4}$$

and the normally distributed irregular, level and slope disturbances, ε_t, η_t and ζ_t, respectively, are mutually independent. When σ_ζ^2 is positive, but $\sigma_\eta^2 = 0$ the trend is an integrated random walk (IRW). HK found IRW trend to be particularly useful for tracking an epidemic and it will be adopted in the applications here. The speed with which a trend adapts to a change depends on the signal-noise ratio, which for the IRW is $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$; the trend is deterministic when $q = 0$.

Allowing γ_t to change over time means that the progress of the epidemic is no longer tied to the proportion of the population infected as it would be if Y_t followed a deterministic growth curve. Instead the model adapts to movements brought about by changes in behaviour and policies. If γ_t falls to zero, the growth in Y_t becomes exponential while a positive γ_t means that the growth rate is increasing.

Additional components, such as day of the week effects, can be added to (3). These may be deterministic or stochastic. Explanatory variables, including interventions, can also be included. Such models can be estimated using techniques based on state space models and the Kalman filter; see Durbin and Koopman (2012) or Harvey (1989). Here the STAMP package of Koopman et al. (2020) is used.

Remark 1 *When the observations are small, a negative binomial distribution for y_t may be appropriate. HK show how the model may be modified to deal with this possibility for a univariate time series. However, the numbers in the applications here are big enough to allow y_t to be treated as lognormal and hence for the distribution of* $\ln g_t$, *conditional on δ_t, to be considered normal.*

2.2 Forecasts

Recursions for making forecasts of future observations in the dynamic Gompertz model are

$$\widehat{g}_{T+\ell|T} = \exp \delta_{T+\ell|T}, \qquad \ell = 1, 2, .. \tag{5}$$

$$\widehat{\mu}_{T+\ell|T} = \widehat{\mu}_{T+\ell-1|T}(1 + \widehat{g}_{T+\ell|T}) \tag{6}$$

so that $\widehat{y}_{T+\ell|T} = \widehat{g}_{T+\ell|T}\widehat{\mu}_{T+\ell-1|T}$ and $\widehat{Y}_{T+\ell|T} = \widehat{\mu}_{T+\ell|T}$; the initial value is $\widehat{\mu}_{T|T} = Y_T$. The prediction $\delta_{T+\ell|T}$ is simply $\delta_{T|T} - \gamma_{T|T}\ell$. Combining (5) and (6) gives

$$\widehat{y}_{T+\ell|T} = Y_T \exp \delta_{T+\ell|T} \prod_{j=1}^{\ell-1}(1 + \exp \delta_{T+j|T}), \quad \ell = 2, .. \tag{7}$$

and $\widehat{y}_{T+1|T} = Y_T \exp \delta_{T+1|T}$.

The basic forecasts are made with the estimates of δ_t and γ_t at the end of the sample. However, alternative scenarios in which γ_t is assumed to evolve in a certain way, perhaps to reflect changing policies, such as the easing of lockdown restrictions, may also be envisaged. Adapting the forecasts to account for such movements is straightforward.

2.3 Comparing different growth curves

The Gompertz growth curve lies behind the notion of setting up time series models in which the logarithm of the growth rate of the cumulative total of a variable follows a trend. It is therefore able to provide insight on how to formulate and interpret models linking several series.

The Gompertz growth curve is

$$\mu(t) = \overline{\mu} \exp(-\alpha e^{-\gamma t}), \qquad \alpha, \gamma > 0, \qquad -\infty < t < \infty, \tag{8}$$

where γ is a growth rate parameter, $\overline{\mu}$ is the upper bound or saturation level and α reflects initial conditions. The associated incidence curve is

$$d\mu(t)/dt = \mu'(t) = \gamma\alpha\mu(t) \exp(-\gamma t),$$

with a peak at $t = \gamma^{-1} \ln \alpha$. Figure 1 shows an incidence curve with a peak at $t = 19.97$, together with the same curve shifted to the right so the peak

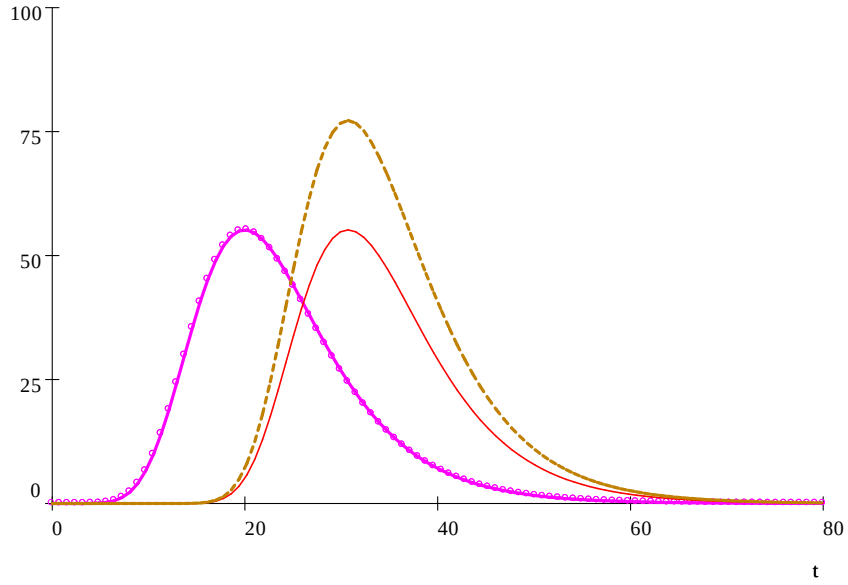

Figure 1: Gompertz incidence curves with $\gamma = 0.15$, $\alpha_1 = 20$ for the left hand curve and $\alpha_2 = 100$ for the right hand curves.

is at 30.71. A curve above the right hand curve is also shown; this is higher because the value of $\overline{\mu}$ is 1400 rather than 1000 as it is for the other two curves. In all cases $\gamma = 0.15$, but for the left hand curve α is 20 whereas for the right hand curves it is 100.

Although the right hand curves in Figure 1 clearly lag the left hand one, it is not immediately evident how to model the relationship. However, the logarithms of the growth rates of $\mu(t)$ are

$$\ln g(t) = \delta - \gamma t, \qquad t \geq 0, \tag{9}$$

where $\delta = \ln \alpha\gamma$; compare (2). Figure 2 shows the two lines for $\ln g(t)$ running in parallel. The distance between them depends on the intercepts, δ, which in turn depend on the initialization parameter, α. The height of the incidence curve, which depends on the saturation level, $\overline{\mu}$, is irrelevant; as a result the lines corresponding to the two right hand incidence curves in Figure 1 are identical. This is important because it means that small populations can be compared with big ones: size does not matter.

