VAR, Error Correction Models and ARIMAX

LECTURE OUTLINE

Last time we studied the following topics:

- ARMA
- ARIMA
- Impulse Response Functions

In this lecture we will study:

- Vector AutoRegression
- Error Correction Models
- ARIMAX

VAR:
Vector Auto-Regression

Simo Alami

VAR Principle

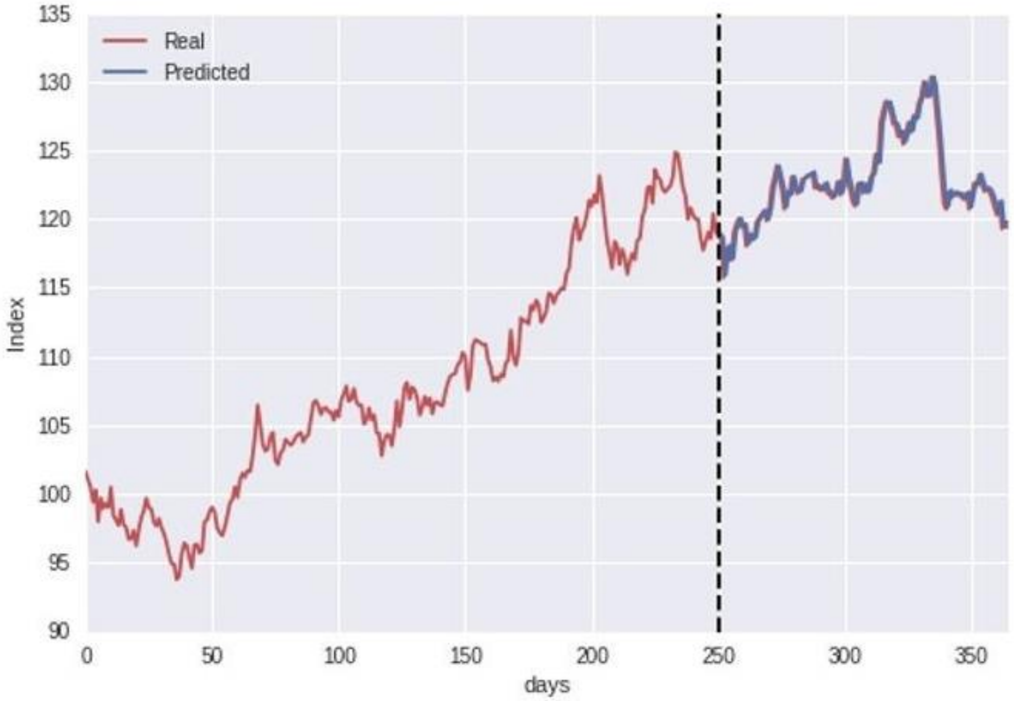
All the models we have seen before have one common and pretty big limitation:

→ They can only use the past data of the series itself

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}$$

→ You might need data other than past data:

- Stock prediction: Sentiment analysis, data from other stocks…
- Energy consumption: weather, time of the day, date…
- Sales: macro-economic data…



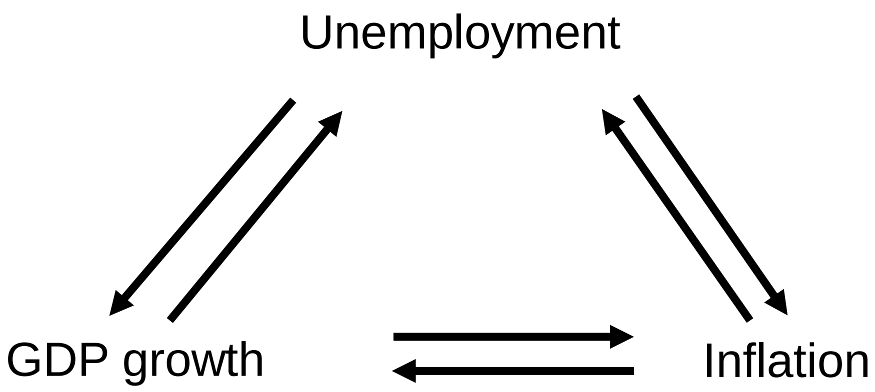
A Vector Autoregression (VAR) model is particularly useful when you have multiple time series that influence each other, and you want to model these relationships.

Example

A VAR model considers the joint dynamics of time series by allowing each variable to be dependent on past values of itself and past values of all other variables in the system

Imagine you are an economist studying the interdependence of three key macroeconomic variables:

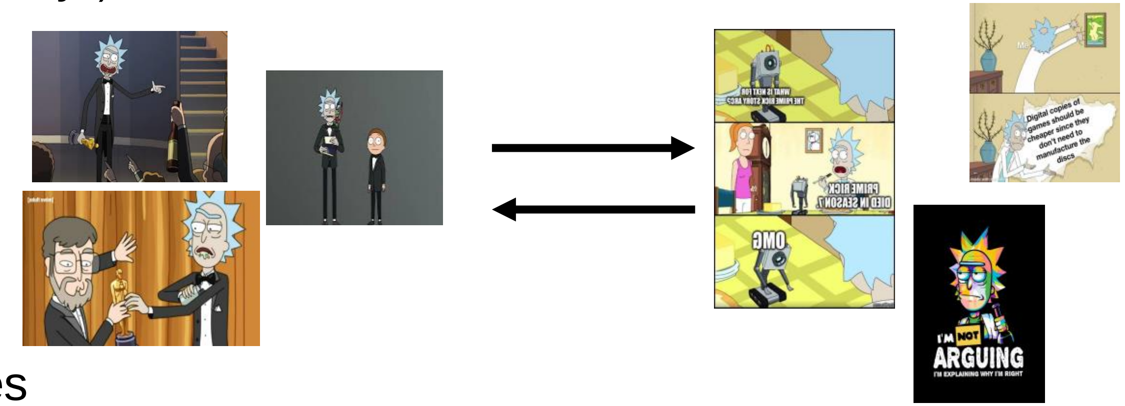
- GDP Growth (economic activity)
- Unemployment Rate (labor market conditions)
- Inflation Rate (price stability)



How does it work (minimal example)?

You work for a streaming platform like Netflix or Hulu. You suspect that the **popularity** of shows on your platform and the frequency of **memes** related to those shows influence each other.

