Level II
of the
CFA® Program

# Quantitative Methods

## EVALUATING REGRESSION MODEL FIT AND INTERPRETING MODEL RESULTS

## *Learning Outcome Statements*

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing **ANOVA table results** and **measures of goodness of fit**.

**LOS :** Formulate **hypotheses on the significance of two or more coefficients** in a multiple regression model and interpret the results of the **joint hypothesis tests**.

**LOS :** *Calculate* and interpret a **predicted value** for the dependent variable, given the estimated regression model and assumed values for the independent variable.

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

## ANOVA Table and Measures of Goodness of Fit

An estimate of R-squared (R2), also known as coefficient of determination, measures how well a regression fits the data.

$$R^2 = \frac{\textbf{Sum of regression squares}}{\textbf{Sum of squares total}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Where:

$n =$ Number of observations.

$Y_i =$ Dependent variable observations.

$\hat{Y}_i =$ Dependent variables predicted value to the independent variable.

$\overline{Y} =$ Dependent variable mean.

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

## Limitations of $R^2$

- $R^2$ ***cannot*** determine the **statistical significance** of coefficients.
- $R^2$ ***cannot*** detect **biases** in coefficients or estimates.
- The better the model, the lower the $R^2$, while the worse the model, the higher the $R^2$, usually because of **overfitting**.

When there are **too many independent variables**, the regression model is **overfitted**.

The coefficients produced by overfitting may not reflect the **true relationship** between the variables.

An alternative measure of goodness of fit is the **adjusted *$R^2$***.

Due to its degree-of-freedom adjustment, adjusted $R^2$ **does not increase automatically when more independent variables are included**.

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

$$\overline{R^2} = 1 - \left[ \frac{\text{Sum of squares error}}{n-k-1} \Big/ \frac{\text{Sum of squares total}}{n-1} \right]$$

Therefore, the relationship between $\overline{R^2}$ and $R^2$ can be mathematically derived as follows:

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

**Note that:**

- **If $k \geq 1$ then $R^2 >$ adjusted $R^2$ the result is that adjusted $R^2$ can be negative while $R^2$ is zero at minimum.**

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

## Consider these factors when adding a new variable to the regression:

- $\overline{R^2}$ increases when the coefficient *t*-statistic is $>|1.0|$.
- $\overline{R^2}$ decreases when the coefficient *t*-statistic is $<|1.0|$.
- At typical significance levels, 5% and 1%, a *t*-statistic with an **absolute value of 1.0** does not indicate the independent variable is different from zero; therefore, the adjusted $R^2$ doesn't demonstrate that it will increase significantly.

## Example: Interpreting Regression Output

Consider the following regression results generated from the multiple regression analysis of the price of the **US Dollar index** on the **inflation rate** and **real interest rate**.

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

| ANOVA | | | |
|---|---|---|---|
| | df | SS | Significance F |
| **Regression** | 2 | 432.2520 | 0.0179 |
| **Residual** | 7 | 200.6349 | |
| **Total** | 9 | 632.8869 | |
| | | | |
| | Coefficients | Standard Error | |
| **Intercept** | 81 | 7.9659 | |
| **Inflation rates** | -276 | 233.0748 | |
| **Real interest Rates** | 902 | 279.6949 | |

Given the above information and the value of **n=10 (no. of observations)**, the regression equation can be expressed as:

$$\boldsymbol{P} = 81 - 276\boldsymbol{INF} + 902\boldsymbol{IR}$$

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

Where:

P = Price of USDX.

INF = Inflation rate.

IR = Real interest rate.

