Introduction to Time Series Data

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics II

What is Time Series Data?

- A time series variable can be defined as a sequence of observations or measurements indexed by time. For example, $y_t, t = 1, 2, \ldots$, where time subscript t is assumed to be discrete.
- Time series data come with a temporal ordering, usually from earliest to latest.
- The time intervals between observations can be regular or irregular (time frequency). We will only focus on regularly measured time series data (for example, at monthly, annual, weekly, daily frequency).
- We must not forget that the past can affect the future, but not vice versa.

What is Time Series Data?

- The samples were randomly drawn from the appropriate population in the cross-sectional data.
- Understanding why cross-sectional data should be viewed as random outcomes is straightforward: a different sample drawn from the population will generally yield different values of the independent and dependent variables.
- Therefore, the OLS estimates computed from different random samples will generally differ, and this is why we consider the OLS estimators to be random variables.
- How should we think about randomness in time series data?
- We do not know which future values a time series (GDP, closing prices of BIST 100 index, etc.) will take on. Since the outcomes of these variables are not foreknown, they should be clearly be viewed as a random variable.

Time series Process or Stochastic Process

Definition: Time series Process or Stochastic Process

Stochastic process or **time series process** is a sequence of random variables indexed by time (t).

- Stochastic means random.
- When we collect a time series data set, we obtain one possible outcome, or realization, of the stochastic process.
- We can only see a single realization, because we cannot go back in time and start the process over again.
- However, if certain conditions in history had been different, we would generally obtain a different realization for the stochastic process.
- This is why we think of time series data as the outcome of random variables. The set of all possible realizations of a time series process plays the role of the population in cross-sectional analysis.

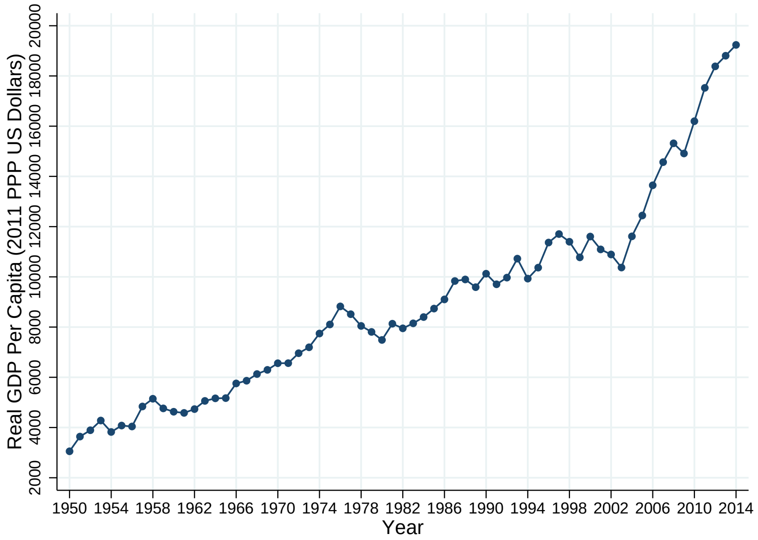
Single Realization of a Time Series

A single realization of a stochastic process can be shown as $\{y_t : t = 1, 2, \ldots, n\}$ or da $\{y_t\}_{t=1}^{n}$. This can be thought of as a a subset doubly infinite series given by:

$$\{y_t\}_{t=-\infty}^{\infty} = \{\ldots, y_{-1}, y_0, \underbrace{y_1, y_2, \ldots, y_{n-1}, y_n}_{\{y_t\}_{t=1}^{n} \text{ realization}}, y_{n+1}, y_{n+2}, \ldots\}$$

- In practice, the time index will always start at 1, but theoretically it can be any integer (or even continuous real number, but we will not cover those).
- If the process can be repeated, then this would result in a different realization (from the same underlying model - or stochastic process).
- In social sciences, economics, finance, and business, we almost always work with single realizations of time series.
- Let's see some examples of time series.

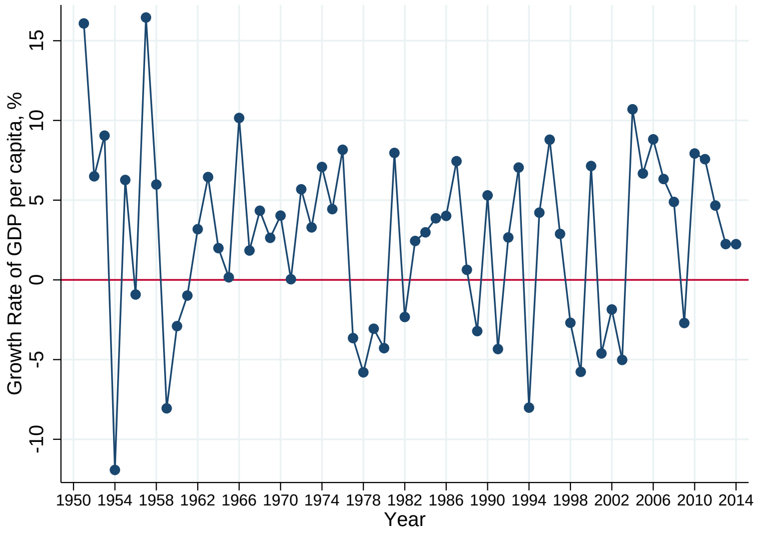
Annual Real GDP per capita in Turkey 1950-2014



Annual Real GDP per capita in Turkey 1950-2014

year	t	GDP_t	GDP_{t-1}	GDP_{t-2}	GDP_{t-3}
1950	1	3054	NA	NA	NA
1951	2	3639	3054	NA	NA
1952	3	3892	3639	3054	NA
1953	4	4279	3892	3639	3054
1954	5	3823	4279	3892	3639
1955	6	4079	3823	4279	3892
1956	7	4042	4079	3823	4279
1957	8	4838	4042	4079	3823
1958	9	5146	4838	4042	4079
⋮	⋮	⋮	⋮	⋮	⋮
2011	62	17525	16198	14914	15317
2012	63	18382	17525	16198	14914
2013	64	18805	18382	17525	16198
2014	65	19236	18805	18382	17525

Growth Rate of Real GDP pc (TURKEY)