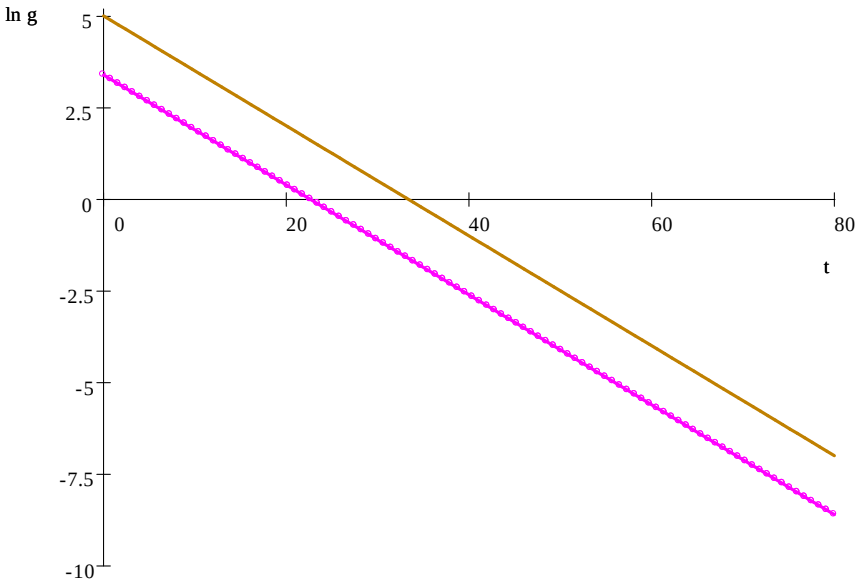

Figure 2: Logarithms of the growth rates for incidence curves in Figure 1; $\gamma = 0.15$, $\alpha_1 = 20$ and $\alpha_2 = 100$ (upper line).

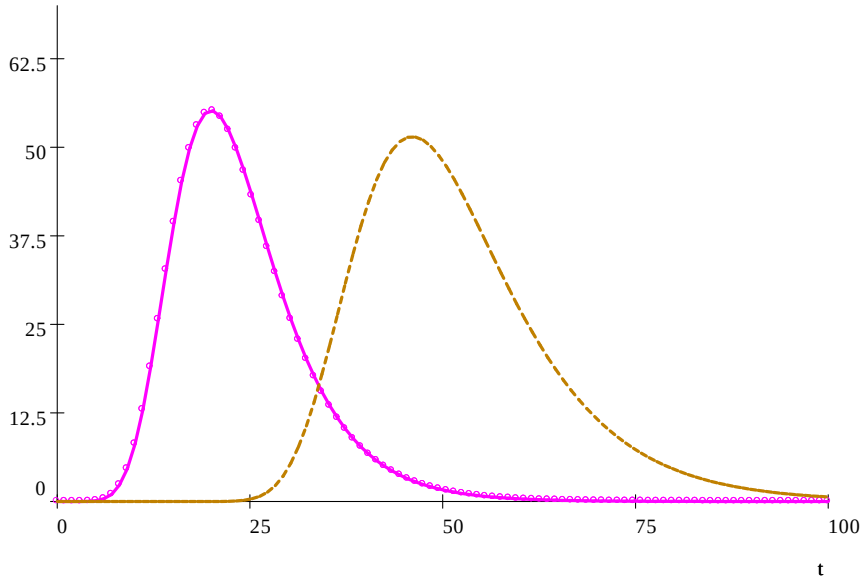

Figure 3: Gompertz incidence curves with $\alpha_1 = 20$ and $\gamma_1 = 0.15$ for the left hand curve and $\alpha_2 = 100$ and $\gamma_2 = 0.1$ for the right hand curve.

When two lines are parallel, the upper line lags the lower one by

$$k = \frac{\delta_2 - \delta_1}{\gamma} = \frac{\ln \alpha_2 - \ln \alpha_1}{\gamma}, \tag{10}$$

where δ_1 and δ_2 are the intercepts of the lower and upper lines respectively and α_1 and α_2 are the corresponding initial conditions. In Figure 2 the lag is $k = 10.73$.

When the $\gamma's$ are different, the epidemic progresses at different speeds, as can be seen in Figure 3. The lines for $\ln g(t)$ are no longer parallel and subtracting the lower one (for variable 1) from the upper one gives $\ln(\alpha_2\gamma_2/\alpha_1\gamma_1) - (\gamma_2 - \gamma_1)t$. First multiplying the lower line by $\beta = \gamma_2/\gamma_1$ removes the time trend. However, when $\beta \neq 1$, the time lag is no longer constant.

3 A model for leading indicators

Now consider observational models of the form (2) for two time series which are on the same growth path because $\gamma_1 = \gamma_2$ but the first series leads the second by k time periods. The observations run from $t = 1$ to T but when the first series is lagged by k time periods, $\ln g_{1,t-k}$ runs from $t = k + 1$ to $T + k$. Subtracting the first series from the second gives

$$\ln g_{2t} = \delta + \ln g_{1,t-k} + \varepsilon_t, \quad t = k+1, ..., T+k \tag{11}$$

where $\delta = \ln(\alpha_2/\alpha_1)$ and the disturbance term is $\varepsilon_t = \varepsilon_{2t} - \varepsilon_{1,t-k}$. The model takes the same form when the trends are stochastic, so long as there is balanced growth.

Allowing for a lag structure in the leading series gives

$$\ln g_{2t} = \delta + \sum_{j=h}^{k} \beta_j \ln g_{1,t-j} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \quad t = k+1, .., T+h, \tag{12}$$

where $h < k$ and $\sum \beta_j = 1$. Daily effects are a complicating factor, but can be introduced into the model as additional terms on the right hand side. The summation restriction is imposed by restricted least squares (RLS) which can be performed very easily by reformulating (12) as

$$\ln g_{2t} - \ln g_{1,t-k} = \delta + \sum_{j=h}^{k-1} \beta_j (\ln g_{1,t-j} - \ln g_{1,t-k}) + \varepsilon_t, \quad t = k+1, .., T+h. \tag{13}$$

All that is required is an OLS regression of $\ln g_{2t} - \ln g_{1,t-k}$ on $\ln g_{1,t-j} - \ln g_{1,t-k}$, $j = h, .., k-1$. The coefficient of $\ln g_{1,t-k}$ is equal to $1 - \sum_{j=h}^{k-1} \beta_j$ and the lag structure should be such that δ is close to zero. Balanced growth continues to hold when ε_t is replaced by any stationary process.

