1. INTRODUCTION

Multiple linear regression is a method used to model the linear relationship between a dependent variable and ***more than one*** independent variables.

Multiple linear regression lets analysts estimate using more complex models with multiple explanatory variables.

Multiple linear regression can be used to predict future returns, improve portfolio construction, and understand security drivers. However, it can produce spurious results and poor predictions if used incorrectly.

How the model works

- The analyst specifies the dependent and independent variables and uses software to estimate the model and generate statistics.
- The software does the estimation and the analyst's task is to interpret the output.

2. USES OF MULTIPLE LINEAR REGRESSION

Financial and economic relations are complex and require models with multiple explanatory variables that must pass rigorous statistical and theoretical scrutiny.

Multiple regression can be used to identify relationships between variables to test current or future theories.

For example, examining multiple factors such as political stability, economic conditions, and ESG considerations on stock returns.

The Regression Process

The objective is to use the variation of two or more independent variables to explain the variation in the dependent variable.

Find out whether the dependent ***variable is continuous***? If the answer is:

- No, use logistic regression
- Yes, use traditional regression model

⬇

Estimate the regression model. Examine the model to see ***if it meets the key assumptions***? If the answer is:

- No, adjust the model
- Yes, examine the model's goodness-of-fit

⬇

Determine whether the ***overall fit is significant***? If the answer is:

- No, adjust the model.
- Yes, determine if this model is the best of all possible models?

⬇

Check ***if the model is the best of all models***? If the answer is:

- No, adjust the model
- Yes, use the model for analysis and prediction.

Refer to: Exhibit 2: The Regression Process' from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

3. THE BASICS OF MULTIPLE REGRESSION

Regression analysis is a tool used to examine whether a variable is useful to explain another variable.

Multiple regression model predicts the value of a dependent variable based on the value of two or more independent variable.

Variation represents the difference between observation and its expected value i.e., how much estimatedI ith value differs from the actual ith value.

A multiple linear regression model has the following general form:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i, i = 1, 2, \ldots n$$

where,

Y_i = i^{th} *observation of dependent variable Y*
X_{ki} = i^{th} *observation of* k^{th} *independent variable X*
β_0 = *intercept term*
β_k = *slope coefficient of* k^{th} *independent variable*
ε_i = *error term of ith observation*
n = number of observations
k = total number of independent variables

- ***A slope coefficient,*** β_j ***is known as partial regression coefficients or partial slope coefficients.*** It measures how much the dependent variable, Y, changes when the independent variable, X_j, changes by one unit, ***holding all other independent variables constant.***

- ***The intercept term*** (β_0) is the value of the dependent variable when the independent variables are all equal to zero.
- A regression equation has k slope coefficients and one intercept i.e., k + 1 regression coefficients.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

4. ASSUMPTIONS UNDERLYING MULTIPLE LINEAR REGRESSION

The five key assumptions of the multiple linear regression model are:

1) Linearity
2) Homoskedasticity
3) Independence of errors
4) Normality
5) Independence of independent variables

1 Assumption 1: Linearity

'Relation between the dependent variable Y and the independent variables (X_1, X_2, . . . ,X_k *)is linear.'*

2 Assumption 2: Homoskedasticity

'The variance of residuals is the same for all observations. It is known as Homoskedasticity (same scatter) assumption.'

3 Assumption 3: Independence of errors

'The observations (pairs of Xs and Ys) are independent of each other, which implies the residuals are uncorrelated across observations. (i.e. ***no serial correlation***).'

4 Assumption 4: Normality

'The regression residuals (error term) must be normally distributed.'

5 Assumption 5: Independence of independent variables

a) Independent variables (X_1, X_2, . . . ,X_k) are not random.
b) No exact linear relation exists between two or more of the independent variables.

Note:

- When an exact linear relationship exists between two or more independent variables, linear regression ***cannot*** be estimated.
- Furthermore, when two or more independent variables are highly correlated, the model can be estimated but its interpretation is ***problematic***.

Normal Q-Q Plot

A Q-Q plot is used to compare the distribution of a variable to a normal distribution.

A Q-Q plot is used in regression to compare the model's standardized residuals to a theoretical standard normal distribution. The residuals should align along the diagonal if they are normally distributed.

Refer to: Exhibit 8: Normal Q-Q Plot of Regression Residuals 'from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' 'and end-of-chapter questions from the CFA Institute's Curriculum and FinQuiz Question-bank.

1. GOODNESS OF FIT

Measures of Goodness of Fit

Goodness of fit (i.e., how well the regression model fits the data) can be measured using **coefficient of determination R^2**.

- The coefficient of determination is the percentage of the total variation in the dependent variable that is explained by the independent variable(s).
- The coefficient of determination is also called R-squared and is denoted as R^2.
- It is descriptive measure.

$$Coefficient\ of\ determination\ (R^2) = \frac{Explained\ Variation (SSR)}{Total\ Variation\ (SST)}$$

$$= \frac{Sum\ of\ squares\ regression\ (SSR)}{Sum\ of\ squares\ total\ (SST)} =$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where, $0 \leq R^2 \leq 1$

In a single independent variable, the coefficient of determination is: $R^2 = r^2$

where,
R^2 = Coefficient of determination
r = Simple correlation coefficient

Example: Suppose coefficient of determination between returns of two assets is 0.64. This means that approximately 64 percent of the variability in the returns of one asset (or dependent variable) can be explained by the returns of the other asset (or indepepnent variable).

R^2 in Multiple Linear Regression

R-squared is not a suitable measure to test the "goodness of fit" in **multiple linear regression** as it always increases or stays the same with an increase in the number of independent variables. It is important to keep in mind that a high R^2does not imply causation.

Problems with Multiple Linear Regression

- R^2 does not provide information regarding the statistical significance of the coefficients.
- R^2 does not provide information on whether calculated coefficients and forecasts are biased.
- R^2 cannot determine if the model's fit is satisfactory. A model may have a high R^2 due to overfitting and biases.

Overfitting occurs when a regression model is too complex due to a high number of independent variables with respect to the sample size. This leads to coefficients on independent variables that do not show true correlations with the dependent variable.

Adjusted R^2

The ***adjusted R^2*** is used to deal with this artificial increase in accuracy. Adjusted R^2 does not automatically increase when another variable is added to a regression; it is adjusted for degrees of freedom. The *adjusted R^2* is given by

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

where,
n = sample size,
k = number of independent variables

- When $k \geq 1$, then R^2 is strictly > Adjusted R^2.
- Adjusted R^2 decreases if the new variable added does not have any significant explanatory power.
- Adjusted R^2 can be negative as well but R^2 is always positive.
- Adjusted R^2 is always $\leq R^2$.
- If the coefficient's t-statistic > |1.0|, $\bar{R}^2$ *increases*
- If the coefficient's t-statistic < |1.0|, $\bar{R}^2$ *decreases*

ANOVA

Analysis of Variance (ANOVA) is a statistical method used to divide the total variance in a study into meaningful pieces that correspond to different sources.

Analysis of Variance Table for Simple Linear Regression

ANOVA	df	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$

Stanley Notes™ 2025

Error	n-2	$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$	$MSE = \frac{SSE}{n-2}$	
Total	n–1	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

Or

Source of Variability	df	Sum of Squares	Mean Sum of Squares
Regression (Explained)	1	SSR	MSR = RSS/1
Error (Unexplained)	n-2	SSE	MSE = SSE/n-2
Total	n-1	SST=SSR + SSE	

Refer to: Exhibit 1 to exhibit 3 and ANOVA table for practicing example from the CFA Institute's

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

Important Things to Consider

- The adjusted R^2 in multiple regression doesn't clearly interpret the proportion of dependent variable explained by independent variable.
- The adjusted R^2 also doesn't indicate if the regression coefficients are significant or if predictions are biased.
- R^2 and adjusted R^2 alone are not suitable for verifying the model's fit. ANOVA and goodness-of-fit metrics are required for this.

AIC and BIC

Including independent variables may increase both R^2 and adjusted R^2, but it can lead to overfitting.

Many statistics can be used to compare model quality and determine which model is the most cost effective. Two of these statistics are AIC and BIC.

1. Akaike's information criterion (AIC)

- AIC is a metric for model parsimony. Lower AIC means a better fit.
- AIC is often included in the output of regression software.
- AIC can also be calculated as follows:

$$\text{AIC} = n \ln\left(\frac{SSE}{n}\right) + 2(k+1)$$

where,
n = sample size
k = no. of independent variables
SSE = sum of square error
Term 2(k + 1) represents the penalty for adding variables

2. Schwarz's Bayesian information criterion (BIC or SBC)

- The BIC model, like the AIC model, provides for the comparison of models with the same dependent variable. The formula is shown below:

$$\text{BIC} = n \ln\left(\frac{SSE}{n}\right) + \ln(n)(k+1)$$

- BIC evaluates a greater penalty for having more parameters in a model than AIC, therefore it will prefer small, more parsimonious models.

Important:

- Both AIC and BIC models allow comparison of models with the same dependent variable.
- AIC is preferred when model is used for prediction purposes.
- BIC is preferred when the best goodness of fit is required.
- It is the relative values of AIC or BIC among the collection of models that are crucial, not their absolute values.

Refer to: Exhibit 4 for practicing example from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

2. TESTING JOINT HYPOTHESES FOR COEFFICIENTS

Significance of two or more coefficients in Multiple Regression

- Interpreting intercept and slope coefficients in multiple regression is similar to linear regression, but with one difference. The intercept in multiple regression predicts the dependent variable if all independent variables are zero.
- The slope coefficient predicts the change in the dependent variable for a one-unit change in an independent variable, while holding other variables constant.
- Tests on a single coefficient in multiple regression are the same as those in simple regression.
- The hypothesis structure and t-test are also identical.
- The null and alternative hypotheses are the same for a two-sided alternative hypothesis that the true coefficient b_j is equal to a hypothesized value B_j.

Null and Alternative Hypotheses

H_0: $b_j = 0$ (no linear relationship)
H_a: $b_1 \neq 0$ (linear relationship does exist)

Unrestricted Models and Restricted (Nested) Models

Consider a general model that have five independent variables:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + b_5X_{5i} + \epsilon_i$$

A model with all independent variables is called ***'unrestricted model'***. This unrestricted model has all five independent variables.

Suppose we want to know if X_4 and X_5 significantly explain the dependent variable. This is done by checking if b_4 and b_5 are equal to zero

H_0: $b_4 = b_5 = 0$

To do so, we must contrast the unrestricted model with this ***'restricted model'*** by limiting the slope coefficients on X_4 and X_5.

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \epsilon_i$$

As the restricted model nests within the unrestricted model therefore this is also called '***nested model'***.

This models' comparison indicates a null hypothesis that restricts two coefficients jointly.

H_0: $b_4 = b_5 = 0$
H_a: $b_4 = b_5 \neq 0$

To compare models, we use a statistic comparing an unrestricted model to a restricted one where one or more slope coefficients are set to zero. This helps examine the impact of jointly omitted variables on the model's ability to explain the dependent variable. We use an F-distributed test statistic for this purpose.

F-Distributed Joint Test of Hypotheses

F-distributed test statistic is used to examine the role of the jointly omitted variables.

$$F = \frac{(Sum\ of\ squares\ error\ restricted\ model - Sum\ of\ squares\ error\ unrestricted)/q}{Sum\ of\ squares\ error\ unrestricted\ model/(n-k-1)}$$

where

q= no. of restrictions = no. of omitted variables

For example, as mentioned above:

Unrestricted model = $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + b_5X_{5i} + \epsilon_i$

Restricted model = $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \epsilon_i$

df numerator = $q = 2$
df denominator = $n - k - 1 = n - 2$

Decision Rule: Reject H_0 if F > F-critical.

Note: A question may arise, why not test individual variables and draw conclusions about the set based on the results?

Reason: Testing individual variables in multiple regression with financial variables may not provide accurate conclusions due to correlation between variables and shared explanatory power.

Steps for Hypothesis Testing

1. State the hypothesis
2. Identify the appropriate test statistic
3. Specify the level of significance
4. State the decision rule
5. Calculate the test statistic
6. Make a decision

Refer to: Exhibit 5, Panel A and Panel B for practicing example from the CFA Institute's Curriculum.

General Linear F-Test

The generic linear F-test can be used to test the null hypothesis that all slope coefficients are equal to zero.

F-Statistic or F-Test evaluates how well a set of independent variables, as a group, explains the variation in the dependent variable.

In multiple regression, the F-statistic is used to test whether at least one independent variable, in a set of independent variables, explains a significant portion of variation of the dependent variable.

The F statistic is calculated as the ratio of the average regression sum of squares to the average sum of the squared errors,

$$\frac{MSR}{MSE} = \frac{(\frac{RSS}{k})}{(\frac{SSE}{n-k-1})}$$

df numerator = k = 1
df denominator = n – k – 1 = n – 2

Decision Rule: Reject H_0 if F > F-critical.

Note:

- F-test is always a one-tailed test.
- In a regression with just one independent variable, the F statistic is simply the square of the t-statistic i.e. $F = t^2$. F-test is most useful for multiple independent variables while the t-test is used for one independent variable.
- When independent variable in a regression model does not explain any variation in the dependent variable, then the predicted value of y is equal to mean of y. Thus, RSS = 0 and F-statistic is 0.

Refer to: Exhibit 6, Panel A and Panel B for practicing example from the CFA Institute's Curriculum.

Model Fit Assessment Using Multiple Regression Statistics

Statistic	Assessment Criteria
Adjusted R^2	The higher the better
AIC	The lower the better
BIC	The lower the better
t-statistic on a slope coefficient	Outside bounds of critical t-value(s) for the selected significance level
F-test for joint test of slope coefficients	Exceeds the critical F-value for the selected significance level

Source: Exhibit 7, CFA Institute's Curriculum

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

3. FORECASTING USING MULTIPLE REGRESSION

Predicting the Dependent Variable in a Multiple Regression Model

The process of calculating the predicted value of dependent variable is the same as simple linear regression except that we need projected values for multiple independent variables.

Prediction equation:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}$$

where,

$\hat{Y}_i$: Estimated or predicted value of Y
b_0: Estimated intercept
b_1, b_2,… & b_k: Estimated slope coefficients

Note:

Using a multiple regression model for forecasting requires caution.

- The model must include all independent variables, even if they are not significant, due to correlations used to calculate slope coefficients.
- The intercept term must also be included for accurate forecasting.

Errors:

- **Model Error ε_i:** The uncertainty in the model is mostly due to the regression residual i.e.,

observations not falling on the predicted line. It is the stochastic component of the model.

- **Sampling Error:** Incorporating forecasts as independent variables can lead to sampling error, which happens when a sample does not accurately represent the population. Such an error results from inaccurate forecasting.

The combined effect of model error and sampling errors leads to a wider prediction interval.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

Practice: End-of-Chapter Questions from the CFA Institute's Curriculum and Questions from FinQuiz Question-bank.

1. MODEL SPECIFICATION ERRORS

Model specification refers to the set of variables that are used in the regression and the functional form of the regression equation.

Principles for Proper Regression Model Specification

	Principle	*Rationale*
Economic Reasoning	Model should be based on economic reasoning.	The model should economically justify variable choices.
Parsimonious	Model should be parsimonious.	Each variable in the regression should play an important role.
Out of Sample Performance	Model should perform well out-of-sample.	The model may only explain the particular dataset it was trained on, which leads to overfitting.
Appropriate functional form	The functional form of the model should be appropriate.	Models should include nonlinear terms if regressors are nonlinear.
Satisfying Regression Assumptions	The model should meet the regression assumptions.	If there is heteroskedasticity, serial correlation, or multicollinearity, the regression variables and/or functional form should be revised.

2. MISSPECIFIED FUNCTIONAL FORM

Failures in Regression Functional Form

Failures	*Explanation*	*Consequence*
Omitted variables	Omitting important variable(s) from the regression.	May lead to heteroskedasticity, serial correlation.
Inappropriate form of variables	Ignoring a nonlinear relation b/w the dependent & independent variable.	May lead to heteroskedasticity.
Inappropriate variable scaling	Variable(s) may need to be transformed before regression estimation.	May lead to heteroskedasticity, or multicollinearity.
Inappropriate data pooling	Regression model pools data from various samples that should not be pooled.	May lead to heteroskedasticity or serial correlation.

a. Omitted Variables

Omitted variables bias: One or more important variables are omitted from regression.

If the omitted variable is:

- ***uncorrelated*** with other independent variables, the error term will reflect the omitted variable's contribution, which can lead to biased intercept values, but coefficients can still be estimated accurately.
- ***correlated*** with the remaining independent variables, the model will be unreliable as the regression coefficients, intercept, and residuals will be biased and inconsistent.

b. Inappropriate Forms of Variables

One or more of the regression variables may need to be *transformed or corrected* (for example, by taking the natural logarithm of the variable) before estimating the regression.

c. Inappropriate Scaling of Variables

Using unscaled data in regressions instead of scaled data can lead to model misspecification. Analysts must decide whether to scale variables (such as common-size statements) before comparing data across firms.

d. Inappropriate Pooling of Data

The regression model pools data from different samples that should not be pooled. This misspecification occurs when the sample spans structural breaks or regime changes in the data such as change in Govt. regulations, low/high market volatility etc.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

3. VIOLATIONS OF REGRESSION ASSUMPTIONS: HETEROSKEDASTICITY

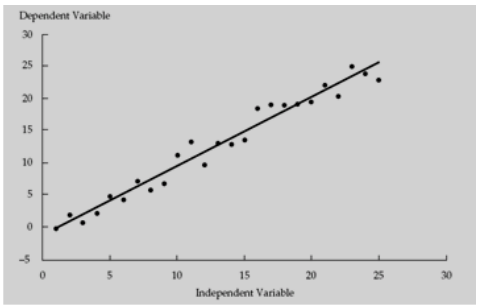
Heteroskedasticity

Heteroskedasticity occurs when the variance of the errors differs across observations i.e., variances of errors are not constant.

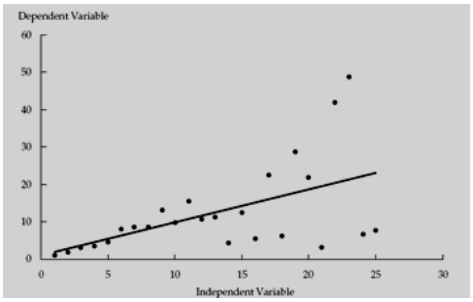
Heteroskedasticity may arise from model misspecifications described in the previous section.

Note: Linear regression model assumes that variance of errors is homoscedastic.

Regressions with Homoskedasticity



Regressions with Heteroskedasticity



The Consequences of Heteroskedasticity

Types of Heteroskedasticity:

1. **Unconditional Heteroskedasticity** - occurs when the error variance is not correlated with independent variables. This violates the assumption of linear regression but does not cause significant issues with regression as the error variance does not systematically increase or decrease with changes in the independent variable.

2. **Conditional Heteroskedasticity -** exists when error variance is correlated with the independent variables. This means that error variance increases as the value of independent variable increases. It is more problematic than unconditional heteroskedasticity.

Note: Conditional heteroskedasticity can make it seem like there are significant correlations where there are none, which can cause more Type I errors.

Testing for Conditional Heteroskedasticity

1. ***Plotting residuals:*** A scatter plot of the residuals versus one or more of the independent variables can describe patterns among observations (as shown below).

2. ***Breusch–Pagan (BP) test***: The Breusch–Pagan test involves regressing the squared residuals from the estimated regression equation on the independent variables in the regression.

H_0 = No conditional Heteroskedasticity exists
H_A = Conditional Heteroskedasticity exists

$$\textbf{Test statistic} = n \times R^2_{residuals}$$

where,

$R^2_{residuals}$ = *R^2 from a second regression of the squared residuals from the first regression on the independent variables*
n = number of observations

- Critical value is calculated from χ^2 distribution table with df = k.
- It is a one-tailed test since we are concerned only with large values of the test statistic.

Decision Rule: When test statistic > critical value, Reject H_0 and conclude that error terms in the regression model are conditionally Heteroskedastic.

- If no conditional heteroskedasticity exists, the independent variables will not explain much of the variation in the squared residuals.
- If conditional heteroskedasticity is present in the original regression, the independent variables will explain a significant portion of the variation in the squared residuals.

Refer to: Exhibit 4 and 5 for practicing example from the CFA Institute's Curriculum.

Correcting for Heteroskedasticity

One method to correct the effects of conditional heteroskedasticity is:

Computing ***robust standard errors*** (heteroskedasticity-consistent standard errors or white-corrected standard errors), corrects the standard errors of the linear regression model's estimated coefficients to deal with conditional heteroskedasticity.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

4. VIOLATIONS OF REGRESSION ASSUMPTIONS: SERIAL CORRELATION

Serial Correlation

When regression errors are correlated across observations, then errors are serially correlated (or auto correlated). Serial correlation typically arises in time-series regressions.

The Types of Serial Correlation

1. ***Positive serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a positive (negative) error for another observation.

2. ***Negative serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a negative (positive) error for another observation.

Note: Positive serial correlation is the most common type and here we assume first-order serial correlation between neighboring observations, meaning that the residual sign tends to continue in a time series.

The Consequences of Serial Correlation

- The principal problem caused by serial correlation in a linear regression is an incorrect estimate of the regression coefficient standard errors.
- When one of the independent variables is a lagged value of the dependent variable, then serial correlation causes all the parameter estimates to be inconsistent and invalid. Otherwise, serial correlation does not affect the consistency of the estimated regression coefficients.
- Serial correlation leads to wrong inferences.
- In case of positive (negative) serial correlation: *Standard errors are underestimated (overestimated) → T-statistics & F-statistics are inflated (understated) → Type-I (Type-II) error increases.*

Testing for Serial Correlation

Durbin-Watson (DW) and Breusch-Godfrey (BG) are the most common tests for serial correlation.

1. The **Durbin Watson statistic** is used to test for serial correlation.

   The DW test measures autocorrelation by comparing squared differences of successive residuals to squared residuals total. However, it only tests for first-order serial correlation.

   The DW statistic tests the null hypothesis of no autocorrelation against the alternative hypothesis of positive (or negative) autocorrelation.

2. The **Breusch-Godfrey (BG) test** is more reliable because it can find autocorrelation up to a pre-set order p, where the error in period t is related to the error in period t – p.

The BG statistic tests the null hypothesis of no autocorrelation up to lag p against the alternative hypothesis of positive (or negative) autocorrelation for at least one lag. Assuming p equals 1, there is one lagged residual independent variable.

Using an F-statistic with n-p-k-1 degrees of freedom and p lags, the resulting P-value can be compared to the significance level to evaluate the hypotheses.

Refer to: Exhibit 8 and 9 for practicing example from the CFA Institute's Curriculum.

Correcting for Serial Correlation

Adjustment Method: To correct the impact of serial correlation, the standard errors of linear regression coefficient estimates are ***adjusted***. This method also corrects for conditional heteroskedasticity. Some of the names given to the adjustment method are:

- serial-correlation consistent standard errors
- serial correlation and heteroskedasticity adjusted standard errors
- Newey-West standard errors
- robust standard errors

Refer to: Exhibit 10 for practicing example from the CFA Institute's Curriculum.

Statistical tests require consideration of serial correlation and heteroskedasticity. Efficient financial market data should ideally not have these issues. However, if they do exist, residual patterns can be used for trading before they are removed by other market participants.

Practice: Question under 'Knowledge Check' 'from the CFA Institute's Curriculum.

5. VIOLATIONS OF REGRESSION ASSUMPTIONS: MULTICOLLINEARITY

Multicollinearity

Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other.

Consequences of Multicollinearity

- A high degree of multicollinearity can make it difficult to detect significant relationships.
- Multicollinearity does not affect the consistency of the estimates of the regression coefficients, but estimates become extremely imprecise and unreliable.
- It does not affect F-statistic.
- The multicollinearity problem does not result in biased coefficient estimates; however, standard errors of regression coefficients can increase, causing insignificant t-tests and wide confidence intervals i.e., Type-II error increases.

Detecting Multicollinearity

- High pairwise correlations among independent variables may not indicate multicollinearity while low pairwise correlations do not prove the absence of multicollinearity. Correlation between independent variables is only useful in indicating multicollinearity in case of two independent variables.
- The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant.
- Variance inflation factor (VIF) is used to quantify multicollinearity problems

Variance Inflation Factor

Multicollinearity can be quantified using the variance inflation factor (VIF) for each independent variable. Suppose there are k independent variable. We begin by regressing one independent variable (suppose X_j) against the other remaining k-1independent variables.

The following equation is then used to compute the VIF of the variable under consideration.

$$VIF_j = \frac{1}{1 - R_j^2}$$

If there is no correlation between the variable and other independent variables, $R_j^2 = 0$ and therefore VIF_j value is 1. The minimum VIF_j value is 1

As correlation increases, VIF_j value also increases. Higher VIF_j means an independent variable can be predicted from other variables, making it redundant.

Rule of Thumb

- $VIF_j > 5$, indicates that further investigation is required
- $VIF_j > 10$, indicates severe multicollinearity that needs to be corrected.

Practice: Question under 'Identifying Multicollinearity as a Problem' 'from the CFA Institute's Curriculum.

Correcting for Multicollinearity

The problem of multicollinearity can be corrected by:

- Increasing the sample size
- Using a different proxy for one of the variables
- excluding one or more of the regression variables

Multicollinearity can be difficult to deal with but experimenting with different independent variables can help find the best solution. If predicting the dependent variable is the only goal, multicollinearity may not be a major concern.

Practice: Questions under 'Knowledge Check' 'from the CFA Institute's Curriculum.

Practice: End-of-Chapter Questions from the CFA Institute's Curriculum and Questions from FinQuiz Question-bank.

Summary of Violations of Assumptions from Model Misspecification

	Heteroskedasticity	Serial Correlation	Multicollinearity
Assumption	• Error terms should be homoscedastic	• Independence of Observations	• Independence of Independent Variables
Violation	• Heteroskedastic error terms i.e., Errors are correlated with earlier independent variable.	• Serial correlation i.e., Errors are correlated with earlier errors	• Multicollinearity i.e., independent variables are strongly correlated with each other
Issue	• Biased estimates of standard errors of coefficients.	• Biased standard errors • Inconsistent estimates of coefficients.	• Inflated standard errors
Detection	• Visual inspection of residuals • BP Test	• BG Test	• VIF
Correction	• Revision of Model • Use robust standard errors	• Revision of Model • Use serial-correlation consistent standard errors	• Revision of Model • Increase sample size

2 INFLUENCE ANALYSIS

Influential Data Points

Influential Observation: An observation whose inclusion can dramatically affect the outcome of a regression.

Two types of influential observations are:

1. **High-leverage point**: Extreme value of an independent variable
2. **Outlier:** Extreme value of the dependent variable

Outliers and high-leverage points are uncommon but not necessarily problematic. A high-leverage point may have independent variable values that differ greatly from other observations but still remain close to the solid regression line.

However, the high-leverage point or outlier point far from the regression line causes problems. Extreme values "push" the predicted regression line toward them, affecting slope coefficients and goodness-of-fit.

Detecting Influential Data Points

- In **simple** linear regression, **'scatterplots'** are used to identify outliers and high-leverage points
- In **multiple** linear regression, a '**quantitative approach'** is required to identify and evaluate extreme values.
  - i. Leverage → a tool to identify high-leverage point.
  - ii. Studentized Residual → a tool to identify outlier.
  - iii. Cook's Distance → a tool to measures the effect of removing i^{th} on regression estimates.

i. Leverage (h_{ii})

Leverage is a measure used to identify high-leverage point.

- Leverage measures the difference between an i^{th} observation's value of an independent variable and the mean value of that variable among all n observations.
- It ranges from 0 to 1.
- Higher leverage indicates greater influence on estimated regression.
- The sum of all leverages for observations equals k + 1. the number of independent variables plus one for the intercept.
- Observations with leverage exceeding $3\left(\frac{k+1}{n}\right)$ can have a significant impact.
- Software tools can easily calculate the leverage value.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

ii. Studentized Residuals t_{i^*}

Outliers - observations with an unusual dependent variable values - are measured using '*studentized residuals*'. The following is the procedure for implementing this measure:

Step 1: Create an initial regression model with n observations, then remove one observation at a time and re-estimate the model using the remaining n-1 observations.

Step 2: Compare the observed Y values (on n observations) with the expected Y values from models with the i^{th} observation deleted (on n – 1 observations).

