Econometrics

By:

Manish Jalan

Director, Samssara Capital Technologies LLP (www.samssara.com)

Introduction

- Introduction
- Econometrics – Basics
- Regression Analysis
- Time Series – Volatility Analysis

Why Econometrics?

- Pattern recognition
- Forecasting
- Game theory e.g.: Chess

Econometrics Defined

- Application of maths and statistics to analyze economic data
- Quantitative analysis of actual economic phenomena based on the concurrent development of theory an observation, related by appropriate methods of inference
- Social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena
- Branch of knowledge concerned with empirical determination of economic laws

Econometrics - Basics

- Economic Measurement
- Application of Mathematical Statistics
- Model to obtain numerical estimates
- Based on Concurrent Theory and observation
- Empirical determination
- Not derived from controlled experiments
- E.g.: Economic data

Statistical vs. Deterministic Relationships

- Statistical – deal with random or stochastic variables
- Greek word meaning " a bull's eye"
- The outcome of throwing darts on dart board is a stochastic process, i.e, the process fraught with misses
- Deal with variables which have probability distributions
- E.g.: Normal Distribution

Deterministic variables

- Classical physics
- Exact relationships
- Deals with non random variables
- Non stochastic variables

Specification of Statistical/ Econometric Model

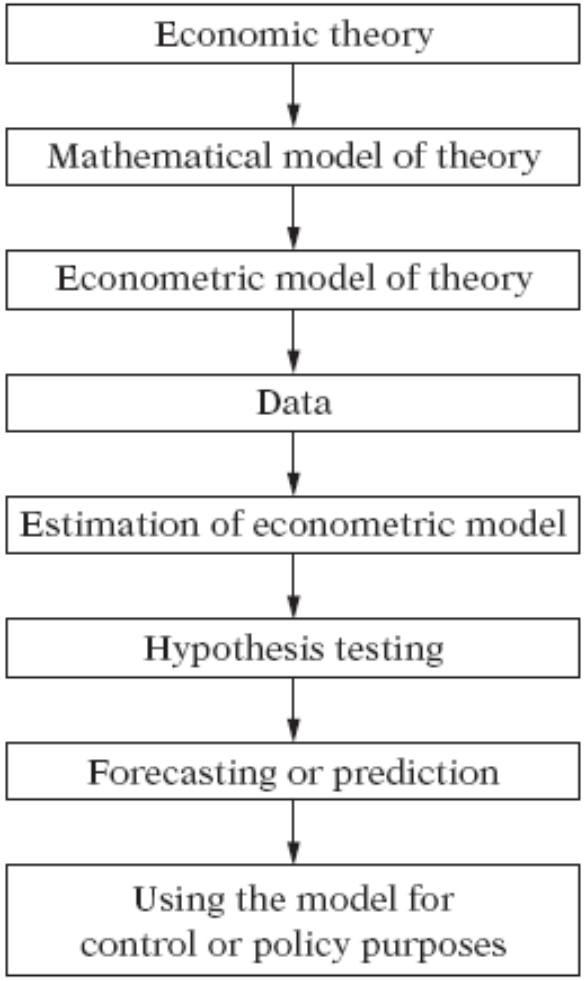
- Mathematical model assumes deterministic relationships
- Economic variables do not behave in exact manner
- Other variables will impact consumption behavior along with income
- There will be some error and it is captured by the disturbance term in the econometric model

$$y(i) = a + b*x(i) + e(i)$$

Choosing Among Competing Models

- Consumption Function
  - Keynesian consumption
  - Permanent Income Hypothesis: Consumption is function of long term income and not of current disposable income
  - Relative Income Hypothesis: Relation to others and individuals' previous attained consumption
  - Life Cycle Hypothesis: Consume constant percent of present value of life income and save the rest
- Production Function
  - Cobb-Doulas Production Function: labor and capital
  - Constant Elasticity of substitution
- New piece of evidence to empirical model (radiation effect and mobile phone purchase)

Anatomy of Econometric Model



Regression definition

- Study of dependence of one variable, dependent variable on one or more independent variables, with a view to estimating and or predicting the (population) mean or average value of the former in terms of the known fixed values(in repeated sampling) values of the latter.
- Used in
  - X to Y relation
  - Y to lagged variable relation (Unit root, ADF etc.)
- Heavy usage in forecasting

Regression- examples

- Scatter diagram – distribution of heights of sons
- Average – regression line
- Height to Age
- MPC (Consumption to Income)
- Monopolist – price elasticity to demand
- Phillips curve – money wages to unemployment

Regression- examples

- Relationship between inflation and k(proportion of people holding money, money/income)
- Elasticity of demand – advertising expenditure and sales relationship
- Crop yield depends on temperature, rainfall, sunshine, fertilizer
- Sales Vs Advertising
- Sales Vs Price of product

Regression Vs Causation

- Kendall and Stuart – A statistical relationship, however strong and however suggestive can never establish causal connection
- Ideas of causation must come from outside statistics, ultimately from some theory or the other
- A statistical relationship in itself cannot logically imply causation
- Variables might have measurement errors, reporting errors etc.
- No cause and effect relation: Crop yield effected by rain-fall and crop-yield cannot effect rain-fall

Regression Vs. Correlation

- Closely related but conceptually very different
- Primary objective of correlation analysis is to measure the strength or degree of linear association between two variables
- Correlation coefficient between smoking & lung cancer
- In regression the dependent variable has no random component

Statement of theory - hypothesis

- Statement of Theory or Hypothesis
- Keynes states that on average, consumers increase their consumption as their income increases, but not as much as the increase in their income (MPC < 1).
- Specification of the Mathematical Model of Consumption (single-equation model)

$$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1 \qquad \text{(I.3.1)}$$

Y = consumption expenditure and (dependent variable)

X = income, (independent, or explanatory variable)

β_1 = the intercept

β_2 = the slope coefficient

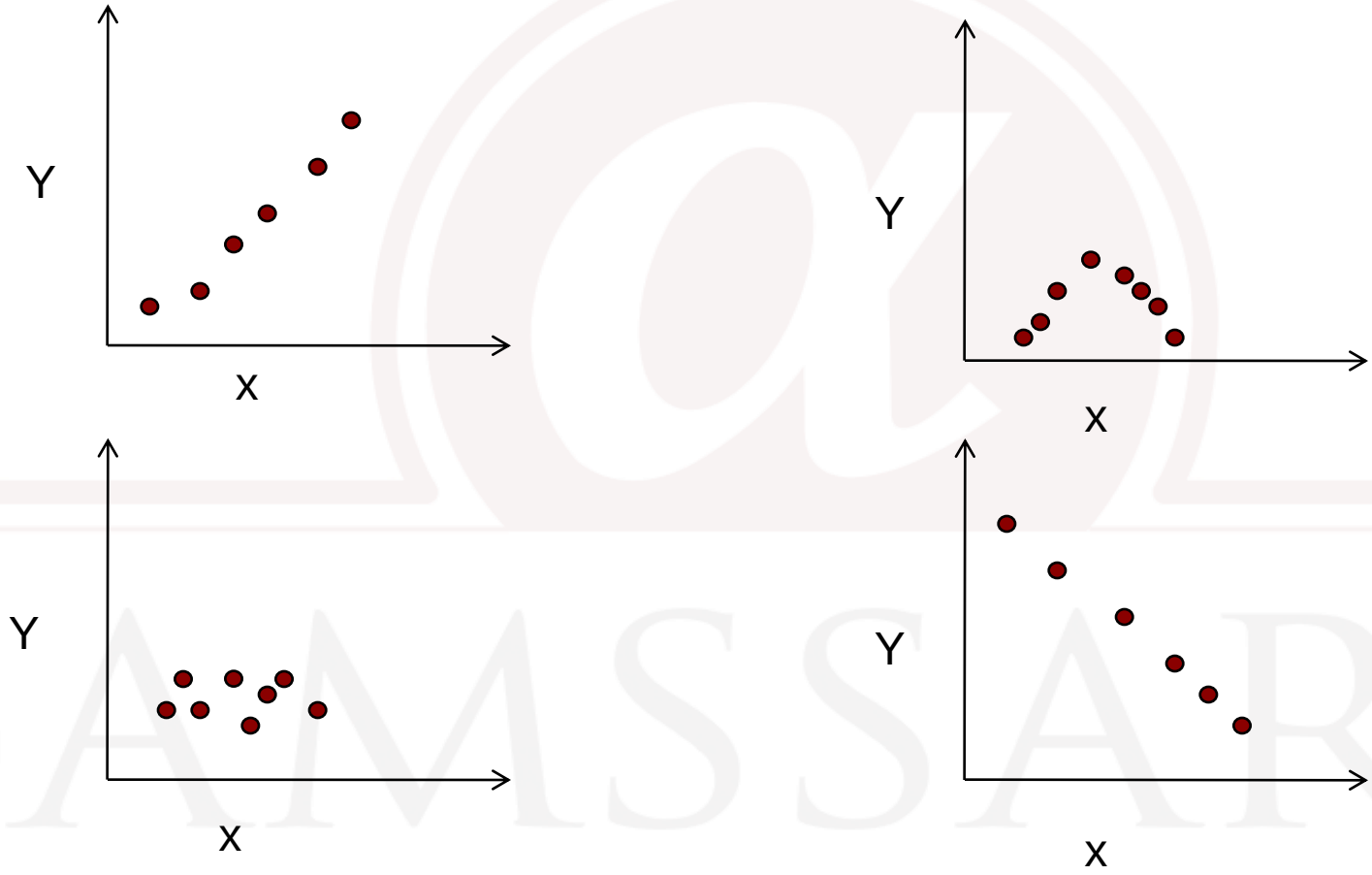
- The slope coefficient β_2 measures the MPC.

Types of Data & Sources

- Time series
- Cross section
- Panel data
- Pooled data
- Sources
- Official- NSSO, CSO, MOF, RBI
- Private – data vendors
- Survey data

Note: Plot and Check for outliers, gaps and jumps

Data Analysis & Plotting

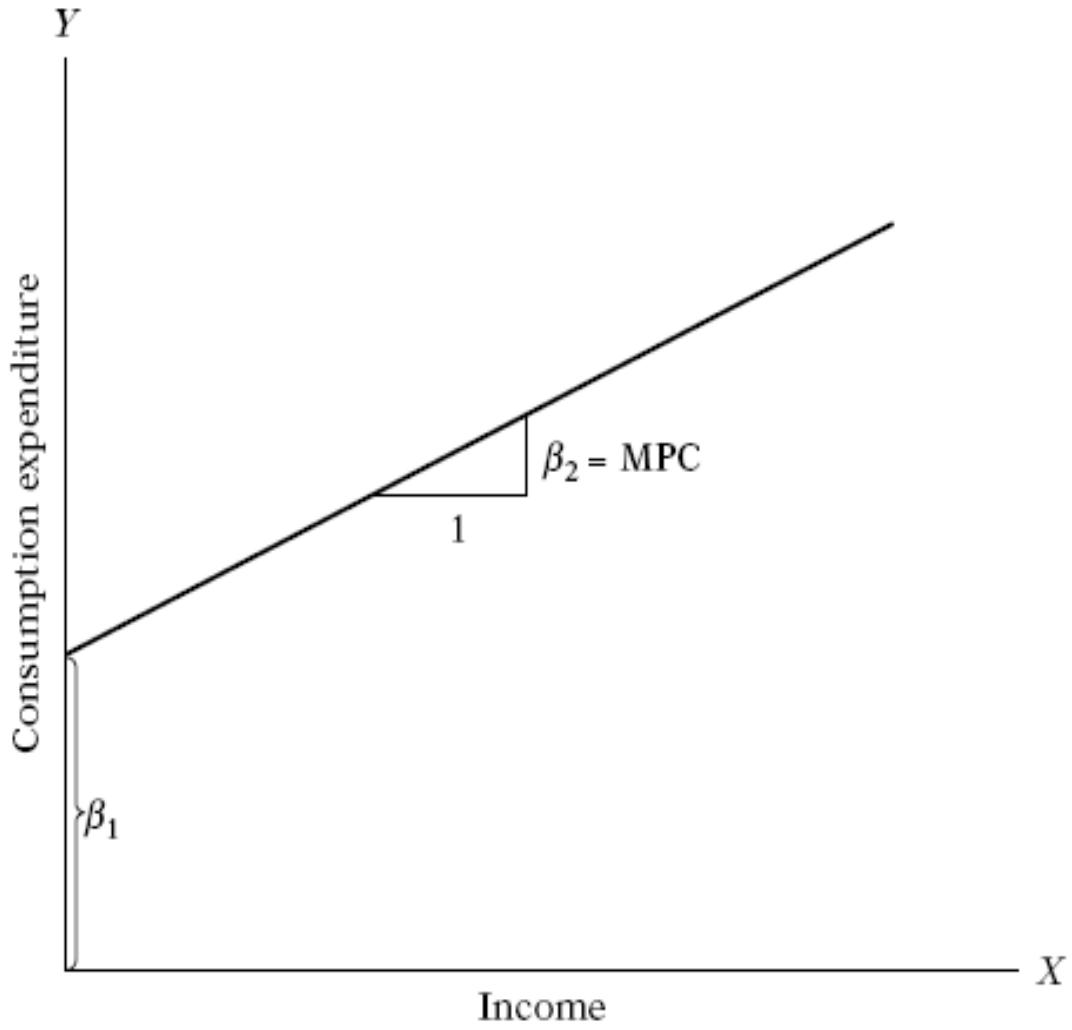


Obtaining data

Year	*Y*	*X*
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062.0
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486.0	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

The ideal hypothesis diagram

- Geometrically,



The estimation of the model

- Regression analysis is the main tool used to obtain the estimates. Using this technique and the data given in Table I.1, we obtain the following estimates of β_1 and β_2, namely, −99.3 and 0.69. Thus, the estimated consumption function is:
- $\hat{Y} = -99.3 + 0.69X_i$
- The regression line fits the data quite well. The slope coefficient (i.e., the MPC) was about 0.70, an increase in real income of 1 dollar led, on average, to an increase of about 70 cents in real consumption.

