# Quantitative Methods

# Learning Module 1
# Basics of Multiple Regression and Underlying Assumptions

**LOS:** Describe the types of investment problems addressed by multiple linear regression and the regression process.

**LOS:** Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients.

**LOS:** Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

Multiple linear regression is a modeling technique that uses two or more independent variables to explain the variation of the dependent variable. A reliable model can lead to a better understanding of value drivers and improve forecasts, but an unreliable model can lead to spurious correlations and poor forecasts.

Several software programs and functions exist to help execute multiple regression models:

| Software | Programs/Functions |
|---|---|
| Excel | Data Analysis > Regression |
| Python | scipy.stats.linregress<br>statsmodels.lm<br>sklearn.linear_model.LinearRegression |
| R | lm |
| SAS | PROC REG<br>PROC GLM |
| STATA | regress |

# Uses of Multiple Linear Regression

**LOS:** Describe the types of investment problems addressed by multiple linear regression and the regression process.

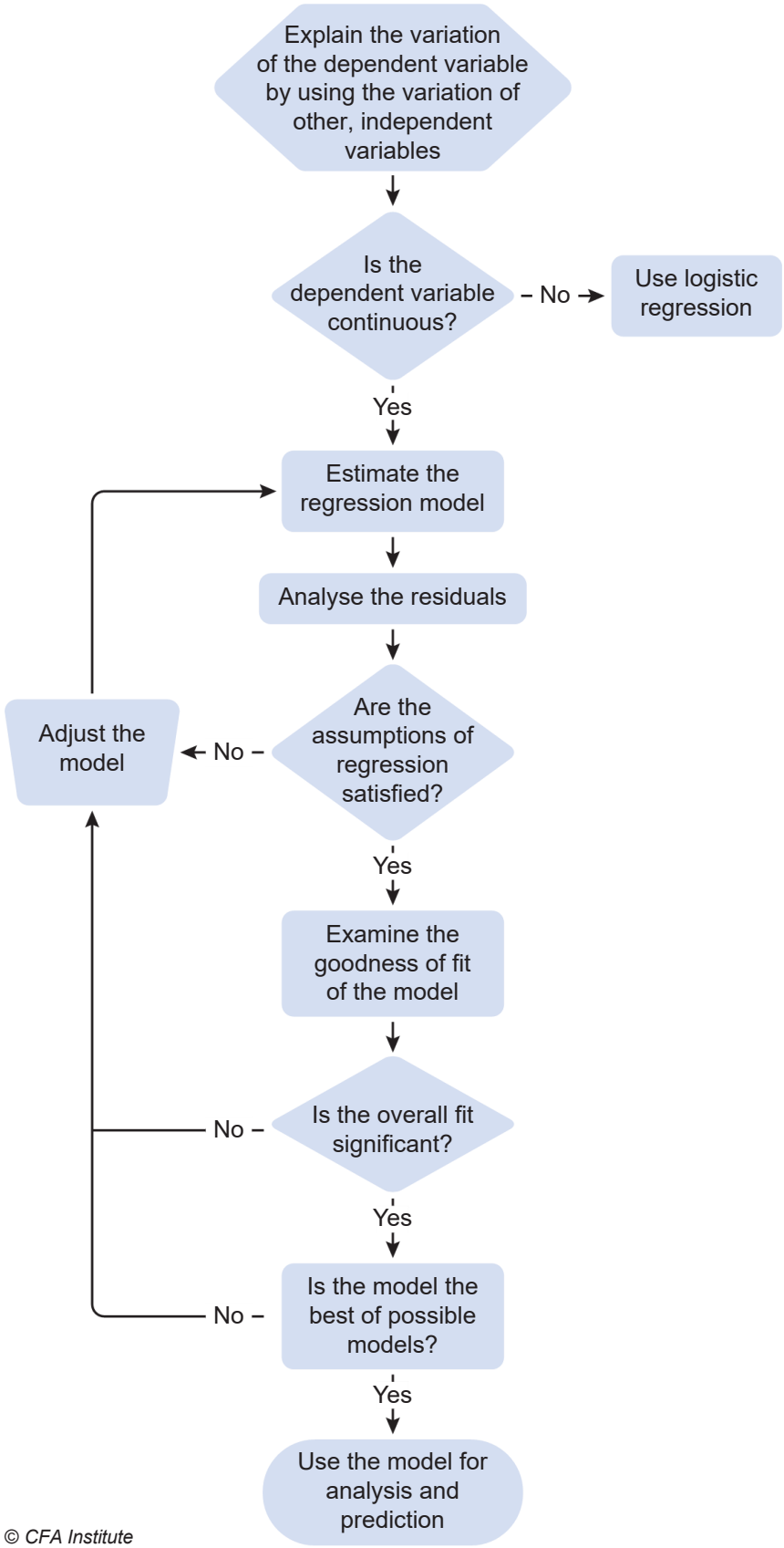
The complexity of financial and economic relationships often requires understanding multiple factors that affect the dependent variable. Some examples where multiple linear regression can be useful include:

- A portfolio manager wants to understand how returns are influenced by underlying factors.
- A financial advisor wants to identify when financial leverage, profitability, revenue growth and changes in market share can predict financial distress.
- An analyst wants to examine the effect of country risk on fixed-income returns.

In all cases, the basic framework of a regression model is as follows:

- Specify a model, including independent variables.
- Estimate a regression model and analyze it to ensure that it satisfies key underlying assumptions and meets the goodness-of-fit criteria.
- Test the model's out-of-sample performance. If acceptable, it can then be used for further identifying relationships between variables, testing existing theories, or forecasting.

**Exhibit 1 Regression process**



# The Basics of Multiple Regression

**LOS:** Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients.

Multiple regression is similar to simple regression where a dependent variable, $Y$, is explained by the variation of an independent variable, $X$. Multiple regression expands this concept into a statistical procedure that evaluates the impact of more than one independent variable on a dependent variable. A multiple linear regression model has the following general form:

**Multiple regression equation**

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

Where:

$Y_i$ = The $i$th observation of the dependent variable $Y$

$X_{ji}$ = The $i$th observation of the independent variable $X_j$, $j = 1, 2, \ldots, k$

$b_0$ = The intercept of the regression

$b_1, \ldots, b_k$ = The slope coefficients for each of the independent variables

$\varepsilon_i$ = The error term for the $i$th observation

$n$ = The number of observations

The slope coefficients, $b_1$ to $b_k$, measure how much the dependent variable, $Y$, changes in response to a one-unit change in that specific independent variable. In our equation, the independent variable $X_1$, holding all other independent variables constant, will change $Y$ by a factor of $b_1$. Here, $b_1$ is called a **partial regression coefficient**, or a partial slope coefficient, because it explains only the part of the variation in $Y$ related to that specific variable, $X_1$.

Note that for any multiple regression equation:

- There are $k$ slope coefficients in a multiple regression.
- The $k$ slope coefficients and the intercept, $b_0$, are all known as regression coefficients.
- There are $k + 1$ regression coefficients in a multiple regression equation.
- The residual term, $\varepsilon_i$, equals the difference between the actual value of $Y$ ($Y_i$) and the predicted value of $Y$ ($\hat{Y}_i$). In terms of our multiple regression equation:

**Residual term**

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

# Assumptions Underlying Multiple Linear Regression

**LOS:** Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

In order to make valid predictions using a multiple regression model based on ordinary least squares (OLS), a few key assumptions must be met.

**Exhibit 2 Multiple linear regression assumptions**

| Assumption | Description | Violation |
|---|---|---|
| Linearity | Dependent and independent variable have linear relationship | Nonlinearity |
| Homoskedasticity | Variance of residuals constant across all observations | Heteroskedasticity |
| Independence of errors | Observations are independent of each other; errors (ie, residuals) uncorrelated across all observations | Serial correlation or autocorrelation |
| Normality | Residuals normally distributed, with expected value of zero | Non-normality |
| Independence of independent variables | Independent variables are not random; no exact linear relation between independent variables | Multicollinearity |