- y_t: Daily viewership numbers of a hit show (e.g., ”Rick and Morty”).
- x_t: Number of memes or social media posts about the show.



Why Use VAR?

- Popular shows generate memes, but memes can also attract new viewers. A VAR model can analyze this two-way relationship.
- The feedback loop can be timed to measure how long it takes for a meme to impact viewership or for a surge in viewership to generate more memes.

Insights:

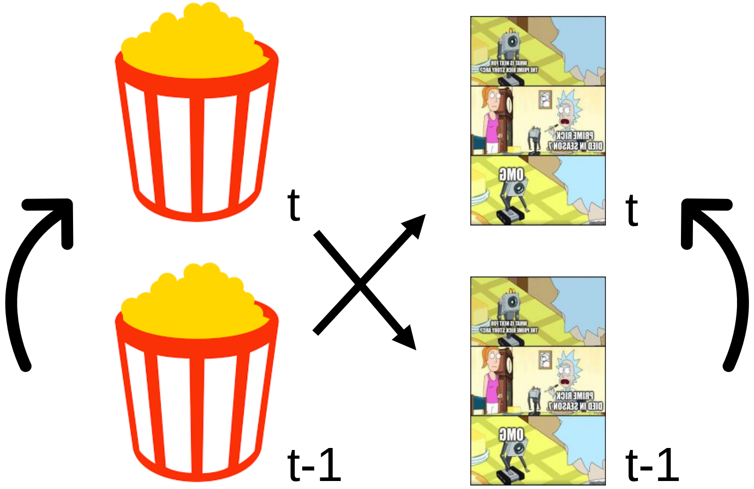
- Plan meme-driven marketing strategies to boost viewership when it’s declining.
- Analyze how memes sustain interest in a series long after its release.
- Check Granger causality to see if mentions predict sales or vice versa.
- Use impulse response functions (IRFs) to simulate what happens to sales if social media mentions suddenly spike (e.g., due to a viral post).

VAR(1) more formally

$$v_t = \phi_{0,1} + \phi_{11} v_{t-1} + \phi_{12} m_{t-1} + \epsilon_{v_t}$$

$$m_t = \phi_{02} + \phi_{21} v_{t-1} + \phi_{22} m_{t-1} + \epsilon_{m_t}$$

$$\underbrace{\begin{bmatrix} v_t \\ m_t \end{bmatrix}}_{y_t} = \underbrace{\begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \end{bmatrix}}_{\phi_0} + \underbrace{\begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}}_{\phi_1} \underbrace{\begin{bmatrix} v_{t-1} \\ m_{t-1} \end{bmatrix}}_{y_{t-1}} + \underbrace{\begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}}_{\epsilon_t}$$



$$f_t = \phi_0 + \phi f_{t-1} + \epsilon_t$$

VAR(2)

$$v_t = \phi_{0,1} + \phi_{11} v_{t-1} + \phi_{12} m_{t-1} + \psi_{11} v_{t-2} + \psi_{12} m_{t-2} + \epsilon_{v_t}$$

$$m_t = \phi_{02} + \phi_{21} v_{t-1} + \phi_{22} m_{t-1} + \psi_{11} v_{t-2} + \psi_{12} m_{t-2} + \epsilon_{m_t}$$

$$\begin{bmatrix} v_t \\ m_t \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} v_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix} \begin{bmatrix} v_{t-2} \\ m_{t-2} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}$$

$$\begin{bmatrix} v_t \\ m_t \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \end{bmatrix} + \begin{bmatrix} \phi^1_{11} & \phi^1_{12} \\ \phi^1_{21} & \phi^1_{22} \end{bmatrix} \begin{bmatrix} v_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} \phi^2_{11} & \phi^2_{12} \\ \phi^2_{21} & \phi^2_{22} \end{bmatrix} \begin{bmatrix} v_{t-2} \\ m_{t-2} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}$$

$$y_t = \phi_0 + \phi y_{t-1} + \psi y_{t-2} + \epsilon_t$$

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

VAR(k)

$$\begin{bmatrix} v_t \\ m_t \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \end{bmatrix} + \begin{bmatrix} \phi_{11}^1 & \phi_{12}^1 \\ \phi_{21}^1 & \phi_{22}^1 \end{bmatrix} \begin{bmatrix} v_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11}^2 & \phi_{12}^2 \\ \phi_{21}^2 & \phi_{22}^2 \end{bmatrix} \begin{bmatrix} v_{t-2} \\ m_{t-2} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11}^k & \phi_{12}^k \\ \phi_{21}^k & \phi_{22}^k \end{bmatrix} \begin{bmatrix} v_{t-k} \\ m_{t-k} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}$$

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \epsilon_t$$

What if we have k different time series?

this is for 2:

$$v_t^1 = \phi_{0,1} + \phi_{11} v_{t-1}^1 + \phi_{12} v_{t-1}^2 + \epsilon_{v_t^1}$$

$$v_t^2 = \phi_{02} + \phi_{21} v_{t-1}^1 + \phi_{22} v_{t-1}^2 + \epsilon_{v_t^2}$$

this is for 3:

$$v_t^1 = \phi_{0,1} + \phi_{11} v_{t-1}^1 + \phi_{12} v_{t-1}^2 + \phi_{13} v_{t-1}^3 + \epsilon_{v_t^1}$$

$$v_t^2 = \phi_{02} + \phi_{21} v_{t-1}^1 + \phi_{22} v_{t-1}^2 + \phi_{23} v_{t-1}^3 + \epsilon_{v_t^2}$$

$$v_t^3 = \phi_{03} + \phi_{31} v_{t-1}^1 + \phi_{32} v_{t-1}^2 + \phi_{33} v_{t-1}^3 + \epsilon_{v_t^3}$$

And for k:

$$v_t^1 = \phi_{0,1} + \phi_{11} v_{t-1}^1 + \phi_{12} v_{t-1}^2 + \cdots + \phi_{1k} v_{t-1}^k + \epsilon_{v_t^1}$$

$$v_t^2 = \phi_{02} + \phi_{21} v_{t-1}^1 + \phi_{22} v_{t-1}^2 + \ldots \phi_{2k} v_{t-1}^k + \epsilon_{v_t^2}$$

$$\ldots$$

$$v_t^k = \phi_{0k} + \phi_{k1} v_{t-1}^1 + \phi_{k2} v_{t-1}^2 + \ldots \phi_{kk} v_{t-1}^k + \epsilon_{v_t^k}$$

