# Spurious Regression Analysis

Abhradipta Ghosh (MD2101)
Mainack Paul (MD2111)
Somnath Bera (MD2122)

April 30, 2023

## 1 Introduction

Testing and allowing for non-stationary time-series data has been one of the major themes in econometrics over the past quarter-century or so. In their influential and relatively early contribution, Granger and Newbold (1974) drew our attention to some of the likely consequences of estimating a spurious regression model. From their studies we would conclude that if a regression equation relating economic variables is found to have strongly autocorrelated residuals, equivalent to a low Durbin-Watson value, the only conclusion that can be reached is that the equation is mis-specified, whatever the value of $R^2$ observed. They argued that the levels of many economic timeseries are integrated or nearly so, and that if such data are used in a regression model a high $R^2$ value is likely to be found even when the series are independent of each other. They also illustrated that the regression residuals are likely to be autocorrelated, as evidenced by a very low value for the Durbin-Watson (DW) statistic. However Phillips (1986) who provided a formal analytical explanation for the behaviour of the Ordinary Least Squares (OLS) coefficient estimator, the associated t-statistics and F-statistic, and the $R^2$ and DW statistics in such models.

Phillips (1986) developed a sophisticated asymptotic theory that he used to prove that in a spurious regression, the DW statistic converges in probability to zero, the OLS parameter estimators and $R^2$ converge weakly to non-standard limiting distributions, and the $t$-ratios and $F$ statistic diverge in distribution as $T$ tends to $\infty$ . Phillips solved the spurious regression problem, and proved that the unfortunate consequences of modelling with integrated data cannot be eliminated by increasing the sample size. This paper uses Phillips asymptotic theory to demonstrate that the pitfalls of estimating a spurious regression extend to the application of standard diagnostic tests for the normality or homoskedasticity of the model error term. We prove that the associated test statistics diverge in distribution as the sample size grows, so that one is led inevitably to the false conclusion that there is a problem with the usual assumptions about the error term. In fact, the real problem is a failure to take account of the non-stationarity of the data when specifying the model. The positive aspect of these results is that they provide us with an extended basis for detecting that we are unwittingly trying to estimate a spurious regression model.

Spurious regression is a statistical model that shows misleading statistical evidence of a linear relationship; in other words, a spurious correlation between independent non-stationary variables. The seminal study of Granger and Newbold showed that when two independent random walks are used in a linear regression, one tends to find a significant relationship between the variables. This phenomenon is termed spurious regression. It occurs when a pair of independent series, but with strong temporal properties, are found apparently to be related according to standard inference in an OLS regression.

In the case of a spurious regression, some statistically significant coefficients are obtained and the $R^2$ is very high. This high $R^2$ and significant $t$-values might mislead us to nonsense regressions. Only the Durbin-Watson (DW) ratio is a clue to detect a nonsense regression because its value is low.

Here we are going to check the behaviours of Spurious regression through simulation studies and also do analysis on a real data.

# 2 Spurious Regression

### How nonsense regression arise

Granger and Newbold (1974) showed in their paper that how the nonsense regression can arise. Let us consider the usual linear regression model with stochastic regressors :

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where $\mathbf{Y}$ is a $(T \times 1)$ vector of observations on a dependent variable, $\beta$ is a $(k \times 1)$ vector of coefficients whose first member $\beta_0$ represents a constant term and $\mathbf{X}$ is a $(T \times K)$ matrix containing a column of ones and $T$ observations on each of $(K-1)$ independent variables which are stochastic, but distributed independently of the $(T \times 1)$ vector of errors $\epsilon$.

It is generally assumed that,

$$\mathbb{E}(\epsilon) = 0 \tag{2}$$

$$\mathbb{E}(\epsilon\epsilon') = \sigma^2 I \tag{3}$$

A test of the null hypothesis that the independent variables contribute nothing towards explaining variation in the dependent variable can be framed in terms of the coefficient of multiple correlation $R^2$. The null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0 \tag{4}$$

and the test statistic

$$F = \frac{(T-K) \cdot R^2}{(K-1) \cdot (1-R^2)}$$

is compared with tabulated values of Fisher F distribution with $(K-1)$ and $(T-K)$ degrees of freedom, normality being assumed. Of course, it is entirely possible that, whatever the properties of the individual time series, there does exist some $\beta$, so

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta$$

satisfies the conditions 2 and 3. However, to the extent that the $Y_t$'s do not constitute a white noise process, the null hypothesis 4 cannot be true, and tests of it are inappropriate.

