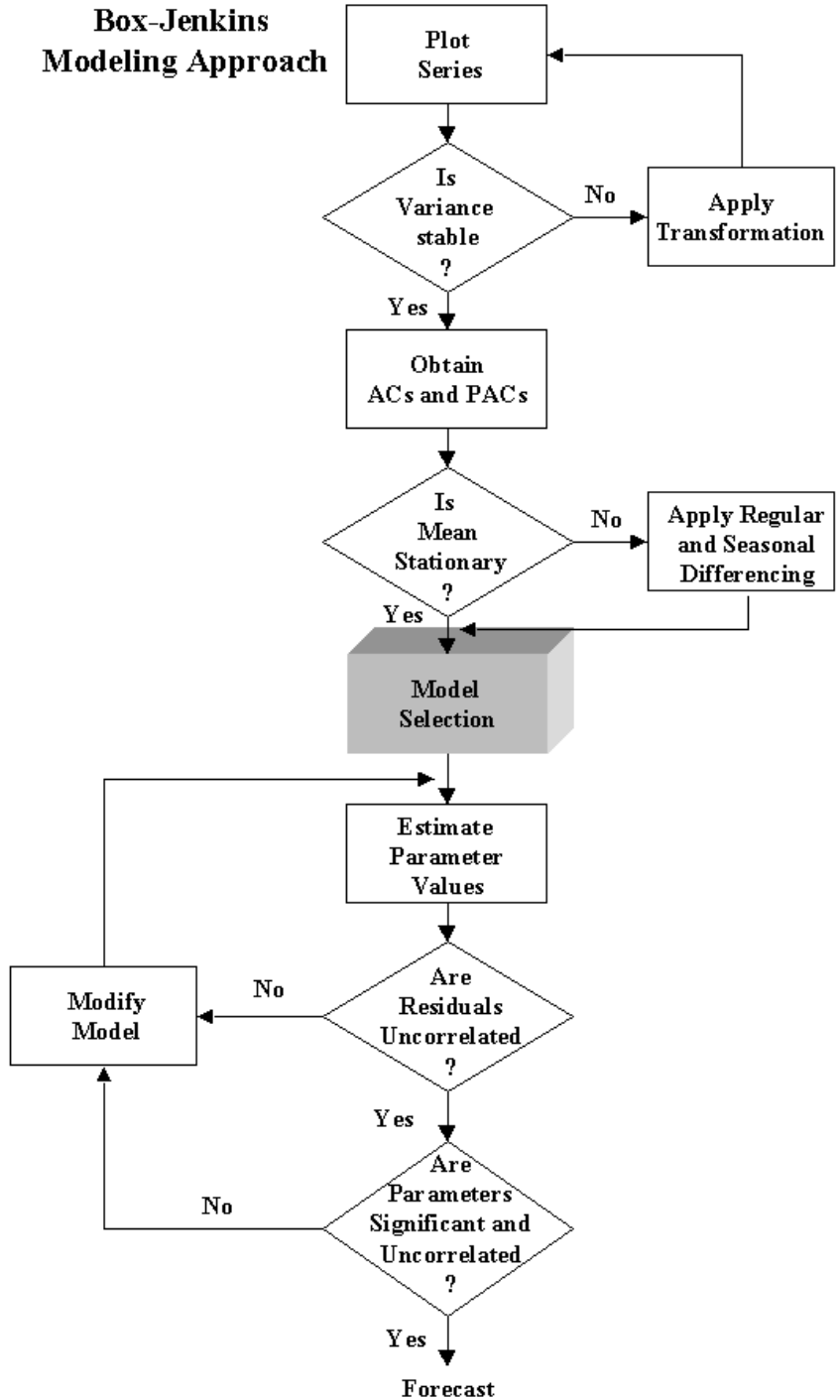
ARIMA Modeling (The Box-Jenkins methodology)

Dr. Khaing S Htun

The Box-Jenkins methodology

- G.E.P.Box and G.M.Jenkins (1970)
- a set of procedures for identifying, fitting, and checking **A**uto **R**egressive **I**ntegrated **M**oving **A**verage (ARIMA) models with time series data
- no independent variables in their construction
- rely heavily on autocorrelation patterns in the data

The Box-Jenkins methodology



ARIMA Modelling

- A time series method that fits a specific, statistical model/equation to past data
- The model fitted is either a moving average (MA), and autogressive (AR) or a mixed (ARMA) equation
- AR(p) model:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

- MA(q) model:

$$Y_t = \mu + e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

- ARMA(p,q) model:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

Auto Regressive Model

- Series current values depend on its own previous values
- AR(p) model:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

- "Auto Regressive" in ARIMA - a linear regression model that uses its own lags as predictors
- Linear Regression models work best with the predictors are not correlated and independent of each other

Autoregression is a time series model that uses observations from previous time steps called lag variables as input to a regression equation to predict the value at the next time step.

