Coronavirus and the Score-driven Negative Binomial Distribution

Andrew Harvey[a,b] and Rutger Lit[b]

[a]Faculty of Economics, Cambridge University

[b]Time Series Lab

May 29, 2020

Abstract

A new class of time series models, developed by Harvey and Kattuman (2020), is designed to predict variables which when cumulated are subject to an unknown saturation level. Such models are relevant for many disciplines, but the applications here are for deaths from coronavirus. When numbers are small a score-driven Negative Binomial model can be used. It is shown how such models can be estimated with the Time Series Lab - Score Edition software package and their specification assessed by statistical tests and graphics.

Key words: Generalized logistic; Gompertz curve; Negative binomial distribution; Time Series Lab.

1 Introduction

Following earlier work by Harvey (1984), Harvey and Kattuman (2020) develop time series models for predicting future values of a variable which when cumulated is subject to an unknown saturation level. Such models are relevant for many disciplines, but the examples here are in epidemiology and concern coronavirus.

The generalized logistic class of growth curves contains the logistic and Gompertz as special cases; see, for example, Panik (2014) and Daley and Gani (2001). They lead to a model in which the increase, $y_{t,}$ at time t depends on the cumulative total Y_t. Specifically

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

where $y_t = Y_t - Y_{t-1}$ and ε_t is a serially independent Gaussian disturbance with mean zero and constant variance, σ_ε^2, that is $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$. The cumulative number follows a logistic curve when $\rho = 2$ and a Gompertz when $\rho = 1$. Estimation is by OLS. Additional flexibility can be introduced into the model by letting the deterministic trend be time-varying. Thus

$$\ln y_t = \rho \ln Y_{t-1} + \delta_t + \varepsilon_t, \quad t = 2, ..., T,$$

where

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

and the normally distributed irregular, level and slope disturbances, ε_t, η_t and ζ_t, respectively, are mutually independent. When $\sigma_\eta^2 = \sigma_\zeta^2 = 0$, the trend is deterministic, that is $\delta_t = \delta - \gamma t$ with $\delta = \delta_0$. When only σ_ζ^2 is zero, the slope is fixed and the trend reduces to a random walk with drift. On the other hand, allowing σ_ζ^2 to be positive, but setting $\sigma_\eta^2 = 0$ gives an integrated random walk (IRW) trend, which when estimated tends to be relatively smooth. Such a model can be handled using the STAMP package.

The Kalman filter can be by-passed by adopting the reduced form, which comes from the innovations form of the Kalman filter so that

$$\ln y_t = \rho \ln Y_{t-1} + \delta_{t|t-1} + \varepsilon_t, \quad t = 3, ..., T, \tag{3}$$

where

$$\begin{aligned} \delta_{t+1|t} &= \delta_{t|t-1} - \gamma_{t|t-1} + \alpha_1 \varepsilon_t \\ \gamma_{t+1|t} &= \gamma_{t|t-1} + \alpha_2 \varepsilon_t, \end{aligned}$$

where α_1 and α_2 are non-negative parameters. Unless ρ is fixed, it may be hard to estimate in small samples. Restrictions on the trend, such as setting $\alpha_1 = 0$, may also be prudent.

2 Small numbers: the negative binomial distribution

When y_t is small, it may be necessary to adopt a discrete distribution, particularly if some observations are zero. The best choice is the negative binomial which, when parameterized in terms of a time-varying mean, $\xi_{t|t-1}$, and a fixed positive shape parameter, υ, has probability mass function (PMF)

$$p(y_t) = \frac{\Gamma(\upsilon + y_t)}{y_t! \, \Gamma(\upsilon)} \xi_{t|t-1}^{y_t} (\upsilon + \xi_{t|t-1})^{-y_t} (1 + \xi_{t|t-1}/\upsilon)^{-\upsilon}, \quad y_t = 0, 1, 2, \ldots,$$

with $Var_{t-1}(y_t) = \xi_{t|t-1} + \xi^2_{t|t-1}/\upsilon$. An exponential link function ensures that $\xi_{t|t-1}$ remains positive and at the same time yields an equation similar to (1):

$$\ln \xi_{t|t-1} = \rho \ln Y_{t-1} + \delta - \gamma t, \quad \rho \geq 1, \qquad t = 3, ..., T. \tag{4}$$

A stochastic trend may be introduced into the model as in (3). The conditional score framework of Creal et al. (2013) and Harvey (2013) suggests

$$\ln \xi_{t|t-1} = \rho \ln Y_{t-1} + \delta_{t|t-1}, \quad t = 3, ..., T, \tag{5}$$

where

$$\begin{aligned} \delta_{t+1|t} &= \delta_{t|t-1} - \gamma_{t|t-1} \\ \gamma_{t+1|t} &= \gamma_{t|t-1} + \alpha u_t, \qquad \alpha \geq 0, \end{aligned}$$

but with $u_t = y_t/\xi_{t|t-1} - 1$, which is the conditional score for $\ln \xi_{t|t-1}$, that is $\upsilon(y_t - \xi_{t|t-1})/(\upsilon + \xi_{t|t-1})$, divided by the information quantity. The dynamic Gompertz model has $\rho = 1$.

Predictions of future observations and the saturation level can be obtained from the recursions

$$\begin{aligned} \widehat{y}_{T+\ell|T} &= \widehat{\mu}^{\rho}_{T+\ell-1|T} \exp(\delta_T) \exp(-\gamma \ell) \\ \widehat{\mu}_{T+\ell|T} &= \widehat{\mu}_{T+\ell-1|T} + \widehat{y}_{T+\ell|T}, \qquad \ell = 1, 2, \ldots \end{aligned} \tag{6}$$

where δ_T is the level at time T and $\widehat{\mu}_{T|T} = Y_T$.