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{k,t} \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \\ \vdots \\ \phi_{0,k} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \cdots & \phi_{1k} \\ \vdots & \ddots & \vdots \\ \phi_{k1} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ \vdots \\ Y_{t-1,k} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \\ \vdots \\ e_{t,k} \end{bmatrix}$$

K time series and N time steps

2 time series with n time steps → n matrices of size 2x2

$$\begin{bmatrix} v_t \\ m_t \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \end{bmatrix} + \begin{bmatrix} \phi_{11}^1 & \phi_{12}^1 \\ \phi_{21}^1 & \phi_{22}^1 \end{bmatrix} \begin{bmatrix} v_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11}^2 & \phi_{12}^2 \\ \phi_{21}^2 & \phi_{22}^2 \end{bmatrix} \begin{bmatrix} v_{t-2} \\ m_{t-2} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11}^n & \phi_{12}^n \\ \phi_{21}^n & \phi_{22}^n \end{bmatrix} \begin{bmatrix} v_{t-n} \\ m_{t-n} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \end{bmatrix}$$

k time series with 1 time step → 1 matrix of size kxk

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{k,t} \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \\ \vdots \\ \phi_{0,k} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \cdots & \phi_{1k} \\ \vdots & \ddots & \vdots \\ \phi_{k1} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ \vdots \\ Y_{t-1,k} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \\ \vdots \\ e_{t,k} \end{bmatrix}$$

k time series with n time step → n matrices of size kxk

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{k,t} \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \\ \vdots \\ \phi_{0,k} \end{bmatrix} + \begin{bmatrix} \phi_{11}^1 & \cdots & \phi_{1k}^1 \\ \vdots & \ddots & \vdots \\ \phi_{k1}^1 & \cdots & \phi_{kk}^1 \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ \vdots \\ Y_{t-1,k} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11}^n & \cdots & \phi_{1k}^n \\ \vdots & \ddots & \vdots \\ \phi_{k1}^n & \cdots & \phi_{kk}^n \end{bmatrix} \begin{bmatrix} Y_{t-n,1} \\ Y_{t-n,2} \\ \vdots \\ Y_{t-n,k} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \\ \vdots \\ e_{t,k} \end{bmatrix}$$

VARMA

- VAR models only incorporate an AR part
- We can do as we did for ARMA and add an MA part → **VARMA**

$$Y_t = \alpha_0 + A_1 Y_{t-1} + e_t + B_1 e_{t-1}$$

$$Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{k,t} \end{bmatrix} = \begin{bmatrix} \phi_{0,1} \\ \phi_{0,2} \\ \vdots \\ \phi_{0,k} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \cdots & \phi_{1k} \\ \vdots & \ddots & \vdots \\ \phi_{k1} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ \vdots \\ Y_{t-1,k} \end{bmatrix} + \begin{bmatrix} e_{t,1} \\ e_{t,2} \\ \vdots \\ e_{t,k} \end{bmatrix} + \begin{bmatrix} \psi_{11} & \cdots & \psi_{1k} \\ \vdots & \ddots & \vdots \\ \psi_{k1} & \cdots & \psi_{kk} \end{bmatrix} \begin{bmatrix} e_{t-1,1} \\ e_{t-1,2} \\ \vdots \\ e_{t-1,k} \end{bmatrix}$$

This is called VARMA(1,1) and there is also VARMA(p,q) of course:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} A_1 Y_{t-i} + e_t + \sum_{j=1}^{q} B_1 e_{t-j}$$

Notes on VAR and VARMA models

- When estimating one target variable ARMA models are more popular that AR models
- When estimating multiple target variable, VAR models are more popular than VARMA models

- While VAR models need more parameters (more lags) to be as predictive as VARMA models, they are way easier to compute.
- VARMA models need iterative MLE due to the MA component and partial derivative computation.
- VAR models on the other hand remain manageable using a simple MLE approach.

Partial derivatives are hard to compute. Example:

- Given an 10 target variables, VAR(20) needs to compute $20 * 10 * 10 + 10 + 55 = 2065$.
- VARMA(1,1) needs $2*10*5+10+55 = 165$. But requires 1622 partial derivatives computations!!

How to choose the correct number of lags?

- For the AR part we use the PACF as usal.

But what about the number of lags of the second series?

> The VAR model from statsmodel gives the coefficients for all regressors until maxlags. You should look at the probs column which is the p-value of each coefficient, and only consider the coefficients with a low p-value.

IRF:

- The IRF measures how a variable in a VAR system reacts over time to a one-unit shock in one of the variables while holding all other shocks at zero.
- It provides insight into the dynamic behavior and relationships among the variables in the system.

- You can also use the correlation between the series of interest and a lagged auxiliary series to get a hint on the right lag value to test

FEVD: Forecast Error Variance Decomposition

Forecast Error Variance

- The forecast error variance represents the uncertainty in predicting a variable's future values within a time series model
- It quantifies the variance of the errors made when forecasting a variable at different time horizons.

Forecast Error:

- The **forecast error** is the difference between the actual value and the predicted value at a future time t + h (for horizon h):

$$\text{Forecast Error} = Y_t - \hat{Y}_t$$

Variance of Forecast Error:

- Over time, the forecast error varies because predictions are influenced by random shocks in the system.
- The forecast error variance measures the spread or dispersion of these forecast errors over repeated predictions.

Role in VAR Models:

- In a VAR model, the forecast error variance arises from shocks to all variables in the system.
- Shocks can propagate across variables, influencing the forecast error variance.

