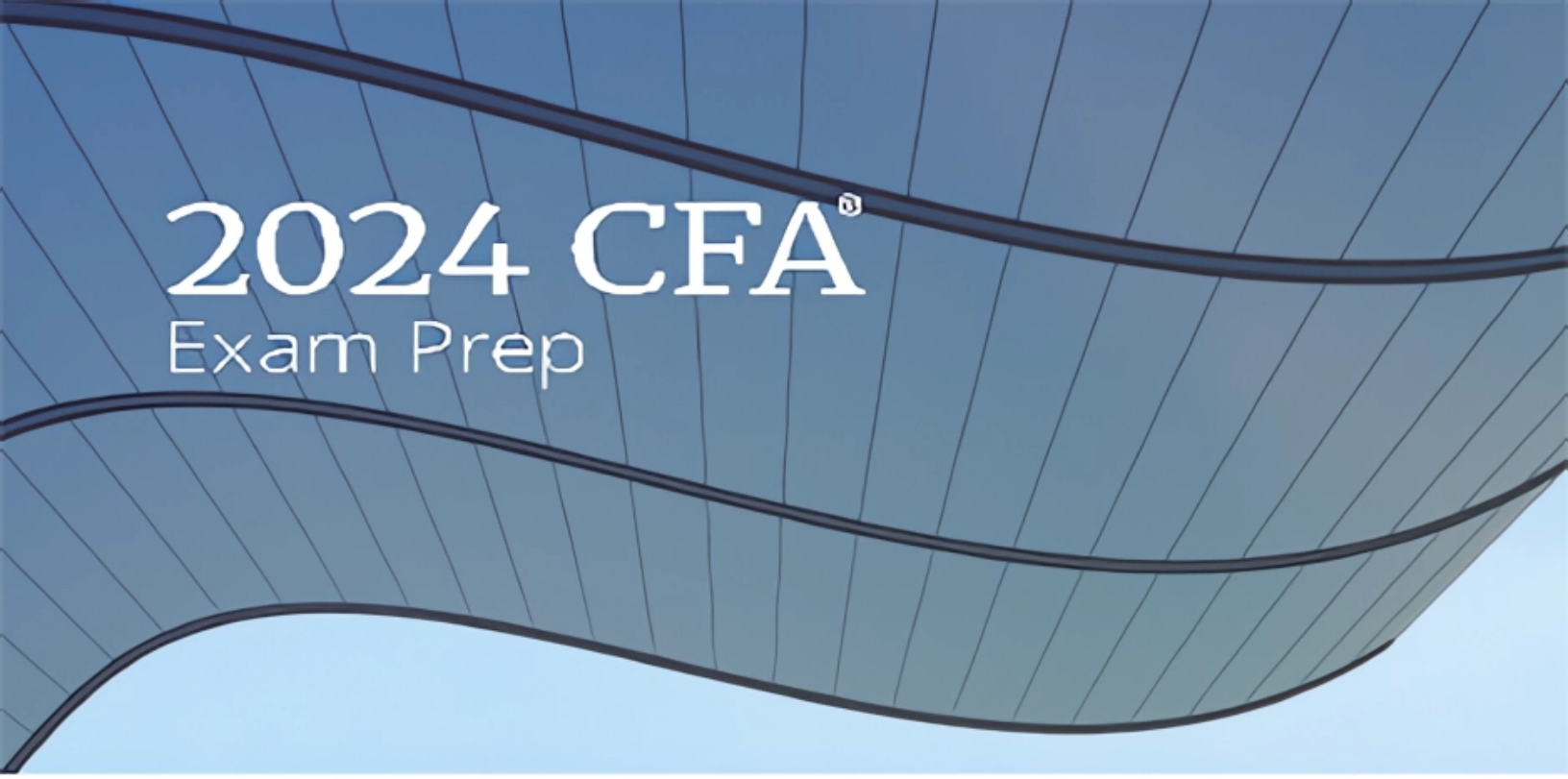
2024 CFA®
Exam Prep

Schweser's Secret Sauce®

LEVEL II



KAPLAN SCHWESER

Schweser's Secret Sauce®

Level II CFA®

2024

SCHWESER'S SECRET SAUCE®: 2024 Level II CFA®

Published in 2024 by Kaplan, Inc.
ISBN: 978-1-0788-3653-1

CONTENTS

FOREWORD

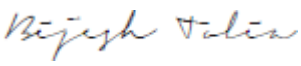
Secret Sauce® offers concise and readable explanations of the major ideas in the Level II CFA curriculum.

