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 statsmodels.lm sklearn.linear_model.LinearRegression
R	lm
SAS	PROC REG 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.

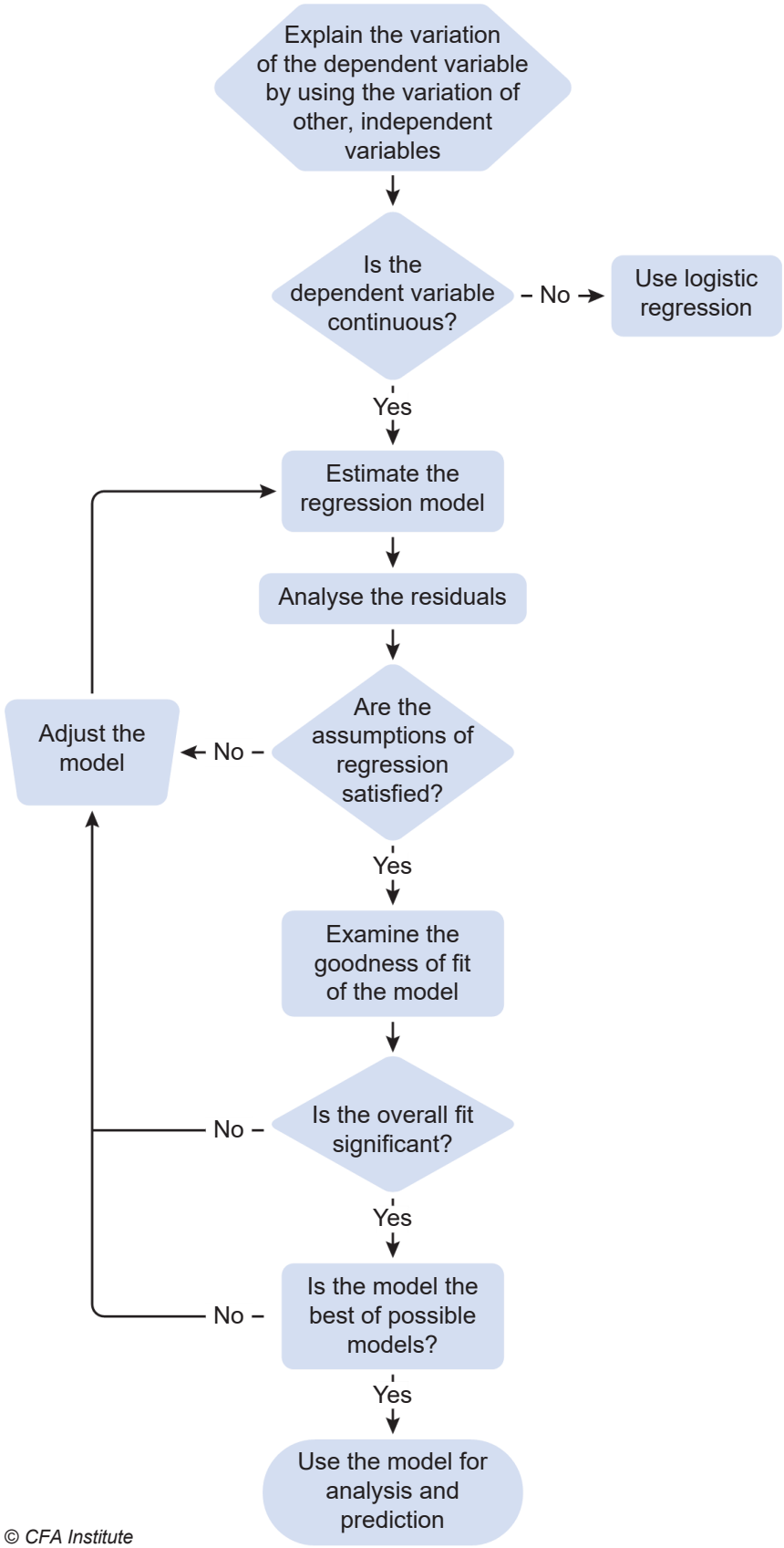
The complexity of financial and economic relationships often requires understanding multiple factors that affect the dependent variable. Some examples where multiple linear regression can be useful include:

- 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 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 regression

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

ε_i = The error term for the ith observation

n = The number of observations

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, ε_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