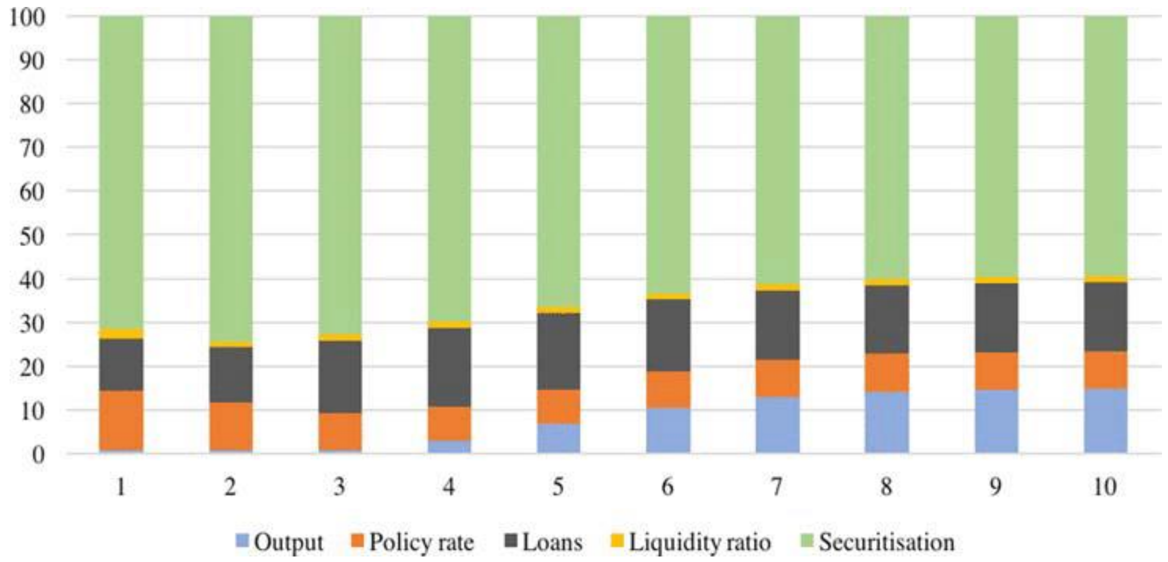
Sources of Variance in a VAR Model

- Own Shocks: Variance contributions from the variable itself.
- Cross-variable Shocks: Variance contributions from shocks to other variables in the system, propagating through dynamic interactions.

Forecast Error Variance Decomposition

- In FEVD, the forecast error variance is decomposed into portions attributable to shocks in each variable.
- This decomposition helps identify how much of the forecast uncertainty (variance) in one variable is explained by shocks to itself and others.

The forecast error variance quantifies the total uncertainty, while FEVD determines which variables contribute most to this uncertainty.



- **Time Horizons**: FEVD is calculated for different forecast horizons. Short-term and long-term decompositions often reveal different dynamics.

- The FEVD results are expressed as percentages (or proportions) indicating the relative importance of each variable’s shock

IRF vs FEVD

	IRF	FEVD
What It Answers	• What happens to variable Y if there's a one-unit shock to variable X? • How does the effect of this shock evolve over time?	• How much of the uncertainty in forecasting Y is caused by shocks to X versus shocks to Y itself or other variables? • Which shocks are most important in explaining the variance in forecasts?
Output	A time series showing the response of each variable to a shock in one specific variable.	A breakdown of the forecast error variance into components attributable to each variable's shocks, expressed as proportions (or percentages).
Example	Shows how GDP responds to a 1% monetary policy shock (e.g., an interest rate change) over the next 10 quarters.	it might show that 70% of GDP's forecast error variance at a 10-quarter horizon is due to shocks in GDP itself, while 20% is due to monetary policy shocks, and 10% to inflation shocks.

Example

Suppose the FEVD output for a VAR(3) model with variables GDP, inflation, and interest rate is:

Horizon (h)	GDP (by GDP)	GDP (by Inflation)	GDP (by Interest Rate)
1	0.90	0.05	0.05
5	0.60	0.25	0.15
10	0.50	0.30	0.20

- At h = 1, GDP's forecast error variance is mostly due to its own shocks (90%).
- By h = 5, inflation shocks become more significant, explaining 25% of the variance, while interest rate shocks explain 15%.
- At h = 10, the influence of GDP's own shocks reduces further to 50%, and inflation's influence grows to 30%.

→ in the short term, GDP is driven mostly by its own dynamics, but in the long run, inflation shocks play a larger role

Last word

- **IRF**: Imagine you’re studying how a ripple in a pond spreads. The IRF tells you how far and in what direction the ripple moves (its dynamics) if you drop a stone.

- **FEVD**: Now imagine you want to understand why the pond’s surface is choppy over time. The FEVD tells you how much of the chop is caused by stones being dropped in specific areas.

Time for Exercise 1

Cointegration

Simo Alami

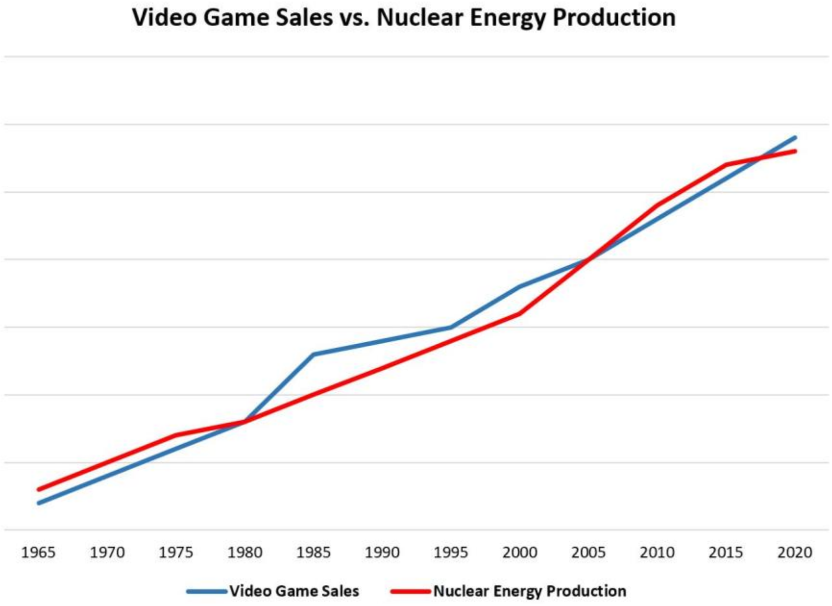
Stationarity and Spurious Regression

- As usual VAR requires the considered series to be stationary before being applied → **Why?**

→ We first need to define spurious regression, which is close to the case of spurious correlations we already mentioned.