For an observation i the residual between the observed Y (Yi) and the predicted Y with the i^{th} observation deleted is as follows.

Residual $e_i^* = Y_i - \hat{Y}_{i^*}$

Step 3: Now, divide the residual value e_i^* calculated in step 2 by the estimated standard deviation s_{e^*}. The outcome is *studentized residual* value.

$$t_{i^*} = \frac{e_i^*}{s_{e^*}} = {}^{e_i}\sqrt{\frac{n-k-1}{SSE(1-h_{ii})-e_i^2}}$$

where,

e_i = residual with i^{th} observation deleted
s_{e^*} = standard deviation of residuals
k = no. of independent variables
SSE initial model's sum of squares errors
h_{ii} = leverage value of i^{th} observation

Important Note:

- Compare studentized residual value to critical t-value to check for ***influential observations***.
- ***Outliers*** can be identified if absolute studentized residual value is over 3.

Refer to: Exhibit 5 from the CFA Institute's Curriculum.

iii. Cook's Distance

Cook's distance is also called Cook's D (D_i). It identifies influential data points by removing the i^{th} observation from regression estimations. It is expressed as:

$$D_i = \frac{e_i^2}{k \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

where,

MSE = mean square error
h_{ii} = leverage value for i^{th} observation
e_i = residual for i^{th} observation
k = no. of independent variables

Key Points

1. Cook's D detects extreme values of both dependent and independent variables by using residuals and leverages.

2. A greater D_i value implies that the ith observation greatly influences the estimated values of the regression. According to the practical guidelines if the value of D_i is > than:

   a. 0.5 → i^{th} observation may be influential (further investigation is required).
   b. 1.0 → it is highly likely that the i^{th} observation is an influential data point.
   c. $\sqrt[2]{k/n}$ → it is highly likely that the i^{th} observation is an influential data point.

3. Cook's D measures the change in all estimated regression values when the i^{th} observation is removed from the sample.

Refer to: Exhibit 8 and 9 for practicing example from the CFA Institute's Curriculum.

Summary of Measures of Influential Observations

Measures	Leverage	Studentized Residual	Cook's Distance
Influence of	• independent variable	• dependent variable	• independent variable & • Dependent variable
Process	• h_{ii} ranges from 0 to 1.	• compare calculated $\|t - statisitc\|$ with critical t-value	• compare calculated Cook's D with $\sqrt[2]{k/n}$
Potentially influential when?	• $h_{ii} > 3\left(\frac{k+1}{n}\right)$,	• calculated $\|t - statisitc\|$ > critical t-value.	• calculated Cook's D > $\sqrt[2]{k/n}$

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

3. DUMMY VARIABLES IN A MULTIPLE LINEAR REGRESSION

Defining Dummy Variable

Dummy variable (also called indicator) is a qualitative variable that takes on a value of 1 if a particular condition is true and 0 if that condition is false. It is used to distinguish between groups or categories such as male or female, month of the year effects, etc.

Some examples of a dummy variable in dataset are as follows:

1. Dummy variable may reflect an inherent property of the data.

   For example, if a company belongs to a particular industry (dummy variable =1), otherwise (dummy variable = 0).

2. Dummy variable may be an identified character of the data. Such as, a binary variable that is either true or false.

   For example, dataset before 2008 (dummy variable = 0), dataset after 2008 (dummy variable = 1).

3. Dummy variable may reflect a condition (either true or false) constructed from some characteristic of the data.

   For example, if condition satisfies e.g., revenue of a company exceeds $1 billion (dummy variable =1, otherwise equals to 0).

Note:

If we want to distinguish among n categories, we need n -1 dummy variables because if we use n dummy variables the assumption that there is no exact linear relationship between two or more independent variables is violated.

Visualizing and Interpreting Dummy Variables

1. **Intercept Dummy**: A dummy intercept adds to or subtracts from the intercept if a certain condition is met.

$$Y_i = b_0 + d_0D_i + b_1X_i + e_i$$

- If D = 0, $\rightarrow Y_i = b_0 + b_1X_i + e$
- If D = 1, $\rightarrow Y_i = (b_0 + d_0) + b_1X_i + e$

2. **Slope Dummy**: A slope dummy allows a changing slope when a particular condition is met.

   The slope dummy variable generates an **interaction term** between the X variable and the condition when D = 1.

$$Y_i = b_0 + b_1X_i + d_iD_iX_i + e_i$$

- If D = 0, $\rightarrow Y_i = b_0 + b_1X_i + e$
- If D = 1, $\rightarrow Y_i = b_0 + (b_1 + d_1)X + e$

3. **Both**: Regressions can use dummies in both slope and intercept. This model allows for a variation in intercept and slope between the two groups.

$$Y_i = b_0 + d_0D_i + b_1X_i + d_iD_iX_i + e_i$$

- If D = 0, $\rightarrow Y_i = b_0 + b_1X_i + e$
- If D = 1, $\rightarrow Y_i = (b_0 + d_0) + (b_1 + d_1)X + e$

Testing for Statistical Significance of Dummy Variables

Refer to: Exhibit 12 and 13 for practicing example from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

4. MULTIPLE LINEAR REGRESSION WITH QUALITATIVE DEPENDENT VARIABLES

Qualitative Dependent Variables

Qualitative dependent variables are dummy variables used as dependent variables instead of independent variables. They can be binary or may fall into more than two categories.

For example, to predict a company's bankruptcy, we need a binary qualitative dependent variable (bankruptcy vs. no bankruptcy) and independent factors such as return on equity, debt-to-equity ratio, or debt rating.

Linear regression cannot estimate this model, but a linear probability model can be used by using Y=1 if bankrupt and 0 if not, in a linear model with three independent variables.

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$$

This form has two problems.

i. Firstly, the predicted value of the dependent variable Y is binary (not discrete) and cannot be greater than 1 or less than 0.
ii. Secondly, the linear relationship between the dependent variable and independent factors may not be realistic.

Logistic Transformation

To solve these issues, a *nonlinear transformation* is applied to the bankruptcy probability using the **logistic transformation**.

The logistic transformation can be calculated using the following formula:

ln[P/(1 − P)]

P/(1-P) is the ratio which represents the probability of an event happening versus not happening.

For example, if the chance of a company going bankrupt is 0.75, the ratio would be 3 calculated as $[\frac{0.75}{1-0.75}]$, indicating that the odds of going bankrupt are three times more likely than not going bankrupt.

The **log odds**, also called the **logit function**, is the natural logarithm (ln) of the odds that something will happen.

Logistic regression (logit) uses the log odds, ln[P/(1 − P)], as the dependent variable:

$$\ln\left[\frac{P}{(1-P)}\right] = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$$

$$P = \frac{1}{1+\exp[-(b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i})]}$$

Logistic transformation helps linearize the relationship between dependent and independent variables. Logistic regression is preferred over linear regression to estimate the likelihood of categorical dependent variables like bankruptcy.

Maximum Likelihood Estimation (MLE): Maximum likelihood estimation (MLE) is used to estimate logistic regression coefficients.

The MLE approach maximizes data likelihood to estimate logistic regression coefficients that make sample choices most likely. It is based on the binomial distribution and requires software to compute. MLE maximizes log likelihood through iteration until the difference between two iterations is negligible.

- Logistic regression uses the logistic transformation of event probability as the dependent variable which makes interpretation of regression coefficients difficult.
- In a logit model, slope coefficient represents the change in log odds of the event occurring per unit change in independent variable.
- The odds ratio is calculated using exponential function to slope coefficients, e^{b_i} which is inverse of natural log.
- Hypothesis testing process in logit regression is same as ordinary least squares regression.

Likelihood Ratio (LR): The likelihood ratio (LR) is a test to assess the fit of logistic regression models that is based on the log-likelihood measure. The LR test statistic is presented below.

LR = -2(Log likelihood restricted model – Log likelihood unrestricted model).

- The LR test assesses logistic regression model fit by comparing the log-likelihoods of restricted and unrestricted models.
- It is similar to the joint F-test in multiple regression but uses chi-squared distribution.
- The log-likelihood metric is always negative therefore a higher log-likelihood value indicates a better fit.
- Log-likelihood metric is not significant alone but helpful for comparing regression models with the same dependent variable.

Suppose, model A is unrestricted model and mode B is with restrictions where $b_2 = b_3 = 0$.

Model A: Unrestricted model

$$= \ln\left[\frac{P}{(1-P)}\right] = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$$

Model B: Restricted model

$$= \ln\left[\frac{P}{(1-P)}\right] = b_0 + b_1X_{1i}$$

H_0: $b_1 = b_2 = 0$
H_a: either b_1 or $b_2 \neq 0$

LR is a joint test of restricted coefficients. LR performs well when sample size is large.

Pseudo- R^2: Logistic regression doesn't use least squares and therefore R^2 doesn't apply. Instead, standard software generates pseudo- R^2 that captures explained variation. However, pseudo- R^2 can only be used to compare model specifications of the same model.

Uses in Machine Learning and for Sentiment Analysis:
Logistic regression is used in machine learning and neural networks for binary classification. It is also used in natural language processing to classify financial news sentiment as positive or negative to aid in investment analysis. The model uses financial text tokens from annual reports, earnings releases, and business announcements as independent variables to classify the sentiment of the news. This helps improve investment valuation.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

Practice: End-of-Chapter Questions from the CFA Institute's Curriculum and Questions from FinQuiz Question-bank.

1. INTRODUCTION TO TIME SERIES ANALYSIS AND

A time series is any series of data that varies over time e.g., the quarterly sales for a company during the past five years or daily returns of a security.

Time-series models are used to:

1. explain the past
2. predict the future of a time-series

Challenges of Working with Times Series

When assumptions of the regression model are not met, we need to transform the time series or modify the specifications of the regression model.

Problems in time series:

1. When the dependent and independent variables are distinct, presence of serial correlation of the errors does not affect the consistency of estimates of intercept or slope coefficients.

But in an autoregressive time-series regression, presence of serial correlation in the error term makes estimates of the intercept (b_0) and slope coefficient (b_1) to be inconsistent.

2. When mean and/or variance of the time series model changeover time and is not constant, then using an autoregressive model will provide invalid regression results.

Because of these problems in time series, time series model is needed to be transformed for the purpose of forecasting.

2. LINEAR TREND MODELS

Linear Trend Models

In a linear trend model, the dependent variable changes at a ***constant rate*** with time.

$$y_t = b_0 + b_1 t + \varepsilon_t$$

where,

y_t = *value of time series at time t (value of dependent variable)*
b_0 = *y-intercept term*
b_1 = *slope coefficient or trend coefficient*
t = *time, the independent or explanatory variable*
ε_t = *random error term*

The predicted or fitted value of y_t in period 1 is:

$$\hat{y}_1 = \hat{b}_0 + \hat{b}_1(1)$$

The predicted or fitted value of y_t in period 5 is:

$$\hat{y}_5 = \hat{b}_0 + \hat{b}_1(5)$$

The predicted or fitted value of y_t in period T + 1 is:

$$\hat{y}_{T+1} = \hat{b}_0 + \hat{b}_1(T+1)$$

NOTE:

Each consecutive observation in the time series increases by $\hat{b}_1$ in a linear trend model.

Practice: Example 1
CFA Institute's Curriculum.

3. LOG-LINEAR TREND MODELS

When time series has exponential growth rate, it is more appropriate to use log-linear trend model instead of linear trend model. Exponential growth rate refers to a ***constant growth*** at a particular rate.

$$y_t = e^{b_0 + b_1 t}$$

where,

t = 1, 2, 3,…,T

Taking natural log on both sides we have:

$\ln y_t = b_0 + b_1 t + \varepsilon_t$

where,

t = 1, 2, 3, …,T

Linear trend model	Log-linear trend model
Predicted trend value of y_t is $\hat{b}_0 + \hat{b}_1 t$,	Predicted trend value of y_t is $e^{\hat{b}_0+\hat{b}_1 t}$ because $e^{\ln y_t} = y_t$.
The model predicts that y_t grows by a ***constant amount*** from one period to the next.	The model predicts a ***constant growth rate*** in y_t of $e^{b_1} - 1$.
A linear trend model is appropriate to use when the residuals from a model are equally distributed above and below the regression line e.g., inflation rate.	A log-linear model is appropriate to use when the residuals of the model exhibit a persistent trend i.e., either positive or negative for a period of time e.g., financial data i.e. stock prices, sales, and stock indices.

Practice: Example 2 & 3
CFA Institute's Curriculum.

Limitation of Trend Models: Trend model is based on only one independent variable i.e. time; therefore, it does not adequately incorporate the underlying dynamics of the model.

4. TREND MODELS AND TESTING FOR CORRELATED ERRORS

In case of presence of serial correlation, both the linear trend model and the log-linear trend model are not appropriate to use. In case of serial correlation, autoregressive time series models represent better forecasting models.

5. AR TIME-SERIES MODELS AND COVARIANCE-STATIONARY SERIES

An autoregressive (AR) model is a time series regression in which the independent variable is a lagged (past) value of the dependent variable i.e.

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$$

First order autoregressive AR (1) for the variable x_t is:

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$$

A pth-order autoregressive AR (p) for the variable x_t is:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2} + \ldots\ldots + b_p x_{t-p} + \varepsilon_t$$

Covariance-Stationary Series

In order to obtain a valid statistical inference from a time-series analysis, the time series must be covariance stationary.

Time series is covariance stationary when:

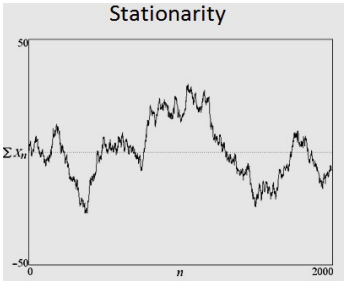
1. The expected value of the time series is constant and finite in all periods.
2. The variance of the time series is constant and finite in all periods.
3. The covariance of the time series with past or future values of itself is constant and finite in all periods.

NOTE:

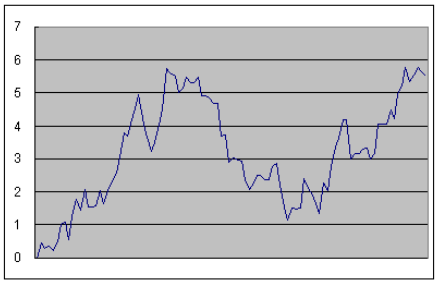
Weakly stationary also refers to covariance stationary.

Stationary Data: When a time series variable does not exhibit any significant upward or downward trend over time.

Nonstationary Data: When a time series variable exhibits a significant upward or downward trend over time.



Non-stationarity (upward trend)



Consequence of Covariance Non-Stationarity: When time series is not covariance stationary, the regression estimation results are invalid because:

- The "*t*-ratios" will not follow a *t*-distribution.
- The estimate of b_1 will be biased and any hypothesis tests will be invalid.

NOTE:
Stationarity in the past does not guarantee stationarity in the future because state of the world may change over time.

6. DETECTING SERIALLY CORRELATED ERRORS IN AN AUTOREGRESSIVE MODEL

An Autoregressive model can be estimated using ordinary least squares model (OLS) when the time series is covariance stationary, and the errors are uncorrelated.

Detecting Serial Correlation in AR models: In AR models, Durbin-Watson statistic cannot be used to test serial correlation in errors. In such cases, t-test is used.

The autocorrelations of time series refer to the correlations of that series with its own past values.

- When autocorrelations of the error term are zero, the model can be specified correctly.
- When autocorrelations of the error term are significantly different from zero, the model cannot be specified correctly.

Example:

Suppose a sample has 59 observations and one independent variable. Then,

S.D = 1 / $\sqrt{T}$ = 1 / $\sqrt{59}$ = 0.1302
Critical value of t (at 5% significant level with df = 59 – 2 = 57) is 2.

Suppose autocorrelations of the Residual are as follows:

Lag	Autocorrelation	Standard Error	t-statistic*
1	0.0677	0.1302	0.5197
2	-0.1929	0.1302	-1.4814
3	0.0541	0.1302	0.4152
4	-0.1498	0.1302	-1.1507

* **t-statistic = Autocorrelations / Standard Error**

It can be seen from the table that none of the first four autocorrelations has t-statistic > 2 in absolute value.

Conclusion: None of these autocorrelations differ significantly from 0 thus, residuals are not serially correlated, and model is specified correctly and OLS can be used to estimate the parameters and the standard errors of the parameters in the autoregressive model.

Correcting Serial Correlation in AR models: The serial correlation among the residuals in AR models can be removed by estimating an autoregressive model by adding more lags of the dependent variable as explanatory variables.

Practice: Example 4
CFA Institute's Curriculum.

7. MEAN REVERSION AND MULTIPERIOD FORECASTS

Mean Reversion

A time series shows mean reversion if it tends to move towards its mean i.e. decrease when its current value is above its mean and increase when its current value is below its mean.

- When a time series equals its mean-reverting value, then the model predicts that the value of the time series will be the same in the next period i.e. $x_{t+1} = x_t$.

Mean reverting level of $x_t = \frac{b_0}{1-b_1}$

- Time series will remain the same if its current value $= \frac{b_0}{1-b_1}$
- Time series will Increase if its current value $< \frac{b_0}{1-b_1}$
- Time series will Decrease if its current value $> \frac{b_0}{1-b_1}$

Multiperiod Forecasts and the Chain Rule of Forecasting

The **chain rule of forecasting** is a process in which a predicted value two periods ahead is estimated by first predicting the next period's value and substituting it into the equation of a predicted value two periods ahead i.e.

The one-period ahead forecast of x_t from an AR (1) model is as follows:

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t$$

Two-period ahead forecast is:

$$\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 x_{t+1}$$

NOTE:

Multiperiod forecast is more uncertain than single period forecast because the uncertainty increases when number of periods in the forecast increase.

Example:

The one-period ahead forecast of x_t from an AR (1) model when $x_t = 0.65$ is as follows:

$$\hat{x}_{t+1} = 0.0834 + 0.8665(0.65) = 0.6466$$

Two-period ahead forecast is:

$$\hat{x}_{t+2} = 0.0834 + 0.8665(0.6466) = 0.6437$$

Practice: Example 6
CFA Institute's Curriculum.

8. COMPARING FORECAST MODEL PERFORMANCE

The accuracy of the model depends on its forecast error variance.

- *The smaller the forecast error variance, the more accurate the model will be.*

In-sample forecast errors: These are the residuals from the fitted time series model i.e., residuals within a sample period.

Out-of-sample forecast errors: These are the residuals outside the sample period. It is more important to have smaller forecast error variance (i.e., high accuracy) for out-of-sample forecasts because the future is always out of sample.

To evaluate the out-of-sample forecasting accuracy of the model, ***Root mean squared error*** (RMSE) is used. RMSE is the square root of the average squared error.

Decision Rule: *The smaller the RMSE, the more accurate the model will be.*

The **RMSE (Root Mean Squared Error)** is used as a criterion for comparing forecasting performance of different forecasting models. To accurately evaluate uncertainty of forecast, both the uncertainty related to the error term and the uncertainty related to the estimated parameters in the time-series model must be considered.

NOTE:

If the model has the lowest RMSE for in-sample data, it does not guarantee that the model will have the lowest RMSE for out-of-sample data as well.

Practice: Example 7 & 8
CFA Institute's Curriculum.

9. INSTABILITY OF REGRESSION COEFFICIENTS

When the estimated regression coefficients in one period are quite different from those estimated during another period, this problem is known as instability or nonstationarity.

The estimates of regression coefficients of the time-series model can be different across different sample periods i.e., the estimates of regression coefficients using shorter sample period will be different from using longer sample periods. Thus, sample period selection is one of the important decisions in time series regression analysis.

- Using longer time periods increase statistical reliability but estimates are not stable.
- Using shorter time periods increase stability of the estimates but statistical reliability is decreased.

NOTE: We cannot select the correct sample period for the regression analysis by simply analyzing the autocorrelations of the residuals from a time-series model. In order to select the correct sample, it is necessary that data should be Covariance Stationary.

Practice: Example 9
CFA Institute's Curriculum.

10. RANDOM WALKS

Random Walks

A. Random walk without drift: In a random walk without drift, the value of the dependent variable in one period is equal to the value of the series in the previous period plus an unpredictable random error.

$x_t = x_{t-1} + \varepsilon_t$

where,

$b_0 = 0$ and $b_1 = 1$.

In other words, the best predictor of the time series in the next period is its current value plus an error term.

The following conditions must hold:

1. Error term has an expected value of zero.
2. Error term has a constant variance.
3. Error term is uncorrelated with previous error terms.

- The equation of a random walk represents a special case of an AR (1) model with $b_0 = 0$ and $b_1 = 1$.
- AR (1) model cannot be used for time series with random walk because random walk has no finite mean, variance and covariance. In random walk $b_0 = 0$ and $b_1 = 1$, so $\frac{b_0}{1 - b_1} = 0 / 0 =$ undefined mean reverting level.
- A standard regression analysis cannot be used for a time series that is random walk.

Correcting Random Walk: When time series has a random walk, it must be converted to covariance-stationary time series by taking the first difference between x_t and x_{t-1} i.e. equation becomes:

$y_t = x_t - x_{t-1} = \varepsilon_t$

- Thus, best forecast of y_t made in period t-1 is 0. This implies that the best forecast is that the value of the current time series x_{t-1} will not change in future.

After taking the first difference, the first differential variable y_t becomes covariance stationary. It has $b_0 = 0$ and $b_1 = 0$ and mean reverting level = 0/1 = 0.

- The first differential variable y_t can now be modeled using linear regression.
- However, modeling the first differential variable y_t with an AR (1) model is not helpful to predict the future because $b_0 = 0$ and $b_1 = 0$.

Consequences of Random Walk: When the model has random walk, its R^2 will be significantly high and at the same time changes in dependent variable are unpredictable. In other words, the statistical results of the regression will be invalid.

B. Random walk with a drift: In a random walk with a drift, dependent variable increases or decreases by a constant amount in each period.

$x_t = b_0 + x_{t-1} + \varepsilon_t$

where,
$b_0 \neq 0$ and $b_1 = 1$.

By taking first difference,
$y_t = x_t - x_{t-1} = b_0 + \varepsilon_t$

NOTE:

All random walks (with & without a drift) have unit roots.

Practice: YEN/US Dollar Exchange Rate Question. CFA Institute's Curriculum.

11. THE UNIT ROOT TEST OF NONSTATIONARITY

AR (1) time series model will be covariance stationary only when the absolute value of the lag coefficients $b_1<1$. (Note that when b_1 is > 1 in absolute value, it is known as explosive root).

Detecting Random Walk: When time series has random walk, the series does not follow t-distribution and t-test will be invalid. Therefore, t-statistic cannot be used to test the presence of random walk because standard errors in an AR model are invalid if the model has a random walk. Thus, **Dickey-Fuller** test is used to detect nonstationarity:

Method 1: Examining Autocorrelations of the AR model
Stationary Time Series:

- Autocorrelations at all lags equals to zero, or
- Autocorrelations decrease rapidly to zero as the number of lags increases in the model.

Nonstationary time series:

- Autocorrelations at all lags are not equal to zero, or
- Autocorrelations do not decrease rapidly to zero as the number of lags increases in the model.

Method 2: Using Dickey-Fuller Test

Subtracting x_{t-1} from both sides of AR (1) equation we have:

$$x_t - x_{t-1} = b_0 + (b_1 - 1)\, x_{t-1} + \varepsilon_t$$

(or)

$$x_t - x_{t-1} = b_0 + g_1 x_{t-1} + \varepsilon_t$$

where,

$g_1 = (b_1 - 1)$.

- If $b_1 = 1$, then $g_1 = 0$. This implies that there is a unit root in AR (1) model.

Null Hypothesis: H_0: $g_1 = 0 \rightarrow$ time series has a unit root and is Nonstationary

Alternative Hypothesis: H_1: $g_1 < 0 \rightarrow$ time series does not have a unit root and is Stationary

- t-statistic is calculated for predicted value of g_1 and critical values of t-test are computed from Dickey-Fuller test table (these critical t-values in absolute value > than typical critical t-values).

Practice: Example 11 and 12 CFA Institute's Curriculum.

12. MOVING-AVERAGE TIME SERIES MODELS

Moving average (MA) is different from AR model. MA is an average of successive observations in a time series. It has lagged values of residuals instead of lagged values of dependent variable.

Smoothing Past Values with an n-Period Moving Average

n-period moving average is used to smooth out the fluctuations in the value of a time series across different time periods.

$$\frac{x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-(n-1)}}{n}$$

Drawbacks of Moving Average:

- It is biased towards large movements in the actual data.
- It is not the best predictor of the future.
- It gives equal weights to all the periods in the moving average.

Distinguishing AR time series from a MA time series:

- Autocorrelations of most AR (p) time series start large and decline gradually.
- Autocorrelations of MA (q) time series suddenly drop to 0 after the first q autocorrelations.

Practice: Example 13 and 14
CFA Institute's Curriculum.

13. SEASONALITY IN TIME-SERIES MODELS

When a time series variable exhibit a repeating patterns at regular intervals over time, it is known as seasonality e.g. sales in Dec. > sales in Jan. A time series with seasonality also has a non-constant mean and thus is not covariance stationary.

Detecting seasonality: In case of seasonality in the data, autocorrelation in the model differs by season. For example, in case of quarterly sales data of a company, if the fourth autocorrelation of the error term differs significantly from 0 → it is a sign of seasonality in the model.

Decision Rule: When t-statistic of the fourth lag of autocorrelations of the error > critical t-value → reject null hypothesis that fourth autocorrelations is 0. Thus, there is seasonality problem.

Correcting Seasonality: This problem can be solved by adding seasonal lags in an AR model i.e., after including a seasonal lag in case of quarterly sales data, the AR model becomes:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-4} + \varepsilon_t$$

In case of monthly sales data, the AR model becomes:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-12} + \varepsilon_t$$

NOTE:

R^2 of the model without seasonal lag will be less than the R^2 of the model with seasonal lag. This implies that when time series exhibit seasonality, including a seasonal lag in the model improves the accuracy of the model.

Practice: Example Seasonality in Sales and Example 16
CFA Institute's Curriculum.

14. AUTOREGRESSIVE MOVING-AVERAGE MODELS (ARMA) AND AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY MODELS (ARCH)

An ARMA model combines both autoregressive lags of the dependent variable and moving-average errors.

Drawbacks of ARMA model:

- Parameters of ARMA models are usually very unstable.
- ARMA models depend on the sample used.
- Choosing the right ARMA model is a difficult task because it is more of an art than a science.

Autoregressive Conditional Heteroskedasticity Models (ARCH)

When regression model has (conditional) heteroskedasticity i.e. variance of the error in a particular time-series model in one period depends on the variance of the error in previous periods, standard errors of the regression coefficients in AR, MA or ARMA models will be incorrect, and hypothesis tests would be invalid.

ARCH model:

ARCH model must be used to test the existence of conditional heteroskedasticity. An ARCH (1) time series is the one in which the variance of the error in one period depends on size of the squared error in the previous period i.e. if a large error occurs in one period, the

variance of the error in the next period will be even larger.

To test whether time series is ARCH (1), the squared residuals from a previously estimated time-series model are regressed on the constant and first lag of the squared residuals i.e.

$$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 \hat{\varepsilon}^2_{t-1} + \mu_t$$

where,

μ_t is an error term

Decision Rule: If the estimate of α_1 is statistically significantly different from zero, the time series is ARCH (1). If a time-series model has ARCH (1) errors, then the variance of the errors in period t+1 can be predicted in period t using the formula:

$$\hat{\sigma}^2_{t+1} = \hat{\alpha}_0 + \alpha_1 \hat{\varepsilon}^2_t$$

Consequences of ARCH:

- Standard errors for the regression parameters will not be correct.
- When ARCH exists, we can predict the variance of the error terms.

Generalized least squares or other methods that correct for heteroskedasticity must be used to estimate the correct standard error of the parameters in the time-series model.

Autoregressive model versus ARCH model:

- Using AR (1) model implies that model is correctly specified.
- Using ARCH (1) implies that model cannot be correctly specified due to existence of conditional heteroskedasticity in the residuals; therefore, ARCH (1) model is used to forecast variance/volatility of residuals.

Practice: Example 17
CFA Institute's Curriculum.

