Similar to other hypothesis tests, the chi-squared test compares the test statistic, χ^2_{n-1}, to a critical chi-squared value at a given level of significance and n – 1 degrees of freedom.

Example: Chi-squared test for a single population variance

Historically, High-Return Equity Fund has advertised that its monthly returns have a standard deviation equal to 4%. This was based on estimates from the 1990–1998 period. High-Return wants to verify whether this claim still adequately describes the standard deviation of the fund's returns. High-Return collected monthly returns for the 24-month period between 1998 and 2000 and measured a standard deviation of monthly returns of 3.8%. **Determine** if the more recent standard deviation is different from the advertised standard deviation.

Answer:

State the hypothesis. The null hypothesis is that the standard deviation is equal to 4% and, therefore, the variance of monthly returns for the population is $(0.04)^2 = 0.0016$. Since High-Return simply wants to test whether the standard deviation has changed, up or down, a two-sided test should be used. The hypothesis test structure takes the form:

$H_0: \sigma^2 = 0.0016$ versus $H_A: \sigma^2 \neq 0.0016$

Select the appropriate test statistic. The appropriate test statistic for tests of variance using the chi-squared distribution is computed as follows:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

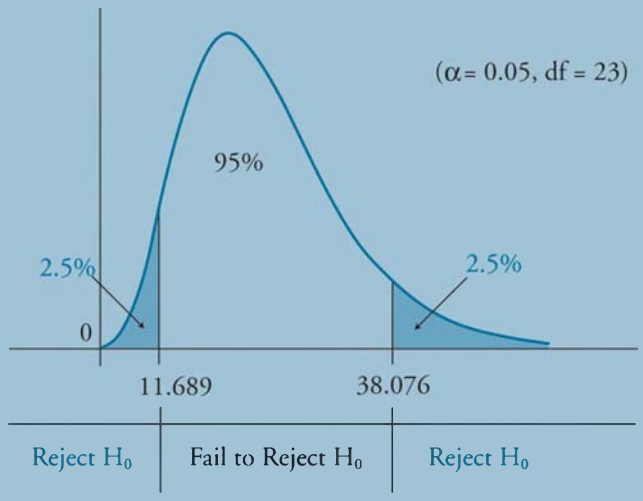
Specify the level of significance. Let's use a 5% level of significance, meaning there will be 2.5% probability in each tail of the chi-squared distribution.

State the decision rule regarding the hypothesis. With a 24-month sample, there are 23 degrees of freedom. Using the table of chi-squared values at the back of this book, for 23 degrees of freedom and probabilities of 0.975 and 0.025, we find two critical values, 11.689 and 38.076. Thus, the decision rule is:

Reject H_0 if: $\chi^2 < 11.689$, or $\chi^2 > 38.076$

This decision rule is illustrated in the following distribution.

Decision Rule for a Two-Tailed Chi-Squared Test of a Single Population Variance



Collect the sample and calculate the sample statistics. Using the information provided, the test statistic is computed as:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(23)(0.001444)}{0.0016} = \frac{0.033212}{0.0016} = 20.7575$$

Make a decision regarding the hypothesis. Since the computed test statistic, χ^2, falls between the two critical values, we fail to reject the null hypothesis that the variance is equal to 0.0016.

Make a decision based on the results of the test. It can be concluded that the recently measured standard deviation is close enough to the advertised standard deviation that we cannot say it is different from 4%, at a 5% level of significance.

The *F*-Test

The hypotheses concerned with the equality of the variances of two populations are tested with an *F*-distributed test statistic. Hypothesis testing using a test statistic that follows an *F*-distribution is referred to as the *F*-test. The *F*-test is used under the assumption that the populations from which samples are drawn are normally distributed and that the samples are independent.

If we let σ_1^2 and σ_2^2 represent the variances of normal Population 1 and Population 2, respectively, the hypotheses for the two-tailed *F*-test of differences in the variances can be structured as:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_A: \sigma_1^2 \neq \sigma_2^2$$

and the one-sided test structures can be specified as:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ versus } H_A: \sigma_1^2 > \sigma_2^2, \text{ or } H_0: \sigma_1^2 \geq \sigma_2^2 \text{ versus } H_A: \sigma_1^2 < \sigma_2^2$$

The test statistic for the *F*-test is the ratio of the sample variances. The *F*-statistic is computed as:

$$F = \frac{s_1^2}{s_2^2}$$

where:

s_1^2 = variance of the sample of n_1 observations drawn from Population 1

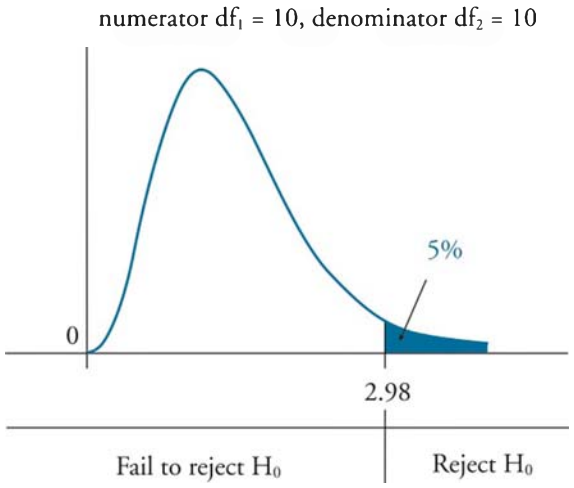
s_2^2 = variance of the sample of n_2 observations drawn from Population 2

Note that $n_1 - 1$ and $n_2 - 1$ are the degrees of freedom used to identify the appropriate critical value from the *F*-table (provided in the Appendix).

Professor's Note: Always put the larger variance in the numerator (s_1^2). Following this convention means we only have to consider the critical value for the right-hand tail.

An *F*-distribution is presented in Figure 10. As indicated, the *F*-distribution is right-skewed and is truncated at zero on the left-hand side. The shape of the *F*-distribution is determined by *two separate degrees of freedom*, the numerator degrees of freedom, df_1, and the denominator degrees of freedom, df_2. Also shown in Figure 10 is that the *rejection region is in the right-side tail* of the distribution. This will always be the case as long as the *F*-statistic is computed with the largest sample variance in the numerator. The labeling of 1 and 2 is arbitrary anyway.

Figure 10: *F*-Distribution



Example: *F*-test for equal variances

Annie Cower is examining the earnings for two different industries. Cower suspects that the earnings of the textile industry are more divergent than those of the paper industry. To confirm this suspicion, Cower has looked at a sample of 31 textile manufacturers and a sample of 41 paper companies. She measured the sample standard deviation of earnings across the textile industry to be $4.30 and that of the paper industry companies to be $3.80. **Determine** if the earnings of the textile industry have greater standard deviation than those of the paper industry.