This book does not cover every Learning Outcome Statement (LOS) and, as you are aware, any LOS is "fair game" for the exam. We focus here on those LOS that are core concepts, have application to other LOS, are complex and difficult for candidates, or require memorization of characteristics or relationships.

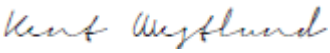
Secret Sauce is easy to carry with you and will allow you to study these key concepts, definitions, and techniques over and over, an important part of mastering the material. When you get to topics where the coverage here appears too brief or raises questions in your mind, this is your cue to go back to your **SchweserNotes** to fill in the gaps in your understanding. There is no shortcut to learning the vast breadth of subject matter covered by the Level II curriculum, but this volume will be a valuable tool for reviewing the material as you progress in your studies over the months leading up to exam day.

Pass rates remain around 50%, and returning Level II candidates make comments such as, "I was surprised at how difficult the exam was." You should not despair because of this, but more importantly, do not underestimate the challenge. Our study materials, mock exams, question bank, videos, classes, and Secret Sauce are all designed to help you study as efficiently as possible, grasp and retain the material, and apply it with confidence on exam day.

Best regards,



Dr. Bijesh Tolia, CFA, CA
Vice President of CFA Education
and Level II Manager
Kaplan Schweser



Kent Westlund, CFA, CPA
Senior Content Specialist

QUANTITATIVE METHODS

Topic Weight on Exam	5%–10%
SchweserNotes™ Reference	Book 1, Pages 1–113

Multiple regression and machine learning are key topics for the Level II exam. For the time series material, the concepts of nonstationarity, unit roots (i.e., random walks), and serial correlation will be important, as well as being able to calculate the mean-reverting level of an autoregressive (AR) time-series model. Understand the implications of seasonality and how to detect and correct it, as well as the root mean squared error (RMSE) as a model evaluation criterion.

MULTIPLE REGRESSION

Cross-Reference to CFA Institute Assigned Reading #1

Multiple regression is a rather important part of the quant material.

You should understand the interpretation of a regression output, how a reduced versus a full model is evaluated using an *F*-test, and the impact of influential observations on the regression results. You should also understand the effect that heteroskedasticity, serial correlation, and multicollinearity have on regression results.

A regression of a dependent variable (e.g., sales) on three independent variables would yield an equation like the following:

$$Y_i = b_0 + (b_1 \times X_{1i}) + (b_2 \times X_{2i}) + (b_3 \times X_{3i}) + \varepsilon_i$$

Multiple Regression: Model Fit

ANOVA is a statistical procedure that attributes the variation in the dependent variable to one of two sources: the regression model or the residuals (i.e., the error term). The structure of an ANOVA table is shown in Figure 1.

Figure 1: Analysis of Variance (ANOVA) Table

Source	df (Degrees of Freedom)	SS (Sum of Squares)	MS (Mean Square = SS/df)
Regression	k	RSS	$MSR = \frac{RSS}{k}$
Error	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$
Total	$n - 1$	SST	

Note that RSS + SSE = SST. The information in an ANOVA table can be used to calculate R^2, the *F*-statistics, and the standard error of estimate (SEE).