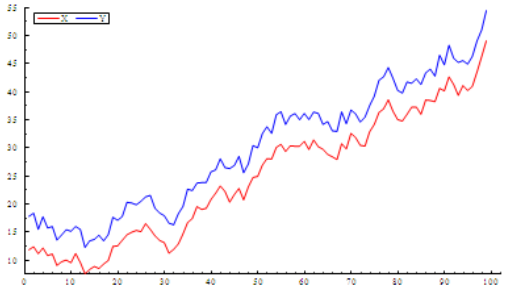
15. REGRESSIONS WITH MORE THAN ONE TIME SERIES

1. When neither of the time series (dependent & independent) has a unit root, linear regression can be used.
2. One of the two time series (i.e. either dependent or independent but not both) has a unit root, we should not use linear regression because error term in the regression would not be covariance stationary.
3. If both time series have a unit root, and the time series are not cointegrated, we cannot use linear regression.
4. If both time series have a unit root, and the time series is cointegrated, linear regression can be used. Because, when two time series are cointegrated, the error term of the regression is covariance stationary and the t-tests are reliable.

Cointegration: Two time series are cointegrated if

- A long term financial or economic relationship exists between them.
- They share a common trend i.e. two or more variables move together through time.

Two Cointegrated Time Series



NOTE:

Cointegrated regression estimates the long-term relation between the two series. Therefore, it is not the best model of the short-term relation between the two series.

Detecting Cointegration: The Engle-Granger Dickey-Fuller test can be used to determine if time series are cointegrated.

Engle and Granger Test:

1. Estimate the regression $y_t = b_0 + b_1 x_t + \varepsilon_t$
2. Unit root in the error term is tested using Dickey-fuller test but the critical values of the Engle-Granger are used.
3. If test fails to reject the null hypothesis that the error term has a unit root, then error term in the regression is not covariance stationary. This implies that two time series are not cointegrated and regression relation is spurious.
4. If test rejects the null hypothesis that the error term has a unit root, then error term in the regression is covariance stationary. This implies that two time series are cointegrated and regression results and parameters will be consistent.

NOTE:

- When the first difference is stationary, series has a single unit root. When further differences are required to make series stationary, series is referred to have multiple unit roots.
- For multiple regression model, rules and procedures for unit root and stationarity are the same as that of single regression.

Practice: Example 18 to 21
CFA Institute's Curriculum.

16. OTHER ISSUES IN TIME SERIES

Suggested Steps in Time-Series Forecasting

Following is a guideline to determine an accurate model to predict a time series.

1. Select the model on the basis of objective i.e., if the objective is to predict the future behavior of a variable based on the past behavior of the same variable, use Time series model and if the objective is to predict the future behavior of a variable based on assumed causal relationship with other variables Cross sectional model should be used.
2. When time-series model is used, plot the series to detect Covariance Stationarity in the data. Trends in the time series data include:

- A linear trend
- An exponential trend
- Seasonality
- Structural change i.e., a significant shift in mean or variance of the time series during the sample period

3. When there is no seasonality or structural change found in the data, linear trend or exponential trend is appropriate to use i.e.

i. Use linear trend model when the data plot on a straight line with an upward or downward slope.
ii. Use log-linear trend model when the plot of the data exhibits a curve.
iii. Estimate the regression model.
iv. Compute the residuals
v. Use Durbin-Watson statistic to test serial correlation in the residual.

4. When serial correlation is detected in the model, AR model should be used. However, before using AR model, time series must be tested for Covariance Stationarity.

- If time series has a linear trend and covariance nonstationary; it can be transformed into covariance stationary by taking the first difference of the data.
- If time series has exponential trend and covariance nonstationary; it can be transformed into covariance stationary by taking natural log of the time series and then taking the first difference.
- If the time series exhibits structural change, two different time-series model (i.e. before & after the shift) must be estimated.
- When time series exhibits seasonality, seasonal lags must be included in the AR model.

5. When time series is converted into Covariance Stationarity, AR model can be used i.e.

- Estimate AR (1) model;

- Test serial correlation in the regression errors; if no serial correlation is found only then AR (1) model can be used. When serial correlation is detected in AR (1), then AR (2) should be used and tested for serial correlation. When no serial correlation is found, AR (2) can be used. If serial correlation is still present, order of AR is gradually increasing until all serial correlation is removed.

6. Plot the data and detect any seasonality. When seasonality is present, add seasonal lags in the model.
7. Test the presence of autoregressive conditional heteroskedasticity in the residuals of the model i.e. by using ARCH (1) model.
8. In order to determine the better forecasting model, calculate out-of-sample RMSE of each model and select the model with the lowest out-of-sample RMSE.

Practice: CFA Institute's Curriculum End of Chapter Questions & FinQuiz Question-bank (Item-sets +Questions).

1. INTRODUCTION

Machine Learning and Investment Management

Machine learning models are affecting investment management processes – from client profiling, to asset allocation, to stock selection, to portfolio construction and risk management, to trading.

Machine learning can provide significant contribution to the asset and wealth management value chain in areas including:

- Retirement savings
- Alpha generation for security selection
- Calculating target portfolio weights

Research demonstrates that machine learning solutions outperform mean-variance optimization in portfolio construction.

Machine learning is already creating better order flow management tools with non-linear trading algorithms which are reducing the costs of implementing portfolio decisions.

2. WHAT IS MACHINE LEARNING?

1 Defining Machine Learning

ML extracts knowledge from large amounts of data without any restrictions. The goal of ML is to automate decision-making processes by learning from known examples to determine an underlying structure in the data.

ML is appropriate to use with problems which have:

- many variables (high dimensionality)
- with a high degree of non-linearity

ML is divided into supervised, unsupervised learning, and deep learning.

2 Supervised Learning

Supervised learning involves applying ML algorithms to labelled data sets (or training data sets contains matched sets of observed inputs and the associated output) to infer patterns between inputs and output, which is further used to train the algorithm.

Training the algorithm: process of inferring a pattern between inputs and output using a ML algorithm.
Note:

- Once algorithm is trained, the inferred pattern can be used to predict output values based on new inputs.
- The fit of the model is evaluated using labeled test data in which predicted targets are compared to actual targets.
- The dependent variable is the target or output while the independent variables are known as features

Supervised learning is suitable for the following two problem categories:

- Regression:
  - Focus on predicting continuous target variables
  - Examples include multiple linear regression models, problems involving the use of historical stock market returns to forecast stock price performance or use of historical corporate financial ratios to forecast the probability of default.
- Classification:
  - Focus on sorting observation into distinct categories.
  - Classifier: Model which relates outcome to the independent variables (features) when the dependent variable is categorial
  - Many classification models are binary classifiers although multi-category classification is not uncommon.

3 Unsupervised Learning

Unsupervised learning does not make use of labeled data. Inputs are used in the analysis without any targets. The algorithm seeks to discover structure within the data themselves.

Note: Unsupervised learning is more appropriate for new data sets as it provides human insight into data that is too big or complex to visualize.

Unsupervised learning is suitable for the following two problem categories:

- Dimension reduction:
  - Involves reducing features but preserving variations across observations to preserve embedded information
  - May be applied to data with a large number of features to produce lower dimensional representation (with fewer features) so that it can fit on a computer screen
  - Also used in quantitative and risk management applications
- Clustering: Sorting observations into empirically determined groups such that observations in a group are similar and distinct from other groups.
  - Note: Clustering has not used by asset managers to sort companies into conventional groups (based on sectors or countries).

4 Deep Learning and Reinforcement Learning

Deep learning (DL): Sophisticated algorithms address highly complex tasks such as image classification, face recognition, speech recognition, and natural language processing.

Reinforcement learning (RL): A computer learns by interacting with itself (or data generated by the same algorithm).

Neural networks: Include highly flexible ML algorithms which have been successfully applied to a wide variety of tasks characterized by non-linearities and interactions among features. DL and RL are based on this.

5 Summary of ML Algorithms and How to Choose Among Them

GUIDE TO ML ALGORITHMS

	Supervised (Target Variable)	**Unsupervised (No Target Variable)**
Continuous	Regression (Linear, Penalized, Logistic) CART Random Forest	Dimension Reduction (PCA) Clustering (K-Means, Hierarchical)
Categorical	Classification (Logistic, SVM, KNN, CART	
Continuous or Categorical	Neural Networks Deep Learning Reinforcement Learning	

Practice: Example 1 from the CFA Institute's Curriculum..

3. EVALUATING ML ALGORITHM PERFORMANCE

Advantages of ML algorithms versus structured statistical approaches in analyzing and exploring the structure of large data sets:

- Able to determine interaction between feature variables and target variable
- Can process large amounts of data quickly
- Can easily capture non-linear relationships
- May capture and predict structural changes between features and the target

ML algorithms rely on non-parametric and non-linear models which allow for more flexibility in inferring relationships.

Drawbacks of using ML algorithms:

- Can produce overly complex models with difficult to interpret results
- May be sensitive to particular changes in data
- May fit the trading data too well (overfitting)

Overfitting: The fitted algorithm does not generalize well to new data and thus does not predict well using new data.

A model which generalizes well is a model that retains its explanatory power when predicting out-of-sample. An overfit model incorporates the noise or random fluctuations in training data into its learned relationship.

Generalization is an objective to model building and overfitting is a challenge to attaining that objective.

1 Generalization and Overfitting

ML model is typically applied to the following data sets:

1. Training sample (or in-sample): used to train the model
2. Validation sample (or out-of-sample): for validating and tuning the model
3. Test sample (or out-of-sample): for testing the model's ability to predict well on new data

Supervised ML models must generalize beyond the training sample and should retain explanatory power when tested out-of-sample. Failure to do so results in overfitting.

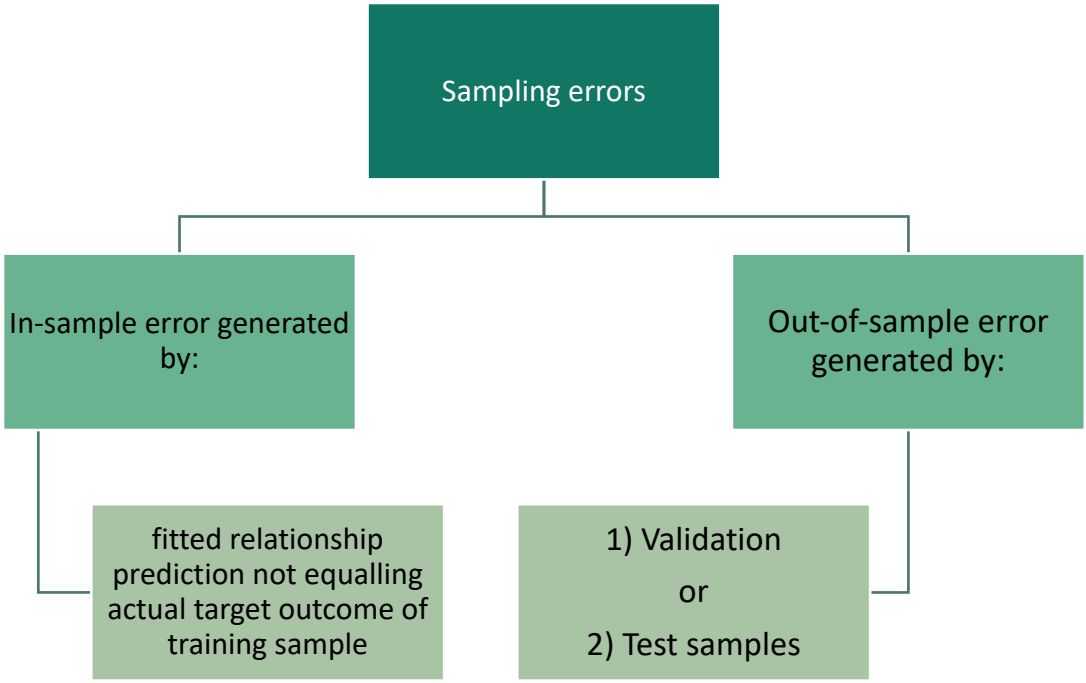
Overfitting increases as:
- noise levels in the data .⇧
- complexity in the model .⇧

Complexity refers to the number of features, terms, or branches in the model and to whether the model is linear (or non-linear). Non-linear models are more complex.

A good fit/robust model fits (in-sample) data well and generalizes well to out-of-sample and in-sample (or training) data, both within acceptable degrees of error.

2 Errors and Overfitting

Data scientists compare in- and out-of-sample error rates as a function of both the data and the algorithm.



Note: Figure for Sample errors:

Low/no in-sample error but large out-of-sample error indicates poor generalization.

Types of total out-of-sample errors:

1. Bias error: the degree to which a model fits the training data. Underfitting and large in-sample errors result when algorithms generate incorrect assumptions ➔ results in high bias with poor approximation .
2. Variance error: change in models in response to new data from validation and test samples. Unstable models pick up noise and produce high variance, causing overfitting and high out-of-sample error.
3. Base error: due to randomness in data.

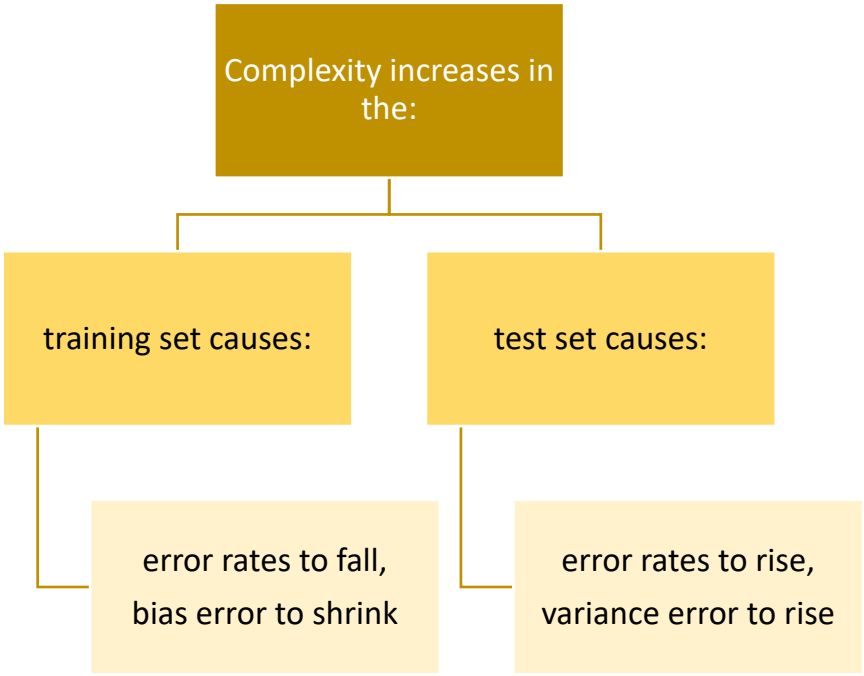
<u>Learning curve</u>
The curve:

- Plots accuracy rate in the validation or test samples on the *y*-axis and number of training samples on the x-axis.
- Describes under- and overfitting as a function of bias and variance errors.

For robust models: As out-of-sample accuracy↑, training sample size ↑. Error rates in the validation or test samples converge towards each other and toward a desired error rate (or base error).

For underfitted models with high bias errors, high error rates will cause convergence between the two rates which is less than the desired accuracy rate. Adding more training samples will not improve the model.
For overfitted models with high variance errors, there is no convergence between the two rates.

Optimal model: Minimizes both bias and variance errors and selects an algorithm with good predictive or classifying power.



- Linear functions are prone to bias error and underfitting
- Non-linear functions are prone to variance error and overfitting
- Optimal point of complexity exists where the bias and variance error curves intersect and in- and out-of-sample error rates are minimized

Fitting curve: shows in- and out-of-sample error rates on the y-axis and model complexity on the x-axis.

Successful generalization occurs when overfitting risk is managed and is the point on the fitting curve just before total error rates start to rise.

3 Preventing Overfitting in Supervised Machine Learning

Methods to reduce overfitting include:

1) Complexity reduction: Prevent the algorithm from getting too complex by reducing number of features and penalizing algorithms which are too complex or flexible
2) Cross validation: Technique for estimating out-of-sample error directly by determining the error in validation samples. Helps avoid sampling bias through proper data sampling.

- Objective is to have a large data set to make training and testing possible on representative samples.
- Note: A small sample size or unrepresentative sample increase its bias
- To reduce sampling bias: Divide data into three group (training sample, validation and test sample) with the goal of positioning model on fresh data from the same domain.

K-fold cross validation:

A cross-validation technique used to mitigate the problem of reducing the training sample too much by excluding holdout samples (data samples not used for model training).

The data samples are shuffled randomly and divided into *k* equal sub-samples. One sample is used as the validation sample and remaining as training samples. Process helps minimize bias and variance.

Practice: Example 2 from the CFA Institute's Curriculum.

4. SUPERVISED MACHINE LEARNING ALGORITHMS

Supervised ML are trained using labeled data and can be divided into two groups:

- Regression for a continuous target variable, which includes:
  - Penalized regression
  - LASSO
- Classification for a categorial or ordinal target variable, which includes:
  - Support vector machine (SVM)
  - *k*-nearest neighbor (KNN)
  - Classification and regression tree (CART) algorithms

The assumption made in the following sections is that there are a number of observations of a target variable, *Y*, and *n* real-valued features that are used to establish a relationship (regression or classification) between *X* and *Y*.

Penalized Regression

The technique is useful for reducing a large number of features into a manageable set which is useful for making predictions in a variety of large data sets, especially when features are correlated.

Penalized regression can help avoid the overfitting problem. Parsimonious models (models in which each variable plays an essential role) are less subject to overfitting.

Includes a constraint that regression coefficients are chosen to minimize the sum of squared residuals and a penalty term that increases in size with number of features. Only important features explaining Y will remain in a penalized regression model while other features will be penalized for being included.

Least Absolute Shrinkage and Selection Operator (LASSO):

- A type of penalized regression
- Penalty term has the following form with $\lambda > 0$:
  - Penalty term = $\lambda \sum_{k=1}^{K} |\bar{b}_k|$
  - Is only added during model building process. Once model is built, it is evaluated using sum of squared residuals generated using the data set.
- Involves minimizing sum of squared residuals and sum of absolute values of the regression coefficient.
- A greater number of included features (variables with non-zero coefficients) will increase the penalty term
- LASSO eliminates less important features
- Ensures that a variable is only included if sum of squared residuals declines by more than the increase in penalty term.

$$\sum_{i=1}^{n}(Y_i - Y_i)^2 + \lambda \sum_{k=1}^{K} |\hat{b}_K|$$

Lambda (λ) determines the balance between fitting the model and keeping the model parsimonious. When λ = 0, LASSO penalized regression = OLS regression.

Regression (in context of penalized regression): Reducing regression coefficient estimates to zero and avoiding complex models and the risk of overfitting.

LASSO has been used in industrial sectors which have scores of features which are collinear. Regularization methods can also be applied to non-linear models.

5. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM)

SVM is a supervised algorithm used for classification, regression, and outlier detection.

Linear classifier: a binary classifier that makes its classification decision based on a linear combination of the features of each data point.

SVM is a linear classifier which determines the hyperplane that separates the observations into two sets of data points. The idea behind SVM is maximizing the

probability of making a correct prediction by determining the boundary that is furthest away from all the observations.

Issues with SVM and how to deal with them:

Observations may be misclassified by the SVM as all data sets may not be linearly separable as may be assumed by the SVM. This problem can be handled by an adaptation to the SVM called soft margin classification which adds a penalty to the objective function for observations in the training set that are misclassified.

Another alternative is that a non-linear SVM algorithm can be run by introducing more advanced, non-linear boundaries reducing the instances of misclassification in the training data sets but will have more features increasing model flexibility.

Uses of SVM:

- SVM is suited for small- to medium-size but highly complex high-dimensional data sets.
- Investors can use SVM to predict company failures to decide which stock to avoid or short sell.
- Can be used to classify text from documents into useful categories for investors.

6. K-NEAREST NEIGHBOR

This represents a supervised learning technique used for classification and sometimes regression. Main idea is to classify a new observation by finding similarities between the new observation and existing data.

Example: Assume a database of corporate bonds classified by credit ratings which also contains detailed information on bond characteristics. If a new bond is issued with no credit rating, existing database bonds of similar issuer and characteristics as the new bond will be used to determine credit rating.

Note:

- KNN results may be sensitive to the inclusion of irrelevant or correlated features so manual selection of features may be necessary
- KNN algorithms work better with a smaller number of features

- Hyperparameter of the model, *k*, must be chosen with care as different values of *k* will lead to different conclusions:
  - High *k* value will lead to too many neighbors
  - Choosing a very small value of *k* will result in high error rate and sensitivity to local outliers
- Application of KNN in investment industry:
  - Bankruptcy prediction
  - Stock price prediction
  - Corporate bond credit rating assignment
  - Customized equity and bond creation

7. CLASSIFICATION AND REGRESSION TREE

Classification and Regression Tree

Classification and Regression Tree (CART) is a supervised ML technique which can be applied to:

- predict a categorial target variable to produce a classification tree,
- predict a continuous target variable to produce a regression tree, or
- to binary classification and regression.

Refer to Exhibit 9 for an illustration of the CART using a binary tree.

When using a binary tree for the CART:

- the classification model is trained from the labeled data.
- the CART algorithm the chooses the feature and cutoff value at each node that generates the widest separation of labeled data to minimize classification error.
- After each decision node, partition of the feature space becomes smaller and smaller

to reduce within-group error for group observations

CART makes no assumptions about the characteristics of the training data and can perfectly learn the data if left unconstrained. To prohibit overfitting, various regularization parameters can be added.

Examples of regularization:

- Defining maximum tree depth, minimum node population, or maximum number of decision nodes.
- Pruning technique used to remove sections of the tree providing little classifying power to reduce the size of a tree

Advantages of CART:

- Provides a visual explanation for the prediction compared to other algorithms
- Can induce robust rules despite noisy data and complex relationships between high numbers of features

Applications of CART in investment management:

- Enhancing detection of fraud in financial statements
- Generating consistent decision processes in equity and fixed-income selection
- Simplifying communication of investment strategies to clients.

8. ENSEMBLE LEARNING AND RANDOM FOREST

Ensemble learning: The practice of combining many predictions from many models and averaging the result to achieve a reduction in noise as the average result converges to an accurate prediction.

Ensemble method: Combination of multiple learning methods

Ensemble learning is divided into the following categories:

- **Category 1**: An aggregation of heterogenous learners (different types of algorithms combined together with a voting classifier)
- **Category 2**: An aggregation of homogenous learners (a combination of the same algorithm using different training data and based on a certain technique such as bootstrap aggregating.

1 *Voting Classifiers*

Majority-voting classifiers will assign the predicted label with the greatest number of votes to a new data point.

The greater the individual models which are trained, the higher the accuracy of aggregated prediction up to a certain point. Beyond this point, performance may deteriorate from overfitting.

Assumption taken by the method: If model predictions are independent, the law of large numbers can be used to achieve a more accurate prediction.

2 Bootstrap Aggregating (Bagging)

The original training data set is used to generate *n* new training data sets or bags of data.

Bagging process:

- Each new bag of data is generated by random sampling with replacement from the initial training set.
- The algorithm can now be trained on *n* independent data sets that will generate *n* new models
- For each new observation, aggregate *n* predictions using a majority-vote classifier for a classification or an average for a regression.

Advantages of bagging include model stability and protection against overfitting the model.

3 Random Forest

Random forest: a collection of a large number of decision trees trained using the bagging method.

Example: A random forest classifier may comprise of a number of decision trees generated by a CART algorithm which, in turn, is trained using of the *n* independent data sets (from the bagging process)

A random reduction in features will help generate additional diversity in the trees and generate more individual predictions.

For any new observation, the random forest undertakes classification by majority vote.

Advantages of the random forest:

- Protects against overfitting on the training data
- Reduces the ratio of noise to signal because errors cancel out across the collection of different classification of trees.

Drawback of the random forest:

- Individual trees cannot be interpreted with relative ease (it is a black-box type algorithm)

Investment applications:

- Used in factor-based investment strategies for asset allocation and investment selection
- Used in predicting whether an IPO will be successful given attributes of IPO and corporate issuer

Practice: Example 3 and 4 from the CFA Institute's Curriculum.

Case Study: Refer to Curriculum for 'Classification of Winning and Losing Funds'.

Practice: Example 5 from the CFA Institute's Curriculum.

9. UNSUPERVISED MACHINE LEARNING ALGORITHMS

Unsupervised ML does not use labelled data (i.e., no target variable) but focuses on finding patterns within the data themselves. Unsupervised ML algorithms include:

- Dimension reduced based on principal components analysis
- Hierarchical clustering

9.1 Principal Component Analysis (PCA)

A widely used type of unsupervised ML which aims to represent a data set with many features (typically correlated) by a smaller set of features that do well in describing the data.

PCA is a statistical method for reducing highly correlated features into a few main, uncorrelated composite variables.

Composite variable: A variable that combines two or more variables that are statistically strongly related to each other

PCA involves two key concepts:

1. Eigenvectors: New, mutually uncorrelated composite variables that are linear combinations of the original features
2. Eigenvalue: Proportion of total variance in the initial data that is explained by each eigenvector

The PCA algorithm orders the eigenvectors from highest to lowest based on their eigenvalues or in terms of their usefulness in explaining total variance in the initial data.

Principal components:

First principal component: the eigenvector with the highest eigenvalue that is selected by PCA which explains the largest proportion of variation in the data set.

Second principal component explains the next largest proportion of variation in the data set remaining after the first principal component.

Principal components are linear combinations of the initial feature set.

Note: When deciding on the number of principal components to retain, there is a tradeoff between maintaining a lower-dimensional, more manageable view of a complex dataset when only a few are selected and the loss of information.

Scree plots: Show the contribution to total variation in the data set by each of the principal components. Practically, the minimum number of components which should be retained is that which the scree plot shows as explaining 85% to 95% of total variance in the data set.

Drawback of PCA: Principal components cannot be easily labelled or directly interpreted by the analyst since they are combinations of the data set's initial features.

Machine learning models are quicker to train, easier to interpret and reduce overfitting if provided with lower dimensional data sets.

10. CLUSTERING

A type of unsupervised ML which organizes data points into groups called clusters.

Cluster: Contains a subset of observations from the data set which are similar.

Good cluster: Observations within a cluster are coherent or similar or close to each other and the observations in two different clusters are very distinct from each other (known as separation).

Investment uses of clustering:

- For grouping companies based on financial statement items or financial ratios, for example
- Improving portfolio diversification

Popular clustering approaches include:

- K-means clustering
- Hierarchical clustering

10. K-Means Clustering

Involves partitioning the data into a fixed number, *k,* of non-overlapping clusters. *k,* the number of clusters, represents a model parameter, which must be established before learning begins.

Each cluster is characterized by its centroid (center) and observations are assigned to the cluster with the centroid to which the observations are closest.

Refer to: Exhibit 10 from the CFA Institute's Curriculum. for an illustration of how the *k*-means clustering groups observations

Note: The *k*-means algorithm seeks to minimize intra-cluster distance (thereby maximizing cohesion) and maximize inter-cluster distance (thereby maximizing separation).

Advantage of the technique: *k*-means algorithm is fast and works well with hundreds of millions of observations. Limitation of the technique: Hyperparameter, *k*, the number of clusters into which to partition the data must be decided before *k*-means is run. Analysts should overcome this problem by:

- Determining how many clusters are reasonable for the problem under investigation and data set being analyzed or
- Run the algorithm using a range of values for k to find the optimal number of clusters – the *k* that minimizes intra-cluster similarity (or cohesion) and maximizes inter-cluster distance (difference).

Investment uses of *k*-means algorithms:

- Used to classify investment vehicles or hedge funds
- Visualize data and detect trends or outliers
- Most used algorithm in investment management in:
  - data exploration for discovering patterns in high dimensional data
  - deriving alternatives to static industry classifications

11. HIERARCHICAL CLUSTERING: AGGLOMERATIVE AND DENDOGRAMS

Hierarchical clustering:

Unlike *k*-means clustering which segment the data into predetermined clusters with no relationship among the clusters, hierarchical clustering results in the creation of intermediate rounds of clusters of increasing (in agglomerative) or decreasing (in conglomerative) size until a final clustering is reached. The latter process creates relationships among clusters.
Agglomerative (or bottom-up) agglomerative clustering involves the following steps:

- Define observations as individual clusters
- Identify the two closest clusters defined by some measure of distance (similarity)

- Combine individual clusters into one large cluster

Divisive (or top-down) hierarchical clustering involves the following steps:

- Start with all the observations belonging to a single cluster
- Divide the observations into two clusters based on a measure of distance
- Divide each cluster into smaller clusters until each cluster contains 1 observation

Summary of differences between agglomerative and divisive hierarchical clustering:

Agglomerative Clustering:	Divisive Clustering:
A bottom-up approach	A top-down approach
Used within large data sets due to algorithm's fast computing speed.	
Well suited for identifying small clusters: Makes clustering decisions based on local patterns initially not considering the global structure of data.	Well suited for identifying large clusters: Makes clustering decisions based on a holistic representation of data with the algorithm designed to account for the global structure of data.