Auto Regressive Model

- AR(p) model:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

p - the order of the autoregressive process

Example:

AR (2,0,0) or AR (2)

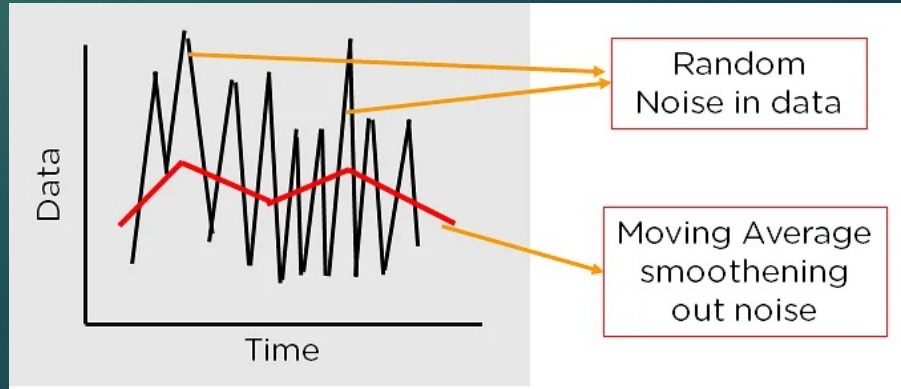
$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$$

Moving Average Model

- The current deviation from mean depends on previous deviations
- MA(q) model:

$$Y_t = \mu + e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

- a statistical method that takes the updated average of values to help cut down on noise
- the average over a specific interval of time



Moving Average Model

- MA(q) model:

$$Y_t = \mu + e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

q - the order of the moving average process

Example:

MV (0,0,1) or MV (1)

$$Y_t = \mu + e_t - \theta_1 e_{t-1}$$

Auto Regressive Moving Average Model

- ARMA(p,q) model:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}$$

where p - the order of the autoregressive part

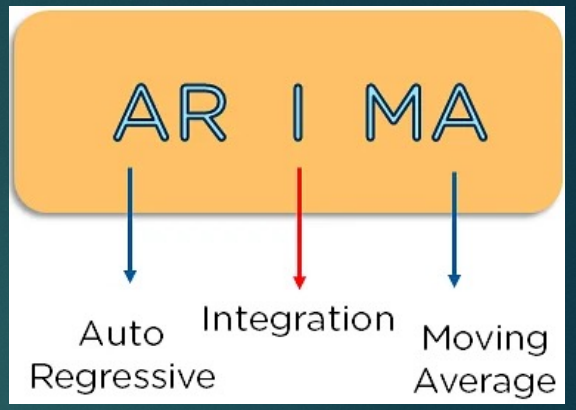
q - the order of the moving average part

Example:

ARMV (1,0,1) or ARMV (1,1)

$$Y_t = \phi_0 + \phi_1 Y_{t-1} e_t - \theta_1 e_{t-1} - \theta_1 e_{t-1}$$

ARMA and ARIMA models



- **ARMA**: Auto Regressive Moving Average
  - ARMA(p,q) = AR(p) & MA(q)

- **ARIMA:** Auto Regressive Integrated Moving Average
  - predict the future values of a time series using its past values and forecast errors
  - **Box-Jenkins** approach
  - can handle any series, with or without seasonal elements
  - ARIMA(p,d,q)
  - Integrated, d – differencing steps required **to make** time series **stationary**

Integration

ARIMA(p,d,q)

parameters are:

- p: Previous lagged values for each time point. Derived from the Auto-Regressive Model.
- q: Previous lagged values for the error term. Derived from the Moving Average.
- d: Number of times data is differenced to make it stationary. It is the number of times it performs integration.

ARIMA models

- ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. Specifically,
  - AR Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
  - I Integrated. The use of differencing of raw observations in order to make the time series stationary.
  - MA Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
- Each of these components are explicitly specified in the model as a parameter.
- Note that **AR** and **MA** are two widely used linear models that work on stationary time series, and **I** is a preprocessing procedure to "stationarize" time series if needed.