- When you have two different time series, you might compute the correlations between them and you will find high correlations.

- the R2 statistic between these two series is very high.



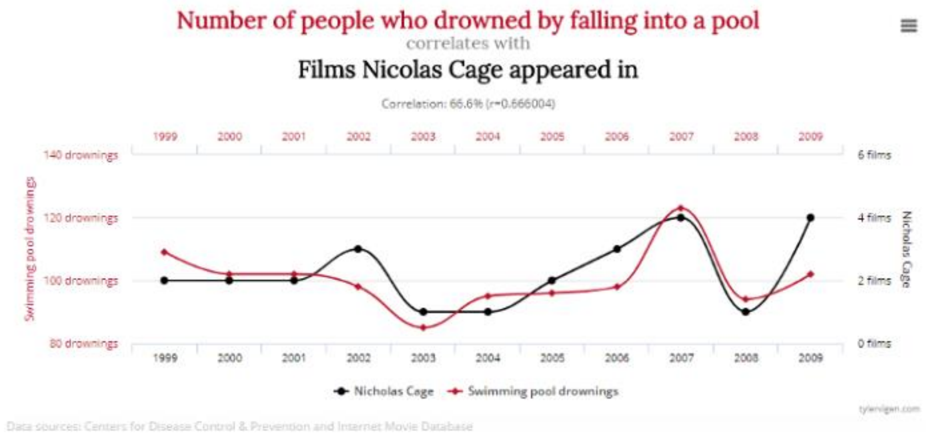


Figure 3-17. Some spurious correlations can look surprisingly convincing. This plot was taken from Tyler Vigen's website of spurious correlations.

R^2 reminder

- The R^2 statistic, also known as the coefficient of determination, measures how well a regression model explains the variability of the dependent variable.

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{Res}}}{\mathrm{SS}_{\mathrm{Total}}}$$

Where:

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- R^2 ranges between 0 and 1:
  - $R^2 = 0$ the model explains none of the variability in Y
  - $R^2 = 1$ the model explains all the variability in Y
- A higher R^2 value indicates a better fit of the model to the data.

- To account for the number of predictors and prevent overfitting, the Adjusted R^2 is used:

$$R^2_{\mathrm{adj}} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

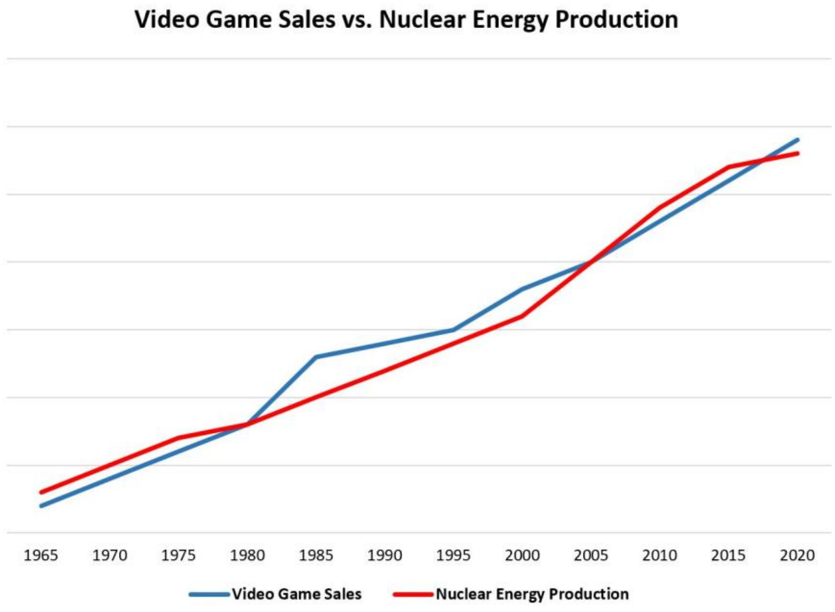
Where:

- n is the number of observations.
- k is the number of predictors.

R^2 and stationarity

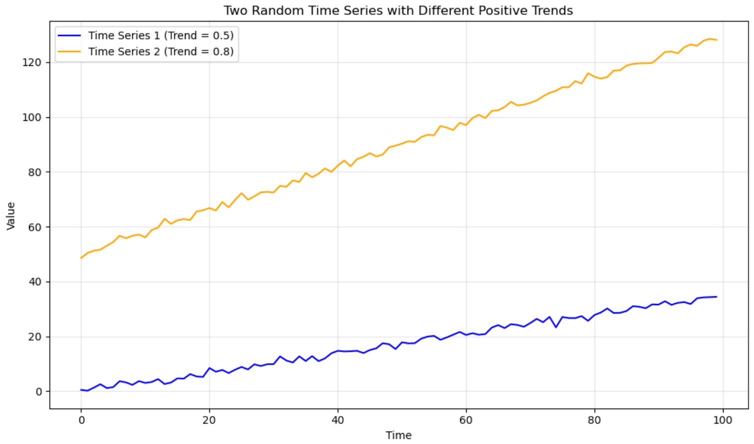
- Obviously nuclear energy production does not explain or rather in a limited fashion the video games sales.

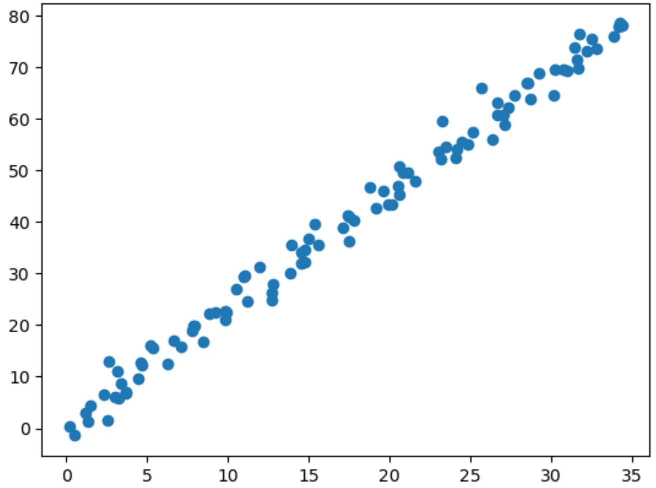
→ The issue is that the two time series are not stationary. **Why does non stationarity induces high R2 statistic?**



- Non-stationary series typically exhibit trends or persistent patterns over time. Even if these trends are entirely unrelated, they may still appear correlated because they move in similar directions over time.