The *coefficient of determination* (R^2) is the percentage of the variation in the dependent variable explained by the independent variables.

$$R^2 = \frac{\text{regression sum of squares (RSS)}}{\text{total sum of squares (SST)}}$$

$$= \frac{\text{SST} - \text{sum of squared errors (SSE)}}{\text{SST}}$$

In multiple regression, you also need to understand *adjusted* R^2. The adjusted R^2 provides a measure of the goodness of fit that adjusts for the number of independent variables included in the model.

Akaike's information criterion (AIC) and **Schwarz's Bayesian information criterion (BIC)** are used to evaluate competing models with the same dependent variable. Lower values indicate a better model under either criteria. AIC is used if the goal is a better forecast, while BIC is used if the goal is a better goodness of fit.

$$\text{AIC} = n \times \ln(\text{SSE}/n) + 2(k+1)$$
$$\text{BIC} = n \times \ln(\text{SSE}/n) + \ln(n) \times (k+1)$$

where:

k = the number of independent variables

Joint Hypothesis Tests

Nested models comprise a full (or unrestricted) model, and a restricted model that uses "q" fewer independent variables. To test whether the "q" excluded variables add to the explanatory power of the model, we test the hypothesis:

$H_0: b_i = b_j = \ldots b_q = 0$
vs.
H_a: at least one of the slope coefficients of excluded variables $\neq 0$

We calculate the *F*-statistic to test this hypothesis as:

$$F = \frac{(\text{SSE}_R - \text{SSE}_U)/q}{\text{SSE}_U/(n-k-1)} \text{ with } q \text{ and } (n-k-1) \text{ degrees of freedom}$$

where:

R and U represent the restricted and unrestricted models, respectively

q = number of excluded variables in the restricted model

k = number of independent variables in the unrestricted model

Decision rule: reject H_0 if *F* (test-statistic) > F_c (critical value)

Tests of all coefficients collectively. For this test, the null hypothesis is that all the slope coefficients simultaneously equal zero. The required test is a one-tailed *F*-test and the calculated statistic is:

$$F = \frac{\text{regression mean square (MSR)}}{\text{mean squared error (MSE)}} \text{ with } k \text{ and } n-k-1 \text{ df}$$

Rejection of the null hypothesis at a stated level of significance indicates that at least one of the coefficients is significantly different than zero, which is interpreted to mean that at least one of the independent variables in the regression model makes a significant contribution to the explanation of the dependent variable.

Potential Problems in Regression Analysis

You should be familiar with the three violations of the assumptions of multiple regression and their effects.

Figure 2: Violation of Regression Assumptions

Violation	Conditional Heteroskedasticity	Serial Correlation	Multicollinearity
What is it?	Residual variance is related to level of independent variables	Residuals are correlated with each other	Two or more independent variables are highly correlated
Effect?	Coefficients are consistent. Standard errors are underestimated. Too many Type I errors	Coefficients are consistent. Standard errors are underestimated. Too many Type I errors (positive correlation)	Coefficients are consistent (but unreliable). Standard errors are overestimated. Too many Type II errors
Detection?	Breusch–Pagan chi-square test	Breusch–Godfrey (BG) *F*-test	Conflicting *t* and *F*-statistics; high variance inflation factors (VIF)
Correction?	Use robust or White-corrected standard errors	Use robust or Newey–West corrected standard errors	Drop one of the correlated variables, or use a different proxy for an included independent variable

Model Misspecification

Figure 3 shows types of model misspecifications and their impact on the regression results.

Figure 3: Model Misspecifications

Misspecification	Description	Effect
Omission of important independent variable(s)	One or more variables that should have been included have been omitted	Biased and inconsistent regression parameters May lead to serial correlation or heteroskedasticity in the residuals
Inappropriate transformation	Linear model trying to fit nonlinear relationships	May lead to heteroskedasticity in the residuals
Inappropriate variable scaling	Variables are not transformed appropriately	May lead to heteroskedasticity in the residuals or multicollinearity
Data improperly pooled	Sample has periods of dissimilar economic environments (i.e., the slope coefficients are unstable)	May lead to heteroskedasticity or serial correlation in the residuals

Influence Analysis and Influential Data Points

Outliers are extreme observations of the dependent or "*Y*" variable, while high-leverage points are the extreme observations of the independent or "*X*" variables. **Influential data points** are extreme observations that when excluded cause a significant change in model coefficients. Influential data points cause the model to perform poorly out of sample.

