

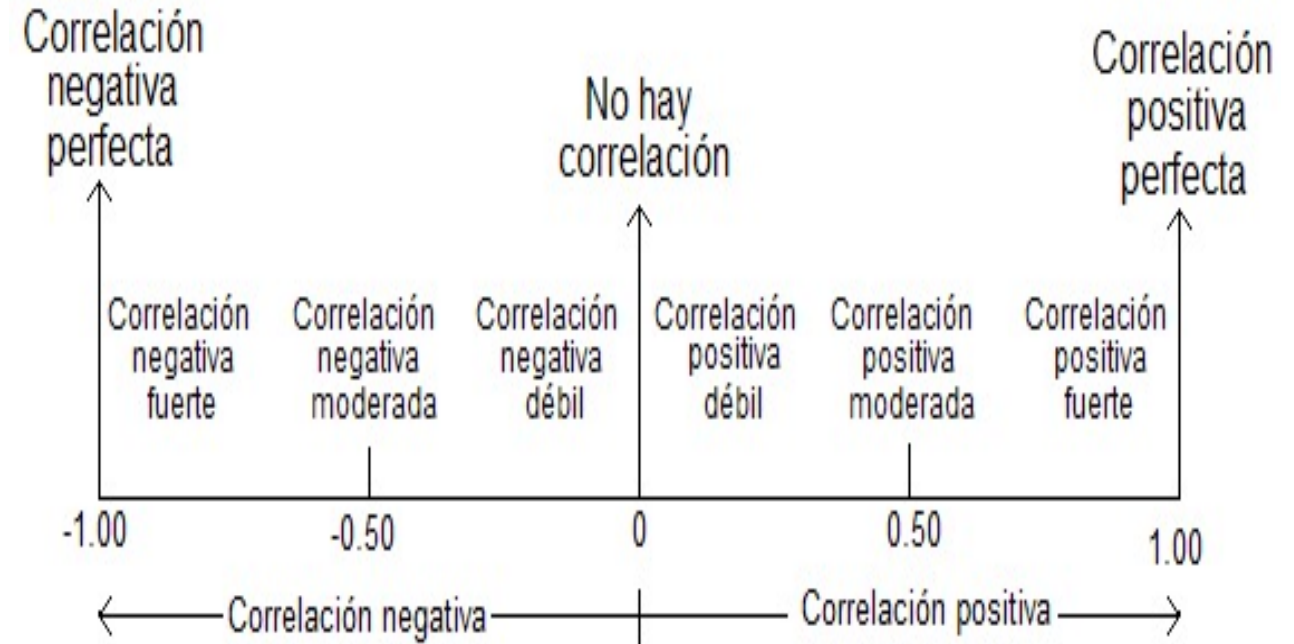
REVIEW

Let's talk about correlation

It measures the 'intensity' of the relationship between variables, in addition to indicating the trend these have.

The correlation coefficient is a value that ranges from **-1 to 1**. The closer to **(+1 or -1)**, the 'stronger' the relationship between variables. If it is positive the relationship is **DIRECT**.

If it is negative the relationship is **INVERSE**.



Linear Model: Ordinary Least Squares

The objective of the OLS method is to minimize the sum of the squares of the residuals (differences between the observed values and the values predicted by the model).

It is a method widely used in linear regression to estimate the coefficients of a model.
Characteristics:

a) Linearity in the Parameters: the estimated coefficients ($\beta_0, \beta_1, \beta_2 \dots \beta_n$) appear linearly in the regression equation.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

That is, the relationship between the dependent variable Y and the independent variables X_i is linear in the coefficients β_i .

The assumptions of the **MC0** Model are:

1) Linearity: The relationship between the dependent and independent variables is linear in the parameters.

2) Exogeneity: The regressors (independent variables) are not correlated with the error term.

$$E(\epsilon_i | X_i) = 0. \quad E(\beta_j) = \beta_j.$$

3) Homoscedasticity: The residuals (errors) have a constant variance.

$$\text{Var}(\epsilon_i) = \sigma^2 \text{ para todos los } i.$$

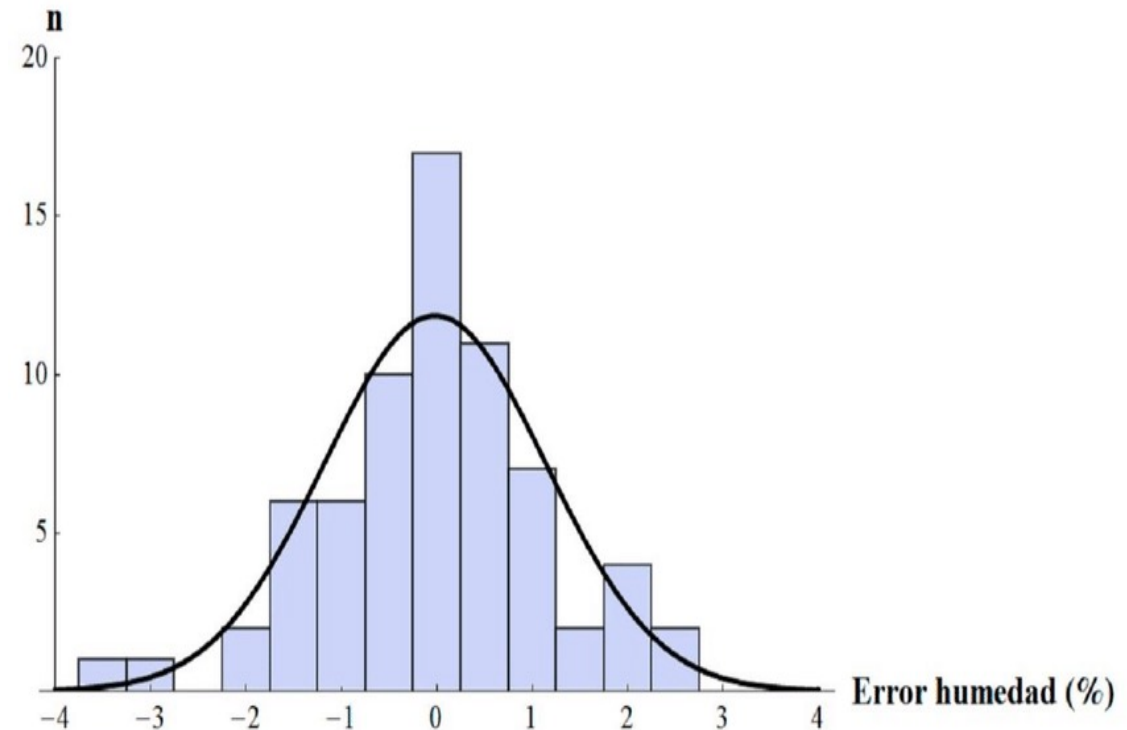
4) No Autocorrelation: The residuals are not correlated with each other.

$$E(\epsilon_i \epsilon_j) = 0 \quad \text{para } i \neq j.$$

5) No Perfect Multicollinearity: There is no perfect linear relationship between the independent variables.

6) Normality: Errors are normally distributed: $\epsilon_i \sim N(0, \sigma^2)$

This assumption is important primarily for statistical inference (e.g., for construction of confidence intervals and hypothesis testing) and model validation.



<i>Estadísticas de la regresión</i>	
Coeficiente de correlación múltiple	0.832412347
Coeficiente de determinación R^2	0.692910315
R^2 ajustado	0.682673993
Error típico	0.355391936
Observaciones	32

indicates the % of the variability of the dependent variable that is explained by the independent variables.

A

It helps determine whether the model “as a whole” is useful in explaining the variability of the dependent variable. $F < 0.05$

ANÁLISIS DE VARIANZA								
	<i>Grados de libertad</i>	<i>Suma de cuadrados</i>	<i>Promedio de los cuadrados</i>	<i>F</i>	<i>Valor crítico de F</i>			
Regresión	1	8.549647154	8.549647154	67.6913309	3.48794E-09			
Residuos	30	3.789102846	0.126303428					
Total	31	12.33875						
	<i>Coeficientes</i>	<i>Error típico</i>	<i>Estadístico t</i>	<i>Probabilidad</i>	<i>Inferior 95%</i>	<i>Superior 95%</i>	<i>Inferior 95.0%</i>	<i>Superior 95.0%</i>
Intercepción	-0.19	0.42	-0.45	0.66	-1.04	0.67	-1.04	0.67
temp	0.15	0.02	8.23	0.00	0.11	0.19	0.11	0.19

B

C

D

C) Propensity

We observe how the dependent variable changes in response to a change in a specific independent variable.