Answer:

State the hypothesis. In this example, we are concerned with whether the variance of the earnings of the textile industry is greater (more divergent) than the variance of the earnings of the paper industry. As such, the test hypotheses can be appropriately structured as:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ versus } H_A: \sigma_1^2 > \sigma_2^2$$

where:

σ_1^2 = variance of earnings for the textile industry

σ_2^2 = variance of earnings for the paper industry

Note: $\sigma_1^2 > \sigma_2^2$

Select the appropriate test statistic. For tests of difference between variances, the appropriate test statistic is:

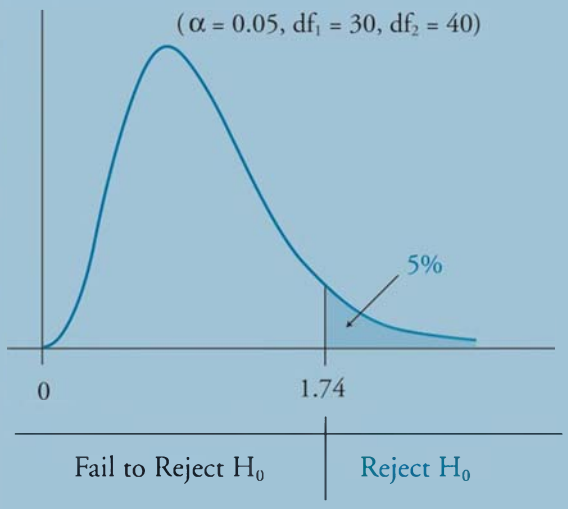
$$F = \frac{s_1^2}{s_2^2}$$

Specify the level of significance. Let's conduct our hypothesis test at the 5% level of significance.

State the decision rule regarding the hypothesis. Using the sample sizes for the two industries, the critical *F*-value for our test is found to be 1.74. This value is obtained from the table of the *F*-distribution at the 5% level of significance with $df_1 = 30$ and $df_2 = 40$. Thus, if the computed *F*-statistic is greater than the critical value of 1.74, the null hypothesis is rejected. The decision rule, illustrated in the distribution below, can be stated as:

Reject H_0 if: $F > 1.74$

Decision Rule for *F*-Test



Collect the sample and calculate the sample statistics. Using the information provided, the *F*-statistic can be computed as:

$$F = \frac{s_1^2}{s_2^2} = \frac{\$4.30^2}{\$3.80^2} = \frac{\$18.49}{\$14.44} = 1.2805$$

Professor's Note: Remember to square the standard deviations to get the variances.

Make a decision regarding the hypothesis. Since the calculated *F*-statistic of 1.2805 is less than the critical *F*-statistic of 1.74, we fail to reject the null hypothesis.

Make a decision based on the results of the test. Based on the results of the hypothesis test, Cower should conclude that the earnings variances of the industries are not statistically significantly different from one another at a 5% level of significance. More pointedly, the earnings of the textile industry are not more divergent than those of the paper industry.

Chebyshev's Inequality

Chebyshev's inequality states that for any set of observations, whether sample or population data and regardless of the shape of the distribution, the percentage of the observations that lie within k standard deviations of the mean is *at least* $1 - 1/k^2$ for all $k > 1$.

Example: Chebyshev's inequality

What is the minimum percentage of any distribution that will lie within ±2 standard deviations of the mean?

Answer:

Applying Chebyshev's inequality, we have:

$$1 - 1/k^2 = 1 - 1/2^2 = 1 - 1/4 = 0.75 \text{ or } 75\%$$

According to Chebyshev's inequality, the following relationships hold for any distribution. At least:

- 36% of observations lie within ±1.25 standard deviations of the mean.
- 56% of observations lie within ±1.50 standard deviations of the mean.
- 75% of observations lie within ±2 standard deviations of the mean.
- 89% of observations lie within ±3 standard deviations of the mean.
- 94% of observations lie within ±4 standard deviations of the mean.

The importance of Chebyshev's inequality is that it applies to any distribution. If we know the underlying distribution is actually normal, we can be even more precise about the percentage of observations that will fall within a given number of standard deviations of the mean.

Note that with a normal distribution, extreme events beyond ±3 standard deviations are very rare (occurring only 0.26% of the time). However, as Chebyshev's inequality points out, events that are ±3 standard deviations may not be so rare for nonnormal distributions (potentially occurring 11% of the time). Therefore, simply assuming normality, without knowing the parameters of the underlying distribution, could lead to a severe underestimation of risk.

Backtesting

LO 19.6: Demonstrate the process of backtesting VaR by calculating the number of exceedances.

The process of **backtesting** involves comparing expected outcomes against actual data. For example, if we apply a 95% confidence interval, we are expecting an event to exceed the confidence interval with a 5% probability. Recall that the 5% in this example is known as the level of significance.

It is common for risk managers to backtest their value at risk (VaR) models to ensure that the model is forecasting losses with the same frequency predicted by the confidence interval (VaR models typically use a 95% confidence interval). When the VaR measure is exceeded during a given testing period, it is known as an **exception** or an **exceedance**. After backtesting the VaR model, if the number of exceptions if greater than expected, the risk manager may be underestimating actual risk. Conversely, if the number of exceptions is less than expected, the risk manager may be overestimating actual risk.

Example: Calculating the number of exceedances

Assume that the value at risk (VaR) of a portfolio, at a 95% confidence interval, is $100 million. Also assume that given a 100-day trading period, the actual number of daily losses exceeding $100 million occurred eight times. Is this VaR model underestimating or overestimating the actual level of risk?

Answer:

With a 95% confidence interval, we expect to have exceptions (i.e., losses exceeding $100 million) 5% of the time. If the losses exceeding $100 million occurred eight times during the 100-day period, exceptions occurred 8% of the time. Therefore, this VaR model is underestimating risk because the number of exceptions is greater than expected according to the 95% confidence interval.

One of the main issues with backtesting VaR models is that exceptions are often serially correlated. In other words, there is a high probability that an exception will occur after the previous period had an exception. Another issue is that the occurrence of exceptions tends to be correlated with overall market volatility. In other words, VaR exceptions tend to be higher (lower) when market volatility is high (low). This may be the result of a VaR model failing to quickly react to changes in risk levels.

Professor's Note: We will discuss VaR methodologies and backtesting VaR in more detail in Book 4.

Key Concepts

LO 19.1

$$\text{Population variance} = \sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}, \text{ where } \mu = \text{population mean and } N = \text{size}$$

$$\text{Sample variance} = s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, \text{ where } \bar{X} = \text{sample mean and } n = \text{sample size}$$

The standard error of the sample mean is the standard deviation of the distribution of the sample means and is calculated as $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, where σ, the population standard deviation, is known, and as $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s, the sample standard deviation, is used because the population standard deviation is unknown.

LO 19.2

For a normally distributed population, a confidence interval for its mean can be constructed using a *z*-statistic when variance is known, and a *t*-statistic when the variance is unknown. The *z*-statistic is acceptable in the case of a normal population with an unknown variance if the sample size is large (30+).

In general, we have:

- $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ when the variance is known
- $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ when the variance is unknown

LO 19.3

The hypothesis testing process requires a statement of a null and an alternative hypothesis, the selection of the appropriate test statistic, specification of the significance level, a decision rule, the calculation of a sample statistic, a decision regarding the hypotheses based on the test, and a decision based on the test results.

The test statistic is the value that a decision about a hypothesis will be based on. For a test about the value of the mean of a distribution:

$$\text{test statistic} = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{standard error of sample mean}}$$

With unknown population variance, the *t*-statistic is used for tests about the mean of a normally distributed population: $t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. If the population variance is known, the appropriate test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ for tests about the mean of a population.

LO 19.4

A two-tailed test results from a two-sided alternative hypothesis (e.g., $H_A: \mu \neq \mu_0$). A one-tailed test results from a one-sided alternative hypothesis (e.g., $H_A: \mu > \mu_0$, or $H_A: \mu < \mu_0$).

LO 19.5

Hypothesis testing compares a computed test statistic to a critical value at a stated level of significance, which is the decision rule for the test.

