Simulation Assignment to udnerstand stochastic trends and spurious regressions

Table of contents

##Regressions with trending variables

0.1 Introduction

In the time series topic of the course, we mainly assumed that all variables were stationary, that is, that their distribution did not change over time. Regressions using stationary variables are "nice" in that the estimates usually are consistent and asymptotically normal.

A common way in which stationarity can be violated is when the variables involved in a regression have trends. When we run a regression with trending variables, we can obtain misleading results. This situation is called "spurious regression", a term that you already have encountered in the context of cross-sectional regression: a large/significant regression coefficient which does not measure a causal effect due to omitted variables. In time series, a "spurious regression" is one in which a large/significant regression coefficient may be a type-I error. That is, there is no correlation in the population, but the sample statistic indicates there is. In this assignment, you will code, run and interpret simulations in R to explore this type of spurious regression.

The assignment is designed so that the information in the questions and your progressive answers are sufficient to learn and understand the concept, causes, and some solutions for spurious regressions in time series. However, if you want additional information, you can also use the following sections from the textbooks as background readings: Wooldridge 18.3 ("Spurious regression"), and Stock & Watson 15.7 ("Nonstationarity I: Trends"; in particular, subsection "Spurious regression").

Consider a simple regression of the time series Y_t on the time series X_t:

$$Y_t = \beta_0 + \beta_1 Xt + U_t, t = 1, 2, ...n \tag{1}$$

where U_t is a white noise error term. The variables Y_t and X_t are generated by the following Throughout the assignment, Y_t and X_t are not related to each other, so that $\beta_1 = 0$. The interest lies in the estimated β_1 and the result of testing $H_0 : \beta_1 = 0$. We can view Equation 1 as a special

case of an ARDL model in which, for simplicity, all coefficients on lagged variables are zero. To make things even simpler, we will also let $\beta_0 = 0$[1]

1 1. Coding a simulation in R

The simulations are similar to the ones you have worked with in the tutorials. You can use those simulations as a guide on how to set up and structure your simulation.

– In particular, you can write your simulations around two nested for() loops: an outer loop that loops over the values of the sample size n; and an inner loop looping over the repeated samples (replications), in each of which you draw n observations of all the variables, run the regressions, and save the estimates or statistics in previously created storage matrices.

– Fix the random seed to '2023' at the beginning of your R script by writing rseed(2023) so that your results are replicable.

– Try not to write the two for() loops in one go. Make sure that the code works for a single sample, then slowly generalise it to the inner for() loop, then finally add the outer loop.– When testing for() loops, use a small number of replications so that the code runs faster. Just remember to change it back when you are sure it works.

– The assignment has many sub-questions, but you don't need to code a new simulation for every sub-question. For instance, you can code one simulation for every DGP, of which there are three (DGP1 = Question 1, DGP2= Question 2, DGP3= Questions 3 & 4) and obtain everything asked for from the inner loop of the respective simulation.

2 Assignment questions

2.1 1. No spurious regression

Consider a baseline scenario where there is no spurious regression problem. Let $Y_t \sim N(01)$ and $X_t \sim N(01)$. Simulate 2,500 repeated samples or replications from this DGP for each of the three sample sizes $n = 50, 100, 200$. In each replication, regress Y_t on X_t, and save $\hat{\beta}_1$ and the p-value of the significance test of $H_0 : \beta_1 = 0$ against the two-sided alternative $H_1 : \beta_1 \neq 0$.

(a) Show a figure with the estimated density of 1 for each of the three sample sizes (all three estimated densities in one single graph). Use the command density() in R to obtain a density estimate which can be graphed via plot() and/or lines().

(b) Report the estimated rejection frequencies for the test performed at the 5% significance level (to three decimal places) for each of the three sample sizes.

(c) Show a time series plot with both Y_t and X_t (that is, a plot with time t on the x-axis) for the last simulated sample/replication with sample size $n = 200$.

(d) Briefly describe your findings (a)-(c). Do the findings conform to your expectations? Explain.

[1]While this is a very simple setup, your findings in this assignment apply broadly to general time series models, such as the ARDL(p,q) and VAR(p) models discussed in the lecture. .

2.2 2. Spurious regression due to deterministic trends

We say that a variable has a linear deterministic trend when its expectation changes by a constant amount from period to period: $E(Y_t - Y_{t-1}) = c$, or equivalently $E(Y_t) = ct$. Clearly, such a variable is not stationary as its mean changes over time. Modify the DGP from Question 1 so that the two time series variables have linear deterministic trends: $Y_t \sim N(0.05t - 1)$ and $X_t \sim N(0.03t - 1)$. Run a new simulation for $n = 50, 100, 200$ of 2,500 replications each.

(a) Show the same figures and report the same statistics as described in Question 1(a)-(c) but for this DGP.

(b) Explain the problems that happen in this setup with reference to your findings in Question 2(a).