Next, let us suppose that the null hypothesis is correct and one attempts to fit a regression of the form 1 to the levels of economic time series. Suppose the series are non-stationary or highly autocorrelated. In such a situation the test procedure just described breaks down, since the quantity F statistic will not follow Fisher F distribution under the null hypothesis 4. This follows since under that hypothesis the residuals from eq.2.

$$\epsilon_t = \mathbf{Y} - \beta_0 \quad , \forall t$$

will have the same autocorrelation properties as the $Y_t$ series. Some idea of the distributional problems involved can be obtained from consideration of the case

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

where it is assumed that $Y_t$, and $X_t$, follow the independent first order autoregressive processes,

$$Y_t = \phi Y_{t-1} + a_t \tag{5}$$
$$X_t = \phi^* X_{t-1} + b_t \tag{6}$$

where $a_t$ , $b_t$ are iid noise and independently distributed. In this case, $R^2$ is simply the square of the ordinary sample correlation between $Y$, and $X$. Kendall (1954) gives

$$var(R) = \frac{1 + \phi\phi^*}{T(1 - \phi\phi^*)}$$

Thus a high value of $R^2$ should not, on the grounds of traditional tests, be regarded as evidence of a significant relationship between autocorrelated series. Also a low value of d strongly suggests that there does not exist $\beta$ such that $\epsilon$ in eq. (2) satisfies eq. (3). So spurious problem arises.

## Concept

Consider a regression of the form

$$\boldsymbol{y}_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$$

Where $y_t$ and $x_t$ might be non-stationary. If there does not exist some population value for $\beta$ for which the residual $u_t = y_t - x_t'\beta$ is $I(0)$, then OLS is quite likely to produce spurious results. A general statement of the spurious regression problem can be made as follows. Let $y_t$, be an $(n \times 1)$

vector of $I(1)$ variables. Define $g = (n-1)$ and partition of $y_t$ as

$$y_t = \begin{bmatrix} y_{1t} \\ \boldsymbol{y_{2t}} \end{bmatrix}$$

where $\boldsymbol{y_{2t}}$, denotes a $(g \times 1)$ vector. Consider the consequences of an OLS regression of the first variable on the others and a constant,

$$y_{1t} = \alpha + \boldsymbol{\gamma}'\boldsymbol{y_{2t}} + u_t$$

Then the OLS coefficient estimates for a sample of size $T$ are given by

$$\begin{bmatrix} \hat{\alpha_T} \\ \hat{\boldsymbol{\gamma_T}} \end{bmatrix} = \begin{bmatrix} T & \sum \boldsymbol{y'_{2t}} \\ \sum \boldsymbol{y_{2t}} & \sum y_{1t}\boldsymbol{y'_{2t}} \end{bmatrix}^{-1} \begin{bmatrix} \sum y_{1t} \\ \sum \boldsymbol{y_{2t}} y_{1t} \end{bmatrix}$$

where $\sum$ indicates summation over $t$ from 1 to $T$. It turns out that even if $y_{1t}$ is completely unrelated to $y_{2t}$, the estimated value of $\boldsymbol{\gamma}$ is likely to appear to be statistically significantly different from zero.

Unless there is some value for $\gamma$ such that $(y_{1t} - \hat{\gamma}'\boldsymbol{y_{2t}})$ is stationary, the OLS estimate $\hat{\boldsymbol{\gamma_T}}$ will produce spurious results, in the sense that the $R^2$ will be high, the F test is virtually certain to reject any null hypothesis if the sample size is sufficiently large and Durbin-Watson statistic DW will be low.