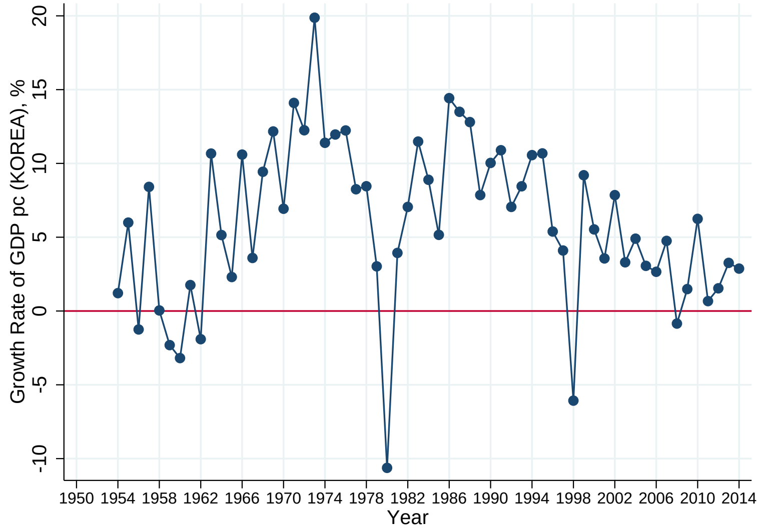
Growth Rate (g_t) of Real GDP pc, Turkey 1950-2014

year	t	GDP_t	g_t	g_{t-1}	g_{t-2}
1950	1	3053.906	NA	NA	NA
1951	2	3639.277	19.17	NA	NA
1952	3	3892.127	6.95	19.17	NA
1953	4	4279.322	9.95	6.95	19.17
1954	5	3823.493	−10.65	9.95	6.95
1955	6	4079.167	6.69	−10.65	9.95
1956	7	4042.11	−0.91	6.69	−10.65
1957	8	4838.391	19.70	−0.91	6.69
1958	9	5146.153	6.36	19.70	−0.91
⋮	⋮	⋮	⋮	⋮	⋮
2011	62	17525.17	8.19	8.61	−2.63
2012	63	18382.36	4.89	8.19	8.61
2013	64	18804.88	2.30	4.89	8.19
2014	65	19236.12	2.29	2.30	4.89

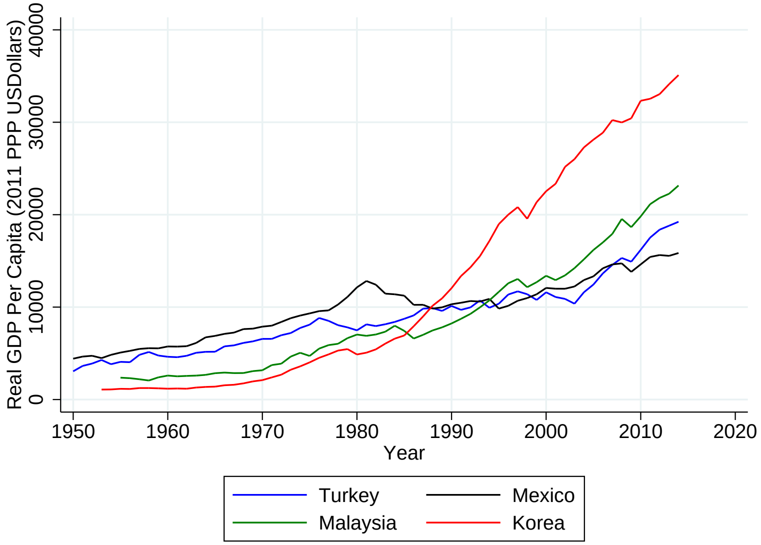
growth rate is defined as $g_t = 100 * \frac{(GDP_t - GDP_{t-1})}{GDP_{t-1}}$

Korean GDP growth rate

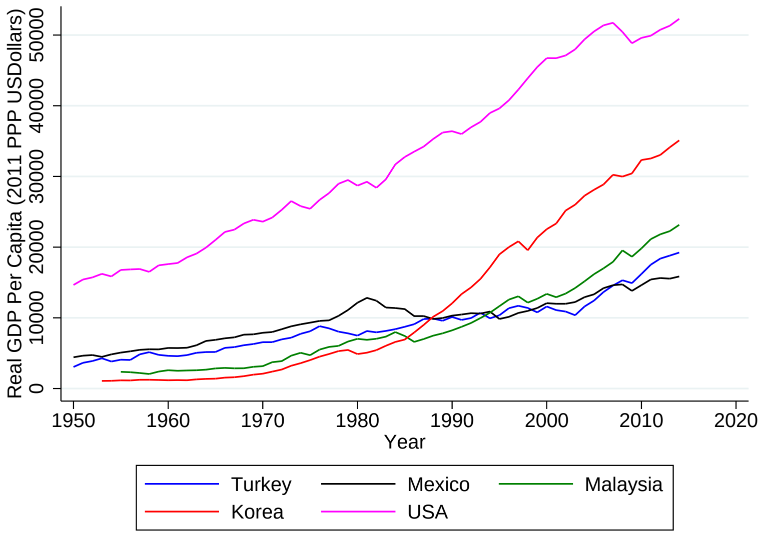
Compare and contrast with Turkey's GDP growth rate



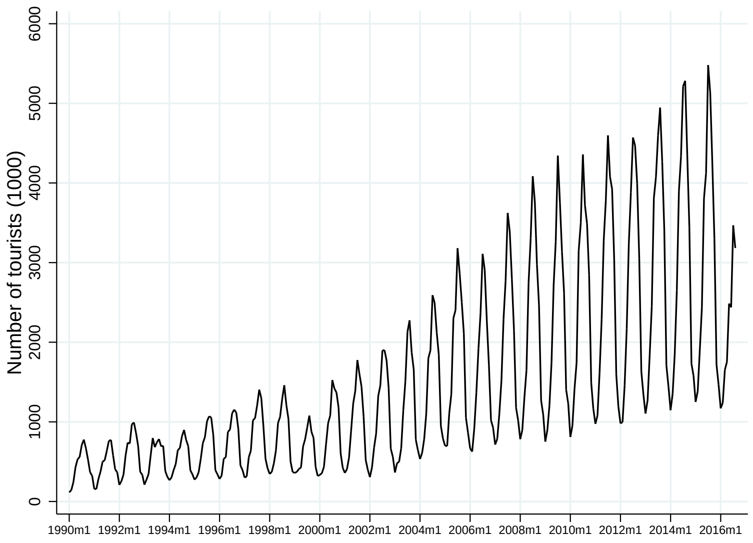
Annual Real GDP per capita in Selected Countries 1950-2014