When the two series are not on the same growth path, the model can be extended by replacing the constant term by a stochastic trend. Thus

$$\ln g_{2t} - \ln g_{1,t-k} = \delta_t + \sum_{j=h}^{k-1} \beta_j (\ln g_{1,t-j} - \ln g_{1,t-k}) + \varepsilon_t, \quad t = k+1, .., T+h, \tag{14}$$

where δ_t is as in (4) and estimation is by the Kalman filter. The interpretation is that the growth path of the target series that derives from balanced growth

with the leading indicator is augmented by a stochastic growth component that might reflect the way the first series and/or the second is measured. A theoretical case for δ_t being a random walk, rather than an IRW, is outlined in sub-sevtion 3.4.

When (13) has been estimated, the residuals may be tested for serial correlation. The implication of the dynamic specification of δ_t in (14) is of a nonstationary alternative. Hence the stationarity test of Kwiatkowski et al (1992) - the KPSS test - can be used; see Harvey and Thiele (2020) for a discussion of these issues in the context of balanced growth.

3.1 Predictions

The predictions for the logarithms of the growth rate in model (14) are

$$\ln \widehat{g}_{2,T+\ell|T} = \delta_{T+j|T} + \sum_{j=h}^{k} \widehat{\beta}_j \ln g_{1,T+\ell-j}, \quad \ell = 1, ..., h, \tag{15}$$

with $\widehat{\beta}_k = 1 - \sum_{j \neq k} \widehat{\beta}_j$ and $\delta_{T+\ell|T} = \delta_{T|T} + \gamma_{T|T}\ell$. When $\delta_t = \delta$, as in (13), $\delta_{T+\ell|T} = \widehat{\delta}$. The predictions for $\widehat{g}_{2,T+\ell|T}$ can be converted into predictions for $y_{2,T+\ell}$ by inserting into (6). Solving the recursions with $\widehat{\mu}_{T|T} = Y_T$ as in (7) yields

$$\widehat{y}_{2,T+\ell|T} = Y_{2,T}\widehat{g}_{2,T+\ell|T} \prod_{j=1}^{\ell-1}(1 + \widehat{g}_{2,T+j|T}), \quad \ell = 2, 3, .. \tag{16}$$

and $\widehat{y}_{2,T+1|T} = Y_{2,T}\widehat{g}_{2,T+1|T}$. The construction of prediction intervals remains a topic for future research.

3.2 Italy and the UK

Figure 4 shows the daily deaths in Italy and the UK from March 1st to June 20th, 2020; data sources are given in the Appendix. Italy clearly leads the UK but the relationship is captured more precisely in Figure 5 which shows the logarithms of the growth rates (LDL) of total deaths.

Subtracting LDL Italy from the LDL UK for data from 16th March and estimating the mean, with daily[2] dummy variables included, gives 0.686. A

[2]The data is for when the deaths are recorded rather than when they occur. Series based on date of death would not have the daily pattern but are difficult to obtain.

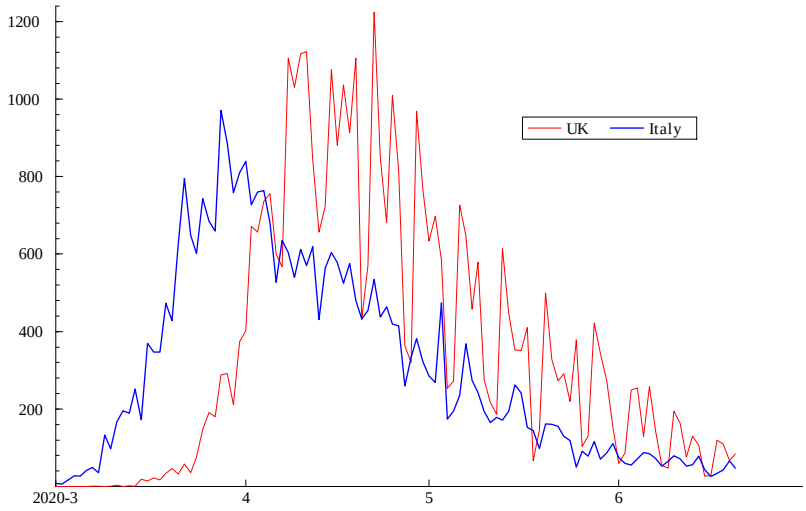

Figure 4: Daily deaths in Italy and UK

rough estimate of the slope, γ, obtained by fitting a time trend to LDL Italy is 0.05, so equation (10) suggests a lag close to 14 days. A lag of 14 is not inconsistent with prior information and it has the attraction of lining up the days of the week in the two countries. Figure 6 shows the LDL series with Italy lagged by 14 days together with the contrast between the two countries obtained by subtracting Italy from the UK.

A regression model was estimated with daily dummy variables included and lags with the constraint that the coefficients sum to one imposed as in (13). A model with lags of 14 and 13 emerged as the best fit. The results were: $\widetilde{\beta}_{13} = 0.567$ (0.099), implying $\widetilde{\beta}_{14} = 0.433$, together with $\widetilde{\delta} = -0.158$ and the following diagnostics[3]: $DW = 2.00$, $Q(14) = 11.89$, $BS = 7.67$ and $H = 2.00$. When lags at 12 and 15 were included they were small and statistically insignificant.

[3]DW is Durbin-Watson, Q(P) is Box-Ljung with P autocorrelations, BS is the Bowman-Shenton normality statistic and H is a heteroscedasticity statistic constructed as the ratio of the sum of squares in the last third of the sample to the sum of squares in the first third. Numbers in parenthesis after estimates are standard errors.

