Modeling Temporal Dependence and Exploring Quantities of Interest Using `dynamac`

CARTS/ITSS Quantitative Methods Workshop

Andrew Q. Philips

University of Colorado Boulder

October 4, 2018

Plan for Today

- Who are you?
- Why care about temporal dependence?
- Diagnosing temporal dependence
- Modeling temporal dependence using `dynamac`
- Since we have limited time, we'll move through the lecture, I will show you the labs on-screen, and we can spend the last 30 minutes in lab format
- Labs will be provided in `R`
- This is a short workshop; time series could take up a full semester course...lot's we're not covering

Course Introductions

- Name
- Department
- What are you studying?

Why Care About Temporal Dependence?

A simulation:

- Generate 20 unrelated random normally distributed variables (e.g., $y \sim N(0,1)$)
- Generate 20 unrelated variables that exhibit time series behavior
- Plot them
- Correlate them. Given that each of the series are unrelated, how many do you think will be statistically significant?

Lab:

- TS correlations.R

Why Care About Temporal Dependence?

- As the lab showed, time series data produced statistically significant correlations—even when the series were unrelated—at rates well above 0.05
- This is because time series data nearly always violate the assumption of independent observations:
  - Unemployment this month is related to last month
  - Snow in the Rockies this season is related to last season
  - Presidential poll taken this week is related to the past week
  - Weather this hour is similar to the previous hour
  - Level of corruption in a country this year is highly related to last year

A Crucial OLS Assumption

$\varepsilon_i \sim N(0, \sigma^2)$

- What does this mean?
- How many assumptions does it entail?
- With time series data, often our OLS assumptions are stretched/not-believable
- This is especially true with the assumptions about the error
- Analogously, any time we fail to correctly specify a dynamic process (when it exists), those omitted variables end up in the error term
- This likely will lead to violations of the OLS assumptions

Quick Note on Notation

- You're probably used to textbooks using cross-sectional notation

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

- Time series data index by t instead of i:

$$y_t = \alpha + \beta x_t + \varepsilon_t \tag{2}$$

- i.e., we're analyzing a single cross-sectional unit, i, over multiple time points t
- As discussed earlier, the assumption that ε_t is independent and identically distributed may be violated regularly in time series:
  - Errors may not be independent (violate assumption of no autocorrelation)
  - Errors may not have constant variance over time (violate assumption of no heteroskedasticity)

Characteristics of Time Series Data

- Data are sampled consistently (i.e., the time intervals between data points are the same)
- Order matters; a time variable is required in a time series dataset
- Different levels of aggregation:
  - highly aggregated to disaggregated: Annual→Quarterly→Monthly→Daily
  - highly disaggregated to aggregated: Hourly→Daily→Weekly→Monthly
- Stretching of frequentist concept of sample-to-population; our inability to "re-run" the world
- Use of secondary data (collected by someone else)
- Missing data are problematic

Lags

- Previously, we said that order matters when working with time series data
- Data that are more proximate to one another tend to be more correlated
- Often, we account for this in our model by using lags: $y_{t-1}, y_{t-2}, y_{t-3}, \cdots, y_{t-s}$, which tell us the value of a series one, two,...,s periods ago
- Lags will become extremely useful in our models as we will use them to condition for the effect from previous time periods

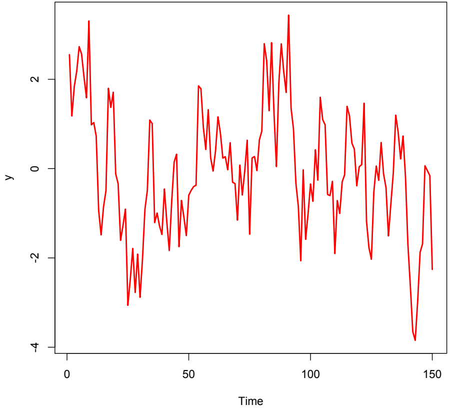
Lags

Time	y_t	y_{t-1}	y_{t-2}
1	3.3	–	–
2	4.5	3.3	–
3	3.2	4.5	3.3
4	6.4	3.2	4.5
5	4.3	6.4	3.2
6	7.8	4.3	6.4
7	8.0	7.8	4.3
8	8.3	8.0	7.8
9	8.2	8.3	8.0
10	9.1	8.2	8.3

Diagnosing Time Series

- We need to identify the characteristics of the series
- This will determine the types of models we can estimate
- Generally:
  1. Plot
  2. Diagnose stationarity
- The very first thing we should do is plot the series