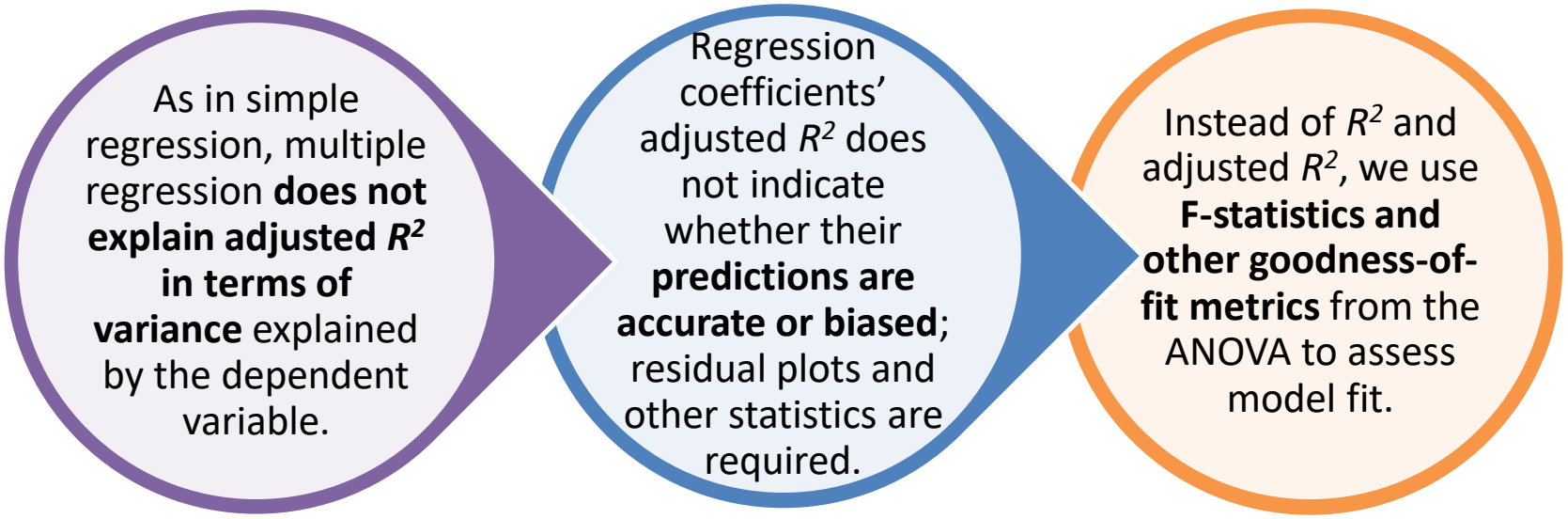
$R^2$ and adjusted $R^2$ can also be calculated as follows:

$$\boldsymbol{R^2 = \frac{RSS}{SST} = \frac{432.2520}{632.8869} = 0.6830 = 68.30\%}$$

$$\textbf{Adjusted } \boldsymbol{R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right) = 1 - \frac{10-1}{10-2-1}(1 - 0.6830)}$$

**Adjusted $R^2$ =0.5924 = 59.24%**

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.



**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

## Joint Hypotheses Testing

A hypothesis test involves **testing a parameter** related to a **population**.

**Null hypotheses** are conditions thought **not to be true**.

**Alternative hypotheses** are accepted when **sufficient evidence exists** against the null hypothesis.

To determine whether the dependent variable is explained by the independent variables, hypothesis testing is performed on the estimated **slope coefficients**.

**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

**To test the significance of individual coefficients in a multiple regression model, the t-statistic is calculated as follows:**

$$t = \frac{\widehat{b_j} - b_{H0}}{S_{\widehat{b_j}}}$$

Where:

$\widehat{b_j}$ **=** Estimated regression coefficient.

$b_{H0}$ = Hypothesized value.

$S_{\widehat{b_j}}$ = Standard error of the estimated coefficient.

*k* is the number of independent variables, and 1 is the intercept term, which makes the test statistic have ***n-k-1*** **degrees of freedom**.

A t-test tests the null hypothesis that the regression coefficient equals a hypothesised value against the alternative hypothesis.

**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

$$H_0: b_j = v \; vs \; H_a : b_j \neq v$$

Where: v = Hypothesized value

F-tests determine whether all independent variables explain the dependent variable.