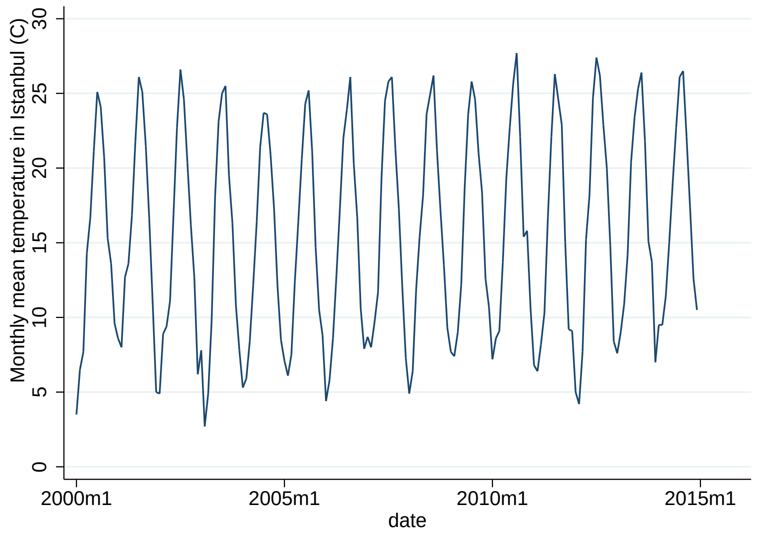
Annual Real GDP per capita in Selected Countries 1950-2014



