Lecture 12. Regression and Spurious Regression

Muyi Li

WISE&SOE, Xiamen University

2020 Spring

- The Autoregressive (AR) model

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + e_t \tag{1}$$

- The regression model:

$$y_t = \alpha + \delta x_t + e_t \tag{2}$$

- The Distributed Lag model

$$y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{3}$$

- The Autoregressive Distributed Lag (ADL) model:

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{4}$$

Core insights

- The coefficients can be estimated by OLS just like a conventional regression.
- The appropriate standard error depends on whether or not the dynamics have been explicitly modeled. In AR and ADL models, the robust standard errors are appropriate as long as the number of lags p is sufficiently large so that the errors are serially uncorrelated.
- In the regression and DL models, the equation error e_t will be serially correlated, so we should use HAC (heteroskedasticity-and-autocorrelation) standard errors.

Core insights: Cont'd

- The number of lags (p and q) inAR and ADL models may be selected by comparing models using the Akaike Information Criterion (AIC). Test statistics (t- and F-statistics) should not be used for model selection.
- In the regression and DL models, we should be greatly concerned about the possibility of a spurious regression: the setting where a regression with two unrelated time series has misleadingly large conventional t-statistics and R^2 values. Understanding the potential for spurious regression and avoid it.

Core insights: Cont'd

- The parameters of time series models are likely to have changed over time. This requires care and attention.
- The concept of Granger causality can be explained within this model, namely that x_t does not Granger-cause y_t if $\delta_1 = \cdots = \delta_q = 0$.
- The ADL model without the contemporaneous regressor x_t may be used to produce one-step ahead point forecasts for y_{n+1}. Point forecasts should be combined with interval forecasts, as the latter convey the degree of uncertainty about future outcomes.
- To produce multi-step (h step) forecasts using a ADL model, we can use the multi-step version

$$y_{t+h} = \alpha + \phi_1 y_t + \cdots + \phi_p y_{t-p+1} + \delta_1 x_t + \cdots + \delta_q x_{t-q+1} + e_t$$

which can also be estimated by OLS.

Standard errors and $t-$statistics

- A critical issue in time series (and econometrics in general) is which standard error to use. There are three popular standard error formulae for applied econometric time series:
  - classical;
  - heteroskedasticity-robust;
  - HAC.
- So-called classical (or homoskedastic) standard errors can not be used in applied econometric practice if there exists serial correlation and heteroscedasticity.
- HAC standard errors (Newey-West) are appropriate for simple (nondynamic) regressions and DL models. They should be used whenever the serial correlation in the error has not been modeled.
- Robust standard errors are the most commonly used in current applied practice. They are appropriate for time-series models that are dynamically well-modeled, including autoregressive and ADL models.

Standard errors and $t-$statistics: Illustrative Examples

- Let gas_t denote the weekly percentage change in U.S. retail gasoline prices, and let oil_t denote the weekly percentage change in the Brendt European spot price for crude oil, for 1991C2016. The estimates are:

$$\begin{aligned} gas_t = \quad & 0.029 && +0.269 oil_t + \hat{e}_t \\ & (0.046) && +(0.011) \\ & (0.046) && +(0.015) \\ & (0.073) && +(0.021) \end{aligned} \tag{5}$$

- The first set of standard errors are classical (homoskedastic), the second are heteroskedasticity-robust, and the third are Newey-West estimates with 12 lags.We can see that the choice of standard error formula matters greatly, with the Newey-West roughly twice themagnitude of the classical.
- Because this is a static regression, the Newey-West are the appropriate choice.

Illustrative example

- Let $return_t$ denote the weekly percentage change in the S&P 500 index for 1950C2016. (For simplicity, we ignore dividends. Also, while daily observations are available, they have extra complications so it is easier to focus on weekly observations.) The estimates are:

$$\begin{aligned} return_t = \quad & 0.16 \quad && -0.032 return_{t-1} + 0.037 return_{t-2} + \hat{e}_t \\ & (0.04) && (0.029) \qquad\qquad (0.025) \end{aligned}$$

- Since the model is dynamic, we we report heteroskedasticity-robust standard errors.
- A simple form of the efficient market hypothesis suggests that stock returns are unpredictable, and thus the AR coefficients should be zero. Thus, the F-statistic for the AR coefficients is a simple test of efficient markets. In this example, the p value for the F-statistic is .12, so we fail to reject the efficient market hypothesis.