Stationarity & Differencing

- Any ARMA(p,q) model **must be fitted to stationary data**
- **Stationary** time series has **no upward or downward trend**
- A **trend** in the data can be **removed by differencing** one or more times (usually once is enough)
- First difference: $\Delta Y = Y_t - Y_{t-1}$
- ARIMA (p,d,q) model: An ARMA (p,q) model where Y has been differenced "d" times
  - ARIMA(1,1,1) – with first differencing

Stationarity

- statistical properties do not change over time
- constant mean and variance
- covariance is independent of time
- a flat looking series

ARIMA modelling process

- Box-Jenkins approach
- Selecting approriate models for estimating and forecasting univariate models
  - "Let my past predict you my future"

1. Identification – test for stationrity and identify the appropriate model
2. Estimation – fit the best model
3. Diagnostics (Model checking) – check residuals are white noise
4. Forecasting – use model equation to calculate forecasts

1. Identification

1. Is it Stationary?
   + If not, need to be differenced to remove the trend
     - If Stantionary, ARMA Model (p,q)
     - If Non stationary: ARIMA (p,d,q)

2. Identify Model
   + Use autocorrelation function ACF and partial autocorrelation function PACF to identify best ARIMA (p,d,q) model

Key Hint:

Parsimony: Adding more variables will increase the model fit (R-squared) at cost of decreasing degree of freedom

2. Estimation

1. Estimate parameters
   - + model coefficients, t-test, p-values
   - + May estimate other model to find best-fitting ones

2. Estimate the variance of the error

$$s^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-r} = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n-r}$$

e_t = the residual at time *t*
n = the number of residuals
r = the total number of parameters estimated

3. Diagnostics

Test the residuals to see of they are “white noise”

- Purely random or uncorrelated errors

Check

- Chi-square test
- Ljung-Box Q statistics
- ACF & PACF

If residuals are not white noise, need to try another model

4. Forecasting

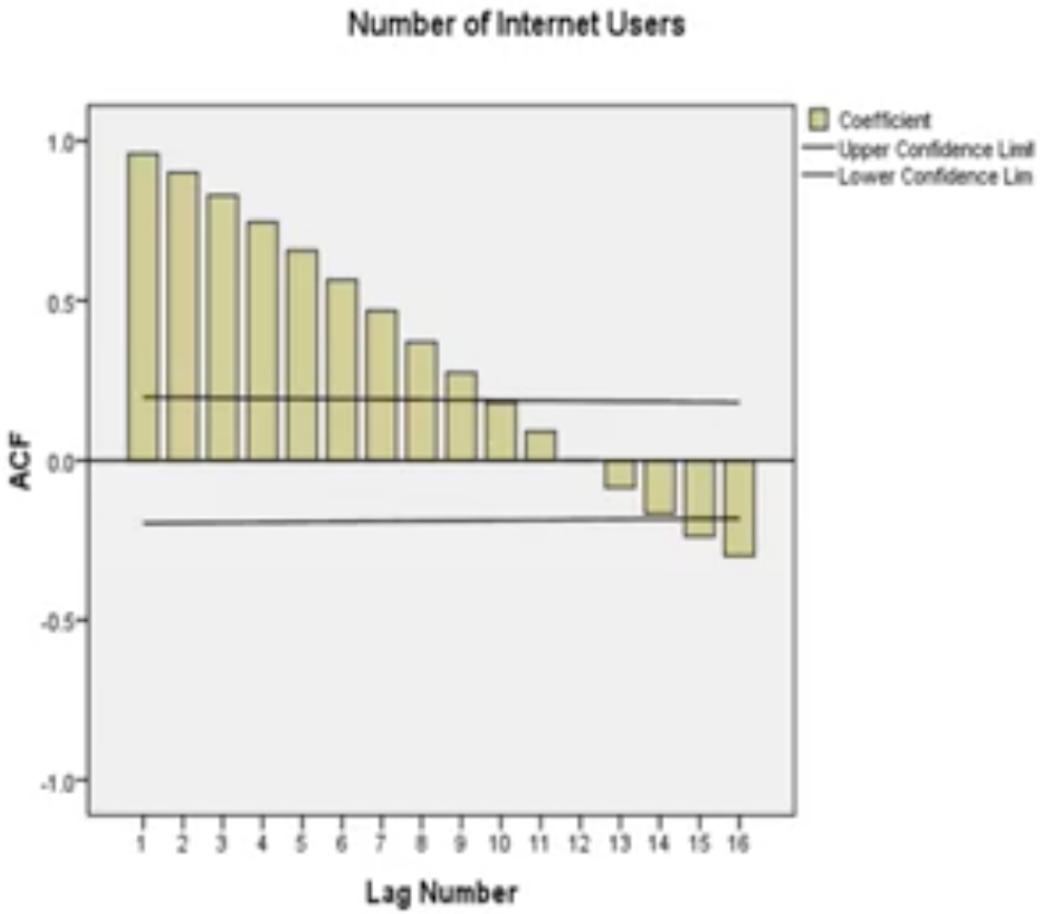
- Use the estimated model to generate forecasts
- the new data may be used to re-estimate the model parameters or, if necessary, to develop an entirely new model.

