Econometrics: Time Series

Diego López Tamayo * Based on MOOC by Erasmus University Rotterdam

Contents

"There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

*El Colegio de México, diego.lopez@colmex.mx

Time series

What is a time serie.

Look at Dates and Times in R Without Losing Your Sanity to understand how to use correctly date labes in R. Datasets to be used:

```
revenue <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset61.csv")
revenue$YEAR<- as.Date(paste0(revenue$YEAR, '-01-01'))

production <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset62.csv")
# We replace the weird "M" before months.
production <- rename(production, date=`YYYY-MM`)
production$date <- gsub("M","-",production$date)
production$date <- as.Date(as.yearmon(production$date))

dataset_training <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer61.csv")
```

Time series data are a specific type of data that need a somewhat special treatment when using econometric methods. The specific aspect of time series variables is that they are sequentially observed. That is, one observation follows after another. The sequential nature of time series observations has important implications for modeling and especially for forecasting and this is different from the cross-sectional data that we have mostly looked at so far.

Think of the shoe size of your next-door neighbor. Now, it is quite unlikely that the very fact that someone lives next to you, implies that this person's shoe size has predictive value for yours. But with time series data, this is different.

Yesterday's sales level, likely has predictive value for today's sales level. Just like last month's inflation has for current inflation and your last year's disposable income for this year's.

A time series variable is observed at a **regular frequency.** This can be once per year, once per month, every day and sometimes, like in some areas of finance, even each millisecond. You can imagine that recent observations on a certain time series variable can have predictive value for future observations. If it is winter-like weather today, it will most likely be so tomorrow. When unemployment is high this month, it probably is still going to be high next month.

So in terms of regression models, you may want to include the past of a variable in order to predict its future. That is, to predict a new observation of y, you can use another variable X, but you can also think of using y one period lagged. $y(-1)$.

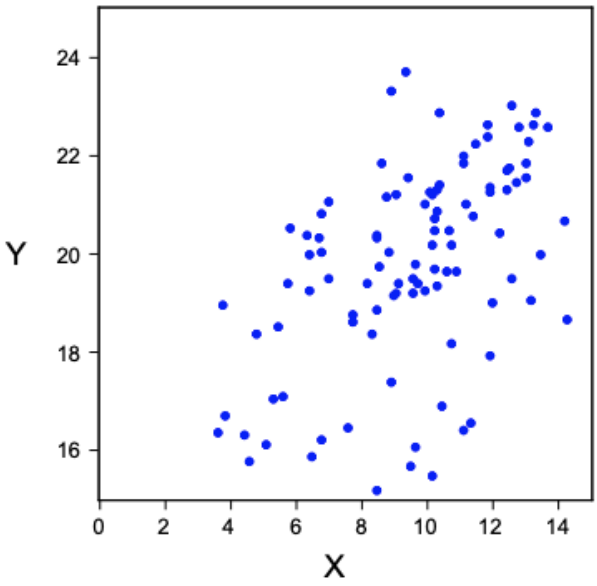
The inclusion of lagged values of the dependent variable in your regression model can also **prevent you from drawing spurious conclusions.** That is, you might think that another variable X helps to predict the variable of interest Y, while in reality, Y one period lagged predicts Y and X is irrelevant.

To illustrate this point consider two variables, X and Y, for which we know that the true data generating process is such that they depend with a factor 0.9 on their own previous value, whereas the variables X and Y are completely uncorrelated.

$$x_t = 1 + 0.9x_{t-1} + \epsilon_{x,t} \text{ and } y_t = 2 + 0.9y_{t-1} + \epsilon_{y,t}$$

$$\text{Two series completely uncorrelated } E(\epsilon_{y,t}, \epsilon_{x,s}) = 0 \forall t, s$$

A scatter of simulated Y and X variables with 100 observations may look like this.



Note that there seems to be some positive connection between the two, while we know that they are completely uncorrelated. You could be tempted to fit a simple regression model. Now suppose you would do so.