3 Germany

The negative binomial model, (4), with ρ set to one, was estimated using data, including some zeroes, from[1] March 11th 2020 up to, and including, May 6th. The result was $\widetilde{\alpha} = 0$ - corresponding to a deterministic trend - and $\widetilde{\gamma} = 0.071$. The fit and the ACF of the scores are shown in Figure 1. Including the daily (seasonal) effect produces the fit in Figure 2; the α coefficient is estimated close to zero and can be fixed at zero to constrain the effect to be deterministic. As might be expected from the excellent fit, the log-likelihood is significantly increased. With the daily seasonal included, the likelihood increases from -278.88 to -267.95. The parameter estimates are $\widetilde{\gamma} = 0.070$, $\widetilde{\delta}_T = -4.14$ and $\widetilde{\upsilon} = 13.25$. The final total is predicted to be 8714. Further estimation details are given in the appendix.

[1] The data is given on the ECDC website.

Figure 1

German deaths with score-driven Negbin model

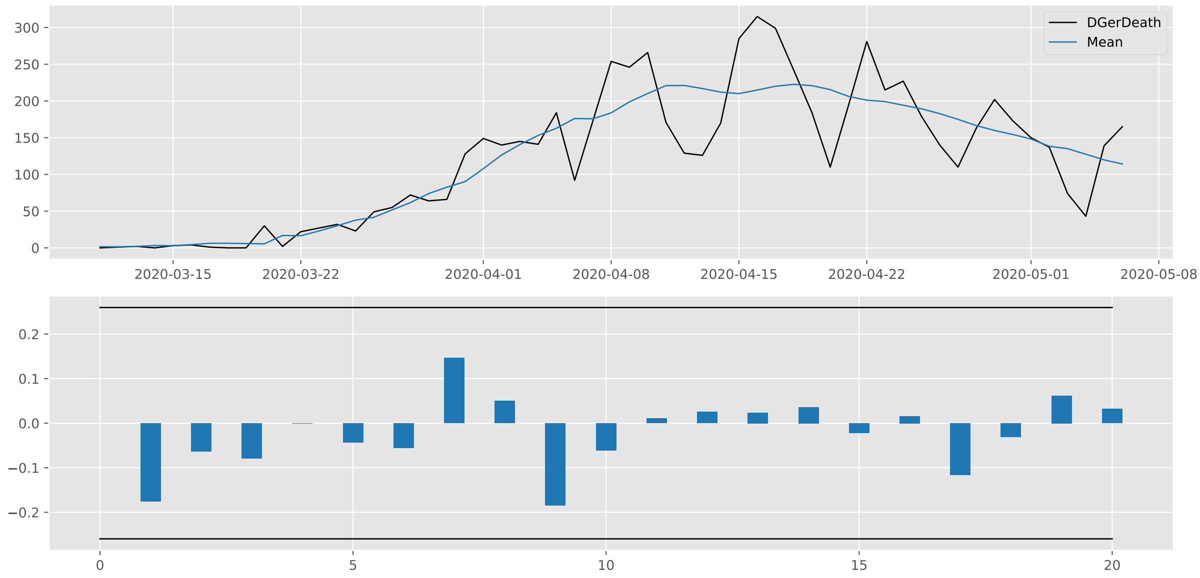
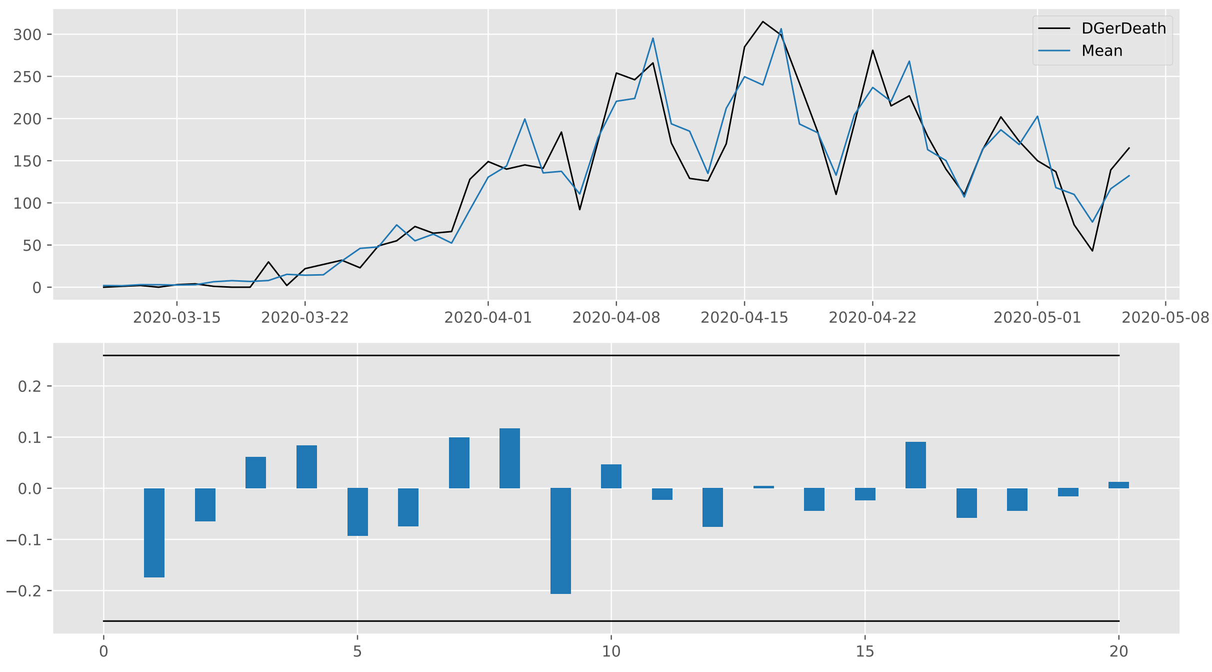