Reading:

Y= liters per day per inhabitant

X= temperature

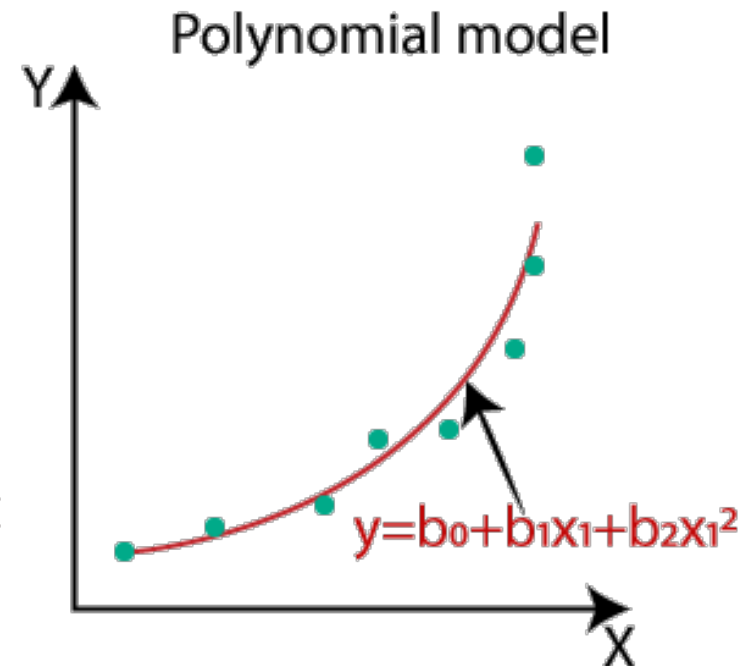
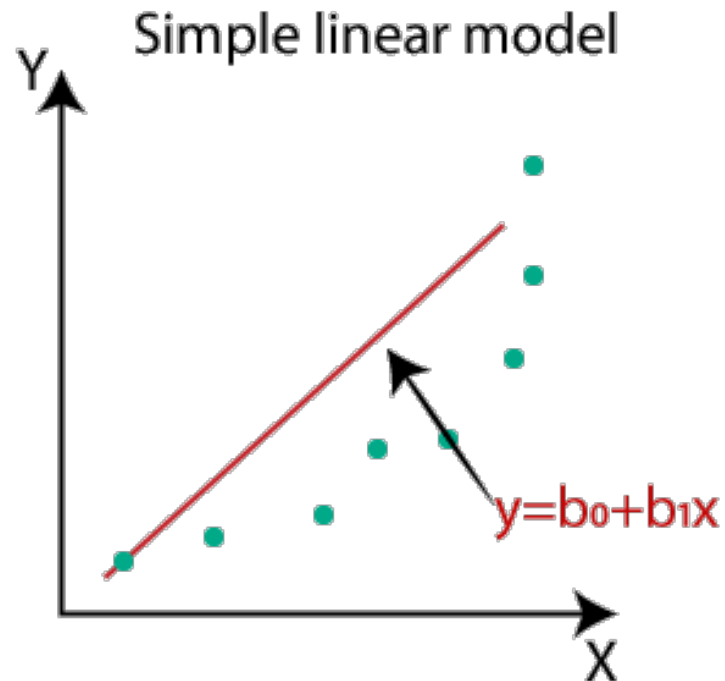
Coeff: Temp: 0.15

If X (independent variable) increases by one unit, Y (dependent variable) increases by 0.15 units (liters/person/day).

For every extra Celsius degree of increase in temperature, soda consumption increases by 0.15 liters per day per inhabitant.

Polynomial model

- It is a type of regression model used to describe the relationship between independent variable(s) and the dependent variable as a polynomial function.
- Unlike a simple linear regression, which fits a straight line to the data, a polynomial regression can fit curves of various shapes to capture the more complex relationship.



When to use it: when the relationship between the independent variable(s) and the dependent variable is not linear but instead follows a more complex pattern.

Curved Relationships: the relationship between age and certain health metrics might not be linear.

Quadratic or Higher-Order Trends: the relationship between the speed of a car and its stopping distance might be quadratic.

Non-Linear Dynamics: In scenarios where the relationship between variables involves non-linear dynamics, such as in certain physical, biological, or economic processes.

Simple
Linear
Regression

$$y = b_0 + b_1x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Polynomial
Linear
Regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$

LASSO & RIDGE. When to use them

These are categorized as linear models, but they have a couple of particularities:

- We use them for Feature selection (best variables).
- Both add a penalty term to the residuals to control the size of coefficients.
- Both are used to prevent overfitting and highly correlated predictors (multicollinearity).
- Both work well with highly correlated variables and high-dimensional datasets.

Estimating the β

Let us remember the ordinary minimum square method to solve the linear model
SCE (Sum of square error):

$$\text{SCE} = \sum (Y - \hat{Y})^2$$

Both methods Ridge and LASSO are trying to penalize the size of the parameters. In the case of LASSO we use an L_1 (absolute value) correction, which defines estimation of the parameters:

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta} \left(\text{SCE} + \lambda \sum_{j=1}^p |\beta_j| \right)$$

Penalization: Absolute value correction

Ridge Model.

Ridge regression also penalizes the size of the parameters through *shrinkage*.

BUT, instead of using LASSO's norm, Ridge uses the L_2 norm instead of the L_1

Remember the solution of Ordinary Least Squares, it is meant to estimate the β parameters that minimize the sum square error (SCE). In the case of Ridge:

$$\hat{\beta}^{\text{Ridge}} = \arg \min_{\beta} \left(\text{SCE} + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

Penalization: L2= squared parameters

Lasso vs Ridge

Ridge regression squares
the variables



The sum of the squared residuals

+

$\lambda \times \text{the slope}^2$

Penalization

Lasso regression takes
the absolute value.



The sum of the squared residuals

+

$\lambda \times |\text{the slope}|$

Penalization

Main problems to solve in cross section

Multicoliniality,
Heterocedasticity,
Error Normality

a) Multicollineality

It occurs when two or more predictor variables in a regression model are highly correlated, to the extent that it becomes difficult to separate their individual effects on the dependent variable.

Detection:

- **Correlation Matrix:**
- **Variance Inflation Factor (VIF):** A VIF above 5 often suggests multicollinearity.

Consequences:

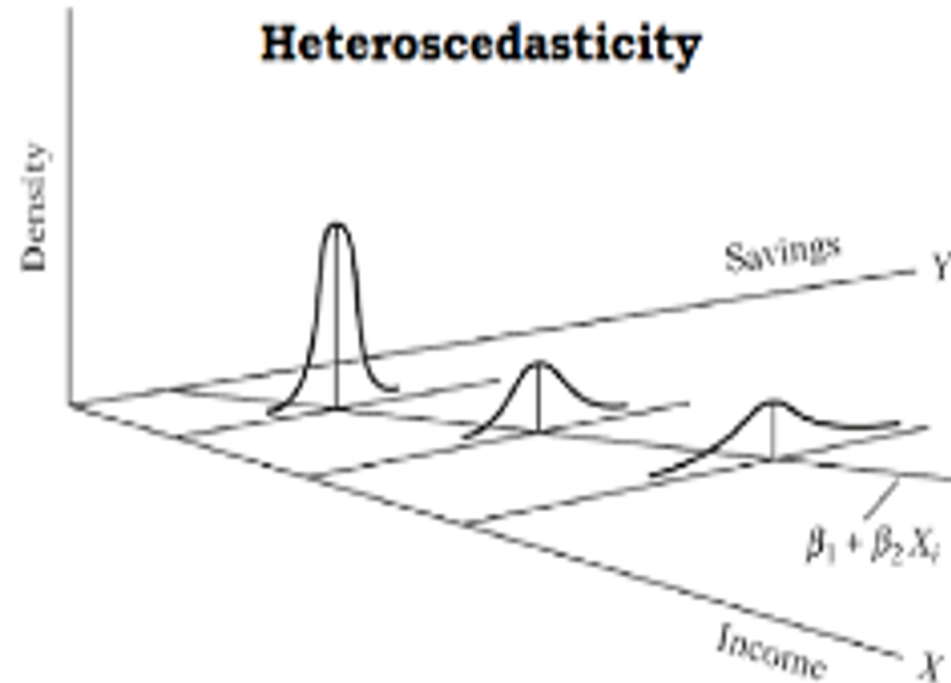
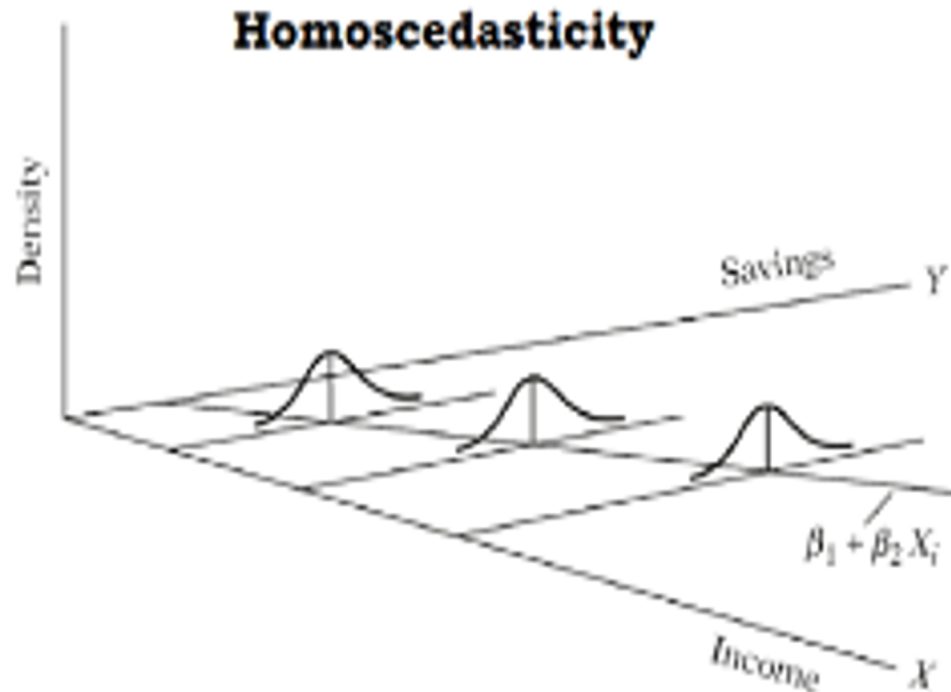
- Inflated standard errors of the coefficients, increase in the variance and covariance, the parameter estimation becomes unstable.