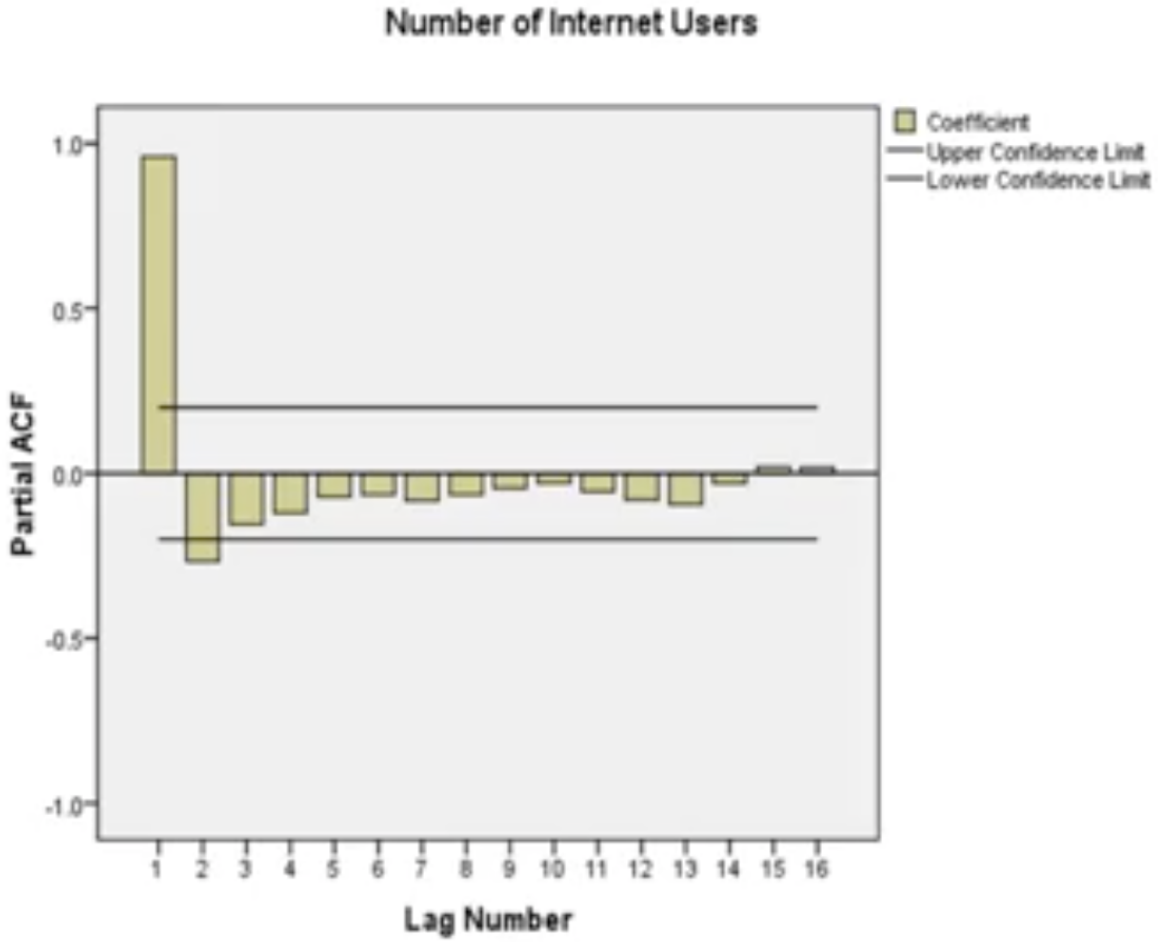
ARIMA Models for Seasonal Data

- an additional seasonal difference
- use same model-building strategy

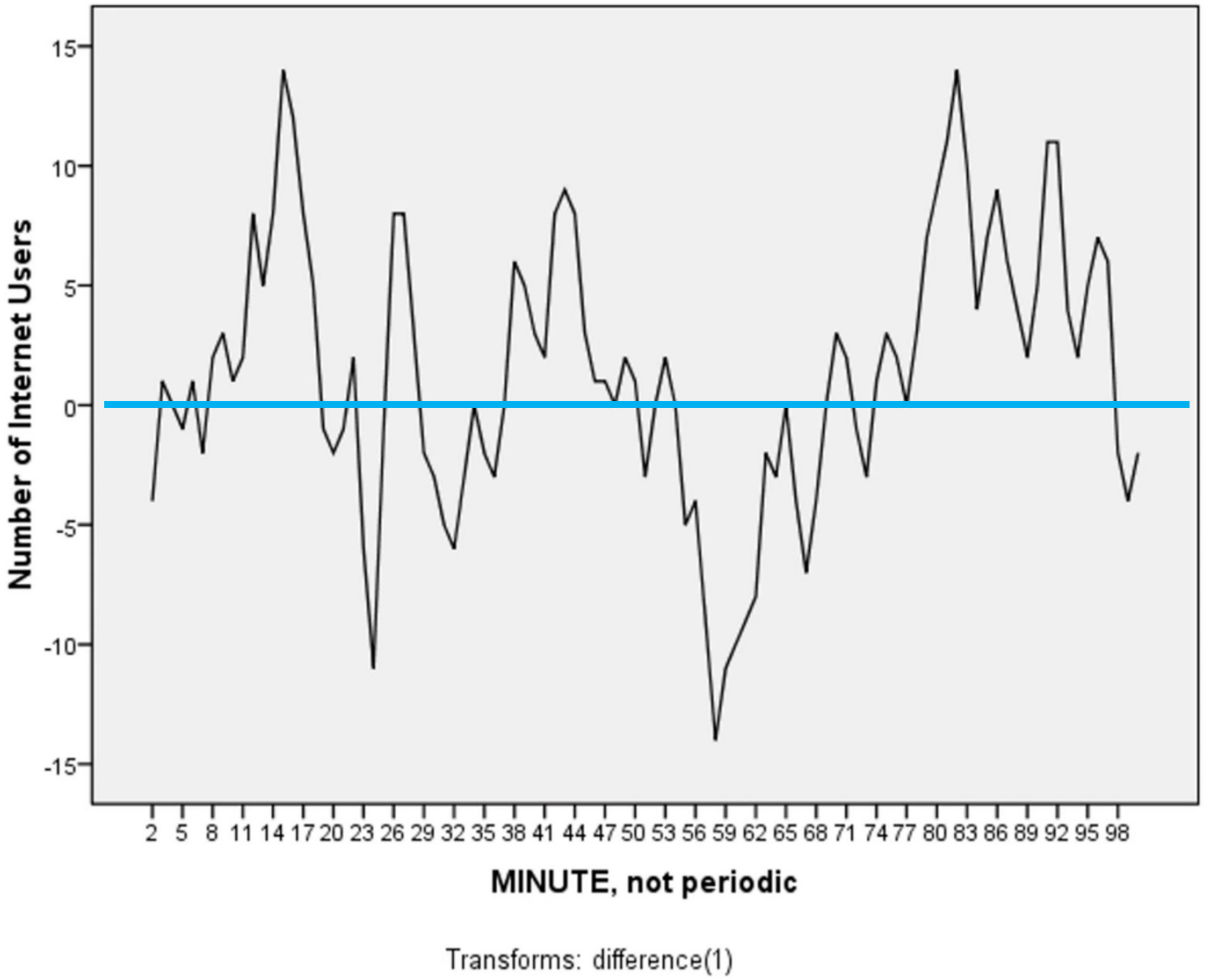
EXAMPLE:
INTERNET
USERS

Identification stage

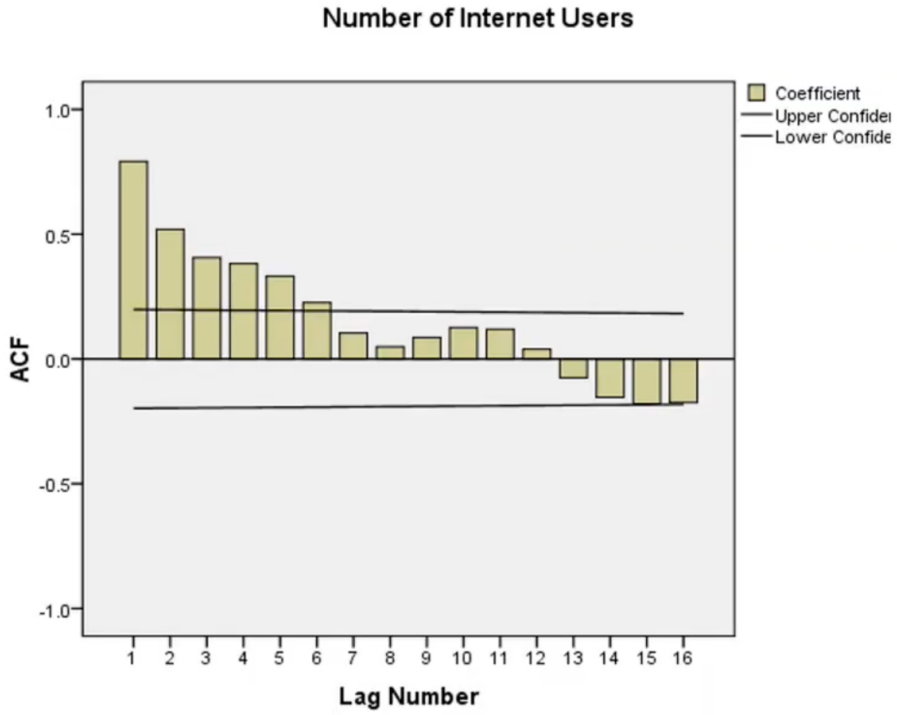