Figure 2

German deaths with Negbin and a daily effect

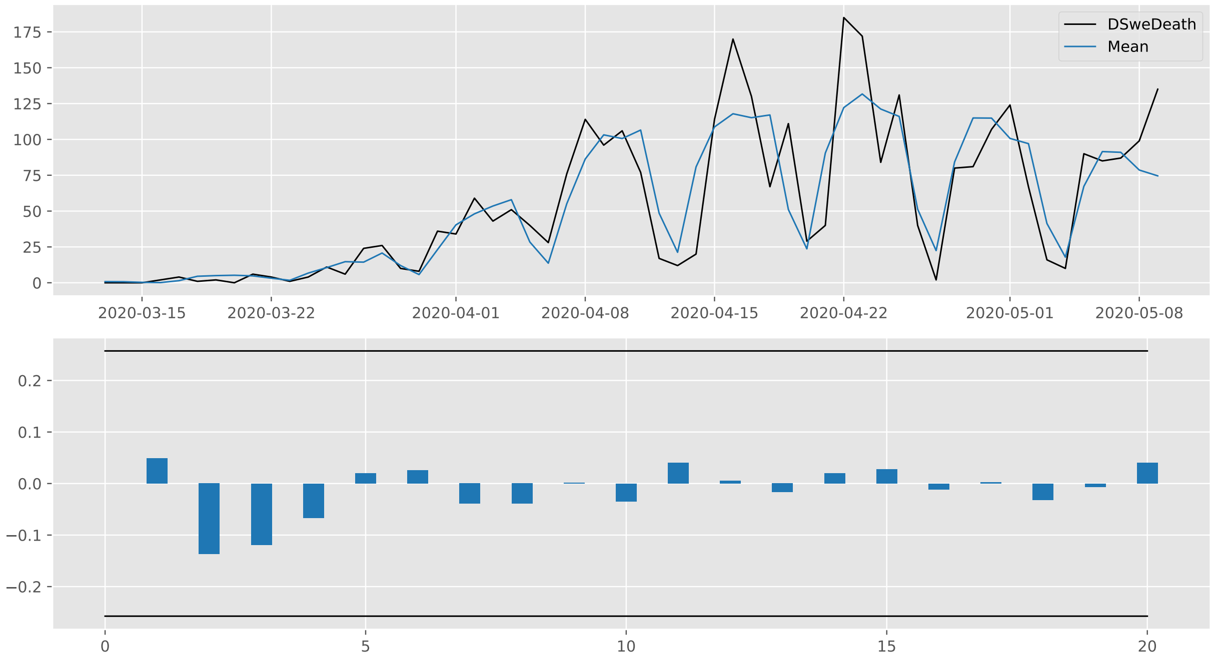


4 Sweden

Our series for deaths in Sweden starts on March 13th and ends on May 9th. There are zeroes near the beginning and the numbers are smaller than those in Germany. Again the trend α effectively is equal to zero. Including the daily effect is essential and as can be seen from Figure 3, the fit is again very good. The parameter estimates are $\widetilde{\gamma} = 0.061$, $\widetilde{\delta}_T = -4.05$ and $\widetilde{\upsilon} = 4.80$. The values of $\widetilde{\gamma}$ are $\widetilde{\delta}_T$ are similar in magnitude to those reported for Germany. Further estimation details are given in the appendix. The final total is predicted to be 4188. Given the much higher population of Germany this is relatively high and it could be ascribed to the less stringent lockdown in Sweden. However, it is not out of line with other countries[2] like Italy and UK.

Figure 3

Deaths in Sweden and Negbin model



[2]'Swedes, especially of the older generation, have a genetic disposition to social distancing anyway.' [Former Swedish PM]

References

Creal, D. D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28(5)*, 777–795.

Daley, D. J. and J. Gani (2001). *Epidemic modelling: an introduction*, Volume 15. Cambridge University Press.

Harvey, A. (1984). Time series forecasting based on the logistic curve. *Journal of the Operational Research Society 35*(7), 641–646.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, Volume 52. Cambridge: Cambridge University Press.

Harvey, A. C. and P. Kattuman (2020). Time series models based on growth curves with applications to forecasting coronavirus. Discussion paper, mimeo.

Koopman, S. J., R. Lit, and A. C. Harvey (2018). Structural time series analyser, modeller and predictor.

Lit, R., S. J. Koopman, and A. C. Harvey (2020). Time Series Lab - Score Edition. https://timeserieslab.com.

Panik, M. J. (2014). *Growth curve modeling: theory and applications*. John Wiley & Sons.

Time Series Lab output Germany

Session started at 2020-05-15 15:48

——————————— MODEL DESCRIPTION ———————————âĂŤ

Database
Model number: TSL001
The database used is: C:/...
The selection sample is: 1 - 57 (N = 1, T = 57 with 0 missings)

Distribution
The dependent variable is DGerDeath
The selected distribution is the Negative Binomial distribution with parameters:

Parameters	**Symbol**	**Time-varying**	**Domain**
Mean	λ	Yes	> 0
Dispersion	r	No	> 0

Parameter specification
λ = exp(Level + Seasonal(7) + Xβ + Score(1))
r = constant

Explanatory variables
Explanatory variable for location is: LGerDeaths_1

Initialisation of intensity
Initialisation component: Level
Type of initialisation: Estimate

——————————— PARAMETER OPTIMIZATION ———————————

Parameter starting values:

Parameter type	**Value**	**Free/Fix**
Log intensity: IRW κ	0.0200	Free
Log intensity: init	4.8098	Free
Log intensity: init slope	0.0000	Free
Log intensity: seasonal κ	0.0000	Fixed
Log intensity: init seasonal 1	0.0000	Free
Log intensity: init seasonal 2	0.0000	Free
Log intensity: init seasonal 3	0.0000	Free
Log intensity: init seasonal 4	0.0000	Free
Log intensity: init seasonal 5	0.0000	Free
Log intensity: init seasonal 6	0.0000	Free
Log intensity: β_LGerDeaths_1	1.0000	Fixed
Dispersion	5.0000	Free

