MODELING AND FORECASTING THE GROSS ENROLLMENT RATIO IN ROMANIAN PRIMARY SCHOOL

MARINOIU CRISTIAN
ASSOC. PROF. PH.D., PETROLEUM-GAS UNIVERSITY OF PLOIESTI, ROMANIA
e-mail: marinoiu_c@yahoo.com

Abstract
The gross enrollment ratio in primary school is one of the basic indicators used in order to evaluate the proposed objectives of the educational system. Knowing its evolution allows a more rigorous substantiation of the strategies and of the human resources politics not only from the educational field but also from the economic one. In this paper we propose an econometric model in order to describe the gross enrollment ratio in Romanian primary school and we achieve its prediction for the next years, having as a guide the Box-Jenkins’s methodology. The obtained results indicate the continuous decrease of this rate for the next years.

Key-words: time series, forecast, ARIMA, gross enrollment ratio

Classification JEL: C50, I21

1. Introduction

The evaluation of the working condition of an educational system is achieved on the basis of a set of indicators which reflect the fundamental aspects of such a system: the educational context, the educational resources, the access and the participation in education and the results of education. The access and the participation in education are quantified through specific indicators: the gross scholar enrollment ratio, the net scholar enrollment ratio, the school life expectancy and so on.

The primary scholar gross enrollment ratio (GER) represents the percentage of pupils included in primary education, regardless of their age, relative to the total population from the official age-group which corresponds to the primary educational system [1]. This percentage can exceed 100% because we can include into the primary school pupils who are outside the official limits of age (either because they were too early or too late registered in school, or because of dropping school).

GER is an important indicator in the process of evaluating the conditions in the educational system from a country or from a region. Predicting its evolution can be useful in order to implement the medium and long term policies and strategies from the human resources field and, specially, from the educational system.

The study presented in this paper is based on the annual values of GER for Romania, which correspond to the period 1986-2011, presented in [2]. The objective is to build an econometric model starting from these data, to validate it and to make predictions about the GER evolution for the next years.

2. The GER time series

The GER values observed annually form a time series and may be considered as realizations of a random process $y_t$. The time series is graphically represented in figure 1 and it has the mean equal to 97.16584 and the variance equal to 43.35884.
The graph does not present clues that suggest an unstable variance. At the same time some increases and decreases are obvious for long periods of time (for example the periods 2000-2005 and 2005-2011), which suggest the fact that the series is not stationary. In order to verify this hypothesis we applied the ADF (Augmented – Dickey-Fuller) and KPSS (Kwiatkowski-Phillips-Schmidt-Shin) tests [3], used in order to detect the non-stationarity, the level of significance being fixed to 0.05. The p-values of the two tests are represented in table 1.

<table>
<thead>
<tr>
<th>Test</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Augmented – Dickey-Fuller(ADF)</td>
<td>0.237</td>
</tr>
<tr>
<td>Kwiatkowski-Phillips-Schmidt-Shin(KPSS)</td>
<td>0.02323</td>
</tr>
</tbody>
</table>

The null hypothesis of ADF test is that the series is not stationary, while the null hypothesis of KPSS test is that the series is stationary. In the case of ADF test we have 0.237>0.05 so we cannot reject the hypothesis that the series is not stationary. In the case of KPSS test we have 0.02323<0.05 so we reject the hypothesis that the series is stationary.

As a consequence, taking into account all these observations, we arrive at the conclusion that the series is not stationary.

### 3. The selection of the best model

Due to the fact that the GER time series is not stationary, we will try to transform it into a stationary time series. We may often obtain a stationary time series $z_t$ from a non-stationary time series $y_t$ through a transformation such as: $z_t = y_t - y_{t-1}$, named the difference of first order [4]. The difference of first order of the time series GER, named $dif1$ is represented in figure 2.
Figure 2. The time series difference of first order (dif1)
Source: made by the author in R

The aspect of the time series dif1 suggests that the series of the differenced data seems to be stationary. By applying the KPSS test to the time series dif1 we obtain the p-value equal to 0.1. Because 0.1>0.05 we cannot reject the hypothesis that the time series of the differences of first order (dif1) is stationary.

The choice of the best model for the GER time series can be done automatically by using the special functions of different software packages or "manually" by selecting from many candidate models the model which has the best values for certain indicators (for example the smallest values for the AIC or BIC statistics). In order to select the most suitable model we preferred to use the automatic variant based on the implementation of the function auto.arima() [5] from the library forecast [6] of the software package R.

The selected model on the basis of this function is, in this case, ARIMA(1,1,0) with drift, the results being as it follows:

Coefficients:
  ar1  drift
  0.5213  -0.2103
s.e.  0.1716   1.1560
sigma^2 estimated as 8.253: log likelihood=-59.04
AIC=124.09   AICc=125.23   BIC=127.75

As a consequence, the proposed model is: \( y_t - y_{t-1} = 0.5213(y_{t-1} - y_{t-2}) - 0.2103 \)

4. The validation of the selected model

The validation of the selected model involves checking the following conditions:
- The residuals of the model are not correlated;
- The residuals of the model are normally distributed with mean 0 and constant variance.

The graphical representation of residuals (fig. 4) suggests that errors are not correlated (we cannot observe a certain pattern of these residuals).

Figure 4. The graph of residuals
Source: made by the author in R
Moreover, the graph of the values of autocorrelation coefficients (fig. 5) shows that all the values of the autocorrelation coefficients do not exceed the significance 95% bounds, with one single exception: the autocorrelation coefficient $r_4$. The exception which refers to the autocorrelation coefficient $r_4$ can be ignored, because its appearance can be explained probably due to chance: we expect one out of 20 autocorrelation coefficients to have an exceeding of the significance 95% bounds. In conclusion, the analysis of the correlogram of the residuals suggests that errors are not correlated.