Cook's D values greater than $\sqrt{k/n}$ indicate that the observation is highly likely to be an influential data point. Influential data points should be checked for input errors; alternatively, the observation may be valid but the model incomplete.

Qualitative Independent Variables

Qualitative independent variables (dummy variables) capture the effect of binary independent variables. When we want to distinguish between n classes, we must use (n – 1) dummy variables. Otherwise, we would violate the regression assumption of no exact linear relationship between independent variables.

A dummy variable can be an intercept dummy, or a slope dummy, or a combination of the two.

Logistic Regression Models

Qualitative dependent variables (e.g., bankrupt versus non-bankrupt) require methods other than ordinary least squares. **Logistic regression (logit) models** use log odds as the dependent variable, and the coefficients are estimated using the maximum likelihood estimation methodology.

TIME-SERIES ANALYSIS

Cross-Reference to CFA Institute Assigned Reading #2

Types of Time Series

Linear Trend Model

The typical time series uses time as the independent variable to estimate the value of time series (the dependent variable) in period *t*:

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

The predicted change in y is b_1 and t = 1, 2, …, T

Trend models are limited in that they assume time explains the dependent variable. Also, they tend to be plagued by various assumption violations. The Durbin-Watson test statistic can be used to check for serial correlation. A linear trend model may be appropriate if the data points seem to be equally distributed above and below the line and the mean is constant. Growth in GDP and inflation levels are likely candidates for linear models.

Log-Linear Trend Model

Log-linear regression assumes the dependent financial variable grows at some constant rate:

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

$$\ln(y_t) = \ln(e^{b_0 + b_1(t)}) \Rightarrow b_0 + b_1(t)$$

The log-linear model is best for a data series that exhibits a trend or for which the residuals are correlated or predictable or the mean is non-constant. Most of the data related to investments have some type of trend and thus lend themselves more to a log-linear model. In addition, any data that have seasonality are candidates for a log-linear model. Recall that any exponential growth data call for a log-linear model.

The use of the transformed data produces a linear trend line with a better fit for the data and increases the predictive ability of the model. Because the log-linear model more accurately captures the behavior of the time series, the impact of serial correlation in the error terms is minimized.

Autoregressive (AR) Model

In AR models, the dependent variable is regressed against previous values of itself.

An autoregressive model of order *p* can be represented as:

$$x_t = b_0 + b_1x_{t-1} + b_2x_{t-2} + \ldots + b_px_{t-p} + \varepsilon_t$$

There is no longer a distinction between the dependent and independent variables (i.e., *x* is the only variable). An AR(p) model is specified correctly if the autocorrelations of residuals from the model are not statistically significant at any lag.

When testing for serial correlation in an AR model, don't use the Durbin-Watson statistic. Use a *t*-test to determine whether any of the correlations between residuals at any lag are statistically significant.

If some are significant, the model is incorrectly specified and a lagged variable at the indicated lag should be added.

Chain Rule of Forecasting

Multiperiod forecasting with AR models is done one period at a time, where risk increases with each successive forecast because it is based on previously forecasted values. The calculation of successive forecasts in this manner is referred to as the *chain rule of forecasting*. A one-period-ahead forecast for an AR(1) model is determined in the following manner:

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

Likewise, a 2-step-ahead forecast for an AR(1) model is calculated as:

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

Covariance Stationary

Statistical inferences based on an autoregressive time series model may be invalid unless we can make the assumption that the time series being modeled is covariance stationary. A time series is covariance stationary if it satisfies the following three conditions:

1. Constant and finite mean.
2. Constant and finite variance.
3. Constant and finite covariance with leading or lagged values.