Start estimation

it0	f=	-30.92671843
it10	f=	-5.40858290
it20	f=	-4.93143812
it30	f=	-4.77494383
it40	f=	-4.71288976
it50	f=	-4.70105712
it58	f=	-4.70103452

Strong convergence using numerical derivatives
Log-likelihood = -267.958968; T = 57

Optimized parameter values:

Parameter type	**Value**	**Free/Fix**
Log intensity: IRW κ	1.2477e-08	Free
Log intensity: init	-0.2255	Free
Log intensity: init slope	-0.0700	Free
Log intensity: seasonal κ	0.0000	Fixed
Log intensity: init seasonal 1	0.1985	Free
Log intensity: init seasonal 2	0.1362	Free
Log intensity: init seasonal 3	0.3596	Free
Log intensity: init seasonal 4	-0.1108	Free
Log intensity: init seasonal 5	-0.1568	Free
Log intensity: init seasonal 6	-0.4515	Free
Log intensity: β_LGerDeaths_1	1.0000	Fixed
Dispersion	13.2604	Free

Estimation process completed in 1.4288 seconds

——————————————— STATE INFORMATION ———————————————âĂŤ

Component intensity	**Initial**	**Time T**
Mean	1.9468	132.1634
Composite signal	0.6662	4.8840
Integrated random walk	-0.2255	-4.1437
Slope	-0.0700	-0.0700
Seasonal	0.1985	0.1985
Xβ	0.6931	8.8292

———————————————âĂŤ DIAGNOSTICS ———————————————

Summary statistics for residuals and score:

Statistic	**Residuals**	**Pearson**	**Score**
Observations	57.000	57.000	57.000
Obs no nan	57.000	57.000	57.000
Mean	0.232	-0.028	-0.031
Variance	683.856	1.578	0.290
Median	-0.634	-0.014	-0.004
Minimum	-55.980	-2.315	-1.000
Maximum	75.269	6.208	2.788
Skewness	0.129	1.881	2.085
Kurtosis	0.384	8.656	11.331

Test for autocorrelation:

Durbin-Watson

Lag	**Score**	**Pearson**	**Critical value**
1	2.278	2.362	1.50 - 2.50

Ljung-Box

Lag	**Score**	**Pearson**	**Critical value**
3	2.294	2.880	3.841
4	2.741	2.958	5.991
5	3.302	4.992	7.815
6	3.661	5.233	9.488
7	4.325	5.456	11.070
8	5.262	7.582	12.592
9	8.253	8.491	14.067
10	8.404	8.659	15.507
11	8.440	8.664	16.919
12	8.868	9.217	18.307
13	8.870	9.217	19.675

Time Series Lab output Sweden

Time Series Lab - Score Edition 1.10, Copyright © 2019-2020 Nlitn

Session started at 2020-05-15 15:48

——————————— MODEL DESCRIPTION ———————————âĂŤ

Database
Model number: TSL002
The database used is: C:/...
The selection sample is: 1 - 58 (N = 1, T = 58 with 0 missings)

Distribution
The dependent variable is DSweDeath
The selected distribution is the Negative Binomial distribution with parameters:

Parameters	**Symbol**	**Time-varying**	**Domain**
Mean	λ	Yes	> 0
Dispersion	r	No	> 0

Parameter specification
λ = exp(Level + Seasonal(7) + Xβ + Score(1))
r = constant

Explanatory variables
Explanatory variable for location is: LSweDeath_1

Initialisation of intensity
Initialisation component: Level
Type of initialisation: Estimate

——————————— PARAMETER OPTIMIZATION ———————————

Parameter starting values:

Parameter type	**Value**	**Free/Fix**
Log intensity: IRW κ	0.0200	Free
Log intensity: init	4.0023	Free
Log intensity: init slope	0.0000	Free
Log intensity: seasonal κ	0.0000	Fixed
Log intensity: init seasonal 1	0.0000	Free
Log intensity: init seasonal 2	0.0000	Free
Log intensity: init seasonal 3	0.0000	Free
Log intensity: init seasonal 4	0.0000	Free
Log intensity: init seasonal 5	0.0000	Free
Log intensity: init seasonal 6	0.0000	Free
Log intensity: β_LSweDeath_1	1.0000	Fixed
Dispersion	5.0000	Free

Start estimation

it0	f=	-25.99268241
it10	f=	-4.35338492
it20	f=	-4.17788034
it30	f=	-4.08485481
it40	f=	-4.07737987
it50	f=	-4.07557205
it60	f=	-4.07555992
it64	f=	-4.07555988

Strong convergence using numerical derivatives
Log-likelihood = -236.382473; T = 58

Optimized parameter values:

Parameter type	Value	Free/Fix
Log intensity: IRW κ	1.3142e-07	Free
Log intensity: init	-0.5833	Free
Log intensity: init slope	-0.0607	Free
Log intensity: seasonal κ	0.0000	Fixed
Log intensity: init seasonal 1	0.3624	Free
Log intensity: init seasonal 2	0.3370	Free
Log intensity: init seasonal 3	-0.4809	Free
Log intensity: init seasonal 4	-1.2662	Free
Log intensity: init seasonal 5	0.1173	Free
Log intensity: init seasonal 6	0.4528	Free
Log intensity: β_LSweDeath_1	1.0000	Fixed
Dispersion	4.7952	Free

Estimation process completed in 1.5124 seconds

——————————— STATE INFORMATION ———————————âĂŤ

Component intensity	Initial	Time T
Mean	0.8018	74.5765
Composite signal	-0.2209	4.3118
Integrated random walk	-0.5833	-4.0448
Slope	-0.0607	-0.0607
Seasonal	0.3624	0.3370
Xβ	0.0000	8.0196

———————————âĂŤ DIAGNOSTICS ———————————

Summary statistics for residuals and score:

Statistic	Residuals	Pearson	Score
Observations	58.000	58.000	58.000
Obs no nan	58.000	58.000	58.000
Mean	0.531	0.050	0.198
Variance	610.383	1.302	3.770
Median	-0.725	-0.127	-0.060
Minimum	-60.800	-1.810	-1.000
Maximum	62.816	5.092	14.254
Skewness	0.261	1.554	6.519
Kurtosis	0.975	4.771	44.376