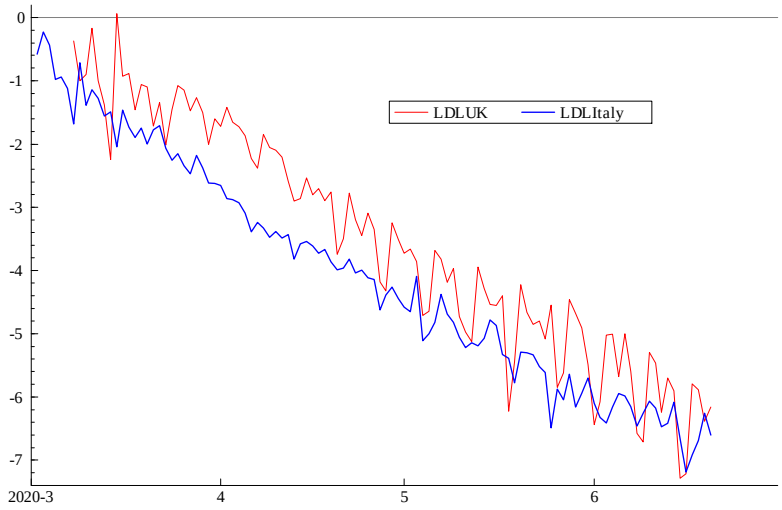

Figure 5: Logarithms of the growth rates (LDL) of total deaths in UK and Italy

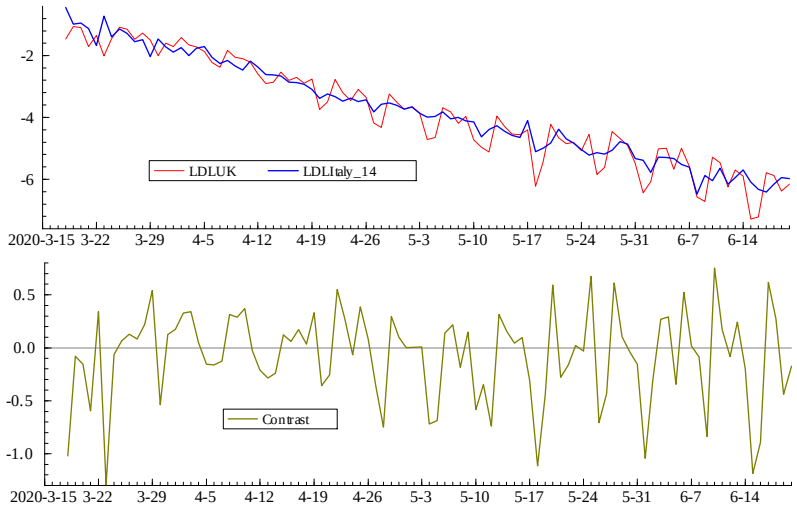

Figure 6: LDL series with Italy lagged by 14 days together with the contrast LDLUK-LDLItaly

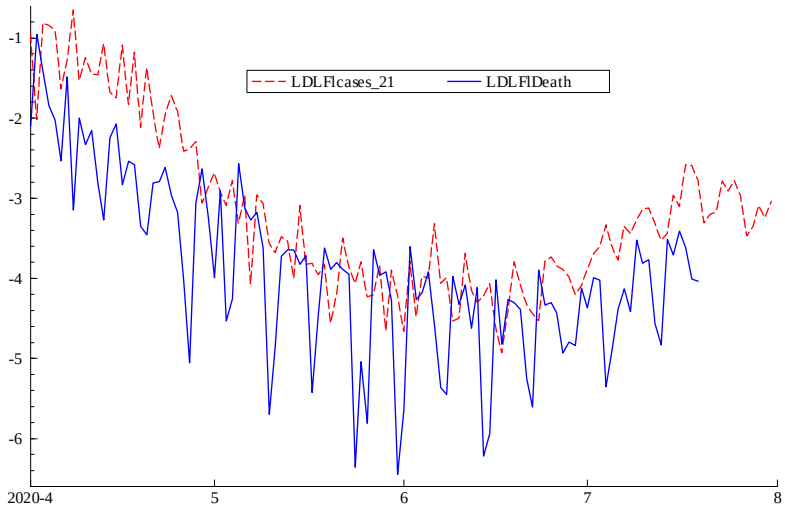

Figure 7: Florida: LDLDeath and new cases lagged 21 days.

3.3 Deaths and New Cases in Florida

Daily cases of Covid-19 in the US state of Florida peaked in early April. There was then a decline following a lockdown during April. After April restrictions were eased and there was a leveling out in May, followed by a sharp rise in June. This second wave poses a challenge for a model in which new cases are used as a leading indicator for deaths.

Aside from the model having to deal with a situation where new cases and deaths have fallen before rising again, there is the problem that the basis on which new cases are recorded changes over time. At the beginning of the pandemic, new cases in many countries were primarily hospital admissions, but over time testing became more widespread. A balanced growth model assumes that the growth rate in deaths is the same as the growth rate in new cases. When this does not hold the inclusion of a stochastic trend in the model offers a way of dealing with the discrepancy. In the case of Florida there was an increase in testing in May, although the growth rate in tests was roughly constant from the end of May onwards. This suggests that the growth rate of confirmed new cases may still be a good indicator of the path of new infections.

Figure 7 shows the logarithms of the growth rates of total new cases and

deaths, with the former lagged by 21 days. The observations are from March 29th to July 19th 2020 inclusive. The choice of 21 days is convenient because of the strong day of the week effect.

After some experimentation with different lags it was found that a lag centred on 18 days is better than one centred on 21. A balanced growth regression model with a constant and lags on either side of 18 gave weights of 0.271 (0.123) for 17 and 0.249 (0.115) for 19, leaving 0.480 for 18. The fit was good but a high Q-statistic - Q(15) was 40.89 - indicates residual serial correlation. The sample autocorrelations are very persistent and slow to die out; see Figure 8. Including a deterministic time trend in the model makes very little difference as the coefficient is very small and statistically insignificant with a p-value of 0.82. By contrast, a stochastic slope, with an estimated q_ζ of 0.00011, removes the serial correlation as Q(16)=7.73. However, there is only a small improvement in the fit as measured by prediction error variance, which is 0.247 with the stochastic slope and 0.259 without. A random walk trend - the stochastic level model - is better in that it reduces the prediction error variance to 0.237. The signal noise ratio, $q_\eta = \sigma_\eta^2/\sigma_\varepsilon^2$, is 0.0041 and the lag coefficients at 17 and 19 are 0.265 (0.122) for 17 and 0.256 (0.113) for 19. Estimating an unrestricted stochastic trend confirms the choice of the random walk because the slope variance is zero and the estimated (constant) slope is small and statistically insignificant.

Figure 9 compares the leading indicator forecasts of the logarithm of the growth rate of deaths in the regression model (without the slope), together with the forecasts made from a univariate model and the actual observations up to, and including, 12th August. Corresponding predictions of daily deaths can be made with (16). The univariate forecasts overshoot because the estimate of γ_T is positive. It needs to be negative for the forecasts to eventually start moving down; see the examples and discussion in HK. The forecasts obtained with the stochastic slope leading indicator model are very similar to those with no slope, as are those with the stochastic level model. The fact that a simple regression model works so well here is rather surprising, but this may not always be the case and the stochastic level model is likely to be the default option.