A hypothesis about a population parameter is rejected when the sample statistic lies outside a confidence interval around the hypothesized value for the chosen level of significance.

LO 19.6

Backtesting is the process of comparing losses predicted by the value at risk (VaR) model to those actually experienced over the sample testing period. If a model were completely accurate, we would expect VaR to be exceeded with the same frequency predicted by the confidence level used in the VaR model. In other words, the probability of observing a loss amount greater than VaR should be equal to the level of significance.

Concept Checkers

1. An analyst observes that the variance of daily stock returns for Stock X during a certain period is 0.003. He assumes daily stock returns are normally distributed and wants to conduct a hypothesis test to determine whether the variance of daily returns on Stock X is different from 0.005. The analyst looks up the critical values for his test, which are 9.59 and 34.17. He calculates a test statistic of 11.40 for his set of data. What kind of test statistic did the analyst calculate, and should he conclude that the variance is different from 0.005?

	Test statistic	Variance ≠ 0.005
A.	*t*-statistic	Yes
B.	Chi-squared statistic	Yes
C.	*t*-statistic	No
D.	Chi-squared statistic	No

Use the following data to answer Questions 2 and 3.

Austin Roberts believes the mean price of houses in the area is greater than $145,000. A random sample of 36 houses in the area has a mean price of $149,750. The population standard deviation is $24,000, and Roberts wants to conduct a hypothesis test at a 1% level of significance.

2. The appropriate alternative hypothesis is:
   A. $H_A: \mu < \$145{,}000$.
   B. $H_A: \mu \pm \$145{,}000$.
   C. $H_A: \mu \geq \$145{,}000$.
   D. $H_A: \mu > \$145{,}000$.

3. The value of the calculated test statistic is closest to:
   A. $z = 0.67$.
   B. $z = 1.19$.
   C. $z = 4.00$.
   D. $z = 8.13$.

4. The 95% confidence interval of the sample mean of employee age for a major corporation is 19 years to 44 years based on a *z*-statistic. The population of employees is more than 5,000 and the sample size of this test is 100. Assuming the population is normally distributed, the standard error of mean employee age is closest to:
   A. 1.96.
   B. 2.58.
   C. 6.38.
   D. 12.50.

Use the following data to answer Question 5.

XYZ Corp. Annual Stock Prices					
1995	1996	1997	1998	1999	2000
22%	5%	–7%	11%	2%	11%

5. Assuming the distribution of XYZ stock returns is a sample, what is the sample standard deviation?
   A. 7.4%.
   B. 9.8%.
   C. 72.4%.
   D. 96.3%.

Concept Checker Answers

1. **D** Hypothesis tests concerning the variance of a normally distributed population use the chi-squared statistic. The null hypothesis is that the variance is equal to 0.005. Since the test statistic falls within the range of the critical values, the test fails to reject the null hypothesis. The analyst cannot conclude that the variance of daily returns on Stock X is different from 0.005.

2. **D** H_A: $\mu > \$145{,}000$.

3. **B** $$z = \frac{149{,}750 - 145{,}000}{24{,}000 / \sqrt{36}} = 1.1875.$$

4. **C** At the 95% confidence level, with sample size n = 100 and mean 31.5 years, the appropriate test statistic is $z_{\alpha/2} = 1.96$. *Note: The mean of 31.5 is calculated as the midpoint of the interval, or (19 + 44) / 2.* Thus, the confidence interval is $31.5 \pm 1.96s_x$, where s_x is the standard error of the sample mean. If we take the upper bound, we know that $31.5 + 1.96s_x = 44$, or $1.96s_x = 12.5$, or $s_x = 6.38$ years.

5. **B** The sample standard deviation is the square root of the sample variance:

$$s = \left\{ \frac{\left[(22-7.3)^2 + (5-7.3)^2 + (-7-7.3)^2 + (11-7.3)^2 + (2-7.3)^2 + (11-7.3)^2 \right]}{6-1} \right\}^{1/2}$$

$$= \left(96.3\%^2 \right)^{1/2} = 9.8\%$$

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. This topic is also covered in:

Linear Regression with One Regressor

Topic 20

Exam Focus

Linear regression refers to the process of representing relationships with linear equations where there is one dependent variable being explained by one or more independent variables. There will be deviations from the expected value of the dependent variable called error terms, which represent the effect of independent variables not included in the population regression function. Typically we do not know the population regression function; instead, we estimate it with a method such as ordinary least squares (OLS). For the exam, be able to apply the concepts of simple linear regression and understand how sample data can be used to estimate population regression parameters (i.e., the intercept and slope of the linear regression).

Regression Analysis

LO 20.1: Explain how regression analysis in econometrics measures the relationship between dependent and independent variables.

A regression analysis has the goal of measuring how changes in one variable, called a **dependent** or **explained** variable can be explained by changes in one or more other variables called the **independent** or **explanatory** variables. The regression analysis measures the relationship by estimating an equation (e.g., linear regression model). The **parameters** of the equation indicate the relationship.

A **scatter plot** is a visual representation of the relationship between the dependent variable and a given independent variable. It uses a standard two-dimensional graph where the values of the dependent, or Y variable, are on the vertical axis, and those of the independent, or X variable, are on the horizontal axis.

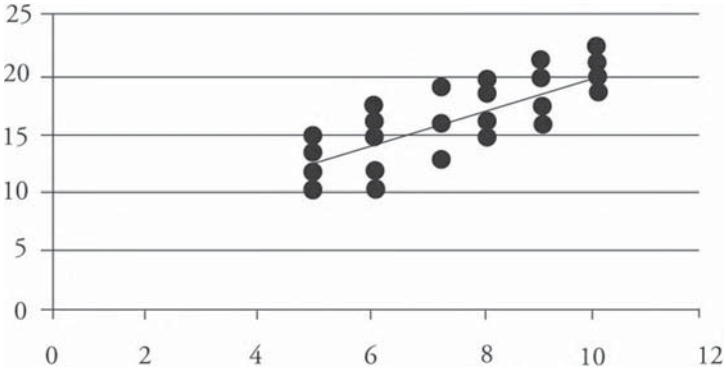
A scatter plot can indicate the nature of the relationship between the dependent and independent variable. The most basic property indicated by a scatter plot is whether there is a positive or negative relationship between the dependent variable and the independent variable. A closer inspection can indicate if the relationship is linear or nonlinear.

As an example, let us assume that we have access to all the returns data for a certain class of hedge funds over a given year. The population consists of 30 hedge funds that follow the same strategy, but they differ by the length of the lockup period. The lockup period is the minimum number of years an investor must keep funds invested. For this given strategy of hedge funds, the lockup periods range from five to ten years. Figure 1 contains the hedge fund data, and Figure 2 is a scatter plot that illustrates the relationship.

Figure 1: Hedge Fund Data

Lockup (yrs)	*Returns (%) per year*					*Average Return*
5	10	14	14	15	12	13
6	17	12	15	16	10	14
7	16	19	19	13	13	16
8	15	20	19	15	16	17
9	21	20	16	20	18	19
10	20	17	21	23	19	20

Figure 2: Return Over Lockup Period



The scatter plot indicates that there is a positive relationship between the hedge fund returns and the lockup period. We should keep in mind that the data represents returns over the same period (i.e., one year). The factor that varies is the amount of time a manager knows that he will control the funds. One interpretation of the graph could be that managers who know that they can control the funds over a longer period can engage in strategies that reap a higher return in any given year. As a final note, the scatter plot in this example indicates a fairly linear relationship. With each 1-year increase in the lockup period, according to the graph, the corresponding returns seem to increase by a similar amount.

