Econometrics I - Intro Time Series

Ryan T. Godwin

University of Manitoba

This is not a comprehensive introduction. This is a short collection of topics that I find important and interesting, namely (i) the inconsistency of LS under lagged dependent variables and autocorrelated errors, and (ii) spurious regressions due to random walks.

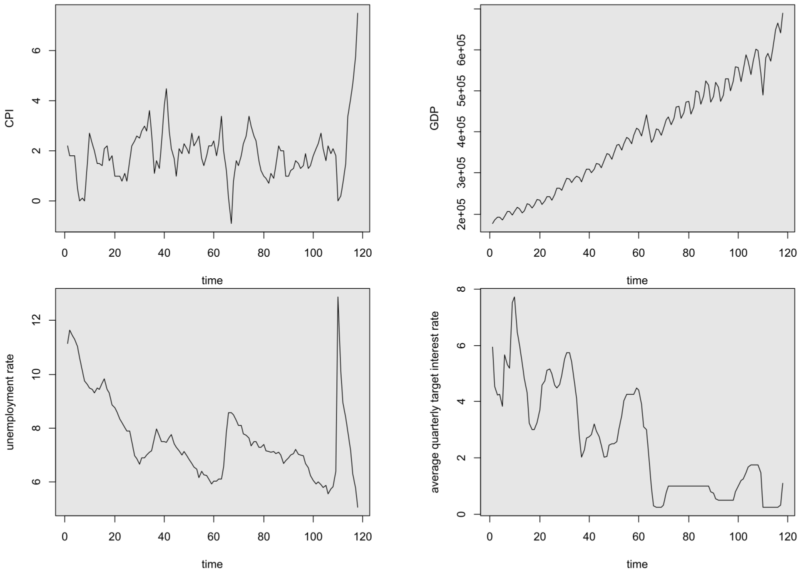
What is a time series

A time series is a single occurrence of a random event. The sequence of observations, $\{y_t\}_{t=-\infty}^{t=+\infty}$, is a time series process. There is no counterpart to repeated sampling for a time series. We observe realizations of this process in a time window, $t = 1, \ldots, T$. The frequency of observations are important, but the length of the window is arguably more important. Asymptotics involves considering an increasingly longer window.

Quarterly time series data on Canadian GDP, CPI, unemployment rate, and average target interest rates from 1993Q1 to 2022Q2 (see Figure 1):

```
can <- read.csv("https://rtgodwin.com/data/canseries.csv")
plot(can$time, can$CPI, type="l")
plot(can$time, can$GDP, type="l")
plot(can$time, can$unemployment, type="l")
plot(can$time, can$interest, type="l")
```

Figure: Various Canadian time series from 1993Q1 to 2022Q2.



Time series models typically seek to *forecast* (predict) future values of the series, and sometimes look to estimate causal effects of policies or interventions (like the effect of interest rates on inflation). In order to model a time series, y_t, it is typical to include:

- contemporaneous factors, x_t
- lagged factors, $x_{t-1}, x_{t-2}, \ldots$
- its own past, $y_{t-1}, y_{t-2}, \ldots$
- a time trend, t
- seasonal dummies, d
- disturbances (innovations), ϵ_t

For example:

$$y_t = \beta_1 + \beta_2 t + \beta_3 d_t + \beta_4 x_t + \beta_5 y_{t-1} + \epsilon_t$$

Autocorrelation

In time series models, the past is very important! Often a very good predictor of y_t are the $y_{t-1}, y_{t-2}, \ldots$. This suggests that the error term, ϵ_t, also depends on its own past values. That is, even after including other time series (x_t) in the regression, there are likely still missing time series in ϵ_t that depend on their own past. This means that the error term will very likely violate A.4. The error term is correlated to its own past values. This is called serial correlation, or autocorrelation.

In the case of autocorrelation, the off-diagonal elements of $V(\boldsymbol{\epsilon})$ will be non-zero. The particular values they take will depend on the form of autocorrelation. That is, they will depend on the pattern of the correlations between the elements of the error vector. For example, the covariance matrix of the error term will look like:

$$V(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma^2 \end{bmatrix}$$

If the errors themselves are autocorrelated, often this will be reflected in the regression residuals also being autocorrelated. That is, the residuals will follow some sort of pattern, rather than just being random.

