Take home exam

Econometrics III - Time Series

PhD in Business Economics

Instructor: Hedibert Freitas Lopes

Starts at 9am of Tuesday, February 23th 2021

Ends at 12pm of Thursday, February 25th 2021

This is an individual exam. You are free to consult books, videos, short-courses or other studying material. You are not suppose though to consult your colleagues or any other individual to assist you throughout the examination.

Modeling daily death counts by COVID-19

Let y_t and x_t denote daily counts of deaths and confirmed cases on day t by COVID-19 for a given country (Brazil, US and UK, in our study), as recorded, for instance, here. You should download the number of confirmed cases and the death counts for Brazil, US and UK from the two files:

```
library("tidyverse")
library("readr")
url="https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/"
url.cases=paste(url,"time_series_covid19_confirmed_global.csv",sep="")
url.deaths=paste(url,"time_series_covid19_deaths_global.csv",sep="")
cases=read_csv(url(url.cases))
deaths=read_csv(url(url.deaths))
```

The first 4 columns of both files are: 1) `Province/State`, 2) `Country/Region`, 3) `Lat` and 4) `Long`. When I downloaded both files, on Sunday, February 14th, they had 273 rows and 393 columns. Therefore, 389 days 345 days in 2020 (starting on January 22nd) and 44 days in 2021 (up to February 13th). Whenever you download the data yourself, let us focus on the most recent 300 days. Your goal is to predict (in-sample) and forecast (out-of-sample) y_{t+h} given information up to time t. Below are two specific modeling scenarios:

A) SARIMA-type models

Basically, you should estimate y_t through various ARIMA$(p,d,q)(P,D,Q)_7$ models. The period of the time series of death counts are (artificially) weekly, therefore the 7 in the seasonal modeling specification. You might want to (learn and) use the automatic search function `auto.arima` available in the R package `forecast`. Learn which SARIMA models fit the number of deaths in Brazil, USA and UK and use the respective models to forecast the next 28 days

(4 weeks). Discuss the similarity and/or idiosyncrasies of each one of the countries. The file `sp.csv` contains cases and deaths for the City of São Paulo from March 2nd 2020 to February 14th 2021. Repeat your analysis for São Paulo's data and discuss your additional findings. Notice that by February 23rd you will have 10 extra days to use to compare to the forecast from the SARIMA models.

```
library("forecast")
data = ts(y,frequency=7)
fit = auto.arima(data)
fit
names(fit)
fit$arma
```

B) AR plus lagged-regressor models

Similar to A), I would like you to model y_t as follows:

$$y_t = \phi_0 + \sum_{i=1}^{14} \phi_i y_{t-h-i} + \sum_{j=1}^{14} \beta_j x_{t-h-j} + \varepsilon_t,$$

where ε_t are i.i.d. $N(0, \sigma^2)$ and $h = 0, 7, 14$. We are essentially running Gaussian multiple linear regressions y_t on z_t^h, where

$$\begin{aligned} z_t^0 &= (1, y_{t-1}, \ldots, y_{t-14}, x_{t-1}, \ldots, x_{t-14})', \\ z_t^7 &= (1, y_{t-8}, \ldots, y_{t-21}, x_{t-8}, \ldots, x_{t-21})', \\ z_t^{14} &= (1, y_{t-15}, \ldots, y_{t-28}, x_{t-15}, \ldots, x_{t-28})'. \end{aligned}$$

The goal is to find the best model that predicts y_t based on information available up to the previous day, or seven days before or even fourteen days before. In matrix notation, the three regressions become

$$y = Z_h \theta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$, y is $(n \times 1)$ and $Z_h = (z_1^h, z_2^h, \ldots, z_n^h)'$ is $(n \times k)$, for $n = 300$, $k = 29$ and $h = 0, 7, 14$. When $h = 0$, the model becomes an AR(14) with additional 14 lags of x_t, which resembles the models fit in A). A chunk of my R code to perform variable selection is as follows.

```
library("leaps")
dataZ = data.frame(y=y,Z=Zh)
# regsubsets runs all possible models and picks the best one
# for each number of regressors from 1 to nvmax.
models = regsubsets(y~Z,data=dataZ,nvmax=28)
summaries = summary(models)
summaries
# Picking the best model according to Mallow's Cp and BIC
```

```
ord = 1:28
cp = summaries$cp
bic = summaries$bic
select.cp = ord[cp==min(cp)]
select.bic = ord[bic==min(bic)]
# Picking in Zh the columns selected by the best models
Z.cp = Zh[,ord[summaries$which[select.cp,]]]
Z.bic = Zh[,ord[summaries$which[select.bic,]]]
# Fitting the best models
fit.cp = lm(y~Z.cp-1)
fit.bic = lm(y~Z.bic-1)
```

Learn which models (columns of Z_h, $h = 0, 7, 14$) best fit the number of deaths in Brazil, USA and UK and use them to forecast the number of deaths the next day ($h = 0$), the next 7 days ($h = 7$) and the next 14 days ($h = 14$), respectively. Discuss the similarity and/or idiosyncrasies of each one of the countries. Repeat your analysis for São Paulo's data and discuss your additional findings. Notice that by February 23rd you will have 10 extra days to use to compare to the forecast from the AR plus lagged regressors model. Compare your findings in A) and B).