To determine whether a time series is covariance stationary, we can:

- Plot the data to see if the mean and variance remain constant.
- Perform the Dickey-Fuller test (which is a test for a unit root, or if $b_1 - 1$ is equal to zero).

If the times series does not satisfy these conditions, we say it is not covariance stationary, or that there is nonstationarity. Most economic and financial time series relationships are not stationary. The degree of nonstationarity depends on the length of the series and the underlying economic and market environment and conditions.

For an AR(1) model to be covariance stationary, the mean reverting level must be defined. Stated differently, b_1 must be less than one.

If the AR model is not covariance stationary, we can often correct it with first differencing.

Mean Reversion

A time series is mean reverting if it tends towards its mean over time. The mean reverting level for an AR(1) model is $\frac{b_0}{(1 - b_1)}$.

The value of the variable tends to fall when above its mean and rise when below its mean.

Unit Root

If the value of the lag coefficient is equal to one, the time series is said to have a unit root and will follow a random walk process. A series with a unit root is not covariance stationary. Economic and finance time series frequently have unit roots. First differencing will often eliminate the unit root. If there is a unit root, this period's value is equal to last period's value plus a random error term and the mean reverting level is undefined.

Random Walk

A random walk time series is one for which the value in one period is equal to the value in another period, plus a random (unpredictable) error. If we believe a time series is a random walk (i.e., has a unit root), we can transform the data to a covariance stationary time series using a procedure called first differencing.

Random walk without a drift: $x_t = x_{t-1} + \varepsilon_t$

Random walk with a drift: $x_t = b_0 + x_{t-1} + \varepsilon_t$

In either case, the mean reverting level is undefined ($b_1 = 1$), so the series is not covariance stationary.

First Differencing

The first differencing process involves subtracting the value of the time series in the immediately preceding period from the current value of the time series to define a new variable, *y*. If the original time series has a unit root, this means we can define y_t as:

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

Then, stating *y* in the form of an AR(1) model:

$$y_t = b_0 + b_1 y_{t-1} + \varepsilon_t$$

where:

$$b_0 = b_1 = 0$$

This transformed time series has a finite mean-reverting level of $\frac{0}{1 - 0} = 0$ and is, therefore, covariance stationary.

First differencing can remove a trend in the data and result in a covariance stationary series.

PROFESSOR'S NOTE

By taking first differences, you model the *change* in the value of the variable rather than the value of the variable.

Seasonality

Seasonality in a time series is tested by calculating the autocorrelations of error terms. A statistically significant lagged error term may indicate seasonality. To adjust for seasonality in an AR model, an additional lag of the variable (corresponding to the statistically significant lagged error term) is added to the original model. Usually, if quarterly data are used, the seasonal lag is 4; if monthly data are used, the seasonal lag is 12. If a seasonal lag coefficient is appropriate and corrects the seasonality, a revised model incorporating the seasonal lag will show no statistical significance of the lagged error terms.

Assessing Forecast Accuracy With Root Mean Squared Error

Root mean squared error (RMSE) is used to assess the predictive accuracy of autoregressive models. For example, you could compare the results of an AR(1) and an AR(2) model. The RMSE is the square root of the average (or mean) squared error. The model with the lower RMSE is better.

Out-of-sample forecasts predict values using a model for periods beyond the time series used to estimate the model. The RMSE of a model's out-of-sample forecasts should be used to compare the accuracy of alternative models.

Structural Change (Coefficient Instability)

Estimated regression coefficients may change from one time period to another. There is a trade-off between the statistical reliability of using a long time series and the coefficient stability of a short time series. You need to ask: Has the economic process or environment changed?

A structural change is indicated by a significant shift in the plotted data at a point in time that seems to divide the data into two distinct patterns. When this is the case, you have to run two different models, one incorporating the data before and one after that date, and test whether the time series has actually shifted. If the time series has shifted significantly, a single time series encompassing the entire period (i.e., encompassing both patterns) will likely produce unreliable results, so the model using more recent data may be more appropriate.