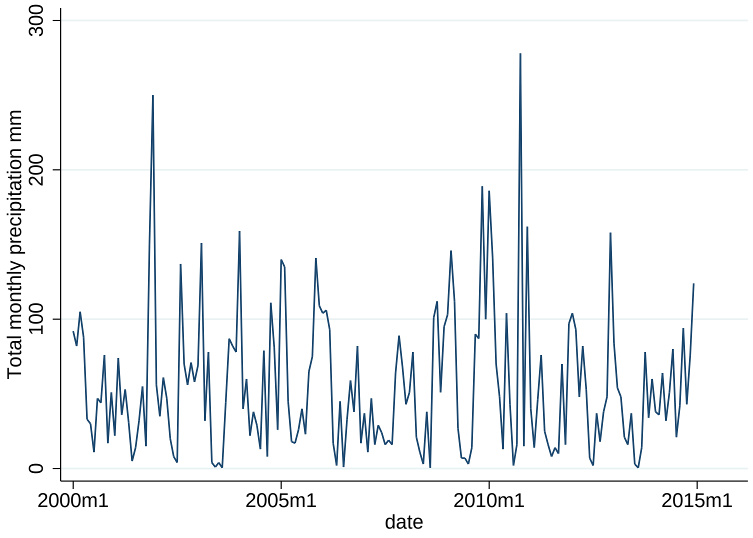
Number of Tourist Arrivals in Turkey



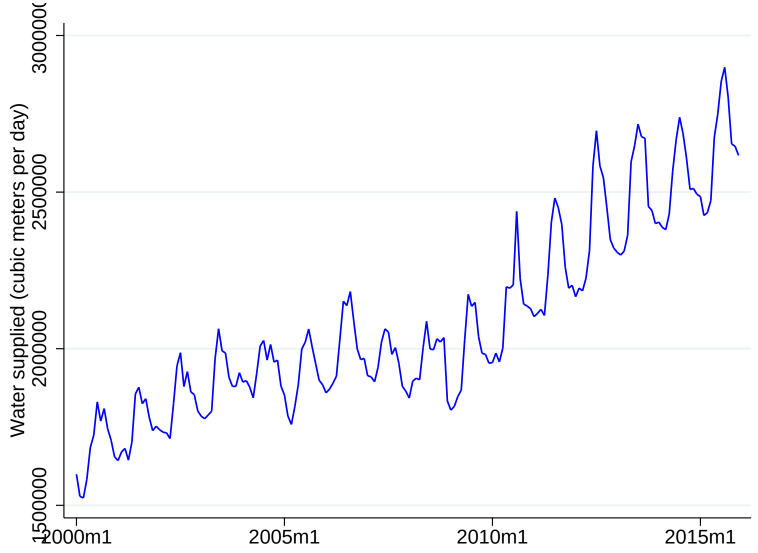
Average Temperature in Istanbul



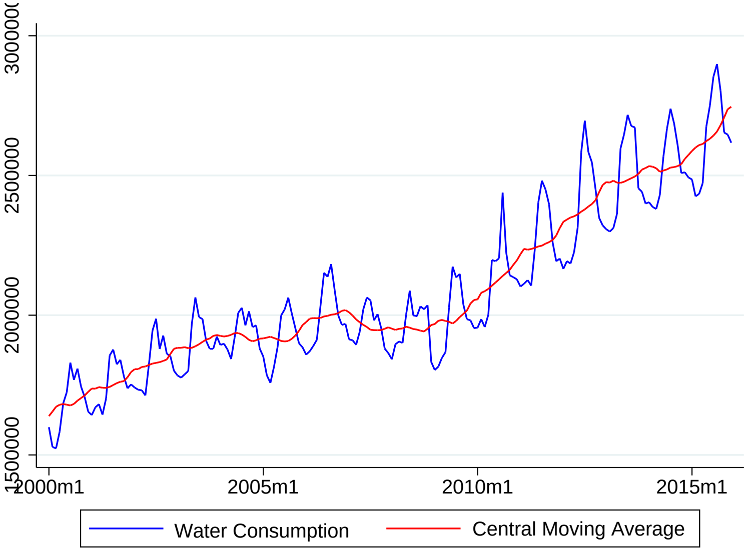
Total Precipitation in Istanbul



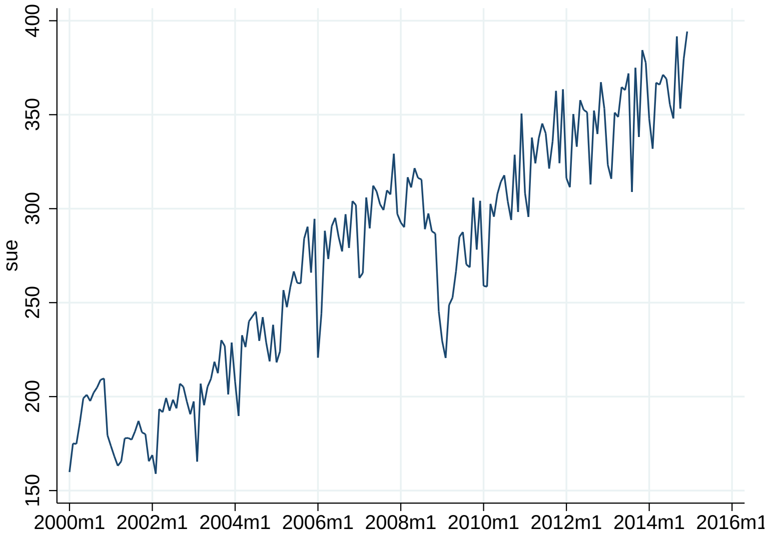
Water Consumption in Istanbul (average pc per day)



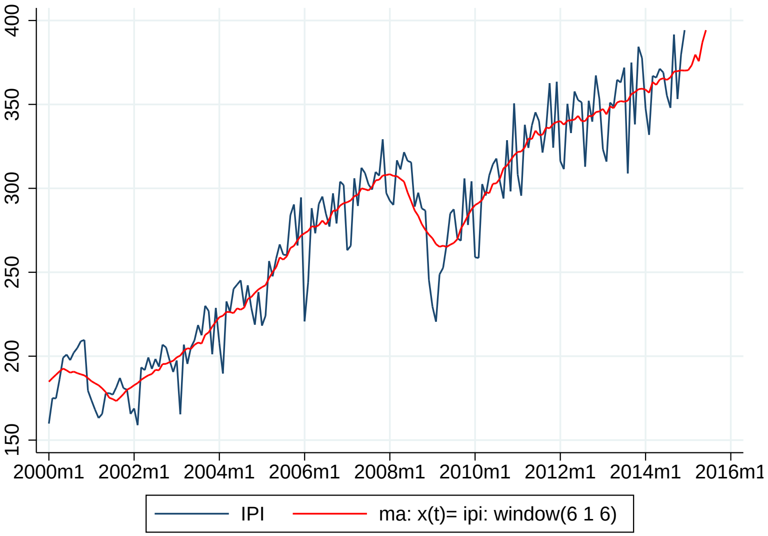
Water Consumption in Istanbul (average pc per day)



Industrial Production Index



Industrial Production Index



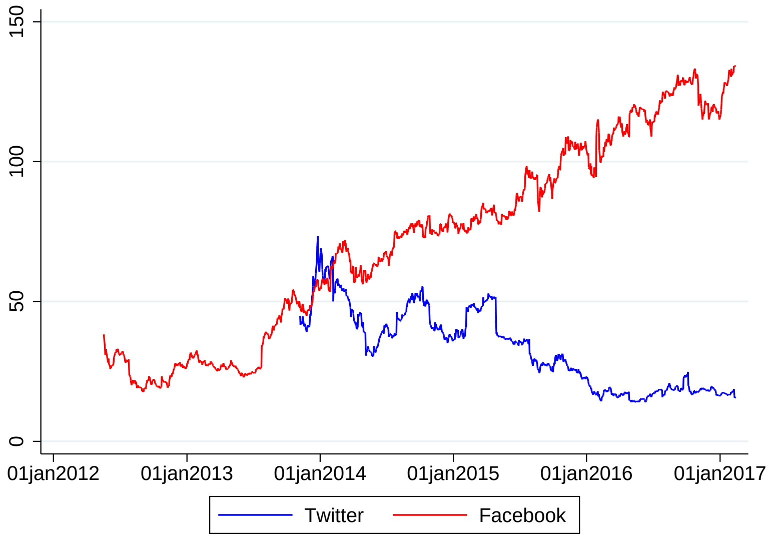
A Note on Index Numbers

- Index numbers are widely used in macroeconometric and financial applications. For example, Industrial Production Index (IPI), Consumer Price Index (CPI).
- Particular values of an index can only be interpreted with the base value at a year. E.g., if the base year is 2000 and base value is 100, then values at other time periods can be interpreted relative to the base period. If the value is 120 in 2003, then we can say that the index increased 20% from 2000 to 2003.
- We can easily change the base period using the following formula

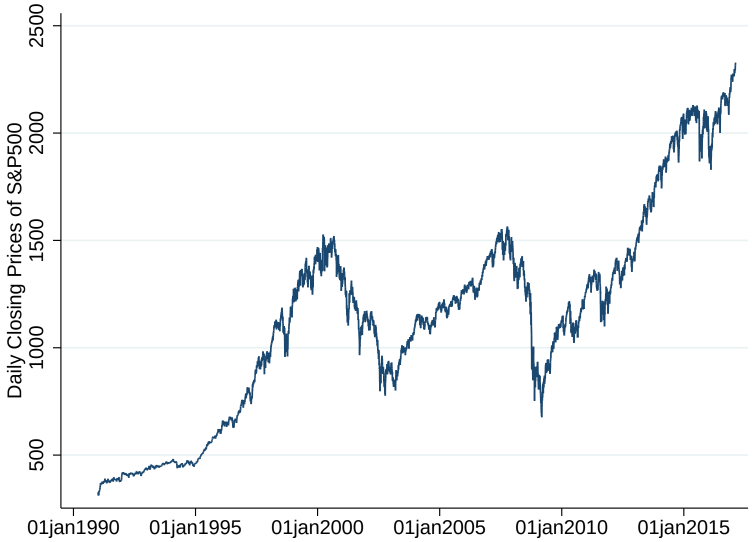
$$new\ index_t = 100 \times \frac{old\ index_t}{old\ index_{newbase}}$$

where $old\ index_{newbase}$ is the original value of the index in the new base year