Regression in Excel

- Regression in Excel [Y = mX + C]
  - Slope: m
  - Intercept: C
  - Index(Linest,2): T-Stat of X

- TODO: Exercise on regression on excel after this slide

Forecasting

- Predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars consumption would be:

  $\hat{Y}_{1997} = -184.0779 + 0.7064\,(7269.8) = 4951.3$

- The *actual value* of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model thus over-predicted the actual consumption expenditure by about 37.82 billion dollars.
- We could say the *forecast error* is about 37.8 billion dollars, which is about 0.76 percent of the actual GDP value for 1997.

Use of model for policy

- Government believes that consumption expenditure of $4900 bln is required to keep the rate of unemployment rate at 4.2% what level of income will guarantee target amount of consumption expenditure.
- *4900 = -184.077 + 0.7064 X*
- Given MPC of 0.71, X = 7197 will produce 4900 expenditure.
- Control variable is income
- Target variable is consumption expenditure

The real world – prone with error

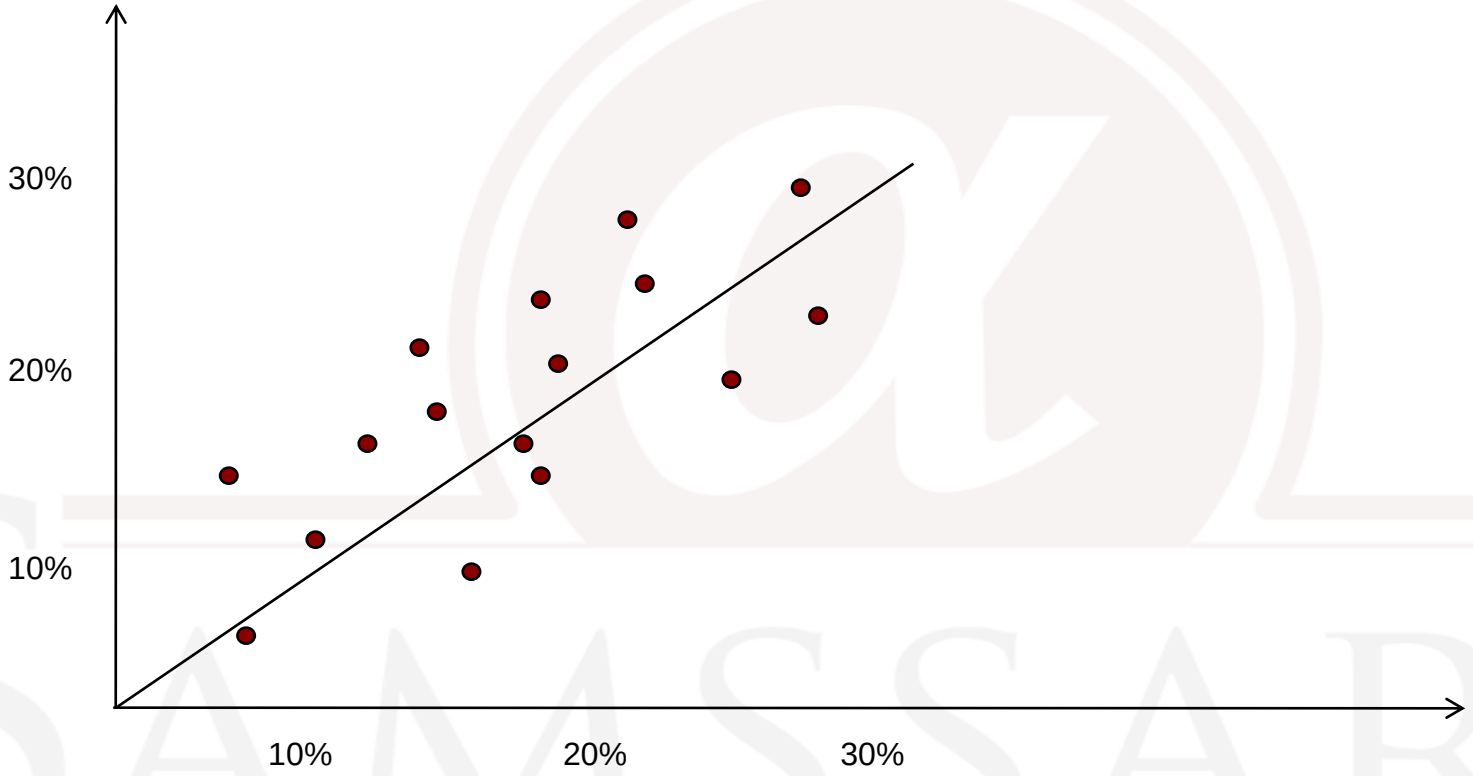
- Relationships are not exact
- Other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion, etc., are likely to exert some influence on consumption
- To allow for the inexact relationships between economic variables, the equation is modified as:
- $Y = \beta 1 + \beta 2X + u$
- where u, known as the disturbance, or error, term, is a random (stochastic) variable that has well-defined probabilistic properties. The disturbance term u may well represent all those factors that affect consumption but are not taken into account explicitly.

The estimation models and assumptions

Assumptions of OLS

- Error in estimation is what the REAL life is…!
- Model is *linear in parameters*
- The data are a *random sample* of the population
- The errors are *statistically independent* from one another
- The expected value of the errors is always zero
- The independent variables are not too strongly *collinear*
- The independent variables are measured *precisely*
- The residuals have *constant variance*
- The errors are normally distributed

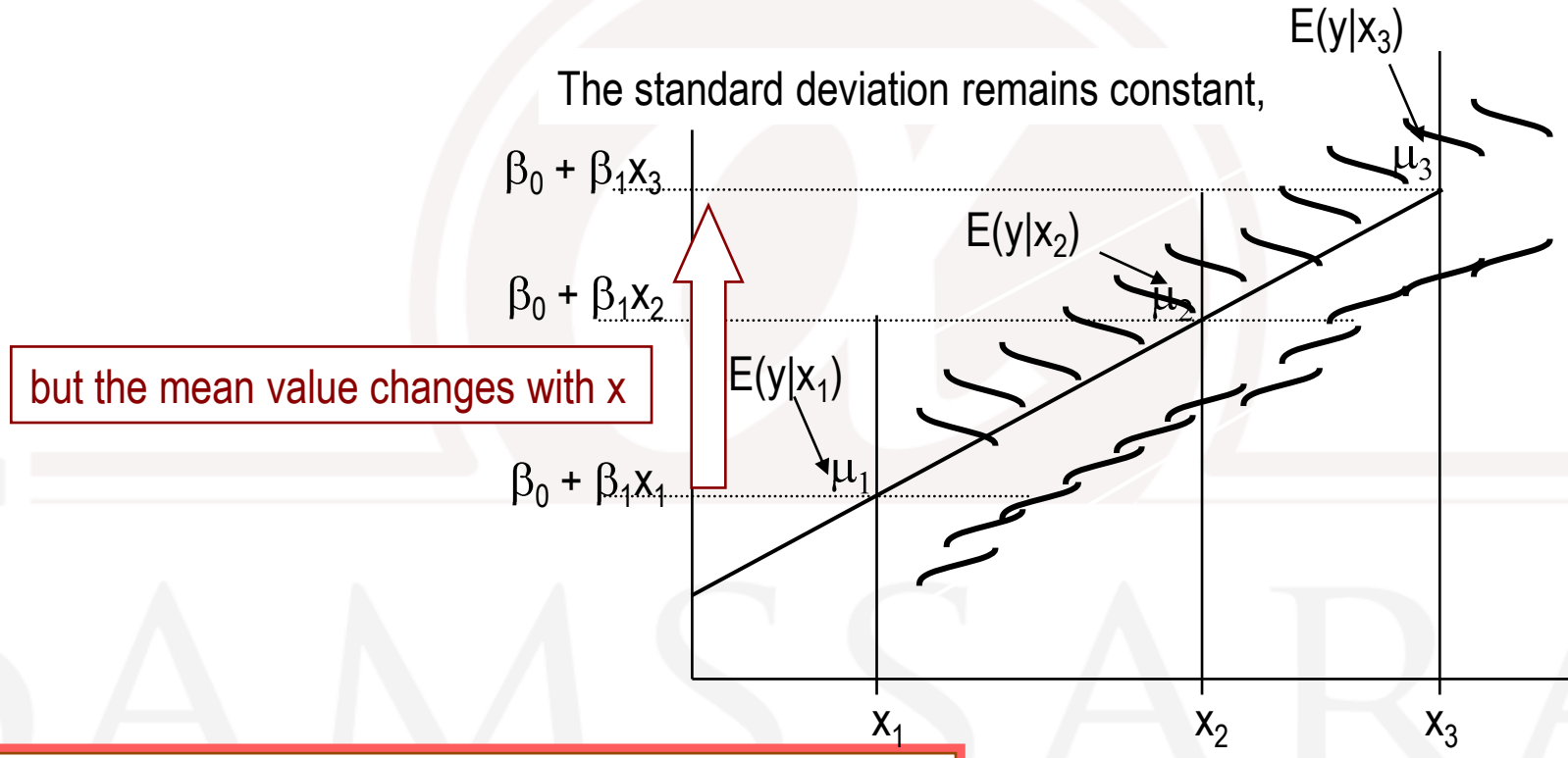
The errors in estimations



TODO: SEE (Standard Error Estimate) = SQRT [SUM (Errors)/ N-2]

Estimate of how much Real Y is away from the Regression Line

The Normality of ε



From the first three assumptions we have:
y is normally distributed with mean
$E(y) = \beta_0 + \beta_1 x$, and a constant standard deviation σ_ε

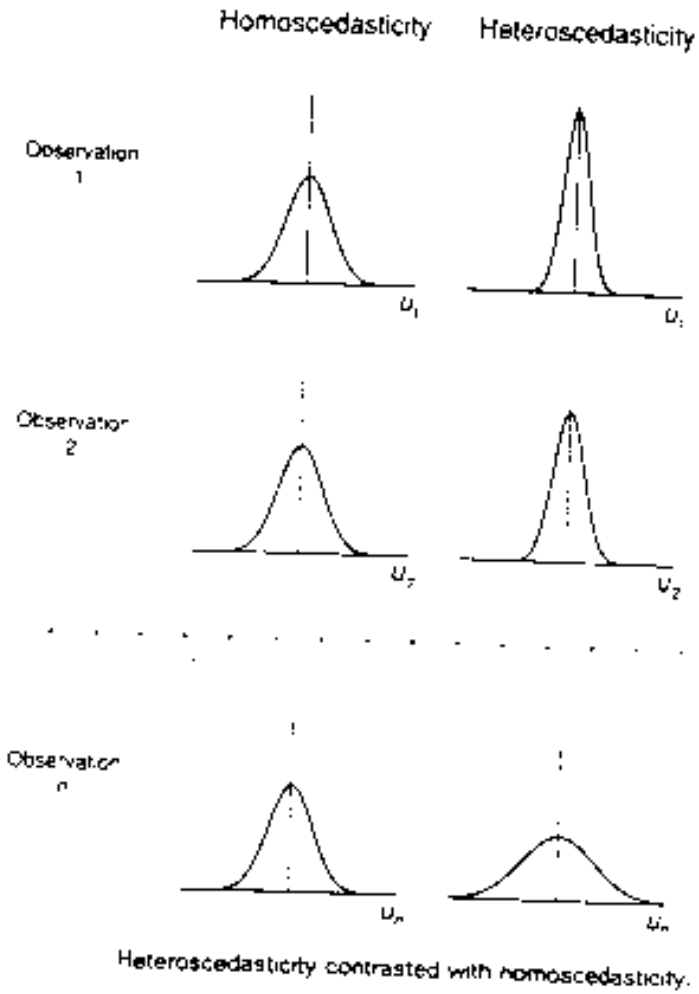
Errors not normally distributed

- **Problem:**
  - Parameter estimates are *unbiased*
  - P-values are *unreliable*
  - Regression fits the mean; with skewed residuals the mean is not a good measure of central tendency
  - The error is not distributed normally. Example, there may be fat tails. Consequence, use of the normal may underestimate true 95 % confidence intervals.