Plot the Series



Plotting allows you to identify missing data, trends, other issues. I'm personally skeptical about the human ability to "diagnose" visually

Diagnose Stationarity

- Stationarity is a very important topic in time series
- A series is stationary, $y_t \sim I(0)$, if it has all of the following characteristics:
  1. Mean reverting: $E[y_t] = \mu$
  2. Constant variance: $E[y_t^2] = \sigma^2$
  3. Constant covariance $E[y_t, y_{t-s}] = \rho \ \forall \ t \neq s$
- We will discuss one of the most common stationary time series before moving onto what a violation of stationarity looks like

Diagnose Stationarity

- One common type of stationary series is an autoregressive process of order 1, or AR(1):

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t \tag{3}$$

where $\varepsilon_t \sim N(0, \sigma^2)$, $|\phi| < 1$

- Past values of y_t are related to current values. To see this, we can write the value of y_{t-1}:

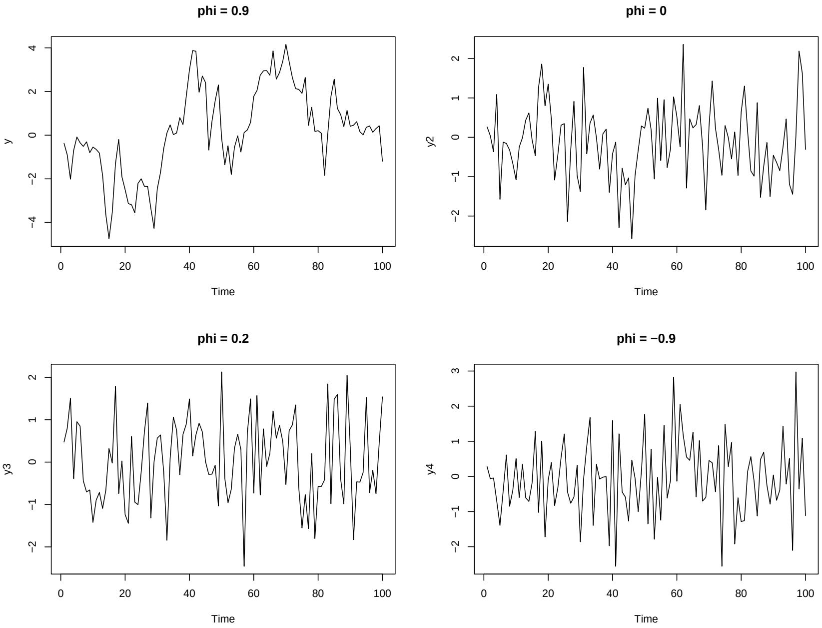
$$y_{t-1} = \alpha + \phi y_{t-2} + \varepsilon_{t-1} \tag{4}$$

- Substituting:

$$y_t = \alpha + \phi(\alpha + \phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = (1+\phi)\alpha + \phi^2 y_{t-2} + \varepsilon_t + \phi \varepsilon_{t-1} \tag{5}$$

All About ϕ

- The value of ϕ is particularly important. Previously we set $|\phi| < 1$ to ensure stationarity
- Positive values of ϕ result in positive autocorrelation; current values of y_t are positively related to past ones
- Negative values result in negative autocorrelation; given a high/positive value in the previous period, we expect a low/negative value in this period
- Larger $|\phi|$ means that the relationship across periods is stronger
- If $\phi = 0$, there is no autocorrelation
- We can have multiple AR terms in our model, although commonly only a single lag is included for each variable


phi = 0.9
phi = 0
phi = 0.2
phi = −0.9
y
y2
y3
y4
Time

Lab

Lab:

- Simulate AR series.R

Unit Root

- Series with constant mean, variance, and covariance are said to be stationary
- A violation of one of these conditions implies the series is non-stationary, or $I(1)$
- One common form of non-stationarity is the unit root:

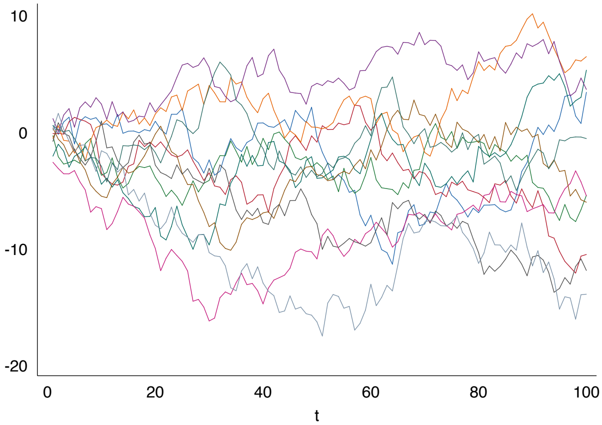
$$y_t = y_{t-1} + \varepsilon_t \tag{6}$$