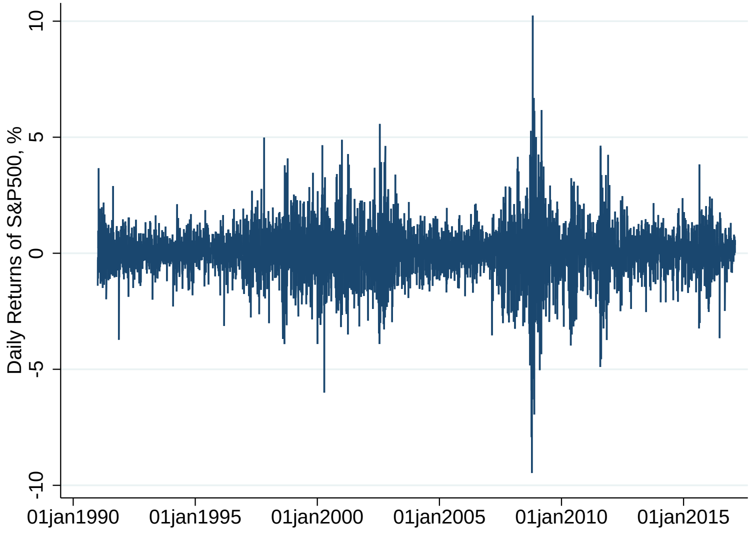
Twitter and Facebook Stock Prices



S&P500 Daily Closing Prices



S&P500 Daily Returns



Vostok (South Pole) Ice Core CO2 Data

- Layers of ice formed over long periods of time can be useful to analyze atmospheric conditions in the past.
- As falling snow becomes part of the ice it entraps chemicals and particulate matter in the air layer by layer.
- Researchers drill ice to examine these layers which contain information on the greenhouse gas concentration, temperature, among other things.
- Ice-cores are drilled at the research locations in the South Pole. These can be up to 3 kilometers long.
- The following data plots carbon dioxide (CO2) concentrations
- The time scale is measured in years Before Present (BP): 417160 - 2342 years BP. For more information visit: https://icecores.org/about-ice-cores and https://cdiac.ess-dive.lbl.gov/trends/co2/vostok.html

Ice-core Drilling

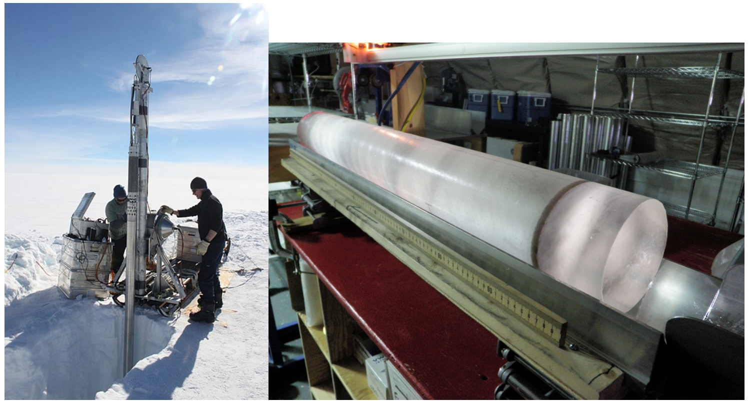
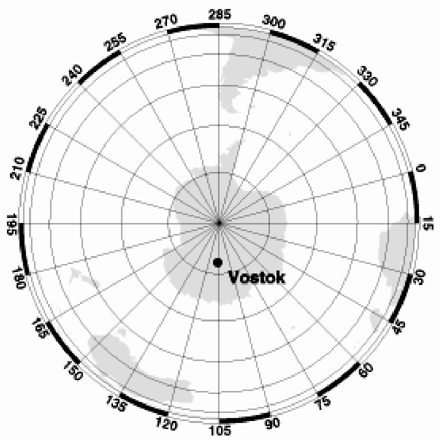


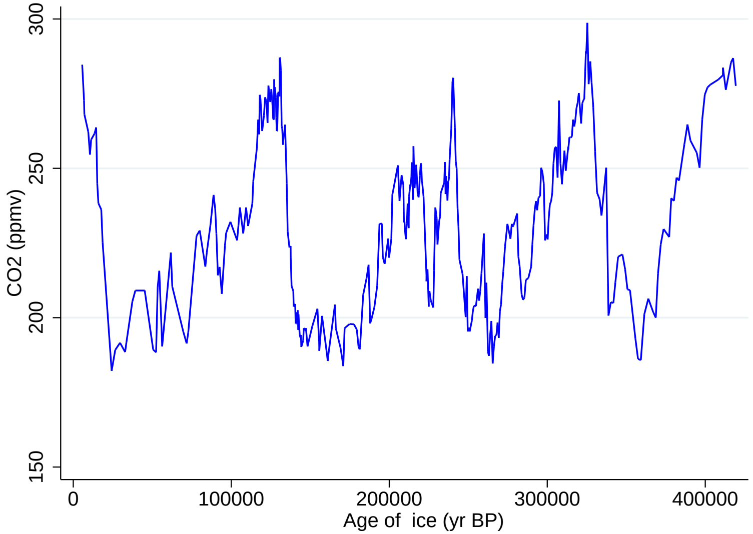
Figure sources: Wikipedia Commons and icecores.org.

Vostok Research Station



Vostok, Antarctica
78°28' S, 106°48'E
3488 m above MSL

Vostok (South Pole) Ice Core CO2 Data



Measuring Dependence in Time Series

- Time series variables tend to be dependent on its own past.
- The nature of this dependence can be useful in modeling time series data.
- Autocovariance and autocorrelation are widely used to measure dependence in time series.
- Let's start with the definition of the sample autocovariance.

Sample Autocovariance

Covariance and correlation coefficient measure the linear association between random variables X and Y. In time series analysis, we are particularly interested in the correlation between the value at time t and the value at previous time periods, say $s = t - h$. In other words covariance and correlation with itself:

Definition: Sample Autocovariance

$$\hat{\gamma}_h = \frac{1}{n} \sum_{t=h+1}^{T} (y_t - \bar{y})(y_{t-h} - \bar{y}) \tag{1}$$

where $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$ is the sample mean.