Homoscedasticity

- Equal variance
- Given the values of X, the variance of Ui is the same for all observations
- Var (u_i / Xi)=s^2 for all i.

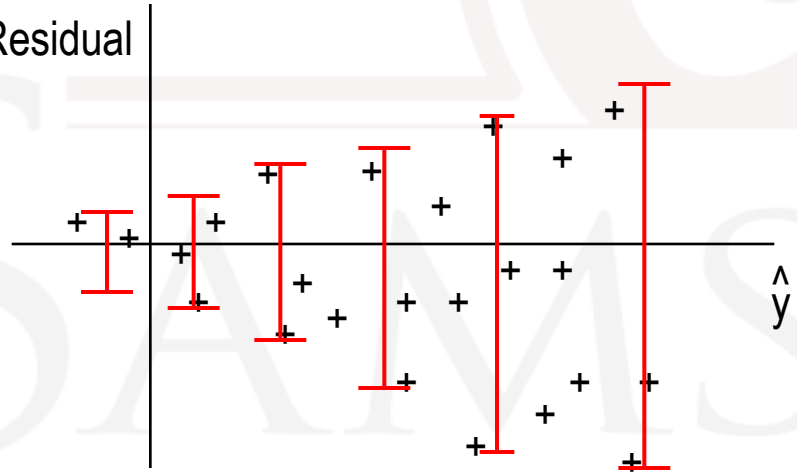
Homoscedasticity / Heteroscedasticity



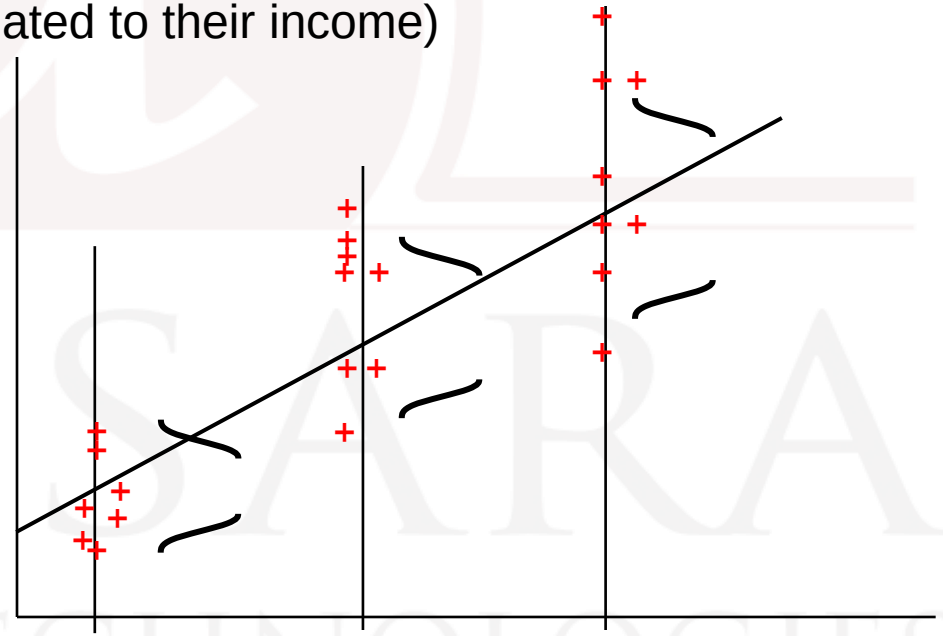
Heteroscedasticity contrasted with homoscedasticity.

Heteroscedasticity

- When the requirement of a constant variance is violated we have a condition of heteroscedasticity.
- E.g.: Higher Income households have higher savings => Erratic consumption and hence more error in estimation
- Lower income households (more closely related to their income)



The spread increases with ŷ

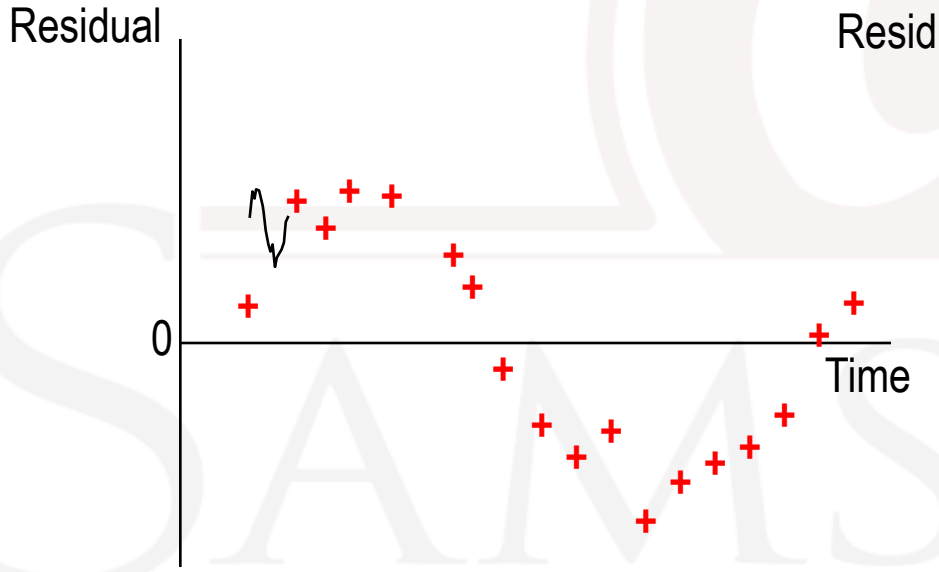


Correlation between errors

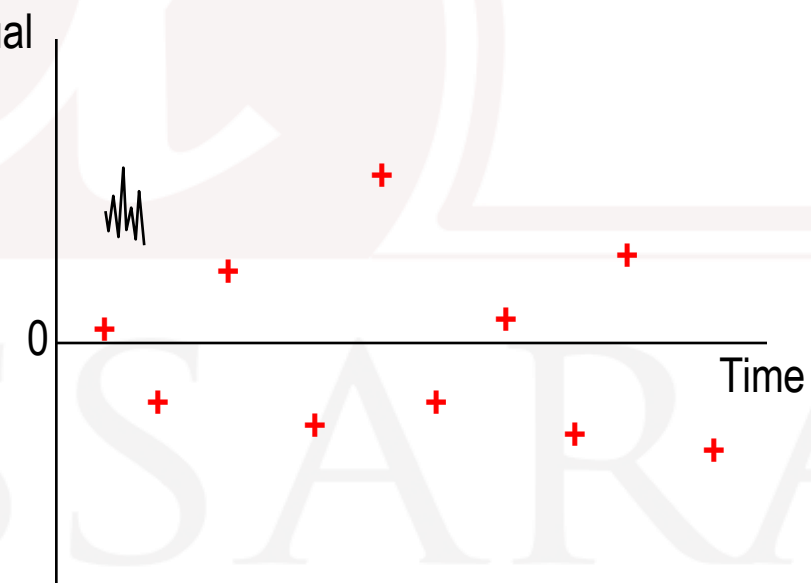
- When the error terms of 2 observations are correlated
  - E.g.: Expenditure (Y) = Income (X) + C + Error
  - Expenditure of n families
  - E.g. of auto-correlation: When one family spends money seeing the other family in their vicinity
  - Observations (Y) are not independent of each other and hence the errors are also related
  - Apart from Income the errors will also follow a pattern and hence yield a higher R-Square – leading to spurious regression
  - Consequence of auto-correlation
  - Variance in the error is under-estimated and hence R-Square is spurious and more

Non Independence of Error Variables

- Given any two X values Xi and Xj, the correlation between any two u and uj is 0
  *u*i and *u*j are independent for all *i* and *j*



Note the runs of positive residuals, replaced by runs of negative residuals



Note the oscillating behavior of the residuals around zero.

No Multicollinearity

- There is no perfect linear relationship among the explanatory variables *xi* and *x*j are independent for all *i* and *j*
- The independent variables may be highly correlated. As a consequence, they do not truly represent separate causal factors, but instead a common causal factor

Example of Multicollinearity

X2	X3	X3*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

It appears that X3i = 5X2i. Therefore, there is a perfect collinearity between X2 and X3 since the coefficient of correlation r_{23} is unity. The variable X3* was created from X3 by simply adding to it the following numbers which were taken from a table of random numbers: 2,0,7,9,2. Now there is no longer perfect collinearity between X2 and X3*. However the two variables are highly correlated because calculations show that the coefficient of correlation between them is 0.9959.

The diagnosis – non-normal errors

- Examining the residuals (or standardized residuals), help detect violations of the required conditions.
- Use Excel to obtain the standardized residual histogram
- E-views – draw histogram with coefficients of skewness and kurtosis
- Examine the histogram and look for a bell shaped. diagram with a mean close to zero.

The diagnosis – Multicollinearity

- Multicollinearity may be suspected if:
  - The t-stats of the coefficients of explanatory variables are not significant
  - The coefficient of determination (r-square) is high.
  - The correlation between the explanatory variable can then be calculated. To see if it is high.
- Collinearity does not destroy the property of minimum variance
- The OLS estimators and their std errors can be sensitive to small changes in the data
- E.g.: Consumption $_i = \beta_1 + \beta_2$ income $+ \beta_3$ Wealth $+ \mu_i$

The diagnosis – Auto correlation

- Autocorrelation: The Durbin-Watson statistic is a scalar index of autocorrelation, with values near 2 indicating no autocorrelation and values near zero indicating autocorrelation
- Examine the plot of the residuals in the view menu of the regression window in EViews
- Examining the residuals over time, no pattern should be observed if the errors are independent
- Autocorrelation can be detected by graphing the residuals against time

Remedial Measures

- "Do nothing" school of thought
- Rule of thumb procedures
- Combining cross-sectional and time-series data
- Dropping a variable/s and specification bias
- Transformation of variables
- Additional or new data
- Reducing collinearity in polynomial regressions
- Other methods
  - Factor analysis
  - Principal components

Correcting autocorrelation

- GLS method – If the coefficient of first order correlation is known – transform the equation by multiplying both sides with row value and apply OLS
- When row is not known – use first difference method – no intercept
- Iterative methods of estimating row value
- Dropping first observation
- Dummy variables and AC
- Savings Y and Income X model with D dummy taking 0 and 1 values
- ARCH and GARCH models

Is Multicollinearity necessarily bad?

