Spurious Regression: Simulation and Theoretical Analysis

Yihui XIE †

School of Statistics, Renmin University of China

Abstract

This paper discusses a classical and instructional problem in linear regression, i.e. spurious regression with respect to integrated processes. The term "spurious regression" was first brought forward by Granger and Newbold (1974), pointing out the high but meaningless R^2 in a regression involving time series; Phillips (1986) provided a further analytical study of their empirical result.

Keywords: Spurious regression; Random walk; Monte Carlo

1. Introduction

It's universally known that integrated time series exists widely in financial and economic data, which can be described as I(1) process. Surely it would be worth our attention examining the characteristics of integrated variables, especially whether the traditional conclusions in regression analysis still hold. The influential paper by Granger and Newbold (1974) has shown some undesirable results through simulation by the regression on two random walk process, which gave a warning to researchers who applied traditional regression on non-stationary time series; but this paper mentioned little about the theoretical analysis. Phillips (1986) made further research based on computer simulation, and put forward an excellent analytical study (asymptotic theory) for the essential of "spurious regression". This paper intends to give a summarization of both the simulation and theory for the phenomenon of "spurious regression".

2. Regression and I(1) Process

In this section we briefly review some basic knowledge about regression analysis and time series process.

2.1. Univariate Linear Regression

For convenience, here we only consider the case of univariate exploratory variable. The traditional regression equation can be expressed as:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1, 2, \cdots, T \tag{1}$$

For the disturbances ε_t, Gauss-Markov condition must be satisfied, and when it comes to the statistical reference (about the coefficients as well as other statistics), there's also a condition on the distribution of ε_t:

$$\begin{cases} E(\varepsilon_t) = 0, \quad t = 1, 2, \cdots, T \\ cov(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \quad (i, j = 1, 2, \cdots, T) \end{cases} \tag{2}$$

$$\varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2), \quad t = 1, 2, \cdots, T \tag{3}$$

Through OLS or MLE, we can easily get the estimator for the coefficients α and β, and here we just focus on β, because usually the intercept is less important than the slope. As we've already known, $\hat{\beta} = \sum_{t=1}^{T}(x_t - \bar{x})y_t / \sum_{t=1}^{T}(x_t - \bar{x})^2$, and $E(\hat{\beta}) = \beta$, $var(\hat{\beta}) = \sigma^2 / \sum_{t=1}^{T}(x_t - \bar{x})^2$. Here we should pay special attention to the variance of $\hat{\beta}$: it shrinks to zero as $T \to \infty$, which indicates a degenerate distribution of $\hat{\beta}$. Later we'll discuss the limiting distribution of $\hat{\beta}$ in the case when y_t and x_t are two random walk process.

2.2. Integrated Process

As mentioned in Section 1, integrated time series, which are non-stationary, will produce a false regression result. Again, we only consider the order one integrated process (i.e. I(1),

† Yihui Xie is postgraduate, School of Statistics, Renmin University of China, 100872 (Email: xieyihui@gmail.com)

also known as "random walk") for convenience. This process can be expressed in the following form:

$$y_t = y_{t-1} + u_t, \quad x_t = x_{t-1} + v_t, \quad t = 1, 2, \cdots \tag{4}$$

where u_t and v_t are independent white noises.

Actually this process can be quite easily simulated by means of a computer. For example, if set $y_0 = x_0 = 0$, we'll get $y_t = \sum_{t=1}^{T} u_t$ and $x_t = \sum_{t=1}^{T} v_t$, then we just have to generate independent random numbers following normal distribution. Figure 1 shows an I(1) process as an example.

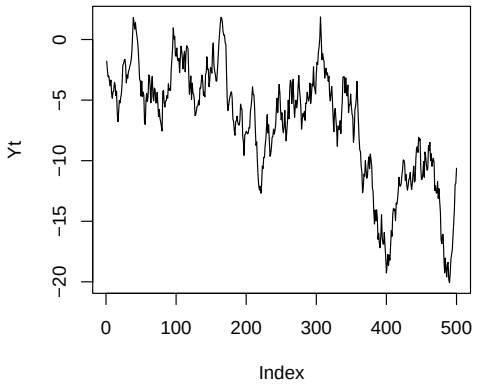

Figure 1: *A simulation of random walk process*

3. Test for Regression Coefficients

It's generally known that regression coefficients are tested using t-statistic, so here we review the theory of construction of t-statistic before carrying on a simulation with respect to the coefficients using random numbers generated from I(1) process.