Sample Autocovariance

For example, $h = 1$ is called the first autocovariance

$$\hat{\gamma}_1 = \frac{1}{n} \sum_{t=2}^{T} (y_t - \bar{y})(y_{t-1} - \bar{y}) \tag{2}$$

, $h = 2$ is the second autocovariance, etc. What about $h = 0$? This is simply the sample variance:

$$\hat{\gamma}_0 = \widehat{\mathsf{Var}(y_t)} = \frac{1}{n} \sum_{t=1}^{T} (y_t - \bar{y})^2 \tag{3}$$

Measuring Dependence: Sample Autocorrelation

Autocorrelation measures the linear relationship between lagged values of a time series y_t. The sample autocorrelation is defined as

Definition: Sample Autocorrelation

$$\hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0} \tag{4}$$

Note that this definition is similar to the definition of correlation coefficient that you learned in statistics classes. It measures the strength of the relationship between y values that are h-period apart. $h = 1$ is called the first autocorrelation, $h = 2$ is the second autocorrelation, etc. By computing all autocorrelations up to a predetermined maximum lag order, we inspect the dependence structure of a time series. The plot is called the sample autocorrelation function (ACF or SACF, or correlogram).

ACF - Correlogram

- The graph (sample) autocorrelations for a set of lags $(h = 1, 2, \ldots, H)$ is known as (sample) correlogram.
- It can be used to display dependence structure in a time series.
- In large sample, using Central Limit Theorem

  $$\hat{\rho}_j \sim N\left(0, \frac{1}{n}\right)$$

  (Note: $\sqrt{n}\hat{\rho}_j \sim N(0, 1)$).
- 95% confidence interval around zero can be found by $\pm\frac{1.96}{\sqrt{n}}$.

White Noise Process

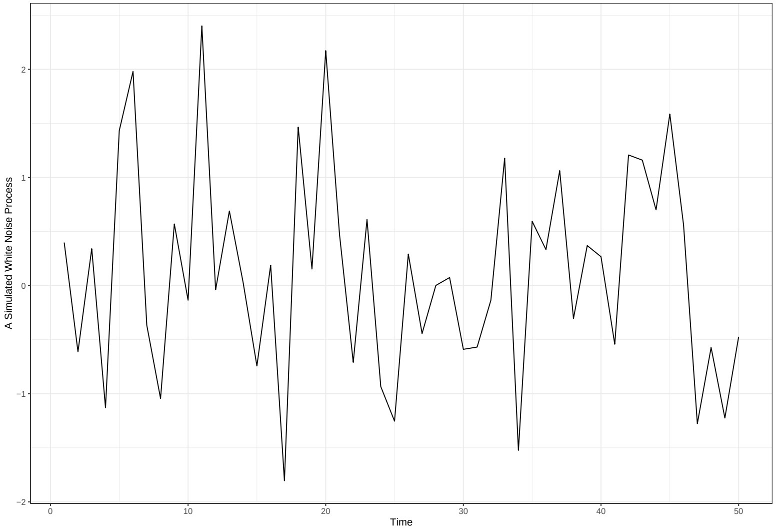
- As an example of a stochastic process, let us consider the "white noise" process. Here is the definition:
- If a stochastic process, $\{\epsilon_t : t = 1, 2, \ldots\}$, has the following properties

$$\begin{aligned} \mathsf{E}[\epsilon_t] &= 0 \quad (5) \\ \gamma_0 = \mathsf{Var}(\epsilon_t) &= \sigma^2 \quad (6) \\ \gamma_{t-s} = \mathsf{Cov}(\epsilon_t, \epsilon_s) &= 0, \quad t \neq s \quad (7) \end{aligned}$$

  Then this process is called the white noise process, denoted $\epsilon_t \sim wn(0, \sigma^2)$.
- Note the white noise process has a zero mean and a constant variance. Mean and variance do not depend on time. Additionally, current value, ϵ_t, does not depend on past values, as indicated by zero autocovariances for any time indices t and s with $t \neq s$

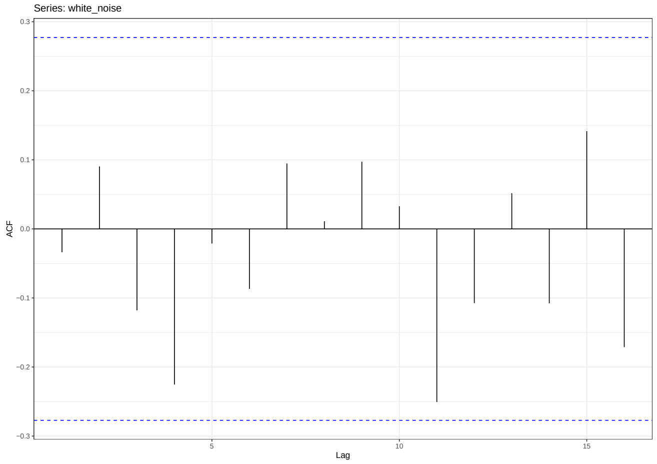
White Noise Process

Here is a simulated example of white noise process

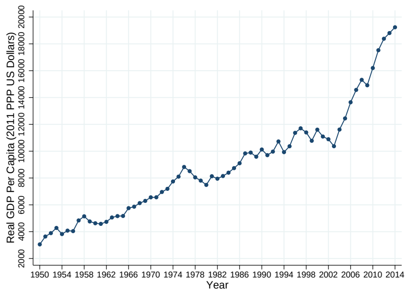


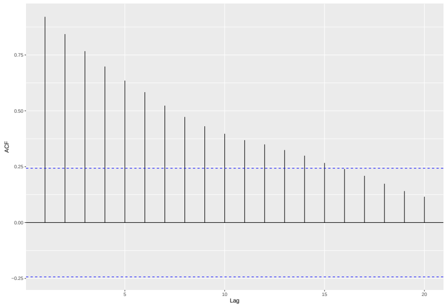
Sample ACF (Correlogram) of a White Noise Process

Note that 95% confidence interval around zero is shown by dotted lines ($\pm\frac{1.96}{\sqrt{n}}$) which contains zero.



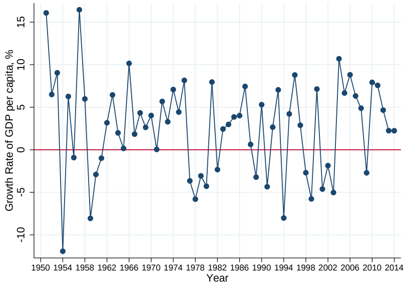
SACF of Turkish GDP pc

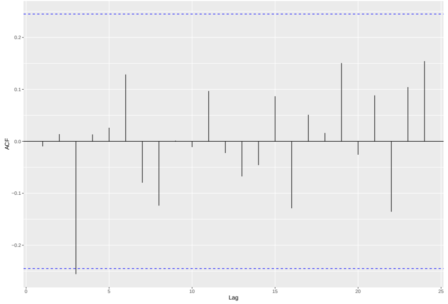




Note that when the series have a trend, as in the GDP graph above, SACF values at low lags tend to be large and positive but declining as lag order (distance) increases.