Dependent variable: Y (sample size $n = 100$)						
	Coef.	t-Stat.	p-value	Coef.	t-Stat.	p-value
Constant	15.99	23.45	0.000	2.91	2.87	0.005
X	0.40	5.78	0.000	0.07	1.53	0.129
Y(-1)	-	-	-	0.82	14.01	0.000
R-squared	0.254			0.753		

At the left-hand side of this table, you see that we estimate the slope parameter to be equal to 0.4 with a p-value of 0.000. So, this suggests that X has predictive value for Y. Now we know of course, this cannot be true given the way we created the data. The right-hand panel of the table shows what happens if we also include the Y variable one period lagged. The coefficient for this lagged variable is 0.82 and it is significant, whereas the coefficient of X is close to 0 and not statistically significant anymore.

You may now wonder whether we should have included not only X, but also X one period lagged. Consider the regression model where Y depends on Y one period lagged, X and also X one period lagged. Do X and its lag have any predictive power?

Dependent variable: Y (sample size $n = 100$)						
	Coef.	t-Stat.	p-value	Coef.	t-Stat.	p-value
Constant	2.88	2.83	0.006	2.69	2.66	0.009
Y(-1)	0.83	14.02	0.000	0.86	17.03	0.000
X	0.15	1.61	0.110	-	-	-
X(-1)	-0.09	-0.99	0.324	-	-	-
R-squared	0.756			0.747		

- Use F-test $F = \frac{(R_1^2 - R_0^2)/g}{(1-R_1^2)/(n-k)} \sim F_{(g,n-k)}$
- Number of restrictions: $g = 2$
- number of observations: $n = 100$
- number of parameters unrestricted model: $k = 4$
- values of R-squared: Unrestricted: $R_1^2 = 0.756$ and Restricted: $R_0^2 = 0.747$
- Substitute these values in formula for F-test: $F = 1.8 < 3.1$

- Joint effect of X and $X(-1)$ on Y is not significant

The larger model contains two extra variables, so the number of restrictions is two. We have 100 observations and the full model has 4 variables. The two R-squared values were reported in the table. Substituting these values in the familiar expression for the F-test gives a value of 1.8, which is smaller than the 5% critical value of 3.1. So even when we include X and one period lagged X, then these variables do not help to predict Y. Recall that the scatter of Y versus X was very suggestive, but proper analysis shows that pictures can sometimes fool us.

Example Airline revenue

Dataset: revenue

Simulated data set on yearly revenue passenger kilometers, 1975-2015 (estimation period 1976-2015, with pre-sample value for 1975).

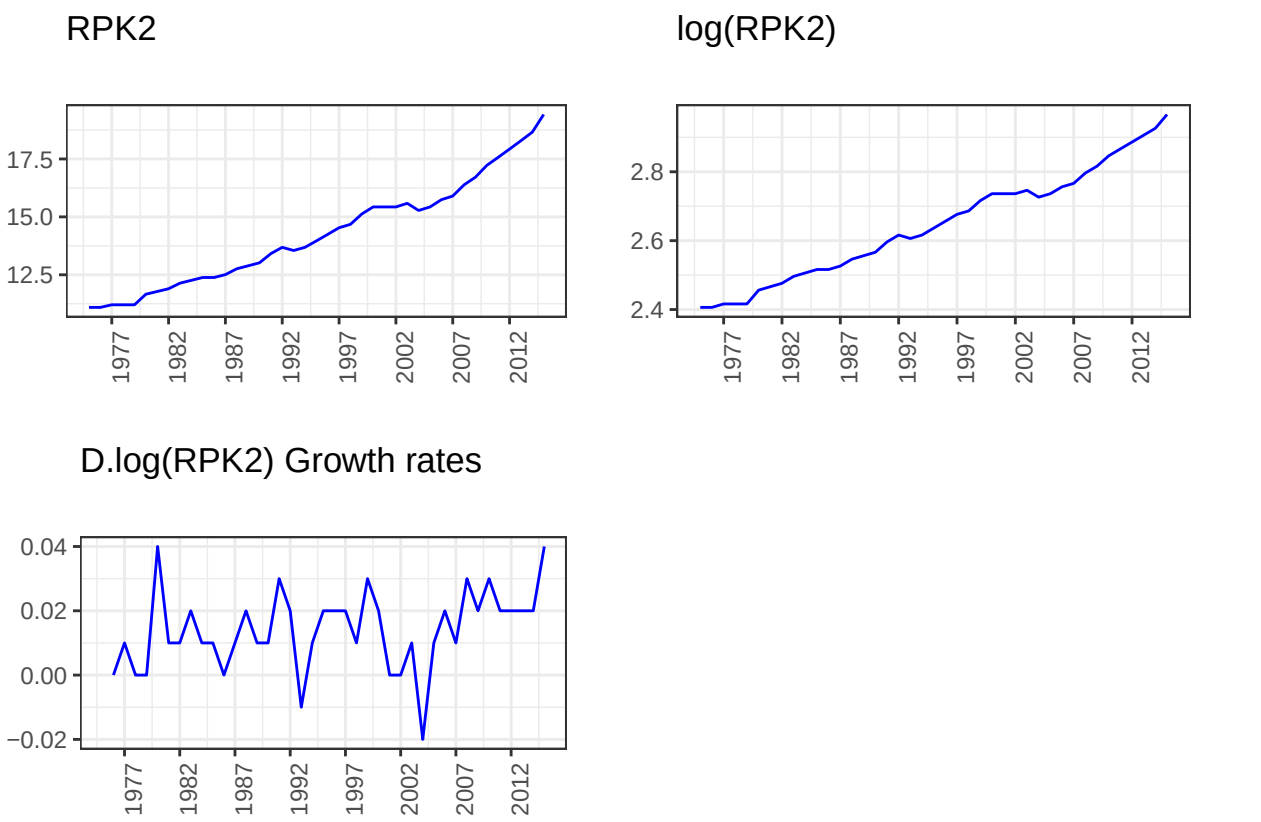
- RPK1: Revenue Passenger Kilometers of company 1 (1975-2015)
- RPK2: Revenue Passenger Kilometers of company 2 (1975-2015)
- X1: log(RPK1) (1975-2015)
- X2: log(RPK2) (1975-2015)
- DX1: first difference of X1, growth rate of RPK1 (1976-2015)
- DX2: first difference of X2, growth rate of RPK2 (1976-2015)
- Year: calendar year