Population Regression Function

LO 20.2: Interpret a population regression function, regression coefficients, parameters, slope, intercept, and the error term.

Assuming that the 30 observations represent the population of hedge funds that are in the same class (i.e., have the same basic investment strategy) then their relationship can provide a **population regression function**. Such a function would consist of parameters called **regression coefficients**. The regression equation (or function) will include an intercept term and one slope coefficient for each independent variable. For this simple two-variable case, the function is:

$$E(\text{return} \mid \text{lockup period}) = B_0 + B_1 \times (\text{lockup period})$$

Or more generally:

$$E(Y_i \mid X_i) = B_0 + B_1 \times (X_i)$$

In the equation, B_0 is the **intercept coefficient**, which is the expected value of the return if $X = 0$. B_1 is the **slope coefficient**, which is the expected change in *Y* for a unit change in *X*. In this example, for every additional year of lockup, a hedge fund is expected to earn an additional B_1 per year in return.

The Error Term

There is a dispersion of *Y*-values around each conditional expected value. The difference between each *Y* and its corresponding conditional expectation (i.e., the line that fits the data) is the **error term** or **noise component** denoted ε_i.

$$\varepsilon_i = Y_i - E(Y_i \mid X_i)$$

The deviation from the expected value is the result of factors other than the included *X*-variable. One way to break down the equation is to say that $E(Y_i \mid X_i) = B_0 + B_1 \times X_i$ is the deterministic or systematic component, and ε_i is the nonsystematic or random component. The error term provides another way of expressing the population regression function:

$$Y_i = B_0 + B_1 \times X_i + \varepsilon_i$$

The error term represents effects from independent variables not included in the model. In the case of the hedge fund example, ε_i is probably a function of the individual manager's unique trading tactics and management activities within the style classification. Variables that might explain this error term are the number of positions and trades a manager makes over time. Another variable might be the years of experience of the manager. An analyst may need to include several of these variables (e.g., trading style and experience) into the population regression function to reduce the error term by a noticeable amount. Often, it is found that limiting an equation to the one or two independent variables with the most explanatory power is the best choice.

Sample Regression Function

LO 20.3: Interpret a sample regression function, regression coefficients, parameters, slope, intercept, and the error term.

The **sample regression function** is an equation that represents a relationship between the *Y* and *X* variable(s) that is based only on the information in a sample of the population. In almost all cases the slope and intercept coefficients of a sample regression function will be different from that of the population regression function. If the sample of *X* and *Y* variables is truly a random sample, then the difference between the sample coefficients

and the population coefficients will be random too. There are various ways to use notation to distinguish the components of the sample regression function from the population regression function. Here we have denoted the population parameters with capital letters (i.e., B_0 and B_1) and the sample coefficients with small letters as indicated in the following sample regression function:

$$Y_i = b_0 + b_1 \times X_i + e_i$$

The sample regression coefficients are b_0 and b_1, which are the intercept and slope. There is also an extra term on the end called the **residual**: $e_i = Y_i - (b_0 + b_1 \times X_i)$. Since the population and sample coefficients are almost always different, the residual will very rarely equal the corresponding population error term (i.e., generally $e_i \neq \varepsilon_i$).

Properties of Regression

LO 20.4: Describe the key properties of a linear regression.

Under certain, basic assumptions, we can use a linear regression to estimate the population regression function. The term "linear" has implications for both the independent variable and the coefficients. One interpretation of the term *linear* relates to the independent variable(s) and specifies that the independent variable(s) enters into the equation without a transformation such as a square root or logarithm. If it is the case that the relationship between the dependent variable and an independent variable is non-linear, then an analyst would do that transformation first and then enter the transformed value into the linear equation as *X*. For example, in estimating a utility function as a function of consumption, we might allow for the property of diminishing marginal utility by transforming consumption into a logarithm of consumption. In other words, the actual relationship is:

$$E(\text{utility} \mid \text{amount consumed}) = B_0 + B_1 \times \ln(\text{amount consumed})$$

Here we let Y = utility and X = ln (amount consumed) and estimate: $E(Y_i \mid X_i) = B_0 + B_1 \times (X_i)$ using linear techniques.

A second interpretation for the term linear applies to the parameters. It specifies that the dependent variable is a linear function of the parameters, but does not require that there is linearity in the variables. Two examples of non-linear relationships are as follows:

$$E(Y_i \mid X_i) = B_0 + (B_1)^2 \times (X_i)$$

$$E(Y_i \mid X_i) = B_0 + (1/B_1) \times (X_i)$$

It would not be appropriate to apply linear regression to estimate the parameters of these functions. The primary concern for linear models is that they display linearity in the parameters. Therefore, when we refer to a linear regression model we generally assume that the equation is linear in the parameters; it may or may not be linear in the variables.

Ordinary Least Squares Regression

LO 20.5: Define an ordinary least squares (OLS) regression and calculate the intercept and slope of the regression.

Ordinary least squares (OLS) estimation is a process that estimates the population parameters B_i with corresponding values for b_i that minimize the squared residuals (i.e., error terms). Recall the expression $e_i = Y_i - (b_0 + b_1 \times X_i)$; the OLS sample coefficients are those that:

$$\text{minimize } \sum e_i^2 = \Sigma[Y_i - (b_0 + b_1 \times X_i)]^2$$

The estimated **slope coefficient** (b_1) for the regression line describes the change in *Y* for a one unit change in *X*. It can be positive, negative, or zero, depending on the relationship between the regression variables. The slope term is calculated as:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

The **intercept** term (b_0) is the line's intersection with the *Y*-axis at X = 0. It can be positive, negative, or zero. A property of the least squares method is that the intercept term may be expressed as:

$$b_0 = \overline{Y} - b_1\overline{X}$$

where:
$\overline{Y}$ = mean of Y
$\overline{X}$ = mean of X

The intercept equation highlights the fact that the regression line passes through a point with coordinates equal to the mean of the independent and dependent variables (i.e., the point, $\overline{X}, \overline{Y}$).

Assumptions Underlying Linear Regression

LO 20.6: Describe the method and three key assumptions of OLS for estimation of parameters.

OLS regression requires a number of assumptions. Most of the major assumptions pertain to the regression model's residual term (i.e., error term). Three key assumptions are as follows:

- The expected value of the error term, conditional on the independent variable, is zero ($E(\varepsilon_i|X_i) = 0$).
- All (X, Y) observations are independent and identically distributed (i.i.d.).
- It is unlikely that large outliers will be observed in the data. Large outliers have the potential to create misleading regression results.

Additional assumptions include:

- A linear relationship exists between the dependent and independent variable.
- The model is correctly specified in that it includes the appropriate independent variable and does not omit variables.
- The independent variable is uncorrelated with the error terms.
- The variance of ε_i is constant for all X_i: $Var(\varepsilon_i|X_i) = \sigma^2$.
- No serial correlation of the error terms exists [i.e., $Corr(\varepsilon_i, \varepsilon_{i+j}) = 0$ for j=1, 2, 3…]. The point being that knowing the value of an error for one observation does not reveal information concerning the value of an error for another observation.
- The error term is normally distributed.

Properties of OLS Estimators

LO 20.7: Summarize the benefits of using OLS estimators.