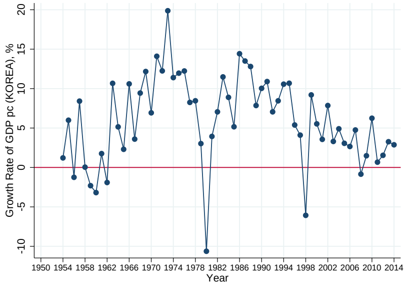
SACF of Turkish GDP pc GROWTH RATE

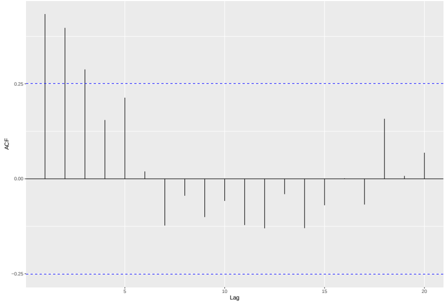




Autocorrelations of the GROWTH RATE of Turkish GDP pc are within the 95% confidence band. This implies that they are uncorrelated.

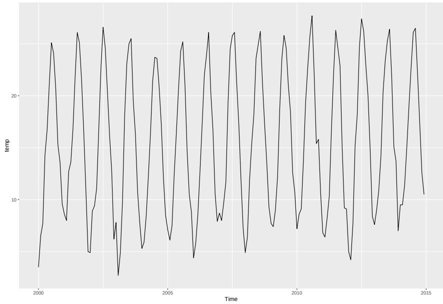
SACF of Korean GDP pc GROWTH RATE

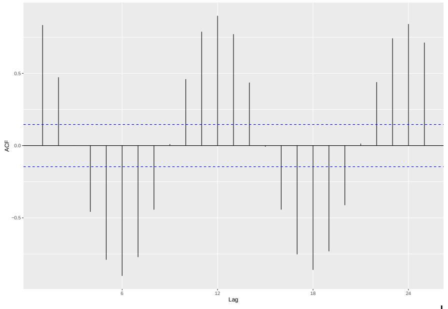




In contrast to Turkish case, Korean SACF values for the lag orders 1,2, 3 seem to be positive and significant.

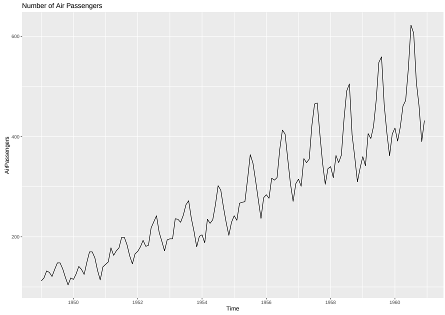
SACF of Monthly Temperature in Istanbul

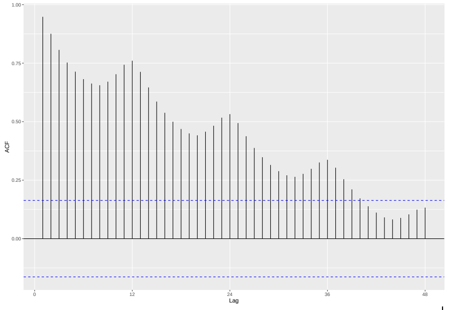




When the time series variable is seasonal (but no trend) then the ACF will have peaks at seasonal lags 12,24,etc, and troughs at 6, 18, etc.

Number of Airline Passengers (monthly)





The number of airline passengers have an increasing trend and seasonal variations. SACF has seasonal peaks at the multiples of 12. SACF decreases as lag order increases (due to increasing trend).

Classical Decomposition of Time Series

Additive Decomposition

$$y_t = Trend_t + Seasonal_t + Irregular_t$$

Multiplicative Decomposition

$$y_t = Trend_t \times Seasonal_t \times Irregular_t$$

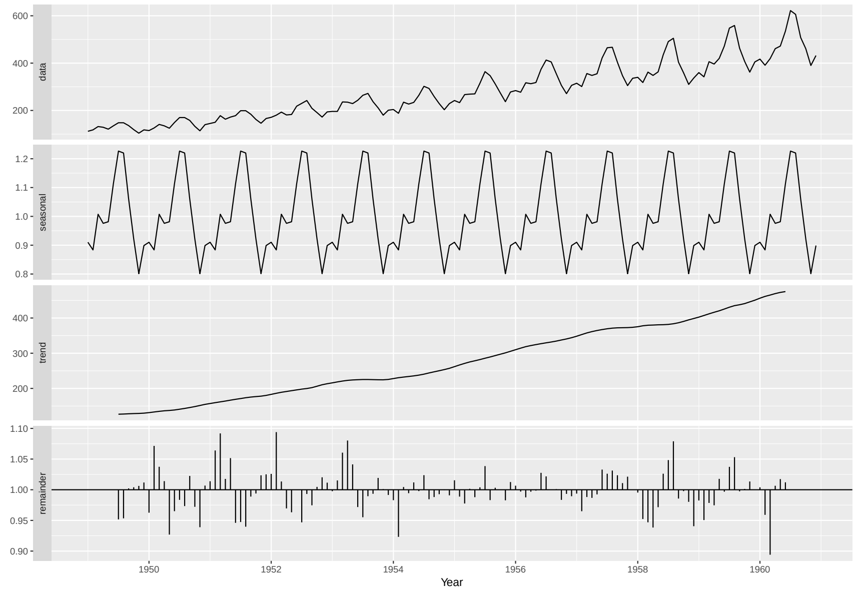
or

$$\log(y_t) = \log(Trend_t) + \log(Seasonal_t) + \log(Irregular_t)$$

Classical Decomposition of Time Series

- $Trend_t$ component is the trend-cycle component (slow-moving long-run and medium-run components including business cycles),
- $Seasonal_t$ component reflects fluctuations that generally repeats during the same time each year or month (or even day, hour, etc.)
- $Irregular_t$ is the remainder component.
- If the seasonal variations around the trend-cycle component remain more or less stable then the additive component may be appropriate.
- Otherwise, the multiplicative component should be used. Take a look at the time series graph of airline passengers in the next slide. As time passes, the variability in the seasonal variations increase. Therefore, a multiplicative decomposition, or, log-additive decomposition would be more appropriate.