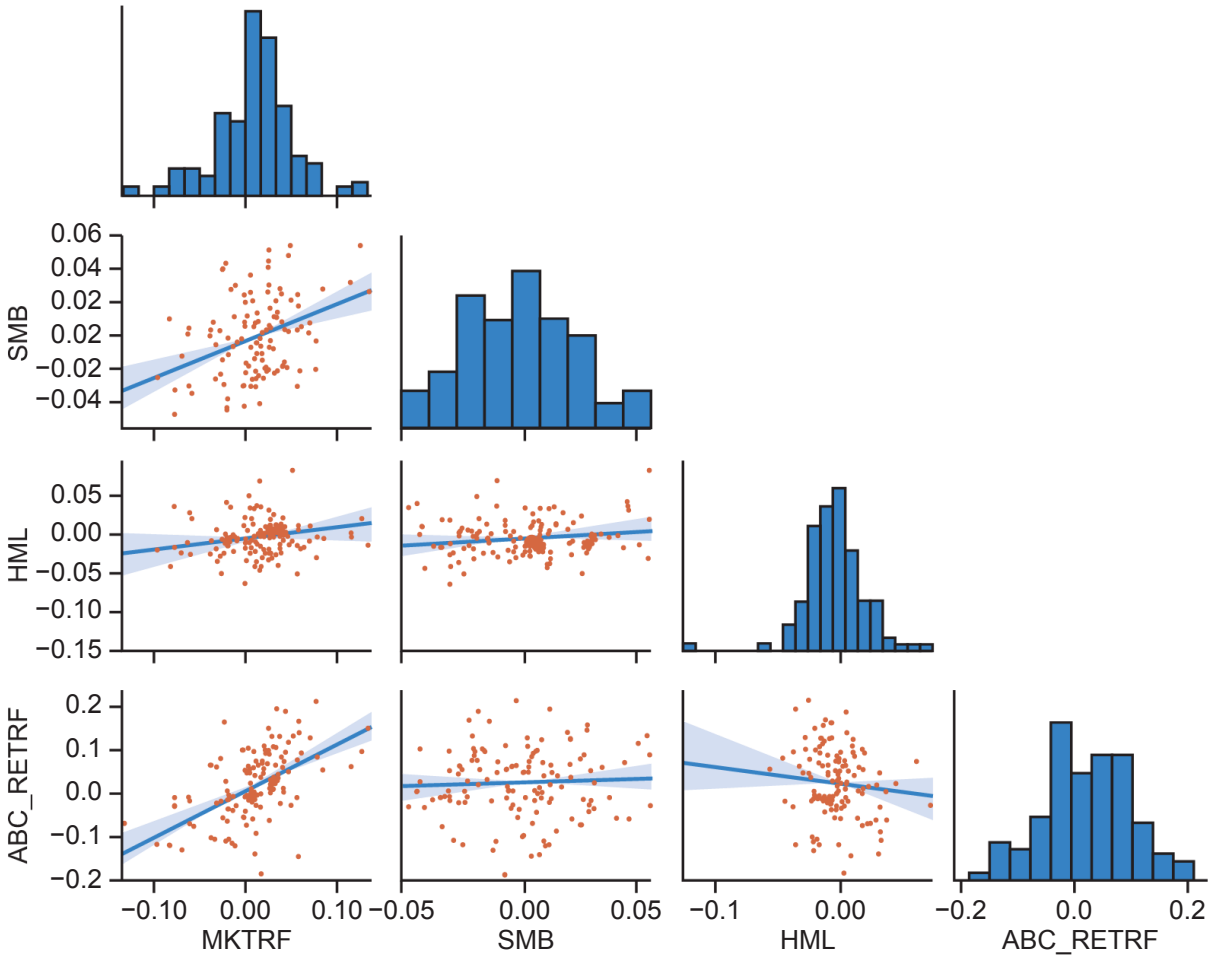
Statistical tools exist to test these assumptions and the model for overall goodness of fit. Most regression software packages have built in diagnostics for this purpose.