- May be not if the objective is prediction only,
- E.g.: Y= number of people employed
- X1- GNP implicit deflator
- X2 = GNP
- X3 = number of people unemployed
- X4 = number of armed forces
- X5 = non institutionalized population over 14 years of age
- X6 = year , equal to 1 in 1947 , 2 in 1948 so on
- Obtain inter-correlations – correlation matrix
- Pair wise correlations high – reveal severe MC problem
- Auxiliary regressions – regress each x on all other xs

Summary of OLS assumptions

Violation	Problem	Solution
Nonlinear in parameters	Can't fit model	NLS
Non-normal errors	Bad P-values	Transform Y; GLM
Heteroskedasticity	Bad P-values	Transform Y; GLM
Nonlinearity	Wrong model	Transform X; add terms
Auto correlation	Biased parameter estimates	GLS
Measurement error	Biased parameter estimates	Hard!!!
Collinearity	Individual P-values inflated	Remove X terms

Regression Coefficients

- β_1 = Y bar - β_2 X bar
- $\beta_2 = \sum Xi\ Yi / \sum Xi^2$
- $Xi = X_i / x$ bar

Coefficient of Determination (r^2)

- TSS = ESS +RSS
- Total Sum of Squares = Explained Sum of Squares + Residual Sum of Squares
- r2 = ESS/ TSS = β (∑Xi2/ ∑ Yi2)
- T-stats and P-value

Basic Model Specification

- Trend Model
- Log linear model
- Semi-log model
- Double-log model
- Linear regression Model

Choice of functional form

- Underlying theory
- Find rate of change and analyse
- A priori expectations – demand function price should have negative coefficient
- If two functional forms give same results, we can take any form
- One should not rely on only r square to choose functional form

Trend Model

- Helps to compute growth rates
- Assumes linear trend
- Specification
- $Y = \alpha + \beta$ (trend)
- If Log Y is dependent variable and time trend is independent variable, beta gives the compounded growth rate of the dependent variable (numerical)

Semi-log models

- Log-Lin model / Lin-Log Model
- E.g.: GNP Vs. Rate of growth of money supply
- GNP = b1 + b2 lnM + e
- If b2 = 2000 => 1% increase in money supply increases GNP by 2000/100 = $20 billion

Log-linear model

- Log-linear model
- lnY = b1 + b2 lnX

$$b_2 = \frac{\Delta(\ln Y)}{\Delta(\ln X)} = \frac{\Delta Y * X}{\Delta X * Y}$$

- 1% increase in Money supply causes GNP Increase by 1%
- Production function
- Both variables are taken in log values
- Also known as log-log or double log models
- Constant elasticity models

Reciprocal model

- Y = b1 + b2 (1/X) + e
- Non linear variables
- X => Increase => 1/X approaches 0
  - Eg: Average Fixed cost declines with increase in output
- Phillips curve – wages and unemployment or inflation and unemployment
- Y = Rate of change of wage
- X = Unemployment rate / Inflation Rate
- b1: Base wage below which one cannot go

Log hyperbola or logrithmic hyperbola

- Log reciprocal
- Short run production model
- Log y is function of 1/x

Functional forms

- Reciprocal $Y_t = \beta_1 + \beta_2(1/t) + U_t$
- Log-Linear $Ln\ (Y_t) = \beta_1 + \beta_{2t} + U_t \quad Y_t > 0$
- Double – log $Ln(Y_t) = \beta_1 + \beta_2\ ln\ (t) + U_t \quad Y_t > 0$
- Logistic $Ln[\ Y_t/1-Y_t] = \beta_1 + \beta_{2t} + U_t \quad 0 < Y_t < 1$

Time series econometrics

Time Series Models violate OLS

- Time Series: Sequence of observation at different time
- In real life observations are auto-correlated
- E.g.: Expenditure in one year is influenced by expenditure in previous year (Maintain lifestyle)
- E.g.: Expenditure influenced by neighbors
- E.g.: GDP forecast higher in this year as compared to last year to “show” the GDP rising higher
- Errors have larger variance then shown by OLS
- Regressing one TS variable on other leads to spurious or non-sense regression

Practical Application of time series

- Gragner Casualty: GDP and Money Supply
- Co-integration / Unit root : Pair/High Frequency
- ARCH/GARCH: Estimation of volatility

Granger Causality Tests

- Can one time series forecast another?
- Powerful tool for cause effect
- Clive Granger won Noble Price in Economics
- X “causes” Y => X and Lagged X forecasts Y
- Which comes first: Chicken or Egg?
  - Result: egg causes chicken. No evidence that chicken causes egg.
- Example: Does the increase of world oil price influence the growth of US economy or vice versa?
  - Hypothesis that world oil price does not influence US economy is rejected. It means that the world oil price does influence US economy

Causality in Economics

- Time does not run backward…
- If event A happens before event B, then it is possible that A is causing B.
- However, it is not possible for B to cause A.
- Events in the past can cause events to happen today
- Future events cannot

The Granger test for causality

- X is said to Gragner cause Y if lagged values of X and lagged values of Y provide statistically significant information about future of Y
- GDP causes Money supply or Money supply causes GDP
- This test assumes that the information relevant to the prediction of the respective variables, GDP, M, is solely in the time series data on these variables
- Assumption of Ui and Uj are uncorrelated
- This is bilateral causality
- Multivariable causality – Vector auto regression

The Granger test

- Postulates that current GDP is related to past values of itself as well as that of M.
- Similar behavior for M
- 4 probable cases
  - Unidirectional causality from M to GDP
  - Unidirectional causality from M to GDP
  - Feedback, or bilateral – when sets of M and GDP coefficients are statistically different from zero
  - Independent – when the sets of M and GDP coefficients are not statistically significant in both regressions.

The Granger test

- Since the future cannot predict the past, if variable X causes Y, then changes in X should precede changes in Y.
- Therefore, in regression of Y on other variables (including its own past values) if we include past or lagged values of X and is significantly improves the prediction of Y, then we can say that X causes Y.

Specification

- $GDP_t = \sum_{i=1}^{n} \alpha_i M_{t-1} + \sum_{j=1}^{n} \beta_j GDP_{t-j} + u_{1j}$

- $M_t = \sum_{i=1}^{n} \lambda_i M_{t-i} + \sum_{j=1}^{n} \delta_j GDP_{t-j} + u_{2t}$

Steps to implement Granger test

- Regress current GDP on all lagged GDP terms and other Variables if any but do not include M variables in this regression- get RSS
- Run regression including the lagged M terms – get RSS
- Ho – lagged M terms do not belong in this regression

Steps – Granger test

- Apply F test

$$F = \frac{\left(\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{p2 - p1}\right)}{\left(\frac{\mathrm{RSS}_2}{n - p2}\right)},$$

  - RSS – Residual Sum of Squares
  - P2 = Lag X + Lag Y +1
  - P1 = Lag X + 1
  - N: Number of observations

- If computed value exceeds the critical value we reject Ho in which case lagged M terms belong in the regression – saying that M causes GDP
- Repeat steps for GDP causes M

Issues to be noted – Granger test

- It is assumed that the two variables are stationary – take first difference form if not
- Number of lagged terms – AIC or Schwarz criteria
- Assumption of error terms uncorrelated- transformation
- Present only F test results not the coefficients
- Lag length is critical
- Exogeneity – weak, strong, super

Single Time Series Analysis

- Variable at time t X(t) is impacted by X(t-1), X(t-2) etc.
- In budgeting Next year's budget -> dependent on this year expenditure and budget
- Production in this year is related to production of last year
- Macroeconomic variables move together

AR Process – Auto Regressive

- Theory of observation depending on lagged variables
- Errors in a observation is dependent on lagged errors
- Variable at time t X(t) is impacted by X(t-1), X(t-2) etc.
- In budgeting Next year's budget -> dependent on this year expenditure and budget
- Production in this year is related to production of last year

AR(1) Process

$$X_t = c + \varphi X_{t-1} + \varepsilon_t$$

- Errors are normal with zero mean and Sigma-square variance
- Process is stationary if
- Process is non-stationary otherwise
- Error is also auto-regressive

$$\mathrm{E}(X_t) = \mathrm{E}(c) + \varphi \mathrm{E}(X_{t-1}) + \mathrm{E}(\varepsilon_t) \Rightarrow \mu = c + \varphi\mu + 0.$$

AR(1) Intuition

- E.g.: A drunkard with his dog tied with a knot – random walk individually but predictive relative path
- Imagine infinite number of X(t) paths
- Weakly stationary: Mean and Variance are constant and does not depend on time t
- Strictly stationary: If none of the distributions depends on time

The unit root and co-integration

Key concepts

- Stochastic process/Random process/white noise process, unit root – synonyms
- Non stationary stochastic process – random walk model
- Asset prices, exchange rate, interest rates – non stationary
- Random walk with and without drift (constant)
- Trend stationary and differenced stationary

Unit root

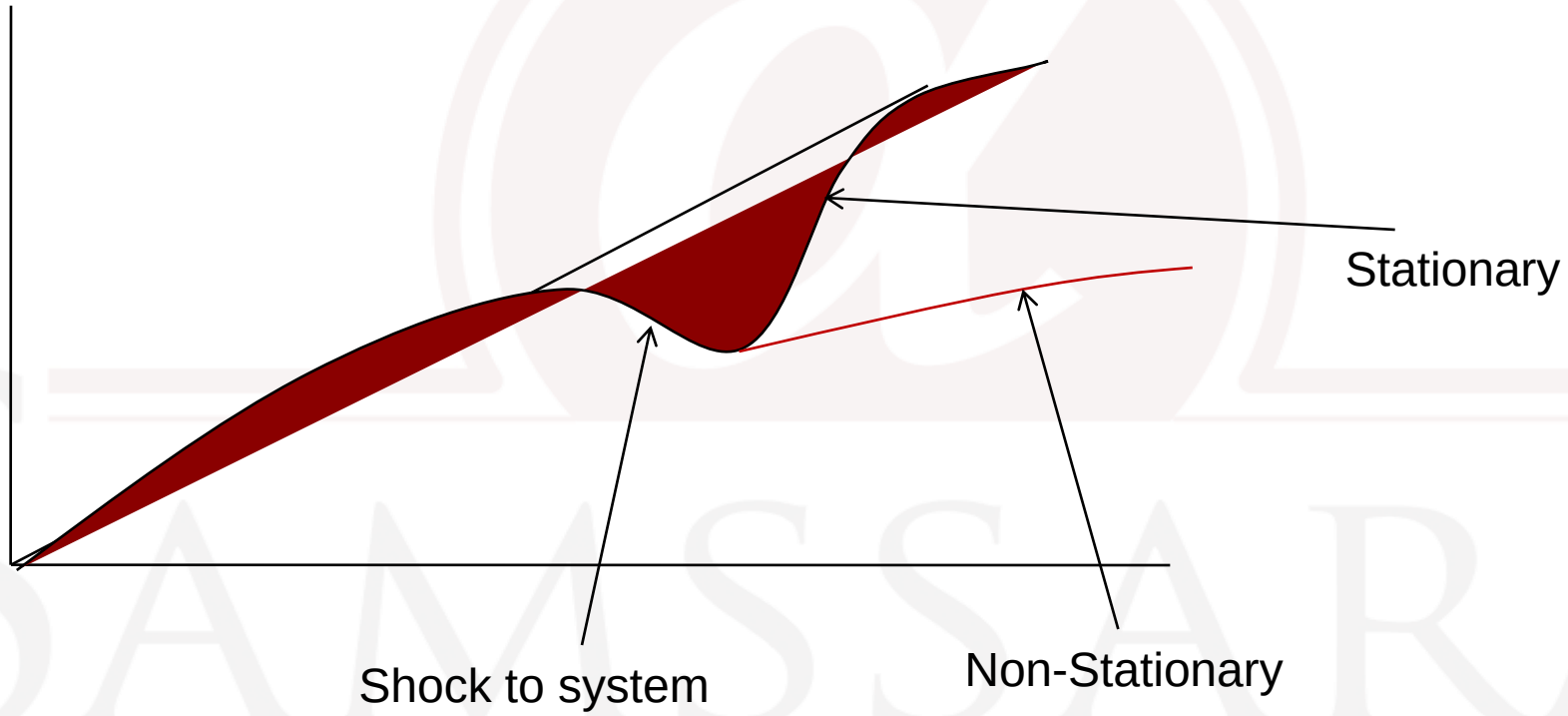
The formula

- Y(t) = Alpha*Y(t-1)+Error
- If Alpha = 1, unit root
- Infinite variance (Variance proportional to t)

Properties:

- Shocks to the system are permanent
- Variance increases with time and tends to Infinity
- Process is not stationary
- But can be differentiated to make it stationary

Unit root – Permanent shocks



Unit Root Tests

- Tests whether a time series variable is non-stationary using an autoregressive model with DF & ADF tests
- Dickey–Fuller test tests whether a unit root is present in an autoregressive model

$$y_t = \rho y_{t-1} + u_t \qquad \nabla y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$$

  - PRow = 1 (Random walk), PRow < 1 (Stationary)
  - If stationary: Tendency to return to constant mean
  - Large values followed by smaller values and vice versa
  - Hence, y(t-1) will significantly predict the next change
  - Augmented Dickey–Fuller test is performed for a larger and more complicated set of time series models
  - *Example: DF*τ statistic −4.57 < −3.50; at 95 per cent level the null hypothesis of a unit root will be rejected