Cointegration

Cointegration means that two time series are economically linked (related to the same macro variables) or follow the same trend and that relationship is not expected to change. If two time series are cointegrated, the error term from regressing one on the other is covariance stationary and the *t*-tests are reliable.

To test whether two time series are cointegrated, we regress one variable on the other using the following model:

$$y_t = b_0 + b_1 x_t + \varepsilon$$

where:
y_t = value of time series y at time t
x_t = value of time series x at time t

The residuals are tested for a unit root using the Dickey-Fuller test with critical *t*-values calculated by Engle and Granger (i.e., the DF-EG test). If the test *rejects* the null hypothesis of a unit root, we say the error terms generated by the two time series are covariance stationary and the two series are cointegrated. If the two series are cointegrated, we can use the regression to model their relationship.

Occasionally, an analyst will run a regression using two time series (i.e., two time series with different variables). For example, to use the market model to estimate the equity beta for a stock, the analyst regresses a time series of the stock's returns on a time series of returns for the market.

- If both time series are covariance stationary, model is reliable.
- If only the dependent variable time series or only the independent time series is covariance stationary, the model is not reliable.
- If neither time series is covariance stationary, you need to check for cointegration.

Autoregressive Conditional Heteroskedasticity (ARCH)

ARCH describes the condition where the variance of the residuals in one time period within a time series is dependent on the variance of the residuals in another period. When this condition exists, the standard errors of the regression coefficients in AR models and the hypothesis tests of these coefficients are invalid.

The ARCH(1) regression model is expressed as:

$$\hat{\varepsilon}_t^2 = a_0 + a_1 \hat{\varepsilon}_{t-1}^2 + \mu_t$$

If the coefficient, a_1, is statistically different from zero, the time series is ARCH(1).

If a time-series model has been determined to contain ARCH errors, regression procedures that correct for heteroskedasticity, such as generalized least squares, must be used in order to develop a predictive model. Otherwise, the standard errors of the model's coefficients will be incorrect, leading to invalid conclusions.

However, if a time series has ARCH errors, an ARCH model can be used to predict the variance of the residuals in following periods. For example, if the data exhibit an ARCH(1) pattern, the ARCH(1) model can be used in period t to predict the variance of the residuals in period t + 1:

$$\hat{\sigma}_{t+1}^2 = \hat{a}_0 + \hat{a}_1 \hat{\varepsilon}_t^2$$

Summary: The Time-Series Analysis Process

The following steps provide a summary of the time-series analysis process. Note that you may not need to go through all nine steps. For example, notice that by step C, if there is no seasonality or structural change and the residuals do not exhibit serial correlation, the model is appropriate.

Step A: Evaluate the investment situation you are analyzing and select a model. If you choose a time series model, follow steps B through I.

Step B: Plot the data and check that it is covariance stationarity. Signs of nonstationarity include linear trend, exponential trends, seasonality, or a structural change in the data.

Step C: If no seasonality or structural change, decide between a linear or log-linear model.

- Calculate the residuals.
- Check for serial correlation using the Durbin-Watson statistic.
- If no serial correlation, model is appropriate to use.

Step D: If you find serial correlation, prepare to use an auto regressive (AR) model by making it covariance stationary. This includes:

- Correcting for a linear trend—use first differencing.
- Correcting for an exponential trend—take natural log and first difference.
- Correcting for a structural shift—estimate the models before and after the change.
- Correcting for seasonality—add a seasonal lag (see step G).

Step E: After the series is covariance stationary, use an AR(1) model to model the data.

- Test residuals for significant serial correlations.
- If no significant correlation, model is okay to use.

Step F: If the residuals from the AR(1) exhibit serial correlation, use an AR(2) model.

- Test residuals for significant serial correlations.
- If no significant correlation, model is okay to use.
- If significant correlation found, keep adding to the AR model until there is no significant serial correlation.

Step G: Check for seasonality.