Dendrograms

Dendrograms highlight the hierarchical relationships among the clusters.

The role of clustering in portfolio diversification: Invest in assets from different clusters to diversify risks as clustering aims to maximize inter-cluster separation which will allow portfolio to reflect a wide variety of characteristics via investments in clusters.

Clustering is useful for important underlying structure in complex data sets.

Case Study: Refer to Curriculum for 'Clustering Stocks Based on Co-Movement Similarity'.

Practice: Example 6 & 7 From the CFA Institute's Curriculum..

13. NEURAL NETWORKS, DEEP LEARNING NETS, AND REINFORCEMENT LEARNING

Neural Networks

Neural networks (also known as artificial neural networks or ANNs) are highly flexible type of ML algorithms which have been successfully applied to tasks characterized by non-linearities and complex interactions among features. They are commonly used for:

- classification learning,
- regression supervised learning, and
- reinforcement (unsupervised) learning

Structure of a neural network

Basic structure includes:

- Layers:
  - Input layer – one node for each feature
  - Hidden layer – where learning occurs in training and inputs are processed on trained nets
  - Output layer – consists of a single node for the target variable *y*. This layer passes information to outside the network
- An additional feature of hidden layers is the transformation of inputs in a non-linear fashion into new values that are then combined into the target value.
- Each node has two functional parts:
  - A summation operator: Multiplies each input value received by nodes in input layer by a weight and sums weighted values to form total net input.
  - An activation function: Transforms total net input into the final output of the node.

Feature inputs in a neural network are typically scaled to account for differences in units of data.

Forward propagation: The process of transmitting output from one set of nodes to another, within or outside the layer, in a neural network.

Note:

- Learning in neural networks takes place through the process of adjusting weights to reduce total error.
- The structure of a neural network allows them to uncover approximate complex non-linear relationships among features

- As more nodes and more hidden layers are specified, a neural network's ability to handle complexity increases as does overfitting risk
- Research indicates that neural networks produce models of equity returns at the individual stock and portfolio level that are superior to models built using traditional statistical methods.
- Neural networks may be better able to cope with non-linear relationships inherent in security prices
- Drawback of using neural networks is their lack of interpretability and amount of data required to train models

14. NEURAL NETWORKS, DEEP LEARNING NETS, AND REINFORCEMENT LEARNING

Deep Learning Nets

Deep learning nets (DLNs) comprise of a minimum of 3 hidden layers but often more than 20 hidden layers.

DLNs are trained based on large data sets and during training the weights are designed to minimize a specified loss function.

DLNs require substantial time to train and systematically varying the hyperparameters until best out-of-sample performance is achieved may not be feasible.

Uses of DLNs:

- Pattern recognition problems
- Credit card fraud detection
- Vision and control problems in autonomous cars
- Natural language processing
- Pricing options
- Predicting corporate fundamental factors and price-related technical factors

Reinforcement Learning

Reinforcement learning (RL) involves an agent that should perform actions with the objective of maximizing its rewards over time taking environmental constraints into consideration.

Unlike supervised learning, RL has neither direct labeled data for each observation nor instantaneous feedback. Algorithm must observe the environment, learn by testing new actions, and reuse its previous experiences. Learning occurs through trial and error.

RL is being applied in markets where the algorithms acts as a virtual trader following certain rules to maximize profits. The success of RL in dealing with market complexities in questionable.

Practice: Example 8 From the CFA Institute's Curriculum..

Case Study: Refer to Curriculum for 'Deep Neural Network-Based Equity Factoral Model'.

Practice: CFA Institute's Curriculum End of Chapter Questions & FinQuiz-bank Item-sets & Questions)

1. INTRODUCTION AND BIG DATA IN INVESTMENT MANAGEMENT

Big data or alternative data encompasses data generated by financial markets, businesses and many other sources.

It is important for portfolio managers and investment analysts to understand how unstructured data can be transformed into structured data suitable as inputs to machine leaning (ML) methods.

BIG DATA IN INVESTMENT MANAGEMENT

Difference between big data and traditional data sources rests on 3 Vs:

1. **Volume** (or quantity of data)
2. **Variety** – array of available data source. Variety includes:

- Traditional transactional data
- User-generated text
- Images
- Videos
- Social media
- Sensor-based data and etc.

3. **Velocity** – the speed at which data are created. Unstructured data is generated at a fast pace. Such information also has implications for real-time predictive analytics in various financial applications.

When using big data for inference of prediction, Veracity relates to the credibility and reliability of different data sources. Veracity become important for big data because of the varied sources of the large datasets of big data.

2. STEPS IN EXECUTING A DATA ANALYSIS PROJECT: FINANCIAL FORECASTING WITH BIG DATA

The increasing use of big data or unstructured data has allowed for faster insights and has enhanced predictive power compared to traditional models which rely on standard data using statistical or mathematical models.

Textual big data provides valuable information including topics and sentiment. Incorporating big data for forecasting purposes and other investment applications requires supplementing traditional data with textual big data.

Traditional ML model building steps using structured data includes:

1. *Conceptualization of the modelling task*. Involves determining

- What the model output should be
- How the model is to be used and by whom
- How the model will be embedded in existing or new business processes

2. *Data collection.* Data used for financial forecasting is:

- numeric data derived from internal and external sources
- presented in a structured tabular format

3. *Data preparation and wrangling*. Includes:

- Cleansing the data: restoring missing values or out-of-range value and so forth
- Preprocessing the data: Involves extracting, aggregating, filtering, and selecting relevant data columns

4. *Data exploration*: Includes exploratory data analysis, feature selection and feature engineering

5. *Model training:* Involves selecting the appropriate ML method, evaluating performance of trained models, and model tuning.

In contrast to structured data, textual data arises from openly available data sources and is unstructured.

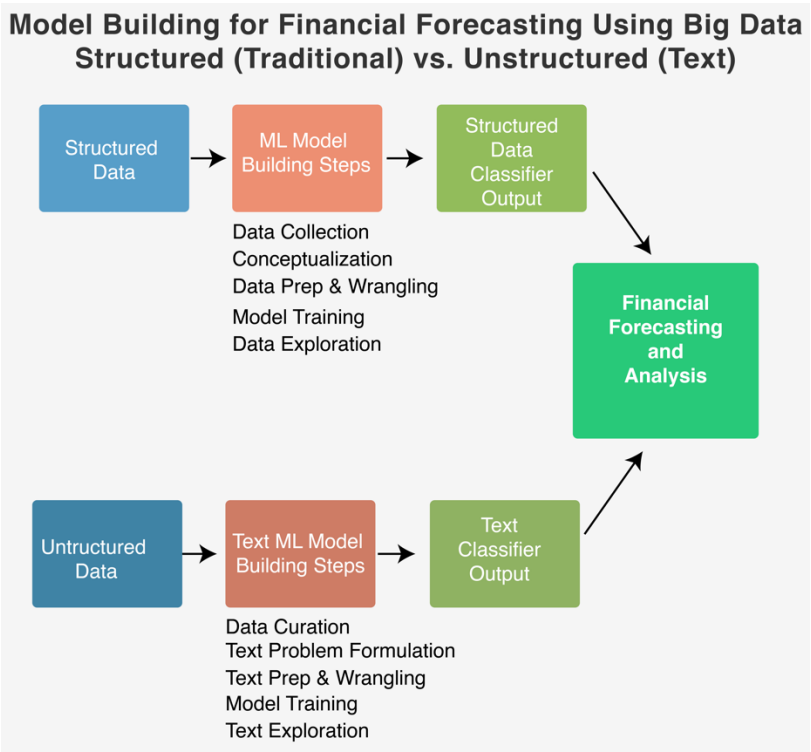
Text ML Model Building steps include four steps:

1. *Text problem formulation:* Identify:
   - exact inputs and outputs for the model
   - Utilization of model's classification output.
2. Data curation. Gather:
   - external data via web services or web spidering programs (involves extracting raw content from a source, typically web pages).
3. *Text preparation and wrangling:*
   - Convert unstructured data into a format usable by traditional modeling methods designed for structured inputs.
4. *Text exploration*: Encompasses:

- text visualization through techniques such as word clouds and
- text feature selection and engineering.

Practice: Example 1 Curriculum, Reading 5.

Output can be used directly for forecasting and/or analysis or combined with other structured variables.



3. DATA PREPARATION AND WRANGLING STRUCTURED DATA

Involves cleansing and organizing raw data so that it is suitable for further analysis and training a machine learning (ML) model. Quality of data affects training of selected ML model.

Data preparation is preceded by data collection which is preceded by:

- stating the problem,
- defining objectives,
- identifying useful data points and
- conceptualizing the model (drawing a modifiable plan that is necessary to initiate the model building process).

Data collection

- Searching for and downloading raw data from one or multiple sources.

- Databases are the most common primary source of internal data
- Data exists in the form of spreadsheets, comma separated values files, text files and other formats.
- Third party vendors can be sources of clean data and this external data can be obtained using:
  - Application programming interface (API) – a set of well-defined methods of communication between various software companies
  - csv files or other formats

Using external data vs. internal data:

External data	Internal data
Saves time & resources	Requires data preparation and wrangling which is time-

External data	Internal data
	consuming and resource intensive
Useful when a project requires generic data (example, demographics of a geographic area)	Useful when a project requires internal data (for example, to understand user traffic on a company website)
Information edge (or alpha) may be lost during the cleansing process	

Data Preparation and Wrangling: Steps of the data preparation and wrangling process vary by nature of data (structured vs. unstructured)

Data Preparation (Cleansing): Examines identifies, and mitigates errors in raw data. Normally, raw data is neither complete no clean to directly train the ML model.

Data Wrangling (Preprocessing): Transforms and processes cleansed data to make data ready for ML model training. Processing involves:

- dealing with outliers,
- extracting useful variables from existing data points, and
- scaling the data.

Structured Data

Data Preparation (Cleansing)

Structured data are organized in a systematic format that is readily searchable and readable by computer operations for processing and analyzing.

In the case of structured data, data cleansing process deals with identifying and mitigating errors such as:

- incomplete
- invalid
- inaccurate
- inconsistent
- non-uniform and
- duplicate data observations.

Refer to: Exhibit 3 from the Curriculum for an example of a raw dataset before cleansing.

Possible errors in a raw dataset include:

1. Incompleteness error: Missing or non-present data. This can be corrected by:
- investigating alternate data sources.
- Missing value and NAs (not applicable or not available values) must be omitted or replaced with NA for deletion or substitution with imputed values (mean, median, mode, or simply assuming zero) during the data exploration stage.

2. *Invalidity error*: Where the data are outside of a meaningful range, resulting in invalid data. This can be corrected by:
- Verifying other administrative data records

3. Inaccuracy error: is where data are not a measure of true value. This can be corrected using:
- business records and
- administrators

4. Inconsistency error: is where data conflicts with corresponding data points or reality. This can be corrected by:
- clarifying with another source

5. Non-uniformity error: is where data is not present in an identical format. This can be resolved by:
- converting data into a standard format

6. Duplication error: is where duplicate observations are present. This can be corrected by:
- removing duplicate entries.

Important points:

- Data cleansing is expensive and cumbersome because it combines automated, rule-based and pattern recognition tools with manual human inspection to check for error
- Data cleansing involves manually inspecting and verifying data and analyzing data using analysis software for investigating errors in data
- Quality of data cleansing depends on value of business project:
  - If data points with errors cannot be resolved due to lack of available resources, data points with errors can be omitted if a project has a large dataset.

Data Wrangling (Preprocessing)

Transformation processes for structured data include:

1. *Extraction*: A new variable can be extracted from the current variable for ease of analyzing and using for training the ML model. Example: Extracting age from date of birth.

2. *Aggregation:* Two or more variables can be aggregated into one variable to consolidate similar variables. Example: Combine salary and other income into a single variable, total income

3. ***Filtration:*** The data rows not needed for the project must be identified and filtered.

4. *Selection*: The data columns that are intuitively not needed for the project can be removed.

5. *Conversion*: The variables can of different types: nominal, ordinal, continuous, and categorial. Variables in the dataset must be converted into appropriate types to further process and analyze them correctly.

- Any values must be stripped out with prefixes and suffixes before conversion.

Note: How to identify and handle outliers?

Outliers can be detected using:

- Standard deviation: a data value outside 3 standard deviations from the mean may be an outlier.
- Interquartile range (IQR): difference between the 75th and 25th percentile data values.
  - If data values outside 1.5 IQR, they are outliers
  - If data values outside of 3IQR, they are extreme values.

Outliers can be handled using:

- Trimming: removing extreme values and outliers from the dataset
- Winsorization: when extreme values and outliers are replaced with the maximum (for large value outliers) and minimum (for small value outliers) values of data points that are not outliers.

Scaling

The process of adjusting the range of a feature by shifting and adjusting the scale of data. For ML model training, all variables should have values in the same range to make dataset homogenous. Outliers must be removed before scaling is undertaken. Two common ways of scaling include:

1. Normalization: The process of rescaling numerical variables in the range of [0,1]. A variable, *X*, is normalized using the following equation:

$$X_{i(normalized)} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Where X_i = value of observation

2. Standardization: The process of centering and scaling the variables.
Centering: subtracting the mean (μ) of the variable from each observation so that the new mean is 0.

Scaling: adjusts the range of data by dividing centered value ($X_i - \mu$) by the standard deviation (σ) of feature X. Resulting standardized variable will have an arithmetic mean of 0 and standard deviation of 1.

Normalization vs. Standardization:

- Normalization is more sensitive to outliers
- Standardization is less sensitive to outliers and more sensitive to mean and standard deviation of data.
- Treatment of outliers is necessary before normalization is performed.
- Standardization requires normally distributed data.

Practice: Example 2 from the CFA Institute Curriculum.

4. UNSTRUCTURED (TEXT) DATA

Text data can be processed by humans but not computers because they are not in a systematic format. Approximately 80% of available data is unstructured data.

For purposes of analysis and training ML models, unstructured data must be transformed into structured data.

Text processing: transforming structured data into structured data and includes two tasks: cleansing and preprocessing.

Text Preparation (Cleansing)

Cleansing is the initial step in text processing and involves cleaning text to remove non-useful elements from raw data.

Raw data are a sequence of characters and contains both useful and non-useful elements and thus requires cleaning prior to preprocessing.

Text cleansing process includes the following basic operations:

I. Remove html tags: Initial task is to remove the html tags that are not part of the actual text using programming functions or regular expressions. Some generic html tags may be kept to maintain certain formatting meaning in the text.

II. Remove punctuations: Involves removing useless punctuations and retaining useful punctuations by possibly relying on regex; the latter are substituted with annotations to preserve grammatical meaning in the text for further processing and analysis.

- **Note:** The context in which periods (dots) are used in the text must be understand and replaced or removed. For example: periods separating sentences should be replaced by the annotation/ end Sentence.

III. Remove numbers: Numbers in the text should be removed or substituted with annotation/number/ often relying on Regex. Such operations are critical for ML model training.

- **For example,** Information extraction is a typical application which may involve extracting monetary values from financial reports. Here, actual number values are critical and so any numbers and decimals must be retained.

IV. Remove white spaces: Extra white spaces, tab spaces, and leading and ending spaces should be removed to keep text intact and clean. Functions in programming languages can be used to remove unnecessary white spaces from text.

Refer to: Exhibit 8 from the Curriculum for an example of text cleansing process.

Note: The sequence and choice of cleaning operations does matter.

Text Wrangling (Preprocessing)

Relevant terms:

Token: equivalent to a word

Tokenization: process of splitting a given text into separate tokens
Text: a collection of tokens

Tokenization can be performed at word or character level but it is most commonly performed at word level.

Refer to: Exhibit 9 from the CFA Institute Curriculum. for an example of four cleansed texts and their word tokens.

Normalization process in text processing includes the following steps:

1. Lowercasing the alphabet involves removing distinctions among same words due to upper and lower cases.

2. Stop words such as 'the', 'is' and 'a' do not carry semantic meaning for the purposes of text analysis and ML training. For ML training purposes, stop words are removed to reduce number of tokens in the training set. Stop words are predefined in programming languages and help with this task.

3. Stemming: A rules-based approach for converting inflected forms of a word into its base word. Porter's algorithm is the most popular method for stemming. Results are not necessarily linguistically sensible.
- Example: stem of the word analyzed and analyzing is analyz.

4. Lemmatization: Process of converting inflected forms of a word into its morphological root (lemma). The approach is computationally more expensive and advanced.
- Example: Lemma of words analyzed and analyzing is analyze.

Note:

- Stemming and lemmatization reduce the repetition of words occurring in various forms and maintain semantic structure of the text data.
- Stemming is common in the English language as it is simpler to perform as it is compared to lemmatization.
- Both stemming and lemmatization decrease data sparseness by aggregating many sparsely occurring words in relatively less sparse stems or lemmas and train less complex ML models.

Normalization is followed by creation of bag-of-words (BOW), which is a procedure for analyzing text using a collection of a distinct set of tokens from all the texts in a simple dataset. BOW is memory efficient and easy to

handle for text analyses but does not capture the position or sequence of words present in the text.

Refer to: Exhibit 10, from the CFA Institute's Curriculum for a BOW representation and transformations occurring in normalization process.

Final stage of text preprocessing is using final BOW after normalization to build a document term matrix (DTM). DTM resembles a data table for structured data and is widely used for text data.

Structure of a DTM:
Rows = number of documents (or text files) in a sample dataset.
Columns = number of tokens from the BOW that is built using all the documents in a sample dataset.

Once DTM is created, unstructured text data are converted to structured data for further processing and training the ML model.

***Drawback of BOW*:** Does not represent word sequences or positions limiting its use for advanced ML training applications. Example: 'No' is treated as a single token but removed during the normalization because it is a stop word which fails to signify the negative meaning of the text.

***How to overcome drawback of BOW*:** Using n-grams, a representation of word sequences. Different n-grams (unigram, b-gram, or trigram, etc.) can be used to build a BOW.

Refer to: Exhibit 12 from the CFA Institute's Curriculum for a combined unigram-to-trigram BOW for the particular text.

Stemming can be applied on the cleansed text before building n-grams and BOW.

Note: Even after implementing N-grams and removing isolated stop words, stop words tend to persist when they are attached to their adjacent words.

Practice: Example 3 from the CFA Institute Curriculum.

5. DATA EXPLORATION OBJECTIVES AND METHODS AND STRUCTURED DATA

Data exploration stage follows the data preparation stage and leads to the model training stage. Domain knowledge plays a vital role in data exploration.

Data exploration involves three tasks:

I. **Exploratory data analysis (EDA) –** the preliminary step in data exploration. Data can be summarized using exploratory graphs, quantitative methods such as descriptive statistics and central tendency measures. Objectives of EDA include:

- Serve as a communication medium among project stakeholders
- Understanding data properties
- Finding patterns and relationships in data
- Inspecting basic questions and hypotheses
- Documenting data distributions and other characteristics, and
- Planning modeling strategies for next steps.

II. **Feature selection:** Selecting relevant features from the dataset for ML model training. Note: Fewer features reduces model complexity and training time.

III. **Feature engineering:** Process of creating new features by changing or transforming existing features.

Note: Model performance heavily depends on II and III.

Structured Data

Explanatory Data Analysis

EDA can be performed on either:

- one dimension with summary statistics (mean, median, quartiles, ranges, standard deviation, skewness and kurtosis) of features can be computed. Each feature of the dataset is summarized using the one-dimension visualization and relying on:
  - histograms,
  - bar charts,
  - box plots, and
  - density plots;
- two dimensions with a summary statistic of relationships such as correlation matrix, can be computed. Two- or more-dimensional visualization explores interactions/relationships

between two or more features in the dataset using scatterplots or line graphs.
  - Note: Scatterplots only provide a starting point for examining a relationship visually. Relationships should be tested further using statistical tests.

Multivariate data: Commonly utilized exploratory visualization designs include:

- stacked bar,
- line charts,
- multiple box plots (multi-box plot charts assess relationships between each features on the x-axis and the target variable of interest on the y-axis), and
- scatterplots

Refer to: Exhibit 15 from the CFA Institute Curriculum for an illustration of a multiple box plot.

Additional points:

- Descriptive statistics can be used in addition to visualization to summarize data such as:
  - Central tendency measures
  - Minimum and maximum values for continuous data
  - Counts and frequencies for categorial data
- EDA is also useful during feature selection and engineering stages
- Possible trends and relationships in data can be used to suggest new features which may improve model training.

Feature Selection

Data columns in table or matrix represent features of structured data.

Key points:

- Removes redundant, irrelevant, and unneeded features in a dataset on which EDA is performed
- Objective is to assist in identifying significant features which when used in a model retain the important patterns and complexities of the larger dataset while requiring fewer data overall.
- Statistical measures can be used to rank features based on importance and eliminate or retain features. Methods include:
  - Chi-squared test
  - Correlation coefficients
  - Information-gain measures

Dimensionality reduction versus feature selection:

Dimensionality reduction identifies features in the data that create greatest variance between observations and allows for the processing of a reduced data volume.

Both feature selection and dimensionality reduction seek to reduce the number of features in a dataset.

Dimensionality reduction creates new combinations of uncorrelated features while feature selection includes and excludes features present in the data without alteration.

Feature Engineering

Helps to optimize and further improve the features such that they can describe the structures inherent in the dataset. This process depends on the context of the project, domain of the data, and nature of the problem.

Feature engineering systemically alters, decomposes, or combines existing features to produce more meaningful features.

A. Engineering structured data:

Structured data are likely to contain quantities which can be engineered to better present relevant patterns in the dataset. An existing feature can be engineered into a new feature or decomposed into multiple features.

B. Engineering continuous data:

A new feature can be created by, for example:

- Taking the logarithm of the product of two or more features
- Using domain knowledge to decompose a feature into classifications of another feature such as decomposing salaries into different tax brackets

C. Engineering categorical data:

A new feature can be a combination of two features or a decomposition of one feature into many.

Example: The categorial feature, educational level, has five values – high school, bachelor's master's, and doctorate – is decomposed into five new features, one for each value (is_highSchool, is_doctorate) and filled with 0s and 1s.

One hot coding: process of converting categorical variables into binary form (0 or 1) for machine reading.

6. UNSTRUCTURED DATA: TEXT EXPLORATION

Exploratory Data Analysis

Application for text data analysis include:

- Text classification – relies on supervised ML to classify texts into different classes
- Fraud detection
- Topic modeling – relies on unsupervised ML approaches to group the texts in the dataset into topic clusters
- Sentiment analysis – predicts sentiment of the texts in a dataset using supervised and unsupervised approaches.

It is useful to perform EDA of text data by computing term frequency on the tokens.

Term frequency = $\frac{Number\ of\ times\ a\ given\ token\ occurs\ in\ all\ texts\ in\ dataset}{Total\ number\ of\ tokens\ in\ dataset}$

Text data: include a collection of texts (or corpus) that are sequences of tokens.
Text statistics: reveal patterns in co-occurrence of words

Text statistics are used in applications of text analytics. Examples:

A. Topic modelling - most informative words are identified by calculating TF of each word. Words with high TF are eliminated as they are likely to be stop words or common vocabulary words making the resulting BOW compact.

B. Sentiment analysis & text classification applications – chi-square measure of word association can be useful for understanding significant word appearances in negative and positive sentences in the text or documents.

Text statistics can be visually comprehended by:

- Bar charts – show words counts or frequency or
- Word clouds – show most commonly occurring words by varying font size and using color for additional dimensions

Refer to: Exhibit 17 from the CFA's Institute Curriculum for Word Cloud Example.

Feature Selection

Involves selecting a subset of terms or tokens in the dataset which serve as features for the ML model training.

Benefits of Feature Selection:

- Decreases the size of vocabulary or BOW making model efficient and less complex
- Eliminate noisy features which otherwise reduce model accuracy

Noisy features include:

- Most frequent tokens, e.g. stop words
- Most sparse (or rare) tokens

An example of ML model underfitting: Frequent tokens strain ML model to choose a decision boundary among the texts.

An example of ML model overfitting: Rare tokens mislead ML model into classifying texts containing the rare terms into a specific class.

Feature selection methods in text data include:

1. ***Frequency*** measures can remove noisy features by filtering tokens with high and low TF values across all texts.

For example, document frequency (DF) is a frequency measure which discards noise features not carrying any specific information about the text class and are present across all texts.

DF of a token = $\frac{Number\ of\ documents\ or\ texts\ containing\ token}{Total\ number\ of\ documents}$

Benefits:
- Simplest feature selection method
- Performs well when many thousands of tokens are present

2. ***Chi-square* test** – tests the independence of two events:
   1) Occurrence of token
   2) Occurrence of class

Test ranks tokens by usefulness to each class in text classification problems. Tokens with highest chi-square test statistic values occur more frequently in texts

associated with a particular class and can be selected for ML model training due to higher discriminatory potential.

3. ***Mutual information (MI)*** measures how much information is contributed by a token to a class of texts.

If MI value = 0, token's distribution in all text classes is the same

If MI = 1, token in a class occurs more frequently in that specific class

Feature Engineering

Goal of this process is to maintain semantic essence of the text while simplifying and converting it into structured data for ML. Techniques include:

1. ***Numbers:*** Different numbers of differing lengths of digits can be converted into different tokens. Example: Numbers with four digits may represent years and replaced with the token "/number4"

2. ***N-grams:*** Multi-word, discriminative patterns can be identified and their connection kept intact. Example: Compared to "market", "stock market" is used in a specific context and can distinguish general texts from finance-related texts. A bi-gram will be useful as it treats two adjacent words as a single token.

3. ***Name*** *entity recognition (NER):* The NER algorithm analyzes individual tokens and their surrounding semantics while referring to its dictionary to tag an object class to the token.

Refer to: Exhibit 19 Example from the CFA Institute's Curriculum.

Advantages of NER tags:

- Can be used as features for ML model training for better performance
- Can identify critical tokens on which lowercasing and stemming can be avoided
- Make features more discriminative

4 Part of speech (POS): Tag every token in the text with a corresponding part of speech. Example: Common POS tags are noun, verb, adjective, and proper noun.

POS tags can be used as features for ML model training and to identify tokens belonging to a POS tag.

Benefits of POS tags:

- Useful for separating verbs and nouns for text analytics and clarify meaning of the text
  - The word 'market' can represent a noun when used as "in the market" or as a verb when used as "to market"
  - The use of 'market' as a verb may indicate the topic relates to marketing and might discuss marketing a product or service

Practice: Example 4 and 5 from the CFA Institute's Curriculum.

7. MODEL TRAINING STRUCTURED AND UNSTRUCTURED DATA, METHOD SELECTION

It is crucial for ML engineers and domain experts to work together in building and training ML models. Model training is an iterative process

ML model training includes the three tasks which may be repeated several times until desired ML model performance is attained:

- Method selection – Decision of which ML model methods to employ is guided by considerations such as classification task, type of data and size of data.
- Performance evaluation – using complementary techniques to quantify and understand a model's performance.
- Tuning – process of undertaking decisions and actions to improve model performance

Refer to: Exhibit 20 for an illustration of model training stages.

Structured and Unstructured Data

ML model training process for structured and unstructured data is typically the same as unstructured data are processed and organized into a structured format in the data preparation stage.

ML model training involves fitting a system of rules on a training dataset to reveal a pattern in the data. Fitting is the degree to which a model can generalize to new data.

A good model fit performs well and can be validated using out-of-sample data.

Types of model fitting:

A. Overfit model may generate no errors with respect to the training data and has best accuracy, it fits training data too well and is unlikely to perform on future test cases

B. Underfit model does not fit the training data well and it produces misclassification errors.

C. A good fit model may fit the training data well but may not generalize well to out-of-sample data.

Model fitting errors can be caused by:

- Dataset size – small datasets may lead to underfitting as small datasets are not sufficient to expose patterns in the data
- **Number of features:** Smaller number of features can lead to underfitting as they may not carry all the characteristics that explain relationships between a target variable and the features. A large number of features can lead to overfitting as they can distort data due to low degrees of freedom. Appropriate feature selection using techniques as chi-square, mutual information minimizes overfitting.