Correction

- **Remove Variables:** Drop the most correlated variables.
- **Ridge or Lasso Regression**

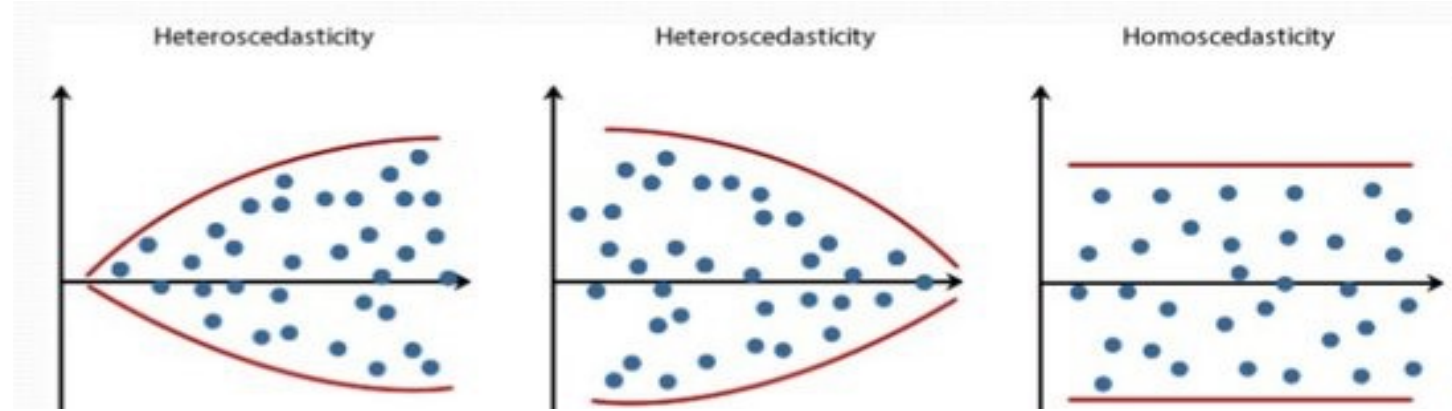
2) Heteroscedasticity

It occurs when the variance of the residuals from a regression model is not constant across observations.



Consequences

- Standard errors are biased, leading to invalid hypothesis tests.
- Inefficiency in estimates, making the model less reliable for inference.

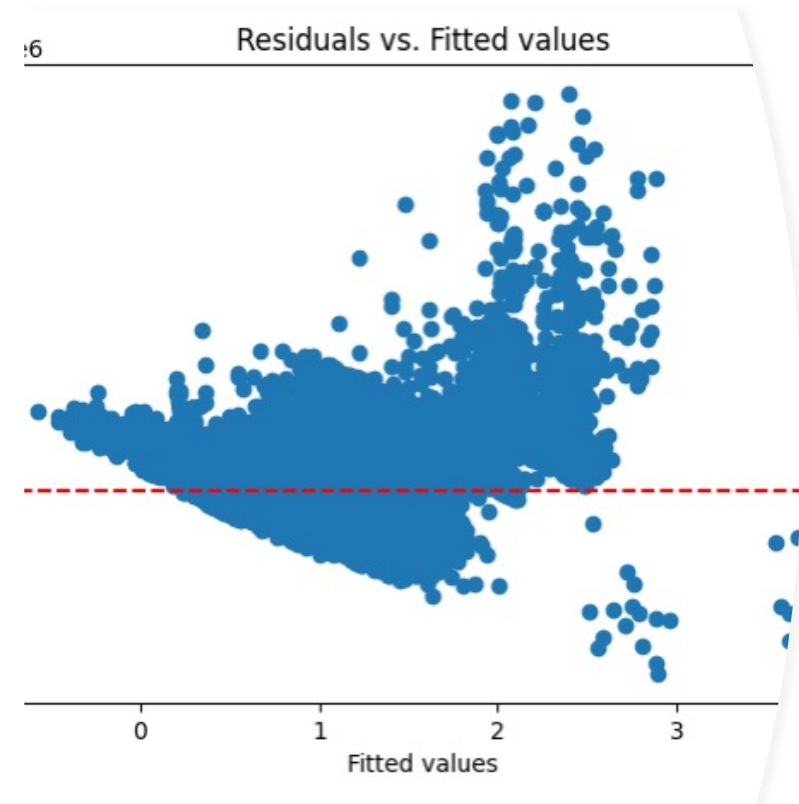


Detection

Breusch-Pagan and White Test:

- **Null Hypothesis (H0):** Homoscedasticity is present (the residuals are distributed with equal variance)
- **Alternative Hypothesis (HA):** Heteroscedasticity is present (the residuals are not distributed with equal variance). P-value < 0.05 heteroscedasticity is present in the regression model.

Residual Plots: Plot the residuals against fitted values.



Correction:

2.1) Transform Variables to Natural log (ln) or log

WE WILL GET AN ELASTICITY

2.2) Generalized Least Squares (GLS): Adjust for heteroskedasticity and autocorrelation.

GLS transforms the original data by multiplying it with the inverse of the square root of the error covariance matrix. This process standardizes the errors to have constant variance and no correlation.

2.3) Weighted Least Squares (WLS) is a special case of GLS

WLS ensures that the regression analysis is more influenced by observations with more reliable estimates.

instead of minimizing the Residual Sum of Square (RSS), we minimize the weighted sum of squares.

Elasticity: How to read when we have transformed variables in (logs)

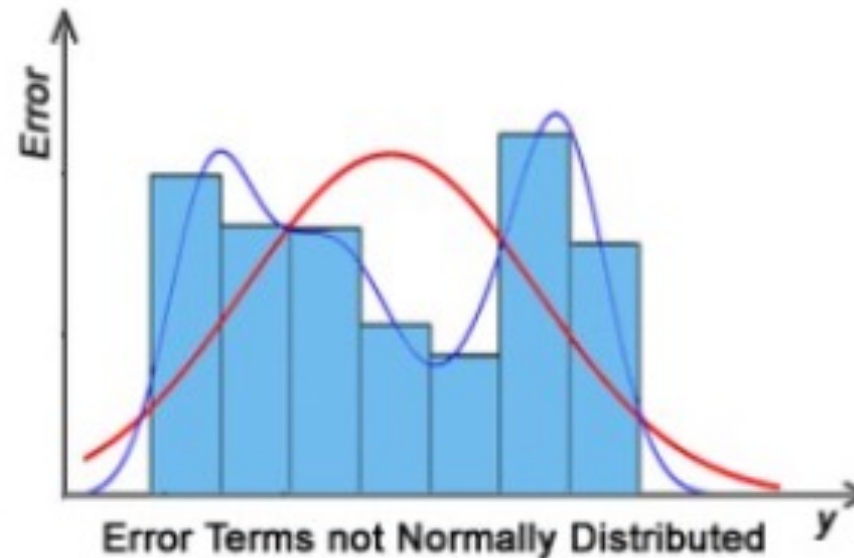
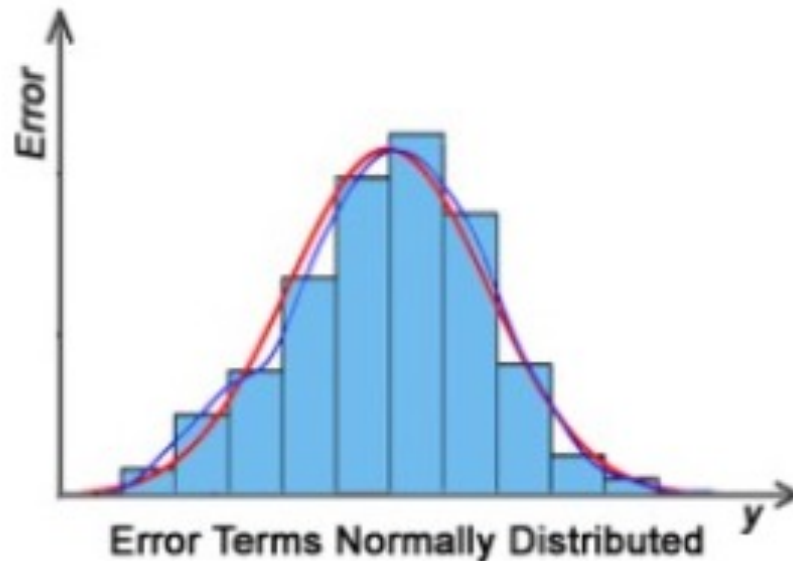
Economic concept used to measure the sensitivity or responsiveness of a product to a change in its price. **Log(Y)= Log(Price)**

	coef	std err	t	P> t
const	13.8131	0.001	1.61e+04	0.000
Property_ln	-0.0223	5.93e-05	-375.334	0.000
Distance_ln	-0.2712	5.21e-05	-5200.685	0.000
Rooms	0.4676	0.001	662.629	0.000
WC	-0.6731	0.001	-928.699	0.000
Garage	0.6040	0.000	4323.131	0.000

- Elasticity: An increase of 1% in the number of properties (region house density) decreases the price of housing by 0.022%.
- **Elasticity:** An increase of 1% in the distance to the centre, decreases the price of housing by 0.27%.
- **Semielasticity:** Every extra Room increases the price of the property in (0.46*100= 46%)
- **Semielasticity:** Every extra space of garage increases the price of the house in 60%

3) Non- normality distribution of residuals

The residuals of a regression should follow a normal distribution with a mean equal to zero. If the error terms are non-normally distributed, suggests that there are a few unusual data points that must be studied closely to make a better model. or close to zero



Identifying Non-Normality:

Histogram of Residuals: Plotting the residuals.