OLS estimators and terminology are used widely in practice when applying regression analysis techniques. In fields such as economics, finance, and statistics, the presentation of OLS regression results is the same. This means that the calculation of b_0 and b_1 and the interpretation and analysis of regression output is easily understood across multiple fields of study. As a result, statistical software packages make it easy for users to apply OLS estimators. In addition to practical benefits, OLS estimators also have theoretical benefits. OLS estimated coefficients are unbiased, consistent, and (under special conditions) efficient. Recall from Topic 16, that these characteristics are desirable properties of an estimator.

LO 20.8: Describe the properties of OLS estimators and their sampling distributions, and explain the properties of consistent estimators in general.

Since OLS estimators are derived from random samples, these estimators are also random variables because they vary from one sample to the next. Therefore, OLS estimators will have their own probability distributions (i.e., sampling distributions). These sampling distributions allow us to estimate population parameters, such as the population mean, the population regression intercept term, and the population regression slope coefficient.

Drawing multiple samples from a population will produce multiple sample means. The distribution of these sample means is referred to as the *sampling distribution of the sample mean*. The mean of this sampling distribution is used as an estimator of the population mean and is said to be an **unbiased estimator** of the population mean. Recall that an unbiased estimator is one for which the expected value of the estimator is equal to the parameter you are trying to estimate.

Given the **central limit theorem**, for large sample sizes, it is reasonable to assume that the sampling distribution will approach the normal distribution. This means that the estimator is also a **consistent estimator**. Recall that a consistent estimator is one for which the accuracy of the parameter estimate increases as the sample size increases. Note that a general guideline for a large sample size in regression analysis is a sample greater than 100.

Like the sampling distribution of the sample mean, OLS estimators for the population intercept term and slope coefficient also have sampling distributions. The sampling distributions of OLS estimators, b_0 and b_1, are unbiased and consistent estimators of population parameters, B_0 and B_1. Being able to assume that b_0 and b_1 are normally distributed is a key property in allowing us to make statistical inferences about population coefficients.

OLS Regression Results

LO 20.9: Interpret the explained sum of squares, the total sum of squares, the residual sum of squares, the standard error of the regression, and the regression R^2.

LO 20.10: Interpret the results of an OLS regression.

The **sum of squared residuals** (SSR), sometimes denoted SSE, for sum of squared errors, is the sum of squares that results from placing a given intercept and slope coefficient into the equation and computing the residuals, squaring the residuals and summing them. It is represented by $\sum e_i^2$. The sum is an indicator of how well the sample regression function explains the data.

Assuming certain conditions exist, an analyst can use the results of an ordinary least squares regression in place of the unknown population regression function to describe the relationship between the dependent and independent variable(s). In our earlier example concerning hedge fund returns and lockup periods, we might assume that an analyst only has access to a sample of returns data (e.g., six observations). This may be the result of the fact that hedge funds are not regulated and the reporting of returns is voluntary. In any case, we will assume that the data in Figure 3 is the sample of six observations and includes the corresponding computations for computing OLS estimates.

Figure 3: Sample of Returns and Corresponding Lockup Periods

	Lockup	*Returns*	$(X - \bar{X})$	$(Y - \bar{Y})$	*Cov(X,Y)*	*Var(X)*
	5	10	–2.5	–6	15	6.25
	6	12	–1.5	–4	6	2.25
	7	19	–0.5	3	–1.5	0.25
	8	16	0.5	0	0	0.25
	9	18	1.5	2	3	2.25
	10	21	2.5	5	12.5	6.25
Sum	45	96	0	0	35	17.50
Average	7.5	16				

From Figure 3, we can compute the sample coefficients:

$$b_1 = \frac{35}{17.5} = 2$$

$$b_0 = 16 - 2 \times 7.5 = 1$$

Thus, the sample regression function is: $Y_i = 1 + 2 \times X_i + e_i$. This means that, according to the data, on average a hedge fund with a lockup period of six years will have a 2% higher return than a hedge fund with a 5-year lockup period.

The Coefficient of Determination

The **coefficient of determination**, represented by R^2, is a measure of the "goodness of fit" of the regression. It is interpreted as a percentage of variation in the dependent variable explained by the independent variable. The underlying concept is that for the dependent variable, there is a total sum of squares (TSS) around the sample mean. The regression equation explains some portion of that TSS. Since the explained portion is determined by the independent variables, which are assumed independent of the errors, the total sum of squares can be broken down as follows:

$$\text{Total sum of squares} = \text{explained sum of squares} + \text{sum of squared residuals}$$

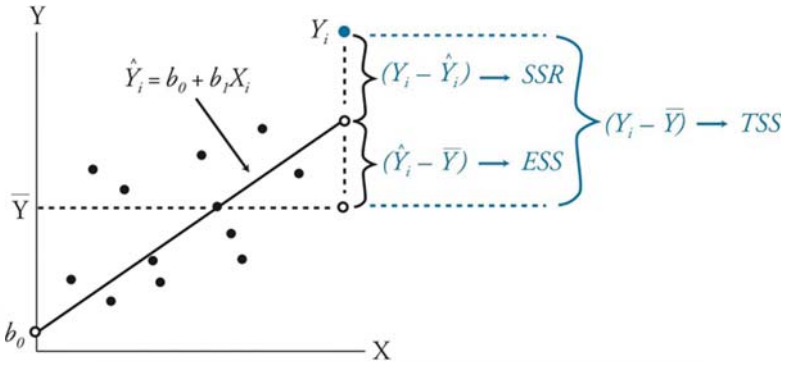
$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y_i - \hat{Y})^2$$

$$\text{TSS} = \text{ESS} + \text{SSR}$$

Professor's Note: As mentioned previously, sum of squared residuals (SSR) is also known as the sum of squared errors (SSE). In the same regard, total sum of squares (TSS) is also known as sum of squares total (SST), and explained sum of squares (ESS) is also known as regression sum of squares (RSS).

Figure 4 illustrates how the total variation in the dependent variable (TSS) is composed of SSR and ESS.

Figure 4: Components of the Total Variation



The coefficient of determination can be calculated as follows:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Example: Computing R^2

Figure 5 contains the relevant information from our hedge fund example where the average of the hedge fund returns was 16% (i.e., $\bar{Y} = 16$). **Compute** the coefficient of determination for the hedge fund regression line.

Figure 5: Computing the Coefficient of Determination

	Lockup	Returns, Y_i	e_i	e_i^2	$\sum(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$\sum(Y_i - \hat{Y}_i)^2$
	5	10	–1	1	36	11	1
	6	12	–1	1	16	13	1
	7	19	4	16	9	15	16
	8	16	–1	1	0	17	1
	9	18	–1	1	4	19	1
	10	21	0	0	25	21	0
Sum	45	96	0	20	90	96	20

Answer:

The coefficient of determination is 77.8%, which is calculated as follows:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{20}{90} = 0.778$$

In a simple two-variable regression, the square root of R^2 is the **correlation coefficient** (r) between X_i and Y_i. If the relationship is positive, then:

$$r = \sqrt{R^2}$$

For the hedge fund data, the correlation coefficient is: $r = \sqrt{0.778} = 0.882$

The correlation coefficient is a standard measure of the strength of the linear relationship between two variables. Initially it may seem similar to the coefficient of determination, but it is not for two reasons. First, the correlation coefficient indicates the sign of the relationship, whereas the coefficient of determination does not. Second, the coefficient of determination can apply to an equation with several independent variables, and it implies a causation or explanatory power, while the correlation coefficient only applies to two variables and does not imply causation between the variables.