## Cures for Spurious Regressions

There are three ways in which the problems associated with spurious regressions can be avoided. Those are as following,

**Approach:1**

The first approach is to include lagged values of both the dependent and independent variable in the regression.

$$y_{1t} = \alpha + \phi y_{1,(t-1)} + \gamma y_{2t} + \delta y_{2,(t-1)} + u_t \tag{7}$$

Here, $\phi = 1$ and $\gamma = \delta = 0$, for which the error term $u_t$, is $I(0)$. It can be shown that OLS estimation of (7) yields consistent estimates of all of the parameters. The coefficients $\hat{\gamma}_t$ and $\hat{\delta}_t$ each individually converge at rate $\sqrt{T}$ to a Gaussian distribution, and the $t$ test of the hypothesis that $\gamma = 0$ is asymptotically $N(0,1)$, as is the $t$ test of the hypothesis that $\delta = 0$. However, an $F$ test of the joint null hypothesis that $\gamma$ and $\delta$ are both zero has a nonstandard limiting distribution. Hence, including lagged values in the regression is sufficient to solve many of the problems associated with spurious regressions, although tests of some hypotheses will still involve nonstandard distributions.

So if we fit regression on the equation (7) using OLS, $\hat{\alpha}$, $\hat{\phi}$, $\hat{\gamma}$ and $\hat{\delta}$ will be consistent estimator.

**Approach:2**

A second approach is to difference the data before estimating the relation in the form,

$$\Delta y_{1t} = \alpha + \gamma \Delta y_{2t} + u_t \tag{8}$$

Clearly, since the regressors and error term $u$, are all $I(0)$ for this regression under the null hypothesis, $\hat{\alpha}$ and $\hat{\gamma}$ both converge at rate $\sqrt{T}$ to Gaussian variables. Any $t$ or $F$ test based on (8) has the usual limiting Gaussian or $\chi^2$ distribution. So basically if we fit the regression using OLS, then $\hat{\alpha}$ and $\hat{\gamma}$ both will be consistent estimator.

Because the specification (8) avoids the spurious regression problem as well as the nonstandard distributions for certain hypotheses associated with the levels regression, many researchers recommend routinely differencing apparently nonstationary variables before estimating regressions.

**Approach:3**

Consider the regression

$$y_{1t} = \alpha + \gamma y_{2t} + u_t \tag{9}$$

A third approach, analyzed by Blough (1992), is to estimate (9) with Cochrane-Orcutt adjustment for first-order serial correlation of the residuals. Let $\hat{u}_t$ denotes the sample residual from OLS estimation, then the estimated autoregressive coefficient $\hat{\rho}_t$ from an OLS regression of $\hat{u}_t$, on $\hat{u}_{t-1}$ converges in probability to unity. Blough showed that the Cochrane-Orcutt GLS regression is then asymptotically equivalent to the differenced regression (8).

## 3 Simulation Study

We assume to have a simple linear regression model:

$$y_t = \alpha + \beta x_t + u_t$$

where $u_t$ is the error term, which is assumed to be $N(0, \sigma^2)$. If we apriori know, that both, $y_t$ and $x_t$ are independent and non-stationary, the estimated regression coefficient should be non-significant. These characteristics we have observed using a simple simulation methodology.

The basic idea is to generate time series data, which are known to be nonstationary and independent. For this purpose, we have used data generating equations. Here, we have analyzed two types of time series models. The Pure Random Walk(PRW)

$$y_t = y_{t-1} + u_t$$

and Random Walk with Drift $\mu$(RWD)

$$y_t = \mu + y_{t-1} + u_t$$

The PRW is a non-stationary process, because with increasing number of observations, the variance of $y_t$ increases, which is a well known fact. The Data Generating Process(DGP) is as follows: the error terms $u_t$ are generated from $N(0, \sigma^2)$, using R and initial values of $y_t$ are set to be zero, i.e, $y_0 = 0$ . For every spurious regression, we have calculated and recorded the following variables:

- Value of $\hat{\beta}$,
- t-statistic for $\hat{\beta}$,
- Durbin Watson(DW) statistics,
- Adjusted coefficient of determination $R^2$,
- Results of Phillips – Perron test for both, $y_t$ and $x_t$.