A slope coefficient test determines whether at least one slope coefficient in a regression is greater or less than zero in order to test for overall significance of the regression.

The F-statistic (which is a one-tailed test) is computed as:

$$F = \frac{\left(\frac{RSS}{k}\right)}{\left(\frac{SSE}{n - (k + 1)}\right)} = \frac{\text{Mean Regression sum of squares } (MSR)}{\text{Mean squared error } (MSE)}$$

Where:

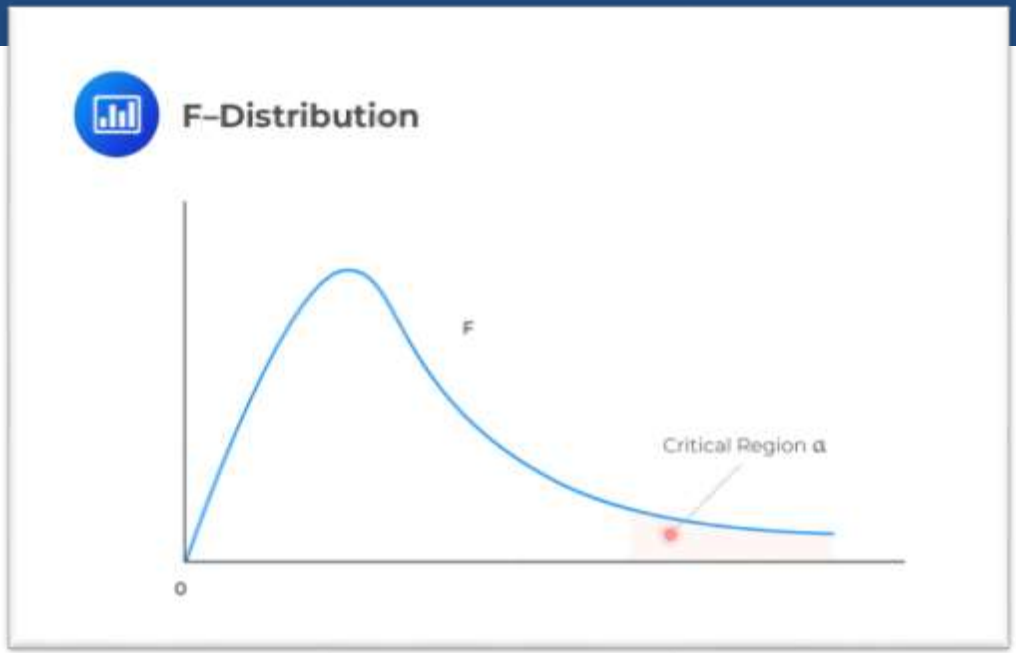
- RSS = Regression sum of squares
- SSE = sum of squared errors
- n = number of observations
- k = number of independent variables

**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

Having a large value of *F* indicates that the regression model is **able to explain variations** in the dependent variable.

In contrast, if the independent variables don't explain the dependent variable, ***F* will be zero**.

As long as the calculated value of ***F* exceeds the upper critical value** of the **one-tailed *F* distribution** with the specified degrees of freedom, we reject the null hypothesis at a given significance level.



**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

**Analyzing multiple regression models for model fit**

| Statistic | Assessing criteria |
|---|---|
| **Adjusted $R^2$** | It is better if it is **higher**. |
| **Akaike's information criterion (AIC)** | It is better if it is **lower**. |
| **Schwarz's Bayesian information criterion (BIC)** | A **lower** number is better. |
| **An analysis of slope coefficients using the t-statistic** | The critical t-value(s) are located **outside the given range** for the selected significance level. |
| **Test of slope coefficients using the F-test** | The F-value for the selected significance level **exceeds the critical value**. |

**Akaike's information** criterion (AIC) is used to evaluate **several models** explaining the same dependent variable.

It can often be calculated from the regression output, but most **regression software** includes it.