Feature engineering prevents underfitting in model training as the technique can include new features which, when engineered properly, can elevate the underlying data points that better explain the interactions of features.

Method Selection

Represents the first step of the ML model training process and is governed by the following factors:

1 *Supervised or unsupervised learning*

Data for training and testing supervised ML models contain **ground truth** - the known outcome (target variable) of each observation in these datasets. Supervised models bring a structure that may or may not be supported by data.

Unsupervised ML modeling is challenging because of the absence of ground truth (i.e., no target variable). Unsupervised models bring no structure beyond which arises from the given data.

2 Type of data

For numerical data: classification and regression tree (CART) methods may be suitable.
For text data: Methods as generalized linear models (GLMs) and SVMs are commonly used.
For image data: Neural networks and deep learning methods tend to perform better than others.
For speech data: deep learning methods can offer promising results.

3 *Size of data*. A dataset has two basic characteristics: 1) Number of instances (i.e, observations) and 2) Number of features, which together determine the method suitable for model training. For example:

- SVMs work better with wider data sets (large number of features) and with fewer instances
- NNs work better on longer datasets where the number of instances is much larger than the number of features

Method-related decisions: need to be made after suitable model training method is selected and constitutes:

- Number of hidden layers in the neural network
- Number of trees in ensemble methods

Prior to model training, supervised learning breaks the master dataset intro three subsets using a random sampling technique:

I. Subset I: A training set used to train the model which should constitute 60% of master dataset
II. Subset II: A cross-validation set used to tune and validate the model which should constitute 60% of master dataset
III. Subset III: Test set for testing the model and uses the remaining data.

In the case of unsupervised learning, no splitting is undertaken due to the absence of labeled training data.

Class imbalance: occurs when the number of instances for a particular class is significantly larger than other classes posing a problem for data used in supervised learning. Balancing the training data can help alleviate the problem by either under-sampling the majority class or oversampling the minority class.

Performance Evaluation

Techniques for measuring model performance or goodness of fit for model validation include:

1 *Error analysis.* For classification problems, error analysis involves computing four basic evaluation metrics: True positive (TP), false positive (FP), true negative (TN), and false negative (FN) metrics.

Important terms:

FP: A type I error
FN: A type II error

Confusion matrix: a grid used to summarize values of these four metrics (Refer to Reading 5, Exhibit 23)
Precision: ratio of correctly predicted positive classes to all predicted positive classes.
Recall (or sensitivity): ratio of correctly predicted positive classes to all actual positive classes.

Note:

- Precision is useful when cost of FP or Type I error is high (example, expensive product is scrapped because it fails quality inspection but is actually perfectly good)
- Recall is useful in situations when cost of FN or Type II error is high (example, when an expensive product passes quality inspection and is sent to the customer but is actually defective)

8. PERFORMANCE EVALUATION

Performance metrics:

F1 score is more appropriate when there is unequal class distribution in the dataset

Accuracy: Percentage of correctly predicted classes out of total predictions

F1 score: harmonic mean of precision and recall.

Note: F1 score is more appropriate when unequal class distribution is in the dataset and it is necessary to measure the equilibrium of precision and recall. Accuracy is an appropriate performance measure when classes are equal in the dataset.

Formulas:

Accuracy = (TP + TN)/(TP + FP + TN + FN)
F1 score = (2*P*R)/(P + R)

High scores on both the measures suggest good model performance.

Receiver Operating Characteristic (ROC): Uses a plotted curve to show trade-off between the false positive rate (x-axis) and true positive rate (y-axis) for various cutoff points.

False positive rate (FPR) = FP/(TN + FP) and

True positive rate (TPR) = TP/(TP + FN), which is same as recall

Example of ROC application: In a logistic regression, if predicted probability (p) for a given observation is greater than the cutoff point, the observation is classified as class = 1, otherwise the observation will be classified as class = 0.

Indicators of model performance

- A more convex curve indicates better model performance
- Area under curve (AUC) close to 1.0 indicates near perfect prediction
- AUC of 0.5 indicates random guessing

3. *Root Mean Squared Error (RMSE)* is:

- Appropriate for continuous data prediction
- Mostly used for regression methods
- Captures all the prediction errors in the data
- Smaller RMSE indicates better model performance
- Formula: $\sqrt{\sum_{i=1}^{n} \frac{(Predicted_i - Actual_i)^2}{n}}$

9. TUNING

Tuning involves taking certain decisions and actions to improve the performance of the model.

Note:

- High prediction on the training set suggests model is underfitting
- If prediction error on the cross-validation set is significantly higher than on the training set, model is overfitting.

- Errors associated with model fitting:
  - Bias error: associated with underfitting and high when a model is overly simplified and does not sufficiently learn from the patterns in training data.
  - Variance error associated with overfitting and model memorizes the training data so much that it performs poorly on new data.

Optimum values of hyperparameters are obtained from:

- Tuning heuristics
- Grid search: Training the ML model using different combinations of hyperparameter values until the optimum set of values are found. Optimum values must result in similar performance of the model on training and CV datasets (i.e, training error and CV error are close). This will ensure that the model can be generalized to test data or to new data and is less likely to overfit.

The fitting curve:

- Plot of training errors for each value of a hyperparameter
- Provide visual insight on model's performance on training and CV datasets
- Helpful for tuning hyperparameters

Refer to: Exhibit 26. Curriculum for Fitting Curve for a regularization hyperparameter.

- With slight (large) regularization:
  - Model complexity is lightly (excessively) penalized
  - Most (Few) features are included in the model resulting in memorization of data (lack of complexity needed to determine underlying patterns in data)
  - Prediction error on the training dataset is small (large)
  - Prediction error on the CV dataset is larger (large)
  - May result in model overfitting (underfitting) with low (high) bias error and high variance error; thus model may perform (not perform) well on the training dataset but (and) generates many FP and FN errors on the CV dataset

Note:

- If high bias or variance error exists after tuning of hyperparameters, either a large number of training examples may be needed or number of features included in the model may need to be decreased (in case of high variance) or increased (in case of high bias). Model is then re-trained and re-tuned using the new training dataset.
- In case of a complex model, ceiling analysis can be performed which entails evaluating different components of model building to determine which can be improved by further model tuning.

10. FINANCIAL FORECASTING PROJECT: CLASSIFYING AND PREDICTING SENTIMENT FOR STOCKS

In the financial services space, robo-readers are being used to examine how views expressed in text relate to future company performance. Robo-readers analyze sentiment polarity – how positive, negative or neutral a phrase or statement is regarding a target. This sentiment news can be combined with structured financial data for forecasting purposes in ML-based models.

Refer to: Curriculum for Sections 10 to 13 for a practical application of how sentiment relating to companies listed in the NASDAW OMX Helsinki can be classified.

Important terms:

- Corpus: A collection of data in any form (list, matrix, or data table forms)
- Sentence length: the number of characters, including spaces, in a sentence.
- Frequency analysis on processed text data helps in filtering unnecessary tokens or features by quantifying how important tokens or features are in the sentience and corpus as a whole.
- Collection frequency/Term frequency at the corpus level: number of times a given word appears in the whole corpus/total number of words in the corpus.

Practice: Example 6 and 7 from the CFA Institute's Curriculum.

Practice: End of Chapter Questions From CFA Institute's Curriculum + FinQuiz Question bank (Item sets + Questions)

2. FOREIGN EXCHANGE MARKET CONCEPTS

Exchange rate: An exchange rate is the price of one currency (***base currency***) in terms of another currency (***price currency***) i.e. the number of units of price currency required to purchase one unit of base currency, quoted as P/B.

- For example, A/ B refers to the number of units of 'Currency A' that can be bought by one unit of 'Currency B'.
- Currency B is the *Base currency* and Currency A is the *Price currency*.
- Typically, exchange rates are quoted to four decimal places; except for yen, for which exchange rate is quoted to two decimal places.

Example:

Suppose, USD/ EUR exchange rate of 1.456 means that 1 euro will buy 1.456 U.S. dollars

- Euro is the base currency
- U.S. dollar is the price currency.

If this exchange rate decreases, then it would mean that fewer U.S. dollars will be needed to buy one euro. It implies that:

- U.S. dollar appreciates against the Euro or
- Euro depreciates against the U.S. dollar.

Spot Exchange-rate: The exchange rate used for ***spot transactions*** i.e. the exchange of currencies settled in two business days after the trade date, is referred to as "T+2 settlement".

Note: For Canadian dollar, spot settlement against the U.S. dollar is on a T + 1 basis.

Two-sided Price: Two-sided Price refers to the buying and selling price of a ***base*** currency quoted by a dealer.

- **Bid Price:** The price at which the dealer is willing to ***buy the Base currency*** i.e. number of units of ***price currency*** that the client will receive by selling 1 unit of base currency to a dealer.
- **Ask or Offer Price:** The price at which the dealer is willing to ***sell the Base currency*** i.e. number of units of ***price currency*** that the client must sell to the dealer to buy 1 unit of base currency.

Bid-offer Spread = Offer price – Bid price

- Bid-offer Spread is the compensation of the counterparty for providing foreign exchange to other market participants.
- The lower the buying rate (bid price) and the higher the selling rate (offer price) ➔ the wider the bid-ask spread, ➔ the higher the profit for a dealer.
- The size of the bid-offer spread (in pips) can vary *widely* across exchange rates and over time.

Important Points:

1) Bid Price is ***always lower*** than Offer Price.
2) The counterparty in the transaction will have the *option (but not the obligation)* to deal at either the bid price (to sell the base currency) or offer price (to buy the base currency) quoted to them by the dealer.

- When counterparty deals at bid price, it is referred to as "hit the bid".
- When counterparty deals at offer price, it is referred to as "paid the offer".

Example:

Suppose, USD/SFr exchange rate = 0.3968/0.3978 ➔ Dealer is willing to pay USD 0.3968 to buy 1 SFr and that the dealer will sell 1 SFr for USD 0.3978.

Interbank Market: It is the market where the dealers (or professional market participants) engage in foreign exchange transactions among themselves. It involves dealing sizes of at least 1 million units of the base currency and trades are measured in terms of multiples of a million units of the base currency.

- The bid-offer spread that dealers receive from the interbank market is generally narrower than the bid-offer spread that they provide to their clients.
- The interbank market facilitates dealers to:
  A. Adjust their inventories and risk positions;
  B. Distribute foreign exchange currencies to end users;
  C. Transfer foreign exchange rate risk to market participants who are willing to assume that risk;
- When the dealer buys (sells) the base currency from (to) a client, the dealer typically enters into an offsetting transaction and sells (buys) the base currency in the interbank market.

Factors that affect the size of the bid/offer spread:

1) **Interbank market liquidity of the underlying currency pair:** The bid-offer spread in the interbank foreign exchange market depends on the liquidity in the interbank market i.e. the greater the liquidity, the narrower the bid-ask spread.

Liquidity in the interbank market depends on the following factors:

a) The currency pair involved: Some currency pairs e.g. USD/EUR, JPY/USD, or USD/GBP have greater liquidity due to greater market participation and as a result narrower bid-offer spreads.

b) The time day: Although FX markets are open 24 hours a day on business days, the interbank FX markets have the greatest liquidity when the major FX trading centers are open. The liquidity in the interbank markets can be quite thin between the time New York closes and the time Asia opens.

c) Market volatility: The more volatile the market is → the more uncertain market participants are about the factors* that influence market pricing → the lower the liquidity and consequently, the wider the bid-offer spreads in both the interbank and broader markets.

*These factors include geopolitical events (e.g. war, civil strife), market crashes, and major data releases (e.g. U.S. nonfarm payrolls).

i. **The size of the transaction:** Generally, the larger the size of the transaction, the more difficult it is for the dealer to lay off foreign exchange risk of the position in the interbank FX market, and as a result, the wider the bid-offer spread. In addition, retail transactions (i.e. dealing sizes of <1 million units of the base currency) tend to have relatively wider bid-offer spreads than that of interbank market.

ii. **The relationship between the dealer and the client:** The dealer may provide a tighter (smaller) bid-offer spot exchange rate quote

- In order to win the client's business for other services besides spot foreign exchange business e.g. transactions in bond and/or equity securities.
- In order to win repeat FX business.
- To a client with a good credit profile.

However, due to short settlement cycle for spot FX transactions, credit risk is not considered as an important factor in determining the client's bid-offer spread on spot exchange rates.

3. ARBITRAGE CONSTRAINTS ON SPOT EXCHANGE RATE QUOTES

The two arbitrage constraints on spot exchange rate quotes are as follows:

1) *The bid quoted by a dealer in the interbank market must be lower than the current interbank offer; and the offer quoted by a dealer must be higher than the current interbank bid*; otherwise, arbitrage opportunities exit.

Example:

Suppose the current spot USD/EUR price in the interbank market is 1.3548/1.3550. But a dealer quoted a price of 1.3551/1.3553. Thus, other market participants would pay the offer in the interbank market i.e. buying EUR at a price of USD 1.3550 and then sell the EUR to the dealer by hitting the dealer's bid at USD 1.3551; hence, making a riskless profit = one pip.

2) *The cross-rate bids (offers) quoted by a dealer must be lower (higher) than the implied cross-rate offers (bids) quoted in the interbank market.*

Cross-Rate Calculations:

Suppose,

Exchange rate for CAD/USD = 1.0460
Exchange rate for USD/EUR = 1.2880

The exchange rate for CAD/EUR is determined as follows:

$$\frac{\text{CAD}}{\cancel{\text{USD}}} \times \frac{\cancel{\text{USD}}}{\text{EUR}} = \frac{\text{CAD}}{\text{EUR}}$$

$$1.0460 \times 1.2880 = 1.3472 \text{ CAD/EUR}$$

Now Suppose,

Exchange rate for CAD/USD = 1.0460
Exchange rate for JPY/USD = 85.50

The exchange rate for JPY/CAD is determined as follows:

$$\frac{\text{CAD}}{\text{USD}} \times \frac{\text{JPY}}{\text{USD}} = \frac{1}{\frac{CAD}{USD}} \times \frac{\text{JPY}}{\text{USD}} = \frac{\cancel{\text{USD}}}{\text{CAD}} \times \frac{\text{JPY}}{\cancel{\text{USD}}} = \frac{\text{JPY}}{\text{CAD}}$$

$$(1/\ 1.0460) \times 85.50 = 81.74 \text{ JPY/CAD}$$

Triangular arbitrage:

The cross-rate quotes ***must be consistent*** with the components' underlying exchange rate quotes. If they are not consistent, then arbitrage opportunities exist.

Suppose, a misguided dealer quotes JPY / CAD rate of 82.00. Hence, profit can be earned by:

- Buying CAD1 at the lower price of JPY81.74.
- Selling CAD1 at JPY82.00.

A riskless arbitrage profit that can be earned by a trader = JPY0.26 per CAD1.

This arbitrage is known as triangular arbitrage because it involves three currencies.

NOTE:

Bid Rate (A per B) = 1 / Ask Rate (B per A)

E.g. Bid Rate (CAD per USD) = 1 / Ask Rate (USD per CAD)
Ask Rate (A per B) = 1 / Bid Rate (B per A)
E.g. Ask Rate (CAD per USD) = 1 / Bid Rate (USD per CAD)

Practice: Example 1 from the CFA Institute's Curriculum.

4. FORWARD MARKETS

Currency forward contracts represent an obligation to buy or sell a certain amount of a specified currency at a *future date* at an exchange rate determined today. Unlike spot transactions, forward contracts involve settlement period longer than the usual "T + 2" settlement for spot delivery.

- The exchange rate used for forwards transactions is called the ***forward exchange rate***.

Example:

Suppose, today is 16 November.

- Spot settlement is for 18 November.
- Three-month forward settlement would be 18 February of the following year.

For details, refer to section 3.1.1 below

Points on a forward rate quote: Typically, forward exchange rates are quoted in terms of points (called pips).

Points on a forward rate quote = Forward exchange rate quote – Spot exchange rate quote

Note:

- In the bid-offer quote, the bid will always be smaller than the offer, even when the forward points are negative.
- **Positive Forward Points:** forward rate > spot rate, indicating that the base currency is trading at a forward premium and price currency is trading at a forward discount.
- **Negative Forward Points:** forward rate < spot rate, indicating that the base currency is trading at a forward discount and price currency is trading at a forward premium.
- The absolute number of forward points is positively related to the term of the forward contract i.e. *the longer the term, the greater the absolute number of forward points*.
-

Example:

Spot exchange rate USD/ EUR =1.2875
One year forward rate USD/ EUR = 1.28485
One year forward point = 1.28485 – 1.2875 = –0.00265

- It is scaled up by four decimal places by multiplying it by 10,000 i.e., -0.00265 × 10,000 = -26.5 points.

Converting forward points into forward quotes:

To convert the forward points into forward rate quote, forward points are scaled down to the fourth decimal place in the following manner:

$$\textbf{\textit{Forward rate}} = \textit{Spot exchange rate} + \frac{\text{Fowrad points}}{10{,}000}$$

$$\textbf{\textit{Forward premium/discount}}\ (\textbf{\textit{in}}\ \%) = \frac{\text{spot exchange rate} - (\text{forward points}/10{,}000)}{\text{spot exchange rate}} - 1$$

- When a market participant is selling (buying) base currency ➔ he/she would use bid (offer) rates for both the spot and the forward points, implying that the market participant will hit the bid (pay the offer).
- It is important to note that quoted points are not annualized because they are already scaled to each maturity.

The spot rate can be converted into a forward quote when points are represented as % as follows:

Spot exchange rate × (1 + % premium)
Spot exchange rate × (1 - % discount)

NOTE:

When exchange rate is quoted to only two decimal places, forward points are divided by 100.

FX swap: FX swap is a combination of an offsetting spot transaction and a new forward contract in the same base currency i.e. the base currency is purchased (sold) spot and sold (purchased) forward. Generally, the mid-market spot exchange rate is used for the swap transaction.

Uses: FX swaps can be used:

- for funding purposes (called swap funding).
- to roll over a forward position into future either for hedging or speculation purposes.
- to eliminate foreign exchange risk.

Example:

Suppose a German based company needs to borrow EUR100 million for 90 days (starting 2 days from today). It can be done in two ways:

i. Borrow EUR100 million starting at T + 2.
ii. Borrow in U.S. dollars and exchange them for Euros in the spot FX market (both with T + 2 settlement) and then sell Euros 90 days forward against the U.S. dollar.

Factors that affect the bid-offer spread for Forward Points:

1. **Interbank market liquidity of the underlying currency pair:** The greater the liquidity, the narrower the bid-ask spread.
2. **Size of the transaction:** The larger the trade size, the lower the liquidity of a forward contract and thus, the wider the bid-ask spread.
3. **Relationship between the client and the dealer** (as explained above).
4. **Term of the forward contract:** The longer the term of the forward contract, the wider the bid-offer spread.

Practice: Example 2 from the CFA Institute's Curriculum.

5. THE MARK-TO-MARKET VALUE OF A FORWARD CONTRACT

Mark-to-market value of Forward Contracts:

The mark-to-market value of forward contracts represent the profit (or loss) that would be realized when the forward position is closed out at current market prices.

- At contract initiation, mark-to-market value of the contract is zero i.e. the forward rate is set such that no cash changes hands at initiation.
- Afterwards, the mark-to-market value of the forward contract changes as the spot exchange rate changes and as interest rates change in either of the two currencies.

Example:

Suppose, an investor originally bought GBP 10 million at an AUD/GBP rate of 1.600 and subsequently sold them at a rate of 1.6200. The 3-month discount rate is 4.80% (annualized).

- Long GBP 10 million at 1.6100 AUD/GBP≡ Short AUD 16,100,000 (10,000,000 × 1.6100) at the same forward rate.

At settlement date:

The net GBP amounts = 0 ➔ i.e. GBP 10 million both bought and sold.

- AUD cash flow = (1.6340 – 1.6100) × 10,000,000 = +AUD 240,000 ➔ cash inflow because the GBP subsequently appreciated (i.e. AUD/GBP rate increased).
- It is important to note that this cash flow will be paid at a settlement date. Thus, PV of the cash inflow is calculated as:

PV of future AUD cash flow = $\frac{AUD\ 240{,}000}{1+0.048\left[\frac{90}{360}\right]}$ = AUD 237,154

The longer the term of the forward contract, the

- lower the liquidity in the forward market.
- greater the exposure to counterparty credit risk.
- higher the price sensitivity to movements in interest rates i.e. the greater the interest rate risk of the forward contract.

Practice: Example 3 from the CFA Institute's Curriculum.

6. INTERNATIONAL PARITY CONDITIONS

1) **Long run versus short run:** Long-term equilibrium values may act as an anchor for exchange rate movement. In the short run, no evident relationship exists between exchange rate movements and economic fundamentals.

- It is important to note that there is no simple formula, model, or approach that can be used to precisely forecast exchange rates.

2) **Expected versus unexpected changes:**

- In an efficient market, prices reflect both market participants' expectations and risk premium (i.e. compensation demanded by investors for exposures to unpredictable outcomes).
- Risk premia primarily depend on confidence and reputation and can change quickly in response to

large, unexpected movements in a variable, leading to immediate, discrete price adjustments.
- In contrast, expectations of long-run equilibrium values tend to change slowly.

3) Relative movements: For determining exchange rates, the *differences* in key factors across countries are more important than the levels or variability of key factors in any particular country.

International Parity Conditions

Parity conditions show relationship between expected inflation differentials, interest rate differentials, forward exchange rates, current spot exchange rates, and expected future spot exchange rates. The key international parity conditions are as follows:

1) Covered interest rate parity
2) Uncovered interest rate parity
3) Forward rates parity
4) Purchasing power parity
5) The international Fisher effect

Assumptions of Parity Conditions:

- Perfect information is available to all market participants.
- Risk neutrality
- Freely adjustable market prices

Implication of Parity Conditions: If parity conditions are held at all times, it implies that no arbitrage opportunities exist i.e. investors cannot exploit profitable trading opportunities.

NOTE:

Parity Conditions are expected to hold in the long-run, but not always in the short term.

7. COVERED INTEREST RATE PARITY, UNCOVERED INTEREST RATE PARITY, & FORWARD RATE PARITY

Covered Interest Rate Parity

According to covered interest rate parity,

The expected return earned on a *fully currency-hedged* foreign money market instrument investment should be **equal** to the return earned an otherwise identical domestic money market investment.

- Covered interest rate parity must always hold because it is enforced by arbitrage.

Assumptions:

- There are zero transaction costs.
- The underlying domestic and foreign money market instruments are identical in terms of liquidity, maturity, and default risk.
- Flow of capital is not restricted.

Explanation:

An investor has two alternatives available i.e.

a) Invest for one period at the domestic risk-free rate i.e. i_d;
- This amount will grow to $(1 + i_d)$ at the end of the investment horizon.

b) Convert 1 unit of domestic currency into foreign currency using the spot rate = $S_{f/d}$. (direct quote)
- Invest this amount for one period at foreign risk-free rate (i.e. i_f) e.g. in the bank deposits.
- The amount invested will grow to $S_{f/d}(1 + i_f)$ at the end of the investment horizon.
- Then, convert this amount to domestic currency using the forward rate i.e. for each unit of foreign currency, investor would obtain $1/F_{f/d}$ units of domestic currency.
- By converting the foreign currency at the forward rate, the investor has eliminated FX risk.

NOTE:

- Both of these alternatives are risk-free and have same risk characteristics.
- The arbitrage relationship holds for any investment horizon.

Covered interest rate parity is stated as follows:

$$(1 + i_d) = S_{\int/d}(1 + i_{\int})\left(\frac{1}{F_{\int/d}}\right)$$

$$F_{\int/d} = S_{\int/d}\left(\frac{1 + i_{\int}}{(1 + i_d)}\right)$$

Forward rate as a % of spot rate can be stated as follows:

$$\frac{F_{\int/d}}{S_{\int/d}} = \left(\frac{1 + i_{\int}}{(1 + i_d)}\right)$$

Using day count convention:

$$\left(1 + i_d\left[\frac{Actual}{360}\right]\right) = S_{f/d}\left(1 + i_f\left[\frac{Actual}{360}\right]\right)\left(\frac{1}{F_{f/d}}\right)$$

$$F_{f/d} = S_{f/d}\left(\frac{1+i_f\left[\frac{Actual}{360}\right]}{1+i_d\left[\frac{Actual}{360}\right]}\right)$$

- The above equation implies that covered (currency-hedged) interest rate differential between the two markets is zero.
- Thus, covered interest rate parity implies that the forward exchange rate ***must be*** the rate at which the holding period returns on these two alternative investment strategies will be exactly the same. Otherwise, investors can sell short lower return investment and invest in higher return investment.

For example,

a) If $(1+ i_d) > [S_{f/d}(1+i_f)(1/F_{f/d})]$

1) Borrow in Foreign currency
2) Buy domestic currency in spot market with foreign currency
3) Lend the domestic currency i.e. invest it at i_d.
4) Sell the domestic currency forward (buy currency of original loan forward i.e. foreign currency)

- The demand for domestic currency-denominated securities causes domestic interest rates to fall, while the higher level of borrowing in foreign currency causes foreign interest rates to rise.

b) If $(1+ i_d) < [S_{f/d}(1+i_f)(1/F_{f/d})]$

1) Borrow in domestic currency
2) Buy foreign currency in spot market with domestic currency
3) Lend the foreign currency i.e. invest it at i_f.
4) Sell the foreign currency forward (buy currency of original loan forward i.e. domestic currency).

Implications of Covered Interest Rate Parity: The forward premium should be approximately equal to the difference in interest rates, implying that any interest rate differential between countries should be offset exactly by the forward premium or discount on its exchange rate.

- When covered interest rate parity holds, the forward exchange rate will be an unbiased forecast of the future spot exchange rate.

Uncovered Interest Rate Parity

According to the uncovered interest rate parity condition,

The expected return on an <u>*uncovered (i.e. unhedged)*</u> foreign currency investment should be ***equal*** to the return on a comparable domestic currency investment.

For a domestic investor, the return on a risk-free domestic money market instrument is known with certainty; however, the domestic investor is exposed to FX risk with regard to an unhedged foreign currency investment.

Uncovered interest rate parity is stated as follows:

$$i_f - \%\Delta S^e{}_{f/d} = i_d$$
$$\%\Delta S^e{}_{f/d} = i_f - i_d$$

where,

- $\%\Delta S^e{}_{f/d}$ **=** Expected change in the foreign currency price of the domestic currency over the investment horizon.
- An increase in $S^e{}_{f/d}$ indicates that the foreign currency is expected to depreciate ➔ resulting in reduction in return for an investor.
- According to uncovered interest rate parity, when return on both unhedged foreign currency investment and domestic investment is equal, investors will be indifferent between both the alternatives, reflecting that investors are ***risk neutral***.
- According to this equation, the change in spot rate over the investment horizon should be, on average, equal to the differential in interest rates between the two countries.
  - E.g., if $i_f - i_d = 5\%$, it indicates that domestic currency is expected to appreciate against the foreign currency by 5% ➔ $\%\Delta S^e{}_{f/d} = 5\%$.
- This implies that on average, the expected appreciation/depreciation of the exchange rate is an unbiased predictor of the future spot rate.

When uncovered interest rate parity holds:

- The currency of a country with the higher (lower) interest rate or money market yield is expected to depreciate (appreciate) such that the higher return offered by the high-yield currency is exactly offset by the depreciation of the high-yield currency.
- The forward exchange rate will be an unbiased forecast of the future spot exchange rate.
- The current exchange rate will NOT be the best predictor unless the interest rate differential is equal to zero.

NOTE:

- Uncovered interest rate parity tends to hold over very long-term periods. It does not hold over short- and medium-term periods. Thus, over short- and medium-term periods, interest rate differentials are poor predictor of future exchange rate changes.

- Unlike covered interest rate parity, uncovered interest rate parity is NOT enforced by arbitrage.

Example:

Suppose, i_f = 10%, i_d = 5%. Consider three cases:

i. The $S_{f/d}$ rate is expected to remain unchanged:

Return on foreign-currency-denominated money market investment = 10% – 0%
= 10%

- Since it is > i_d, investor would prefer the foreign-currency-denominated money market investment.

ii. The domestic currency is expected to appreciate by 10%.