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

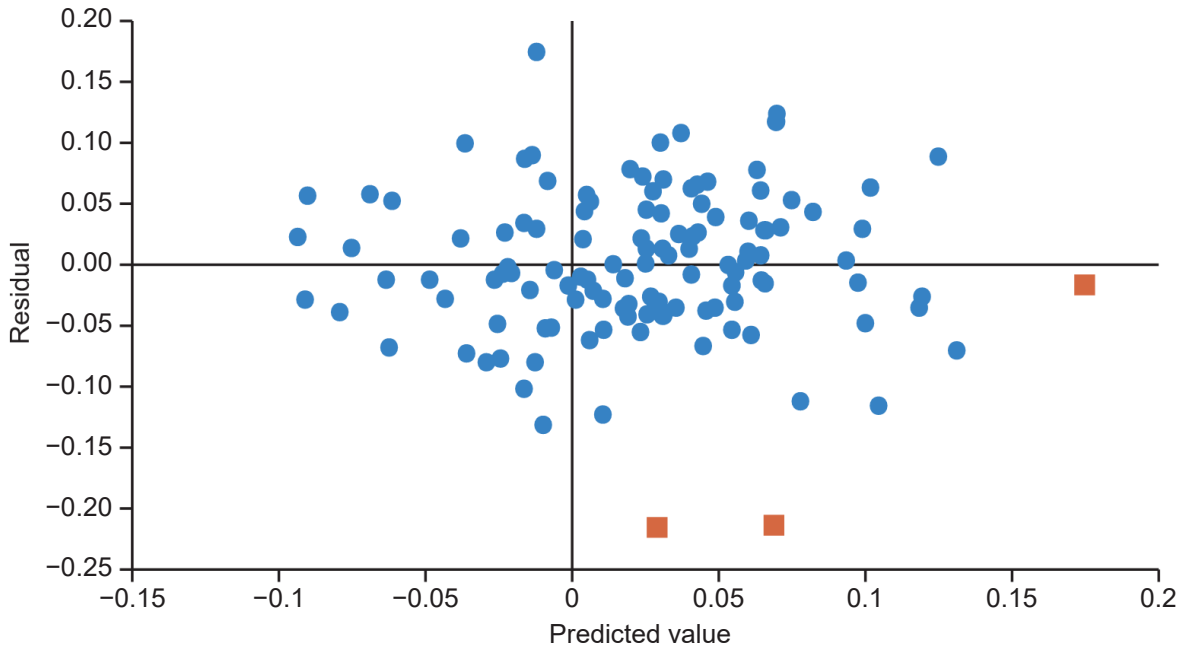


In the lower set of scatterplots of Exhibit 3, there is a positive relationship between ABC's return and the market risk factor (MKTRF), no apparent relationship between ABC's return and the size factor (SMB) and a negative relationship between ABC's return and the value factor (HML).

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:

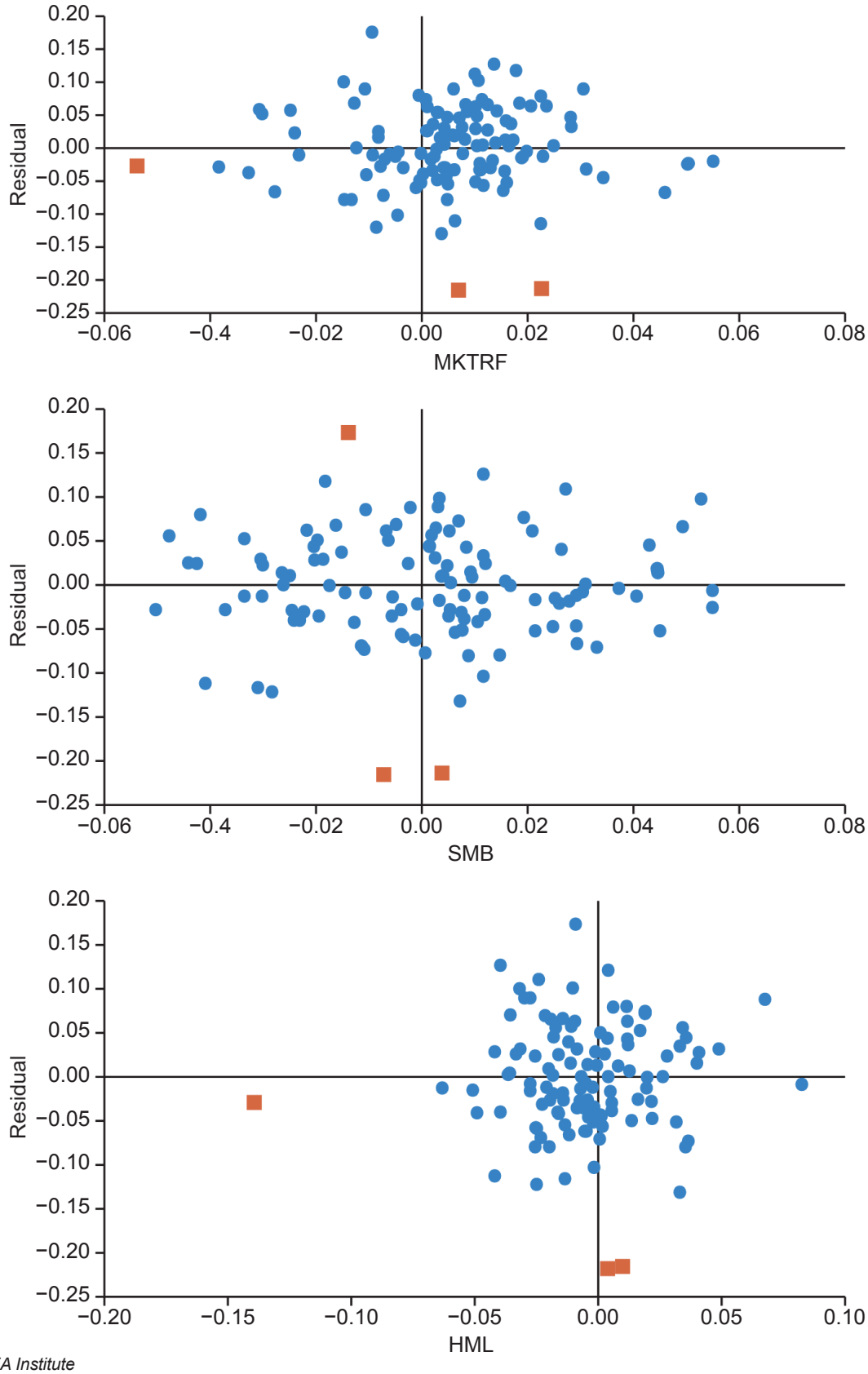
Exhibit 4 Residual plot



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.