Let us now look at how time series in economics and business can look like. Here is an example of passenger revenue data for an airline. The variable of interest is revenue passenger kilometers, which is the sum total over one year of the distance in kilometers traveled by each passenger on each flight of this airline company.

The left-hand graph gives the actual total number of kilometers traveled. The middle graph is obtained when taking natural logs and the right-hand graph shows the yearly growth rates.

```
# We create the log and the growth rate of both series
revenue <- revenue %>% mutate(X1=log(RPK1),DX1=c(NA,diff(log(RPK1))),X2=log(RPK2),DX2=c(NA,diff(log(RPK

plot_a <- ggplot(data=revenue, aes(x=YEAR)) +
  geom_line(aes(y=RPK2),col="blue") +
  labs(x = "", y = "", title = "RPK2",
       subtitle = ("")) +
    scale_x_date(date_breaks = "5 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_b <- ggplot(data=revenue, aes(x=YEAR)) +
  geom_line(aes(y=X2),col="blue") +
  labs(x = "", y = "", title = "log(RPK2)",
       subtitle = ("")) +
    scale_x_date(date_breaks = "5 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_c <- ggplot(data=revenue, aes(x=YEAR)) +
```

```
  geom_line(aes(y=DX2),col="blue") +
  labs(x = "", y = "", title = "D.log(RPK2) Growth rates",
       subtitle = ("")) +
    scale_x_date(date_breaks = "5 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.5, .20),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

grid.arrange(plot_a, plot_b, plot_c, nrow = 2)
```