There is clearly scope for experimenting with more models within the same framework, perhaps with data from other regions. For example, notwithstanding the earlier comments about the constancy of the lag(s) being dependent on a balanced growth model, it may be worth investigating the forecasting performance of an unrestricted regression model. On the data

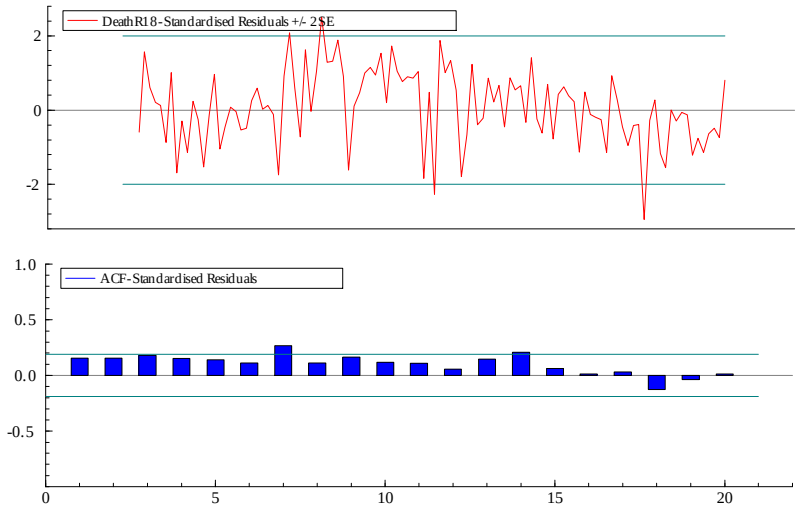

Figure 8: Residuals and associated correlogram from balanced growth regression model fitted to new case and deaths in Florida

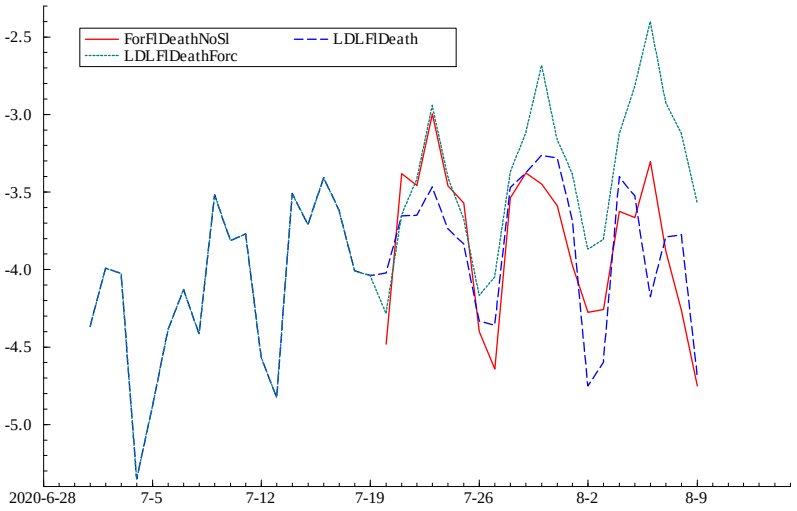

Figure 9: Logarithms of growth rate of deaths in Florida with leading indicator forecasts without the slope and univariate forecasts

side, there is scope for bringing other variables, such as the total tested, into the model.

3.4 Nowcasting and forecasting of the growth path and R

So far the leading indicator models have been used to forecast future observations of a target series, but often what is required is a nowcast or forecast of $g_{y,t}$, its growth rate. This is important for tracking the direction in which an epidemic is moving and, if required, it can be translated into an estimate of the instantaneous reproduction number R_t; see Wallinga and Lipsitch (2008). In a univariate model, Harvey and Kattuman (2020b) use filtered estimates of $g_{y,t} = g_t - \gamma_t$ to track the progress of an epidemic. A corresponding estimator of R_t can be constructed in a number of ways, but the most practical are

$$\widetilde{R}_{t|t,\tau} = 1 + \tau \widehat{g}_{y,t|t} \quad \text{and} \quad \widetilde{R}^{e}_{t|t,\tau} = \exp(\tau \widehat{g}_{y,t|t}), \tag{17}$$

where τ is the generation interval, which is the number of days that must elapse before an infected person can transmit the disease.

Here the aim is to find a way of estimating $g_{y,t}$ for the period $t = T - k, ..., T$ by combining the information in the target series with that in the leading indicator. More specifically the trend in the target series, deaths, gives a coherent measure of the daily growth rate, and hence R_t, but one subject to a delay. The information in confirmed cases is more up to date but it may not yield a consistent time series. The hope is that by combining the two series a better measure of the current growth rate can be extracted.

The model is

$$\begin{aligned} \ln g_{1t} &= \delta_{2t} + \delta_t + \varepsilon_{1t}, \quad t = 1, ..., T, \\ \ln g^*_{2t} &= \delta_{2t} + \varepsilon_{2t}, \qquad\quad t = 1, ..., T, \end{aligned} \tag{18}$$

where $\ln g^*_{2,t} = \ln g_{2,t+k}$, $t = 1, ..., T$. Thus the last k observations on $\ln g^*_{2,t}$ are missing. The stochastic trend, δ_{2t}, models the underlying movements in $g_{y,t}$ by an IRW, as in Harvey and Kattuman (2020b), whereas δ_t is a stochastic component that captures the deviations of the first series from balanced growth. All disturbances, including ε_{1t} and ε_{2t}, are Gaussian and assumed to be mutually as well as serially independent.

A convenient model for δ_t is the first-order autoregression, $\delta_t = \phi\delta_{t-1} + \zeta_t$, where ζ_t is $NID(0, \sigma^2_\zeta)$. When $|\phi| < 1$, the series are co-integrated with

balanced growth as $\ln g_{2t}^* - \ln g_{1t} = -\delta_t + \varepsilon_{2t} - \varepsilon_{1t}$; note the similarity (apart from the sign of δ_t) to (14). However, the RW, when $\phi = 1$, will be the most likely option. Estimation of (18) is by state space methods as in Harvey and Chung (2000). The Kalman filter and smoother provides smoothed estimates of δ_{2t} and γ_{2t} with the (filtered) estimates at $t = T$ giving the nowcast, $\widehat{g}_{y,T|T}$. As new observations become available the nowcast may be updated by the Kalman filter. The hope is that the IRW specification for δ_{2t} enables it to separated from the movements in δ_t when the latter is a random walk.

Harvey and Kattuman (2020b) show that CIs for g_{yt} and R_t may be constructed for a univariate model when the growth rate of the total, g_t, is relatively small and the same is true here. Unfortunately g_t cannot be ignored when the epidemic is growing rapidly as might be the case in a second wave.

4 The effects of policy interventions

This section shows how the time series models can be used to assess the effects of policy. The first example uses univariate time series modelling, while the second builds on the analysis of bivariate series in sub-section 2.4.

4.1 What if lockdown in the UK had been a week earlier?

The UK went into full lockdown on March 23rd. Can we estimate how many deaths could have been saved if it had been a week earlier?