Shapiro-Wilk Test: This test evaluates whether the residuals come from a normal distribution.

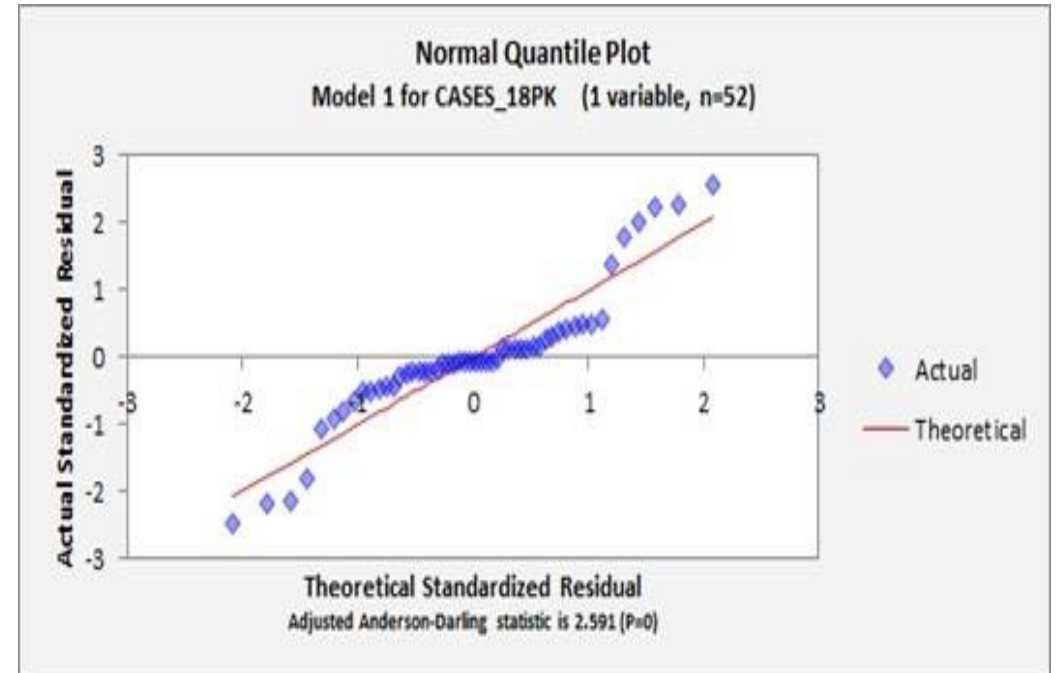
Kolmogorov-Smirnov Test: It compares the sample distribution of residuals with a specified distribution, such as normal.

Null Hypothesis (H0): The sample comes from a normal distribution.

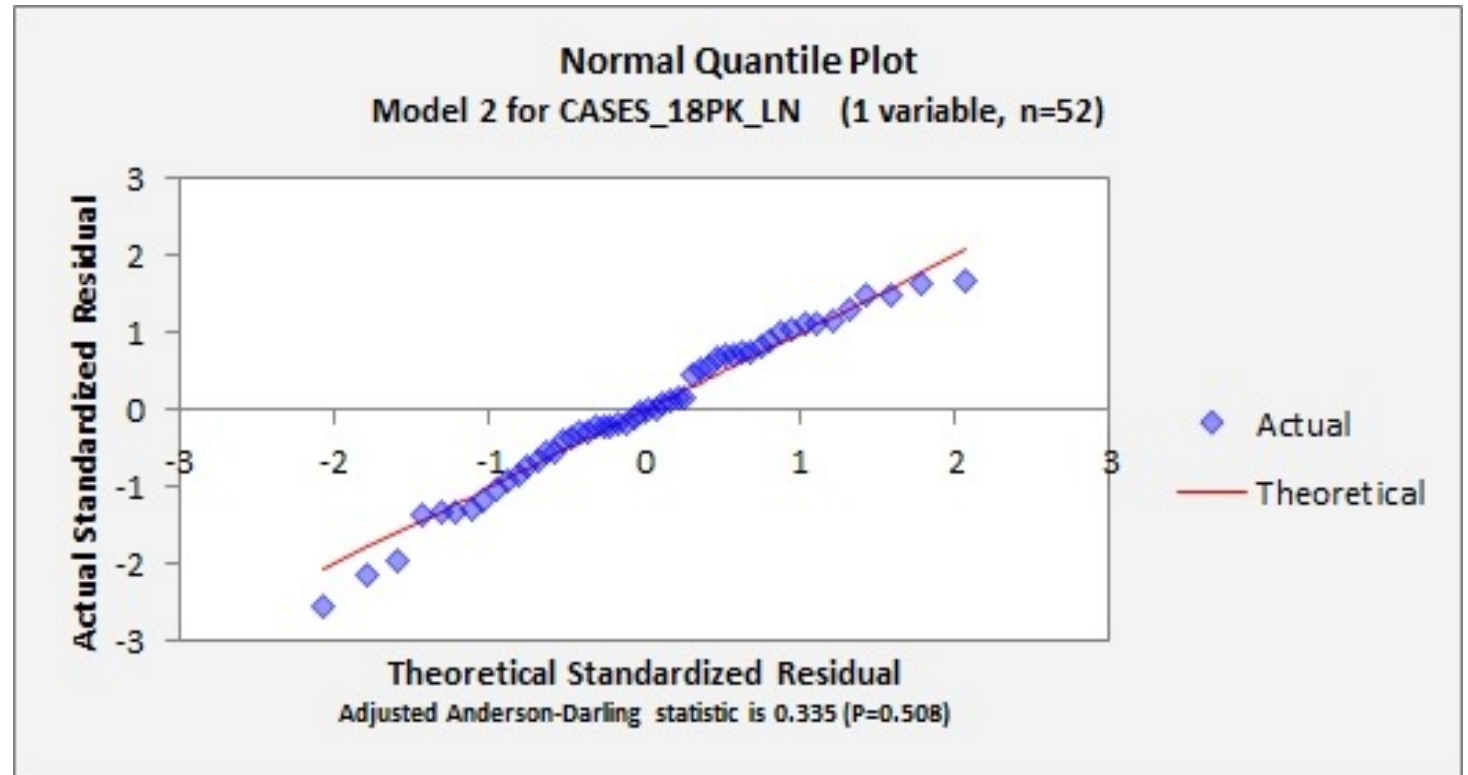
Alternative Hypothesis (H1): The sample does not come from a normal distribution.

If $p > 0.05$: Fail to reject the null hypothesis (suggests normality)

If $p \leq 0.05$: Reject the null hypothesis (suggests non-normality).



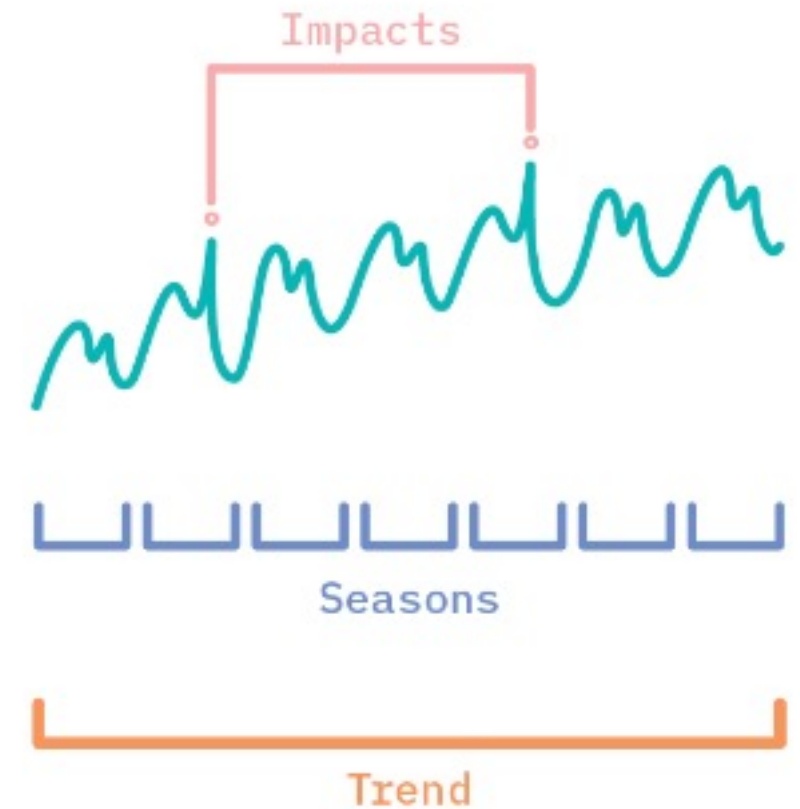
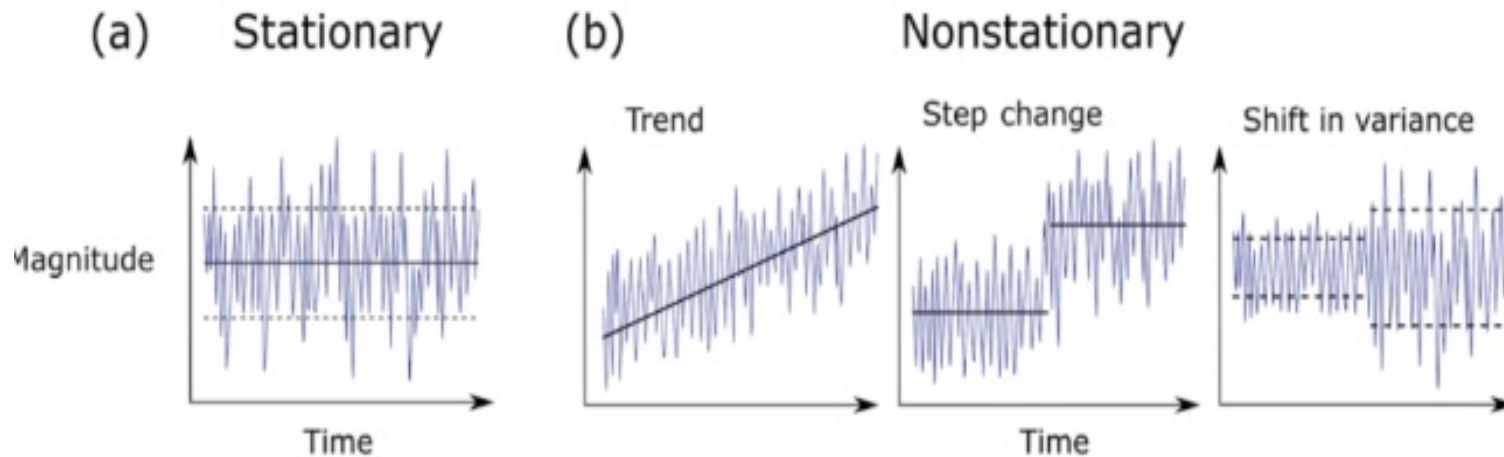
- Solving Non-Normality:
- **Transforming the Data: Log Transformation:** Apply a logarithmic transformation to the dependent variable or predictors.
- **2. Robust Regression:** WSL or GSL.
- **3. Outlier Detection and Treatment:** Identify and remove outliers by hand.