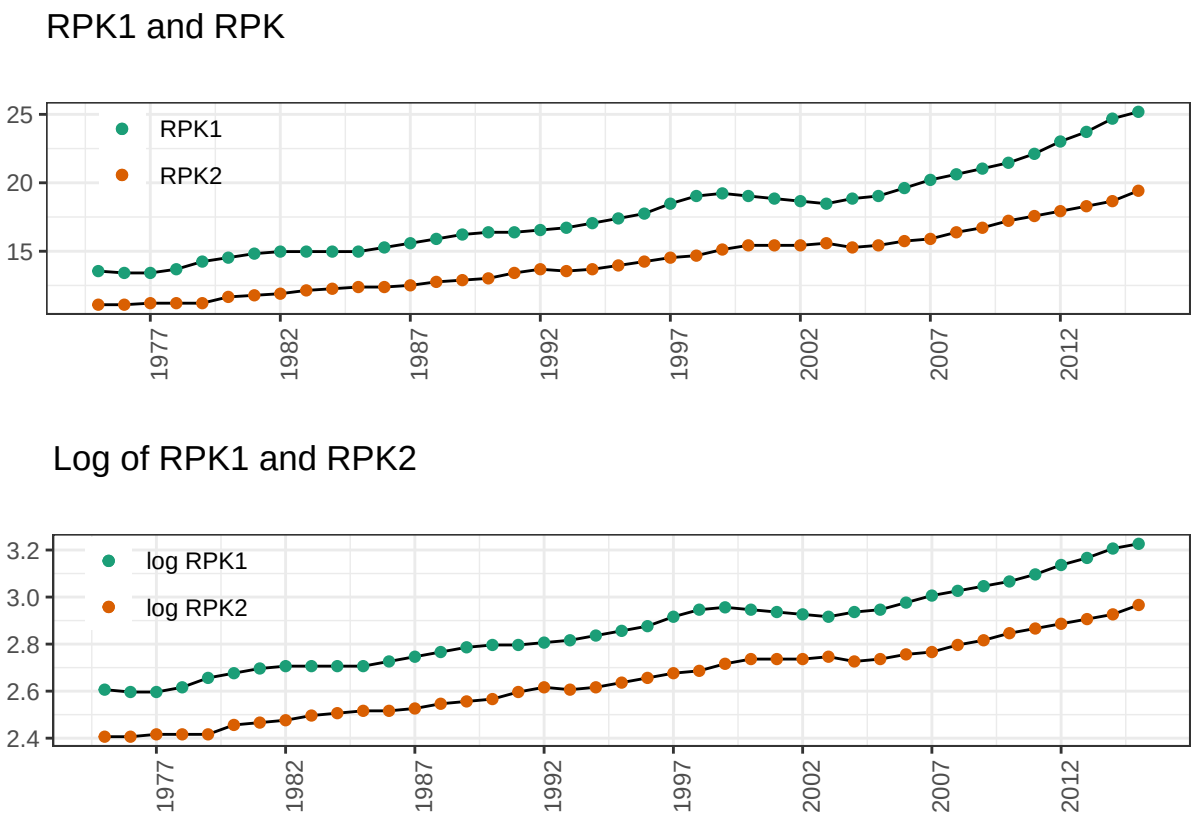
- RPK: Revenue Passenger Kilometers (in billions) yearly totals 1976-2015, trend somewhat exponential
- $log(RPK)$: more linear trend
- $D.log(RPK) = log(RPK_t) - log(RPK_{t-1}) \approx \frac{RPK_t - RPK_{t-1}}{RPK_{t-1}}$ yearly growth rate of RPK

The raw data on the left seems somewhat exponentially increasing, whereas the trend for the log of the time series seems more linear. The yearly growth rates fluctuate between minus 2% and plus 4%. The two leftmost graphs show that the data have a pronounced upward trend. When this occurs, it is not reasonable to assume that the mean of the data is constant over time. In fact, the mean increases with each new observation.

In the next section, we will deal with this important issue in more detail, as for proper statistical analysis, we need data with constant mean. **A constant mean is one aspect of what we call stationarity.** For a stationary time series like in the $D.log(RPK1)$ graph here, we have a straightforward modeling strategy. But for non-stationary time series, we will first need to get rid of this non-stationarity.

This issue of trends is even more important when two time series show similar trending behavior. Look at this graph that depicts the revenue passenger kilometers of two airlines.

```
plot_d <- ggplot(data=revenue, aes(x=YEAR)) +
  geom_line(aes(y=RPK2)) + geom_point(aes(y=RPK2,col="RPK2")) +
  geom_line(aes(y=RPK1)) + geom_point(aes(y=RPK1,col="RPK1")) +
  labs(x = "", y = "", title = "RPK1 and RPK",
       subtitle = ("")) +
    scale_x_date(date_breaks = "5 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

plot_e <- ggplot(data=revenue, aes(x=YEAR)) +
  geom_line(aes(y=X2)) + geom_point(aes(y=X2,col="log RPK2")) +
  geom_line(aes(y=X1)) + geom_point(aes(y=X1,col="log RPK1")) +
  labs(x = "", y = "", title = "Log of RPK1 and RPK2",
       subtitle = ("")) +
    scale_x_date(date_breaks = "5 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")

grid.arrange(plot_d, plot_e, nrow = 2)
```



Clearly, they seem to have the same trend, especially when you take logs. This feature can be useful for forecasting in the following way. You may use both time series to estimate the common trend, then you can

forecast the trend. And finally, derive the individual forecast for each of the airlines. In case of a single or univariate time series, you can use its own past to make forecasts. When you have several or multivariate time series like in this example, you can try to use the other series to improve your forecasts.