- When you run a regression, the model attributes this shared trend to a ”relationship” between the series, inflating the R^2 and statistical significance.





R^2 and random walk

- Many non-stationary series, such as financial data, can be modeled as random walks.

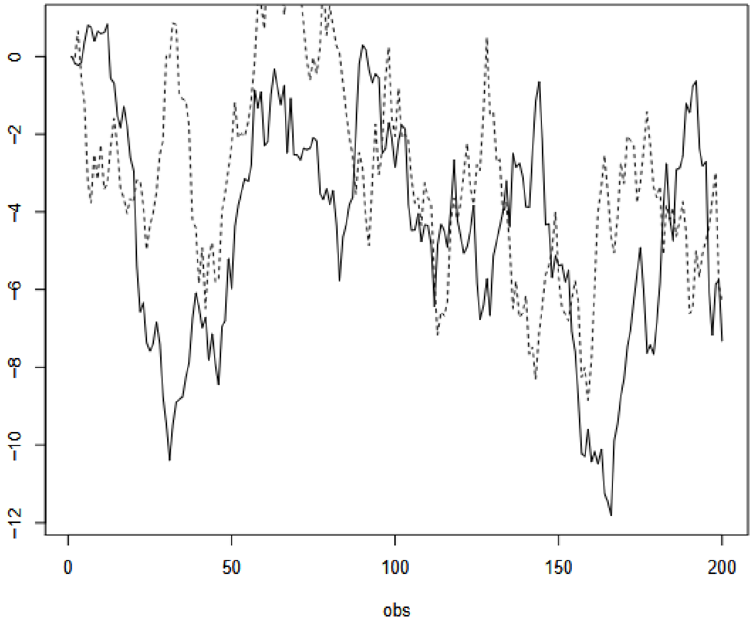
$$X_t = X_{t-1} + \epsilon_t, \quad \text{where } \epsilon_t \sim \text{i.i.d. noise}$$

- Even if two series Xt and Yt are independent random walks, their paths will often cross or trend in similar ways purely by chance

- Sometimes their local trends are similar, giving rise to the spurious regression

→ **High R^2**: The regression captures coincidental co-movement rather than any meaningful relationship.
→ Misleading t-statistics and p-values.

⚠ Just because two series move together does not mean they are related!



- **So what can we do?**

Can't we just make the time series stationary through differencing?
Unfortunately no

Why differencing does not work for long term

Objective: model the **long term** relationship between two series.

- Differencing removes the underlying trend or level information that is crucial to identifying long-run relationship

Example: GDP investment may both be non-stationary but move together over time due. Differencing them would remove the information about this connection.

- Using VAR on differenced time series would remove long term relationships
- Differenced VAR only captures short term dynamics

→ long-term equilibrium behavior, can be critical in many economic, financial, and policy analyses.

We do not care if two series deviate in the short run, we want to assess if they share a stable relationship in the long run.

Example: Imagine two time series: Y_t the consumption and X_t the income. Both are non-stationary, but they move together over the long run. If you difference both series to make them stationary:

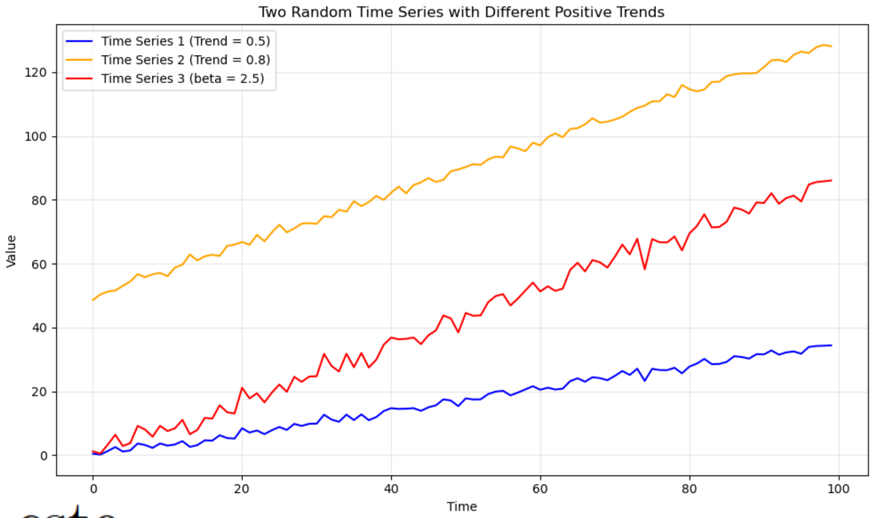
- the resulting relationship reflects short-term changes,
- but you lose the meaningful insight that they share a stable, long-term equilibrium.

Cointegration definition

- When time series are individually non-stationary but have a stable, long-term equilibrium relationship, they are said to be cointegrated

Different slopes, increasing distance

Same slope, parallel time series



Multiply β, to decrease the distance, and make time series parallel to each other.

Parallel = the difference between time series is constant: $y_t - \beta x_t = \text{cste}$

→ The difference equals something that is stable in time: **looks like stationarity**

Two non-stationary series X_t and Y_t are cointegrated if there exists a coefficient β such that:

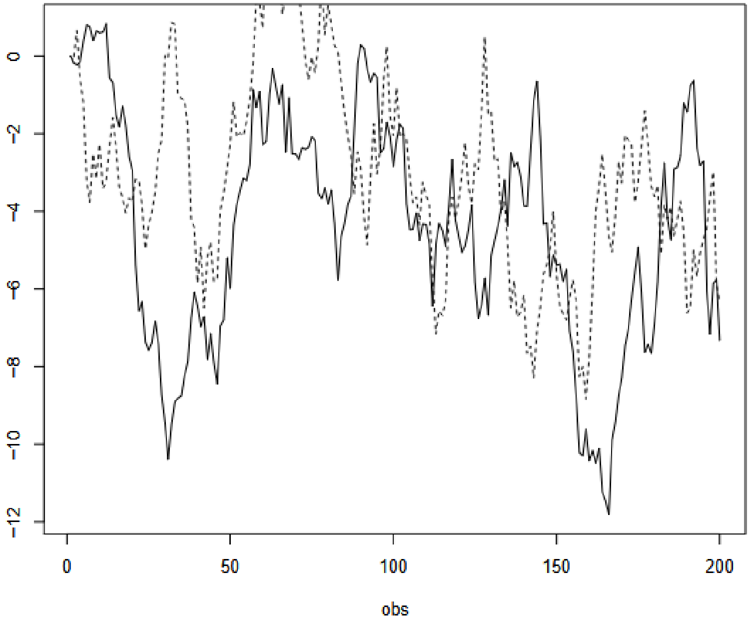
$$Z_t = Y_t - \beta X_t$$

is stationary.