TIME SERIES

Stationary Time Series

- Stationary series are characterized by maintaining their mean, variance and autocovariance stable over time.
- That's why it is important to eliminate **seasonality**.



ARMA: AR(p) MA(q) (Univariate model)

The **ARMA(pq)** model is a time series forecasting technique. This model can predict future values based on past values and has two parameters:

1. Autoregressive (AR) part (p): This component models the relationship between an observation and *a number of lagged observations (previous values)*.



2. Moving Average (MA) part (q): This component models the relationship between an observation and a *number of lagged forecast errors or residuals*.

2. The number of lags in the time series estimation is identified as **(p)**

The size of the moving average is **(q)**

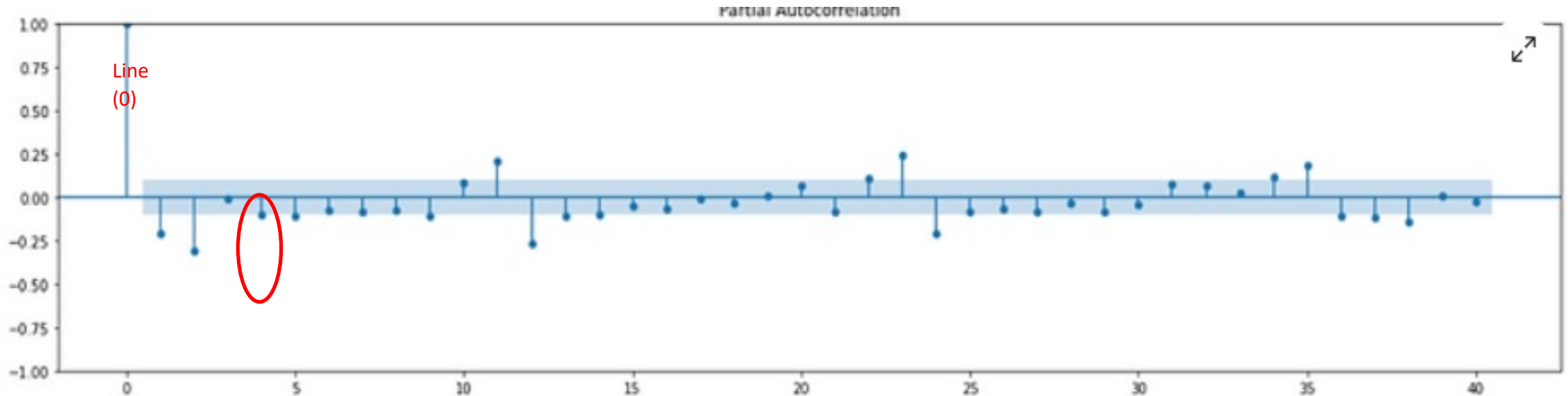
You can only predict with ARMA, *when the serie is already stationary*

How do we identify the **Autoregressive (AR(p))** in a serie?: We use visual inspection on a PACF (Patial Autocorrelation Function) plot for determining **p** values.

How to determine the p value?

PACF visualizes the direct contribution of the past observation to the current observations.

For example, the PACF below when lag = 3 the PACF is roughly -0.60, which reflects the impact of lag 3 on the original data point, while the compound factor of lag 1 and lag 2 on lag 3 are not explained in the PACF value. **The p values for the AR(p) model is then determined by when the PACF drops to below significant threshold (blue area) for the first time,** i.e. $p = 4$ in this example below.



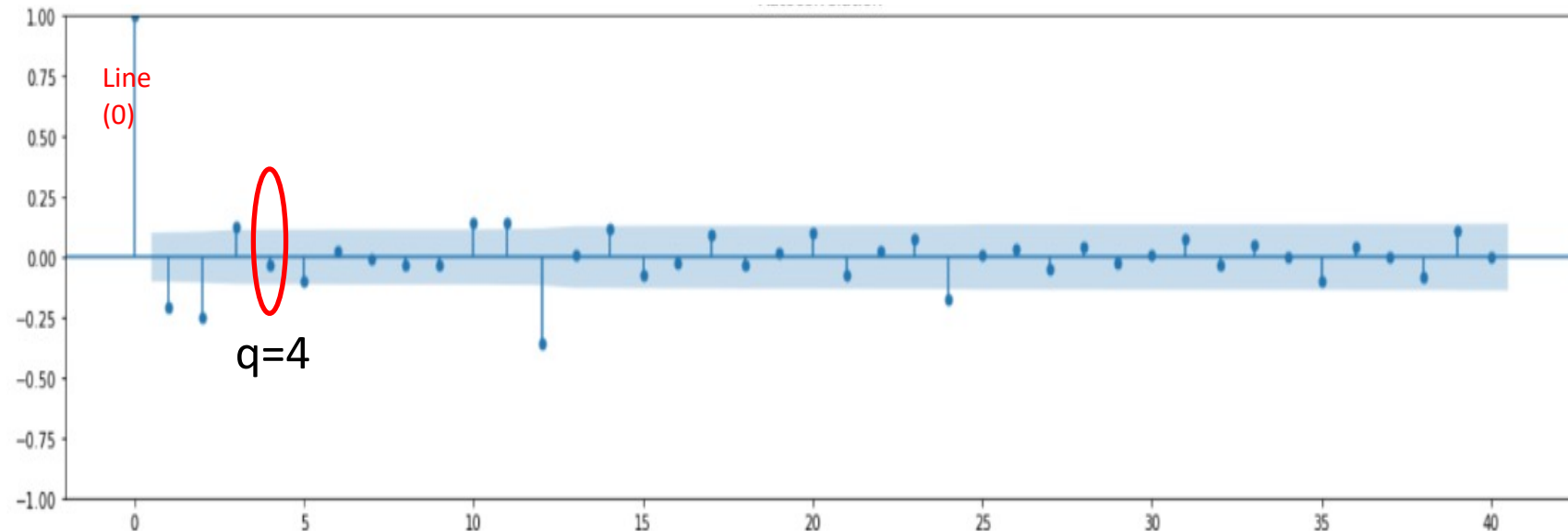
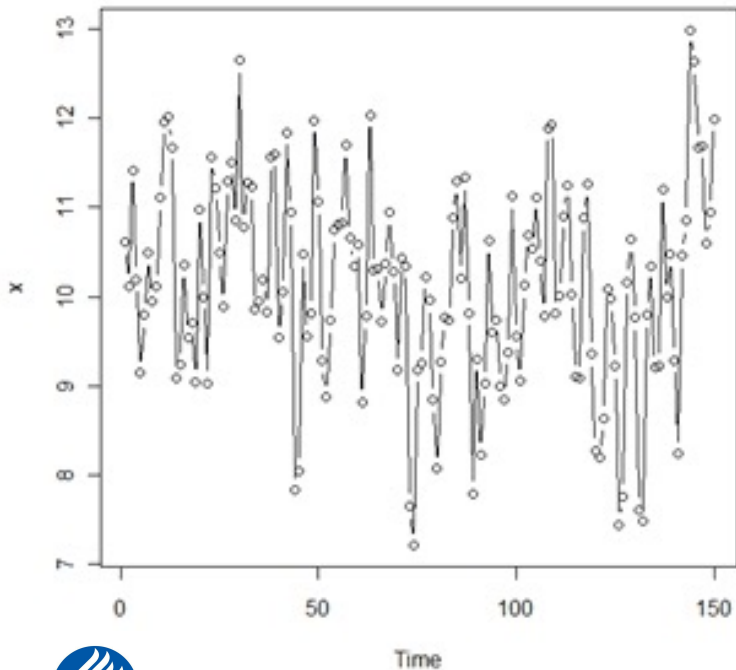
P= el PACF cae por debajo del umbral significativo (área azul) por primera vez.