Example Industrial Production

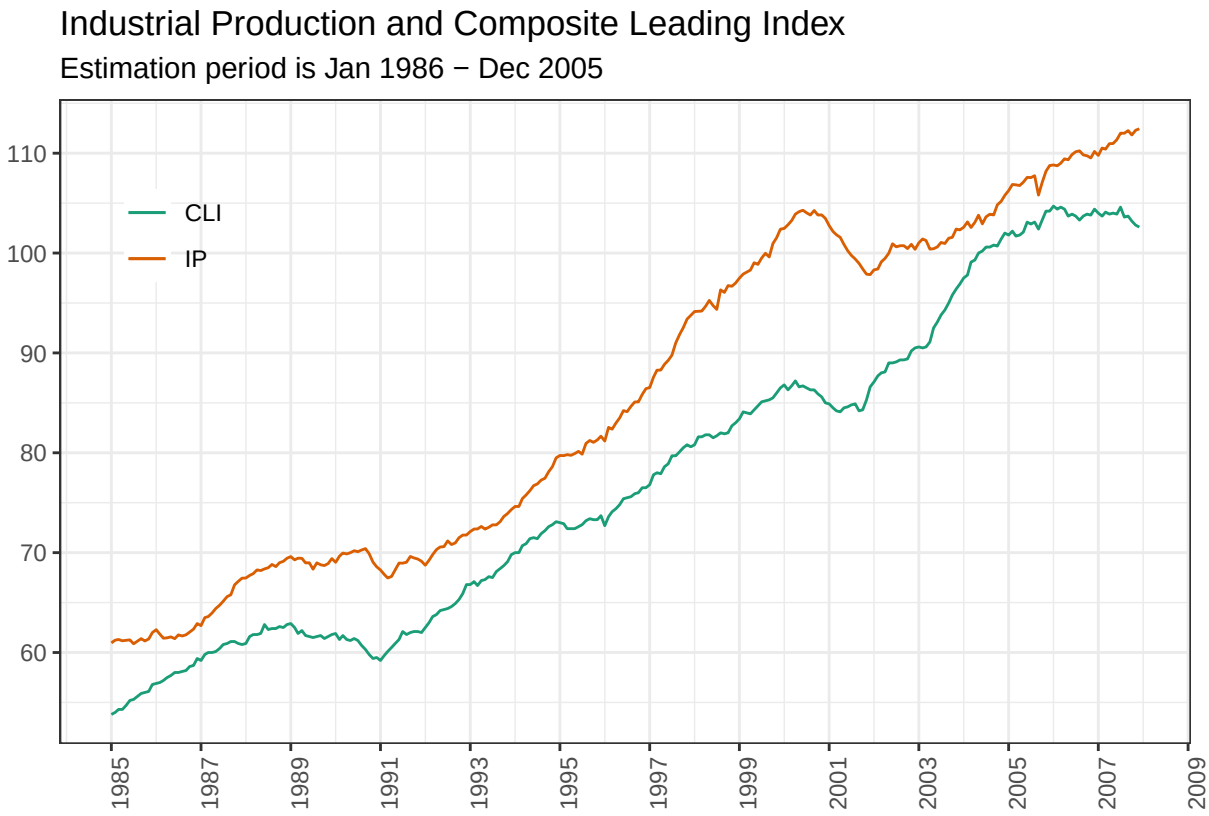
Dataset: production

Data set on Industrial Production and the Composite Leading Index for the USA, monthly data Jan 1985 - Dec 2007 (Source: Conference Board, USA). Estimation period is Jan 1986 - Dec 2005 (pre-sample values in 1985). Forecast evaluation period is Jan 2006 - Dec 2007.