(c) One way of addressing the problems with this regression is to remove the deterministic trends from the variables in order to render them stationary before using them in a regression analysis. This is called de-trending and consists of replacing the variable with its residuals from a regression of the variable on a linear trend. For instance, to de-trend Y_t, we estimate the model $Y_t = \beta_0 + \beta_1 t + U_t$ by OLS, and then construct the de-trended variable $Y_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 t (= U_t)$. Redo Question 2(a) but de-trend Y_t and X_t before performing the regressions.

(d) Describe your findings in Question 2(c). Does the strategy of de-trending adequately address the problem of spurious regression caused by deterministic trends? Explain.

2.3 3. Spurious regression due to stochastic trends

We say that a variable has a stochastic trend when the change in its expectation from period to period is random: $E(Y_t - Y_{t-1}) = t$. Thus, an AR(1) time series with $\rho = 1$ has a stochastic trend. Such an AR(1) model is also called a "unit-root process" and a variable that follows it is said to "have a unit root". It is also called a "random walk" since the variable changes over time by taking a 'step' from its previous realisation in a random direction. This often results in prolonged periods of increases or decreases that change eventually.

Modify the DGP so that both variables have stochastic trends. Specifically, let both time series be AR(1) with $\rho = 1$, $Y_t = Y_{t-1} + t$ and $X_t = X_{t-1} + t$. One can show that unit root processes can be represented as the sum of all past errors; formally, $Y_t = Y_1 + \sum_{j=2}^{t} j$. Assume that $Y_0 = 0$ and $X_0 = 0$. Then, to implement this in R, draw Y_t and X_t as

y <- cumsum(rnorm(n)) x <- cumsum(rnorm(n))

Run a new simulation of this DGP for $n = 50, 100, 200$ of 2,500 replications each.

(a) Show the same figures and report the same statistics as described in Question 1(a)-(c) but for this DGP.

(b) Explain the problems that happen in this setup with reference to your findings in Question 3(a).

(c) One way of addressing the problems with this regression is to replace the variables (which contain unit roots) with their first-differences (which don't, since $Y_t = Y_t - Y_{t-1} = t \sim N(0,1)$ and the same holds for X_t). Redo Question 3(a) but run the regressions in first differences; that is, regress Y_t on X_t.

(d) Describe your findings in Question 3(c). Does the strategy of using first differences adequately address the problem of spurious regression caused by stochastic trends? Explain.

2.4 4. Testing for stochastic trends

In practice, of course, we don't know if the variables we want to use in a time series regression contain unit roots. In this question, you will learn how to test for a unit root in a variable, Y_t. Throughout this question, let $Y_t = \delta + Y_{t-1} + t$, with $\delta = 0$, $\rho = 1$ and $t \sim N(0,1)$. Run three simulations of 2,500 replications each where you regress Y_t on Y_{t-1} using samples of size $n = 50, 100, 200$ respectively.

(a) Present a figure with estimates of the density of $\hat{\delta}$ for each sample size.

(b) Discuss the figure. Does it show evidence of being consistent? Asymptotically normal? Explain. If $\hat{\delta}$ is not asymptotically normal, what are the implications for testing $H_0 : \delta = 1$? Explain.

(c) To test $H_0 : \delta = 1$, the AR model $Y_t = \delta + Y_{t-1} + t$ is usually transformed by subtracting Y_{t-1} from both sides of the equation. Writing the transformed model in regression form gives $Y_t = \delta + Y_{t-1} + t$.

Write down the null hypothesis of $H_0 : \delta = 1$ and the alternative hypothesis of $H_1 : \delta < 1$ in terms of δ. (The case $\delta > 1$ is not considered because such a process would be explosive and is unlikely to be observed in practice.) Suppose you run the regression and use a t-statistic of t to test H_0 against H_1. If the t-statistic were normally distributed, what would be the critical value for a test at the 5% significance level?

(d) Run a simulation of the regression from Question 3(c) for $n = 50, 100, 200$ of 2,500 replications each. In each regression, construct and save the t-statistic for $\hat{\delta}$. Present the rejection frequencies of the corresponding one-sided test of H_0 against H_1 at the 5% significance level using the critical value from a normal distribution. Comment on your results.

(e) The test on from regression (3) you described in Question 3(c) is called a Dickey Fuller test, after Dickey and Fuller (1979, JASA), who derived the distribution of the t statistic for when $\delta = 1$. This distribution is not normal even in large samples. The 5% critical value of this distribution in large samples is –2.86. Redo the tests from Question 3(d), this time using the correct Dickey-Fuller critical value instead of the normal critical value used in 3(d). Present the rejection frequencies and comment.

3 Submission guidelines

Before submitting, check that your assignment is complete. An assignment should include a single file containing

(1) the text answering the assignment questions (possibly including mathematical notation), (2) figures, and (3) your R code used to produce all your results. The requirements on your R code are : your R script needs to replicate all your simulation results and figures without the person who replicates it needing to manually change anything. Figures should be numbered and have an appropriate, descriptive title. In general, a figure should be self-explanatory, not contain unnecessary information, and be simply formatted. Self-explanatory means that the figure can be read without referring back to the description in the text. Further information necessary for understanding the figure should be provided as notes below the table. Good examples of figures can be found in the journal American Economic Review