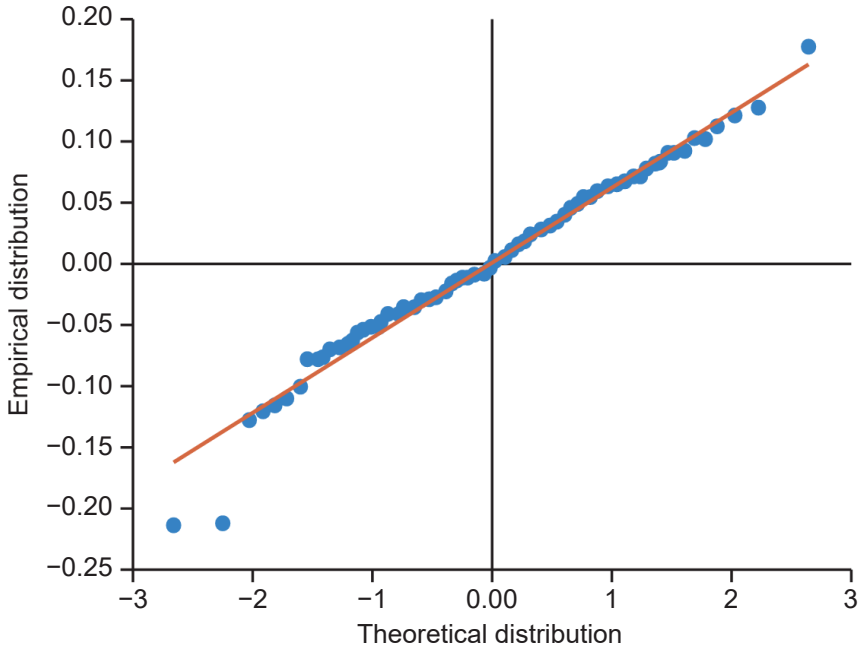
Exhibit 5 Regression residuals versus each of the three factors



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.

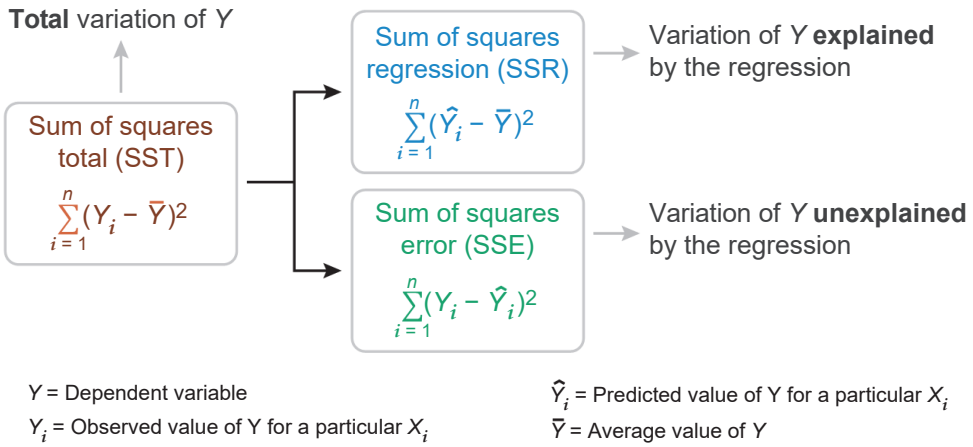
LOS: Calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.

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 ith 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.

Learning Module 3
Model Misspecification

LOS: Describe how model misspecification affects the results of a regression analysis and how to avoid common forms of misspecification.

LOS: Explain the types of heteroskedasticity and how it affects statistical inference.

LOS: Explain serial correlation and how it affects statistical inference.

LOS: Explain multicollinearity and how it affects regression analysis.

Model Specification Errors

LOS: Describe how model misspecification affects the results of a regression analysis and how to avoid common forms of misspecification.