Return on foreign-currency-denominated money market investment = 10% – 10%
= 0%

- Since it is < i_d, investor would prefer the domestic investment.

iii. The domestic currency is expected to appreciate by 5%.

Return on foreign-currency-denominated money market investment = 10% – 5%
= 5%

- Since it is = i_d, the uncovered interest rate parity holds.

Forward Rate Parity

According to forward rate parity, forward rates are unbiased predictor of future exchange rates. Forward rates may not be a perfect forecast therefore, they may overestimate or underestimate the future spot rates but on average they are equal to the future spot rates. Two other parity conditions important for building forward rate parity are:

1. Covered interest rate parity
2. Uncovered interest rate parity

The **forward premium or discount** is calculated as follows:

For one year horizon,

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{i_f - i_d}{1+i_d}\right) \cong S_{f/d}(i_f - i_d)$$

Using day count convention:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{\left[\frac{Actual}{360}\right]}{1+i_d\left[\frac{Actual}{360}\right]}\right)(i_f - i_d)$$

where,

f = foreign or price currency
d = domestic or base currency

- The domestic currency will trade at a forward premium (i.e. $F_{f/d} > S_{f/d}$), if and only if, the foreign risk-free interest rate > domestic risk-free interest rate (i.e. $i_f > i_d$). In other words,
  - *Currency with the higher interest rate will always trade at a discount in the forward market.*
  - *Currency with the lower interest rate will always trade at a premium in the forward market.*
- The forward premium or discount is ***proportional*** to the spot exchange rate ($S_{f/d}$), interest rate differential ($i_f - i_d$) between the markets, and *approximately proportional* to the time to maturity (actual/360).

Forward discount or premium as % of spot rate:

$$\frac{F_{f/d} - S_{f/d}}{S_{f/d}} \cong (i_f - i_d)$$

If uncovered interest rate parity holds,
Forward premium or discount

$$= \frac{F_{f/d} - S_{f/d}}{S_{f/d}} = \%\Delta S^e{}_{f/d} \cong (i_f - i_d)$$

- It follows that the forward exchange rate = Expected future spot exchange rate ➔ $F_{f/d} = S^e{}_{f/d}$
- *Thus, when **both covered and uncovered interest rate parity hold**, the forward exchange rate will be an unbiased forecast of the future spot exchange rate.*

If the forward rate > (<) speculator's expected future spot rate ➔ risk-neutral speculators will buy the domestic currency in the spot (forward) market and simultaneously sell it in the forward (spot) market ➔ generating a profit if expectations are correct.

- Unfortunately, despite being unbiased, forward exchange rates are ***poor predictors*** of future spot exchange rate due to the high volatility in exchange rate movements.
- When it is assumed that exchange rate movements follow a random walk (i.e. $E_t[S_{t+1}] = S_t$), then the current spot exchange rate will be the best predictor of future spot rates.

Under Uncovered interest rate parity:

- When $i_d < i_f$ ➔ domestic (foreign) currency must trade at a forward premium (discount) ➔

expected appreciation of the home/domestic currency.
- When $i_d > i_f$ ➔ domestic (foreign) currency must trade at a forward discount (premium) ➔ expected depreciation of the home/domestic currency.

Under Covered interest rate parity:

- When $i_d < i_f$ ➔ domestic (foreign) currency must trade at a forward premium (discount).

- When $i_d > i_f$ ➔ domestic (foreign) currency must trade at a forward discount (premium).

Practice: Example 4 from the CFA Institute's Curriculum.

8. PURCHASING POWER PARITY

PPP is based on ***law of one price,*** which states that in competitive markets (free of transportation costs and official barriers to trade), identical goods sold in different countries ***must*** sell for the same price when their prices are measured in the same currency.

Hence, according to PPP, the nominal exchange rates would adjust for inflation so that identical goods (or baskets of goods) will have the identical price in different markets i.e. foreign price of a good X should be equal to the exchange rate-adjusted price of the identical good in the domestic country.

$$P^x_f = S_{f/d} \times P^x_d$$

Absolute version of PPP: According to absolute version of PPP, foreign price of a basket of goods and services should be equal to the exchange rate-adjusted price of the identical basket of goods and services in the domestic country.

$$P_f = S_{f/d} \times P_d$$

Nominal exchange rate = Foreign broad price index / Domestic broad price index
i.e.

$$S_{f/d} = P_f / P_d$$

Assumptions of Absolute version of PPP:

- All domestic and foreign goods are freely tradable internationally i.e. transaction costs are zero.
- Identical bundle of goods and services are consumed with equal proportions (same weights) across different countries.

When the above assumptions do not hold, absolute PPP probably doesn't hold precisely in the real world. However, if assumptions do not hold, but transaction costs and other trade impediments are assumed to be ***constant*** over time, then *changes* in exchange rates may be equal to the *changes* in national price levels.

Relative version of PPP: It focuses on ***actual*** changes in exchange rates caused by ***actual*** differences in national inflation rates in a given time period. According to relative version of PPP,

% change in the spot exchange rate = Foreign inflation rate – Domestic inflation rate

$$\%\Delta S_{f/d} = \pi_f - \pi_d$$

- The relative version of PPP implies that the currency of the high-inflation country should depreciate relative to the currency of the low-inflation country.
- If domestic price level rises by 10%, then domestic currency will fall by 10%.

Example:

Suppose, foreign inflation rate is 10% while the domestic inflation rate is 5%, then the $S_{f/d}$ exchange rate must rise by 5% in order to maintain the relative competitiveness of the two regions.

Ex ante version of PPP: The ex-ante version of PPP focuses on ***expected*** changes in the spot exchange rate caused entirely by ***expected*** differences in national inflation rates. According to ex-ante PPP, *currency of a country that is expected to have persistently high (low) inflation rates tends to depreciate (appreciate) over time*. Ex ante PPP can be expressed as:

$$\%\Delta S^e_{f/d} = \pi^e_f - \pi^e_d$$

where,

$\%\Delta S^e_{f/d}$ = Expected % change in the spot exchange rate
π^e_f = Foreign inflation rates expected to prevail over the same period
π^e_d = Domestic inflation rates expected to prevail over the same period

PPP does not hold when:

- There are different baskets of goods for price indexes.
- Goods and services are non-tradable.
- There are barriers to trade.
- There are transportation costs.
- Adjustment involves longer time.

- Over longer time horizons, nominal exchange rates tend to move towards their long-run PPP equilibrium values. Thus, PPP can be used as a long-run benchmark exchange rate and to make meaningful international comparisons of economic data.

Important to Note:

- If the currency is overvalued on a PPP basis ➔ it should depreciate.
- If the currency is undervalued on a PPP basis ➔ it should appreciate.

9. THE FISHER EFFECT, REAL INTEREST RATE PARITY AND TYING THE INTERNATIONAL PARITY CONDITIONS TOGETHER

The Fisher Effect and Real Interest Rate Parity

When the Fisher effect holds, Nominal interest rate in a country = Real interest rate + Expected Inflation rate i.e.
$i_d = r_d + \pi^e_d$
$i_f = r_f + \pi^e_f$

Foreign-domestic nominal yield spread= Foreign-domestic real yield spread + Foreign-domestic expected inflation differential

$i_f - i_d = (r_f - r_d) + (\pi^e_f - \pi^e_d)$
$(r_f - r_d) = (i_f - i_d) - (\pi^e_f - \pi^e_d)$

Important to Note:

- If uncovered interest rate parity holds, then the nominal interest rate spread = expected change in the exchange rate.
- If ex-ante PPP holds, then the difference in expected inflation rates = expected change in the exchange rate.

When both uncovered interest rate parity and ex-ante PPP hold, real yield spread between the domestic and foreign countries will be zero regardless of expected changes in the spot exchange rate i.e.

$$(r_f - r_d) = \%\Delta S^e_{f/d} - \%\Delta S^e_{f/d} = 0$$

Reflecting that,

Nominal interest rate differential between two countries = Difference between the expected inflation rates

$$i_f - i_d = \pi^e_f - \pi^e_d$$

- This relationship is referred to as ***International Fisher effect***. It is based on both real interest rate parity* and ex ante PPP.

***Real Interest rate parity:** According to real interest rate parity, the level of real interest rates in the domestic country will converge to the level of real interest rates in the foreign country.

According to International Fisher effect, the exchange rate of a country with a higher (lower) interest rate than its trading partner should depreciate (appreciate) by the amount of the interest rate difference to maintain equality of real rates of return.

IMPORTANT:

If all the key international parity conditions are held at all times, then the expected % change in the spot exchange rate would be equal to

- The forward premium or discount (in %)
- The nominal yield spread between countries
- The difference in expected national inflation rates

Implying that when all the key international parity conditions hold, no profitable arbitrage opportunities on exchange rate movements would exist for a global investor.

International Parity Conditions: Tying All the Pieces Together

International parity conditions provides support to long-term exchange rate movements as in the long-run, there is as unclear interaction among nominal interest rates, exchange rates and inflation rates

To summarize, following are six international parity conditions

1. According to **Covered interest rate parity:** Arbitrage ensures that Nominal interest rate spreads = % forward premium or discount
2. According to **Uncovered interest rate parity:** Nominal interest rate spread should reflect the expected % Δ of the spot exchange rate.
3. Forward exchange rate will be unbiased predictor of future spot exchange rate when **covered and**

uncovered interest rate parity hold i.e. nominal yield spread = forward premium or discount = expected % Δ in spot exchange rate

4. According to ***ex ante PPP*** expected Δ in the spot exchange rate = expected difference b/w domestic & foreign inflation rate.
5. According to **International Fisher effect**, assuming fisher effect and interest rate parity holds*, then nominal yield spread b/w domestic & foreign markets = domestic-foreign expected inflation difference.
   *i) Nominal interest rate = real interest rate + expected inflation & ii) real interest rates are same across all markets]
6. If **ex ante PPP and Fisher Effect hold** then expected inflation differentials = expected Δ in exchange rate = nominal interest rate differentials..

Practice: Example 5 & 6 from the CFA Institute's Curriculum.

10. THE CARRY TRADE

Foreign Exchange (FX) Carry trade Strategy: This strategy involves going ***long*** a basket of ***high-yielding*** currencies and simultaneously going ***short*** a basket of ***low-yielding*** currencies (also called funding currencies).

During periods of low volatility, carry trades tend to generate positive excess returns. However, during periods of high volatility, carry trades are exposed to significant losses as during such times:

In times of uncertainty, the risk of adverse exchange rate movements increases sharply. High-yield currencies experience selling pressures as investors shift preference to low-interest rate currencies that considered safe. Consequently, the realized returns on long high-yield currency positions are likely to significantly decline, while those on low-yield currencies tend to rise.

If uncovered interest rate parity holds at all times, then using carry trade strategy is not profitable

***Argument for persistence of the Carry Trade*:** It is argued that high-yield currencies represent a risk premium paid for more risky markets and unstable economy; whereas low-yield currencies represent less risky markets.

Reason behind risk of large losses: A Carry trade is a leveraged trade i.e. it involves borrowing in the funding currency and investing in the high-yield currency. Like all leverage, it increases the volatility in the investor's return on equity.

Properties of Carry trade returns:

1) The distribution of carry trade returns is *more peaked* than a normal distribution i.e. tends to generate a larger number of trades with small gains/losses.
2) The distribution of carry trade returns tends to have *fatter tails* and is *negatively skewed* i.e. tends to generate more frequent and larger losses.

Practice: Example 7 from the CFA Institute's Curriculum.

11. THE IMPACT OF BALANCE OF PAYMENTS FLOWS:

A country's BOP can have a significant impact on the level of its exchange rate and vice versa.

Balance of Payment (BOP): The balance of payments represents the record of the **flow** of all of the payments between the residents of a country and the rest of the world in a given year.

BOP = current account + capital account + official reserve account = 0

The three components of BOP are:

Current Account: represents part of economy engaged in actual production of goods and services.
Capital Account: reflects financial flows

Decisions of trade flows (current accounts) and financial flows (capital accounts) are made by different entities and changes in exchange rates aligned these decisions.

Current Account Surplus: Countries with +ve current account balance where Exports>Imports.

Current Account Deficits: Countries with –ve current account balance where Imports>Exports (must attract funds from abroad to keep balance).

Note:
In the long run, countries with persistent current account: surpluses (deficits) often exhibit currency appreciation (depreciation).

At least in the short-to-intermediate term, exchange rate movements are primarily determined by investment/financing decisions (i.e. capital account balance) because:

1) Prices of real goods and services tend to adjust quite slowly than exchange rates and other asset prices.
2) Production of real goods and services takes place over time and demand decisions suffer from substantial inertia. In contrast, in liquid financial markets, financial flows are instantly redirected.
3) Current spending/production decisions only reflect purchase/sales of current production, whereas the investment/financing decisions reflect both the financing of current expenditures and reallocation of existing portfolios.
4) The actual exchange rate is very sensitive to perceived currency values because the expected exchange rate movements can lead to large short-term capital flows.

Current Account Imbalance and the Determination of Exchange Rates

Trends in current account balance affect the exchange rates through following three channels:

- ✓ The flow of supply/demand channel
- ✓ The Portfolio Balance Channel
- ✓ The Debt Sustainability Channel

1) The flow supply/demand channel:

It is based on the fact that supply of domestic currency is driven by the country's demand for foreign goods and services while the demand for domestic currency is driven by foreign demand for a country's goods and services.

- Current account surplus (deficit) implies ➔ higher (lower) demand for domestic currency ➔appreciation (depreciation) of the domestic currency against foreign currencies.
- However, when the domestic currency reaches some particular level, appreciation (depreciation) of the currency leads to deterioration (improvement) in the trade balance of the surplus (deficit) country.

The change in exchange rates that is needed to restore current account balance depends on the following factors:

1) **The initial gap between imports and exports:** When the initial gap between imports and exports is relatively wide for a ***deficit*** nation, then relatively higher growth in exports than growth in imports is needed to narrow the current account deficit.
2) **The sensitivity of import and export prices to changes in the exchange rate:** Typically, depreciation of the deficit country's currency should result in:

- An increase in import prices in domestic currency terms.
- A decrease in export prices in foreign currency terms.

However, it has been experienced that changes in exchange rates have very limited pass-through affect on prices of traded goods and services. Thus,

- The limited (greater) the pass-through of exchange rate changes into traded goods/services prices, the more (less) substantial changes in exchange rates are required to narrow a trade imbalance.

3) **The sensitivity of import and export demand to the changes in import and export prices:** For a deficit nation, when import demand is more price elastic than export demand, then its currency needs to be depreciated by a substantial amount to restore the current account balance.

2) The portfolio balance channel:

According to this channel, current account imbalances shift wealth from deficit nations to surplus nations that can lead to shifts in global asset preferences.

- For example, countries running large current account surpluses against a deficit country may seek to reduce their holdings of deficit country's currency to a desired level; as a result, the value of deficit country's currency is negatively affected.

3) The debt sustainability channel:

According to this channel, running a large and persistent current account deficit ultimately leads to a continuous rise in external debt as a % of GDP. Thus, to narrow the current account deficit and to stabilize the external debt at some sustainable level, a deficit country's currency needs to be depreciated by a substantial amount; and consequently, the currency's real long-run equilibrium value declines.

- For surplus countries, opposite occurs.

Practice: Examples below this from the CFA Institute's Curriculum.

12. CAPITAL FLOWS

The importance of global financial flows in determining exchange rates, interest rates and broad asset price trends has increased with an increase in financial integration of the world's capital markets and free flow of capital.

Excessive Capital inflows in Emerging Markets (EM) fuel boom-like conditions (before crises), such as:

- EM currencies appreciation
- Overinvestment in risky projects
- Asset bubble
- Consumption binge & huge domestic credit
- Huge external indebtedness
- Huge credit/credit account deficit

When Crises Occurs (boom-like conditions suddenly reverse)

- Major economic downturn
- Soverign Default
- Serious Banking Crises
- Significant Currency Depreciation

EM Governments take preventive actions such as:

- Using capita controls to resist huge capital inflows
- Selling domestic currency in the FX market

Equity Market Trends and Exchange Rates

In the long-run, the correlation between exchange rates and equity markets is fairly close to zero.

In the short-to medium term periods, the correlation between exchange rates and equity markets is ***unstable*** i.e. it tends to fluctuate from being highly positive to being highly negative based on the market conditions. Hence, it is difficult to forecast expected exchange rate movements based solely on expected equity market performance.

- When investors' appetite for risk is high (i.e. investors are less risk reverse) ➔ the market is said to be in "risk-on" model; as a result,
  - Investors' demand for risky assets (e.g. equities) tends to increase ➔ which increases the prices of those assets.
  - Investors' demand for safe haven assets (e.g. dollar) tends to decline, which decreases the price of those assets.
- When investors' appetite for risk is low (i.e. investors are more risk reverse) ➔ the market is said to be in "risk-off" model; as a result,
  - Investors' demand for risky assets (e.g. equities) tends to decline ➔ which lowers the prices of those assets.
  - Investors' demand for safe haven assets (e.g. dollar) tends to increase, which increases the price of those assets.

Practice: Example 8 from the CFA Institute's Curriculum.

13. MONETARY AND FISCAL POLICIES

The Mundell-Fleming Model

According to the Mundell-Fleming model, changes in monetary and fiscal policy affect exchange rates through their impact on interest rates and economic activity (output) within a country.

- The model is based on aggregate demand only as it assumes that economy is undercapitalized such that supply can be adjusted without any significant changes in price level or inflation rate.
- In other words, in a Mundell-Fleming model, changes in the price level and/or the inflation rate do not play any role

Easy or expansionary *monetary* policy: Expansionary monetary policy involves reducing interest rates, increasing investment & consumption.

Expansionary monetary policy (lowering interest rates) with flexible exchange rates, induce capital outflows to higher-yielding markets and cause currency depreciation.

Easy or expansionary *Fiscal* policy: Expansionary fiscal policy involves reducing taxes and/or increasing government spending exerts upward pressures on interest rates.

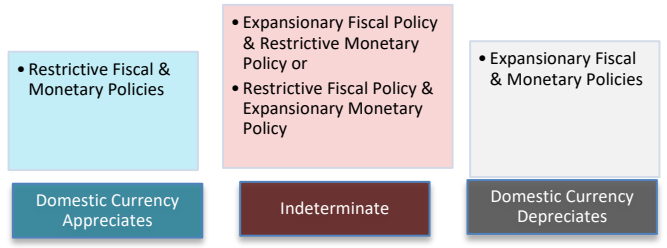
Expansionary fiscal policy with Flexible Exchange Rates leads to:

- An upward pressure on interest rates to finance larger budget deficits.

- Increase in capital inflows from lower-yielding markets and the consequent appreciation of the currency.
- If capital flows are immobile or insensitive to interest rate differentials, the increase in demand increase imports and worsen the trade balance and downward pressure on currency.

Under high capital mobility, what happens to exchange rates for the following Monetary-Fiscal Policy Mix

2) Under low capital mobility, what happens to exchange rates for the following Monetary-Fiscal Policy Mix



When capital mobility is high: Changes in monetary and fiscal policies affect exchange rates mainly through ***capital flows*** rather than trade flows.

- *The combination of expansionary fiscal policy and a restrictive monetary policy* is extremely bullish for a currency.
- *The combination of a restrictive fiscal policy and expansionary monetary policy* is extremely bearish for a currency.
- *The effect of combination of expansionary fiscal policy and expansionary monetary policy* on currency is ambiguous.
- *The effect of combination of restrictive fiscal policy and restrictive monetary policy* on currency is ambiguous.

When capital mobility is low: Changes in monetary and fiscal policies affect exchange rates mainly through ***trade flows*** rather than capital flows.

- *The combination of restrictive fiscal policy and restrictive monetary policy* is bullish for a currency because it tends to improve trade balance.
- *The combination of expansionary fiscal policy and expansionary monetary policy* is bearish for a currency because it tends to deteriorate trade balance.
- *The combination of expansionary fiscal policy and a restrictive monetary policy* has an ambiguous effect on AD and trade balance and hence on currency.
- *The combination of a restrictive fiscal policy and expansionary monetary policy* has an ambiguous effect on AD and trade balance and hence on currency.

Monetary Models of Exchange Rate Determination

Unlike Mundell-Fleming model, under the Monetary models of exchange rate determination,

- Output is fixed.
- The monetary policy affects exchange rates through its impact on the price level and the inflation rate.

The Monetary Approach states that:

- Changes in domestic price levels are primarily determined by changes in domestic money supply i.e. an X% increase (decrease) in the domestic money supply will lead to an X% increase (decrease) in the domestic price level.
- A money supply-induced increase (decrease) in domestic prices relative to foreign prices will lead to a proportional decline (increase) in domestic prices relative to currency's value.

In summary:

- A **relative increase** in a domestic's money supply causes its currency to **depreciate.**
- A **relative decrease** in a domestic's money supply causes its currency to **appreciate.**

Limitation of Pure monetary approach: Since the model assumes that PPP holds at all the times (i.e. both the short and long run), it does not provide a realistic explanation of the impact of monetary factors on the exchange rates.

The Dornbusch Overshooting Model

The Dornbusch Overshooting Model is a ***modified*** monetary model of the exchange rate. This model is free from the limitation of ***pure*** monetary approach model.

Assumptions of the model:

- In the short-run, prices are fixed or have limited flexibility.
- In the long-run, prices are fully flexible, implying that any increase in the domestic money supply will give rise to a proportional increase in domestic prices and will depreciate the domestic currency.

According to the model,

In the short-run when domestic price level is ***inflexible*** and capital is highly mobile, any increase in the nominal

money supply leads to decrease in domestic interest rate➔increase in capital outflow to higher-yielding countries ➔as a result, the nominal exchange rate ***over-depreciates*** (overshoot its long-run PPP level)➔ giving a signal that the domestic currency is so undervalued that it is expected to appreciate in the future.

Practice: Example 9 from the CFA Institute's Curriculum.

Portfolio Balance Approach

Mundell-Fleming model determines exchange rates in short-term but fails to capture the long-term effects of budgetary imbalances. Portfolio Balance approach resolve this limitation.

Portfolio Balance Approach:

According to this approach, exchange rate is a function of relative supplies of domestic and foreign bonds; so, the exchange rate is determined by equilibrium in global asset market.

- Global investors prefer to hold a ***diversified*** portfolio of domestic and foreign assets, including bonds and the desired allocation of each investor depends on expected return and risk.
- A persistent increase in government budget deficit leads to a steady increase in the supply of domestic bonds outstanding
- Investors will held these bonds if they are compensated for returns such as:
  - Higher interest rates and/or higher risk premium
  - Depreciation of the currency to a level sufficient to generate expected profit from subsequent appreciation of the currency.
  - Some combination of the two.

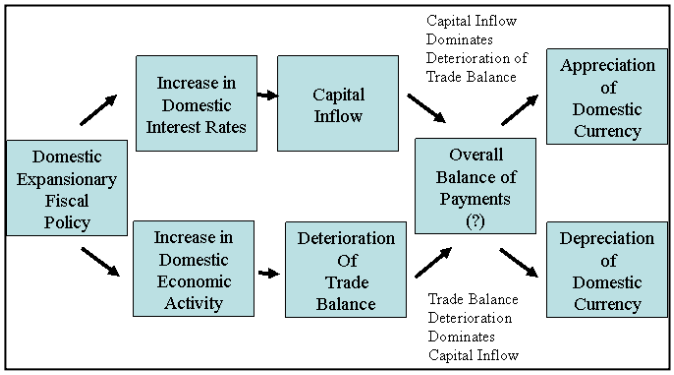
Important to Note:

In the long run, currencies of countries that run large budget deficits on a persistent basis eventually depreciate.

The combination of Mundell-Fleming and portfolio balance models:

When capital is highly mobile:

- In the short-run, expansionary fiscal policy leads to an increase in domestic real interest rates relative to other countries → which leads to appreciation of the domestic currency.
- In the long-run, expansionary fiscal policy may cause the amount of debt to increase to an unsustainable level; thus, either
  the central bank is forced to ***monetize the debt*** i.e. print additional money to buy the government's debt. Consequently, the domestic currency depreciates; or
  fiscal stance will become restrictive i.e. seeks to reduce the public deficit and debt levels by issuing fewer government bonds. As the supply of bonds fall → bonds yields decrease and consequently, the domestic currency depreciates.



Practice: Example 10, From the CFA Institute's

14. EXCHANGE RATE MANAGEMENT: INTERVENTION AND CONTROLS

Capital inflows can affect an economy positively or negatively. An increase in capital inflows can increase a country's economic growth and asset values. As a result currency appreciates and foreign investors earn healthier returns.
On the other hand, capital inflows can worsen a country's asset price bubble or overvaluation of its currency. When short-term capital inflow eventually reverse, foreign investors suddenly draw their capital back, economy suffers significant drop in asset prices and huge currency depreciation.

Increase in capital inflows are caused by a combination of 'pull' and 'push' factors.

Pull Factors, may stem from both public or private sectors, represent positive developments in an economy that attract overseas capital into that economy. For example,

- Expected decline in inflation and inflation volatility
- More-flexible exchange rate regimes
- Improved fiscal positions

- Privatization of state-owned entities
- Liberalization of financial markets
- Removal of foreign exchange regulations and controls

As a result, growth in private sector attracts foreign investment, healthy export sector improves current account balance and strong FX reserves support against future speculative attacks.

Push Factors: They represent a set of developments in *other economies* that causes overseas capital to flow to a particular economy. For example,

- The low interest rate policies in other economies (specially industrial countries) may encourage investors to shift capital to high-yielding economies.
- Changes in asset allocation over time, which eventually increase the share of funds allocated to a particular economy e.g. high allocation in emerging market (EM) countries.
- Studies have shown that EM countries better handled the global financial crises of 2008, as a result, capital flows to EM countries rose.
- Ultra low interest rates in economies such as U.S., Euro area and Japan have encouraged investors to invest in high-yielding EM economies.

Governments directly intervene to resist excessive inflows and currency bubbles. Some forms of capital control may include:

- Preventing banks from selling local currency in a FC transaction.
- Limiting FC transaction by imposing higher tax rates.
- Requiring that a portion of investment must be deployed in a term bank deposit.
- Limiting foreign ownership of local institutions.
- Preventing foreign investors from repatriating funds from the sale of financial assets.
- Making local currency available at exchange rates linked to the usage of foreign currency.

Many participants believe that capital control distort global trade and finance, deflect capita flows from economies and complicate monetary and exchange rate policies in those economies. Despite these concerns IMF asserts that the resultant benefits of capital control may exceed the related costs as capital controls prevent countries from future financial deterioration, asset bubble formation and overshooting of exchange rates.

NOTE:

The higher the ratio of **central bank FX reserves holding to average daily FX turnover in the domestic currency**, the greater their firepower, and thus, the greater the ability to affect the level and path of exchange rates.

The ratio is negligible in developed countries but is quite sizeable in EM economies. Therefore, the effectiveness of intervention is:

- limited in developed markets
- more mixed in EM economies as it lowers the exchange rate volatility.

15. Warning Signs of CURRENCY CRISES

When capital inflows unexpectedly stop, the economy contracts, asset value fall, the currency sharply depreciates and it may result in financial crises.

- As selling pressure begins, repositioning of portfolios & liquidation of vulnerable positions by investors and borrowers, to avoid excessive capital losses, intensify the currency crises.
- The major issue with such currency crises is that they are difficult to anticipate adequately because their underlying causes differ greatly.
- Developing early warning signs are challenging because views on primary causes of crises varies.

Two School of thoughts with regard to Currency crisis anticipation:

- According to the first school of thought, the major cause of the currency crises is the deteriorating and weak economic fundamentals, implying that if an economy is facing weak and deteriorating fundamentals, it gives a warning sign that in the near future its currency may be vulnerable to speculative attacks.
- According to the second school of thought, there is no particular factor that may precipitate currency crisis i.e. it can occur out of the blue. Under this school of thought, an economy with sound and strong economic fundamentals may suffer from speculative attacks on its currency because of
  - an abrupt adverse shift in market sentiment, completely unrelated to economic fundamentals or
  - contagion or spillover effects arising from crises developments in other markets.