Let's take a look at the quarterly Canadian GDP series. We know from a previous section how to de-seasonalize (and de-trend) a time series:

```r
# Create quarterly dummies
n <- 118
can$q4 <- can$q3 <- can$q2 <- can$q1 <- 0
can$q1[seq(1, n, 4)] <- 1
can$q2[seq(2, n, 4)] <- 1
can$q3[seq(3, n, 4)] <- 1
can$q4[seq(4, n, 4)] <- 1

# Regress GDP on a time trend and quarterly dummies
gdp.mod <- lm(can$GDP ~ can$time + can$q2 + can$q3 + can$q4)

# Collect the residuals, which are de-trended, de-seasonalized GDP
gdp.resid <- gdp.mod$residuals
```

Now that we have de-trended and de-seasonalized GDP, let's try to explain GDP in terms of its past values. We'll estimate the equation (where $GDP^\star$ is the de-trended, de-seasonlaized series):

$$GDP^\star_t = \beta_0 + \beta_1 GDP^\star_{t-1} + \beta_2 GDP^\star_{t-2} + \beta_3 GDP^\star_{t-3} + \beta_4 GDP^\star_{t-4} + \beta_5 GDP^\star_{t-5} + \epsilon$$

We need to regress GDP on it's lagged values (notice that we lose 5 observations):

```
gdp.mod.lag <- lm(gdp.resid[6:118] ~ gdp.resid[5:117]
  + gdp.resid[4:116] + gdp.resid[3:115] + gdp.resid[2:114]
  + gdp.resid[1:113])
summary(gdp.mod.lag)
```

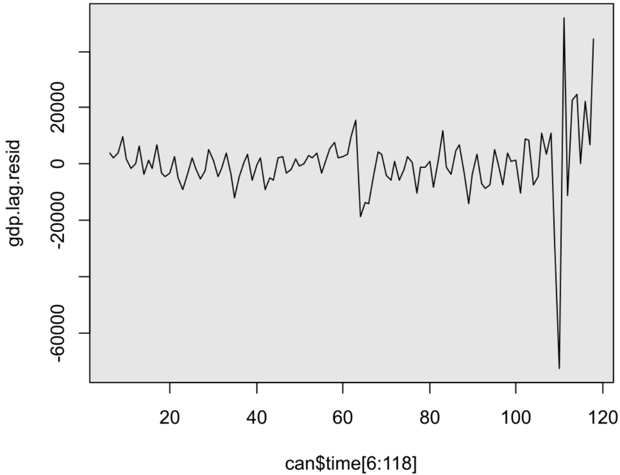
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.7951  1181.7694  -0.005  0.99610
gdp.resid[5:117]    0.9013     0.1003   8.985    1e-14 ***
gdp.resid[4:116]   -0.3254     0.1336  -2.435  0.01655 *
gdp.resid[3:115]    0.2336     0.1389   1.682  0.09551 .
gdp.resid[2:114]    0.1455     0.1343   1.083  0.28111
gdp.resid[1:113]   -0.3140     0.1050  -2.989  0.00347 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
    1

Residual standard error: 12330 on 107 degrees of freedom
Multiple R-squared:  0.5273,  Adjusted R-squared:  0.5052
F-statistic: 23.87 on 5 and 107 DF,  p-value: 4.666e-16
```

Now, we'll once again get the residuals from this model, and plot them over time (see Figure 2):