Together, we had 18 groups of different types of data, which were formed as follows. First, we used various types of regressions(TR):

- The type 1 - were the cases with $y_t$ and $x_t$. being $I(1)$processes.
- The type 2 - were the cases with $y_t$ being $I(1)$ processes and $x_t$ being $I(1)$ with drift processes.
- The type 3 - were both $y_t$ and $x_t$ are $I(1)$ with drift processes.

Secondly, we are also interested in the possible dependence of recorded variables upon the number of observations, we analysed samples with following sizes: $n = 50, 200, 1000$. We also replicated these simulations using time series with differences. By using level variables and differences, three types of sample sizes and three TR, the above mentioned 18 groups were formed. In every group, we performed 10000 regressions (replications).

Additionally, in the type 3 regressions, we fixed the drift value in $y_t$ and increased the drift value in $x_t$. The question we are trying to answer is, whether there is a systematical effect of increased drift on the recorded variables. Rather than answering this question analytically, we incorporated it into the design of type 3 regressions.

### Note

Errors of $y_t$ and $x_t$ are assumed to be coming from $N(0, 1)$ and $N(0, 2)$ respectively. In type 2, for the drift in $x_t$,we have selected $\mu = 2$. In type 3, for the drift in $y_t$,we have selected $\mu = 1$ and for the drifts in $x_t$,we have taken $\mu$ from 1 to 10000 increasing by 1, for each of the 10000 repeatitions.

## Results

The results are presented in the following next two tables. The first table reports the type I error of falsely rejecting the null hypothesis $H_0 : \hat{\beta} = 0$ . As can be seen, in all types of processes the error of rejecting the null hypothesis is high. For example, in the type 3 regressions, where both the dependent and independent variables were non-stationary with drift, we have rejected the null hypothesis in 100% from 10000 cases. Special attention should be addressed to the type 3 regressions, where independent variables had different drift parameters. Our results suggest that this had no effect on the results. The mean of Adjusted R$^2$has been rounded off to 4 decimal places.

Table 1: Results from simulations

| **DGP** | **Type 1** | | | **Type 2** | | | **Type 3** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sample** | n=50 | n=200 | n=1000 | n=50 | n=200 | n=1000 | n=50 | n=200 | n=1000 |
| | **Type I Error** | | | | | | | | |
| **Rejected** | 66.7% | 82.9% | 92.2% | 81.6% | 91.4% | 96.3% | 100% | 100% | 100% |
| **Rejected*** | 5.1% | 5% | 5.1% | 5.1% | 5% | 5.1% | 5.1% | 5% | 5.1% |
| **Adjusted R$^2$** | | | | | | | | | |
| **Mean** | 0.226 | 0.236 | 0.241 | 0.413 | 0.429 | 0.435 | 0.983 | 0.995 | 0.999 |
| **SD** | 0.23 | 0.23 | 0.23 | 0.302 | 0.301 | 0.301 | 0.012 | 0.0026 | 0.0005 |
| **Mean*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **SD*** | 0.029 | 0.007 | 0.001 | 0.029 | 0.007 | 0.001 | 0.029 | 0.007 | 0.001 |
| **DW Statistic** | | | | | | | | | |
| **Mean** | 0.334 | 0.089 | 0.018 | 0.381 | 0.101 | 0.02 | 0.381 | 0.101 | 0.021 |
| **SD** | 0.196 | 0.056 | 0.011 | 0.195 | 0.056 | 0.011 | 0.196 | 0.055 | 0.011 |
| **Mean*** | 1.998 | 1.997 | 2.001 | 2.002 | 1.997 | 2.0008 | 2.003 | 1.997 | 2 |
| **SD*** | 0.277 | 0.14 | 0.06 | 0.28 | 0.14 | 0.06 | 0.28 | 0.14 | 0.064 |

**Note:** symbol * denotes those results, where time series in differences was applied.

In contrast to the results of non-stationary data, the regressions with stationary data had a very low rejection rate at about 5% of all the time. A similar result may be observed looking at the adjusted $R^2$ .