Model specification refers to the set of variables included in the regression and the regression equation's functional form. A good regression model will:

- Be grounded in economic reasoning
- Be concise: each variable included in the model is essential
- Perform well out of sample
- Have an appropriate functional form (for example, if a nonlinear form is expected, then it should use nonlinear terms)
- Satisfy regression assumptions, without heteroskedasticity, serial correlation, or multicollinearity

Misspecified Functional Form

Exhibit 1 illustrates four ways a model's functional form may fail:

Exhibit 1 Functional form failures

Failures in regression functional form	Explanation	Possible consequence
Omitted variables	One or more important variables are omitted from the regression	Heteroskedasticity or serial correlation
Inappropriate form of variables	Ignoring a nonlinear relationship between the dependent and the independent variable	Heteroskedasticity
Inappropriate variable scaling	One or more regression variables may need to be transformed before estimating the regression	Heteroskedasticity or multicollinearity
Inappropriate data pooling	Regression model pools data from different samples that should not be pooled	Heteroskedasticity or serial correlation

Omitted Variables

The omitted variable bias is the bias resulting from the omission of an important independent variable.

For example, assume the true regression model is defined as:

$$Y_1 = b_0 + b_1X_1i + b_2X_2i + \varepsilon_i$$

But the model was estimated as:

$$Y_1 = b_0 + b_1X_1i + \varepsilon_i$$

In this case, the model would be misspecified by the omission of X_2. If the omitted variable is uncorrelated with X_1, then the residual would be $b_2X_2i + \varepsilon_i$. This means that the residual would not have an expected value of 0 nor would it be independent and identically distributed. As a result, the estimate of the intercept would be biased, even if the X_1 were estimated correctly.

If, however, the omitted variable X_2 is correlated with X_1, then the error term in the model would now be correlated with X_1 and the estimated values in the model would be biased and inconsistent with b_1, so the intercept and residuals would also be incorrect.

Inappropriate Form of Variables

An example is when the analyst fails to account for nonlinearity in the relationship between the dependent variable and one or more independent variables. The analyst should consider whether the situation suggests a nonlinear relationship and should confirm nonlinearity by plotting the data. Sometimes, misspecification can be fixed by taking the natural logarithm of the variable.

Inappropriate Scaling of Variables

Using unscaled data when scaled data is more appropriate can result in a misspecified model. This can happen, for example, when looking at financial statement data across companies. This misspecification can be addressed by using common-size financial statements, allowing analysts to quickly compare trends such as profitability, leverage, and efficiency.

Inappropriate Pooling of Data

Inappropriate pooling of data occurs when a sample spans structural breaks in the behavior of the data, such as changes in government regulations or a change from a low-volatility period to a high-volatility period. In a scatterplot, this type of data can appear in discrete, widely separated clusters with little or no correlation. This can be fixed by using the subsample most representative of conditions during the forecasting period.

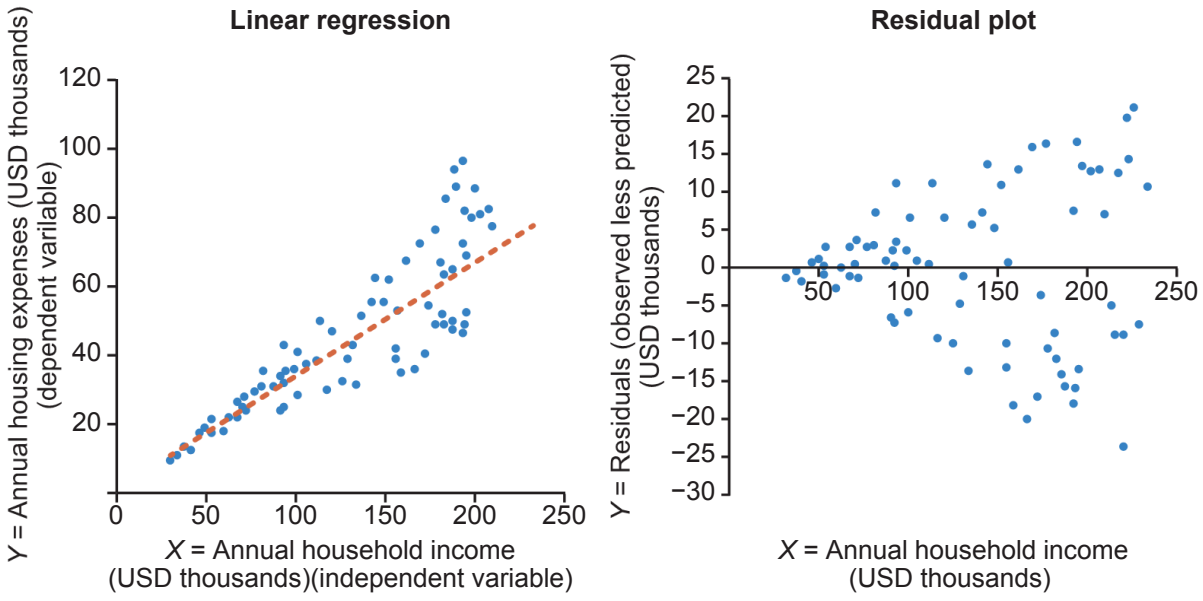
Violations of Regression Assumptions: Heteroskedasticity

LOS: Explain the types of heteroskedasticity and how it affects statistical inference.