Test for autocorrelation:

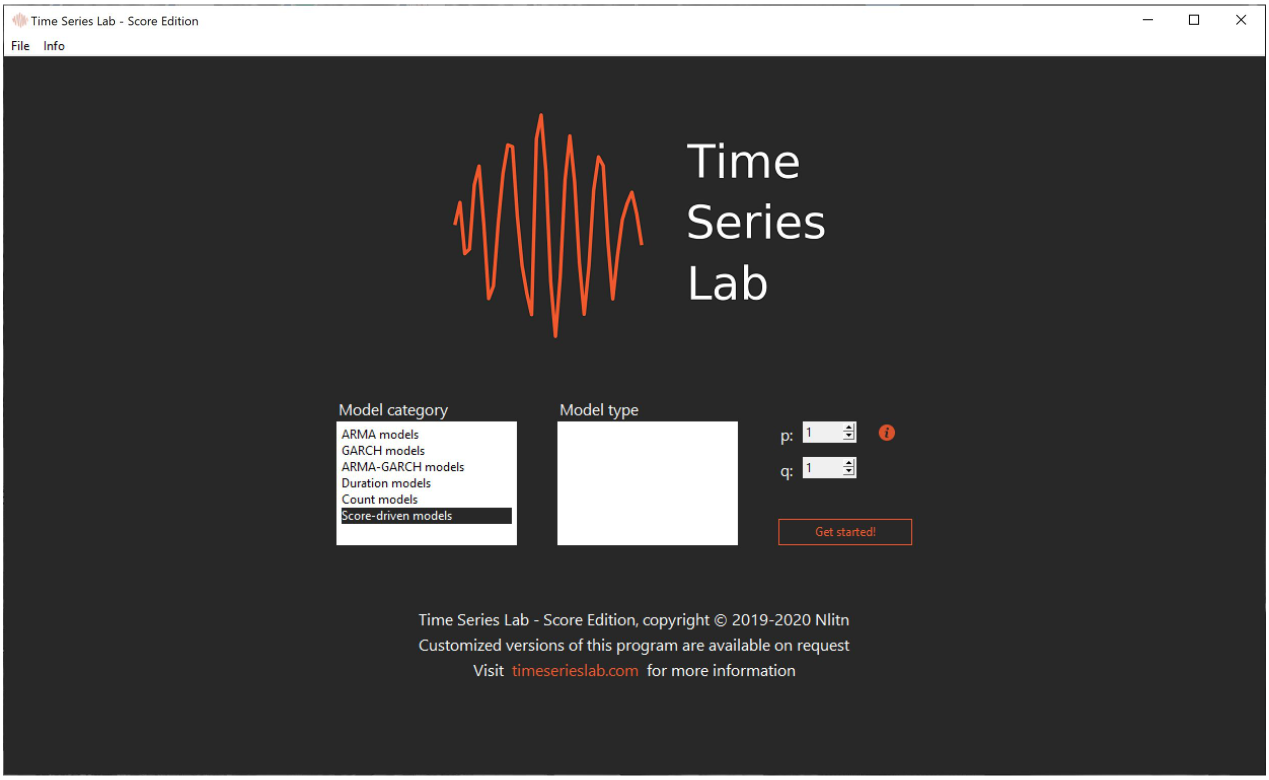
Durbin-Watson

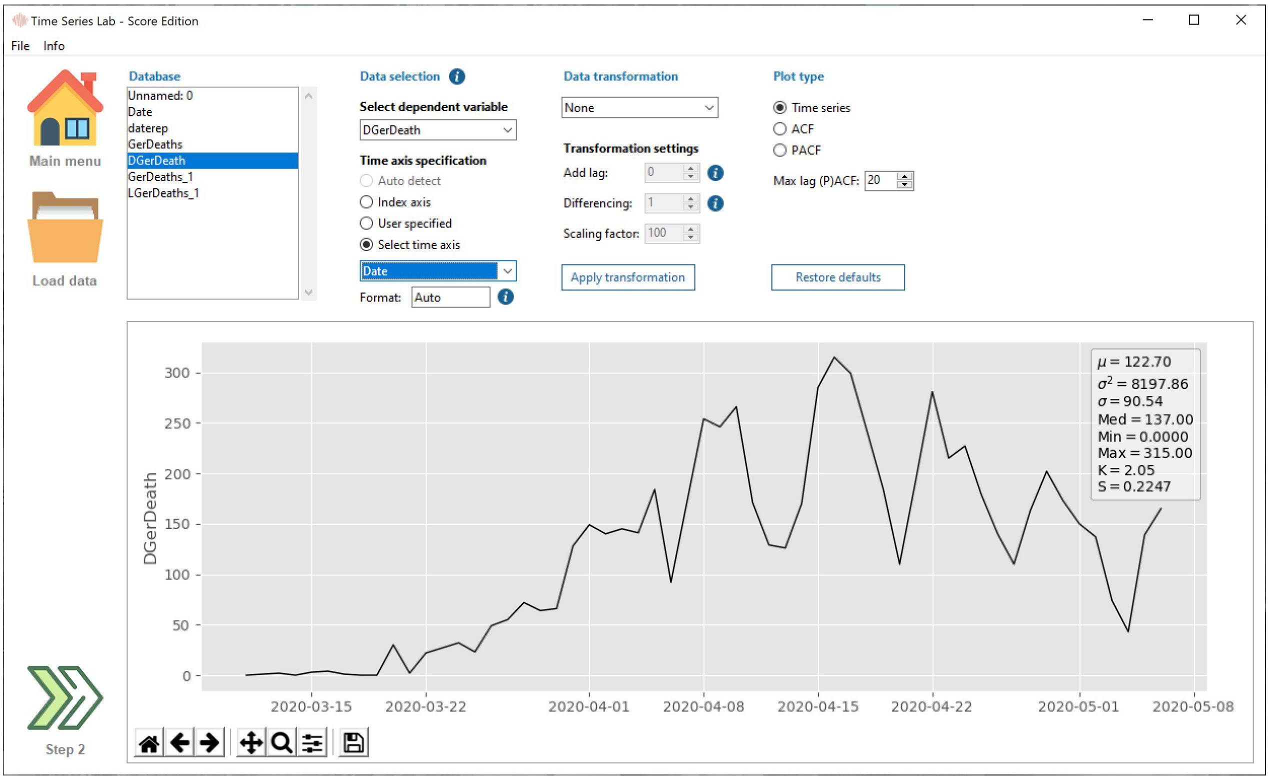
Lag	Score	Pearson	Critical value
1	1.874	1.606	1.50 - 2.50