- CLI: Composite Leading Index (based on 10 leading indicators)
- IP: Industrial Production (index, seasonally adjusted)
- LOGCLI: logarithm of CLI
- LOGIP: logarithm of IP
- GRCLI: monthly growthrate of CLI, first difference of LOGCLI
- GRIP: monthly growthrate of IP, first difference of LOGIP

Here is another pair of time series that are clearly related over time. These are the monthly industrial production index for the United States of America and the so-called **composite leading indicator or CLI.**

```
plot_f <- ggplot(data=production, aes(x=date)) +
  geom_line(aes(y=IP,col="IP")) +
  geom_line(aes(y=CLI,col="CLI")) +
  labs(x = "", y = "", title = "Industrial Production and Composite Leading Index ",
       subtitle = ("Estimation period is Jan 1986 - Dec 2005")) +
    scale_x_date(date_breaks = "2 year", date_labels = "%Y") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = c(.1, .8),
        legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot_f
```



The CLI is constructed by The Conference Board based on a set of ten variables like manufacturer's new orders, stock prices and consumer expectations. All these variables are forward looking. And therefore, they are believed to have predictive value for future macroeconomic developments. And for that reason, it may be useful to consider the CLI in case you want to forecast a variable like industrial production.

As with the airlines, the trends in industrial production and the Composite Leading Index seem to follow a similar pattern, which here associates with the business cycle. In our last section on time series, you will see if industrial production can indeed be predicted by means of this index.

Example spurious regression

Dataset: dataset_training

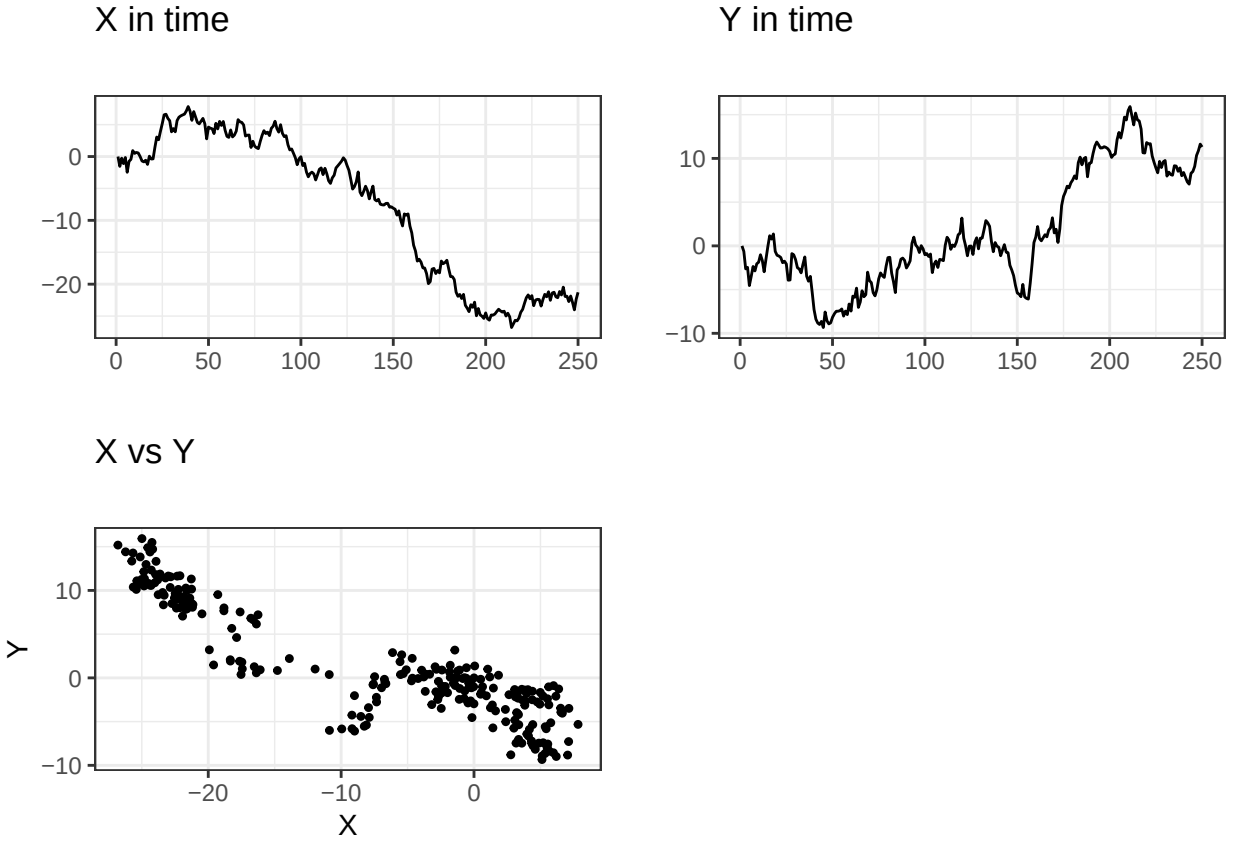
- epsx: sample of 250 values from normally and independently distributed white noise with mean 0 and variance 1 (independent of ϵ_{yt})
- epsy: sample of 250 values from normally and independently distributed white noise with mean 0 and variance 1 (independent of ϵ_{xt})
- x: random walk generated from epsx: $x_1 = 0$, and $x_t = x_{t-1} + \epsilon_{xt}$
- y: random walk generated from epsy: $y_1 = 0$, and $y_t = y_{t-1} + \epsilon_{yt}$

