UWorld | FINANCE

SimpleSheets™+

CFA® Exam Formulas | Level 2

2025 Edition

SimpleSheets

Formulas at Your Fingertips

Quantitative Methods

Multiple Regression

- Multiple Regression Equation

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

where:

Y_i = the ith observation of the dependent variable Y
X_{ji} = the ith observation of the independent variable X_j, $j = 1,2,\ldots, k$
b_0 = the intercept of the equation
$b_1, \ldots, b_k$ = the slope coefficients for each of the independent variables
ε_i = the error term for the ith observation
n = the number of observations

Residual Term

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \ldots + \hat{b}_kX_{ki})$$

- *F*-statistic

$$F\text{-stat} = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]}$$

- Evaluating Regression Model Fit

$$R^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

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

$$AIC = n \ln\left(\frac{\text{Sum of squares error}}{n}\right) + 2(k + 1)$$

$$BIC = n \ln\left(\frac{\text{Sum of squares error}}{n}\right) + \ln(n)(k + 1)$$

- Testing Joint Hypotheses for Coefficients

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

Testing for Heteroskedasticity—The Breusch-Pagan (BP) Test

$\chi^2 = nR^2$ with k degrees of freedom.

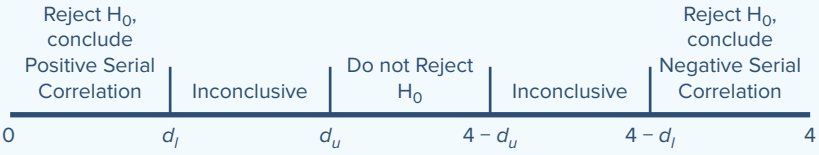
n = Number of observations
R^2 = Coefficient of determination of the **second regression** (the regression when the squared residuals of the original regression are regressed on the independent variables)
k = Number of independent variables

Testing for Serial Correlation–The Durbin-Watson (DW) Test

DW ≈ 2(1 − r); where r is the sample correlation between squared residuals from one period and those from the previous period.

Value of Durbin-Watson Statistic

(H_0: No serial correlation)



Detecting Multicollinearity

$VIF_J = 1/(1 - R_J^2)$.

- $VIF_J > 5$ warrants further investigation of the given independent variable.
- $VIF_J > 10$ indicates serious multicollinearity requiring correction.

Problems in Linear Regression and Solutions

Problem	Effect	Solution
Heteroskedasticity	Incorrect standard errors	Use robust standard errors (corrected for conditional heteroskedasticity)
Serial correlation	Incorrect standard errors (additional problems if a lagged value of the dependent variable is used as an independent variable)	Use robust standard errors (corrected for serial correlation)
Multicollinearity	High R^2 and low t-statistics	Remove one or more independent variables; often no solution based in theory

- Influence Analysis

  Studentized Residual

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

  In the equivalent formula (on the right) the terms are based on the initial regression with n observations where:
  e_i^* is the residual with the i^{th} observation deleted
  s_{e^*} is the standard deviation of the residuals
  k is the number of independent variables
  MSE_i is the mean squared error of the regression with the ith observation eliminated
  h_{ii} is the leverage value for the ith observation

- Cook's Distance

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

  Where:
  e_i^* is the residual for observation i
  k is the number of independent variables
  MSE is the mean squared error of the estimated regression model
  h_{ii} is the leverage value for observation i

  Practical guidelines for using Cook's D are the following:

  - If D_i is greater than 0.5 the i^{th} observation may be influential and merits further investigation.
  - If D_i is greater than 1.0 observation is highly likely to be an influential data point.
  - If $D_i > \sqrt{k/n}$ the i^{th} observation is highly likely to be an influential data point.

- Logit Model

  $$\ln\left(\frac{p}{1 - p}\right) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$$

- Event Probability

  $$p = \frac{1}{1 + \exp[-(b_0 + b_1X_1 + b_2X_2 + b_3X_3)]}$$

Time-Series Analysis

- Linear Trend Models

  $$y_t = b_0 + b_1t + \varepsilon_t, \qquad t = 1, 2, \ldots, T$$

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

- Log-Linear Trend Models

  A series that grows exponentially can be described using the following equation:

  $$y_t = e^{b_0 + b_1t}$$

  where:
  y_t = the value of the time series at time t (value of the dependent variable)
  b_0 = the y-intercept term
  b_1 = the slope coefficient
  t = time = 1, 2, 3, ..., T

  We take the natural logarithm of both sides of the equation to arrive at the equation for the log-linear model:

  $$\ln y_t = b_0 + b_1t + \varepsilon_t, \qquad t = 1, 2, \ldots, T$$

- Autoregressive (AR) Time-Series Models

  A first-order autoregressive model is represented as:

  $$x_t = b_0 + b_1x_{t-1} + \varepsilon_t$$

  A pth order autoregressive model is represented as:

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

- Detecting Serially Correlated Errors in an AR Model

  $$t\text{-stat} = \frac{\text{Residual autocorrelation for lag}}{\text{Standard error of residual autocorrelation}}$$

  where:
  Standard error of residual autocorrelation = $1/\sqrt{T}$
  T = Number of observations in the time series

- Mean Reversion

  $$x_t = \frac{b_0}{1 - b_1}$$

- Multiperiod Forecasts and the Chain Rule of Forecasting

  $$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1x_t$$

- Random Walks

  $x_t = x_{t-1} + \varepsilon_t$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t\varepsilon_s) = 0$ if $t \neq s$

  The first difference of the random walk equation is given as:

  $y_t = x_t - x_{t-1} = x_{t-1} + \varepsilon_t - x_{t-1} = \varepsilon_t$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$

- Random Walk with a Drift

  $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$
  $b_1 = 1$, $b_0 \neq 0$, or
  $x_t = b_0 + x_{t-1} + \varepsilon_t$, $E(\varepsilon_t) = 0$

  The first-difference of the random walk with a drift equation is given as:
  $y_t = x_t - x_{t-1}$, $y_t = b_0 + \varepsilon_t$, $b_0 \neq 0$

- The Unit Root Test of Nonstationarity

  $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$
  $x_t - x_{t-1} = b_0 + b_1x_{t-1} - x_{t-1} + \varepsilon_t$
  $x_t - x_{t-1} = b_0 + (b_1 - 1)x_{t-1} + \varepsilon_t$
  $x_t - x_{t-1} = b_0 + g_1x_{t-1} + \varepsilon_t$

  - The null hypothesis for the Dickey-Fuller test is that $g_1 = 0$ (effectively means that $b_1 = 1$) and that the time series has a unit root, which makes it nonstationary.
  - The alternative hypothesis for the Dickey-Fuller test is that $g_1 < 0$, (effectively means that $b_1 < 1$) and that the time series is covariance stationary (i.e., it does not have a unit root).

- Seasonality

  $$x_t = b_0 + b_1x_{t-1} + b_2x_{t-n} + \varepsilon_t$$

  Where n = number of periods in the seasonal pattern

- Moving Average Models

  $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$

  $x_t = \varepsilon_t + \theta\varepsilon_{t-1} + \ldots\theta_q\varepsilon_{t-q}$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$

- Autoregressive Moving Average (ARMA) Models

  $x_t = b_0 + b_1x_{t-1} + \ldots + b_px_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}$
  $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$

- Autoregressive Conditional Heteroskedasticity Models (ARCH Models)

$\hat{\varepsilon}_t^2 = a_0 + \hat{a}_1\hat{\varepsilon}_{t-1}^2 + u_t$

The error in period t+1 can then be predicted using the following formula:

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

Machine Learning

ML Algorithm Type	Supervised/Unsupervised	When to Use?
Classification and Regression Tree (CART)	Supervised	Most commonly applied to binary classification or regression.
Deep Learning Net	Both	A form of neural network with three or more ”hidden” layers
Ensemble Learning	Supervised	The use of a combination of algorithms to describe the data.
Hierarchical	Unsupervised	A form of clustering data (separating observations into groups) into different and final levels of clusters based on relationships between clusters.
K-Means	Unsupervised	A form of clustering data into a predetermined number of groups.
K-Nearest Neighbor (KNN)	Supervised	Mainly used for classification, by classifying new observations based on existing data.
LASSO	Supervised	A type of penalized regression that also eliminates the least important features of the regression model.
Neural Networks	Both	Commonly used for regression and classification in which input features (similar to regression independent variables) are connected to the output (target) variable by “hidden” layers of relationships.
Penalized Regression	Supervised	Regression technique to avoid overfitting by penalizing data features that make insufficient contribution to the regression model.
Principal Components Analysis (PCA)	Unsupervised	Used to help reduce the features in a data set to a manageable level.
Random Forest	Supervised	Type of ensemble learning using collection of decision trees.
Reinforcement Learning	Unsupervised	An algorithm that uses the experience of millions of trials and errors to maximize future success.
Support Vector Machine (SVM)	Supervised	Used for classification, regression, and outlier detection by finding the optimal boundary between sets of data points.

- LASSO Penalized Regression Constraint

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \lambda\sum_{k=1}^{K}\left|\hat{b}_k\right|$$

where:
λ = hyperparameter set by researcher prior to learning
b_k = regression coefficient of kth feature (factor)

Big Data Projects

- Normalization

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

- Standardization

$$X_{i\,(standardized)} = \frac{X_i - \mu}{\sigma}$$

- Performance Evaluation

Accuracy = (TP + TN)/(TP + FP + TN + FN)

FI score = (2 * P * R)/(P + R)

Precision (P) = TP/(TP + FP)

Recall (R) = TP/(TP + FN)

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