The Standard Error of the Regression

The standard error of the regression (SER) measures the degree of variability of the actual Y-values relative to the estimated Y-values from a regression equation. The SER gauges the "fit" of the regression line. The smaller the standard error, the better the fit.

The SER is the standard deviation of the error terms in the regression. As such, SER is also referred to as the standard error of the residual, or the standard error of estimate (SEE).

In some regressions, the relationship between the independent and dependent variables is very strong (e.g., the relationship between 10-year Treasury bond yields and mortgage rates). In other cases, the relationship is much weaker (e.g., the relationship between stock returns and inflation). SER will be low (relative to total variability) if the relationship is very strong and high if the relationship is weak.

Key Concepts

LO 20.1

Regression analysis attempts to measure the relationship between a dependent variable and one or more independent variables.

A scatter plot (a.k.a. scattergram) is a collection of points on a graph where each point represents the values of two variables (i.e., an X/Y pair).

LO 20.2

A population regression line indicates the expected value of a dependent variable conditional on one or more independent variables: $E(Y_i \mid X_i) = B_0 + B_1 \times (X_i)$.

The difference between an actual dependent variable and a given expected value is the error term or noise component denoted $\varepsilon_i = Y_i - E(Y_i \mid X_i)$.

LO 20.3

The sample regression function is an equation that represents a relationship between the *Y* and *X* variable(s) using only a sample of the total data. It uses symbols that are similar but still distinct from that of the population $Y_i = b_0 + b_1 \times X_i + e_i$.

LO 20.4

In a linear regression model, we generally assume that the equation is linear in the parameters, and that it may or may not be linear in the variables.

LO 20.5

Ordinary least squares estimation is a process that estimates the population parameters B_i with corresponding values for b_i that minimize $\sum e_i^2 = \Sigma[Y_i - (b_0 + b_1 \times X_i)]^2$. The formulas for the coefficients are:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

LO 20.6

Three key assumptions made with simple linear regression include:

- The expected value of the error term, conditional on the independent variable, is zero.
- All (X, Y) observations are independent and identically distributed (i.i.d.).
- It is unlikely that large outliers will be observed in the data.

LO 20.7

OLS estimators are used widely in practice. In addition to practical benefits, OLS estimators exhibit desirable properties of an estimator.

LO 20.8

Since OLS estimators are random variables, they have their own sampling distributions. These sampling distributions are used to estimate population parameters. Given that the expected value of the estimator is equal to the parameter being estimated and the accuracy of the parameter estimate increases as the sample size increases, we can say that OLS estimators are both unbiased and consistent.

LO 20.9

Explained sum of squares (ESS) measures the variation in the dependent variable that is explained by the independent variable.

Total sum of squares (TSS) measures the total variation in the dependent variable. TSS is equal to the sum of the squared differences between the actual Y-values and the mean of Y.

Sum of squared residuals (SSR) measures the unexplained variation in the dependent variable.

The standard error of the regression (SER) measures the degree of variability of the actual Y-values relative to the estimated Y-values from a regression equation.

The coefficient of determination, represented by R^2, is a measure of the "goodness of fit" of the regression.

LO 20.10

Assuming certain conditions exist, an analyst can use the results of an ordinary least squares regression in place of an unknown population regression function to describe the relationship between the dependent and independent variable.

Concept Checkers

1. If the value of the independent variable is zero, then the expected value of the dependent variable would be equal to the:
   A. slope coefficient.
   B. intercept coefficient.
   C. error term.
   D. residual.

2. The error term represents the portion of the:
   A. dependent variable that is not explained by the independent variable(s) but could possibly be explained by adding additional independent variables.
   B. dependent variable that is explained by the independent variable(s).
   C. independent variables that are explained by the dependent variable.
   D. dependent variable that is explained by the error in the independent variable(s).

3. What is the most appropriate interpretation of a slope coefficient estimate equal to 10.0?
   A. The predicted value of the dependent variable when the independent variable is zero is 10.0.
   B. The predicted value of the independent variable when the dependent variable is zero is 0.1.
   C. For every one unit change in the independent variable the model predicts that the dependent variable will change by 10 units.
   D. For every one unit change in the independent variable the model predicts that the dependent variable will change by 0.1 units.

4. A linear regression function assumes that the equation must be linear in:
   A. both the variables and the coefficients.
   B. the coefficients but not necessarily the variables.
   C. the variables but not necessarily the coefficients.
   D. neither the variables nor the coefficients.

5. Ordinary least squares refers to the process that:
   A. maximizes the number of independent variables.
   B. minimizes the number of independent variables.
   C. produces sample regression coefficients.
   D. minimizes the sum of the squared error terms.

Concept Checker Answers

1. **B** The equation is $E(Y \mid X) = b_0 + b_1 \times X$. If $X = 0$, then $Y = b_0$ (i.e., the intercept coefficient).

2. **A** The error term represents effects from independent variables not included in the model. It could be explained by additional independent variables.

3. **C** The slope coefficient is best interpreted as the predicted change in the dependent variable for a 1-unit change in the independent variable. If the slope coefficient estimate is 10.0 and the independent variable changes by one unit, the dependent variable will change by 10 units. The intercept term is best interpreted as the value of the dependent variable when the independent variable is equal to zero.

4. **B** Linear regression refers to a regression that is linear in the coefficients/parameters; it may or may not be linear in the variables.

5. **D** OLS is a process that minimizes the sum of squared residuals to produce estimates of the population parameters known as sample regression coefficients.

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. This topic is also covered in:

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

Topic 21

Exam Focus

As shown in the previous topic, the classical linear regression model requires several assumptions. One of those assumptions is homoskedasticity, which means a constant variance of the errors over the sample. If the assumptions are true, the estimated coefficients have the desirable properties of being unbiased and having a minimum variance when compared to other estimators. It is usually assumed that the errors are normally distributed, which allows for standard methods of hypothesis testing of the estimated coefficients. For the exam, be able to construct confidence intervals and perform hypothesis tests on regression coefficients, and understand how to detect heteroskedasticity.

Regression Coefficient Confidence Intervals

LO 21.1: Calculate, and interpret confidence intervals for regression coefficients.

Hypothesis testing for a regression coefficient may use the confidence interval for the coefficient being tested. For instance, a frequently asked question is whether an estimated slope coefficient is statistically different from zero. In other words, the null hypothesis is H_0: $B_1 = 0$ and the alternative hypothesis is H_A: $B_1 \neq 0$. If the confidence interval at the desired level of significance does not include zero, the null is rejected, and the coefficient is said to be statistically different from zero.

The confidence interval for the regression coefficient, B_1, is calculated as:

$$b_1 \pm \left(t_c \times s_{b_1}\right), \text{ or } \left[b_1 - \left(t_c \times s_{b_1}\right) < B_1 < b_1 + \left(t_c \times s_{b_1}\right)\right]$$

In this expression, t_c is the critical two-tailed *t*-value for the selected confidence level with the appropriate number of degrees of freedom, which is equal to the number of sample observations minus 2 (i.e., n – 2).

The standard error of the regression coefficient is denoted as s_{b_1}. It is a function of the SER: as SER rises, s_{b_1} also increases, and the confidence interval widens. This makes sense because SER measures the variability of the data about the regression line, and the more variable the data, the less confidence there is in the regression model to estimate a coefficient.

Professor's Note: It is highly unlikely you will have to calculate s_{b_1} on the exam. It is included in the output of all statistical software packages and should be given to you if you need it.