![Figure 5](Image)

**Figure 5.** The correlogram of residuals. The 95% significance bounds are marked by dashed lines. 
Source: made by the author using the functions from R.

The Ljung-Box test [7]-[8] is one of the tests through which we can verify the null hypothesis:

$H_0$: residuals are not correlated
against

$H_a$: residuals are correlated.

The results of using of this test at the level of significance 0.05 are represented in figure 6. We may observe that all the p-values which correspond to the first 10 lags are superior to 0.05 (value marked in the figure by a dashed line), which signify that we accept the hypothesis that residuals are not correlated. Consequently, the results of Ljung-Box test confirm the fact that residuals are not correlated.

![Figure 6](Image)

**Figure 6.** In this picture, the p-values of the test Ljung-Box for the first 10 lags are marked with circles. The dashed line emphasizes the value 0.05 under which the null hypothesis is rejected. 
Source: made by the author in R

In order to verify the normality of residuals, we will analyze the histogram of residuals, the graph qqplot of residuals and the results of the Shapiro-Wilk test [9].

The histogram of residuals (fig. 7 left) suggests that we can accept the fact that residuals are normally distributed. The statement is also sustained by the graph qqplot (fig. 7 right): the points of the graph are situated, with some exceptions, along the line which passes through the quartiles 1 and 3.
Using the Shapiro-Wilk test we verify the null hypothesis
\( H_0 \): residuals are normally distributed against
\( H_a \): residuals are not normally distributed.

The p-value is equal to 0.1807. Because 0.1807>0.05 (level of significance) we cannot reject the null hypothesis \( H_0 \). Taking into account the histogram of residuals, the graph qqplot of residuals and the results of the test Shapiro-Wilk we conclude that residuals are normally distributed.

Because the residuals of the model obtained \( ARIMA(1,1,0) \) are not correlated and normally distributed we conclude that the model obtained is valid.

5. The forecast of the GER time series values

The fact that the model \( ARIMA(1,1,0) \) is valid allows us to proceed to the next step, predicting the GER time series values for the next years and building prediction intervals for these forecasts. The results presented in Table 4 are obtained using the function `forecast` from R.

<table>
<thead>
<tr>
<th>Year</th>
<th>Forecast</th>
<th>Lo 80</th>
<th>Hi 80</th>
<th>Lo 95</th>
<th>Hi 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>95.68132</td>
<td>91.99957</td>
<td>99.36307</td>
<td>90.05057</td>
<td>101.3121</td>
</tr>
<tr>
<td>2016</td>
<td>95.35964</td>
<td>88.65700</td>
<td>102.06227</td>
<td>85.10883</td>
<td>105.6104</td>
</tr>
<tr>
<td>2017</td>
<td>95.09130</td>
<td>85.68373</td>
<td>104.49887</td>
<td>80.70367</td>
<td>109.4789</td>
</tr>
<tr>
<td>2018</td>
<td>94.85077</td>
<td>83.05095</td>
<td>106.65059</td>
<td>76.80450</td>
<td>112.8970</td>
</tr>
<tr>
<td>2019</td>
<td>94.62473</td>
<td>80.69938</td>
<td>108.55008</td>
<td>73.32775</td>
<td>115.9217</td>
</tr>
<tr>
<td>2020</td>
<td>94.40625</td>
<td>78.57241</td>
<td>110.24009</td>
<td>70.19048</td>
<td>118.6220</td>
</tr>
<tr>
<td>2021</td>
<td>94.19171</td>
<td>76.62401</td>
<td>111.75941</td>
<td>67.32422</td>
<td>121.0592</td>
</tr>
<tr>
<td>2022</td>
<td>93.97922</td>
<td>74.81870</td>
<td>113.13974</td>
<td>64.67574</td>
<td>123.2827</td>
</tr>
<tr>
<td>2023</td>
<td>93.76780</td>
<td>73.12963</td>
<td>114.40597</td>
<td>62.20445</td>
<td>125.3312</td>
</tr>
<tr>
<td>2024</td>
<td>93.55694</td>
<td>71.53642</td>
<td>115.57746</td>
<td>59.87947</td>
<td>127.2344</td>
</tr>
</tbody>
</table>

Source: Results obtained by the author using functions from R.

In figure 8 the values of the GER time series, the forecasts for the period 2015-2024 and their 80%, and respectively 90% prediction intervals are plotted. We can notice the decreasing tendency of the time series values for the next years.
Figure 8. The graph of the GER time series values, the forecasts for the years 2015-2024 (the line included in the shaded area) as well as the 80% (the dark shaded area) and the 95% prediction intervals (the light shaded area). Source: made by the author using the functions from R.

6. Conclusions

This paper introduced an econometric model for the gross enrollment ratio in Romanian primary school (GER), having as a guide the methodology Box-Jenkins. In order to process the data and to analyze the results we used the software package R. The predictions realized on the basis of the model obtained show a continuous decrease of GER values in the next years. The results obtained can be a good starting point for any study which has as a theme the establishment of personnel policies in the educational system and the preparation of the Romanian labour force.

7. Bibliography

[3]***https://www.otexts.org/fpp/8/1
[4]***http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm
[5]***http://www.inside-r.org/packages/cran/forecast/docs/auto.arima
[6]***http://www.jstatsoft.org/v27/i03/paper
[9]***http://www.answers.com/topic/shapiro-wilk-test
[10]***http://altons.github.io/r/2013/05/12/arima-models-in-a-nutshell/