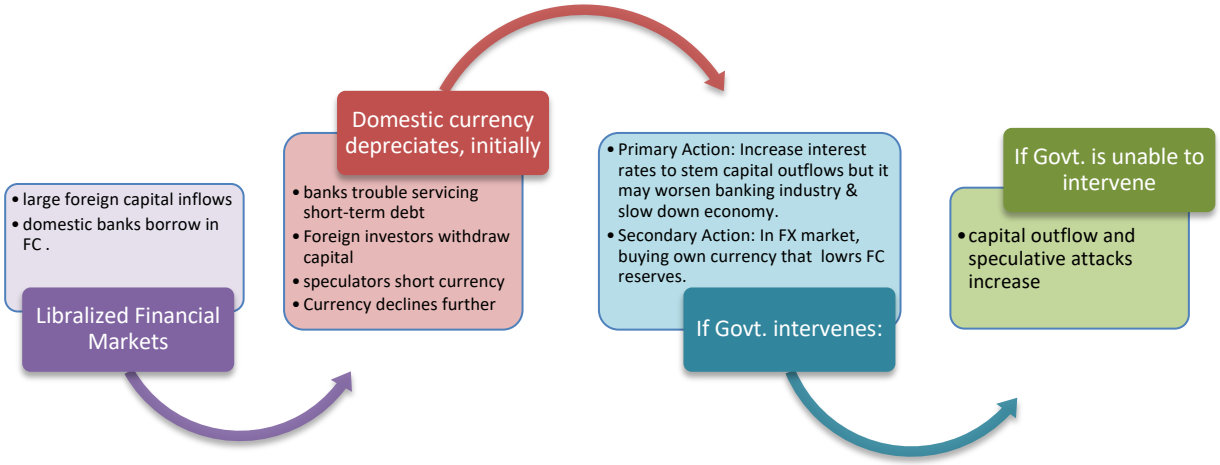
Features of an Ideal Early Warning System: An ideal warning system should

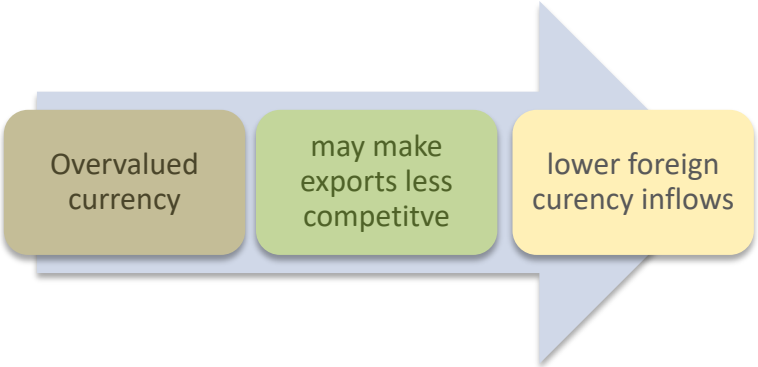
i. Have a strong record both with regard to predicting actual crises and preventing the frequent issuance of false signals.
ii. Be based on macroeconomic indicators with readily available data without any long time lags.
iii. Provide an early warning signal *well in advance* of actual currency crises to provide market participants sufficient time to adjust or hedge their portfolios.
iv. Be broad-based i.e. covers a wide range of indicators of currency crises.

Number of variables used to anticipate currency crises:
Although variables or methodologies differ from one study to the next, following are some conditions identified in one or more studies.

1. Liberalized capital markets i.e. free flow of capital prior to a currency crisis.
2. Pre crises period exhibits large foreign capital inflows (relative to GDP). Foreign currency denominated short-term funding is particularly problematic.
3. A currency crisis often leads to banking crises.
4. Countries with fixed or partially fixed exchange rates are more prone to currency crises than countries with floating exchange rates.
5. As crises approaches, foreign exchange reserves decline sharply.
6. Pre-crises period exhibits excessive appreciation of the currency as compared to its historical mean.
7. Prior to crises, the term of the trade ($ratio\ of\ \frac{Exports}{Imports}$) deteriorates.
8. Prior to crises, broad money growth and the ratio of 'M2 money supply to bank reserves' rise.
9. Excessive rise in inflation has been observed in pre crises period compared with tranquil period.

- *These factors are highly interrelated and often one factor leads to another. The two diagrams below show how the above-mentioned factors are interrelated*





Though unable to accurately predict upcoming currency crisis, models –combined with other analytical judgment and analysis, can help in preparing for and assessing exposures.

Practice: Example 11 from the CFA Institute's Curriculum.

Practice: CFA Institute's Curriculum End of Chapter Questions & FinQuiz Question-bank (Item-sets + Questions)

1. AN INTRODUCTION TO GROWTH IN THE GLOBAL ECONOMY

Actual growth rate of GDP v/s Potential GDP growth rate: In the long-run, the actual growth rate of GDP should be equal to the growth rate of potential GDP.

- The growth rate of potential GDP acts as an upper limit to growth i.e. it is the maximum amount of output that an economy can sustainably produce without creating any upward pressures on the inflation rate.

As the growth rate of potential GDP increases:

- The level of income rises: Due to compounding effect, even small changes in the growth rate lead to large changes in the level of income over time.
- The level of profits rises.
- The living standard of the population increases.

A country's standard of living and level of economic development can be measured by estimating GDP and per capita GDP.

Economic growth = Annual % change in real GDP or in real per capita GDP

Real GDP: Growth in real GDP reflects the expansion in total economy over time.

Real per capita GDP: Growth in real per capita GDP reflects the increase in the average standard of living in each country. It indicates that growth in real GDP is greater than that of population. Countries with high (low) per capita GDP are said to be developed (developing) countries.

Converting GDP in currency terms: The GDP data can be converted into currency (e.g. in dollars) in two ways i.e.

1) Using current market exchange rates:

Country's GDP measured in its own currency × Current market exchange rates

Limitations: This method is inappropriate for two reasons i.e.

i. Market exchange rates are highly volatile and thus, even small changes in the exchange rate can translate into large changes in estimated value of GDP even if there is little or no growth in the country's economy.
ii. Since market exchange rates are determined by trade and financial flows, they do not incorporate differences in prices of non-tradable goods and services across countries. As a result, the standard of living of consumers in developing countries is understated.

2) Using exchange rates implied by PPP:

Country's GDP measured in its own currency × Exchange rates implied by PPP

- This method is more appropriate to compare living standards across time and/or across countries because at PPP implied exchange rates, the cost of a typical basket of goods and services is the same across all countries.

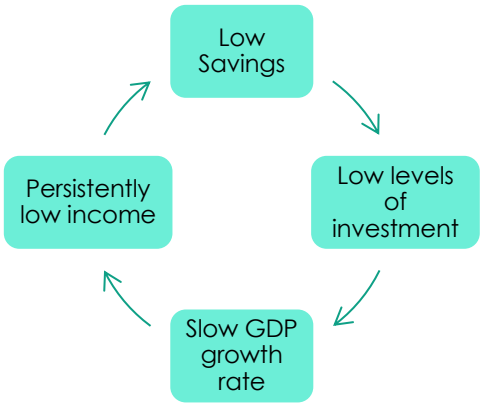
2. FACTORS FAVORING AND LIMITING ECONOMIC GROWTH

Savings and Investment

Capital can be accumulated through private and public sector investment.

- The higher the level of disposable income → the higher the savings → the greater the capital accumulation → the higher the per capita GDP.

Vicious Cycle of Poverty:



- The vicious cycle of low savings can be broken by attracting greater foreign investment.
- Besides level of savings, an economy's growth also depends on how efficiently saving is allocated within the economy.

1 Financial Markets and Intermediaries

Financial markets act as an intermediary between savers and borrowers. Better-functioning financial markets facilitate countries to grow at a faster rate. Financial markets and intermediaries (i.e. banks) can promote growth in at least three ways:

1. The financial markets and intermediaries channel financial capital (savings) from savers to those investment projects that are expected to generate the highest risk-adjusted returns.
2. By creating attractive investment instruments that facilitate risk transfer and diversification, the financial markets and intermediaries encourage savers to invest and assume risk.
3. The existence of well-developed and better-functioning financial markets and intermediaries facilitate corporations to finance their capital investments.

However, financial sector intermediation that results in declining credit standards and/or increasing leverage will increase risk.

2 Political Stability, Rule of Law, and Property Rights

Key ingredients for economic growth are:

- Stable and effective government;
- Well-developed legal and regulatory system that establishes, protects and enforces property rights;
- Enforcement and respect for property rights that govern the protection of private property, intellectual property;

All the above mentioned factors encourage domestic households and companies to invest and save.

Factors that increase investment risk, discourage foreign investment, and weaken growth:

- Wars
- Military coups
- Corruption
- Political instability

3 Education and Health Care Systems

Adequate education at all levels is a key component of a sustainable growth for all the economies.

- Physical and human capital are often complementary; thus, the productivity of existing physical capital can be enhanced by increasing human capital via improving education through both formal schooling and on-the-job training.
- The economic growth also depends on the efficiency of allocation of education spending among different types and levels i.e. primary, secondary and post-secondary.

The impact of education spending varies among developing and developed countries i.e.

a) Developed countries are on the ***leading edge of technology***; thus, they need to invest in ***post-secondary education*** to promote innovation and growth. In such countries, incremental spending on post-secondary education will have a greater impact on growth.

b) Developing countries mostly apply and imitate technology developed elsewhere; thus, they need to invest in primary and secondary education. In such countries, incremental spending on primary and secondary education will have a greater impact on growth as it improves a country's ability to absorb new technologies and to perform tasks more efficiently.

4 Tax and Regulatory Systems

Tax and regulatory policies play an important role in the growth and productivity of an economy, particularly at the company level.

- The limited regulations promote entrepreneurial activity → attract new companies → increase productivity levels.
- In addition, the lower the administrative start-up costs, the greater the entrepreneurial activity.

5 Free Trade and Unrestricted Capital Flows

Opening an economy to capital and trade flows has a significant impact on economic growth.

Benefits of an open Economy: In an open economy,

- Domestic investment can be financed using world savings.
- World savings can facilitate an economy to break the vicious cycle of property (explained above).
- An economy can attract foreign investment, which helps an economy to break the vicious cycle of poverty, by increasing savings, physical capital stock, productivity, employment and wages.

Types of Foreign investment:

1) **Foreign direct investment (FDI):** It refers to the direct investment by foreign companies in a domestic country in the form of building or buying property, plant, and equipment. FDI facilitates developing countries to have access to technology developed and used in developed countries.

2) **Foreign indirect investment:** It refers to an indirect investment by foreign companies and individuals in a domestic economy in the form of purchase of securities (equity & fixed income) issued by domestic companies.

Benefits of Free Trade:

- Reducing tariffs on foreign imports (capital goods, in particular) and removing restrictions on foreign direct and indirect investments tend to lead to higher economic growth.
- By reducing tariffs and restrictions on trade, domestic residents can have access to a variety of goods at relatively lower costs.
- Free trade promotes competition among domestic companies by decreasing their pricing power and provides them an access to larger markets.

Summary of Factors Limiting Growth in Developing Countries

Factors that negatively impact growth:

- Low rates of saving and investment
- Poorly developed financial markets
- Weak and/or corrupt legal systems
- Lack of enforcement of laws
- Lack of property rights
- Unstable political system
- Inadequate and poor public education and health services
- Tax and regulatory policies discouraging entrepreneurial activity
- Restrictions on international trade and flows of capital

Pre-conditions for Economic Growth:

1) Well-functioning and well-developed markets
2) Clearly defined property rights and rule of law
3) No restrictions on international trade and flows of capital
4) Adequate public education and health services
5) Tax and regulatory policies that encourage entrepreneurial activity
6) Adequate investment in infrastructure that increases stock of physical capital, labor productivity and growth.

It is important to understand that *an economy needs to have a **sustained** (not one time) increase in growth rates to become a high-income country and to improve its standard of living*.

Obstacles to growth in the developing countries:

- Inadequate education level
- ***"Brain drain***" problem i.e. departure of most highly educated individuals in developing country to the developed countries.
- Lack of appropriate institutions;
- Poor legal and political environment
- Lack of physical, human, and public capital
- Little or no innovation
- Poor health
- Lower life expectancy rates

Practice: Example 1, From the CFA Institute's Curriculum.

3. WHY POTENTIAL GROWTH MATTERS TO INVESTORS

The potential risk and return associated with long-term investments in the securities of companies located or operating in that country can be evaluated based on an economy's long-term economic growth.

Relationship between economic growth and stock prices:

- Equity values reflect anticipated growth in aggregate earnings, which in turn depend on expectations of future economic growth.
- Generally, earnings growth rate of companies operating in an economy is lower than the earnings growth rate for the overall economy.
- However, when the ratio of corporate profits to GDP increases over time ➔ company's earnings will grow at a rate greater than the of GDP growth rate.
- It must be stressed that earnings growth rate cannot exceed GDP growth rate on a persistent basis, implying that *in the long-run, real earnings growth cannot exceed the growth rate of potential GDP*.
- The economic growth and the long-run growth of aggregate earnings depend on same factors.

The performance of stock market depends on an economy's performance, measured by its GDP.

$$P = GDP\left(\frac{E}{GDP}\right)\left(\frac{P}{E}\right)$$

where,

P = Aggregate value (price) of equities
E = Aggregate corporate earnings
GDP = can be real or nominal with a corresponding real or nominal interpretation of the other variables.

Expressing in terms of logarithmic rates:

(1/T) % ΔP = (1/T) % ΔGDP + (1/T) %Δ (E / GDP) + (1/T) % Δ(P / E)

% change in stock market value = % change in GDP + % change in the share of earnings (profit) in GDP + % change in the price-to-earnings multiple

where,

T = time horizon

- ***Over short to immediate horizons***, the stock market value is affected by all of the three factors.
- ***In the long run***, the stock market value ***majorly*** depends on the growth rate of GDP.
  - The ratio of earnings to GDP can neither rise nor decline forever, implying that in the long-run, % change in the share of earnings (profit) in GDP must be approximately zero.
  - Similarly, the P/E ratio can neither rise nor decline forever, implying that in the long-run, % change in the price-to-earnings multiple must be approximately zero.
- Hence, in the long-run, changes in the earnings-to-GDP and P/E ratios largely affect the volatility of the market, not its return.
- A country's GDP growth rate is not constant; rather, it can and does change (i.e. increase or decrease) over time.
- Factors and policies that affect potential growth rate of an economy by a small amount lead to large changes in the standards of living and the future level of economic activity due to effect of compounding.
- A ***persistent*** increase in the rate of labor productivity growth increases the sustainable economic growth rate, resulting in increase in the earnings growth and potential return on equities.

Relationship between Fixed income returns and Economy's potential growth rate:

Fixed income returns are mainly based on the relationship between actual and potential growth.

- When an actual GDP > (<) potential GDP → inflation increases (decreases) → nominal interest rates increase (decrease) and consequently, bond prices fall (rise).
  - However, it does not imply that there is a long-run trade-off between growth and inflation.
- The level of real interest rates and real asset returns also depend on the growth rate of potential GDP of an economy.
- The real return that consumers/savers demand for forgoing present consumption is the real interest rate. Thus,
  - The higher the potential GDP growth rate → the higher the real interest rate → the more consumers save and the higher the expected real asset returns, in general.
  - In addition, when the rate of potential GDP growth increases → the general credit quality of fixed income securities improves because such securities are mostly backed by a flow of income.
- Monetary policy decisions also depend on output gap (i.e. difference between an economy's estimated potential output ***level*** and its actual operating level) and difference between ***growth rate*** of actual GDP and potential GDP.
  - When forecasted actual GDP growth < (>) growth in potential GDP → output gap widens (narrows) → an economy slows down (heats up) → downward (upward) pressure on inflation → inflationary expectations reduce (increase).
  - To close this output gap → the central bank may need to pursue an easy (tight) monetary policy by lowering (raising) short-term interest rates; as a result, bond prices rise (fall).
- The growth rate of potential GDP is also used by credit rating agencies to evaluate the credit risk of sovereign or government-issued debt i.e. the higher the estimated potential GDP growth rate, the lower the perceived risk of such bonds, all else equal.
- Fiscal policy decisions also depend on output gap and difference between growth rate of actual GDP and potential GDP i.e. typically,
  - During recessions (i.e. when output gap widens) → a government may need to pursue an easy fiscal policy, leading to increase in budget deficits.
  - During expansions (i.e. when output gap narrows) → a government may need to pursue a tight fiscal policy, leading to decrease in budget deficits.

Volatility of Equity market v/s Long-term real GDP growth:

- Due to high volatility associated with equity market, it is very difficult to predict equity returns using historical equity returns. In contrast, since long-term real GDP growth rate depends on slowly evolving fundamental economic factors, it tends to exhibit relatively low volatility, particularly in developed countries.
- Similarly, countries with prudent monetary policies tend to have less volatile inflation rates compared to stock prices.

Practice: Example 2 from the CFA Institute's Curriculum.

4. DETERMINANTS OF ECONOMIC GROWTH

1. Production Function

Inputs to economic growth:

1. Labor
2. Physical
3. Human capital
4. Technology
5. Natural resources
6. Public infrastructure

A two-factor aggregate production function:

$$Y = AF(K, L)$$

where,

Y = Level of aggregate output in the economy
L = Quantity of labor or number of workers or hours in the economy
K = Stock of capital used to produce goods and services
A = Total Factor Productivity (TFP)

NOTE:

Capital and labor can be employed in various combinations to produce output.

Total Factor Productivity (TFP): It reflects the general level of productivity, innovation or technology in the economy.

- Increase in TFP implies a proportionate increase in output for any combination of inputs.
- Increase in TFP does not imply a ***change*** in the relative productivity of the inputs.
- Changes in TFP are estimated using a growth accounting method, explained below.

Cobb-Douglas Production Function: It is stated as follows:

$$F(K, L) = K^{\alpha} L^{1-\alpha}$$

where,

α = Share of output or GDP paid by companies to suppliers of capital*
$1 - \alpha$ = Share of output or GDP paid by companies to suppliers of labor

NOTE:

The value of α lies between 0 and 1.

*Under the Cobb-Douglas production function, MPK is stated as:

$$\textbf{MPK} = \alpha AK^{\alpha-1} L^{1-\alpha} = \alpha Y/K$$

Since in a competitive economy, profit is maximized when Marginal product of capital (MPK) = Rental price of capital (r) and Marginal product of labor = Real wage rate,

$\alpha Y/K = r \rightarrow \alpha = r(K) / Y$ = Capital income / Output or GDP

Output per worker or Average labor productivity (Y/L or y): It refers to the average amount of goods that can be produced by a unit of labor. It is estimated as follows.

GDP/Labor input = TFP × capital-to-labor ratio × share of capital in GDP

Or

$$y = Y/L = Ak^{\alpha}$$

Capital-to-labor ratio (k):

It reflects the amount of capital available for each worker.

- Due to lack of human and physical capital, developing countries have relatively *higher labor productivity growth* and relatively small impact of diminishing marginal returns than developed countries but *low levels* of productivity compared to developed countries.

Two important Properties of Cobb-Douglas Production Function:

1) The Cobb-Douglas production function exhibits ***constant returns to scale*** i.e. if all the inputs into the production process are increased by the X %, then output will also increase by X %.

- Constant returns to scale implies that $\alpha + (1-\alpha)$ must always = 1.
- This indicates that if both inputs are increased proportionately, then there are no diminishing marginal returns.

2) The Cobb-Douglas production function exhibits ***diminishing marginal productivity*** with respect to each individual input i.e. the marginal (incremental) output produced by employing each additional unit of *variable* input, keeping the other inputs unchanged, will decline.

- The diminishing marginal productivity implies that adding more and more capital to a fixed number of workers increases per capita output *but at a decreasing rate*.

Significance of diminishing marginal returns to capital depends on the importance of capital in production i.e. value of "α":

- When the value of "α" is close to 0, reflecting that capital is relatively unimportant, diminishing marginal returns to capital will be ***very significant (rapid)*** and any addition to capital will NOT have a considerable impact on growth. In other words, as capital increases, each additional unit of capital will result in progressively smaller increase in output.
- When the value of "α" is close to 1, reflecting that capital is very important, diminishing marginal returns to capital will be ***small (slow)*** and any addition to capital will have a considerable impact on growth. In other words, the marginal output produced by employing next unit of capital will nearly the same as that of previous unit of capital.

2. Growth Accounting

Three sources of growth:

1) Growth in Labor
2) Growth in Capital
3) Technological Progress (TFP)

Growth Accounting Equation based on Solow Approach:

Growth rate of output = Rate of technological change + (α × Growth rate of capital) + (1- α) × Growth rate of labor

$$\Delta Y / Y = \Delta A / A + \alpha \Delta K/K + (1 - \alpha) \Delta L/ L$$

where,

α = Elasticity of output with respect to capital i.e. 1% increase in capital leads to α% increase in output. It also represents relative shares of capital in national income.

(1 – α) = Elasticity of output with respect to labor. It also represents relative shares of labor in national income.

A = Growth in TFP i.e. entrepreneurial ability, education, roads, technology, natural resources etc.

- It measures the amount of output that cannot be explained by growth in capital or labor.
- It is not directly measured; rather, it must be estimated as a ***residual*** in the above equation e.g. using a time-series, econometric model.

TFP = Growth in output – Growth in the factor inputs or weighted growth rates of these inputs

- TFP estimates highly depend on the measurement of the labor and capital inputs.

Uses of Growth Accounting Equation: It can be used to

- Estimate the contribution of technological progress to economic growth
- Analyze and decompose the sources of growth in an economy.
- Quantify the contribution of each factor to long-term growth in an economy i.e. contribution of capital and labor, impact of TFP etc.
- Estimate potential output.

Labor productivity growth accounting equation: It is an alternative method of measuring potential GDP.

Growth rate in potential GDP = Long-term growth rate of labor force + Long-term growth rate in labor productivity

Advantages:

- It is a simple method relative to Solow approach.
- It does not require estimating the capital input and TFP.

Disadvantage: Under this method, it is difficult to directly analyze and predict impacts of capital deepening and TFP progress.

3. Extending the Production Function

An Extended Production Function includes the following inputs:

1. Raw materials and natural resources i.e. oil, lumber and available land (N).
2. Quantity of labor i.e. the number of workers in the country (L).
3. Human capital i.e. education and skill level of the workers (H).
4. Information, computer, and telecommunications (ICT) capital i.e. computer hardware, software, and communication equipment (K_{IT}).
5. Non-ICT capital i.e. transport equipment, metal products and plant machinery other than computer hardware and communications equipment, and non-residential buildings and other structures (K_{NT}).
6. Public capital i.e. infrastructure owned and provided by the government (K_P).
7. Technological knowledge i.e. the production methods used to convert inputs into final products, reflected by TFP (i.e. A).

$$Y = AF (N, L, H, K_{IT}, K_{NT}, K_P)$$

5. CAPITAL DEEPENING VS. TECHNOLOGICAL PROGRESS

There are two sources of Growth in per capita output:

1) **Capital deepening:** It refers to an increase in the economy's stock of capital (i.e. plant and equipment) relative to its workforce. It is reflected by increase in the capital-to-labor ratio.

As Savings increase → the amount of income available for investment increases → gross investment increases →net investment increases → eventually stock of capital increases.

Similarly,

When restrictions are removed →foreign investment increases → stock of capital increases

- The increase in capital deepening is represented by the ***movement along*** the production function i.e. from A to B in exhibit 4.
- As long as the growth rate of capital (net investment) > growth rate of labor, ➔ the capital-to-labor ratio increases.
- However, capital deepening cannot result in a sustained growth in per capita income i.e.
  - As the capital-to-labor ratio reaches a maximum value (i.e. at point B), diminishing marginal returns to capital triggers i.e. MPK declines as more capital is added to the labor input. It is represented by a movement to point D (exhibit 4).
- The point where MPK = MC, profit is maximized and no further capital is added by producers i.e. capital-to-labor ratio will stop increasing.

Contribution of Capital Deepening = Labor productivity growth rate – Total Factor Productivity

- The larger the difference between the productivity growth measures, the greater the contribution of capital deepening to the economic growth.

2) **Improvement in technology or technological progress:** It refers to the economy's ability to produce more output without using any more inputs i.e. capital or labor.

- An improvement in TFP is represented by an ***upward shift*** in the entire production function i.e. from point B to C (exhibit 4, page 593).
- Improvement in TFP also increases the marginal product of capital relative to its marginal cost and results in a *permanent (sustained)* increase in per capital output growth rate even in the steady state.

In other words, *in the absence of technological progress, a country cannot permanently increase per capital GDP growth simply by indefinitely increasing its capital stock*.

Contribution of Improvement in technology = Labor productivity growth rate – Capital Deepening

In developing countries, potential GDP growth rate can be increased through both capital deepening and technological progress; while in developed countries, the improvements in potential GDP growth rate largely depend on technological progress.

Refer to: Exhibit 3 from the CFA Institute's Curriculum.

Practice: Example 3 from the CFA Institute's Curriculum.

5. NATURAL RESOURCES

There are two categories of natural resources:

1. **Renewable resources:** These resources are the resources that can be continuously replaced by nature i.e. forest, trees.

2. **Non-renewable resources:** These resources are finite resources i.e. once they are used up, they cannot be replaced by nature e.g. oil and coal.

- Although access to natural resources is important, *ownership and production of natural resources does not necessarily imply a higher economic growth*.

- Indeed, sometimes, access to natural resources may even impede growth, resulting in a "***resource curse***" i.e. when countries rich in natural resources lack the economic institutions necessary for growth.
- Sometimes, countries rich in resources may suffer the ***Dutch disease*** i.e. strong export demand for resources leads to currency appreciation which makes other segments of the economy (particularly manufacturing), globally uncompetitive.
- Due to the shifts towards a services-based economy, the relative share of natural resources in national income has decreased in many countries.

Practice: Example 4 from the CFA Institute's Curriculum.

7. LABOR SUPPLY

Growth in the number of people available for work (i.e., quantity of workforce) is an important source of economic growth. Growth in labor input depends on four factors:

1. Population growth
2. Labor force participation
3. Net migration
4. Average hours worked

1. Population growth

Long-term anticipated growth in labor supply majorly depends on the growth of the working age population.

- Working age population growth depends on fertility rates and mortality rates. Developing countries have higher population growth rates compared to developed countries.
- It must be stressed that *population growth may result in increase in the growth rate of the overall economy but it does not affect the rate of increase in per capita GDP*.

Besides population growth rate, economic growth also depends on the age mix of the population i.e.

- Countries face a ***demographic burden*** when the share of non-working elders (i.e. over 65) in the population exhibits an increasing trend.
- Countries receive a ***demographic boost*** when the share of population below the age of 16 exhibits a declining trend.

2. Labor force participation

Labor force participation rate refers to the percentage of the *working age population* in the labor force.

- Labor force participation rate can be increased with an increase in the female labor force participation rates.
- In the short run due to changes in the labor force participation rate, the growth rate of the labor force may *not* be equal to the population growth.
- Unlike increase in population, an increase in the labor force participation rate may result in an increase in the growth of per capita GDP.
- It is important to understand that increase (or decrease) in labor force participation rate (e.g. by reducing unemployment rate) simply indicates a temporary change in the level of participation; it does not imply a truly permanent rate of change.

3. Net migration

A significant increase in immigration may offset the slow domestic labor force growth rate in an economy.

4. Average Hours Worked

Potential size of the labor input is measured by the total number of hours available for work.
Total number of hours available for work = Labor force × Average hours worked per worker

Labor force = Working age population (ages 16 to 64) that is either employed or available for work but not working (i.e. unemployment)

- The average hours worked per worker is highly sensitive to the business cycle and varies significantly across countries i.e. the average number of hours worked tends to reduce during recession.
- In most developed countries, the average number of hours worked per year has been declining, leading to shorter workweek as workers prefer leisure time to labor income due to following factors:
  - Legislation
  - Collective bargaining agreements
  - Growth of part-time and temporary work, particularly due to increase in female labor force participation rates
  - Wealth effect
  - High tax rates on labor income

NOTE:

Typically, Growth rate of labor input should be = Population growth rate + Net change in population due to immigration.

Practice: Example 5 & 6 from the CFA Institute's Curriculum.

5. Labor Quality: Human Capital

Human capital: It refers to improvement in labor quality via accumulated knowledge and skills acquired from education, on-the-job training or life experience and investment in human beings. It is considered as an important source of growth for an economy.

- Generally, the better educated and more skilled workers →the higher the productivity of the workers, the more adaptable workers to changes in technology or changes in market demand and supply → the higher the economic growth.
- Human capital can be increased by investing in education and on-the-job training and also by improving health of the population.

