Regression with Time Series Data

Tamer Çetin

Preparation of Time Series Data

- Time series data has observations that refer to the same unit but different points in time.

$$y_t^E = \alpha + \beta x_t + e_t$$

- To run a regression of y on x in time series data, the two variables need to be at the same time series frequency.
- When the time series frequencies of y and x are different, we need to adjust one of them.
- Most often that means aggregating the variable at higher frequency (e.g., from weekly to monthly).

Preparation of Time Series Data

- Once variables are at the same time series frequency, the next step is to plot the time series: extreme values and time gaps!
- With two time series variables y and x, we may show the two on the same graph.

Trend and Seasonality

- A fundamental feature of time series data is that variables evolve with time.
- They may hover around a stable average value, or they may drift upwards or downwards.
- A variable in time series data follows a **trend** if it tends to change in one direction; in other words, it has a tendency to increase or decrease.
- Because of such systematic changes, later observations of trending variables tend to be different from earlier observations.
- A time series variable follows a positive trend if its change is positive on average.
- It follows a negative trend if its change is negative on average.
- A change in a variable, also called the first difference (FD) of the variable, is denoted by Δ ($\Delta y_t = y_t - y_{t-1}$).

Trend and Seasonality

- A positive trend and a negative trend are defined as follows:
  - Positive trend: $E[\Delta y_t] > 0$
  - Negative trend: $E[\Delta y_t] < 0$
- We say that the trend is **linear** if the change is the same on average as time passes. We say the
- trend is **exponential** if the relative change in the variable is the same on average as time passes.
- Variables in business and economics often follow an exponential trend due to the multiplicative nature of the mechanisms behind them.

Trend and Seasonality

- Besides a potential trend, another important property of many time series variables is **seasonality**.
- Seasonality means that the value of the variable is expected to follow a cyclical pattern, tracking the seasons of the year, days of the week, or hours of the day.
- Similarly to trends, seasonality may be linear, when the seasonal differences are constant.
- Seasonality may also be exponential, if relative differences (that may be approximated by log differences) are constant.
- Many economic activities follow seasonal variation over the year, and through the week or day.

Stationarity, Non-stationarity, Random Walk

- Understanding trends and seasonality is important because they make regression analysis challenging.
- They are examples of a broader concept, non-stationarity.
- **Stationarity** means stability; non-stationarity means the lack of stability.
- Stationary time series variables have the same expected value and the same distribution at all times.
- Trends and seasonality violate stationarity because the expected value is different at different times.
- With a positive trend, the expected value increases over time; with a negative trend, the expected value decreases over time; with seasonality, the expected value follows a regular cycle.

Stationarity, Non-stationarity, Random Walk

- Another common kind of non-stationarity is the **random walk**.
- Time series variables that follow a random walk change in random, unpredictable ways.
- Whatever the current value of the variable and whatever its previous change was, the next change may be anything:
  - positive or negative, small or large.
- As a formula, y_t follows a random walk if its value at time t equals its previous value plus a completely random term, which we denote by e_t:

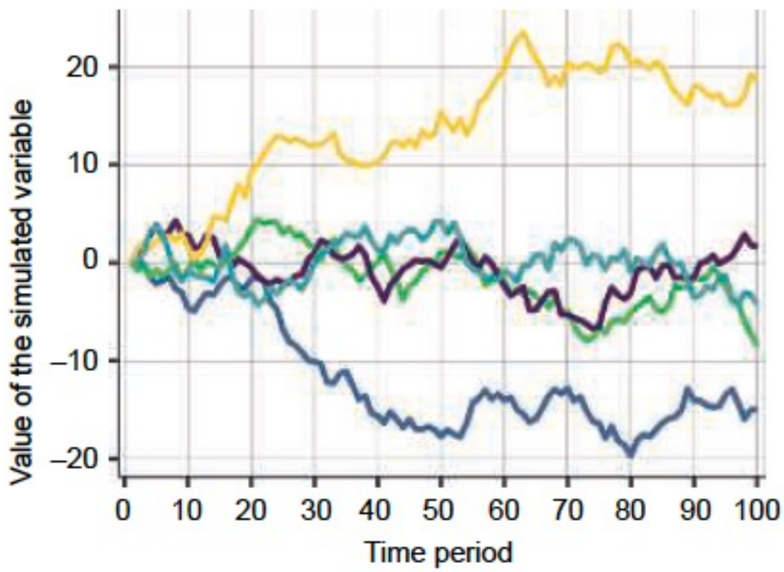
$$y_t = y_{t-1} + e_t$$

Stationarity, Non-stationarity, Random Walk

- The random term (e_t) is independent across time:
  - its value next time is independent of what it was this time.
- But the random walk variable itself, y_t, is not independent across time.
- Its value next time is heavily dependent on its value last time.
- We cannot predict this behavior because random walks are non-stationary as they violate another aspect of stability:
  - their spread increases with time.
- As each step is completely random, independent of all previous steps, there is no way to use information from the past to predict the next steps.

Stationarity, Non-stationarity, Random Walk

This figure illustrates the non-stationary nature of random walks.
The standard deviation calculated over an interval or variance increases with time.
In general, time series variables that have this random walk-like property are called variables with a **unit root**.



Stationarity, Non-stationarity, Random Walk

- The problem with non-stationarity is that the ingredients of correlation coefficients (covariance, standard deviation) may change over time.
- This means that, if a time series variable is non-stationary, its first-order, second-order, and higher-order correlation coefficients may change with time:
  - No consistency
- In contrast, if a time series variable is stationary, its first-order correlation is the same across time, its second-order correlation is also the same across time, and so on.