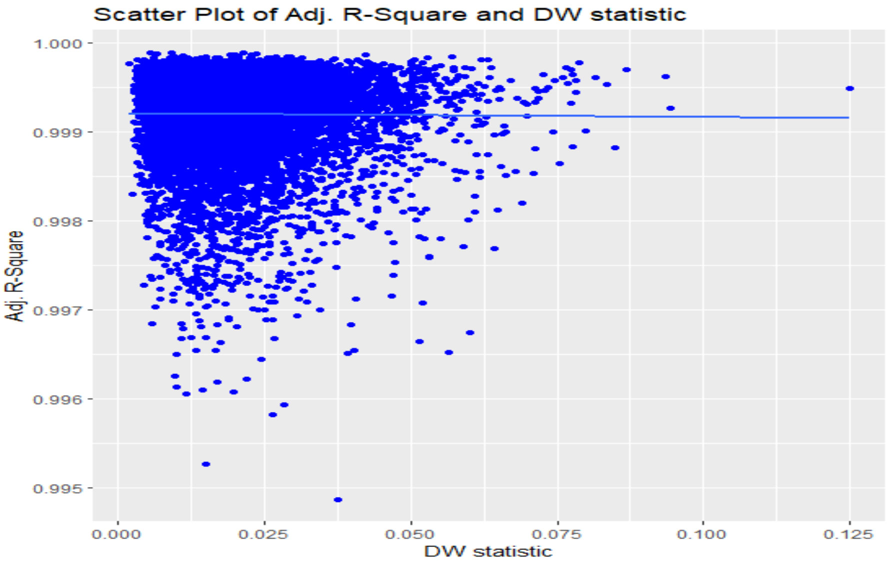
Table 2: Results from the PP test – rejection rate of $H_0$ in %

| DGP | **Type 1** | | | **Type 2** | | | **Type 3** | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample | n=50 | n=200 | n=1000 | n=50 | n=200 | n=1000 | n=50 | n=200 | n=1000 |
| $y$ | 5.4% | 6.5% | 5.8% | 5.4% | 6.4% | 5.8% | 5.4% | 6.5% | 5.8% |
| $x$ | 5.5% | 6.8% | 5.7% | 5.5% | 6.5% | 5.7% | 5.5% | 6.8% | 5.6% |
| $y^*$ | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| $x^*$ | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

**Note:** symbol * denotes those results, where time series in differences was applied.

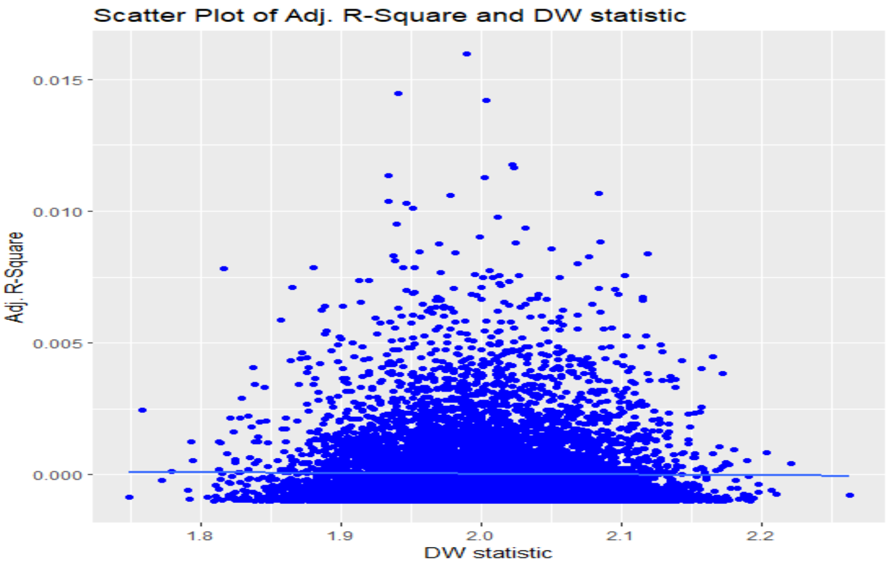
The second phenomenon of our interest was the increasing sample sizes. The observed results suggest that the effect is different with regard whether we regress stationary or non-stationary data. With the increase of sample sizes the rejection rate increases regardless of the TR used in the regression. There can be various statistical explanations for this effect. An intuitive non-statistical explanation may be that increasing the number of spurious observations increases the spuriousness of the dataset, thus making the phantom relationships more convincing. The more "bad" data are used, the more are we fooled. One last and interesting fact had been observed. A good "rule of thumb" of identifying incorrect regression results as a consequence of non-stationarity is a high coefficient of determination and a low Durbin – Watson statistic of autocorrelation. We were interested whether the "rule of thumb" was present also in our short study.