Classical Decomposition of Airline Passengers



Examples of Time Series Models

Now we discuss some examples of time series model useful in empirical time series analysis estimated by OLS.

Static Model

Suppose that we have time series data available on two variables, say y and z dated contemporaneously. A static model relating y to z is

$$y_t = \beta_0 + \beta_1 z_t + u_t \quad t = 1, 2, ..., n$$

A static model can also be postulated in first-differences:

$$\Delta y_t = \beta_1 \Delta z_t + \Delta u_t \quad t = 1, 2, ..., n$$

Static Phillips Curve

- The static Phillips curve is an example of the static model, given by

Static Phillips Curve

$$inf_t = \beta_0 + \beta_1 unemp_t + u_t$$

- inf_t: inflation rate and $unemp_t$: unemployment rate
- This form of the Phillips curve assumes a constant **natural rate of unemployment** and constant **inflationary expectations**.
- It can be used to study the **contemporaneous tradeoff** between inf_t and $unemp_t$.

Finite Distributed Lag Models, FDL models

- In a finite distributed lag model (FDL model) we allow one or more variable to affect y with a lag. For example, for annual observations, consider the model

Effect of Tax Exemption on Fertility

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t$$

- gfr_t: Fertility rate, children born per 1000 women.
- pe_t: The real dollar value of the personal tax exemption (some kind of incentive to have a child).

Finite Distributed Lag Models, FDL models

- The fertility rate may depend on on the tax value of a child, the effect may have a lag. The following model is an FDL of order two.

FDL model of order two

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$$

- suppose that z is a constant, equal to c, in all time periods before time t. At time t, z increases by one unit to $c + 1$ and then reverts to its previous level at time $t + 1$. (the increase in z is temporary.)

  $$..., z_{t-2} = c, \quad z_{t-1} = c, \quad z_t = c + 1, \quad z_{t+1} = c \quad z_{t+2} = c, ...$$

- To focus on the ceteris paribus effect of z on y, we set the error term in each time period to zero. The ceteris paribus effect on y is called the **impact multiplier or impact propensity**.

Calculating Impact Multiplier

- Calculating the impact multiplier of FDL model of order 2:

$$y_{t-1} = \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,$$

$$y_t = \alpha_0 + \delta_0 (c + 1) + \delta_1 c + \delta_2 c,$$

$$y_{t+1} = \alpha_0 + \delta_0 c + \delta_1 (c + 1) + \delta_2 c,$$

$$y_{t+2} = \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 (c + 1),$$

$$y_{t+3} = \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,$$

- From the first two equations, $y_t - y_{t-1} = \delta_0$.
- δ_0 is the immediate change in y due to the one-unit increase in z at time t.
- δ_0 is usually called the impact propensity or impact multiplier.

Calculating Impact Multiplier

- Similarly, $y_{t+1} - y_{t-1} = \delta_1$ is the change in y one period after the temporary change,
- And $y_{t+2} - y_{t-1} = \delta_2$ is the change in y two periods after the change.
- At time $t+3$, y has reverted back to its initial level: $y_{t+3} = y_{t-1}$
- This is because we have assumed that only two lags of z appear in the FDL model of order 2.
- When we graph the δ_j as a function of j, we obtain the **lag distribution**,which summarizes the dynamic effect that a temporary increase in z has on y

Long Run Propensity, LRP

- We are also interested in the change in y due to a permanent increase in z.
- Before time t, z equals the constant c. At time t, z increases permanently to $c+1$.
- Again, setting the errors to zero, we have

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\
y_t &= \alpha_0 + \delta_0(c+1) + \delta_1 c + \delta_2 c, \\
y_{t+1} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2 c, \\
y_{t+2} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2(c+1),
\end{aligned}$$

Long Run Propensity, LRP

- With the permanent increase in z, after one period, y has increased by $\delta_0 + \delta_1$, and after two periods, y has increased by $\delta_0 + \delta_1 + \delta_2$. There are no further changes in y after two periods.
- This shows that the sum of the coefficients on current and lagged z,$\delta_0 + \delta_1 + \delta_2$, is the long-run change in y given a permanent increase in z. and is called the long-run propensity (LRP) or long-run multiplier.
- This effect is called the **long run multiplier or long run propensity**.
- The LRP is often of interest in distributed lag models.

Long Run Propensity, LRP

Effect of Tax Exemption on Fertility

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t$$

- In this model, δ_0, measures the immediate change in fertility due to a one-dollar increase in pe. There are reasons to believe that δ_0, is very small or zero.
- But δ_1 and δ_2 or both, might be positive. If pe permanently increases by one dollar, then, after two years, gfr will have changed by $\delta_0 + \delta_1 + \delta_2$.
- This model assumes that there are no further changes after two years. Whether or not, this is actually the case is an empirical matter.

FDL model of order q

- A finite distributed lag model of order q is written as

FDL(q)

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + ... + \delta_q z_{t-q} + u_t$$

- The FDL models is useful for testing whether there are lagged effects of independent variable z on dependent variable y.
- The impact propensity is always the coefficient on the contemporaneous z_t, δ_0.
- The long-run propensity, LRP is the sum of all coefficients of z_{t-j}.

$$LRP = \delta_0 + \delta_1 + ... + \delta_q$$

FDL model of order q, FDL(q)

- Because of the often substantial correlation in z at different lags due to multicollinearity, it can be difficult to obtain precise estimates of the individual δ_j. Interestingly, even when the δ_j cannot be precisely estimated, we can often get good estimates of the LRP.
- We can have more than one explanatory variable appearing with lags, or we can add contemporaneous variables to an FDL model.
- **Question:** In an equation for annual data, suppose that

FDL(2)

$$\widehat{\mathrm{int}_\mathrm{t}} = 1.6 + 0.48inf_t - 0.15inf_{t-1} + 0.32inf_{t-2}$$

- where int is an interest rate and inf is the inflation rate, what are the impact and long-run propensities?