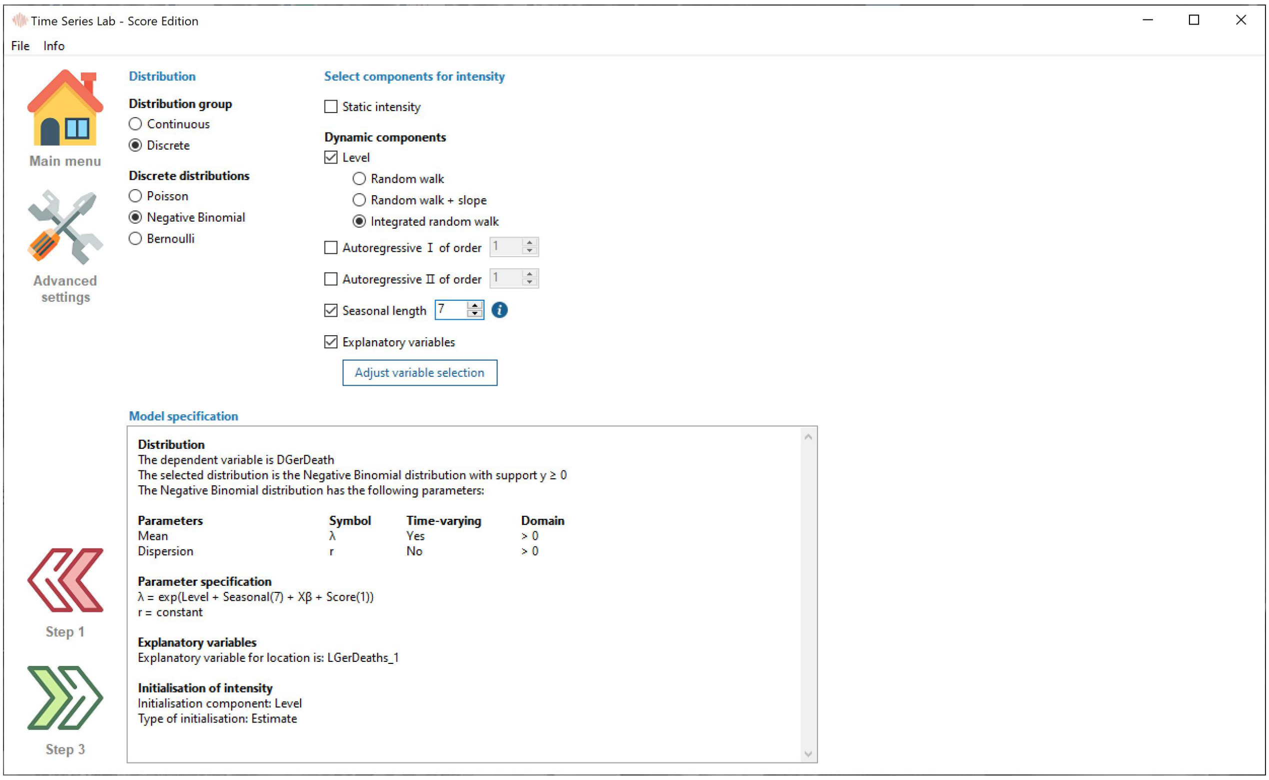
Ljung-Box

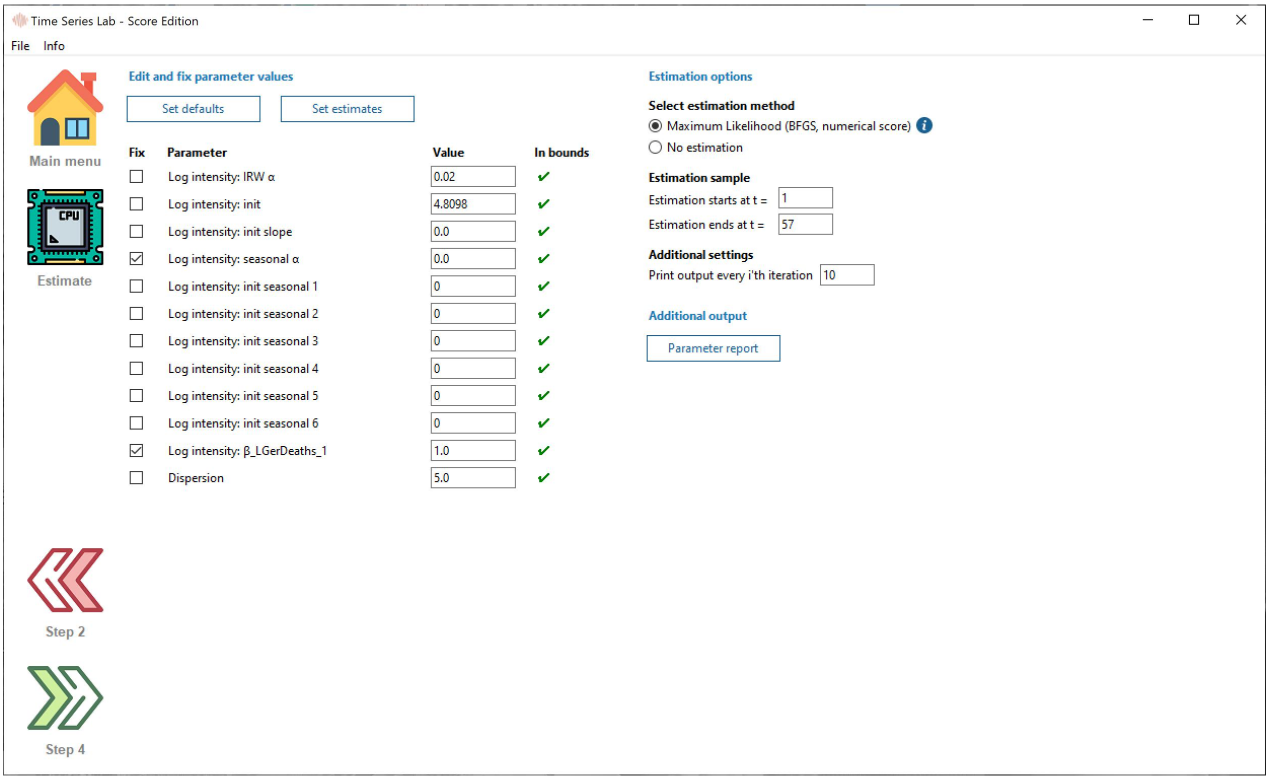
Lag	Score	Pearson	Critical value
3	2.202	4.181	3.841
4	2.490	5.094	5.991
5	2.516	5.481	7.815
6	2.559	5.992	9.488
7	2.662	7.504	11.070
8	2.768	9.383	12.592
9	2.768	9.409	14.067
10	2.856	9.763	15.507
11	2.974	9.847	16.919
12	2.976	9.903	18.307
13	2.997	9.983	19.675