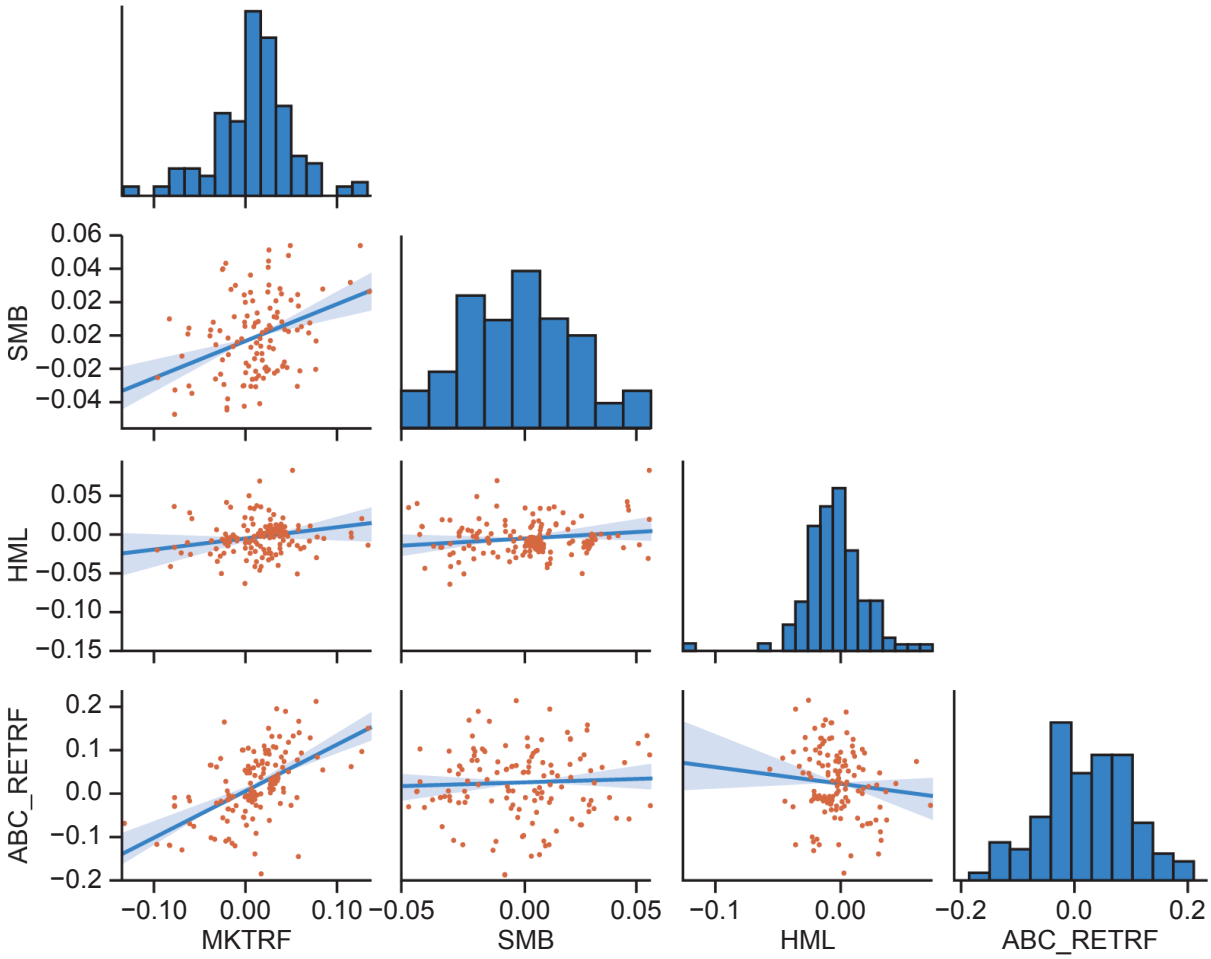
Statistical tools exist to test these assumptions and the model for overall goodness of fit. Most regression software packages have built in diagnostics for this purpose.

To better illustrate this, consider a regression to analyze 10 years of monthly total excess returns of ABC stock using the Fama-French three-factor model. This model uses market excess return (MKTRF), size (SMB), and value (HML) as explanatory variables.

$$ABC_{return_t} = b_0 + b_1\,MKTRF_t + b_2 SMB_t + b_3 HML_t + \varepsilon_t$$

The software produced the following set of scatterplots to test the relationship between the three independent variables:

**Exhibit 3 Scatterplots for three independent variables**



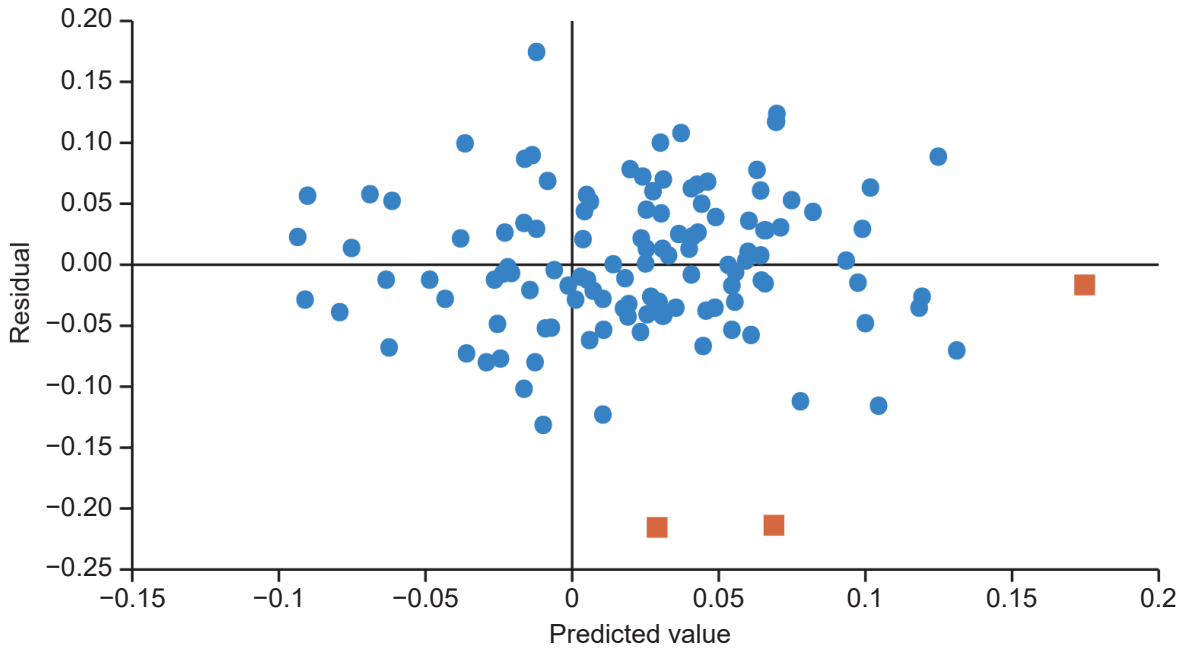



In the lower set of scatterplots of Exhibit 3, there is a positive relationship between ABC's return and the market risk factor (MKTRF), no apparent relationship between ABC's return and the size factor (SMB) and a negative relationship between ABC's return and the value factor (HML).

In the second-to-last (penultimate) level we can see little relationship between SMB and HML. This suggests independence between the variables, which satisfies the assumption of independence.

Then, compare the predicted values, or $\hat{Y}_i$, with the actual values of ABC RETRF$_t$ in the residual plot in Exhibit 4:

**Exhibit 4 Residual plot**

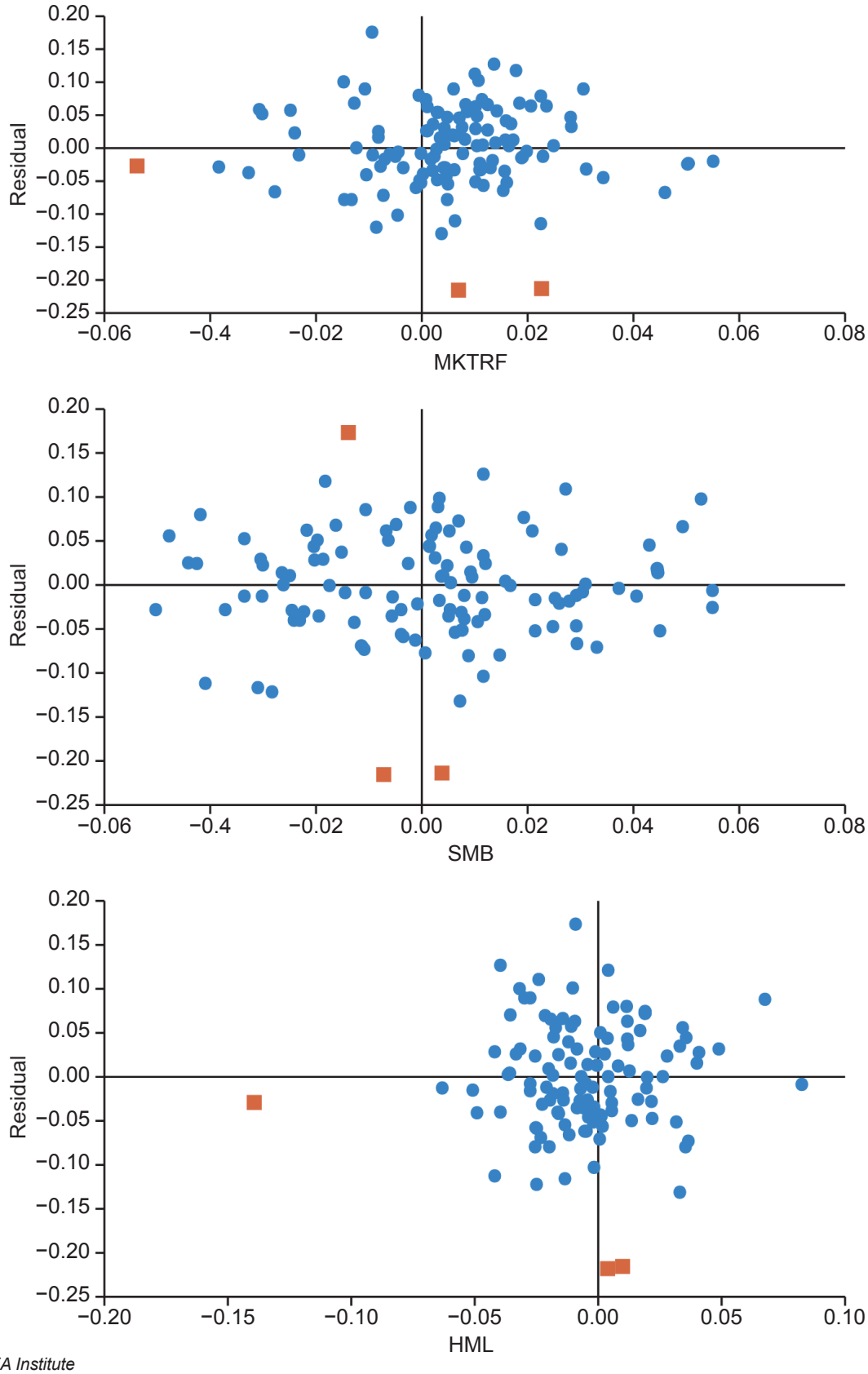



*Potential outliers indicated with square markers*



Exhibit 4 shows the relationship between the residuals and the predicted values. A visual inspection does not show any directional relationship, positive or negative, between the residuals and the predicted values from the regression model. This also suggests that the regression's errors have a constant variance and are uncorrelated with each other. There are however, three residuals (square markers) that may be outliers.

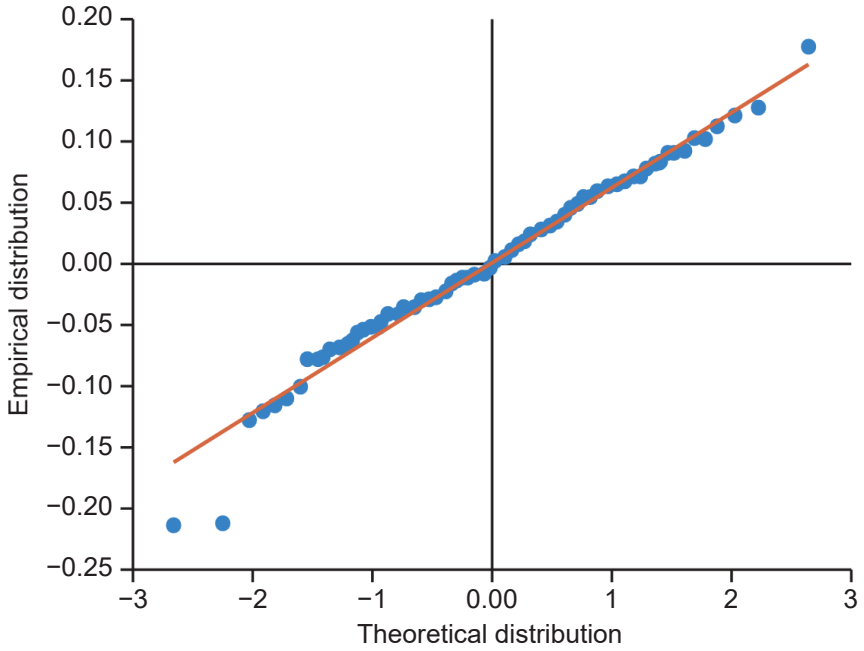
**Exhibit 5 Regression residuals versus each of the three factors**