```r
gdp.lag.resid <- gdp.mod.lag$residuals
plot(can$time[6:118], gdp.lag.resid, type="l")
```

Figure: The residuals from our time series model are autocorrelated.



The figure shows some sort of patern in the residuals, indicating that the error term is autocorrelated. The time series model that we estimated is simple, and we could include explanatory $x_t, x_{t-1}, \ldots$ variables. Even when we do so, however, we are unlikely to be able to completely account for the autocorrelation apparent in the residuals.

If the errors of our model are autocorrelated, then the OLS estimator of $\boldsymbol{\beta}$ usually will be unbiased and consistent, but it will be inefficient. In addition $V(\boldsymbol{\beta})$ will be computed incorrectly, and the standard errors, etc., will be inconsistent. (Same situation as with heteroskedasticity). In general time series models are concerned with testing for the presence/absence of autocorrelation, estimating the form of autocorrelation (the $V(\boldsymbol{\beta})$ matrix), and then estimating models where the errors are autocorrelated. In this introduction, we will not look at these methods. We will consider two ways of modelling autocorrelation: an AR process and an MA process. We will also consider a limiting form of an AR process.

Autoregressive process

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right] \quad ; \quad |\rho| < 1$$

This is an AR(1) model for the error process. More generally:

$$\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \cdots + \rho_p\epsilon_{t-p} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right]$$

This is an AR(p) model for the error process. [e.g., p = 4 with quarterly data.] An AR process can be used to model any time series, not just the error term. Notice that we used an AR(5) process in our GDP example above.

Moving average process

$$\epsilon_t = u_t + \phi u_{t-1} \quad ; \quad u_t \sim \text{ i.i.d.N } \left[0, \sigma_u^2\right]$$

This is an MA(1) model for the error process. More generally:

$$\epsilon_t = u_t + \phi_1 \epsilon_{t-1} + \cdots + \phi_q u_{t-q} \quad ; \quad u_t \sim \text{ i.i.d. } N \left[0, \sigma_u^2\right]$$

This is an MA(q) model for the error process. We can combine both types of process into an ARMA(p, q) model:

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \cdots + \rho_p \epsilon_{t-p} + u_t + \phi_1 u_{t-1} + \cdots + \phi_q u_{t-q} \quad ; \quad u_t \sim \text{ i.i.d. } N$$

Stationarity

Note that in the AR(1) process, we said that $|\rho| < 1$. This condition is needed to ensure that the process is "stationary." Suppose that the following three conditions are satisfied:

1. $E[\epsilon_t] = 0 \quad ; \quad$ for all t
2. $\text{var}.[\epsilon_t] = \sigma^2 \quad ; \quad$ for all t
3. $\text{cov}\cdot[\epsilon_t, \epsilon_s] = \gamma_{|t-s|} \quad ; \quad$ for all $t, s; t \neq s$

Then we say that the time-series sequence, ϵ_t, is "Covariance Stationary"; or "Weakly Stationary".

Unless a time-series is stationary, we can't identify and estimate the parameters of the process that is generating its values.

Let's see how this notion relates to the AR(1) model, introduced above. We have:

$$\begin{aligned}
\epsilon_t &= \rho\epsilon_{t-1} + u_t \\
\epsilon_t &= \rho\left[\rho\epsilon_{t-2} + u_{t-1}\right] + u_t \\
&= \rho^2\epsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2\left[\rho\epsilon_{t-3} + u_{t-2}\right] + \rho u_{t-1} + u_t \\
&= \rho^3\epsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t
\end{aligned}$$

Continuing in this way, eventually, we get:

$$\epsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots \tag{1}$$

This is an infinite order MA process. The value of ϵ_t embodies the entire past history of the u_t values.

From equation 1, $\mathrm{E}(\epsilon_t) = 0$, and:

$$\begin{aligned} \operatorname{var}(\epsilon_t) &= \operatorname{var}.(u_t) + \operatorname{var}.(\rho u_{t-1}) + \operatorname{var}.\left(\rho^2 \epsilon_{t-2}\right) + \cdots \\ &= \sigma_u^2 + \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + \cdots \end{aligned}$$

Question: Under what conditions will this series *converge*?

The series will converge to $\sigma_u^2 \left(1 - \rho^2\right)^{-1}$, as long as $\left|\rho^2\right| < 1$, and this in turn requires that $|\rho| < 1$.

This is a necessary condition needed to ensure that the process ϵ_t is stationary, because if this condition isn't satisfied, then $\operatorname{var}(\epsilon_t)$ is infinite.

If we have a (stationary) AR(1) process, for example, then it can be shown that the covariance matrix for $\boldsymbol{\epsilon}$ is:

$$V(\boldsymbol{\varepsilon}) = \sigma_u^2\Omega = \frac{\sigma_u^2}{(1-\rho^2)} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \ddots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \ldots & 1 \end{bmatrix} \tag{2}$$

So, if we can estimate ρ (for example by looking at the residuals), we can estimate the entire covariance matrix and use FGLS to obtain an estimator that is efficient and has consistent standard errors.

Inconsistency of LS with AR errors and lagged dependent variables

In the presence of autocorrelation, in general $\boldsymbol{b}$ will still be a consistent estimator. However, there is one important situation where it will be inconsistent. This will be the case if the errors are autocorrelated, and one or more lagged values of the dependent variable enter the model as regressors. A quick way to observe that inconsistent estimation will result in this case is as follows. Suppose that:

$$\begin{aligned} y_t &= \beta y_{t-1} + \epsilon_t \quad ; \quad |\beta| < 1 \\ \epsilon_t &= \rho\epsilon_{t-1} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } \left[0, \sigma_u^2\right] \quad ; \quad |\rho| < 1 \end{aligned} \tag{3}$$

Now subtract ρy_{t-1} from the expression for y_t in equation 3:

$$(y_t - \rho y_{t-1}) = (\beta y_{t-1} + \epsilon_t) - \rho\left(\beta y_{t-2} + \epsilon_{t-1}\right) \tag{4}$$

or:

$$\begin{aligned} y_t &= (\beta + \rho) y_{t-1} - \beta\rho y_{t-2} + (\epsilon_t - \rho\epsilon_{t-1}) \\ &= (\beta + \rho) y_{t-1} - \beta\rho y_{t-2} + u_t \end{aligned} \tag{5}$$

So, if we estimate the model with just y_{t-1} as the only regressor, then we are effectively omitting a relevant regressor, y_{t-2}, from the model. This amounts to imposing a false (zero) restriction on the coefficient vector, and we know that this causes OLS to be not only biased, but also inconsistent.

Random walk

Working with non-stationary data is dangerous. A non-stationary time series where:

$$y_t = y_{t-1} + \epsilon_t$$

is said to be integrated of order one I(1), or is said to follow a *random walk*. The danger arises as we can easily find what is called a "spurious relationship" between two unrelated random walks. For example, suppose that x_t also follows a random walk:

$$x_t = x_{t-1} + \varepsilon_t$$

Remember that both of these series have infinite variance! Now, what happens if we regress one random walk on another unrelated random walk? We shouldn't find any relationship, because the two series are unrelated, right? Right?? Granger and Newbold[1] performed a simple simulation showing that such a regression yields spurious results.