- What is ϕ here?
- Knowing that $y_{t-1} = y_{t-2} + \varepsilon_{t-1}$, and so on:

$$y_t = y_0 + \sum_{t=1}^{T} \varepsilon_t \tag{7}$$

- In other words, y_t is the cumulative sum of all past innovations over time. Shocks never wear off...or the series has "full memory"

Unit Roots



Testing For Unit Roots

- Non-stationary series cannot enter into our model due to the spurious regression problem (see "TS correlations.R", though exceptions exist (De Boef and Keele 2008; Philips 2018))
- Instead, we must test to see if a series is non-stationary
- A common test is the Dickey-Fuller. In brief:
  - H_0: Series contains a unit root
  - H_a: Series is stationary
- If we fail to reject the null, we take the first-difference to render the series stationary:

$$\Delta y_t = y_t - y_{t-1} \tag{8}$$

since if $y_t \sim I(1) \rightarrow y_t = y_{t-1} + \varepsilon_t$, then $\Delta y_t = y_{t-1} - y_{t-1} + \varepsilon_t \sim I(0)$
- Taking the first-difference implies moving from the series in levels to the series in changes. This can affect our interpretation/limits our analyses

`dynamac`

- `R` and Stata package used to model time series and explore statistical and substantive significance of our results
- Uses a stochastic simulation procedure similar to `Zelig` (`R`) or Clarify (Stata)
- Plot predicted values with associated measures of uncertainty in response to a change in x
- Available on CRAN and hosting site: https://andyphilips.github.io/dynamac/
- Also contains a cointegration test (not enough time to discuss this ☹)

dynamac Steps

Steps in using dynamac:

- 1. Test each variable for stationarity. If I(1), take first-difference (we can also account for this in dynamac)
- 2. Create model specification. A very common and flexible parameterization is the autoregressive distributed lag ARDL(1,1):

$$y_t = \beta_0 + \phi y_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + \varepsilon_t \tag{9}$$

In dynamac:

```
ardl.model <- dynardl(y~x, data = data, lags = list("y" =
1, "x" = 1), levels = c("x"))
```

`dynamac` Steps

- 3. Create and use dynamic simulations. Steps:
  - Draw 1000 $\hat{\boldsymbol{\beta}}^*_s \sim N(\hat{\boldsymbol{\beta}}, V\hat{C}V)$
  - Draw simulated $\hat{\sigma}^*_s$ (from inv-χ^2, coming from $\hat{\sigma}^2$ scaled by $n-k$)
  - Simulate predicted value at time $t = 1$ (setting $\mathbf{x}$ to means) for all s simulations:

$$y_{s1} = \bar{y} + \mathbf{x}\hat{\boldsymbol{\beta}}^*_s + u_{s1} \qquad u_{s1} \sim N(0, \hat{\sigma}^*_s) \tag{10}$$

  - Average across simulations to create $\bar{y}_1$. Then, at $t = 2$, this becomes the value for y_{t-1}. The average prediction at $t = 2$ then becomes the value for y_{t-1} for $t = 3$, and so on...
  - At a point in time, change one x variable for one period and see how this affects y_t
  - Obtain means and percentile confidence intervals, and plot the results

Lab

Lab:

- Estimating models using dynamac.R

Conclusions

- Time series data are quite common, and they nearly always exhibit some form of dependence
- Failing to specify this dependence correctly often violates our regression assumptions
- You have learned enough to start modeling dynamics when they are present. You *should always* be doing this 😎
- LOTS of topics we did not cover today

Conclusions

- Some useful textbooks:
  - Asteriou, Dimitrios and Stephen G. Hall. 2011. Applied Econometrics. 2nd Edition. Palgrave.
  - Pickup, Mark. 2014. Introduction to Time Series Analysis. SAGE Publications. Quantitative Applications in the Social Sciences. 1st Edition.
  - Box-Steffensmeier, Janet M., John R. Freeman, Matthew P. Hitt, and Jon C.W. Pevehouse. 2015. Time series analysis for the social sciences. Cambridge University Press.
  - Enders, Walter. 2010. Applied Econometric Time Series. 3rd Edition. John Wiley & Sons.
  - Shumway, Robert H. and David S. Stoffer. 2017. Time series analysis and its applications. Springer.

Today's Labs

Lab:

- TS correlations.R
- Simulate AR series.R
- Estimating models using dynamac.R