A slope intervention in (2) enables the effect of a policy which changes γ to be evaluated. Thus

$$\ln g_t = \ln Y_{t-1} + \delta - \gamma t - \beta t w_t + \varepsilon_t, \quad t = 1, ..., T, \tag{19}$$

where w_t are intervention dummies. When the full effect is realized, the slope on the time trend will have moved from γ to $\gamma + \beta$. A positive β lowers the growth rate, g_t, the peak of the incidence curve and the final level. The intervention dummies can be constructed from a logistic CDF. This yields a response curve $W(t) = 1/(1 + \gamma_0^I \exp(-\gamma^I(t - t^I))$, where t^I is the median. With t^L and t^U denoting the beginning and the end of the time span during which the response to the intervention occurs, $w_t = 0$ for $t < t^L$, $w_t = W(t)$ for $t = t^L$, $t^L + 1, .., t^I, .., t^U$ and $w_t = 1$ for $t = t^U + 1, .., T$.

HK fitted the static model in (19) to new cases in the UK, with an intervention starting on March 26th and ending on April 12th, using data from the beginning of March up to April 29th. The result was an estimate of β equal to 0.020 (0.004) and an estimate of γ also equal to 0.020. The overall effect[4] is a new slope of 0.041. The trend, with the intervention included, is shown by the dashed line in Figure 10.

The effect of implementing lockdown restrictions a week earlier can be estimated by shifting the intervention response forward by one week so it starts on March 19th, rather than on March 26th. The adjusted trend in the logarithm of the growth rate is then

$$\ln g_t^* = \delta - \gamma t - \beta t w_{t+7}, \quad t = 1, ..., T. \tag{20}$$

Once the effect of the intervention has worked itself through, the new slope is the same as before, as can be seen in the solid line in Figure 10.

The predicted final total is

$$\overline{\mu} \simeq \mu_T \exp(\exp \delta_{T|T}/(\exp \gamma_{T|T} - 1))$$

where T is April 12th. For the actual data, μ_T can be approximated by Y_T. For the early lockdown scenario, μ_T will be smaller because the growth rate falls earlier. This implies that the level on March 18th is multiplied by $\exp(\sum g_t^*)$, where the summation is over the period from March 19th to April 12th. To ensure comparability, the actual level on April 12th is best estimated in the same way, rather than by Y_T. Thus an estimate of the ratio of the total number of cases for a hypothetical early lockdown to the actual total is given by

$$\frac{Hypothetical}{Actual} = \frac{\exp(\sum g_j^*)}{\exp(\sum g_j)} = \frac{\exp(\sum \exp(\delta - \gamma t - \beta t w_{t+7}))}{\exp(\sum \exp(\delta - \gamma t - \beta t w_t))}.$$

This ratio is 0.551 implying that the number of infections, as measured by data on daily coronavirus hospital admissions, could have been almost halved by an earlier lockdown. If a constant proportion of those admitted die, the implication is that deaths could have been almost halved by an earlier

[4]When the slope was allowed to be stochastic, the estimate of β was reduced to 0.014 (0.006), but with such a small sample size, a stochastic slope risks some confounding with the intervention variable.

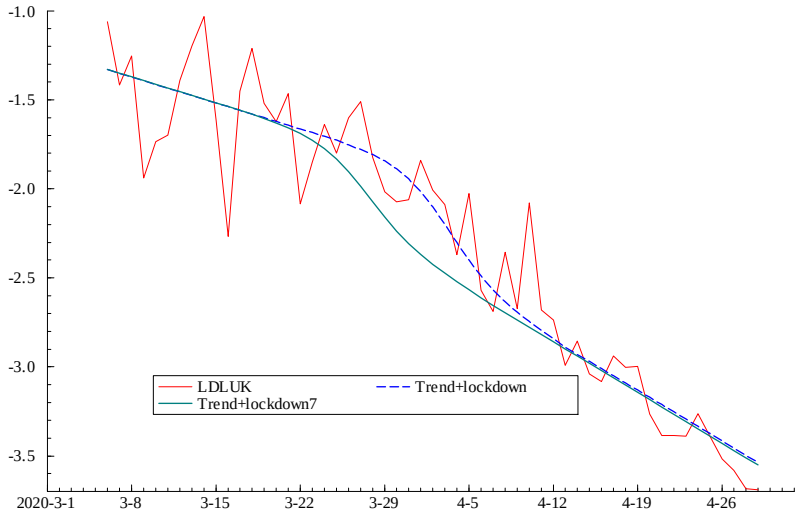

Figure 10: Estimates of logarithm of growth rate of total cases in UK with a logistic intervention and a daily effect

lockdown[5]. This conclusion is not too different from ones obtained by other methods. For example, the BBC reported on 10th June that Professor Neil Ferguson of Imperial College told a committee of MPs: 'Had we introduced lockdown measures a week earlier, we would have reduced the final death toll by at least a half.'

4.2 Fewer deaths in Sweden with a full lockdown ?

Sweden did not opt for the full lockdown that other European countries imposed in March. Restrictions were minimal: the government recommended frequent handwashing, working from home, self-isolation for those who felt ill or were over 70 and social distancing[6]; see, for example, Kamerlin and Kasson (2020). Did this policy lead to the number of deaths being significantly higher

[5]It should be stressed that these findings relate specifically to the effect of the full lockdown of March 2020. A full lockdown imposed now is unlikely to have the same impact because the environment is different in that socal distancing restrictions are in place, behaviour has changed and the risk to care homes is better understood,

[6]Carl Bildt, a former prime minister, was quoted as saying "Swedes, especially of the older generation, have a genetic disposition to social distancing anyway."

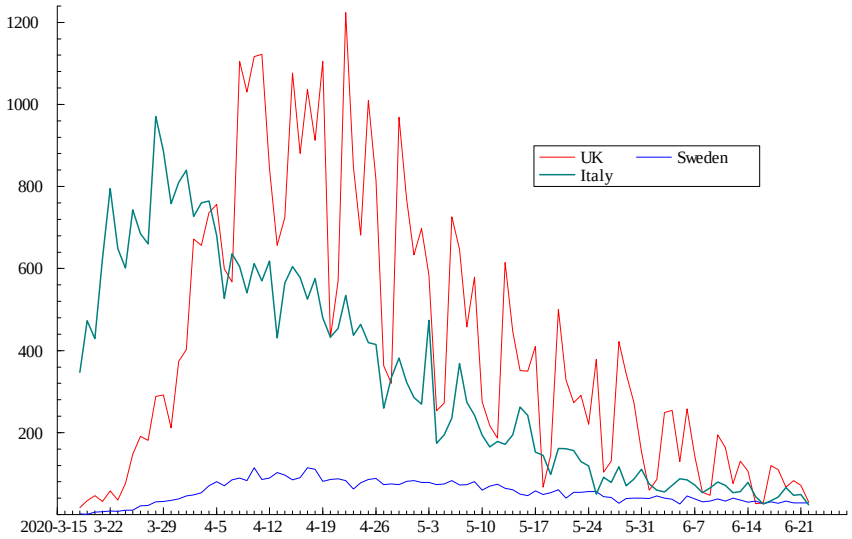

Figure 11: Daily deaths in Sweden (lower line), UK (highest line) and Italy from 18th March to 22nd July

than it might have been under a full lockdown? To answer this question we need to determine the growth path that Sweden would most likely have followed under a hard lockdown.