- Plot data.
- Check seasonal residuals (autocorrelations) for significance.
- If residuals are significant, add the appropriate lag (e.g., for monthly data, add the 12th lag of the time series).

Step H: Check for ARCH.

Step I: Test the model on out-of-sample data.

MACHINE LEARNING

Cross-Reference to CFA Institute Assigned Reading #3

Supervised Machine Learning, Unsupervised Machine Learning, and Deep Learning

The goal of machine learning is use data to automate decision-making.

- **Supervised learning.** Inputs and outputs are identified for the computer, and the algorithm uses this labeled training data to model relationships.
- **Unsupervised learning.** The computer is not given labeled data; rather, it is provided unlabeled data that the algorithm uses to determine the structure of the data.
- **Deep learning algorithms.** Algorithms such as neural networks and reinforced learning learn from their own prediction errors and are used for complex tasks such as image recognition and natural language processing.

Overfitting and Methods of Addressing It

In supervised learning, overfitting results from having a large number of independent variables (features), resulting in an overly complex model which may have generalized random noise that improves in-sample forecasting accuracy. However, overfit models do not generalize well to new data (i.e., low out-of-sample R-squared).

To reduce the problem of overfitting, data scientists use **complexity reduction** and **cross validation**. In complexity reduction, a penalty is imposed to exclude features that are not meaningfully contributing to out-of-sample prediction accuracy. This penalty value increases with the number of independent variables used by the model.

Supervised Machine Learning Algorithms

Supervised learning algorithms include the following:

1. **Penalized regression.** This reduces overfitting by imposing a penalty on—and reducing—the nonperforming features.
2. **Support vector machine.** This is a linear classification algorithm that separates the data into one of two possible classifiers based on a model-defined hyperplane.
3. **K-nearest neighbor.** This is used to classify an observation based on nearness to the observations in the training sample.
4. **Classification and regression tree.** This is used for classifying categorical target variables when there are significant nonlinear relationships among variables.
5. **Ensemble learning.** This combines predictions from multiple models, resulting in a lower average error rate.
6. **Random forest.** This is a variant of the classification tree whereby a large number of classification trees are trained using data bagged from the same data set.

Unsupervised Machine Learning Algorithms

Unsupervised learning algorithms include the following:

1. **Principal components analysis.** This summarizes the information in a large number of correlated factors into a much smaller set of uncorrelated factors called

eigenvectors.

2. **K-means clustering.** This partitions observations into *k* non-overlapping clusters; a centroid is associated with each cluster.
3. **Hierarchical clustering.** This builds a hierarchy of clusters without any predefined number of clusters.

Neural Networks, Deep Learning Nets, and Reinforcement Learning

Neural networks comprise an input layer, hidden layers (which process the input), and an output layer. The nodes in the hidden layer are called neurons, which comprise a summation operator (that calculates a weighted average) and an activation function (a nonlinear function).

Deep learning nets are neural networks with multiple hidden layers, useful for pattern, speech, and image recognition.

Reinforcement learning agents seek to learn from their own errors maximizing a defined reward.

BIG DATA PROJECTS

Cross-Reference to CFA Institute Assigned Reading #4

Steps in a Data Analysis Project

The steps involved in a data analysis project include the following: conceptualization of the modeling task, data collection, data preparation and wrangling, data exploration, and model training.

Preparing and Wrangling Data

Data cleansing deals with missing, invalid, inaccurate, and non-uniform values, as well as with duplicate observations. **Data wrangling** or preprocessing includes data transformation and scaling. Data transformation types include extraction, aggregation, filtration, selection, and conversion of data. **Scaling** is the conversion of data to a common unit of measurement. Common scaling techniques include normalization and standardization. **Normalization** scales variables between the values of 0 and 1, while standardization centers the variables at a mean of 0 and a standard deviation of 1. Unlike normalization, standardization is not sensitive to outliers, but it assumes that the variable distribution is normal.