Heteroskedasticity occurs when the variance of the error term in the regression is not constant across observations; it results from a violation of the assumption of homoskedasticity, that is, that there is no systematic relationship between the regression residuals, or the vertical distances between the data points and the regression line, and the independent variable.

Heteroskedasticity can result from any kind of model misspecification. Exhibit 2 shows the scatterplot and regression line for a model with heteroskedasticity. Notice that the regression residuals appear to increase in size as the value of the independent variable increases.

Exhibit 2 Example of heteroskedasticity (violation of the homoskedasticity assumption)



Consequences of Heteroskedasticity

Heteroskedasticity comes in two forms:

- **Unconditional heteroskedasticity** occurs when the heteroskedasticity of the variance in the error term is not related to the independent variables in the regression. Unconditional heteroskedasticity does not create major problems for regression analysis even though it violates a linear regression assumption.
- **Conditional heteroskedasticity** occurs when the heteroskedasticity in the error variance is correlated with the independent variables in the regression. While conditional heteroskedasticity creates problems for statistical inference, such as unreliable *F*-tests and *t*-tests, it can be easily identified and corrected. Conditional heteroskedasticity will tend to find significant relationships when none actually exist and lead to Type I errors, or false positives.

Testing for Conditional Heteroskedasticity

The most common test for heteroskedasticity is the Breusch-Pagan (BP) test. The BP test is best explained as a three-step process requiring a regression of the squared residuals from the original estimated regression equation, where the dependent variable is regressed on the independent variables in the regression.

If conditional heteroskedasticity does not exist, the independent variables will not explain much of the variation in the squared residuals from the original regression. However, if conditional heteroskedasticity is present, the independent variables will explain the variation in the squared residuals to a significant extent. Because, in this case, each observation's squared residual is correlated with the independent variables, the independent variable will affect the variance of the errors.

The test statistic for the BP test is approximately chi-square distributed, and is calculated as:

Chi-square test statistic

$$\boldsymbol{X^2_{BP,k} = nR^2}$$

Where:

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

The null hypothesis is that the original regression's squared error term is uncorrelated with the independent variables, or no heteroskedasticity is present.

The alternative hypothesis is that the original regression's squared error term is correlated with the independent variables, or heteroskedasticity is present.

The BP test is a one-tailed chi-square test, because conditional heteroskedasticity is only a problem if it is too large.

Example 1 Testing for heteroskedasticity

An analyst wants to test a hypothesis suggested by Irving Fisher that nominal interest rates increase by 1% for every 1% increase in expected inflation. The Fisher effect assumes the following relationship:

$$i = r + \pi_e$$

Where:

- i = Nominal interest rate
- r = Real interest rate (assumed constant)
- π_e = Expected inflation

The analyst specifies the regression model as: $i_i = b_0 + b_1 \pi_e + \varepsilon_i$.

Since the Fisher effect basically asserts that the coefficient on the expected inflation (b_1) variable equals 1, the hypotheses are structured as:

$$H_0: b_1 = 1$$

$$H_a: b_1 \neq 1$$

Quarterly data for 3-month T-bill returns, the nominal interest rate, are regressed on inflation rate expectations over the last 25 years. The results of the regression are:

	Coefficient	Standard error	*t*-statistic
Intercept	0.04	0.0051	7.843
Expected inflation	1.153	0.065	17.738
Residual standard error	0.029		
Multiple *R*-squared	0.45		
Observations	100		
Durbin-Watson statistic	0.547		

To determine whether the data support the assertions of the Fisher relation, we calculate the *t*-stat for the slope coefficient on expected inflation as:

$$t = \frac{\hat{b}_1 - b_1}{\text{Standard error}_{\hat{b}_1}} = \frac{1.153 - 1}{0.065} \approx 2.35$$

The critical *t*-values with 98 degrees of freedom at the 5% significance level are approximately −1.98 and +1.98. The test statistic is greater than the upper critical *t*-value, so we reject the null hypothesis and conclude that the Fisher effect does not hold, since the coefficient on expected inflation appears to be significantly different from 1.

However, before accepting the validity of the results of this test, we should test the null hypothesis that the regression errors do not suffer from conditional heteroskedasticity. A regression of the squared residuals from the original regression on expected inflation rates yields $R^2 = 0.193$.

The test statistic for the BP test is calculated as:

$$X^2 = nR^2 = 100 \times 0.193 = 19.3$$

The critical X^2 value at the 5% significance level for a one-tailed test with one degree of freedom is 3.84. Since the *t*-statistic (19.3) is higher, we reject the null hypothesis of no conditional heteroskedasticity in the error terms. Since conditional heteroskedasticity is present in the residuals (of the original regression), the standard errors calculated in the original regression are incorrect, and we cannot accept the result of the *t*-test above (which provides evidence against the Fisher relation) as valid.