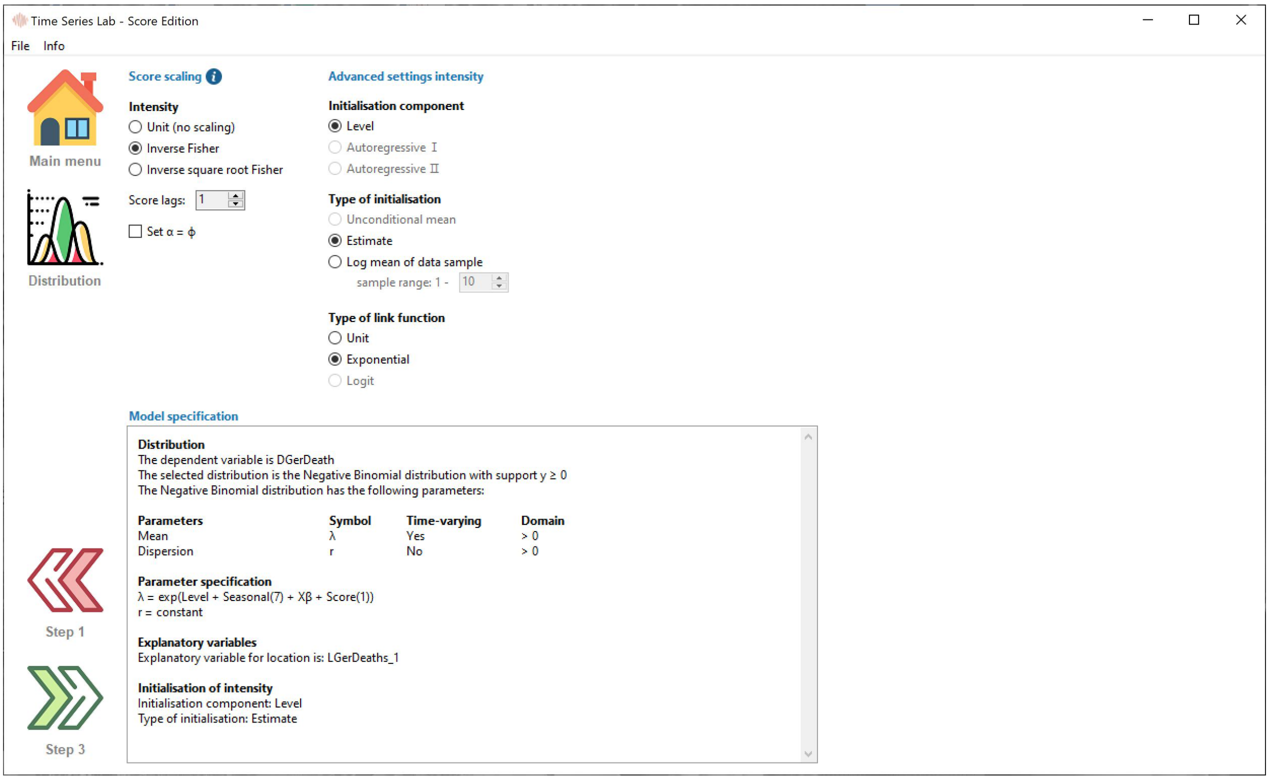
Time Series Lab - Screen Shots



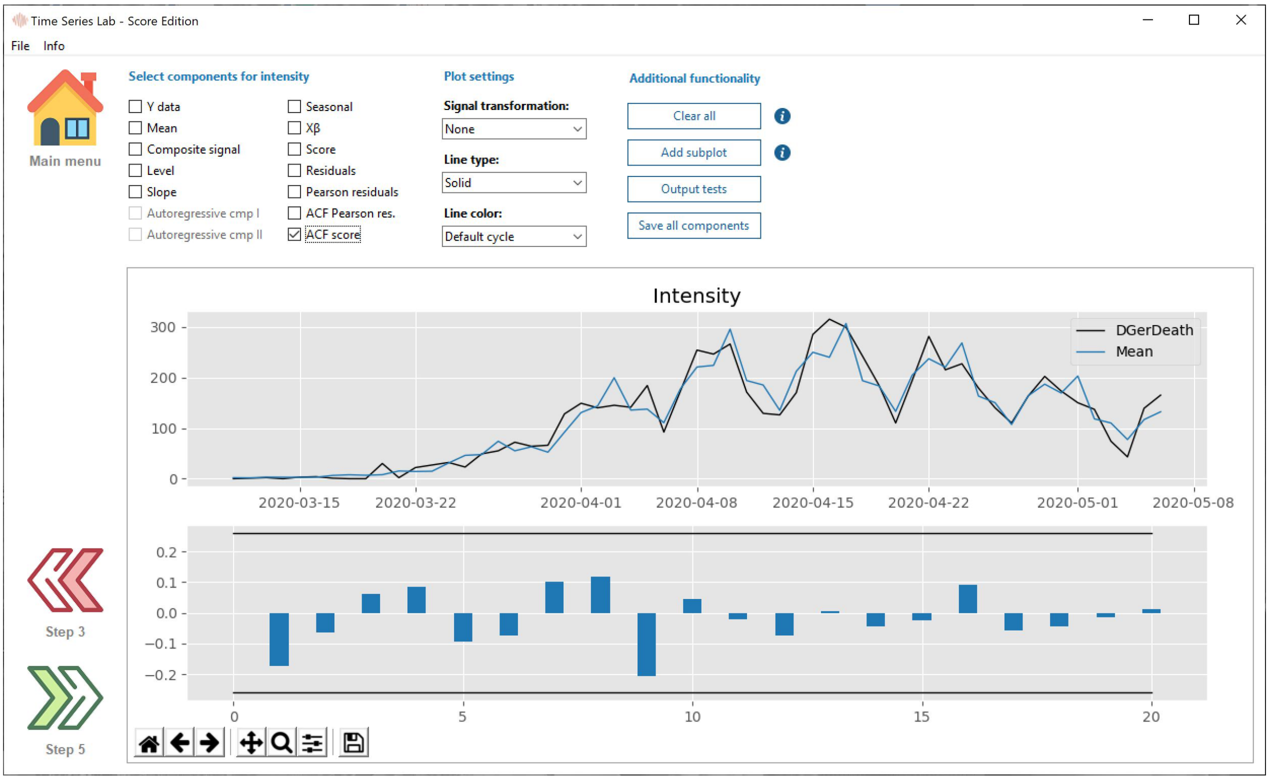



Time Series Lab - Score Edition
File Info
Main menu
Advanced settings
Step 1
Step 3
Distribution
Distribution group
Continuous
Discrete
Discrete distributions
Poisson
Negative Binomial
Bernoulli
Select components for intensity
Static intensity
Dynamic components
Level
Random walk
Random walk + slope
Integrated random walk
Autoregressive I of order 1
Autoregressive II of order 1
Seasonal length 7
Explanatory variables
Adjust variable selection
Model specification
Distribution
The dependent variable is DGerDeath
The selected distribution is the Negative Binomial distribution with support y ≥ 0
The Negative Binomial distribution has the following parameters:
Parameters Symbol Time-varying Domain
Mean λ Yes > 0
Dispersion r No > 0
Parameter specification
λ = exp(Level + Seasonal(7) + Xβ + Score(1))
r = constant
Explanatory variables
Explanatory variable for location is: LGerDeaths_1
Initialisation of intensity
Initialisation component: Level
Type of initialisation: Estimate


Time Series Lab - Score Edition
File Info
Main menu
Estimate
Step 2
Step 4
Edit and fix parameter values
Set defaults
Set estimates
Fix Parameter Value In bounds
Log intensity: IRW α 0.02
Log intensity: init 4.8098
Log intensity: init slope 0.0
Log intensity: seasonal α 0.0
Log intensity: init seasonal 1 0
Log intensity: init seasonal 2 0
Log intensity: init seasonal 3 0
Log intensity: init seasonal 4 0
Log intensity: init seasonal 5 0
Log intensity: init seasonal 6 0
Log intensity: β_LGerDeaths_1 1.0
Dispersion 5.0
Estimation options
Select estimation method
Maximum Likelihood (BFGS, numerical score)
No estimation