Example: Calculating the confidence interval for a regression coefficient

The estimated slope coefficient, B_1, from a regression run on WPO stock is 0.64 with a standard error equal to 0.26. Assuming that the sample had 36 observations, **calculate** the 95% confidence interval for B_1.

Answer:

The confidence interval for b_1 is:

$$b_1 \pm (t_c \times s_{b_1}), \text{ or } \left[b_1 - (t_c \times s_{b_1}) < B_1 < b_1 + (t_c \times s_{b_1})\right]$$

The critical two-tail *t*-values are ± 2.03 (from the *t*-table with n – 2 = 34 degrees of freedom). We can compute the 95% confidence interval as:

$$0.64 \pm (2.03)(0.26) = 0.64 \pm 0.53 = 0.11 \text{ to } 1.17$$

Because this confidence interval does not include zero, we can conclude that the slope coefficient is significantly different from zero.

Regression Coefficient Hypothesis Testing

LO 21.3: Interpret hypothesis tests about regression coefficients.

A *t*-test may also be used to test the hypothesis that the true slope coefficient, B_1, is equal to some hypothesized value. Letting b_1 be the point estimate for B_1, the appropriate test statistic with n – 2 degrees of freedom is:

$$t = \frac{b_1 - B_1}{s_{b_1}}$$

The decision rule for tests of significance for regression coefficients is:

$$\text{Reject } H_0 \text{ if } t > +t_{critical} \text{ or } t < -t_{critical}$$

Rejection of the null means that the slope coefficient is *different* from the hypothesized value of B_1.

To test whether an independent variable explains the variation in the dependent variable (i.e., it is statistically significant), the hypothesis that is tested is whether the true slope is zero ($B_1 = 0$). The appropriate test structure for the null and alternative hypotheses is:

$H_0: B_1 = 0$ versus $H_A: B_1 \neq 0$

Example: Hypothesis test for significance of regression coefficients

Again, suppose that the estimated slope coefficient for the WPO regression is 0.64 with a standard error equal to 0.26. Assuming that the sample has 36 observations, **determine** if the estimated slope coefficient is significantly different than zero at a 5% level of significance.

Answer:

The calculated test statistic is $t = \frac{b_1 - B_1}{s_{b_1}} = \frac{0.64 - 0}{0.26} = 2.46$.

The critical two-tailed *t*-values are ± 2.03 (from the *t*-table with df = 36 – 2 = 34). Because $t > t_{critical}$ (i.e., 2.46 > 2.03), we reject the null hypothesis and conclude that the slope is different from zero. Note that the *t*-test and the confidence interval lead to the same conclusion to reject the null hypothesis and conclude that the slope coefficient is statistically significant.

LO 21.2: Interpret the p-value.

Comparing a test statistic to critical values is the preferred method for testing statistical significance. Another method involves the computation and interpretation of a ***p*-value**. Recall from Topic 19, the *p*-value is the smallest level of significance for which the null hypothesis can be rejected.

For two-tailed tests, the *p*-value is the probability that lies above the positive value of the computed test statistic *plus* the probability that lies below the negative value of the computed test statistic. For example, by consulting the *z*-table, the probability that lies above a test statistic of 2.46 is: (1 – 0.9931) = 0.0069 = 0.69%. With a two-tailed test, this *p*-value is: 2 × 0.69% = 1.38%. Therefore, the null hypothesis can be rejected at any level of significance greater than 1.38%. However, with a level of significance of, say, 1%, we would fail to reject the null.

A very small *p*-value provides support for rejecting the null hypothesis. This would indicate a large test statistic that is likely greater than critical values for a common level of significance (e.g., 5%). Many statistical software packages for regression analysis report *p*-values for regression coefficients. This output gives researchers a general idea of statistical significance without selecting a significance level.

Predicted Values

Predicted values are values of the dependent variable based on the estimated regression coefficients and a prediction about the value of the independent variable. They are the values that are *predicted* by the regression equation, given an estimate of the independent variable.

For a simple regression, the predicted (or forecast) value of Y is:

$$\hat{Y} = b_0 + b_1 X_p$$

where:
$\hat{Y}$ = predicted value of the dependent variable
X_p = forecasted value of the independent variable

Example: Predicting the dependent variable

Given the regression equation:

$$\widehat{WPO} = -2.3\% + (0.64)\ (\widehat{S\&P\ 500})$$

Calculate the predicted value of WPO excess returns if forecasted S&P 500 excess returns are 10%.

Answer:

The predicted value for WPO excess returns is determined as follows:

$$\widehat{WPO} = -2.3\% + (0.64)(10\%) = 4.1\%$$

Confidence Intervals for Predicted Values

Confidence intervals for the predicted value of a dependent variable are calculated in a manner similar to the confidence interval for the regression coefficients. The equation for the confidence interval for a predicted value of Y is:

$$\hat{Y} \pm (t_c \times s_f) \Rightarrow \left[\hat{Y} - (t_c \times s_f) < Y < \hat{Y} + (t_c \times s_f)\right]$$

where:
t_c = two-tailed critical t-value at the desired level of significance with df = n – 2
s_f = standard error of the forecast

The challenge with computing a confidence interval for a predicted value is calculating s_f. It's highly unlikely that you will have to calculate the standard error of the forecast (it will probably be provided if you need to compute a confidence interval for the dependent variable). However, if you do need to calculate s_f, it can be done with the following formula for the variance of the forecast:

$$s_f^2 = SER^2\left[1+\frac{1}{n}+\frac{(X-\bar{X})^2}{(n-1)s_x^2}\right]$$

where:

SER^2 = variance of the residuals = the square of the standard error of the regression
s_x^2 = variance of the independent variable
X = value of the independent variable for which the forecast was made

Example: Confidence interval for a predicted value

Calculate a 95% prediction interval on the predicted value of WPO from the previous example. Assume the standard error of the forecast is 3.67, and the forecasted value of S&P 500 excess returns is 10%.

Answer:

The predicted value for WPO is:

$$\widehat{WPO} = -2.3\% + (0.64)(10\%) = 4.1\%$$

The 5% two-tailed critical *t*-value with 34 degrees of freedom is 2.03. The prediction interval at the 95% confidence level is:

$$\widehat{WPO} \pm (t_c \times s_f) \Rightarrow [4.1\% \pm (2.03 \times 3.67\%)] = 4.1\% \pm 7.5\%$$

or

–3.4% to 11.6%

This range can be interpreted as, given a forecasted value for S&P 500 excess returns of 10%, we can be 95% confident that the WPO excess returns will be between –3.4% and 11.6%.

Dummy Variables

Observations for most independent variables (e.g., firm size, level of GDP, and interest rates) can take on a wide range of values. However, there are occasions when the independent variable is binary in nature—it is either "on" or "off." Independent variables that fall into this category are called **dummy variables** and are often used to quantify the impact of qualitative events.

Professor's Note: We will address dummy variables in more detail when we demonstrate how to model seasonality in Topic 25.

What is Heteroskedasticity?