```r
n <- 100
y <- x <- 0
for(i in 2:n){
 y[i] <- y[i - 1] + rnorm(1)
 x[i] <- x[i - 1] + rnorm(1)
}
plot(y, type = "l", col = "red", ylim=c(min(x,y),max(x,y)))
points(x, type = "l", col = "blue")
summary(lm(y ~ x))
```

As the sample size increases, the t-statistic on β_1 goes to infinity, and the R^2 goes to 1, even though there is no relationship! This is because both series have infinite variance.

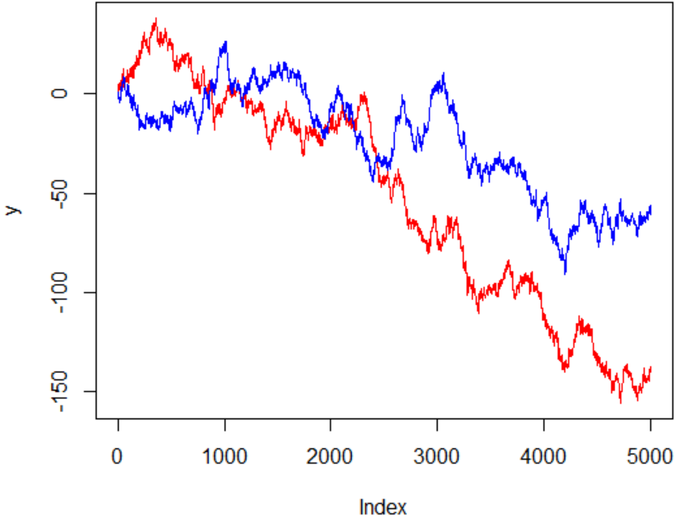
[1] Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. Journal of econometrics, 2(2), 111-120.

```r
set.seed(7010)
n <- 5000
y <- x <- 0
for(i in 2:n){
  y[i] <- y[i - 1] + rnorm(1)
  x[i] <- x[i - 1] + rnorm(1)
}
plot(y, type = "l", col = "red", ylim = c(min(x,y),max(x,y))
    )
points(x, type = "l", col = "blue")
summary(lm(y ~ x))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.94998    0.58101  -18.85   <2e-16 ***
x             1.72176    0.01613  106.74   <2e-16 ***
---
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.54 on 4998 degrees of freedom
Multiple R-squared:  0.6951,  Adjusted R-squared:  0.695
F-statistic: 1.139e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Figure: Two random walks that produce a spurious relationship.



A common way to deal with a random walk is to *first difference* the data. This will ensure that we don't find a spurious relationship (as long as we have the order of integration correct):

```
yd <- y[2:5000] - y[1:4999]
xd <- x[2:5000] - x[1:4999]
summary(lm(yd ~ xd))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02783    0.01414  -1.968   0.0491 *
xd           0.01928    0.01418   1.359   0.1741
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9998 on 4997 degrees of freedom
Multiple R-squared:  0.0003697, Adjusted R-squared:
    0.0001696
F-statistic: 1.848 on 1 and 4997 DF,  p-value: 0.1741
```