Illustrative example

- To repeat our message about the importance of using robust standard errors, if instead we had used the old-fashioned (homoskedastic) formula, the standard errors on the AR coefficients would be 0.17, the second lag would have a t-statistic of 2.2, and the F-statistic for the two AR coefficients would have a p value of .01, which would incorrectly suggest rejection of the efficient market hypothesis. Indeed, using the correct standard error formula makes a huge difference and alters inference.

Illustrative example: Distributed lag models

- Distributed lag models are useful when we want to estimate the impact of one variable upon another. As an example, consider the effect of crude oil prices upon retail gasoline prices, using the data from the earlier section on standard errors and t-statistics. A distributed lag model with a contemporaneous effect and six lags takes the form:

$$\begin{aligned} gas_t = & \underset{(0.057)}{-0.009} + \underset{(0.016)}{0.243 oil_t} + \underset{(0.012)}{0.112 oil_{t-1}} + \underset{(0.011)}{0.063 oil_{t-2}} \\ & + \underset{(0.013)}{0.064 oil_{t-3}} + \underset{(0.010)}{0.030 oil_{t-4}} + \underset{(0.011)}{0.032 oil_{t-5}} + \underset{(0.012)}{0.018 oil_{t-6}} + \hat{e}_t \end{aligned}$$

- Here, the standard errors are computed using the Newey-West formula with 12 lags.

Illustrative example: Distributed lag models

- Under the assumption of strict exogeneity, the coefficients of a DL model are the effects of the regressor on the dependent variable. In this case, we see that a 1 percent change in crude oil prices leads to a contemporaneous (within one week) change in retail gasoline prices of 0.24 percent, followed by further increases over the following six weeks.
- The following equivalent regression can be used to estimate the cumulative multipliers:

$$
\begin{aligned}
gas_t = \quad & -0.009 && +0.243\Delta oil_t + 0.355\Delta oil_{t-1} + 0.418\Delta oil_{t-2} \\
& (0.057) && (0.016) \quad (0.024) \quad (0.028) \\
+ \quad & 0.482\Delta oil_{t-3} && +0.512\Delta oil_{t-4} + 0.544\Delta oil_{t-5} + 0.562\Delta oil_{t-6} + \hat{e}_t \\
& (0.037) && (0.040) \quad (0.045) \quad (0.047)
\end{aligned}
$$

Illustrative example: Distributed lag models

- These results show that a 1 percent change in crude oil prices lead to a 0.56 percent cumulative change in retail gasoline prices after six weeks, with most of the change incorporated within the first four weeks. These estimates show how quickly changes in crude oil prices translate into retail prices.

Spurious Regression

- The traditional statistical theory holds when we run regression using (weakly or covariance) stationary variables.
- For example, when we regress one stationary series onto another stationary series, the coefficient will be close to zero and insignificant if the two series are independent.
- That is NOT the case when the two series are two independent random walks, which are nonstationary.

Spurious Regression

- The regression is spurious when we regress one random walk onto another independent random walk. It is spurious because the regression will most likely indicate a non-existing relationship:
  (1) The coefficient estimate will not converge toward zero (the true value). Instead, in the limit the coefficient estimate will follow a non-degenerate distribution
  (2) The t value most often is significant.
  (3) R^2 is typically very high.

Spurious Regression

Simulation
by construction y and x are two independent random walks
n=1000
y = rep(0,n)
x = rep(0,n)
ey = rnorm(n)
ex = rnorm(n)
rhoy = 1
rhox = 1
for (i in 2:n) {
y[i] = rhoy*y[i-1] + ey[i]
x[i] = rhox*x[i-1] + ex[i]
lm(formula = y$\sim$ x)
}

Time plot of two independent random walk processes

Lessons from Spurious Regression

Lesson 1: always check the stationarity of the residual. The regression is spurious if the residual is nonstationary (cannot reject the null hypothesis of the unit root test).
Lesson 2: just because two series move together does not mean they are related!
Lesson 3: use extra caution when you run regression using nonstationary variables; be aware of the possibility of spurious regression! Check whether the residual is nonstationary.

Spurious correlation

- Consider the regression model:

$$y_t = \beta_0 + \beta_1 x_{t-d} + e_t$$

where $x's$ are i.i.d. random variables with variance σ_x^2 and the $e's$ are WN(0,σ_e^2) and are independent of the X.