Figure 11 shows daily deaths in Sweden, UK and Italy (lagged 14 days) from 18th March to 22nd July; by the end of July numbers had become small. A comparison of actual and potential growth paths has to be based on the logarithms of growth rates of the cumulative total for the reasons discussed earlier. Because the day of the week effect is very strong, particularly in the UK, the logarithms of growth rates were smoothed with a seven day moving average, centred on the fourth day. The graph in Figure 12 shows that Sweden initially fell at the same rate as the UK and Italy but then started to diverge around 24th April, about a month after the UK lockdown began on March 23rd.

If Sweden had kept on the same growth path as the UK and Italy there would have been fewer deaths. An estimate of the number of deaths under this alternative scenario is given by reference to the forecasting equations, (5) and (6). Let $t = m$ denote the date of divergence and let $\widehat{\delta}_t$ denote the

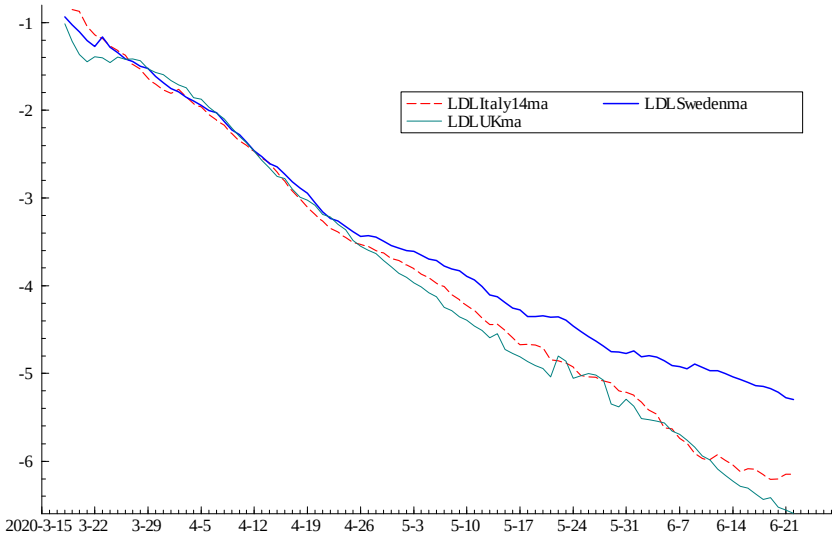

Figure 12: Seven day moving averages of the logarithms of the growth rate from March 18th to July 22nd

values of δ_t estimated for the lockdown growth path using the data on UK and Italy. Since the moving averages are quite smooth, $\widehat{\delta}_t$ was constructed as a simple average of the two countries[7], rather than by RLS as in Harvey and Thiele (2020). Then

$$\widehat{\mu}_{m+j} = \widehat{\mu}_{m+j-1}(1+\widehat{g}_{m+j}) \simeq \widehat{\mu}_{m+j-1}\exp\widehat{\delta}_{m+j}, \qquad j = 1, 2, .., T-m. \tag{21}$$

The initial value is $\widehat{\mu}_m = Y_m$, or a weighted average around that point. Solving the recursion gives

$$\widehat{Y}_T = \widehat{\mu}_T = Y_m \prod_{j=1}^{T-m}(1+\widehat{g}_{m+j}) \simeq Y_m \exp\sum_{j=1}^{T-m}\widehat{\delta}_{m+j} \tag{22}$$

as the estimated total number of deaths, up to time T, under the lockdown scenario. The estimated number of deaths after time m is $\widehat{Y}_T - Y_m$ while the

[7]The general methodology, as set out in Harvey and Thiele (2020), is to select a set of controls from a donor pool by using the KPSS test to determine which series are on a balanced growth path with the target. The control group weighting is then determined by RLS.

actual is $Y_T - Y_m$. Here T is July 22nd; the number of deaths after that is relatively small.

The total on April 24th was 2236 and using formula (22) gives an estimate of 4062 for July 22nd as opposed to an actual figure of 5722, a difference of 1660. The sensitivity to the initial value can be gauged by noting that the estimates using the totals two days before and two days after April 24th are 3808 and 4378 respectively.

One way of reducing the dependence on the starting value is to estimate the underlying total for Sweden using formula (22) with $\widehat{g}_{m+j}$ replaced by the actual Swedish values. This gave a total of 5657. The ratio of $\widehat{Y}_T$ for the lockdown control group to that of Sweden is $1.816/2.530 = 0.718$. For $\widehat{Y}_T - Y_m$ it is $0.816/1.530 = 0.533$. This implies that the actual increase from April 24th, which was 3486, could have been 1902. The first method gave $4062 - 2236 = 1826$. The overall conclusion is that, between April 24th and July 22nd, there were perhaps forty to forty-five per cent more deaths than there might have been under a more stringent lockdown of the kind implemented in the UK and Italy.

It is worth noting that although Sweden may have had more deaths under its soft lockdown, this does not mean a higher death rate than countries which had a hard lockdown. On Sept 4th, the figures for deaths per one million for Sweden were 577 as against 611 for the UK and 587 for Italy. The rates for Denmark, Norway and Finland were 108, 49 and 61 respectively, but this should not lead one to infer that the soft Swedish lockdown resulted in a death rate of perhaps ten times what it might have been.

The number of deaths in Denmark is too small to allow a full analysis based on the logarithms of growth rates. The variability is high and after mid-May there are often days when no deaths occur. Numbers in Norway and Finland are lower still. However, up to the end of April the logarithm of the growth rate for Denmark is informative. Figure 13 shows the logarithms of the growth rates for Sweden, Italy, UK and Denmark. Denmark is on a similar growth path to that of the other countries but it is lower than the UK because coronavirus may have arrived earlier and lockdown was imposed on March 13th; the gap is consistent with Denmark leading the UK by about a week. During this period deaths in Denmark were much lower than in Sweden even though they were on the same growth path until close to the end of April. This difference therefore seems to be for reasons not directly connected to the policies of the two countries on lockdown.