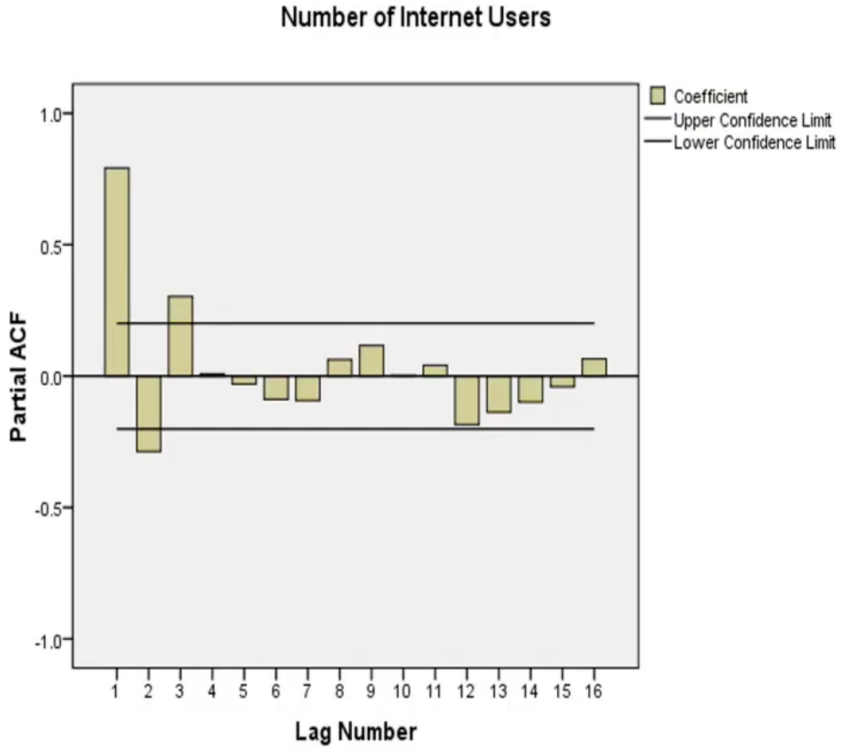


Identification stage



Identification stage





So, ARIMA(3,1,0) model seems appropriate

Estimation stage

ARIMA Model Parameters

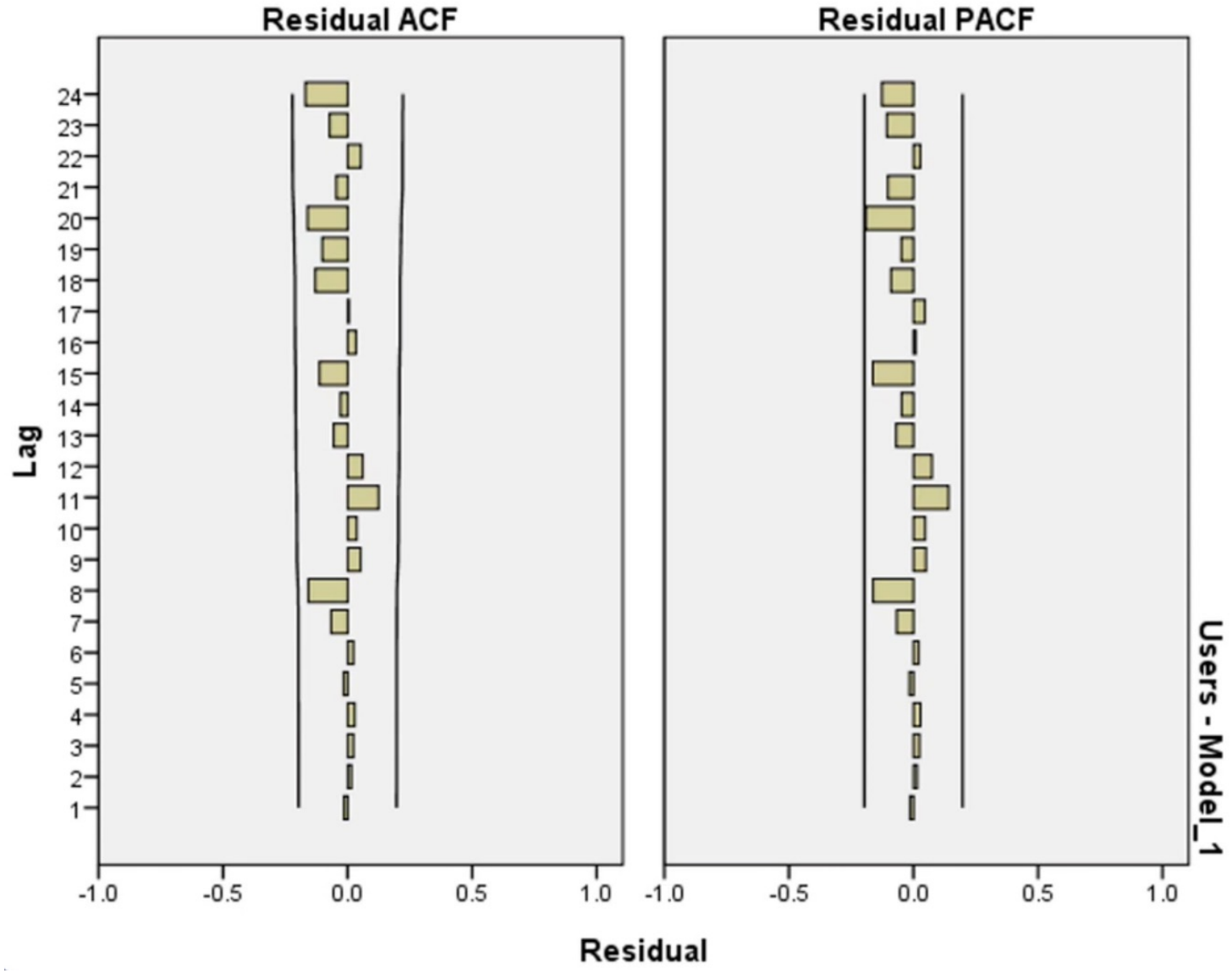
					Estimate	SE	t	Sig.
Number of Internet Users-Model_1	Number of Internet Users	No Transformation	AR	Lag 1	1.151	.097	11.931	.000
				Lag 2	-.661	.137	-4.827	.000
				Lag 3	.341	.096	3.547	.001
			Difference		1			

$$(Y_t - Y_{t-1}) = 1.151(Y_{t-1} - Y_{t-2}) - 0.661(Y_{t-2} - Y_{t-3}) + 0.341(Y_{t-3} - Y_{t-4})$$

(11.93) (-4.83) (3.55) (t-stat)

(0.000) (0.000) (0.001) (p-value)

Diagnostics stage



Forecasting stage

Forecast

Model		101	102	103
Number of Internet Users-Model_1	Forecast	220	219	218
	UCL	226	234	241
	LCL	213	205	196