How do we identify the **Moving Average (MA(q))** in a series?: We use visual inspection on a **ACF** (Autocorrelation Function) plot.

How to determine the q value?

It is typically selected as the first lagged value of which the ACF drops to nearly 0 for the first time. For example, we would choose **q=4** based on the ACF plot below.

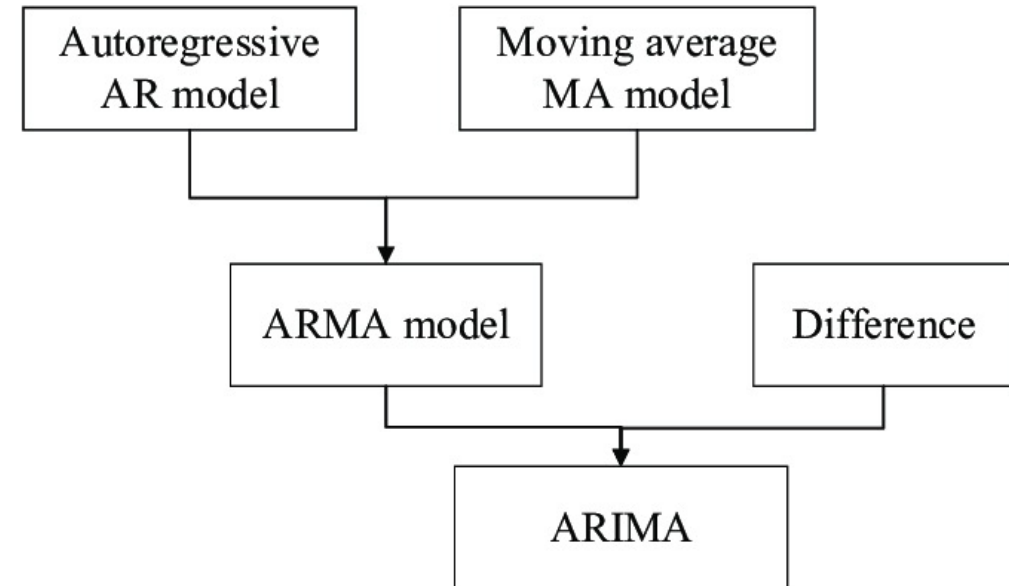
Simulated MA(1) data



An **ARIMA** (pdq) model is defined as a time series analysis and projection model that has:

- Autoregressive Components **AR**(p)
- Components of Moving Average **MA**(q)
- Unit Root Components **I**(1)... d

If an integrated order d process, is differentiated d times and their result is an **ARMA**(p,q) process, then we say that has an ARIMA process(p,d,q).



ARMA or ARIMA?

Augmented Dickey Fuller (ADF test) helps us determine whether a given time series is **stationary or not**.

Results of Dickey-Fuller Test:

Test Statistic	2.228722
p-value	0.998905
#Lags Used	12.000000
Number of Observations Used	280.000000
Critical Value (1%)	-3.453922
Critical Value (5%)	-2.871918
Critical Value (10%)	-2.572300
dtype:	float64

To test stationarity, we will use Augmented Dickey Fuller Test :

- **Null Hypothesis** : It assumes that the time series is non-stationary.

P-value > 0.05

Test statistic > critical value

- **Alternate Hypothesis** : If the null hypothesis is rejected, then the time series is **stationary**.

P-value < 0.05

Test statistic < critical value

SARIMA is an extension of the ARIMA model that incorporates **seasonality** in the data (Remember: single variable time series).

- It includes additional seasonal components to capture recurring patterns at fixed intervals (e.g., monthly, quarterly, yearly).
- SARIMA models are useful for time series with **significant seasonal variations**.

$$SARIMA \underbrace{(p, d, q)}_{non-seasonal} \underbrace{(P, D, Q)_m}_{seasonal}$$

SARIMAX

(Seasonal AutoRegressive
Integrated Moving Average with
exogenous regressors)

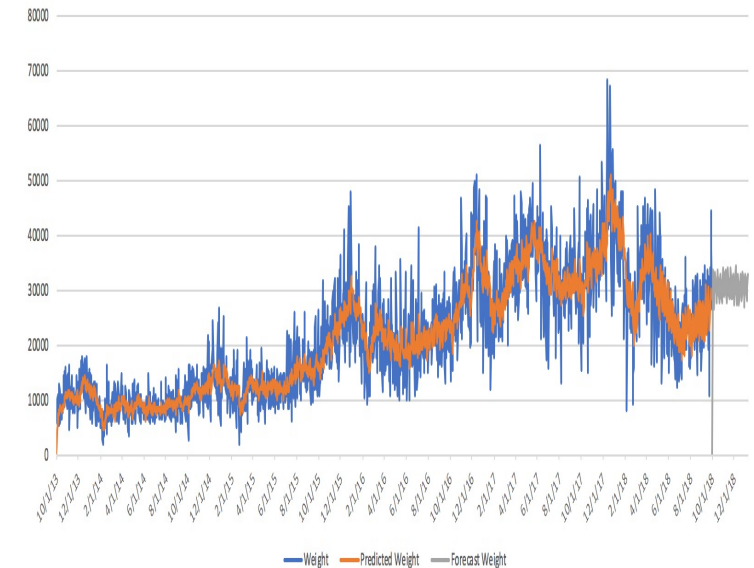
- It is an extension of ARIMA, but, it incorporates **both seasonal components and the effects of exogenous variables.**
- It is used when your time series data exhibit both seasonal patterns and are influenced by external variables.
- **Multiple variables (exogenous).**



SARIMAX

It has the ability to integrate exogenous variables as forecasting factors alongside the primary time series under consideration.

The only requirement for including an exogenous variable **is that we need to know the value of the variable during the forecast period.**



VAR (Vector autoregressive)

- “Multivariate Regression” (Y(dep (no more)))



We predict multiple target variables called “**endogenous**” at the same time through a system of equations.

$$\begin{cases} y_t = a_{11}y_{t-1} + a_{12}x_{t-1} + \epsilon_{1t} \\ x_t = a_{21}y_{t-1} + a_{22}x_{t-1} + \epsilon_{2t} \end{cases}$$

Here, a_{11} , a_{12} , a_{21} , and a_{22} are the coefficients, and ϵ_{1t} and ϵ_{2t} are the error terms.

In a VAR model we regress a vector of time series variables on lagged vectors of these variables.

When to use it:

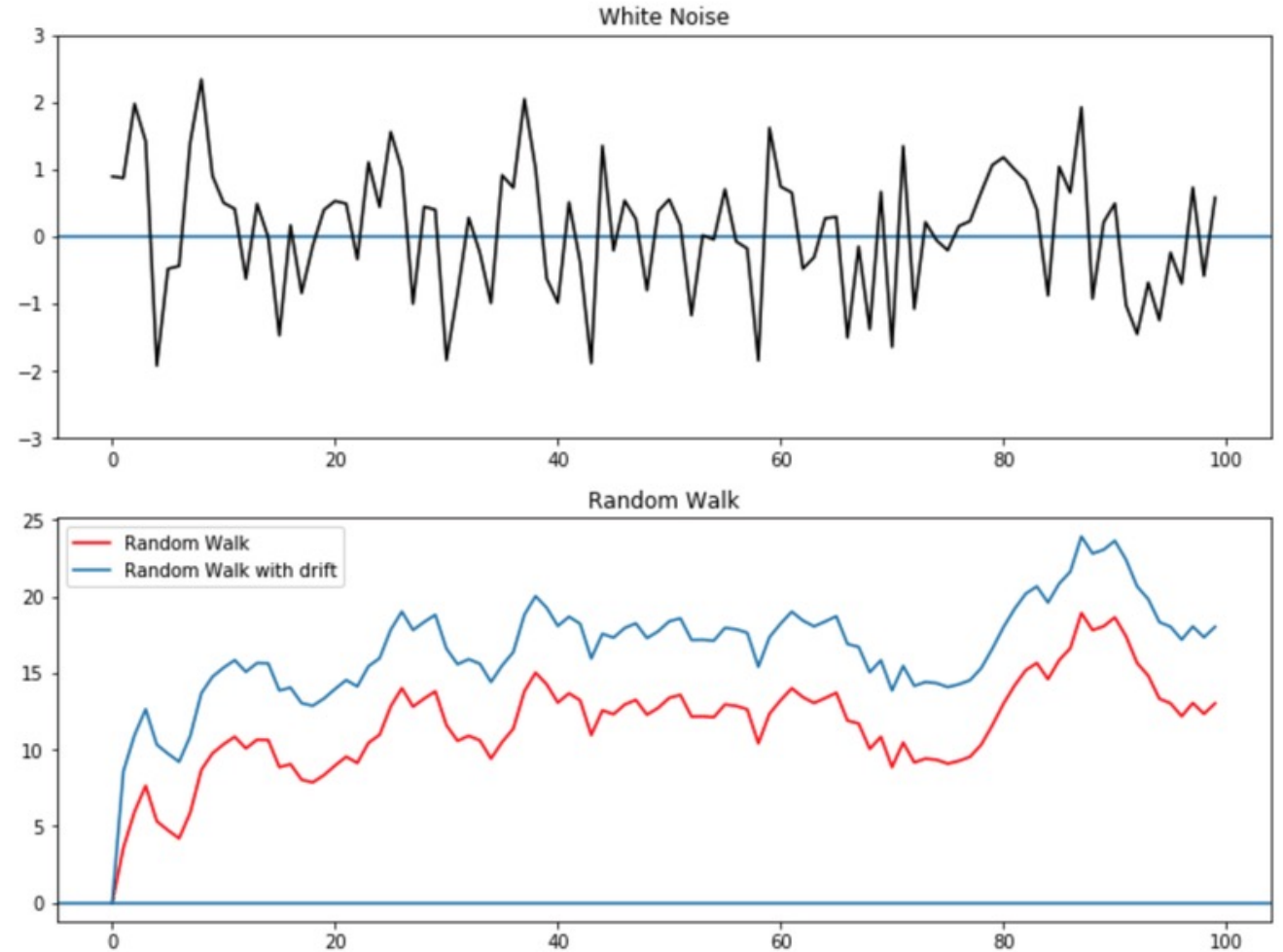
- **Macroeconomics:** To model the relationships between economic indicators.
- **Finance:** To analyze the interdependencies among asset, prices or returns.
- **Marketing:** To study the impact of marketing actions on sales and other performance metrics.