Testing for cointegration (Engle-Granger)

- In the example, there is no way to make the distance between both time series stationary

→ they are not co-integrated, the correlation between them is spurious.



If two series are cointegrated, it might be ok to regress one on the other

- How to test for co-integration?
  1. We need to find β. First you fit a linear regression:

$$y_t = \hat{\alpha} + \hat{\beta} x_t + \epsilon_t$$

  2. Check if the residuals ϵt are stationary → ADF test.

→ Now you tested your series, and found that they are co-integrated, what can you do?

Error Correction Models

Error Correction Models

Assumtions:
- x_t, y_t are two non stationary time series
- We suppose that they are I(1) → their first difference is stationary
- They are cointegrated → in the long run: $y_t = \alpha + \beta x_t$

- We considered that we could regress the first differences on one another: $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + \epsilon_t$
- This only represents the short term relationship, how to add long term relationships?

$$y_t = \alpha + \beta_1 X_t + \beta_2 X_{t-1} + \gamma y_{t-1} + \epsilon_t$$

- y_t is non stationary so let's difference it to make it stationary:

$$\underbrace{y_t - y_{t-1}}_{\text{stationary}} = \underbrace{\alpha + \beta_1 X_t + \beta_2 X_{t-1} - (1-\gamma) y_{t-1} + \epsilon_t}_{\text{Non stationary}}$$

$$y_t - y_{t-1} = \alpha + \beta_1 X_t - \beta_1 X_{t-1} + \beta_1 X_{t-1} + \beta_2 X_{t-1} - (1-\gamma) y_{t-1} + \epsilon_t$$

$$\Delta y_t = \alpha + \beta_1 \Delta X_t - \lambda (y_{t-1} - \phi_0 - \phi_1 X_{t-1}) + \epsilon_t \quad \text{With } \lambda = (1-\gamma),\ \phi_1 = \frac{\beta_1 + \beta_2}{1-\gamma}$$

Error Correction Models

$$\Delta y_t = \underbrace{\alpha + \beta_1 \Delta X_t}_{\text{Short term}} - \underbrace{\lambda(y_{t-1} - \phi_0 - \phi_1 X_{t-1})}_{\text{Long term}} + \epsilon_t$$

What is happening?

- ΔX_t and Δy_t are I(0)

- $\lambda(y_{t-1} - \phi_0 - \phi_1 X_{t-1})$ appeals to the long term relationship: if a long term relationship between the series exists, then this term will be cointegrated

2 possible situations:

- $y_{t-1} > \phi_0 - \phi_1 X_{t-1}$ **:** y_{t-1} is above the equilibrium value → $\underbrace{\text{We multiply it by } -\lambda \text{ to bring it back towards the}}_{\text{This is why we call it an error correction model}}$ equilibrium
- Same logic in the other case.

$\boxed{\lambda \text{ the speed at which the series adjust to any sort of disequilibrium.}}$

Error Correction Models

More generally, the error correction model takes the following form:

$$\Delta y_t = \alpha + \sum_{i=0}^{P} \beta_i \Delta X_{t-i} + \sum_{j=0}^{K} \gamma_j \Delta y_{t-j} - \lambda \underbrace{(y_{t-1} - \phi_0 - \phi_1 X_{t-1})}_{e_{t-1}} + \epsilon_t$$

To fit this model you have to:

1. First assess cointegration and fit ϕ_0 and ϕ_1 so that: $y_t = \phi_0 + \phi_1 X_t + e_t$. Therefore:

$$y_{t-1} - \phi_0 - \phi_1 X_{t-1} = e_{t-1}$$

2. Fit the model

3 cases:

- **Stationary Series**: A standard VAR model can be directly applied
- **Non-Stationary Series Without Cointegration:** you can use differencing to make them stationary before applying a VAR model. This results in a VAR in differences.
- **Cointegrated Series**: If the series are non-stationary but cointegrated, applying a VAR model without considering the cointegration relationship would lead to misspecification

Time for Exercise 2

ARIMAX

Simo Alami

The issue with ARIMA models

- In the feature engineering lecture, we studied ways to introduce new variables (datetime features…) that we don't want to forecast but that may encapsulate information that we would want to leverage.

  → The AR models we studied only handle one variable at once.

ARIMA Model:

- Models the relationship between a dependent variable and its own past values (autoregression) and past forecast errors (moving average).
- Includes a differencing step to make the time series stationary (integration).
- It only uses endogeneous variables

$$y_t = \phi_0 + \sum_{i}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{j}^{q} \theta_j \epsilon_{t-j}$$

- An **ARIMAX** (Autoregressive Integrated Moving Average with Exogenous Variables) model is a powerful extension of the ARIMA model that incorporates additional information (exogenous variables) into the forecasting process.

ARIMAX

Exogenous Variables:
- External variables believed to influence the dependent variable.
- These are provided as additional predictors in the model.

→ Examples:
  - Day of the week: Sales may be higher on weekends.
  - Temperature: Energy consumption may depend on weather conditions.
  - Marketing campaigns: Sales may spike during promotions.

How to include exogeneous variables?

→ These variables are incorporated as direct predictors, improving model accuracy when their influence is significant.