Benefits of Education:

- Education improves the quality of labor and increases the stock of human capital.
- Although investment in education is costly but it generates high return e.g. the higher the education, the higher the wages.
- In addition, investment in education produces a spillover or externality effect i.e. increasing the educational level of one person increases the output for the whole economy.
- Education can cause a sustained increase in the growth of an economy by promoting innovation and technological progress.

8. ICT AND NON-ICT, TECHNOLOGY AND PUBLIC INFRASTRUCUTRE

Capital: ICT and Non-ICT

Net investment in physical capital stock = Gross investment – Depreciation

- As long as the net investment in physical capital stock is positive, the physical capital stock increases over time.
- Although an economy cannot achieve a long-term sustainable growth simply by capital deepening due to diminishing marginal productivity, however, there is a high positive correlation between investment spending and economic growth i.e.
  - The higher the rate of net investment → the higher the investment to GDP ratio → the higher the growth rates of physical capital stock and → the higher the GDP growth rate.
  - However, if population is increasing, then impact of growth of per capita GDP will be small.

Rationale behind positive correlation between Investment spending and Economic Growth:

i. Despite diminishing marginal productivity of capital, investment-driven economic growth may last for a considerable period of time in capital-poor countries.

ii. The positive impact of investment spending on economic growth depends on the existing physical capital stock which varies significantly across countries i.e.

- The smaller the amount of existing physical capital per worker, the greater the positive impact of changes in physical capital stock on growth.
- For countries with a large physical capital stock, the changes in physical capital stock will have a major impact on growth only when there will be a sustained high level of investment over many years.

iii. Economic growth and productivity also depend on the composition of investment spending and the stock of physical capital.

Categories of Investment Spending:

1) ICT Capital: ICT capital refers to information, computer, and telecommunications capital e.g. computer hardware, software, and communication equipment.

- ICT capital spending measures the impact of information technology sector on economic growth.
- Use of IT equipment in various industries has also generated network externalities i.e. internet and email by interconnecting people have facilitated them to work more productivity.

2) Non-ICT Capital: Non-ICT capital includes transport equipment, metal products and plant machinery and non-residential buildings and other structures.

- Non-ICT capital spending measures the impact of capital deepening on economic growth.
- Non-ICT capital spending tends to have relatively less impact on potential GDP growth than ICT capital spending.

Technology

Technology is considered the most important source of growth for an economy.

Technological progress refers to the ability to produce more and/or higher-quality and new variety of goods and services with the same resources or inputs. Technological progress results in an upward shift in the production function.

- Changes in technology are represented by human capital (knowledge, organization, information, and experience base) and/or in new machinery, equipment, and software.
- Technology progress requires countries to innovate through expenditures, both public and private on research and development (R&D).
  - Typically, developed countries tend to have high ratio of R&D spending to GDP.
  - In contrast, since developing countries imitate or rely on technology developed in advanced countries, they tend to have lower ratio of R&D spending to GDP.
  - It is important to note that although high R&D spending increases output and productivity in the long-run; in the short-run, it may cause a cyclical slow down in growth as new technologies and processes substitute old companies and workers.

Practice: Example 7 from the CFA Institute's Curriculum.

Public Infrastructure

Public infrastructure investment is an important source of economic growth and productivity. It includes investment in roads, bridges, municipal water, dams, and electric grids etc.

- Public capital tends to have few substitutes.
- Like technology, public infrastructure investment generates an externality effect in the economy because it acts as a complement to the production of private sector goods and services.

9. SUMMARY OF ECONOMIC GROWTH DETERMINANTS

Practice: Example 8 & 9 from the CFA Institute's Curriculum.

10. THEORIES OF GROWTH

Three theories of Economic Growth:

1. The Classical Model
2. Neoclassical Model
3. Endogenous Growth Model

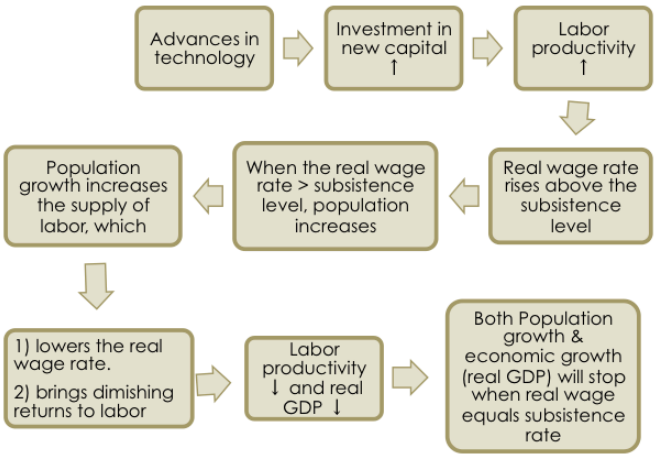
1. Classical Model

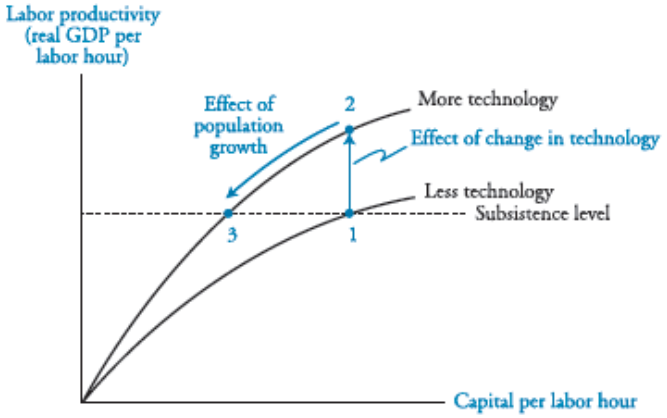
It is commonly known as the Malthusian theory. According to this theory, growth rate in real GDP per capita is temporary because an exploding population with limited resources brings an economic growth to an end.

Inputs to Production Function:

i. Land as a fixed factor
ii. Labor as a variable factor

The Basic Idea:





Implication:

- Under the classical model, in the long run, changes in technology result in a larger NOT richer population.
- Even with technological progress, an economy's standard of living is constant over time and per capita output cannot grow.

Criticism of Classical Theory: It has been observed that:

- Population growth rate is not strongly associated with increase in income per person.
- Population growth does not push income to revert back to subsistence level.
- Per capita income can grow with technological progress, which can offset the impact of diminishing marginal returns.

2. Neoclassical Model

Under Neoclassical growth theory (also known as ***Solow growth*** model), the economic growth and growth in real GDP per person depends solely on exogenous technological progress i.e. as long as technology keeps advancing, real GDP per person will persistently increase.

Inputs to Production Function (based on Cobb-Douglas production function):

i. Capital as variable factor subject to diminishing marginal productivity
ii. Labor as variable factor subject to diminishing marginal productivity

Assumptions:

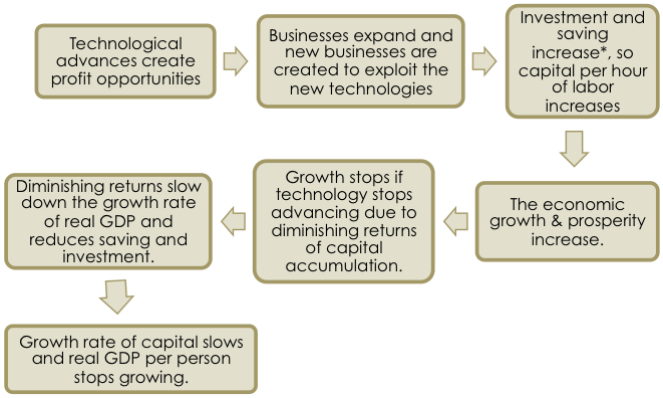
- Economic growth rate depends on the rate of technological change.
- Technological change is exogenous and results from chance.

The Basic Idea:

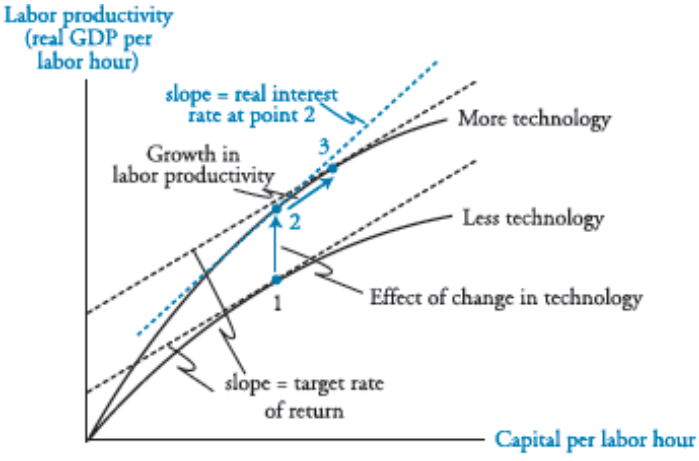
- The population growth rate is independent of real GDP and the real GDP growth rate.

Population growth rate = Birth rate – Death rate

- The birth rate is determined by the opportunity cost of a woman's time i.e. as women's wage rates ↑, the opportunity cost of having children ↑and the birth rate ↓.
- The death rate is determined by the quality and availability of health care services i.e. as the quality and availability of health care improves, the death rate ↓.
- The decrease in both the birth rate and the death rate offset each other and thus make the population growth rate independent of the level of income.



*As long as rate of return (real interest rate) > target return → people have an incentive to save. When rate of return < target return, savings decrease → resulting in decrease in investment.



Balanced or Steady State Rate of Growth in Neoclassical Growth Theory

In a closed economy: There is no international trade or capital flows; thus,

Domestic investment = Domestic savings

And

Growth in physical capital stock = $\Delta K = sY - \delta K$

where,

s = Fraction of income that is saved
sY = Gross investment → Increases in gross investment results in increase in physical capital stock.
δ = A constant rate at which the physical capital stock depreciates → Depreciation results in decrease in physical capital stock.

According to the neoclassical growth theory, an economy moves to an equilibrium position over time i.e. it reaches the balanced or steady state rate of growth.

In the steady state:

- The growth rate of capital per worker = growth rate of output per worker i.e.
  $\Delta k / k = \Delta y / y = \Delta A / A + \alpha \Delta k / k$
- The output-to-capital ratio is constant.
- Capital-to-labor ratio (k) and output per worker (y) grow at the same rate i.e.
  Growth rate of capital per worker = Growth rate of output per worker = $\frac{TFP}{1-\alpha}$ → Steady state growth rate of labor productivity
- The marginal product of capital is also constant and is equal to α (Y/K), which in turn is equal to real interest rate in the economy.
- The increase in the capital-to-labor ratio (i.e. by capital deepening) does not affect the marginal product of capital and growth rate of the economy; rather, the potential growth rate of the economy is affected by only changes in growth rates of TFP and in the labor share of output.

Growth rate of Total output = $\Delta Y / Y$
= Growth rate of TFP scaled by labor force share + Growth rate in the labor force = $\frac{\theta}{1-\alpha} + n$

Steady state Output-to-capital ratio = $\frac{Y}{K}$
$= \left(\frac{1}{s}\right)\left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right] = \Psi$

Gross investment = $\left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right]k$

Refer to: Exhibit 13, Volume 1, Reading 7

- The ***Straight line*** represents the amount of investment required to maintain the physical capital stock at the required rate.
  Slope of straight line = $[\delta + n + \theta / (1 - \alpha)]$
- The ***curved line*** represents the amount of actual investment per worker. The curve reflects diminishing marginal returns to capital.
- Steady state equilibrium occurs where the straight line intersects the curved line.
- Over time when capital-to-labor ratio rises, TFP increases, the actual investment curve shifts upward, the equilibrium moves upward and to the right along the straight line.

During the transition to the steady state growth path, the exogenous factors i.e. labor supply and TFP are fixed and

Growth rates of output per capita = $\Delta y / y$
$= \left[\left(\frac{\theta}{1-\alpha}\right) + as\left(\frac{Y}{K} - \Psi\right)\right] = \left(\frac{\theta}{1-\alpha}\right) + as\ (y/k - \Psi)$

Capital-to-labor ratio = $\Delta k / k$
$= \left[\left(\frac{\theta}{1-\alpha}\right) + s\left(\frac{Y}{K} - \Psi\right)\right] = \left(\frac{\theta}{1-\alpha}\right) + s\ (y/k - \Psi)$

When the actual saving/investment > required investment (e.g. due to low capital-to-labor ratio or high TFP).

- Output-to-capital ratio > equilibrium level
- Growth rates of output per capita and the capital-to-labor ratio will be above the steady state rate.
- Since α< 1, it indicates that growth in capital > output growth rate and the output-to-capital ratio is falling.
- However, with passage of time, the growth rates of both output per capita and the capital-to-labor ratio decline to the steady state rate.

When the actual saving/investment < required investment (e.g. due to high and unsustainable capital-to-labor ratio or low TFP).

- Output-to-capital ratio < equilibrium level
- Growth rates of output per capita and the capital-to-labor ratio will be below the steady state rate.
- Output falls.
- However, with the passage of time, output grows faster than capital and both output per capita and the capital-to-labor ratio rise to the steady state rate.

Refer to: Exhibit 15 from the CFA Institute's Curriculum 'Dynamics in the Neoclassical Model'

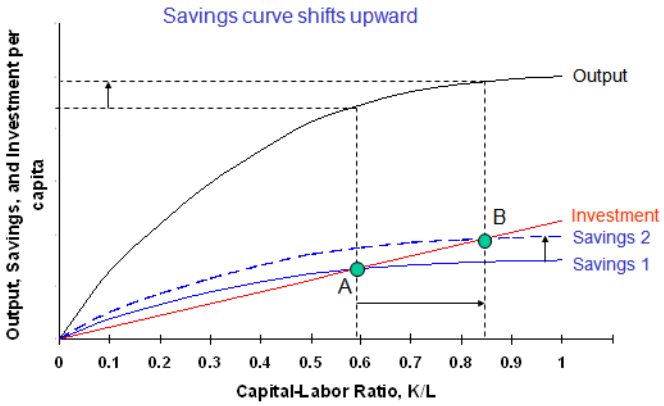
Practice: Example 10 from the CFA Institute's Curriculum.

Impact of parameters:

A. Saving rate (s): When saving rate ↑→ saving/investment at every level of output ↑, → capital-to-labor ratio and output per worker ↑→ saving/investment curve shifts ***upward*** to a new equilibrium level at higher capital-to-labor ratio and output per worker. *See exhibit 14.*

- The saving rate only changes the level of output per

worker; it does not permanently change the growth rate of output per worker i.e. the steady state growth rates of output per capita or output remain unchanged.
- However, the higher the saving rates → the higher the level of per capita output and capital-to-labor ratio, and the higher the level of labor productivity.



B. Labor force growth (n): When labor force growth rate ↑, slope of the required investment line increases → the straight line intersects the supply of saving/ investment curve at new equilibrium point with *lower* capital-to-labor and output per worker ratios.

- The labor force growth rate only changes the level of output per worker; it does not permanently change the growth rate of output per worker.

C. Depreciation rate (δ): When depreciation rate ↑, → net capital accumulation falls at a given rate of gross saving, → slope of the required investment line increases → and it intersects the supply of saving/ investment curve at new equilibrium point with *lower* capital-to-labor and output per worker ratios.

- The depreciation rate only changes the level of output per worker; it does not permanently change the growth rate of output per worker.

***D.* Growth in TFP (θ):** When growth rate of TFP ↑, → in the future, output per worker will grow faster BUT at present with a given supply of labor and a given level of TFP, output per worker will fall, → slope of the required investment line increases → and it intersects the supply of saving/ investment curve at new equilibrium point with *lower* capital-to-labor and output per worker ratios. *See exhibit 15.*

- Due to changes in TFP, the capital-to-labor ratio and output per capita are not constant even in steady state, implying that *changes in the growth rate of TFP can permanently change the growth rate of output per worker*.

Important to Note:

- When the capital-to-labor ratio increases but the output-to-capital ratio declines, a greater fraction of savings is required to maintain the capital-to-labor ratio; as a result, a smaller fraction is left for capital deepening.
- Proportional impact of the change in parameter on the capital-to-labor ratio and per capita income over time is estimated as follows:

$$\frac{k_{new}}{k_{old}} = \left[\frac{\left(\frac{Y}{K}\right)_{new}}{\left(\frac{Y}{K}\right)_{old}} \right]^{\frac{1}{\alpha-1}}$$

$$\frac{y_{new}}{y_{old}} = \left[\frac{k_{new}}{k_{old}} \right]^{\alpha}$$

11. IMPLICATIONS OF THE NEOCLASSICAL MODEL

Implications of the Neoclassical Model:

1) Higher rates of investment (capital accumulation) cannot ***permanently*** increase the rate of per capita growth in an economy i.e. per capita growth in the economy will stop increasing at some point, reaching the steady state of growth.
2) Capital deepening can raise per capita growth only when

- Economy is operating below the steady state; and
- MPK> Marginal cost of capital (MC).

3) When the rate of growth of capital stock > growth rate of labor productivity, the return to investment in an economy should decline over time.
4) Changes in saving and investment only have a transitory impact on growth i.e. steady state rate of economic growth is unrelated to the rate of saving and investment.
5) Because of diminishing marginal returns to capital, potential GDP per capita can sustainably grow only through technological change or growth in TFP.
6) Due to lack of physical capital and hence high marginal productivity of capital and potentially higher saving rates in developing countries, growth rates &

income levels per person of developing countries should converge to the developed countries.

Criticism of Neoclassical Growth Theory:

1. In the neoclassical theory, the technology is treated as exogenous factor; thus, the theory does not explicitly explain the determinants of technological progress or changes in TFP over time.
2. The historical evidence shows that convergence among countries is slow and the poor countries are not catching up.
3. It has also been observed that in developed countries, with rate of growth of capital stock > growth rate of labor productivity, the return to investment has not declined over time.

Practice: Example 11 from the CFA Institute's Curriculum.

12. EXTENSIONS OF NEOCLASSICAL MODEL

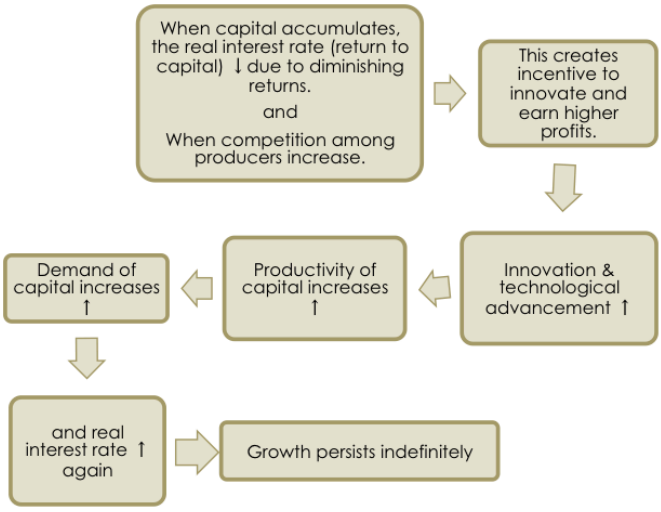
Augmented Solow Approach:

This approach is an extension of the neoclassical model. Under this approach:

- The portion of growth associated with the technological progress (TFP) is relatively small compared to neoclassical model.
- Besides physical capital, investment includes human capital, research and development, and public infrastructure.
- In addition to level of capital spending, the economic growth also depends on the ***composition*** of capital spending i.e. the higher the capital spending on high-technology goods relative to physical capital, the higher the productivity and the higher the growth.
- However, even a broadly defined capital investment is subject to diminishing marginal returns; implying that in the long-run, an economy will ultimately revert towards a steady state growth rate.

13. ENDOGENOUS GROWTH MODEL

The Basic Idea: According to new growth theory, the growth rate depends on ability of people to innovate. This implies that as long as incentives and motives of rising profit exist in an economy, growth can be sustained indefinitely i.e.



Important to Note:

Technological progress is an endogenous factor i.e. it depends on ability and willingness of people to innovate.

Inputs to the production function:

- Capital
- Labor
- Knowledge or human capital
- R&D spending

These factors of production are financed through savings.

Production function in the endogenous growth model:

$$y_e = f(k_e) = ck_e$$

where,

ye = output per worker
ke = stock of capital per worker
c = constant marginal product of capital in the aggregate economy
e = endogenous growth model

- Unlike neoclassical production function, endogenous growth production function represents a straight line.
- The output-to-capital ratio is fixed; as a result, growth rate of output per worker will always be equal to the growth rate of capital per worker.

Growth rate of output per capita = $\Delta y_e/y_e$
$= \Delta k_e/k_e = sc - \delta - n$

- This implies that *permanently higher growth rate in an economy can be achieved through a higher saving rate.*

Following Two factors play a key role in endogenous growth theory:

1. Discoveries are a public capital good.
2. R&D expenditures and human capital (i.e. knowledge) are not subject to diminishing returns i.e. increasing knowledge increase the productivity of both labor and capital; rather, they may have increasing returns to scale due to large positive externalities or spillover effects because spending by companies on R&D and knowledge capital generates benefits to the economy as a whole.

Implication of the Endogenous Growth Model:

1) Higher rates of investment (through higher savings) in capital stock (i.e. pure capital deepening), knowledge and in new, innovative products and processes can result in a permanently higher growth rates.
2) The incomes of developed and developing countries do not necessarily converge over time because developed economies with constant or even increasing returns to knowledge capital can continue to grow as fast as, or faster than, the developing countries.

According to endogenous growth theory, the increase in growth is a perpetual process.

Practice: Example 12 from the CFA Institute's Curriculum.

Differences between theories:

- According to ***classical theory***, increase in population negatively affects economic growth.
- According to ***endogenous growth theory***, increase in population positively affects economic growth because TFP progress depends on ability and willingness of people to innovate.
- According to ***classical theory***, population explosion results in decrease in real GDP.
- According to ***neo-classical theory***, diminishing returns to capital results in decrease in real GDP.
- Both ***classical and neo-classical*** theories consider technology as an exogenous factor that occurs by chance.
- Under an ***Endogenous growth theory***, technology is viewed as an endogenous factor that depends on the ability and capacity of human resources to innovate.

14. CONVERGENCE HYPOTHESIS

According to the Convergence hypothesis, over time, countries with low per capita incomes (i.e., developing countries) should grow at a faster rate than countries with high per capita incomes (i.e., developed countries); so that the per capita income in developing countries will converge toward the same level of per capita income.

Convergence between the developed and developing countries can occur in two ways:

1) Through capital accumulation and capital deepening.
2) By imitating or adopting technology developed in the advanced countries. In addition, the higher the capital spending on technological progress, the narrower the income gap between developed and developing countries.

However, the evidence on convergence is mixed.

Important to Note: If the convergence hypothesis is correct, it implies that the growth rate in per capita GDP is inversely related to the initial level of per capita real GDP.

Types of Convergence under the Neoclassical growth theory:

1) Absolute Convergence: According to absolute convergence, regardless of their particular characteristics, per capita incomes in poor countries will ***grow*** at the same rate as that of rich economies such that all economies will eventually converge to a ***common steady state.***

- However, it does not imply that the ***level*** of per capita income will be the same in all countries regardless of underlying characteristics.

2) Conditional Convergence: According to conditional convergence, countries with low per capita incomes

will catch up the countries with high per capita incomes *ONLY* if they have similar socio-economic characteristics e.g. population growth rate, savings per capita, depreciation and capital stock. Such that

- Only ***homogenous*** economies will converge to the same level of per capita output as well as the same steady state growth rate;
- While ***heterogeneous*** economies will converge to different level of per capita output and steady state growth rate, depending on their human capital endowment and other socio-economic characteristics.

Club convergence: According to club convergence, only *rich and middle-income* countries that are member of the club should converge to the income level of richest countries in the world. Under club convergence,

- The lowest per capita income countries in the club should grow at the fastest rate.
- Per capita income of Non-member countries should continue to decline.
- Poor countries can become members of the convergence club by making appropriate institutional changes e.g., appropriate legal, political, and economic institutions, labor market reforms and trade policy*.

Implication of convergence and/or Club convergence on Equity investment: If convergence and particularly club convergence does occur, then in the long-run,

- Corporate profits, earnings and stock prices in *lower* per capita incomes countries that are members of the convergence club should grow at a faster rate (note that risk will also be higher).
- This implies that in the long-run, investors can earn higher rate of return by investing in lower per capita incomes countries that are members of the club than investing in higher-income countries.

***NOTE:**

Import substitution policies may initially improve growth but if maintained for a long period, they may negatively affect growth.

15. GROWTH IN AN OPEN ECONOMY

Effects of Opening up the economy to trade and financial flows on Economic growth rate: In an open economy,

1) A country can fund its domestic investment by borrowing funds in global markets instead of just relying on domestic savings.
2) Countries can increase their overall productivity by reallocating resources into industries in which they have a comparative advantage away from industries in which they have a comparative disadvantage.
3) Companies have access to a larger, global market for their products so that they can better exploit any economies of scale and have incentives to innovate.
4) Countries can increase their rate of TFP progress by importing technology from other countries.
5) A country can increase its physical capital stock through capital inflows (i.e. by borrowing funds globally), which results in higher productivity growth rate and higher per capita incomes despite low domestic savings.

- Since capital flows must be matched by equal and offsetting trade flows, this implies that *capital-poor countries tend to run a trade deficit*.

6) As global trade increases, competition in the domestic market increases, leading to better quality and low-priced products.

According to the neoclassical model, convergence should occur *more quickly* when:

- Economies are open;
- There is no trade or capital flow restrictions;
- International borrowing and lending is allowed;

Implication of *Capital-to-labor ratios* on rate of return on investment:

- The lower (higher) the capital-to-labor ratio → the higher (lower) the marginal product of capital → the higher (lower) the rate of return on investments.
  - This implies that investors should invest in capital-poor countries to earn higher returns on investments.
- However, as physical capital stock in the capital-poor country increases over time → the return on investments reduces → the rate of investment declines → the size of the country's trade deficit declines → the growth rate will slow down and will revert toward the steady state rate of growth.
  - ➔Consequently, investment < level of domestic savings → country's trade deficit will convert into a

trade surplus and will become a capital-exporting country.

Neoclassical model v/s Endogenous growth model with respect to an open economy:

- In the Solow or neoclassical model, opening up an economy to trade and financial flows will not cause any increase in the rate of growth in an economy i.e. countries will always grow at the steady state rate of growth.
- In contrast, in the Endogenous growth model, opening up an economy to trade and financial flows can permanently increase the rate of economic growth.

Under Endogenous Growth Model, increase in global trade positively affects global output in following three ways:

a) Selection Effect: When due to increased competition from foreign companies, less efficient domestic companies exit the market whereas more efficient domestic companies innovate and discover new technologies to lower their costs and increase profits, the efficiency of the overall national economy tends to increase. This is referred to as selection effect.

b) Scale Effect: When opening up an economy provides companies an access to a larger, global market for their products, they are better able to fully exploit economies of scale and have incentives to innovate, such that spending on R&D and human capital increases and causes the efficiency of the overall national economy to increase. This is referred to as scale effect.

- Typically, scale effect tends to benefit smaller countries.

c) Backwardness Effect: When opening up an economy facilitates less advanced countries or sectors of an economy to import or imitate technology developed in more advanced countries or sectors, it generates knowledge spillover effects and is referred to as backwardness effect.

- Typically, backwardness effect tends to benefit poorer, less developed countries.

NOTE:

However, trade may also hurt growth of countries (particularly small countries) that lack TFP progress.

Practice: Example 13 from the CFA Institute's Curriculum.

Two contrasting strategies for economic development:

1) Inward-oriented policies: Policies that restrict imports to develop and/or support domestic industries and put limits on investment from abroad are referred to as inward-oriented policies. These policies promote production of domestic substitutes despite their higher production costs. These policies are also known as *import substitution policies*.

2) Outward-oriented policies: Policies that focus on promoting integration with the world economy by promoting exports, eliminating trade restrictions and attracting foreign investments are referred to as outward-oriented policies. These policies are basically trade-oriented policies and often referred to as ***export-led growth strategies***. It has been evidenced that countries that pursue outward-oriented policies tend to:

- Have high rates of GDP growth and convergence with developed countries compared to countries that pursue inward-oriented policies.
- Enjoy positive effects of foreign direct investment.

Practice: Example 14 & 15 from the CFA Institute's Curriculum.

Practice: CFA Institute's Curriculum End of Chapter Questions & FinQuiz Question-bank (Item-sets + Questions)