Figure 1: ScatterPlot of Adjusted $R^2$and DW Statistic for non-stationary case



In this above figure, we compared the ordered pairs of adjusted $R^2$ and Durbin-Watson statistics for the type 3 regressions with the sample size of 1000. As can be clearly seen, the "rule of thumb" holds. In this case, the spurious regression was present, and hence we observed much higher values of adjusted $R^2$ and much lower values of DW statistics (close to zero).

Figure 2: ScatterPlot of Adjusted $R^2$and DW Statistic for stationary case



In this above figure, we compared the ordered pairs of adjusted $R^2$ and Durbin-Watson statistics for the type 3 regressions with the sample size of 1000. In the case of non-spurious regression, we observed much lower values of adjusted $R^2$ and DW statistics were close to 2 where differenced time series was applied.

# 4 Real Data Analysis

### Data Description

In this study we have considered the daily closing prices of the 4 indices from CEE Markets namely Hungarian BUX, Polish WIG, Czech PX and Slovakian SAX. We have covered the time period from Jan 1, 2022 to Jan 1,2023. We compare several measures of this dataset with that of the corresponding logarithmic differences.

For a given time series $\{P_t\}$,t=1(1)T, the series of logarithmic differences is given as following –

$$LP_t = log(P_t) - log(P_{t-1}), t = 2(1)T$$

Data Source: `http://stooq.com`

## Analysis

In the figures below, we observe the 4 indices in levels and in logarithmic differences –