Random walk – stationary and non-stationary series

- Test for unit root $\nabla y_t = \delta y_{t-1} + u_t$
- Random Walk with Drift $\nabla y_t = a_0 + \delta y_{t-1} + u_t$
- Drift and deterministic Trend $\nabla y_t = a_0 + a_1 t + \delta y_{t-1} + u_t$
- Non-stationary data can be transformed into stationary data using 1 difference method

Introduction to co-integration

- In Real life – stationary time series rarely exists
- Most time series are random
- Hence stationary time series needs to be constructed out of randomness
- E.g.: A drunkard with his dog tied with a knot – random walk individually but predictive relative path

Co-integration

- Co-integration: Stationary mean and variance
- Time series is stationary when
  - The mean is constant
  - The variance is constant
- Test for co-integration
  - If |r| < 1, the series is stationary
  - If |r| = 1, it is non-stationary (Random walk)

$$y_t = ry_{t-1} + e_t$$

- Most popular test: ADF (Augmented Dickey Fuller)
- If ADF < -3.2 (95% probability of co-integrated series)

Co-integration

- Two or more time series are co integrated if to a limited degree they share a certain type of behavior in terms of their long-term fluctuations
- A class of the linear combination of the unit root processes known as co integrated process
- Example: Short & long term interest rate, though meandering individually, track each other quite closely
- Tests:
  - Unit root analysis using DF and ADF tests
  - Residual-based Test for Co integration
  - Error Correction Model

Residual-based Test for Co integration

$$Y_t = \beta x_t + u_t, t = 1, ...T$$

$x_t = (x_{1t}, x_{2t}, ...x_{kt})'$ is the k-dimensional I(1) regressors

$$y_t \& x_t \, to \, be \, co - \text{int} \, egrated, u_t \, must \, be \, I(0)$$

Test is done in 2 steps,

Step 1: Run the OLS regression of (1) and obtain the residuals by

$$\hat{u}_t = y_t - \hat{\beta} x_t, t = 1....T$$

Step 2: Apply unit root test to the U(t)

$$\hat{u}_t = \phi \hat{u}_{t-1} + \varepsilon_t$$

That is, do the DF t-test of $H_0 : \phi = 1 against \, H_1 : \phi < 1$

Error Correction Model

- This time series modeling should describe both short-run dynamics and the long-run equilibrium simultaneously
- Define the error term by $\xi_t = y_t - \beta x_t$,
- β = co integrating coefficient, long term parameter

$$\alpha \text{ and } \gamma \text{ are called short} - \text{run parameters}$$

- ECM is simply defined as $\Delta y_t = \alpha \xi_{t-1} + \gamma \Delta x_{t-1} + u_t$

Time series modeling

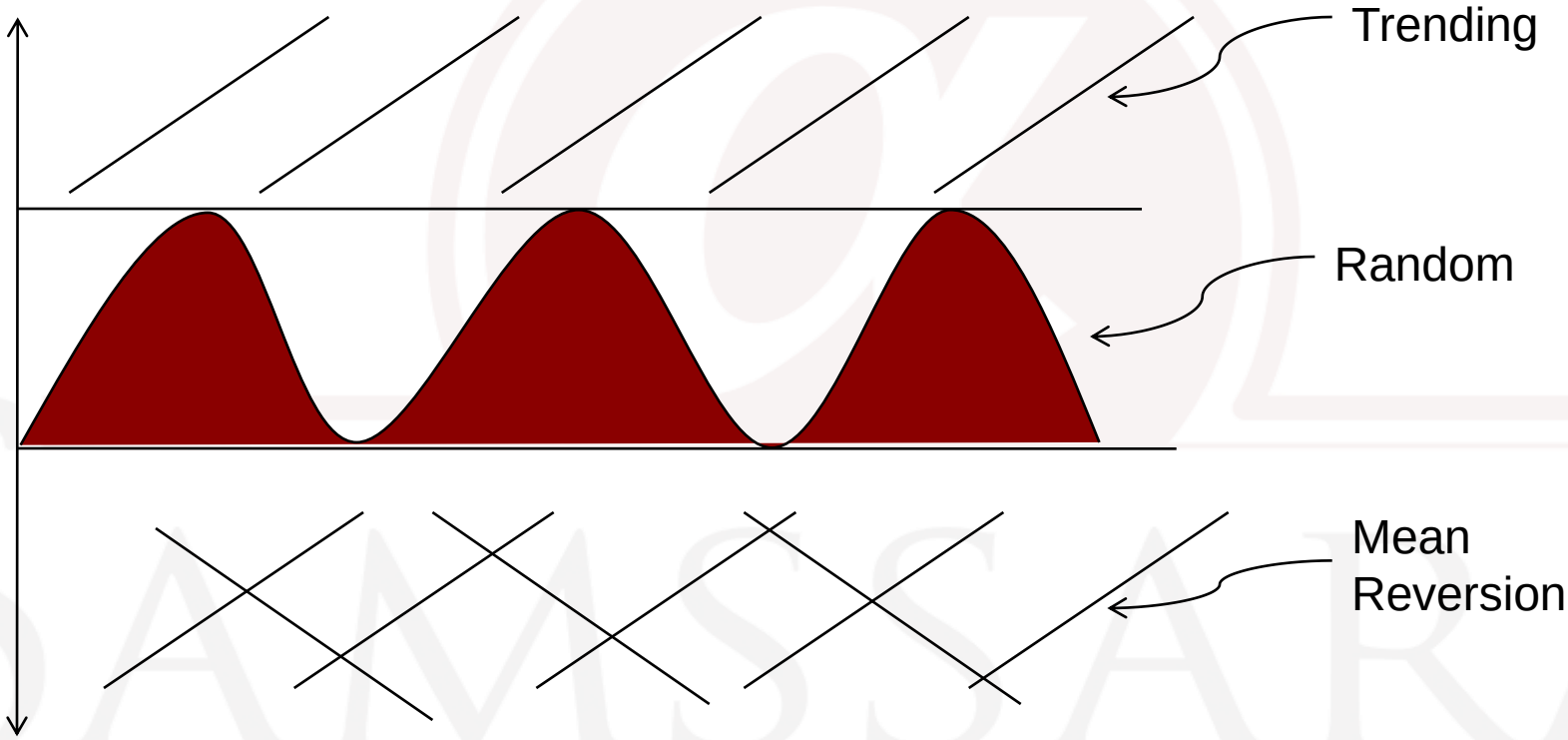
- Behavior of the time-series of data
  - Mean reverting, Trending or Random Walk
  - 50-60% time series is random walk
  - Focus should be on the other 40%
- Key elements: Mean and Variance
- Different behaviors
  - Mean reverting (E.g.: Pairs Trading)
  - Non-mean reverting (E.g.: Trend)
  - Constant variance (E.g.: Pairs Trading)
  - Increasing variance (E.g.: Trend)

The Mean and Variance in Time Series

The Random Walk Hypothesis

- Stock market price – Are they really random?
- Take 100 stocks and their daily prices
- Are all 100 stock prices random?
- Random => No Bull, No Bear and No Range (Is that true?)

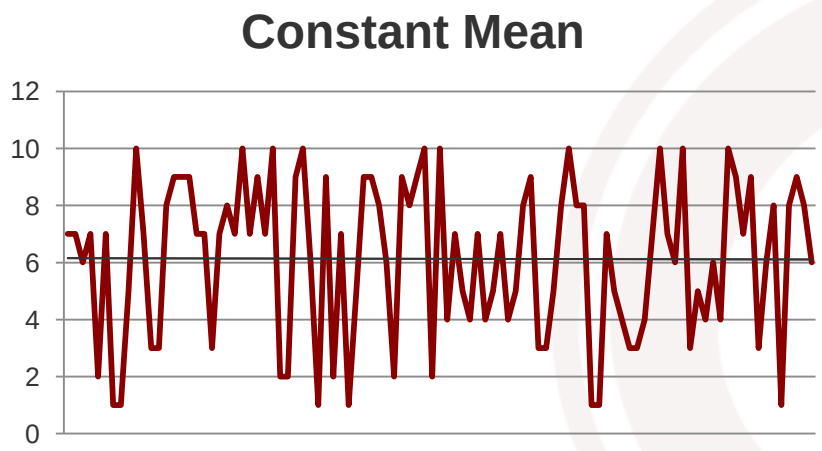
Extracting Non-Randomness

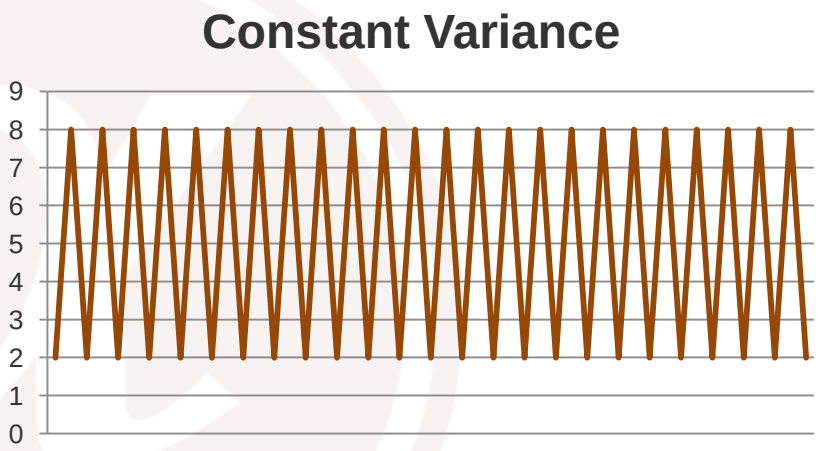


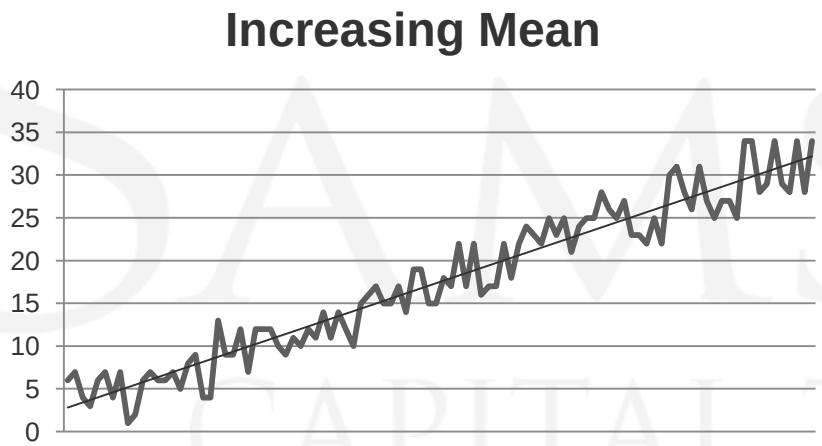
The Mean and Variance

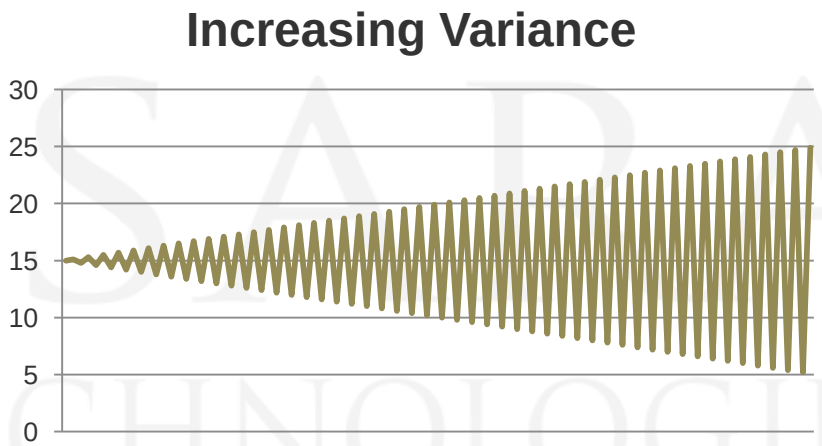
- The Mean
  - Constant Mean
  - Moving Mean
  - Mean across different periods
- The Variance
  - Increasing, Decreasing or remaining constant
  - Variance over different time periods
  - 1 Period, 2 Period…. N Period variance