The datafile contains values of four series of length 250. Two of these series are uncorrelated white noise series denoted by $\epsilon_{x,t}$ and $\epsilon_{y,t}$ where both variables are $NID(0,1)$ and $E(\epsilon_{y,t}, \epsilon_{x,s}) = 0 \forall t, s$. The other two series are so-called random walks constructed from these two white noise series by $x_t = x_{t-1} + \epsilon_{xt}$ and $y_t = y_{t-1} + \epsilon_{yt}$.

As ϵ_{xt} and ϵ_{yt} are independent for all values of t and s, the same holds true for all values of x_t and y_t The purpose of this exercise is to experience that, nonetheless, the regression of y on x indicates a highly significant relation between y and x if evaluated by standard regression tools. This kind of result is called **spurious regression** and is caused by the trending nature of the variables x and y.

- a) Graph the time series plot of x_t against time t, the time series plot of y_t against time t, and the scatter plot of y_t against x_t. What conclusion could you draw from these three graphs?

```
# We add an index column to the dataset for the time t
dataset_training <- dataset_training %>% mutate(time = row_number())
plot_1 <- ggplot(data=dataset_training, aes(x=time)) +
  geom_line(aes(y=X)) +
  labs(x = "", y = "", title = "X in time",
       subtitle = ("")) +
  theme_bw() +
  theme(legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot_2 <- ggplot(data=dataset_training, aes(x=time)) +
  geom_line(aes(y=Y)) +
  labs(x = "", y = "", title = "Y in time",
       subtitle = ("")) +
  theme_bw() +
  theme(legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot_3 <- ggplot(data=dataset_training, aes(x=X,y=Y)) +
  geom_point(shape=20) +
  labs(x = "X", y = "Y", title = "X vs Y",
       subtitle = ("")) +
  theme_bw() +
  theme(legend.background = element_rect(fill = "transparent")) +
  scale_color_brewer(name= NULL, palette = "Dark2")
grid.arrange(plot_1, plot_2, plot_3, nrow = 2)
```



The two variables X,Y have completly random movements up and down. And the scatter plot seems to have a negative relation, so we could use X to forecast Y, but we know that this is not the case, the scatterplot is **misleading** in this sense.

- b) To check that the series ϵ_{xt} and ϵ_{yt} are uncorrelated, regress ϵ_{yt} on a constant and ϵ_{xt}. Report the t-value and p-value of the slope coefficient.

We use the summ() function to output our regression.

```
lm1 <- lm(EPSY ~ EPSX, data=dataset_training)
summ(lm1, digits = 3)
```

Observations	250
Dependent variable	EPSY
Type	OLS linear regression

F(1,248)	1.736
R^2	0.007
Adj. R^2	0.003

	Est.	S.E.	t val.	p
(Intercept)	0.031	0.064	0.484	0.629
EPSX	-0.088	0.067	-1.318	0.189

Standard errors: OLS

The t-value of the coefficient is around -1.32 and the p-value around 0.19, this shows that ϵ_{xt} and ϵ_{yt} have no significant relation.