Stationarity:

A time series has **stationarity** when the observations are not dependent on the time. The process is called “white noise”.

Problem:

When a time series is **non-stationary** or follows a “**Random Walk model**” or has a **unit root**, the mean and variance, and covariance keep on changing, so it will not be possible to find an accurate inference.



Vector Autoregression (VAR) model

Features:

- **Multiple Time Series:** VAR predict future values based on **their own past values and the past values of all other time series (ARMA, ARIMA AND SARIMA can't do this).**
- **Lagged Values:** The model includes lagged (past) values of each variable as predictors.
- **Equations System:** A VAR model is essentially a system of equations, one for each time series variable.
- **Needed Stationarity:** Statistical properties such as *mean and variance do not change over time.*

Important steps to make a VAR

1) Check stationarity with the **Augmented Dickey-Fuller** test.

In the **ADF** test:

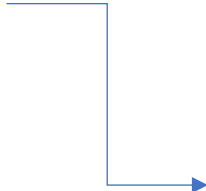
Ho: Time series is considered non-stationary.

Ha: Time series is stationary. p-value of the test should be < 0.05

2) Differentiate variables in order to turn non stationary variables into **stationary**, also use ADF

3) Determine the lag order (p): Apply

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).



Using AIC and BIC for selecting the lag order in VAR is essential for obtaining a model that adequately captures the dynamics between the time series and ensures that the model not only fits well to historical data but also performs well in future predictions.

The **VECM** model gives us estimates of short-run behaviour, **long-run cointegrating relationship** as well as short-run adjustment coefficients.



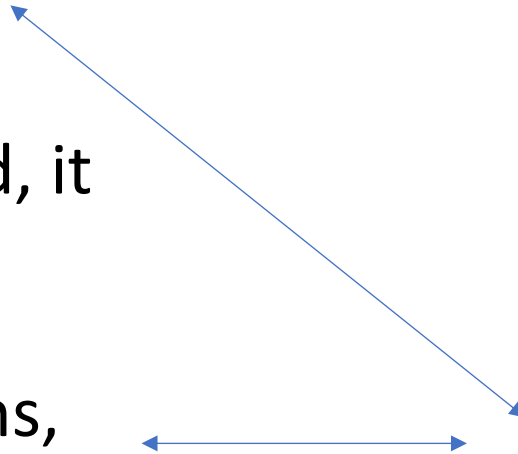
It's used when you have multiple non-stationary but cointegrated time series.

Short-Term Dynamics: It captures short-term deviations from the long-term equilibrium.

Long-Term Equilibrium: It corrects these deviations to maintain the long-term relationship between the series.

Cointegration

When 2 or more time series are cointegrated, it means that, **these individually are non-stationary** (their means, variances, or covariances are not constant over time), but there is a **linear combination of them that is stationary.**



VECM



Vector Error Correction Model

VAR X

Johansen test of cointegration (trace)

The Johansen trace test determines the number of cointegration relationships among several non-stationary time series. **Cointegrating vectors (r).**

a) Testing for $r=0$ (no cointegration relationships):

- Null Hypothesis (H_0): There are no cointegration relationships.
- Alternative Hypothesis (H_a): There is at least one cointegration relationship ($r > 0$).

b) Testing for $r \leq 1$ (at most one cointegration relationship):

- Null Hypothesis (H_0): There is at most one cointegration relationship among the variables
- Alternative Hypothesis (H_a): There are more than one cointegration relationships ($r > 1$).

c) Testing for $r \leq 2$ (at most two cointegration relationships):

- Null Hypothesis (H_0): There are at most two cointegration relationships among the variables ($r \leq 2$).
- Alternative Hypothesis (H_a): There are more than two cointegration relationships among the variables ($r > 2$).

Steps to make a VECM

Check for Cointegration:

- Use tests like the Johansen test to check if your time series are cointegrated.

Determine the Number of Lags:

- Use criteria like AIC or BIC to select the optimal number of lags.

Fit the VECM:

- Use software like Python's statsmodels to fit the VECM with the identified lags and cointegration rank.

PROBLEMS IN TIME SERIES

1) Non normality in residuals

- Diagnosis : Shapiro-Wilk test
- Null Hypothesis (H_0): The data are normally distributed.
- **p-value > 0.05**
- Alternative Hypothesis (H_a): The data are not normally distributed.
- p-value < 0.05

What to do:

Transform the variables into:

- a) Logs,*
- b) Root square,*
- c) Box-Cox transformation to normalize distribution*
- d) Use differences,*
- e) Use robust models such as: (ARIMA and SARIMA (univariate variable), SARIMAX or VAR (multivariate time series))*
- f) VECM**

2) Serial autocorrelation:

The value of a variable at one point in time is correlated with its values at previous points in time.

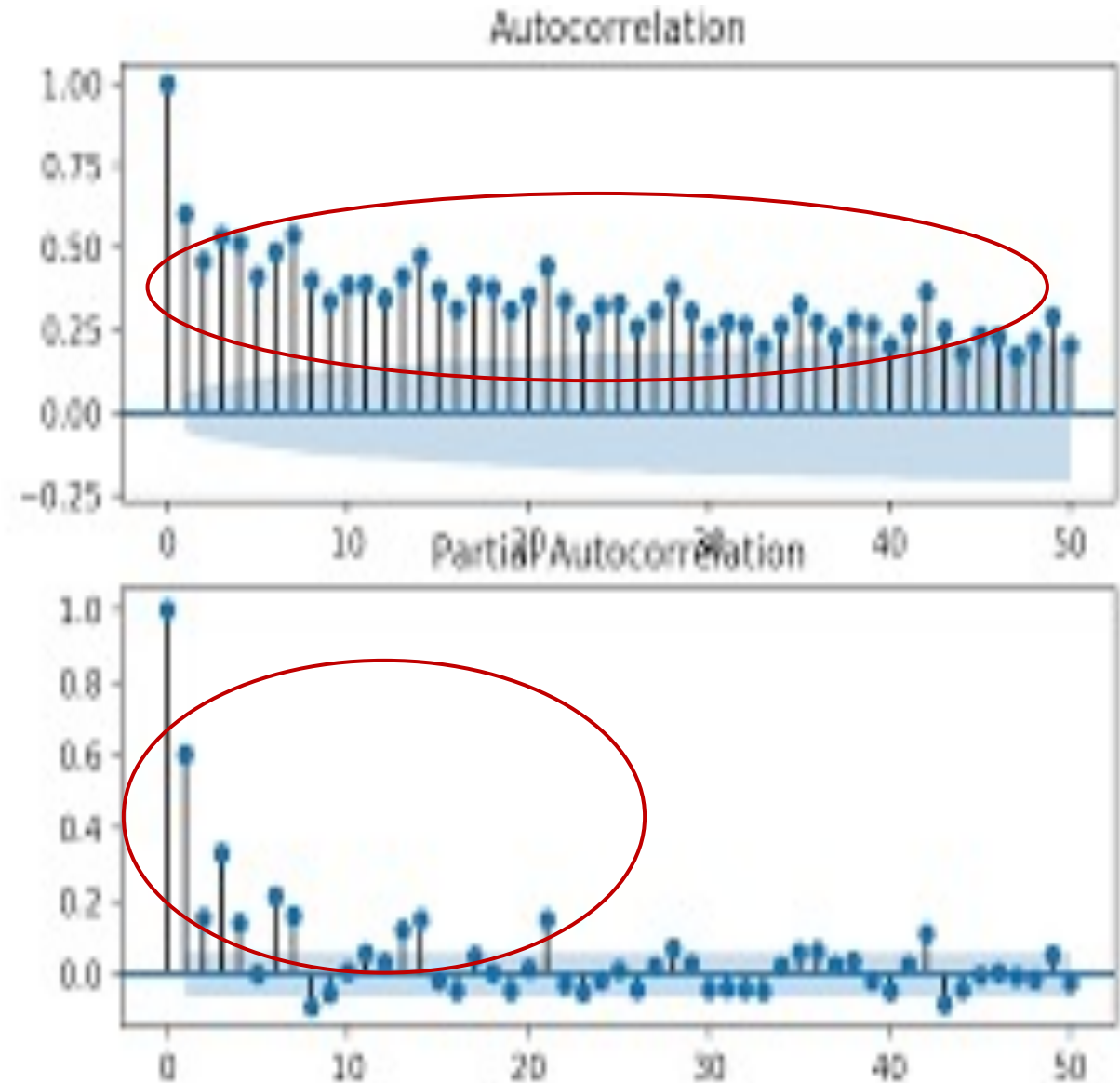
Why is it bad?