Mean and Variance









Variance Ratio Test

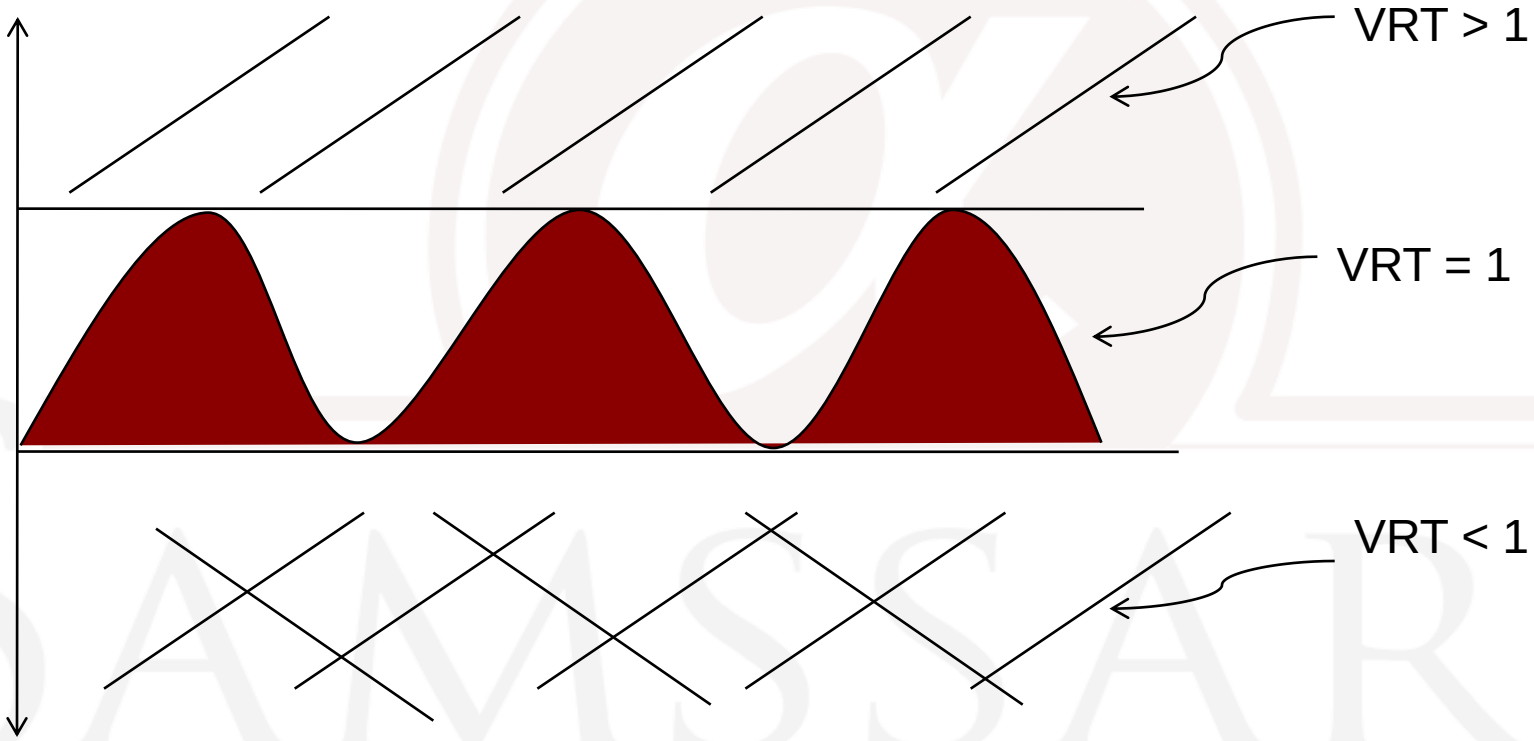
- Variance Ratio Test: Test for variance alone
- Useful when mean is varying w.r.t to the time

$$VR(k) = \frac{Variance(r_{k\Delta t})}{k \times Variance(r_{\Delta t})}$$

Generic time series modeling

- Variance
  - 1 Period vs. 5 Period
  - 1 Period vs. 10 Period
  - 1 Period vs. N Period
- To identify the variance profiles in specific periods

Extracting Non-Randomness



Ornstein-Uhlenbeck Process

- Ornstein-Uhlenbeck Process: Test for mean reversion alone
- Case1: x(t) < Mean, x(t) > Mean
- Assume: Theta, Sigma > 0

$$dx_t = \theta(\mu - x_t)dt + \sigma dW_t$$

Mean Reversion Rate

Mean

Volatility

Brownian Motion

Application of OU Process

- Modeling of
  - interest rate futures
  - Currency exchange rates
  - Commodity prices

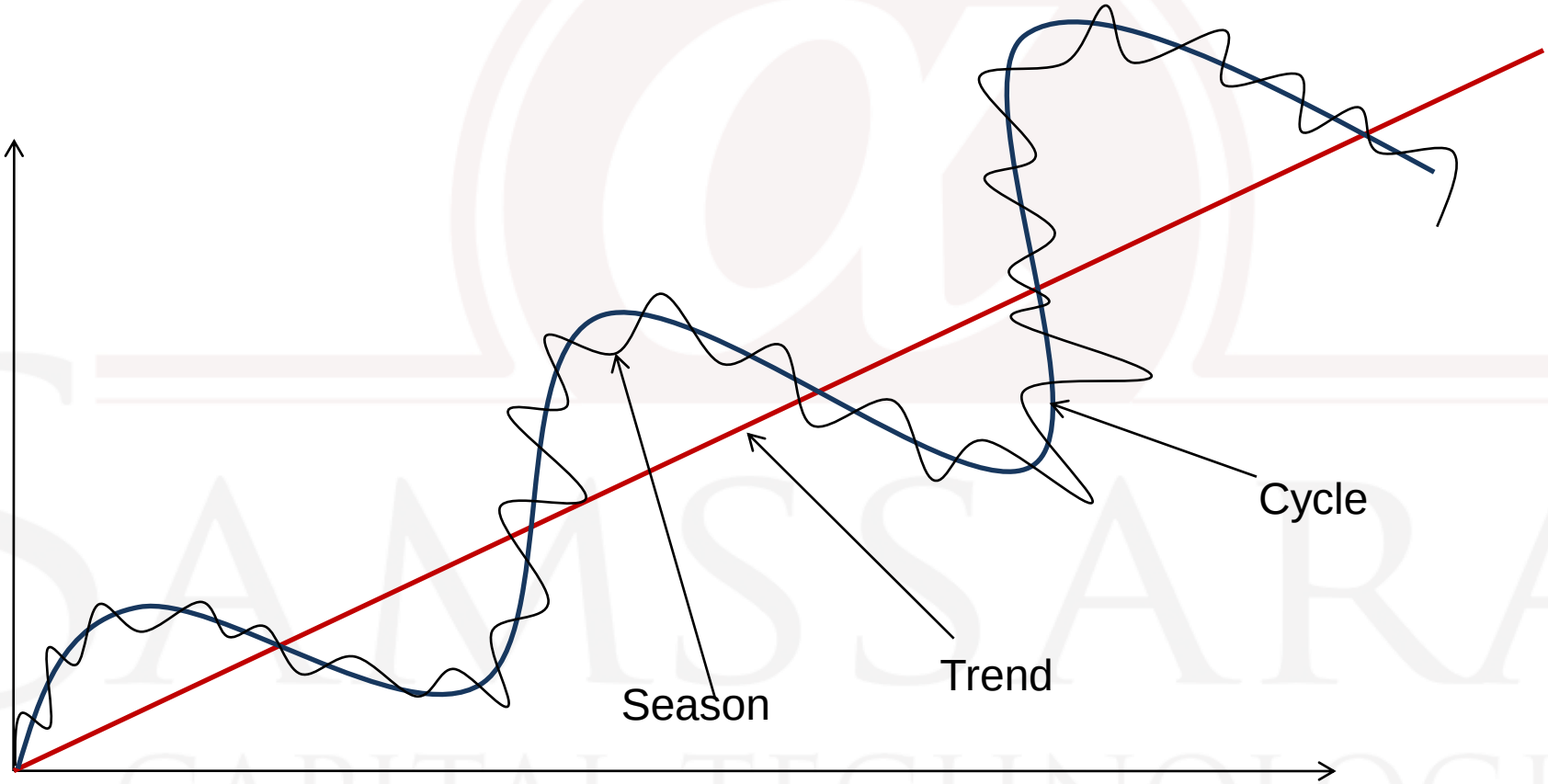
The forecasting techniques

Forecasting of time series

- Time series forecasting is important in
  - Sales forecast
  - Stock price movement
  - GDP and economic data forecasting

The Trend, Seasonal, Cyclical

- Y = TCS + e



The Trend, Seasonal, Cyclical

Year	Qtr	Sales Forecast
2009	1	10.2
	2	12.4
	3	14.8
	4	15.0
2010	1	11.2
	2	14.3
	3	16.4
	4	18.0

How do we predict the Sales Forecast for Q1 of 2011?

The Trend, Seasonal, Cyclical

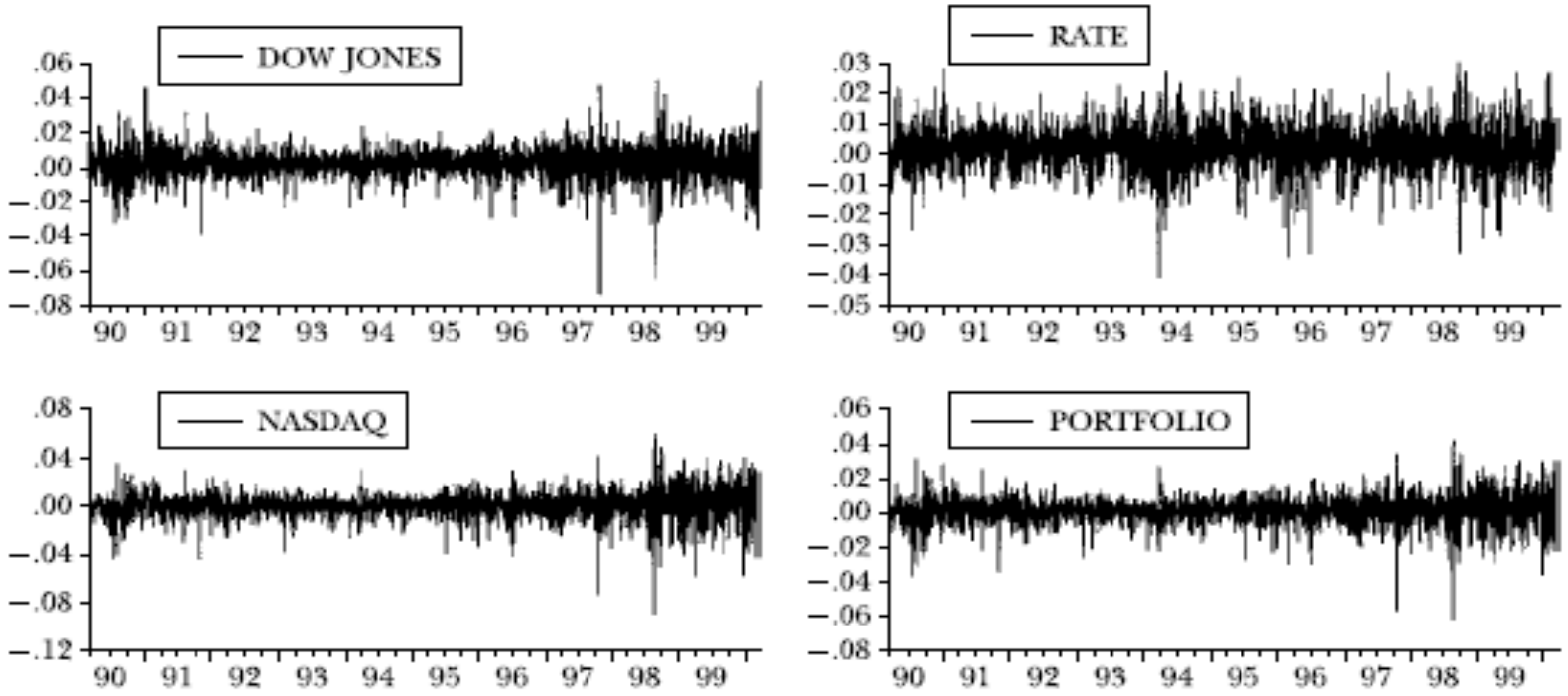
- S = Y/T
- Forecast for a trend and season

  => Y(Forecasted) = Trend * Season + e

ARCH-GARCH

- Modeling of heteroskedasticity
- Variance of error terms are modeled
- Volatility: High volatility followed by Low Volatility period and vice versa – condition of heteroskedasticity
- Y = Return of asset and variance in error => variance in return
- Application:
  - Risk analysis
  - Portfolio selection
  - Derivatives pricing (Implied vs. forecasted vol)
- Calculation of VaR (Value at Risk)