Each plot shows the relationship of the residual output versus the value of each independent variable to look for directional relationships related to that specific factor. In this example, none of the three plots indicate any direct relationship between the residuals and the explanatory variables, which suggests that there is no violation of multiple regression assumptions. Furthermore, in all four graphs, the outliers identified are the same.

Exhibit 6 is a normal Q-Q plot used to visualize the distribution of a variable compared with a theoretical normal distribution.

**Exhibit 6**




*Superimposed on the plot is a linear relation*



In this plot, the red line represents a normal distribution with a mean of 0 and a standard deviation of 1. The green dots are the model residuals fit to a normal distribution, or the empirical distribution on the vertical axis of Exhibit 5. These are superimposed over the red theoretical distribution line to visualize how consistent the normalized residuals are with a standard normal distribution. The same three outliers remain, but the rest of the residuals closely align with a normal distribution, which is the desired outcome.

# Learning Module 2
# Evaluating Regression Model Fit and Interpreting Model Results

**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.

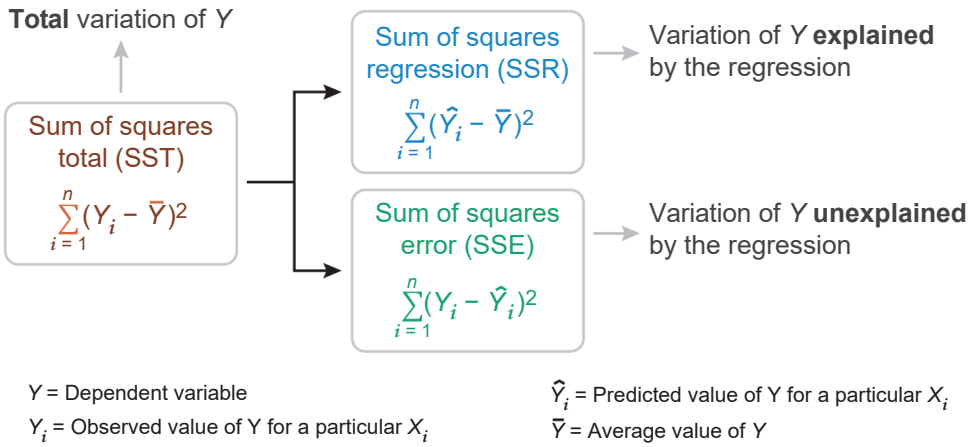
**LOS:** Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

## Goodness of Fit

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

The **coefficient of determination** measures a regression's **goodness of fit, known as** the $R^2$ statistic: how much of the variation in the dependent variable is captured by the independent variables in the regression. Exhibit 1 shows how a regression model explains the variation in the dependent variable:

**Exhibit 1 Regression model seeks to explain the variation of Y**




$Y$ = Dependent variable
$Y_i$ = Observed value of Y for a particular $X_i$
$\hat{Y}_i$ = Predicted value of Y for a particular $X_i$
$\bar{Y}$ = Average value of $Y$

$R^2$ is calculated as:

**Coefficient of determination**

$$R^2 = \text{Total variation} - \text{Unexplained variation}$$

$$R^2 = \frac{\text{Sum of squares regression}}{\text{Sum of squares total}}$$

$$R^2 = \sum_{i=0}^{n} (\hat{Y}_i - \bar{Y}_i)^2 / \sum_{i=0}^{n} (Y_i - \bar{Y}_i)^2$$

Where:

$n$ is the number of observations in the regression

$Y_i$ is an observation of the $Y$ variable

$\hat{Y}$ is the predicted value of the dependent variable

$\bar{Y}$ is the average of the dependent variable

A major concern with using $R^2$ in multiple regression analysis is that as more independent variables are added to the model, the total amount of unexplained variation will decrease as the amount of explained variation increases. As such, each successive $R^2$ measure will appear to reflect an improvement over the previous model. This will be the case as long as each newly added independent variable is even slightly correlated with the dependent variable and is not a linear combination of the other independent variables already in the regression model.