- Autocorrelated residuals can result in underestimated standard errors.
- leading to inflated t-statistics and misleading inferences.
- It can create the illusion of relationships between variables when none exist.
- This can lead to incorrect conclusions about causality and significance.

How to know if there is autocorrelation in a Time Series Model:

Autocorrelation Function (ACF) residuals plot shows the correlation of the time series model with its own lagged values. Check significant spikes outside the confidence intervals indicating autocorrelation.

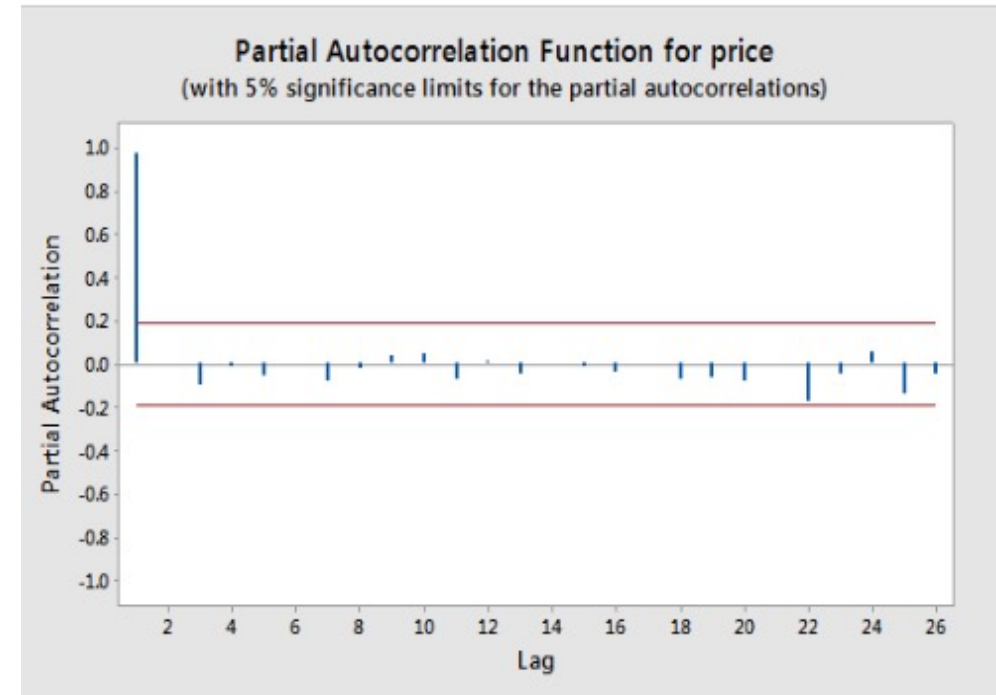
Partial Autocorrelation Function (PACF) plot shows the partial correlation of the time series with its own lagged values, controlling for the values of the time series at all shorter lags.



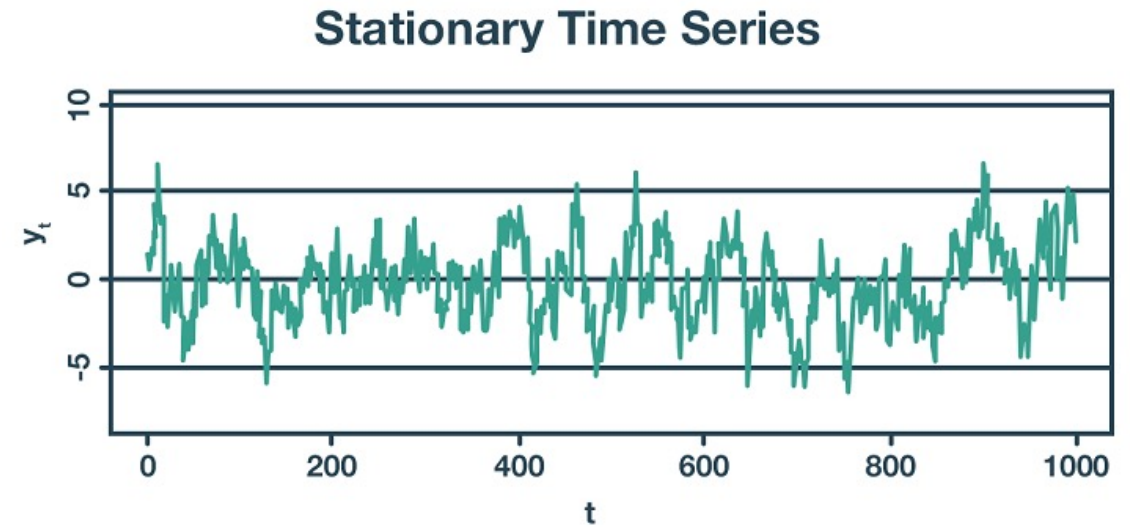
How to solve autocorrelation in Time Series?

- **Differencing:** Removes trends and achieves stationarity.
- **Seasonal Decomposition:** Removes seasonal components.
- **Using ARIMA model:** Captures autocorrelation by including AR and MA terms.
- Use **Generalized Least Squares Model (GLS)**: It transforms residuals to achieve:
 - 1) homoscedasticity (the variance of the errors is not constant over time. and
 - 2) non autocorrelation (when the residuals or the observations in a time series are correlated with each other).

It Corrects for biases in standard errors, leading to more reliable hypothesis tests and confidence intervals.



3) NON STATIONARITY IN VARIABLES





How to know if a time series is stationary or non-stationary:

a) Plot the Time Series: Look for trends & seasonality.

b) Apply Augmented Dickey-Fuller (ADF) Test:

A unit root test to check for stationarity.

Ho: Time series has a unit root (is non-stationary).

Ha: Time series is stationary when p-value of the test < 0.05

c) Apply KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test only for univariate time series.

Ho: Serie is trend stationary or series has no unit root. P value < 0.05

Ha: Serie is non-stationary, or series has a unit root.

How to handle non-stationary variables?

- **Log transformation to stabilize variance**

- **Differentiate time series** : 2 cases, apply:
ARIMA-SARIMA for univariate models

ARIMAX, VAR Model for multivariate time series

- **Look for cointegration:** Apply **Vector Error correction Model (VECM)**, but first make the **Johansen Test for Cointegration**.

Helps solve
Non Normality

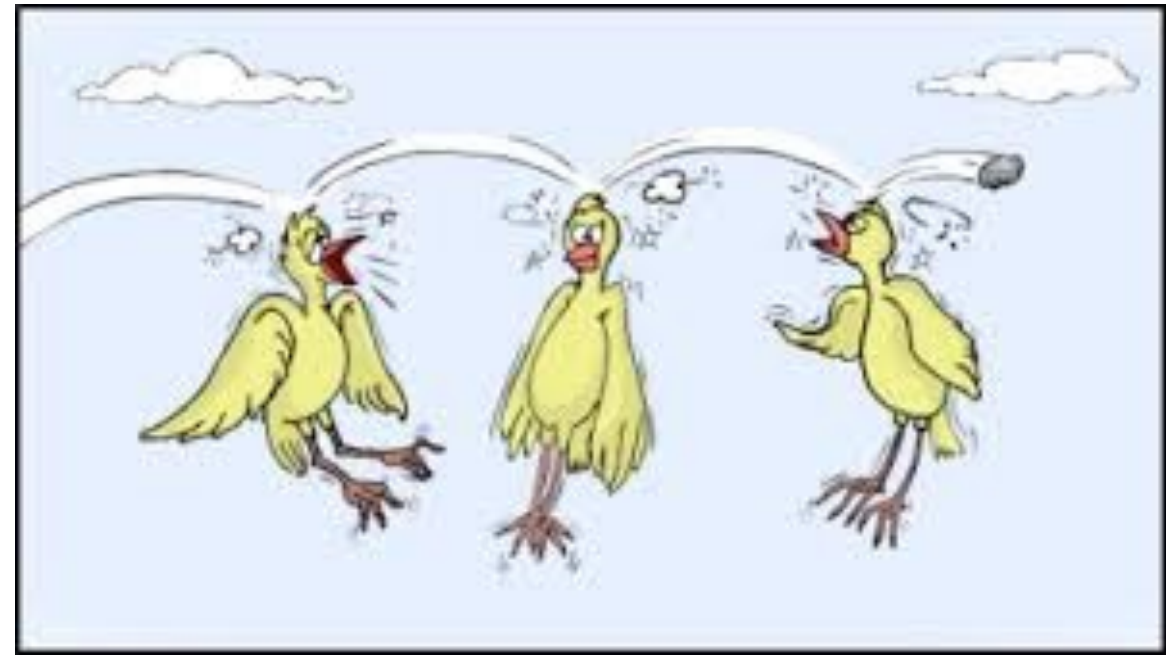


a) *Box-Cox transformation to normalize distribution... Use robust models*

b) *Use differences*

c) *Use robust models such as: (ARIMA and SARIMA (univariate variable), SARIMAX or VAR (multivariate time series)*

d) *VECM model-Johansen test of cointegration*



Helps solve
Non Stationarity

Helps solve
autocorrelation with
“right number of lags”

THANK YOU