The general form of an ARIMAX model is:

$$y_t = c + \underbrace{\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}}_{\text{AR part}} + \epsilon_t + \underbrace{\theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}}_{\text{MA part}} + \underbrace{\beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t}}_{\text{X part}}$$

- The ARIMA part models the autocorrelations in yt, and the regression part explains y_t using the external predictors$x_{1,t}, x_{2,t}, \dots, x_{k,t}$.
- The exogenous variables can be contemporaneous (affecting yt immediately) or lagged (affecting y_t with a delay)

ARIMAX in 4 steps

Step 1: Stationarity and Differencing

- Just like ARIMA, the dependent variable y_t may need to be differenced to achieve stationarity.
- The exogenous variables $x_{i,t}$ typically do not require differencing, but they **must be stationary or properly preprocessed.**

Step 2: Incorporating Exogenous Variables

- Exogenous variables are treated as additional predictors in a regression-like fashion.
- These variables should be aligned in time with the dependent variable y_t.

Step 3: Model Estimation

- The coefficients ϕ (AR), θ (MA), and β (exogenous variables) are estimated simultaneously.
- Estimation is typically performed using MLE.

Step 4: Forecasting

- Future values of y_t are predicted using:
  - The past values of AR terms.
  - The past errors MA terms.
  - Known or forecasted future values of the exogenous variables $x_{i,t}$.

When forecasting, you must provide future values of the exogenous variables:
1. If the exogenous variables are known or deterministic (e.g., day of the week), input them directly into the model.
2. If the exogenous variables are not known (e.g., future temperature), **you must forecast them separately.**

Is Differencing Necessary for Exogeneous Variables?

Generally not necessary: Exogeneous variables are treated as predictors rather than part of the time series being directly modeled for stationarity.

Not required when:

- **Already Stationary Variables:** Categorical variables are considered stationary by design

- the relationship between the dependent variable and exogenous variables is only meaningful in their original form → **Use cointegration**

- **Cointegrated Relationships:** If the exogenous variables are non-stationary but have a cointegrated relationship with the dependent variable (a stable long-term equilibrium exists), differencing may not be necessary. Instead, a cointegration-aware modeling approach can be used. For instance a VAR model.

Scenarios where it might be required:

- **Non-Stationary Exogenous Variables**: If an exogenous variable is non-stationary. **Example:** steadily increasing variable, like population growth over time or a variable with a strong seasonal trend, such as monthly temperature averages.

- **Consistency with the Dependent Variable**: If the dependent variable has been differenced to achieve stationarity, and the exogenous variables exhibit a similar trend, differencing the exogenous variables ensures they are on the same ”stationarity level”

How to handle categorical variables

- Categorical variables like ”day of the week” are considered inherently stationary because they represent a fixed set of repeating labels (e.g., Monday, Tuesday, etc.) that has no statistical property
- To use them effectively in a time series model like ARIMAX, you need to encode them numerically since models cannot directly interpret categorical data → One hot encoding

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding →

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Correlated Exogeneous Variables

- **Multicollinearity:** Highly correlated exogenous variables, two or more exogenous variables are strongly linearly related.

→ **Unstable Coefficient Estimates:** it becomes difficult for the model to determine the unique effect of each variable on the dependent variable. This can result in large standard errors for the coefficients, making them highly sensitive to small changes in the data. reducing their statistical significance. Even variables with strong true effects might appear insignificant due to inflated standard errors.

→ **Reduced Interpretability**: It becomes challenging to interpret the effect of individual exogenous variables because their contributions overlap. **Example:** if temperature and time of year are both included and strongly correlated, determining their independent effects on sales becomes ambiguous.

→ **Overfitting**

How to detect and Address Multicolinearity

- Correlation Matrix
- **Variance Inflation Factor** (VIF): Calculate the VIF fo each variable. A VIF > 5 (sometimes > 10) suggests high multicollinearity.

$$VIF = \frac{1}{1 - R^2}$$

where R^2 is the coefficient of determination when the variable is regressed on all other exogenous variables.

How to Address Multicollinearity:

- Remove Redundant Variables
- Combine Correlated Variables: Aggregate correlated variables into a single composite variable (e.g., average)
- Orthogonalization: Transform correlated variables into orthogonal components using techniques like PCA before including them in the model

Example: Suppose you are building an ARIMAX model to predict sales (Y_t) with the following exogenous variables: Advertising spend (X_1) and Social media activity (X_2)

→ Check the correlation matrix or VIF: If cor(X_1, X_2) > 0.8, consider action.
→ Combine them: Create a new variable, e.g., ”marketing effort” = $X_1 + X_2$.

How to quantify the dependance of the dependent variable on the exogeneous variables?

- **Coefficient Estimates**: The coefficients in the ARIMAX model measure the direct impact of the exogenous variables on the dependent variable.
  → absolute value reflects the strength of the effect of x_k on y_t
  → The sign indicates whether x_k has a positive or negative relationship with y_t

Caution: Scale the variables to have zero mean and unit variance, then compare the magnitudes of the coefficients to assess relative importance.

- Partial correlation
- Granger Causality Test
- **Variance Decomposition**: Variance decomposition quantifies how much of the variance in y_t is explained by each x_k
- Statistical Significance of Coefficients
- Impulse Response Function

Number of lags for Exogeneous Variables

Instead of using only x_t, include x_{t-1}, x_{t-2}, . . . as additional predictors in the model.

$$y_t = \phi_0 + \phi_1 \cdot y_{t-1} + \theta_1 \cdot \epsilon_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \epsilon_t$$

How to Determine the Number of Lags?

- Cross correlation analysis
- Domain knowledge
- Information criterions (AIC, BIC)

Caution:

- Including too many lags can cause overfitting

- **Multicollinearity**: lagged values of exogeneous variables may be highly correlated

Cross Correlation

Definition: Cross-correlation analysis measures the similarity or relationship between two time series as a function of the lag of one relative to the other. It’s often used to determine whether one time series influences or is related to another over time.
→ **Equivalent to the ACF but between different time seris**

The cross-correlation at lag k between two time series X_t and Y_t is computed as:

$$\mathrm{CCF}(k) = \frac{\sum_t (X_t - \bar{X})(Y_{t+k} - \bar{Y})}{\sqrt{\sum_t (X_t - \bar{X})^2 \sum_t (Y_t - \bar{Y})^2}}$$

Limitations:

- **Stationarity Requirement**: Both series should be stationary for reliable results.
- **Spurious Correlations**: High cross-correlation can occur due to common trends or seasonality rather than causality.
- **Non-linear Relationships**: Cross-correlation measures only linear relationships.

Time for Exercise 3