Other limitations to using $R^2$:

- It does not tell the analyst whether the coefficients are statistically significant
- It does not indicate whether there are biases in the coefficients or predictions
- It can misread the fit due to bias and overfitting

**Overfitting** can result from an overly complex model with too many independent variables relative to the number of observations. In such cases, the model does not properly represent the true relationship between the independent and dependent variables.

Therefore, analysts typically use adjusted $R^2$, or $\bar{R}^2$, which does not increase whenever another variable is added to the regression since it is adjusted for degrees of freedom.

**Adjusted $R^2$**

$$R^2 = \frac{\text{Sum of squares error} / (n - k - 1)}{\text{Sum of squares total} / (n - k - 1)}$$

Where: $k$ = Number of independent variables

A few things to note when comparing $R^2$ to $\bar{R}^2$

- If $k = 1$, $R^2 > \bar{R}^2$
- $\bar{R}^2$ will decrease if the inclusion of another independent variable in the regression model results in a nominal increase in explained variation (RSS) and $R2$.

- $\bar{R}^2$ can be negative (in which case we consider its value to equal 0) while $R^2$ can never be negative.
- If $\bar{R}^2$ is used to compare two regression models, the dependent variable must be identically defined in the two models and the sample sizes used to estimate the models must be the same.

Additionally, if the *t*-statistic > |1| then $\bar{R}^2$ will increase; conversely, values < |1| will decrease $\bar{R}^2$

There are cases where both the $R^2$ and $\bar{R}^2$ can increase when more independent variables are added. For these cases there are several statistics used to compare model quality, including **Akaike's information criterion (AIC)** and **Schwarz's Bayesian information criterion (BIC)**.

AIC is used to evaluate a collection of models that explain the same dependent variable. Even though this will generally be provided in the output for regression software, we can also calculate it as:

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

Where:
$k$ = Number of independent variables
$n$ = Sample size

A lower AIC indicates a better-fitting model. Note that AIC depends on the sample size ($n$), the number of independent variables ($k$), and the sum of the squares error (SSE). The term at the end, $2(k + 1)$, is a penalty term that increases as more independent variables, $k$, are added.

Similarly, BIC allows comparison of models with the same dependent variable:

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

Where:
$k$ = Number of independent variables
$n$ = Sample size

With BIC, there is a greater penalty for having more parameters than with AIC. BIC will tend to prefer smaller models because ln($n$) is greater than 2, even for very small sample sizes. AIC is preferred if the model is for prediction purposes, and BIC is preferred for evaluating goodness of fit.

The AIC and the BIC alone are not telling, however, and should be compared across models using a combination of factors. Example 1 shows the goodness-of-fit measures for a model that incorporates five independent variables (factors):

**Example 1 Goodness of fit evaluation**

| | $R^2$ | Adjusted $R^2$ | AIC | BIC |
|---|---|---|---|---|
| Factor 1 only | 0.541 | 0.531 | 19.079 | 22.903 |
| Factors 1 and 2 | 0.541 | 0.531 | 21.078 | 26.814 |
| Factors 1, 2, and 3 | 0.562 | 0.533 | 20.743 | 28.393 |
| Factors 1, 2, 3, and 4 | 0.615 | 0.580 | *16.331* | 25.891 |
| Factors 1, 2, 3, 4, and 5 | 0.615 | 0.572 | 18.251 | 29.687 |

Note that:

$R^2$ increases or stays the same as more factors are added

$\bar{R}^2$ either increases or decreases as each new factor is added

AIC is minimized when the first four factors are used

BIC is minimized when only the first is used

Using the results, we would select the four-factor model if we were using it to make predictions, but would use the first model if we were just measuring goodness of fit.

## Testing Joint Hypotheses for 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.

In a multiple regression, the intercept is the value of the dependent variable if all independent variables are 0. The slope coefficient of each of the independent variables is the change in the dependent variable for a change in that independent variable if all other independent variables remain constant.

Tests for individual coefficients in multiple regression are identical to tests for individual coefficients in simple regression. The hypothesis structure is the same and the *t*-test is the same.

For a two-sided test of whether a variable is significant in explaining the dependent variable's variation, the hypotheses are:

$$H_0: b_i = B_i$$

$$H_a: b_i \neq B_i$$

Where $b$ is the true coefficient for the $i$th independent variable and $B$ is a hypothesized slope coefficient for the same variable.