The volatility clustering



Example of volatility clustering on daily returns of a portfolio

Portfolio = 0.5*Nasdaq + 0.3*Dow + 0.2*Bonds

Standard Approach to Estimating Volatility

- Define σ_n as the volatility per day between day n-1 and day n, as estimated at end of day n-1
- Define S_i as the value of market variable at end of day i
- Define $u_i = \ln(S_i/S_{i-1})$

$$\sigma_n^2 = \frac{1}{m-1}\sum_{i=1}^{m}(u_{n-i} - \bar{u})^2$$

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m}u_{n-i}$$

Weighting Scheme

- Instead of assigning equal weights to the observations we can set

$$\sigma_n^2 = \sum_{i=1}^{m} \alpha_i u_{n-i}^2$$

where

$$\sum_{i=1}^{m} \alpha_i = 1$$

ARCH(m) Model

- In an ARCH(m) model we also assign some weight to the long-run variance rate, VL:

$$\sigma_n^2 = \gamma V_L + \sum_{i=1}^{m} \alpha_i u_{n-i}^2$$

where

$$\gamma + \sum_{i=1}^{m} \alpha_i = 1$$

EWMA Model

- In an exponentially weighted moving average model, the weights assigned to the u2 decline exponentially as we move back through time
- This leads to

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)u_{n-1}^2$$

Attractions of EWMA

- Relatively little data needs to be stored
- We need only remember the current estimate of the variance rate and the most recent observation on the market variable
- Tracks volatility changes
- RiskMetrics uses l = 0.94 for daily volatility forecasting

GARCH

- Forecast Variance (t+1) =

  A * Forecasted variance (t)

  + B * Daily return at t

  + C * Long run Variance

GARCH (1,1)

- In GARCH (1,1) we assign some weight to the long-run average variance rate

$$\sigma_n^2 = \gamma V_L + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2$$

- Since weights must sum to 1

$$\gamma + \alpha + \beta = 1$$

GARCH (1,1) (Contd…)

Setting $\omega = \gamma V$ the GARCH (1,1) model is

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2$$

and

$$V_L = \frac{\omega}{1-\alpha-\beta}$$

Illustration – GARCH (1,1)

- Suppose

$$\sigma_n^2 = 0.000002 + 0.13u_{n-1}^2 + 0.86\sigma_{n-1}^2$$

- The long-run variance rate is 0.0002 so that the long-run volatility per day is 1.4%

Illustration (Contd..)

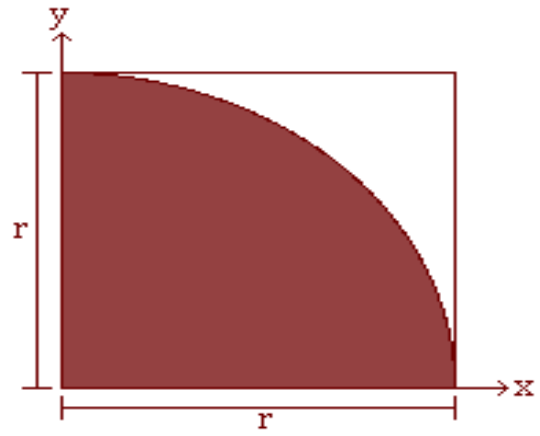
- Suppose that the current estimate of the volatility is 1.6% per day and the most recent percentage change in the market variable is 1%.
- The new variance rate is

$$0.000002 + 0.13 \times 0.0001 + 0.86 \times 0.000256 = 0.00023336$$

The new volatility is 1.53% per day

Monte Carlo Methods

Estimation of Pi using Monte Carlo



Throwing darts randomly at the above figure:

$$\frac{\#\text{ darts hitting shaded area}}{\#\text{ darts hitting inside square}} = \frac{\text{area of shaded area}}{\text{area of square}}$$

Estimation of Pi using Monte Carlo

Hence to estimate Pi we use:

$$\frac{\#\text{ darts hitting shaded area}}{\#\text{ darts hitting inside square}} = \frac{\frac{1}{4}\pi r^2}{r^2} = \frac{1}{4}\pi$$

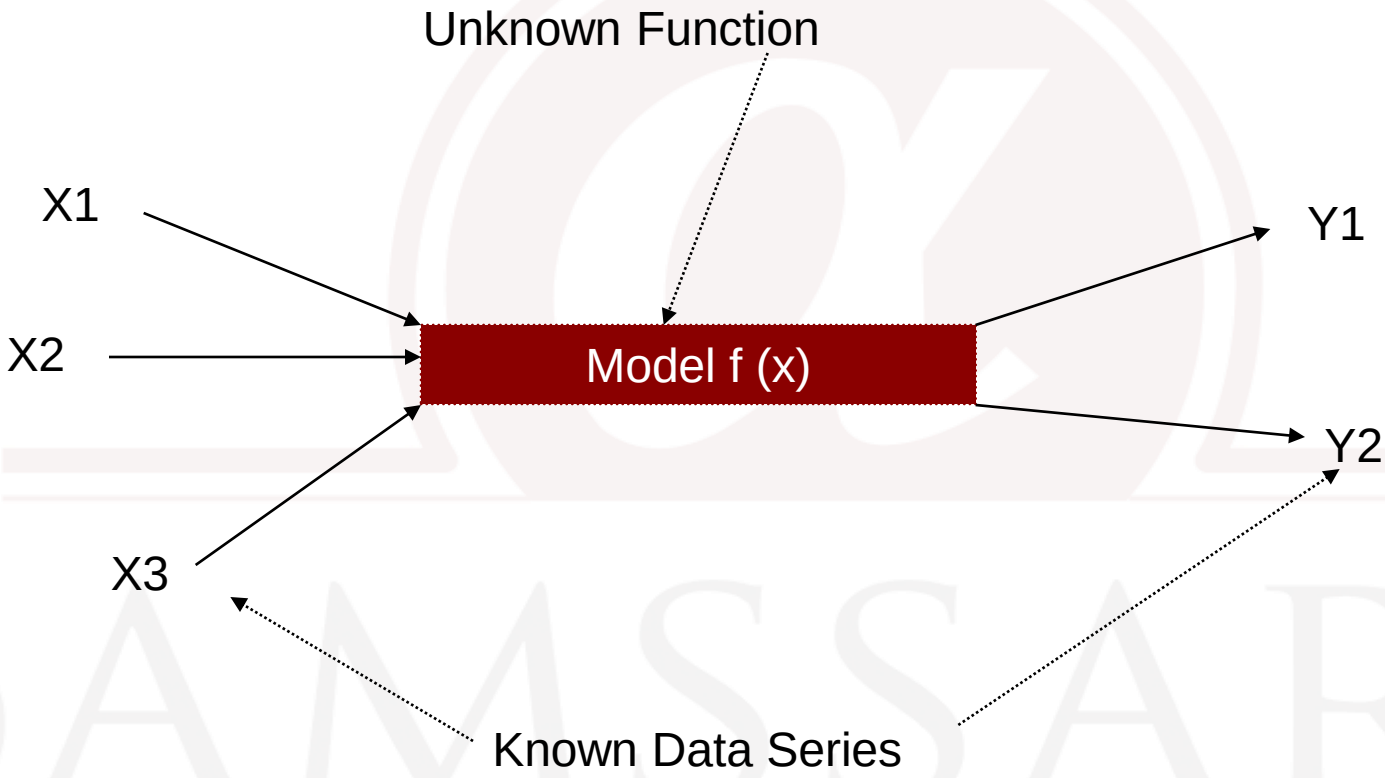
or

$$\pi = 4\,\frac{\#\text{ darts hitting shaded area}}{\#\text{ darts hitting inside square}}$$

Monte Carlo methods

- Monte Carlo methods can help us solve problems that are too complicated to solve using equations, or problems for which no equations exist
- They are useful for problems which have lots of uncertainty
- They can also be used as an alternate way to solve problems that have equation solutions
- However, they have drawbacks: Monte Carlo methods are less accurate for lower dimension problems
- Studying systems with large degree of freedom
- When exact equation / result is not known
- Means of last resort in modeling
- After Monte Carlo: AI and Neural Networks

Monte Carlo Techniques



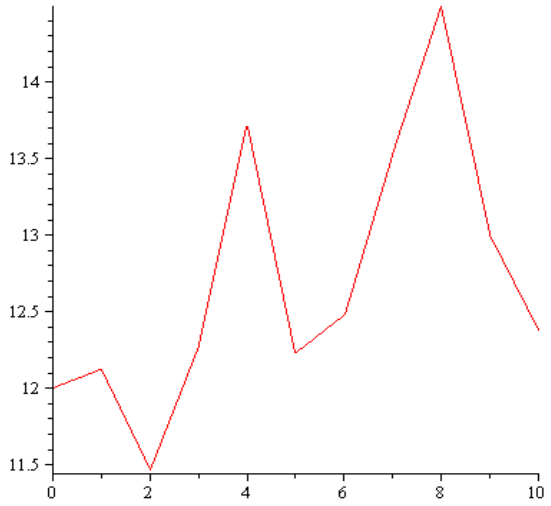
E.g.: Factor Modeling using Value and Growth

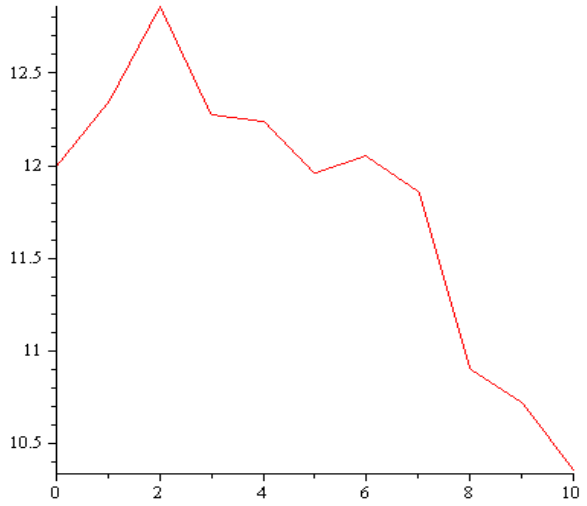
Monte Carlo Techniques

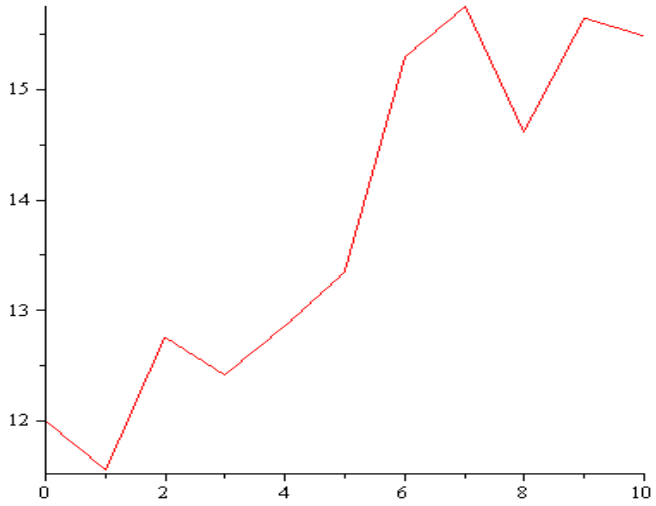
- Applications :
  - Corporate Finance (No longer deterministic)
    - Y = NPV
    - X = Cash Flow, Interest rates, Competition
    - Result compiled into Normal distribution of NPV
  - Options Pricing
  - Back testing of Models
  - Weather Forecasting
  - Estimation of number Pi

A Monte Carlo Method for Financial Derivatives

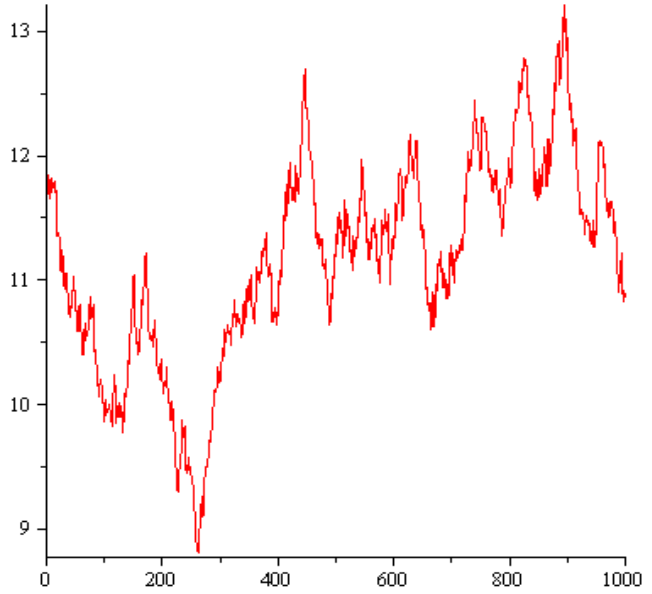
- We don’t know what a stock will do in the future over time T.
- Blackscholes assumes a Brownian motion to figure out what is most likely to happen to the price of a stock after time T.
- if we figure out a way to approximate a possible brownian walk of a stock over some period of time T, we can approximate the blackscholes value by:
  - Doing many sample walks of the stock
  - Computing the value of a derivative for each sample walk
  - Averaging the trial derivative values together to come up with an expected value for the derivative







If we do many random sample walks of a stock over time T, we can use that data to gain information about what the stock is most likely to do over time T.