Correcting Heteroskedasticity

With efficient markets, heteroskedasticity should not exist in financial data. However, when it can be observed, an analyst should not only look to correct for heteroskedasticity, but also understand it and try to capitalize on it.

There are two ways to correct for conditional heteroskedasticity in linear regression models. The first is to use robust standard errors, also known as White-corrected standard errors or heteroskedasticity-consistent standard errors, to recalculate the *t*-statistics for the original regression coefficients. The other method is to use generalized least squares, where the original regression equation is modified to eliminate heteroskedasticity.

Example 2 Using robust standard errors to adjust for conditional heteroskedasticity

The analyst corrects the standard errors from the initial regression of 3-month T-bill returns (nominal interest rate) on expected inflation rates for heteroskedasticity and obtains the following results:

	Coefficient	Standard error	*t*-statistic
Intercept	0.04	0.0048	8.333
Expected inflation	1.153	0.085	13.565
Residual standard error	0.029		
Multiple R-squared	0.45		
Observations	100		

Compared with the regression results in Example 1, notice that the standard error for the intercept does not change significantly, but the standard error for the coefficient on expected inflation increases by about 30% (from 0.065 to 0.085). Further, the regression coefficients remain the same (0.04 for the intercept and 1.153 for expected inflation).

Using the adjusted standard error for the slope coefficient, the test statistic for the hypothesis test is calculated as:

$$t = \frac{\hat{b}_1 - b_1}{\text{Standard error}_{\hat{b}_1}} = \frac{1.153 - 1}{0.085} \approx 1.8$$

When we compare this test statistic to the upper critical *t*-value (1.98), we fail to reject the null hypothesis since the upper value is greater than the test statistic. The conditional heteroskedasticity in the data was so significant that the result of our hypothesis test changed once the standard errors were corrected for heteroskedasticity. We can conclude that the Fisher effect holds since the slope coefficient of the expected inflation independent variable does not significantly differ from 1.

Violations of Regression Assumptions: Serial Correlation

LOS: Explain serial correlation and how it affects statistical inference.

Serial correlation (autocorrelation) occurs when regression errors are correlated, either positively or negatively, across observations, typically in time-series regressions.

The Consequences of Serial Correlation

Serial correlation results in incorrect estimates of the regression coefficients' standard errors. If none of the regressors, or independent variables, is a lagged value of the dependent variable, it will not affect the consistency of the estimated regression coefficients.

For example, when examining the Fisher relation, if we were to use the T-bill return for the previous month as an independent variable (even though the T-bill return that represents the nominal interest rate is actually the dependent variable in our regression model), serial correlation would cause all parameter estimates from the regression to be inconsistent.

- **Positive serial correlation** occurs when:
  - a positive residual from one estimate increases the likelihood of a positive residual in the next observation.
  - a negative residual from one observation raises the probability of a negative residual resulting from the next observation.
  - In either case, positive serial correlation will result in a stable pattern of residuals over time.
- **Negative serial correlation** occurs when a positive residual in one instance increases the likelihood of a negative residual in the next.

Positive serial correlation is the most common type found in regression models. Positive serial correlation does not affect the consistency of the estimated regression coefficients, but it does have an impact on statistical tests. It will cause the *F*-stat, which is used to test the overall significance of the regression, to be inflated because MSE will tend to underestimate the population error variance.

In addition, it will cause the standard errors for the regression coefficients to be underestimated, which results in larger *t*-values. Consequently, analysts may incorrectly reject null hypotheses, making Type I errors, and attach significance to relationships that are in fact not significant.

Testing for Serial Correlation

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

The DW test is a measure of autocorrelation that compares the squared differences of successive residuals with the sum of the squared residuals. This test is somewhat limited, however, because it only applies to first-order serial correlation.

The BG test is more robust because it can detect autocorrelation up to a pre-designated order *p*, where the error in period *t* is correlated with the error in period *t-p*. The null hypothesis of the BG test is that there is no serial correlation in the model's residuals up to lag *p*. The alternative hypothesis is that the correlation of residuals of at least one of the lags is different from zero and that serial correlation exists up to that order. The test statistic is approximately *F*-distributed with *n-p-k*-1 degrees of freedom where *p* is the number of lags.

Correcting Serial Correlation

There are two ways to correct for serial correlation in the regression residuals:

1. Adjust the coefficient standard errors to account for serial correlation: The regression coefficients remain the same but the standard errors change. This also corrects for heteroskedasticity. After correcting for positive serial correlation, the robust standard errors are larger than they were originally. Note that the DW stat still remains the same.
2. Modify the regression equation to eliminate the serial correlation.

Example 3 Correcting for serial correlation

The table shows the results of correcting the standard errors of the original regression for serial correlation and heteroskedasticity using Hansen's method:

	Coefficient	Standard error	*t*-statistic
Intercept	0.04	0.0088	4.545
Expected inflation	1.153	0.155	7.439
Residual standard error	0.029		
Multiple *R*-squared	0.45		
Observations	100		
Durbin-Watson statistic			0.547

Note that the coefficients for the intercept and slope are exactly the same (0.04 for the intercept and 1.153 for expected inflation) as in the original regression (Example 1). Further, note that the DW stat is the same (0.547), but the standard errors have been corrected (they are now much larger) to account for the positive serial correlation.