If the hypothesis test is simply to test the significance of the variable's predictive power, the hypotheses would be: $H_0$: $B_j = 0$ and $H_a$: $B_j \neq 0$

There are times to test a subset of variables in a multiple regression, for example, when comparing the Fama-French three-factor model (MKTRF, SMB, HML) to the Fama-French five-factor model (MKTRF, SMB, HML, RMW, CMA) to determine which model is more concise or to find the factors that are most useful in explaining the variation in the dependent variable. In other words, it may be that not all the factors in such a model are actually required for the model to have predictive power.

The full model, using all independent variables, is called the **unrestricted model**. This model is compared with a **restricted model**, which effectively includes fewer independent variables since coefficients for each unneeded variable are set to 0. A restricted model is also called a nested model since its independent variables form a subset of the variables in the unrestricted model.

Unrestricted five-factor model:

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

Restricted two-factor model:

$$Y_i = b_0 + b_1X_{1i} + b_4X_{4i} + \varepsilon_i$$

The hypothesis test in this example would be to test whether the coefficients of $X_2$, $X_3$, and $X_5$ are significantly different than 0. To compare the unrestricted model to the nested model, perform an *F*-test to test the role of the jointly omitted variables:

$$F = \left( \frac{\text{Sum of squares error}_{\text{(Restricted model)}} - \text{Sum of squares error}_{\text{(Unrestricted model)}}}{q} \Big/ \frac{\text{Sum of squares error}_{\text{(Restricted model)}}}{n - k - 1} \right)$$

Where: $q$ = Number of variables omitted in the restricted model

The role of the *F*-test determines whether the change in the sum of squared errors (SSE) caused by including the variables from the unrestricted model is significant enough to compensate for the decrease in degrees of freedom. In the example shown here, there is a loss of three degrees of freedom since there are only two independent variables instead of five.

- The null hypothesis is that the slope of the omitted factors is equal to 0:
- The alternative hypothesis is that at least one is not equal to 0: $H_a$: at least one of the factors ≠ 0.

If the *F*-statistic is less than the critical value, then we fail to reject the null hypothesis. This means that the added predictive power of the variables omitted in the restricted model is not significant and the restricted model fits the data better.

Exhibit 2 summarizes the desired values of a multiple regression test:

**Exhibit 2 Assessing model fit using multiple regression statistics**

| Statistic | Criterion to use in assessment |
|---|---|
| Adjusted $R^2$ | The higher the better |
| Akaike's information criterion (AIC) | The lower the better |
| Schwarz's Bayesian information criterion (BIC) | The lower the better |
| *t*-statistic on a slope coefficient | Outside the bounds of critical *t*-value(s) for the selected significance level |
| *F*-test for joint tests of slope coefficients | Exceeds the critical *F*-value for the selected significance level |



## Forecasting Using Multiple Regression

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

Predicting the value of the dependent variable in a multiple regression is similar to the prediction process for a simple regression. However, in the case of multiple independent variables, the predicted value is the sum of the product of each variable and its coefficient, plus the intercept:

$$\hat{Y}_f = \hat{b}_0 + \hat{b}_1 X_{1f} + \hat{b}_2 X_{2f} + \cdots \hat{b}_k X_{kf}$$

For example, given the following formula:

$$\hat{Y}_i = 3.546 + 3.235X_1 + 7.342X_2 - 7.234X_3$$

Assume the values of $X_1$, $X_2$, and $X_3$ are:

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| 3.8 | 8.3 | 5.9 |

With this information the predicted value of $Y_i$ is calculated as:

$$\hat{Y}_i = 3.546 + (3.235 \times 3.8) + (7.342 \times 8.3) - (7.234 \times 5.9) = 34.097$$

It should be noted that the estimate should include all the variables, even those that are not statistically significant, since these variables were used in estimating the value of the slope coefficient.

As with simple linear regression, in multiple linear regression there will often be a difference between the actual value and the value forecasted by the regression model. This is the error term, or the $\varepsilon 1$ term of the regression equation: the difference between the predicted value and the actual value. This is the basic uncertainty of the model known as the model error.

Models using estimated independent variables add another source of error. These out-of-sample data introduce sampling error to the model and will increase the error contributed by the model error.