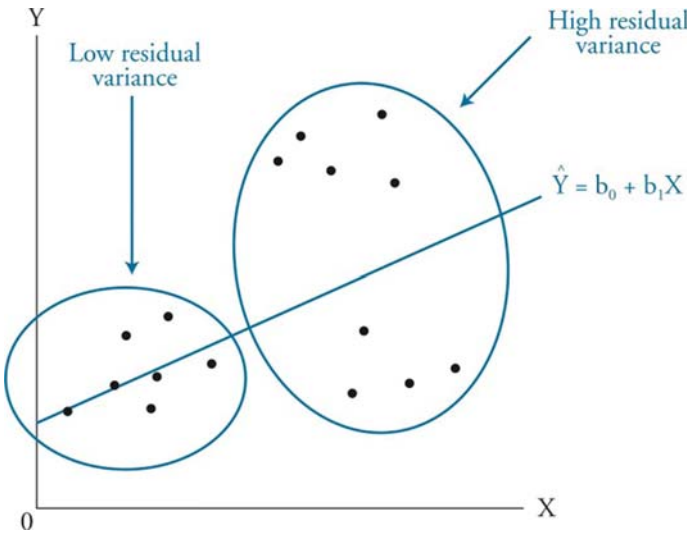
LO 21.4: Evaluate the implications of homoskedasticity and heteroskedasticity.

If the variance of the residuals is constant across all observations in the sample, the regression is said to be **homoskedastic.** When the opposite is true, the regression exhibits **heteroskedasticity**, which occurs when the variance of the residuals is not the same across all observations in the sample. This happens when there are subsamples that are more spread out than the rest of the sample.

Unconditional heteroskedasticity occurs when the heteroskedasticity is not related to the level of the independent variables, which means that it doesn't systematically increase or decrease with changes in the value of the independent variable(s). While this is a violation of the equal variance assumption, *it usually causes no major problems with the regression.*

Conditional heteroskedasticity is heteroskedasticity that is related to the level of (i.e., conditional on) the independent variable. For example, conditional heteroskedasticity exists if the variance of the residual term increases as the value of the independent variable increases, as shown in Figure 1. Notice in this figure that the residual variance associated with the larger values of the independent variable, *X*, is larger than the residual variance associated with the smaller values of *X*. Conditional heteroskedasticity *does create significant problems for statistical inference.*

Figure 1: Conditional Heteroskedasticity



Effect of Heteroskedasticity on Regression Analysis

There are several effects of heteroskedasticity you need to be aware of:

- The standard errors are usually unreliable estimates.
- The coefficient estimates (the b_j) aren't affected.
- If the standard errors are too small, but the coefficient estimates themselves are not affected, the *t*-statistics will be too large and the null hypothesis of no statistical significance is rejected too often. The opposite will be true if the standard errors are too large.

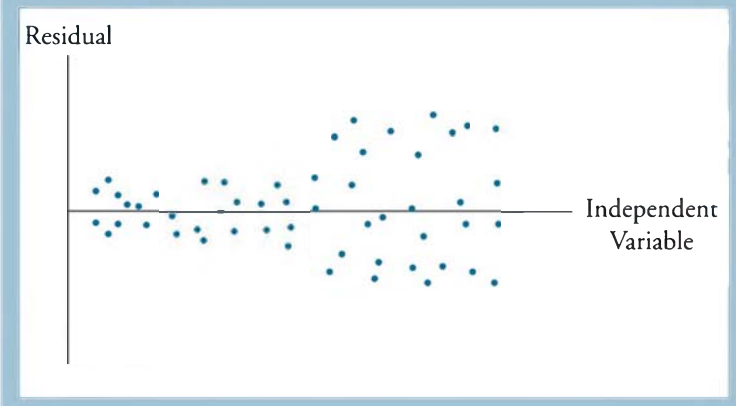
Detecting Heteroskedasticity

As was shown in Figure 1, a scatter plot of the residuals versus one of the independent variables can reveal patterns among observations.

Example: Detecting heteroskedasticity with a residual plot

You have been studying the monthly returns of a mutual fund over the past five years, hoping to draw conclusions about the fund's average performance. You calculate the mean return, the standard deviation, and the portfolio's beta by regressing the fund's returns on S&P 500 index returns (the independent variable). The standard deviation of returns and the fund's beta don't seem to fit the firm's stated risk profile. For your analysis, you have prepared a scatter plot of the error terms (actual return – predicted return) for the regression using five years of returns, as shown in the following figure. **Determine** whether the residual plot indicates that there may be a problem with the data.

Residual Plot



Answer:

The residual plot in the previous figure indicates the presence of conditional heteroskedasticity. Notice how the variation in the regression residuals increases as the independent variable increases. This indicates that the variance of the fund's returns about the mean is related to the level of the independent variable.

Correcting Heteroskedasticity

Heteroskedasticity is not easy to correct, and the details of the available techniques are beyond the scope of the FRM curriculum. The most common remedy, however, is to calculate **robust standard errors**. These robust standard errors are used to recalculate the *t*-statistics using the original regression coefficients. On the exam, use robust standard errors to calculate *t*-statistics if there is evidence of heteroskedasticity. By default, many statistical software packages apply *homoskedastic* standard errors unless the user specifies otherwise.

The Gauss-Markov Theorem

LO 21.5: Determine the conditions under which the OLS is the best linear conditionally unbiased estimator.

LO 21.6: Explain the Gauss-Markov Theorem and its limitations, and alternatives to the OLS.

The **Gauss-Markov theorem** says that if the linear regression model assumptions are true and the regression errors display homoskedasticity, then the OLS estimators have the following properties.

1. The OLS estimated coefficients have the minimum variance compared to other methods of estimating the coefficients (i.e., they are the most precise).

2. The OLS estimated coefficients are based on linear functions.

3. The OLS estimated coefficients are unbiased, which means that in repeated sampling the averages of the coefficients from the sample will be distributed around the true population parameters [i.e., $E(b_0) = B_0$ and $E(b_1) = B_1$].

4. The OLS estimate of the variance of the errors is unbiased [i.e., $E(\hat{\sigma}^2) = \sigma^2$].

The acronym for these properties is "BLUE," which indicates that OLS estimators are the best linear unbiased estimators.

One limitation of the Gauss-Markov theorem is that its conditions may not hold in practice, particularly when the error terms are heteroskedastic, which is sometimes observed in economic data. Another limitation is that alternative estimators, which are not linear or unbiased, may be more efficient than OLS estimators. Examples of these alternative estimators include: the weighted least squares estimator (which can produce an estimator with a smaller variance—to combat heteroskedastic errors) and the least absolute deviations estimator (which is less sensitive to extreme outliers given that rare outliers exist in the data).

Small Sample Sizes

LO 21.7: Apply and interpret the t-statistic when the sample size is small.

The central limit theorem is important when analyzing OLS results because it allows for the use of the *t*-distribution when conducting hypothesis testing on regression coefficients. This is possible because the central limit theorem says that the means of individual samples will be normally distributed when the sample size is large. However, if the sample size is small, the distribution of a *t*-statistic becomes more complicated to interpret.

In order to analyze a regression coefficient *t*-statistic when the sample size is small, we must assume the assumptions underlying linear regression hold. In particular, in order to apply and interpret the *t*-statistic, error terms must be homoskedastic (i.e., constant variance of error terms) and the error terms must be normally distributed. If this is the case, the *t*-statistic can be computed using the default standard error (i.e., the homoskedasticity-only standard error), and it follows a *t*-distribution with n – 2 degrees of freedom.

In practice, it is rare to assume that error terms have a constant variance and are normally distributed. However, it is generally the case that sample sizes are large enough to apply the central limit theorem meaning that we can calculate *t*-statistics using homoskedasticity-only standard errors. In other words, with a large sample size, differences between the *t*-distribution and the standard normal distribution can be ignored.