Given these new and more accurate coefficient standard errors, test the null hypothesis that the coefficient on the expected inflation independent variable equals 1. The test statistic for the hypothesis test is calculated as:

$$t = \frac{\hat{b}_1 - b_1}{\text{Standard error } \hat{b}_1} = (1.153 - 1)/0.155 \approx 0.98$$

The critical *t*-values with 98 degrees of freedom at the 5% significance level are approximately −1.98 and +1.98. Comparing the test statistic (0.987) with the upper critical t-value (+1.98) we fail to reject the null hypothesis and conclude that the Fisher effect holds as the slope coefficient on the expected inflation independent variable does not significantly differ from 1.

Note that this result is different from the result of the test we conducted using the standard errors of the original regression (which were affected by serial correlation and heteroskedasticity) in Example 2. Further, the result is the same as from the test conducted on White-corrected standard errors (which were corrected for heteroskedasticity) in Example 2.

Violations of Regression Assumptions: Multicollinearity

LOS: Explain multicollinearity and how it affects regression analysis.

Multicollinearity occurs when two or more independent variables, or combinations of independent variables, are highly correlated with each other. Multicollinearity can also be present even when there is an approximate linear relationship between two or more independent variables. This is a particular problem with financial and economic data because linear relationships are common.

Consequences of Multicollinearity

While multicollinearity does not affect the consistency of OLS estimates and regression coefficients, it does make them inaccurate and unreliable, as it becomes increasingly difficult to isolate the impact of each independent variable on the dependent variable. This results in inflated standard errors for the regression coefficients, which results in *t*-stats becoming too small and less reliable in rejecting the null hypothesis.

Detecting Multicollinearity

An indicator of multicollinearity is a high R^2 and a significant *F*-statistic, both of which indicate that the regression model overall does a good job of explaining the dependent variable, coupled with insignificant *t*-statistics of slope coefficients. These insignificant *t*-statistics indicate that the independent variables individually do not explain the variation in the dependent variable, although the high R^2 indicates that the model overall does a good job: a classic case of multicollinearity. The low *t*-statistics on the slope coefficients increase the chances of Type II errors: failure to reject the null hypothesis when it is false.

The **variance inflation factor (VIF)** can quantify multicollinearity issues. In a multiple regression, a VIF exists for each independent variable. Assume *k* independent variables and regress one independent variable on the $k - 1$ independent variables to obtain R^2 for the regression explained by the other $k - 1$ independent variables. The VIF for X_j is:

$$\text{VIF}_i = \frac{1}{1 - R_j^2}$$

For a given independent variable, X_j, the minimum VIF_j is 1, which occurs when R^2_j is 0. The minimum VIF_i means that there is no correlation between X_j and the remaining independent variables. However, VIF increases as the correlation increases. So the higher a variable's VIF is, the more likely it is that the variable can be predicted using another independent variable in the model, making it more likely to be redundant.

The following are useful rules of thumb:

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

Bear in mind that multicollinearity may be present even when we do not observe insignificant *t*-stats and a highly significant *F*-stat for the regression model.

Example 4 Testing for multicollinearity

An individual is trying to determine how closely associated her portfolio manager's investment strategy is with the returns of a value index and the returns of a growth index over the last 60 years. She regresses the historical annual returns of her portfolio on the historical returns of the S&P 500/BARRA Growth Index, S&P 500/BARRA Value Index, and the S&P 500. The results of her regression are:

Regression coefficient	*t*-statistic
Intercept	1.250
S&P 500/BARRA Growth Index	−0.825
S&P 500/BARRA Value Index	−0.756
S&P 500 Index	1.520
F-statistic	35.17
R^2	82.34%
Observations	60

Evaluate the results of the regression.

Solution

The absolute values of the *t*-stats for all the regression coefficients—the intercept (1.25), slope coefficient on the growth index (0.825), slope coefficient on the value index (0.756), and slope coefficient on the S&P 500 (1.52)—are lower than the absolute value of t_{crit} (2.00) at the 5% level of significance (df = 56). This suggests that none of the coefficients on the independent variables in the regression are significantly different from 0.

However, the *F*-stat (35.17) is greater than the *F* critical value of 2.76 ($\alpha = 0.05$, $df = 3, 56$), which suggests that the slope coefficients on the independent variables do not jointly equal 0 (at least one of them is significantly different from 0). Further, the R^2 (82.34%) is quite high, which means that the model as a whole does a good job of explaining the variation in the portfolio's returns.

This regression, therefore, clearly suffers from the classic case of multicollinearity as described earlier.