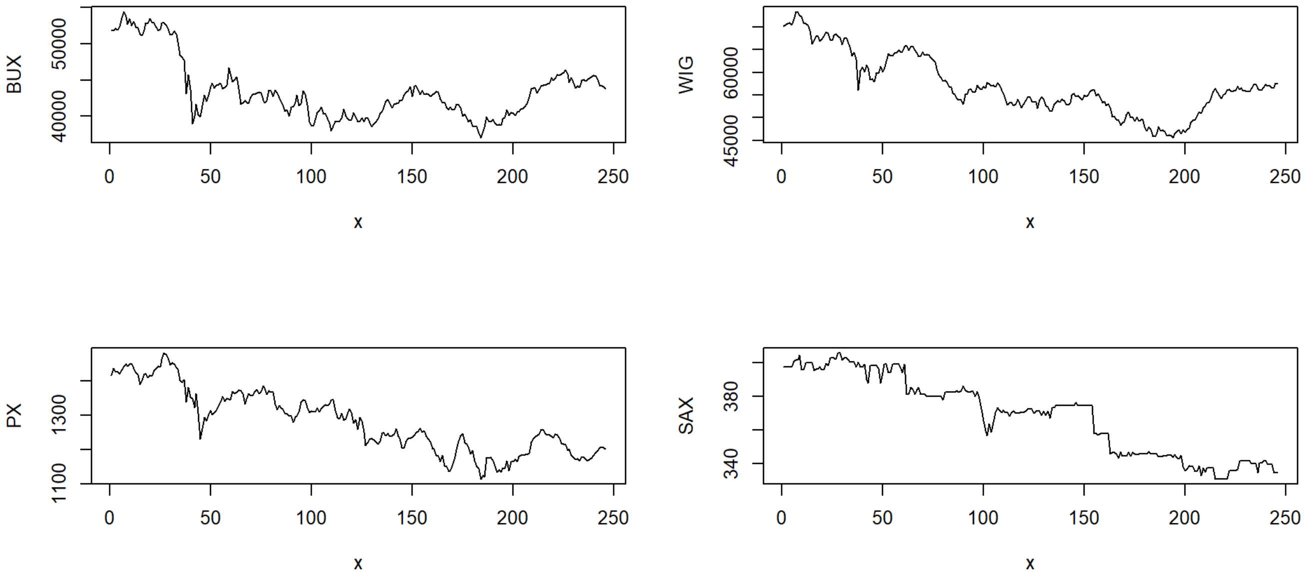


Figure 1: Visual Presentation of indices in level

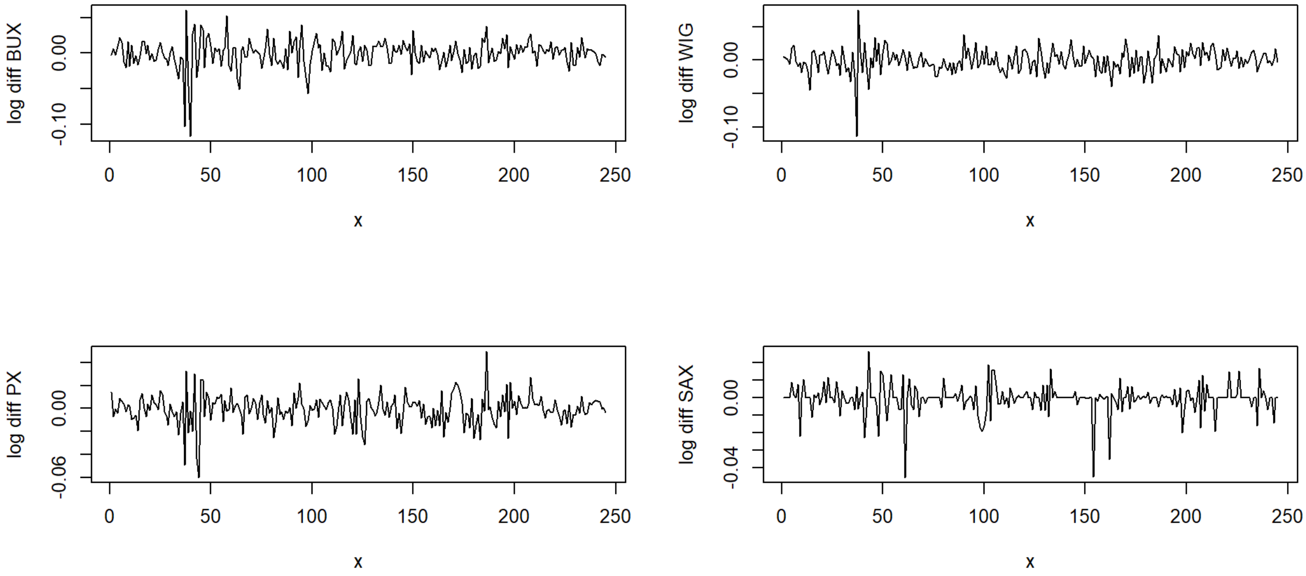


Figure 2: Visual Presentation of log-indices in level

We take origin of time period as 0 .

In the above plots we can see a decreasing trend in the indices at level thus indicating a non stationarity, whereas , in the series of log differences stationarity is apparent.