The **Bayesian Information Criterion** can be used to compare models with **identical dependent variables** (BIC).

BIC prefers models with **fewer parameters** because more parameters incur a penalty.

For very small sample sizes, ***ln(n)* exceeds 2**.

**LOS :** Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.

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

Where:

- $n =$ Sample size.
- $k =$ Number of independent variables.
- $2(k+1) =$ The model is penalized when independent variables are included.

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

**LOS :** Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

## The Use of Multiple Regression for Forecasting

As shown below, multiple regression uses more items summed up than simple regression for predicting the dependent variable's value:

$$\widehat{Y_f} = \hat{b}_0 + \hat{b}_1 X_{1f} + \hat{b}_2 X_{2f} + \cdots + \hat{b}_k X_{kf} = \hat{b}_0 + \sum_{j=1}^{k} \hat{b}_j X_{jf}$$

Where:

- $\hat{Y}_f$ = Predicted (forecasted) value of the **dependent variable**.
- $\hat{b}_j X_f$ = This value is the estimated **slope of the coefficient** multiplied by the assumed value of the variable.
- $\hat{b}_0$ = Estimated **intercept coefficient**.

**LOS :** Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

If a regression model is estimated using all five independent variables, any prediction of the dependent variable must include all five independent variables, **even if they are not statistically significant**.

**Correlations** between variables were taken into account when **estimating slope coefficients**.

Any prediction of the dependent variable **must include the intercept term**.

## Example: Calculating the Predicted Value of a Dependent Variable

Here is a regression equation for USDX on inflation and real interest rates.

$$P = b_0 + b_1 INF + b_2 IR + \epsilon_t$$

**LOS :** Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

**The following table gives the regression results:**

| Regression Statistics | | | | |
|---|---|---|---|---|
| Multiple R | | | 0.8264 | |
| $R^2$ | | | 0.6830 | |
| Adjusted $R^2$ | | | 0.5924 | |
| Standard Error | | | 5.3537 | |
| Observations | | | 10 | |
| | **Coefficients** | **Standard Error** | **t Stat** | **P-value** |
| Intercept | 81 | 7.9659 | 10.1296 | 0.0000 |
| Inflation rates | -276 | 233.0748 | -1.1833 | 0.2753 |
| Real interest Rates | 902 | 279.6949 | 3.2266 | 0.0145 |

Use the estimated regression equation above to *calculate* the predicted price of the US dollar index (USDX), assuming the **inflation rate is 3.5%** and the **real interest rate is 4%**.

**LOS :** Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

**The following table gives the regression results:**

| Regression Statistics | | | | |
|---|---|---|---|---|
| Multiple R | | | 0.8264 | |
| $R^2$ | | | 0.6830 | |
| Adjusted $R^2$ | | | 0.5924 | |
| Standard Error | | | 5.3537 | |
| Observations | | | 10 | |
| | **Coefficients** | **Standard Error** | **t Stat** | **P-value** |
| Intercept | 81 | 7.9659 | 10.1296 | 0.0000 |
| Inflation rates | -276 | 233.0748 | -1.1833 | 0.2753 |
| Real interest Rates | 902 | 279.6949 | 3.2266 | 0.0145 |

**Solution**

$$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1}\widehat{X_{1i}} + \widehat{b_2}\widehat{X_{2i}} + \cdots + \widehat{b_k}\widehat{X_{ki}}$$

$$\widehat{Y_i} = 81 + (-276 \times 0.035) + (902 \times 0.04) = 107.42$$

## *Learning Outcome Statements*

**LOS :** Evaluate how well a multiple regression model explains the dependent variable by analyzing **ANOVA table results** and **measures of goodness of fit**.

**LOS :** Formulate **hypotheses on the significance of two or more coefficients** in a multiple regression model and interpret the results of the **joint hypothesis tests**.

**LOS :** *Calculate* and interpret a **predicted value** for the dependent variable, given the estimated regression model and assumed values for the independent variable.