3.1. t-statistic in Linear Regression

The classical theory of regression analysis constructs a t-statistic to test the regression coefficients under the null hypothesis that these (or one of) coefficients equals to zero. For the coefficient β of equation (1) in section 2.1, we know its estimator $\hat{\beta}$ follows a normal distribution under the assumption (3):

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}) \tag{5}$$

Therefore under the null hypothesis "$H_0 : \beta = 0$", the t-statistic can be constructed as

$$t = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2 / \sum_{t=1}^{T}(x_t - \bar{x})^2}} \tag{6}$$

where $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$. Under the null hypothesis $\beta = 0$, t follows t distribution with $n-2$ degrees of freedom.

3.2. A Simulation Indicating Spurious Regression

By common sense we know regress a random walk process on the other is usually meaningless, no matter how "significant" the coefficients are. However, from simulation we can get an amazing result, i.e. the rejection rate of null hypothesis will reach up to approximately 76%! From a traditional view, this stands for a significant relationship between independent and dependent variables, which is certainly "spurious" from a realistic view.

Next we generate two independent I(1) processes (both of their lengthes are 100) and do regression analysis on them. Through 1000 times of Monte Carlo simulation (see appendix A for the program code), we find a rejection rate of 76.3%, which agrees with the result by Granger and Newbold (1974).

Sure, merely a rejection rate doesn't provide sufficient information for us to research the characteristics of the regression using I(1) processes, especially the large sample features. Below we make further examination on four statistical values: 1) $\hat{\beta}$; 2) t-statistic; 3) R^2; 4) DW-statistic. Please note that from now on we change the length of I(1) process to be 1000 (a number large enough) instead of 100 in order to explore the asymptotic characteristics.

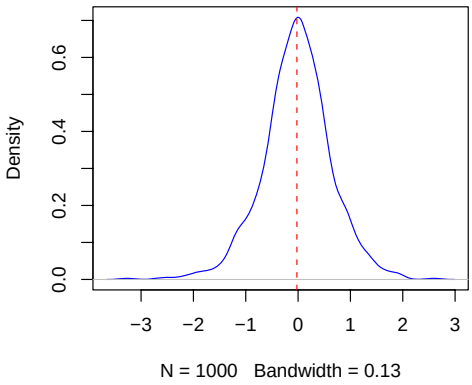

Figure 2: *Empirical distribution of* $\hat{\beta}$

Figure 2 shows that the distribution of $\hat{\beta}$ is approximately normal, but by further calculation we find that the mean is near to 0 (-0.022), which is reasonable, but the standard deviation is 0.66 (unreasonably large). In section 2.1 we know, as $T \to \infty$ (in the simulation $T = 1000$), the standard deviation should decrease to zero.

Figure 3 presents the distribution of the so-called "t-statistic". From traditional regression theory we know the confidence interval when $\alpha = 0.05$ will be $[-qt(0.975, 998), qt(0.975, 998)] = [-1.96, 1.96]$, where the function qt stands for quantile function for t-distribution. Obviously the rejection rate will be extremely high because the confidence interval is too "narrow".

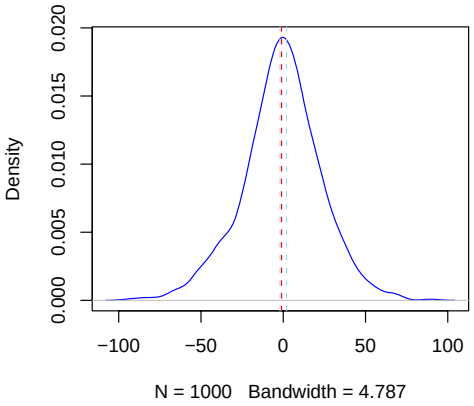

Figure 3: *Empirical distribution of t-statistic*

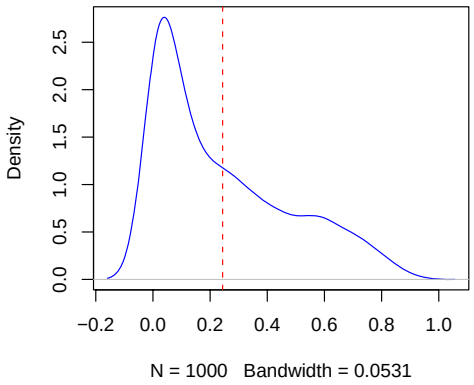

Figure 4: *Empirical distribution of* R^2

From Figure 4 we know the average R^2 is about 0.24, and the maximum R^2 is near to 1; Figure 5 tells us that the value of *DW*-statistic is quite small, indicating a strong serial autocorrelation.

This simulation give an intuitive proof for spurious regression. Now we know sometimes "statistical significance" just means nothing at all. Naturally, the next question will be: how does spurious regression arise?