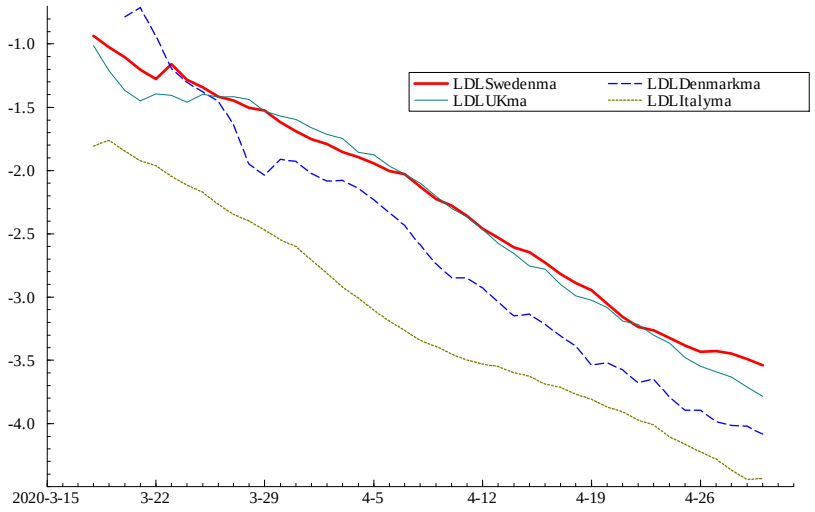

Figure 13: Seven day moving averages of the logarithms of the growth rate from March 18th to April 30th.

On April 30th 2714 deaths had been recorded in Sweden as against 443 in Denmark, a ratio of 6.13. On April 24th the figures were 2236 and 394, a ratio of 5.68. (But bear in mind that the population of Sweden is 1.76 times that of Denmark so in per capita terms the ratio is closer to three.) On July 22nd the ratio of Swedish to Danish deaths had risen to 9.36. However, the ratio of the lockdown estimate of 4062 to the 611 Danish deaths is only 6.64 which is not far from the ratio at the end of April. Thus the estimate of the number of deaths obtained using the control group seems quite plausible. The conclusion is that for reasons unconnected with lockdown policy the death rate per head in Sweden was about three and a half times that in Denmark. The less stringent lockdown then raised this ratio to nearly five and a half.

5 Conclusion

The main aim of this article has been to provide a methodological framework for the statistical analysis of the relationship between time series of the kind that are relevant for tracking and forecasting epidemics and analysing the

effects of policy.

The growth path of an epidemic is best captured by the logarithm of the growth rate of the cumulated series. This may be modelled by a stochastic trend. When two series are on a balanced growth path, leading indicator regression models estimated by restricted least squares can be used to forecast. The relationship between deaths from coronavirus in the UK and Italy provides a good example of balanced growth with deaths in Italy being able to provide forecasts for deaths in the UK up to thirteen days ahead. The balanced growth model may be extended by including a stochastic trend component. The stochastic trend, best specified as a random walk, removes the residual serial correlation found in the regression model linking deaths to new cases in Florida. However, for both models, the forecasts made before the downturn in the series are remarkably successful in picking up the subsequent downward movement. From the practical point of view, such models may be useful for forecasting hospital admissions[8] as well as deaths.

Leading indicators can also be used to improve estimates of the daily growth rate of an epidemic and the associated R_t. A bivariate state space model is proposed with a study of its effectiveness being a topic for future research.

Policy evaluation can be carried out by using some series as control groups for others. A common trend or, better still, balanced growth is the key ingredient. The evaluation of the Swedish policy response to coronavirus provides an example of the methodology. It is shown that the average of the growth paths of deaths in the UK and Italy yields a suitable control group for deaths in Sweden. The Swedish growth path is initially the same as those of the UK and Italy but it diverges as the effectf of the lockdowns in the UK and Italy start to impact daily deaths. The difference in the growth paths then enables the implications of the Swedish soft lockdown policy to be assessed. The analysis suggests[9] an increase in the number of deaths of the order of forty to forty-five per cent compared with what there might have been under a more stringent lockdown of the kind implemented in the UK and Italy.

[8] In the UK, and many other countries, the data on new cases in the early months of the epidemic was only for the most severe cases, primarily those admitted to hospital. It is only with more widespread testing that new cases start to have potential for predicting hospital admissions.

[9] A degree of caution is needed because of revised figures and different definitions of what constitutes a Covid-19 death.

References

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Harvey A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A.C. and C.H. Chung (2000). Estimating the Underlying Change in Unemployment in the UK, with discussion *Journal of the Royal Statistical Society Series A*, 163, 303-39.

Harvey, A. and P. Kattuman. (2020a). Time Series Models Based on Growth Curves with Applications to Forecasting Coronavirus. *Harvard Data Science Review*. Special issue 1 - COVID -19. https://hdsr.mitpress.mit.edu/pub/ozgjx0yn

Harvey, A. and P. Kattuman. (2020b). A Farewell to R: Time Series Models for Tracking and Forecasting Epidemics. *CEPR working paper*, Issue 51, 7th October. https://cepr.org/content/covid-economics

Harvey, A.C. and S. Thiele (2020). Co-integration and Control: assessing the impact of events using time series data. *Journal of Applied Econometrics* (to appear)

Kamerlin, S. C. L. and P. M. Kasson (2020). Managing Coronavirus Disease 2019 Spread With Voluntary Public Health Measures: Sweden as a case study for pandemic control. *Clinical Infectious Diseases*, ciaa 864. https://doi.org/10.1093/cid/ciaa864

Koopman, S.J., R. Lit, and A.C. Harvey (2020). *STAMP 8.4 Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants Ltd.

Kwiatkowski, D., Phillips, P.C.B., Schmidt P. and Y. Shin (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 44, 159-78.

Maddala, G.S. and I-M. Kim (1998). *Unit Roots, Cointegration, and Structural Change*. Cambridge: Cambridge University Press.

Onder, G. (2020). Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *Journal of the American Medical Association*, 323, 1775-6.

Stock, J. and M. Watson (1988). Testing for Common Trends. *Journal of the American Statistical Association*, 83, 1097-1107.

Wallinga, J. and M. Lipsitch (2006). How Generation Intervals Shape the Relationship Between Growth Rates and Reproductive Numbers. *Proc.R.*

Soc. B., 274, 599-604.

A Data sources

The data for European countries was obtained from the European Centre for Disease Prevention and Control (ECDC) website, https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide,. For Florida the source was : https://covidtracking.com/data. The data were obtained at the end of August and the beginning of September. Data can be subject to revisions. For example the UK definition of deaths was changed in August to include only people who had a laboratory-confirmed positive COVID-19 test and had died within 28 days of the date the test result was reported. Before that it included anybody who had ever tested positive for COVID-19 no matter how long before the actual death.

Case-fatality statistics in Italy are based on defining COVID-19–related deaths as those occurring in patients who test positive for SARS-CoV-2 viaRTPCR, independently of preexisting diseases that may have caused death. This method may have resulted in overestimation; see Onder (2020).