Data Exploration

Data exploration involves exploratory data analysis, feature selection, and feature engineering. Exploratory data analysis looks at summary statistics describing the data and any patterns or relationships that can be observed. Feature selection involves choosing only those features that meaningfully contribute to the model's predictive power. Feature engineering optimizes the selected features.

Model Training

Before model training, the model is conceptualized: machine learning engineers work with domain experts to identify data characteristics and relationships. Machine learning seeks to identify patterns in the training data such that the model is able to generalize to out-of-sample data. Model fitting errors can be caused by using a small training sample or by using an inappropriate number of features. Too few features may underfit the data, while too many features can lead to the problem of overfitting.

Model training involves model selection, model evaluation, and tuning.

Preparing, Wrangling, and Exploring Text-Based Data for Financial Forecasting

Text processing involves removing HTML tags, punctuations, numbers, and white spaces. Text is then normalized by lowercasing of words, removal of stop words, and stemming/lemmatization. Text wrangling involves tokenization of text. **N-grams** is a technique that defines a token as a sequence of words and is applied when the sequence is important. A **bag-of-words (BOW)** procedure then collects all the tokens in a document. A **document term matrix** organizes text as structured data: documents are represented by words and tokens by columns. Cell values reflect the number of times a token appears in a document.

Extracting, Selecting, and Engineering Features From Textual Data

Summary statistics for textual data includes term frequency and co-occurrence. A word cloud is a visual representation of all the words in a BOW such that words with higher frequency have a larger font size. This allows the analyst to determine which words are contextually more important. Feature selection can use tools such as document frequency, a Chi-square test, and mutual information. Feature engineering for text data includes identification of numbers, usage of N-grams, **name entity recognition (NER)**, or **parts-of-speech (POS)** tokenization.

Evaluating the Fit of a Machine Learning Algorithm

Model performance can be evaluated by using error analysis. For a classification problem, a confusion matrix is prepared and evaluation metrics such as precision, recall, accuracy score, and F1 score are calculated:

precision (P) = true positives / (false positives + true positives)

recall (R) = true positives / (true positives + false negatives)

accuracy = (true positives + true negatives) / (all positives and negatives)

F1 score = $(2 \times P \times R) / (P + R)$

The **receiver operating characteristic (ROC)** plots a curve showing the tradeoff between false positives and true positives.

Root mean squared error (RMSE) is used when the target variable is continuous.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Predicted}_i - \text{actual}_i)^2}{n}}$$

Model tuning involves balancing bias error versus variance error, and selecting the optimal combination of hyperparameters.

ECONOMICS

Topic Weight on Exam	5%–10%
SchweserNotes™ Reference	Book 1, Pages 115–187

CURRENCY EXCHANGE RATES: UNDERSTANDING EQUILIBRIUM VALUE

Cross-Reference to CFA Institute Assigned Reading #5

Currency Cross Rates

A *cross rate* is the rate of exchange between two currencies implied by their exchange rates with a common third currency.

Suppose we are given three currencies: A, B, and C. We can have three pairs of currencies (i.e., A/B, A/C, and B/C).

Rules:

$$\left(\frac{A}{C}\right)_{bid} = \left(\frac{A}{B}\right)_{bid} \times \left(\frac{B}{C}\right)_{bid}$$

$$\left(\frac{A}{C}\right)_{offer} = \left(\frac{A}{B}\right)_{offer} \times \left(\frac{B}{C}\right)_{offer}$$

To flip quotes, take reciprocals; recognize that bid is the reciprocal of offer and vice versa:

$$\left(\frac{B}{C}\right)_{bid} = \frac{1}{\left(\frac{C}{B}\right)_{offer}}$$

$$\left(\frac{B}{C}\right)_{offer} = \frac{1}{\left(\frac{C}{B}\right)_{bid}}$$

To calculate the profits from a *triangular arbitrage*, follow the process demonstrated in the example below.

EXAMPLE: Triangular arbitrage

The following quotes are available from the interbank market:

Quotes:

USD/EUR 1.271 – 1.272

EUR/GBP 1.249 – 1.250