Time Series Regression

- Regression with time series data is defined and estimated the same way as with other data.
- But we add something to our usual notation here:
  - the **time index** of the variables, such as y_t and x_t.
- This additional notation serves two purposes.
- First, it reminds us that the regression is on time series data, which will allow for additional interpretation of its coefficients.
- Second, later we'll add observations from other time periods to the regression, and there it's essential to record the time period to know what's what.
- With this new notation, a linear regression with one explanatory variable on time series data is the following:

$$y_t^E = \alpha + \beta x_t$$

$$\Delta y_t^E = \alpha + \beta \Delta x_t$$

$$\Delta y_t = y_t - y_{t-1}$$

Time Series Regression

- Trends and seasonality in both y and x in a regression introduce a spurious association.
- Transforming variables into first differences helps trends and random walks.
- Use season dummies for seasonality or time-over-time differences.

Serial Correlation

- After trend and seasonality, the third property of many times series we consider in this chapter is **serial correlation**.
- Serial correlation means correlation of a variable observed at some time with the previous values of the same variable.
- Previous observations of a time series variable are called its **lagged values**, or simply its **lags**.
- Thus, serial correlation is correlation between the observation of a variable and a lagged observation of the same variable.
- Serial correlation does not violate stationarity.
- It is perfectly possible for a time series to have serial correlation but have no trend, seasonality, or any other kind of non-stationary feature.

Serial Correlation

- Serial correlation is characterized by the correlation coefficient.
- The first-order serial correlation coefficient of variable y is the correlation between its value at time t, y_t, and its first lag, y_{t-1}:

$$\rho_1 = Corr[y_t, y_{t-1}]$$

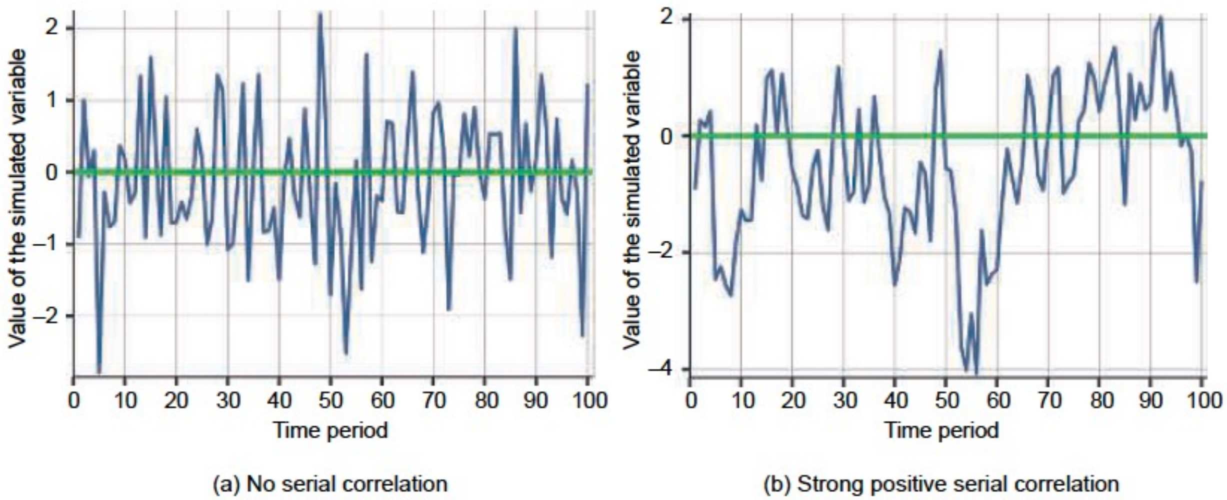


Figure 12.9 Two simulated time series variables with different serial correlation

Serial Correlation

- Serial correlations does not necessarily mean an issue in time series analysis.
- For instance, when a variable is serially correlated, its present value may help predict its value in the near future.
- Here serial correlation can be very useful.
- But sometimes it is an issue.
  - Violation of Independence Assumption
  - Incorrect Model Estimation
  - Model Misspecification
  - Inefficient Use of Information

Dealing with Serial Correlation in Time Series Regressions

- The first solution is to estimate standard errors that are robust to serial correlation.
- All coefficients are the same, only the standard errors are different.
- The second solution is including the **lagged dependent variable** in the right-hand side of the regression.
- The idea is to specify a regression that incorporates all serial correlation of the left-hand-side variable in the regression and produces residuals that are serially uncorrelated.

Lags of x in a Time Series Regression

- An important advantage of regression analysis using time series data is its ability to uncover associations across time.
- For example, we may ask what we can expect to happen for a dependent variable one or two time periods after an explanatory variable changes.
- These are called **lagged associations**, or if they show true effects, **lagged effects**.
- Having lags is very useful when it takes time for an effect to materialize.
- A technology investment, a jump in oil prices, and/or a change in social policy.
- The time series regression in first differences with a single right-hand-side variable x and its two lags has the form

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + \beta_2 \Delta x_{t-2}$$

The Process of Time Series Regression Analysis

- First, we need to understand the time series frequency of each variable we'll work with.
- We need to deal with gaps if necessary, and we need to make sure that all variables have the same time series frequency.
- This may require aggregating variables to a lower frequency.
- The second step is exploratory data analysis.
- The next step is to specify the regression.
  - Check seasonality, if lags are needed,
- Finally, we have to get the standard errors right.
  - In time series data that means handling serial correlation on top of heteroskedasticity.
- Interpretation of the results.