4. Theoretical Analysis on Spurious Regression

Phillips (1986) developed precise theories for integrated processes (not only order one!); here we just make a simple introduction to part of his theories. Actually his assumptions are much weaker than what we required in equation 4. For a sequence $\{\xi_t\}_1^\infty$ of random n-vectors, let $S_t = \sum_{j=1}^t \xi_j$ be the partial sum process and set $S_0 = 0$. The assumptions are:

1. $\mathrm{E}(\xi_t) = 0$ for all t;
2. $\sup_{i,t} \mathrm{E}|\xi_{it}|^{\beta+\varepsilon}$ for some $\beta > 2$ and $\varepsilon > 0$;
3. $\Sigma = \lim_{T\to\infty} T^{-1} E(S_T S_T')$ exists and is positive definite;

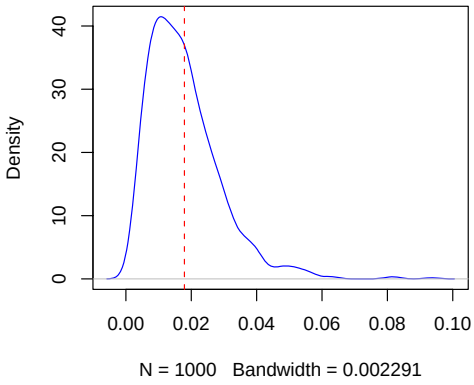

Figure 5: *Empirical distribution of DW-statistic*

4. $\{\xi_t\}_1^\infty$ is strong mixing;

In our example, $\{\xi_t\} = \{(u_t, v_t)\}$. Part of Phillips' results are listed below:

$$\hat{\beta}_T \Rightarrow \sigma_u \sigma_v^{-1} \frac{\varsigma_{UV}}{\varsigma_{VV}} \tag{7a}$$

$$T^{-1/2} t_{\hat{\beta}_T} \Rightarrow \frac{\varsigma_{UV}}{(\varsigma_{UU}\varsigma_{VV} - \varsigma_{UV})^{1/2}} \tag{7b}$$

$$R^2 \Rightarrow \frac{\varsigma_{UV}^2}{\varsigma_{UU}\varsigma_{VV}} \tag{7c}$$

$$dw \xrightarrow{p} 0 \tag{7d}$$

where $\varsigma_{ab} = \int_0^1 a(r)b(r)\mathrm{d}r - \int_0^1 a(r)\mathrm{d}r \int_0^1 b(r)\mathrm{d}r$; $U(r)$ and $V(r)$ are independent Brownian motions (Wiener processes).

From equation 7a, we know as $T \to \infty$, $\hat{\beta}_T$ does NOT converge in probability to a constant, and it has a *non-degenerate* limiting distribution. Equation 7b shows that the traditional t-statistic no longer follows t distribution and in fact it has no limiting distribution, just diverging as $T \to \infty$, which indicating an increasing rejection rate as the sample size increases. Equation 7c and 7d shows respectively that R^2 has a non-degenerate distribution and *DW*-statistic converges to 0 in probability.

The proof of the above results involves with Functional Central Limit Theorem as well as some knowledge of stochastic process (especially stochastic integration). Further details can be found in the appendices of Phillips' paper.

5. Conclusion

Note that we only discuss the univariate case in the paper, however, the results can be generalized to multivariate occasion. An empirical study through simulation will tell us that

as the sample size or the dimension increases, the rejection rate will be higher and higher. The instruction we get from this phenomenon is that statistical analysis is, after all, a tool to help us explore relationships between variables, but it's not the last resort. More practical experiences are required when analyzing with statistics.

Appendix A: R Code for Regression on Two I(1) Processes

Here "tms" record the times of $\hat{\beta}$ being significant according to the significance level $1-\alpha=0.95$; the meaning of the rest variables are obvious, e.g. "Rsquare" means R^2.

```
tms=0;betahat=NULL;tstat=NULL;
Rsquare=NULL;dwstat=NULL
for (i in 1:1000){
  # I(1) process with length of 100
  y=cumsum(rnorm(1000))
  x=cumsum(rnorm(1000))
  # Regression: y=alpha+beta*x+epsilon
  res=lm(y~x)
  summ=summary(res)
  tms=tms+(summ$coefficients)[2,4]<0.05)
  betahat=c(betahat,(summ$coefficients)[2,3])
  Rsquare=c(Rsquare,summ$r.squared)
  dwstat=c(dwstat,dwtest(res)$statistic)
}
# Times of beta being significant
cat('The rejection rate is ',tms/10,'%\n')
```

References

[1] C. W. J. Granger and P. Newbold, Spurious regressions in econometrics, Journal of Econometrics, Volume 2, Issue 2, , July 1974, Pages 111-120.

[2] P.C.B. Phillips, Understanding spurious regressions in econometrics, Journal of Econometrics, Volume 33, Issue 3, December 1986, Pages 311-340.

[3] Sheldon M. Ross, 1983, Stochastic Processes, John Wiley & Sons, Inc