- c) Extend the analysis of part (b) by regressing ϵ_{yt} on a constant, ϵ_{xt}, and three lagged values of ϵ_{yt} and of ϵ_{xt}. Perform the F-test for the joint insignificance of the seven parameters of ϵ_{xt} and the three lags of ϵ_{xt} and ϵ_{yt}. Report the degrees of freedom of the F-test and the numerical outcome of this test, and draw your conclusion. Note: The relevant 5% critical value is 2.0.

```
lm2 <- lm(EPSY ~ lag(EPSY,1) + lag(EPSY,2) + lag(EPSY,3) + EPSX + lag(EPSX,1) + lag(EPSX,2) + lag(EPSX,3
summ(lm2, digits = 3)
```

Observations	247 (3 missing obs. deleted)
Dependent variable	EPSY
Type	OLS linear regression

F(7,239)	0.546
R^2	0.016
Adj. R^2	-0.013

The $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = 0$ and the degrees of freedom of F test $df = (g, n-k)$ where g is the number of parameter restrictions on the null, n is the number of observations and k is the number of variables in the unrestricted model. In this case we have:

- $g = 7$ All 7 restrictions equal to 0.
- $n = 247$ Because 3 observarions are lost because the 3 lag values, so the first available observation is in $t = 4$.
- $k = 8$ Due to the 7 coefficient plus the constant term.

	Est.	S.E.	t val.	p
(Intercept)	0.046	0.066	0.698	0.486
lag(EPSY, 1)	0.025	0.064	0.387	0.699
lag(EPSY, 2)	-0.016	0.065	-0.244	0.807
lag(EPSY, 3)	-0.047	0.064	-0.734	0.464
EPSX	-0.097	0.069	-1.405	0.161
lag(EPSX, 1)	0.020	0.070	0.284	0.777
lag(EPSX, 2)	-0.060	0.070	-0.857	0.392
lag(EPSX, 3)	0.009	0.068	0.138	0.890

Standard errors: OLS

We can see the F-statistic at the top of the output or calculate it by hand $F = \frac{(R_1^2 - R_0^2)/g}{(1-R_1^2)/(n-k)} \sim F_{(g,n-k)}$ where $R_0^2 = 0$ because is a model with only a constant term. $F = 0.55$ and as is smaller of the critical value of 2. We do NOT reject the H_0. This is correct as the value of ϵ_{yt} is independent of all other observations.

- d) Regress y on a constant and x. Report the t-value and p-value of the slope coefficient. What conclusion would you be tempted to draw if you did not know how the data were generated?

```
lm3 <- lm(Y ~ X, data=dataset_training)
summ(lm3, digits = 3)
```

Observations	250
Dependent variable	Y
Type	OLS linear regression

F(1,248)	1090.611
R^2	0.815
Adj. R^2	0.814

	Est.	S.E.	t val.	p
(Intercept)	-2.487	0.214	-11.606	0.000
X	-0.515	0.016	-33.024	0.000

Standard errors: OLS

It seems by looking at the large t-value of X that X has a relevant explanatory power over Y. We know that this is not the case, so the regression is misleading, due to the trending nature of both variables. Look again at the scatterplot of a), it happens that X moves downward for long periods as Y moves upwards for long periods. This is why it seems to be a negative relation.

- e) Let e_t be the residuals of the regression of part (d). Regress e_t on a constant and the one-period lagged residual e_{t-1}. What standard assumption of regression is clearly violated for the regression in part (d)?

```
# We add the residuals of lm3 into the dataset
dataset_training <- dataset_training %>% mutate(lm3.res = resid(lm3))
lm4 <- lm(lm3.res ~ lag(lm3.res,1), data=dataset_training)
summ(lm4, digits = 3)
```

This coefficient is significant at 99%, this shows that the residuals are very strongly correlated. Therefore violates the standar regression assumption A7 that the error terms should be uncorrelated.

Observations	249 (1 missing obs. deleted)
Dependent variable	lm3.res
Type	OLS linear regression

F(1,247)	1457.056
R^2	0.855
Adj. R^2	0.854

	Est.	S.E.	t val.	p
(Intercept)	0.001	0.067	0.008	0.993
lag(lm3.res, 1)	0.925	0.024	38.171	0.000

Standard errors: OLS

Representing time series