To get statistical evidence of stationarity or non-stationarity of the data, we conduct Augmented Dickey Fuller (ADF) Test and Phillips Perron (PP) Test and report the corresponding p-values in the following table.

| | ADF Test | PP Test |
|---|---|---|
| BUX | 0.48 | 0.71 |
| WIG | 0.64 | 0.87 |
| PX | 0.14 | 0.07 |
| SAX | 0.15 | 0.02 |
| log df BUX | 0.01 | 0.005 |
| log df WIG | 0.007 | 0.003 |
| log df PX | 0.000 | 0.000 |
| log df SAX | 0.001 | 0.000 |

In both the tests, the Null hypothesis is that, the series is nonstationary. From the above reported p-values, we can see all the indices at their level are accepted to be non stationary (1% level) , whereas at log differences they are stationary. Now, we will proceed to the Spurious Regression problem . To observe that, we perform Simple Linear Regression between each pair of indices at their level and at their log differences respectively. We use the method of OLS with HAC estimators to avoid the effect of autocorrelation on the significance of the estimators. We calculate the t statistic values, the R-square values and the Durbin-Watson test statistic values for each of these tests. The results are reported in the following table.

| DEPENDENT VARIABLE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| IN LEVELS | | | | | | IN LOG DF | | | |
| | | BUX | WIG | PX | SAX | BUX | WIG | PX | SAX |
| BUX | t-stat | . | 39.49 | 31.30 | 25.19 | . | 2.92 | 3.19 | -1.27 |
| | R-sq | . | 0.99 | 0.99 | 0.99 | . | 0.16 | 0.10 | 0.01 |
| | DW | . | 0.09 | 0.07 | 0.06 | . | 2.02 | 2.08 | 2.26 |
| WIG | t-stat | 38.90 | . | 45.89 | 30.45 | 3.37 | . | 5.57 | -1.36 |
| | R-sq | 0.99 | . | 0.99 | 0.99 | 0.16 | . | 0.16 | 0.01 |
| | DW | 0.09 | . | 0.07 | 0.06 | 1.93 | . | 2.14 | 2.27 |
| PX | t-stat | 32.97 | 46.65 | . | 88.86 | 2.64 | 3.10 | . | -0.45 |
| | R-sq | 0.99 | 0.99 | . | 0.99 | 0.10 | 0.16 | . | 0.00 |
| | DW | 0.07 | 0.07 | . | 0.16 | 2.02 | 2.16 | . | 2.30 |
| SAX | t-stat | 28.38 | 31.38 | 90.85 | . | -1.09 | -1.18 | -0.41 | . |
| | R-sq | 0.99 | 0.99 | 0.99 | . | 0.01 | 0.01 | 0.00 | . |
| | DW | 0.06 | 0.06 | 0.16 | . | 1.96 | 2.06 | 2.07 | . |

In the above table we can see for in level tests, test statistics are significant in all the cases with high value of R-Square and very low value of Durbin Watson Test Statistic. This implies presence of high autocorrelation in the error terms. Since we applied HAC covariance matrix, it has asymptotically no effect on the reported significance of regression coefficients.

We know that the series are non stationary in levels which already make the results misleading.This spurious significance can be interpreted as the measure of trend of both indices, not the relationship between closing prices. A rule of thumb to identify the spurious regression problem is high R-square value and low DW Statistic which applies in this case also.

From the inefficiency of Slovakian Stock Market, one would expect very weak relationship between SAX and any other index. This is evident when the time series are observed at logarithmic differences. Here the coefficient is not significant and $R^2$ is close to 0. The significance of test statistic is retained for other pairs of indices.

## 5 Conclusion

In the beginning of this document, we have described how spurious regression can arise and the remedies of such problem. Both from the results of simulation study and the data study, we could see how spurious regression can induce wrong impression of relationship between two time series especially when they are non-stationary. The thumb rule of spurious regression has been again verified as high coefficient of determination and low Durbin-Watson statistic.

## 6 References

Baumohl, E., & Lyocsa, S. (2009). Stationarity of Time Series and the Problem of Spurious Regression. SSRN Electronic Journal. doi:10.2139/ssrn.1480682

Giles, E. A. D. (2007). Spurious Regressions with Time-Series Data: Further Asymptotic Result. University of Victoria, B.C. Canada

Granger, C. W. J., Newbold, P. (1974). Spurious regressions in econometrics. Journal of Econometrics 2: 111- 120.

Hamilton, J. B. (1994). Time Series Analysis. Princeton NJ; Princeton University Press.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics, Journal of Econometrics 33: 311-340.