Applying the Monte Carlo method to financial derivative

- We know that $ds = S*\mu*dt + S*\sigma*dx$
- Where:
  - ds = change in stock price (St – St-1)
  - μ = “drift” – Trend drift, polynomial drift etc – Deterministic
  - dt = change in time
  - σ = the volatility of the stock
  - dx = a random normally distributed variable with mean 0 and standard deviation SQRT(dt)

Applying the Monte Carlo method to financial derivative

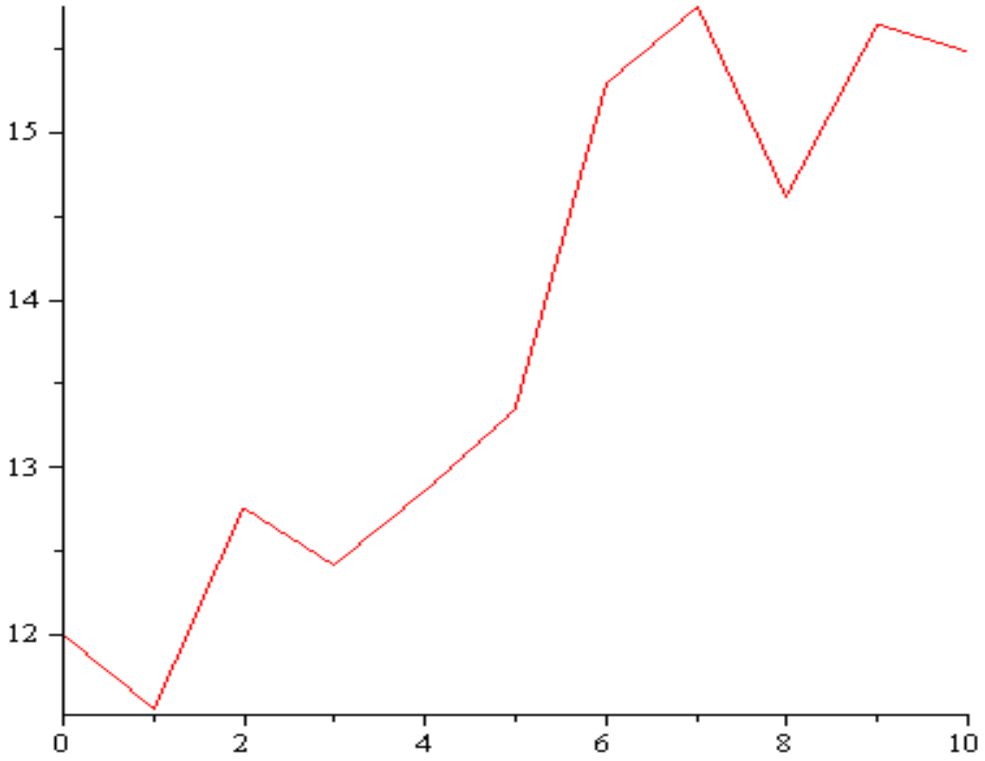
- Since ds = change of stock price from time a to b, we can replace ds with (S_b-S_a)
- We can then take this formula and solve for S_b, the value of the stock at some time in the future, as a function of a bunch of stuff that we know:

  $(S_b-S_a)/S_a = \mu*dt + \sigma *dx$ so,
  $(S_b-S_a) = S_a* \mu*dt + S_a* \sigma *dx$
  $S_b=S_a+ S_a* \mu*dt + S_a* \sigma *dx$

- We can use this formula:

  $S_2=S_1+ S_1* \mu*dt + S_1* \sigma *dx$

- To come up with subsequent values of S in a sample random walk of a stock.



- A sample run of a stock can be generated by using the formula (S2=S1+ S1* μ*dt + S1* σ *dx) to generate subsequent points in the path of a stock.
- We can take the values of a financial derivative (like a call, or a put) over many such sample paths and average them together to get an expected value for the financial derivative.

$S1=S0+ S0* \mu*dt + S0* \sigma *dx$

$S2=S1+ S1* \mu*dt + S1* \sigma *dx$…

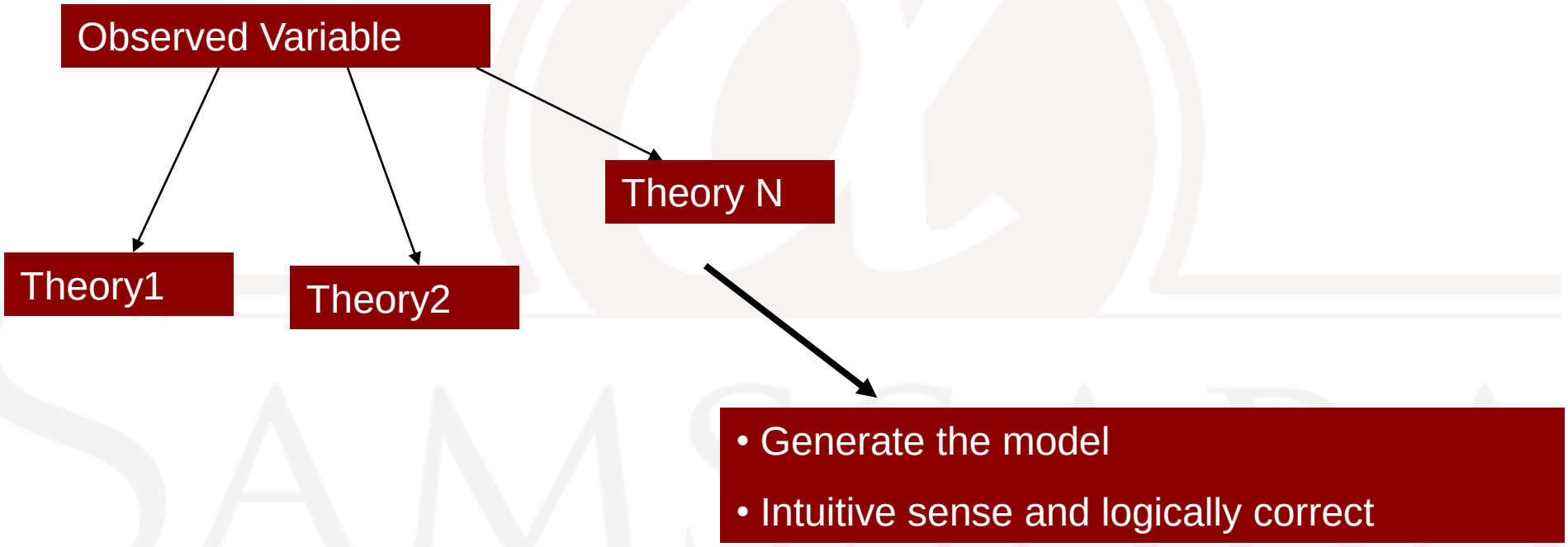
$S_N=S_{N-1}+S_{N-1}* \mu*dt + S_{N-1}* \sigma *dx$

Sensitivity Analysis

- Identifying how sensitive a particular factor is
- Change is delta X => How much change in delta Y

Application of models

Selection of the model



The diagnostic

- Generate the Model
- Check for accuracy
- Diagnose the problems / errors
- Eliminate terms and re-model
- Apply on difference data set
- Check for consistency and variance

The diagnostic

Standard techniques for modeling are well-known

From application view-point use standard techniques

Stochastic: Economic variables, AR process, GARCH

Non stochastic: Social choice theory (Price Inc => Demand Dec)

Fortunately: Most popular models are already known, hence application is more important

Macroeconomic data

- Most macroeconomic data involves using
  - OLS
  - Granger Causality
  - Dependence of one variable on another
  - E.g.: Onion Prices dependent on Inflation, M2, income of people
- The GDP growth of a country is highly mean reverting at times
- Most growth and YoY change data are mean-reverting

Commodity Prices Modeling

- Commodities prices are often correlated
- E.g.: Gold, Crude Oil, pork Bellies, Orange juice
- Trend following systems in commodities are the most common
- Hence, increasing variance w.r.t to the time is trend following system

Stock Price/Currency Modeling

- Black Scholes path modeling using Monte Carlo
- Mean reversion modeling using co integration and variance ratio test
- Volatility modeling using ARCH and GARCH

Conclusion

- Intuitive / Economic sense to develop models is the key
- Clean and precise data is extremely useful in generating effective models
- Choosing the right mathematical model – using simple regression techniques is very useful in macroeconomic analysis
- Understanding the mean reversion, trending and random walk behavior can help build profitable models
- Keeping the models “simple” and understanding “intuitively” can help save from the “black-box-syndrome”
- Avoid data fitting and relying heavily on R-Square and T-Stats which does not make “economic” sense

Recommended referrals

Prop trading

- High-Frequency Trading: A Guide to Algorithmic Strategies and Trading Systems by Irene Aldridge
- Statistical Arbitrage: Algorithmic Trading Insights and Techniques by Andrew Pole
- The Encyclopedia of Trading Strategies by Jeffrey Owen and Donna McCormick

Agency trading

- Algorithmic Trading and DMA: An introduction to direct access trading strategies by Barry Johnson
- Quantitative Trading: How to Build Your Own Algorithmic Trading Business by Ernset P. Chan

Web forums

- Wilmott forum: www.wilmott.com
- Nuclear Phynance: www.nuclearphynance.com

About Samssara Capital Technologies LLP

COMPANY BACKGROUND

- Samssara Capital Technologies LLP ("Samssara") is an investment solutions firm focused solely on developing automated algorithmic and quantitative trading and investment strategies
- It was launched in 2010 by a team of IIM Ahmedabad and IIT Bombay graduates - Rajesh Baheti, Manish Jalan and Kashyap Bhargava
- Samssara caters to its clients' needs of providing an alternative asset management vehicle, with the focus on 100% automated and quantitative trading strategies
- The team at Samssara works on mathematical models and statistics that identify repetitive patterns in equity, commodity and currency markets
- The addressable market for Samssara is global - as the firm can develop and build models which can function in both developing markets with limited competition and developed markets with strong competition
- Samssara's client base includes the leading international and domestic banks, international and domestic stock brokers, family offices, corporate treasuries and HNIs

PRODUCTS OFFERED

- Samssara's products vary from pair trading (statistical arbitrage), factor models, Nifty Index beating products to very high frequency trading strategies
- samCAP, a key product offered by Samssara, is a factor model, where the model identifies a basket of stocks in Nifty that tend to outperform the index and takes a long position in these stocks. Alongside, the product also hedges the investor's portfolio using Nifty futures – whenever the market turns bearish
- Other products offered include samTREND - a trend following strategy in equities, commodities & currencies and samWILLS – a long-short strategy based on statistical arbitrage
- Samssara also develops in-house products which are used by investors like HNI's, corporate treasuries, Prop houses of brokers and investors who wants an alternative vehicle for investment apart from equities and fixed income.
- The products are designed to generate consistent returns and ride the volatility of the markets with systematic approach
- Additionally, Samssara works on providing high end services and strategy development consultancy to hedge funds and International Banks globally

Contact us

Manish Jalan

M: +91 98678 32726
D: +91 22 6748 7720
E: manish@samssara.com

Tarun Soni

M: +91 98692 17190
D: +91 22 6748 7720
E: tarun@samssara.com

Head Office:

208/209, Veena Chambers
21 Dalal Street
Mumbai – 400 001

Development Office:

207, Business Classic,
Behind H P Petrol Pump,
Chincholi Bunder Road, Malad (W)
Mumbai – 400 064

For more information do visit : www.samssara.com