- Recall the cross-correlation function (CCF) $\rho_k(x, y)$ is zero except for lag $k = -d$, where

$$\rho_{-d}(x, y) = \frac{\beta_1 \sigma_x}{\sqrt{\beta_1^2 \sigma_x^2 + \sigma_e^2}}$$

- In this case, the theoretical CCF is nonzero at lag $-d$, reflecting the fact that x is leading y by d units of time.
- The CCF can be estimated by the sample CCF

$$r_k x, y = \frac{\sum(x_t - \bar{x})(y_{t-k} - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}$$

Spurious correlation

The covariate X is independent of Y *if and only if* $\beta_1 = 0$, in which case the SCCF $r_{X,Y}(k)$ is approximately $\mathcal{N}(0, \frac{1}{n})$, where n is the sample size — the number of pairs of (x_t, y_t) available. SCCF that are larger than $1.96/\sqrt{n}$ in magnitude are then deemed significantly different from zero.

- **Simulation**.

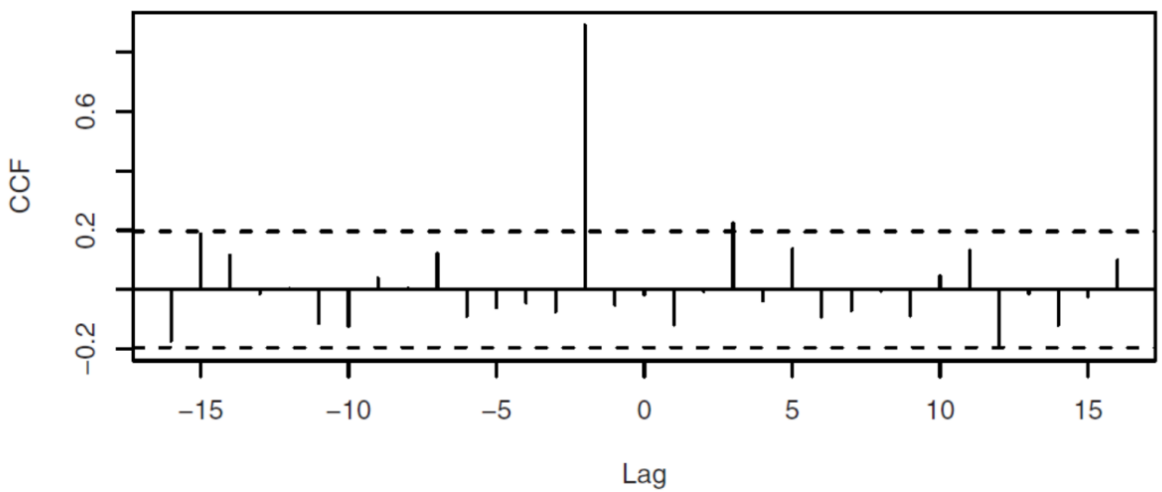

Fig. 1: Sample Cross-Correlation from Equation (1) with $d = 2$. Here, $\beta_0 = 0, \beta_1 = 1$. X and e are independent, and $X \sim \mathcal{N}(0, 1)$ and $e \sim \mathcal{N}(0, 0.25)$. 100 pairs (x_t, y_t).

In this example, X and Y are each WN series, Even though x_{t-2} correlates with y_t.

Spurious correlation

A more useful regression model

$$y_t = \beta_0 + \beta_1 x_{t-d} + z_t,$$

where z_t may follow some ARIMA(p, d, q) model. The variance of $\sqrt{n}\, r_{X,Y}(k)$ is approximately

$$1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y).$$

Particularly, if X and Y are both AR(1) processes with AR(1) coefficients ϕ_X and ϕ_Y, respectively. Then

$$\sqrt{n} r_{X,Y}(k) \xrightarrow{d} \mathcal{N}\left(0, \frac{1+\phi_X\phi_Y}{1-\phi_X\phi_Y}\right).$$

When both AR(1) coefficients are close to 1, the ratio of the sampling variance of $r_{X,Y}(k)$ to the nominal value of $1/n$ approaches infinity. Thus, the unquestioned use of the $1/n$ rule in deciding the significance of the SCCF may lead to many more false positives than the nominal 5% error rate, even though the response and covariate time series are independent of each other. The problem of inflated variance of the SCCF becomes more acute for nonstationary data.