Estimation sample
Estimation starts at t = 1
Estimation ends at t = 57
Additional settings
Print output every i'th iteration 10
Additional output
Parameter report



Time Series Lab - Score Edition
File Info
Main menu
Distribution
Step 1
Step 3
Score scaling
Intensity
Unit (no scaling)
Inverse Fisher
Inverse square root Fisher
Score lags: 1
Set α = ϕ
Advanced settings intensity
Initialisation component
Level
Autoregressive I
Autoregressive II
Type of initialisation
Unconditional mean
Estimate
Log mean of data sample
sample range: 1 - 10
Type of link function
Unit
Exponential
Logit
Model specification
Distribution
The dependent variable is DGerDeath
The selected distribution is the Negative Binomial distribution with support y ≥ 0
The Negative Binomial distribution has the following parameters:
Parameters Symbol Time-varying Domain
Mean λ Yes > 0
Dispersion r No > 0
Parameter specification
λ = exp(Level + Seasonal(7) + Xβ + Score(1))
r = constant
Explanatory variables
Explanatory variable for location is: LGerDeaths_1
Initialisation of intensity
Initialisation component: Level
Type of initialisation: Estimate



Time Series Lab - Score Edition
File Info
Main menu
Step 3
Step 5
Select components for intensity
Y data
Mean
Composite signal
Level
Slope
Autoregressive cmp I
Autoregressive cmp II
Seasonal
Xβ
Score
Residuals
Pearson residuals
ACF Pearson res.
ACF score
Plot settings
Signal transformation:
None
Line type:
Solid
Line color:
Default cycle
Additional functionality
Clear all
Add subplot
Output tests
Save all components
Intensity
DGerDeath
Mean
300
200
100
0
2020-03-15
2020-03-22
2020-04-01
2020-04-08
2020-04-15
2020-04-22
2020-05-01
2020-05-08
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20