# Part I: Hypothesis Testing (假設檢定)

## BASIC CONCEPTS FOR HYPOTHESIS TESTING

- Definitions

Population (母數): all the items in a group of data.

Sample (樣本): a selections of items from a population that is representative of the population.

Statistics (統計值): descriptive characteristics of sample. Below are some important statistics.

Mean/Expected Value (平均值/期望值):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Variance (變異數):

Population Variance: $\sigma_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$

Sample Variance: $S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

Standard Deviation (標準差):

Population Standard Deviation: $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$

Sample Standard Deviation: $S_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$

Normal Distribution : Normal distribution is a widely used distribution which is a bell-shaped, symmetric distribution that can easily be described by two parameters—mean and standard deviation (μ,σ). A normal distribution has 68.2% chance that a value falls within the area of μ ± σ, 95.4% falls within the area of μ ± 2σ, and 99.7% that falls within the area of μ ± 3σ.

Normal Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

where

μ is the mean
σ is the standard deviation

- Symbols:

| | Population (Parameter) | Sample (Statistics) |
|---|---|---|
| Number of items in Group | $N$ | $n$ |
| Arithmetic Average | $\mu_X$ | $\overline{X}$ |
| Variance | $\sigma_X^2$ | $S_X^2$ |
| Standard Deviation | $\sigma_{\overline{X}}$ | $S_X$ |
| Standard Error of Estimate | $\sigma_{\overline{X}}$ | $S_{\overline{X}}$ |

- Confidence Intervals

$$(1-\alpha)\ Confidenc\ \ Interval = Point\ \ Estimate \pm (Reliability\ \ Factor \times Standard\ \ Error)$$

**Example**: an analyst selected 5 companies traded on Taiwan Stock Exchange. The average stock price return for the year is 7.6%, with a standard deviation of 4.8%. Estimate, with 95% of confidence, a range that will include the true average stock return of TWSE companies. The t-value for 95% confidence is 2.776.

**Answer:** The true population standard deviation is unknown thus

$$S_{\overline{X}} = \frac{S_X}{\sqrt{n}} = \frac{4.8\%}{\sqrt{5}} = 2.15\%$$

The confidence interval is calculated as follows:

$$95\%\ \ Confidence\ \ Interval\ \ = 7.6\% \pm 2.776(2.15\%) = 1.63\%\ \ to\ \ 13.6\%$$

**PERFORMING A SIMPLE HYPOTHESIS TEST**

A hypothesis test is a procedure based on statistics to infer whether the Hypothesis (statement) is likely to be true or false, given a specified *degree of confidence*.

A hypothesis test itself has seven steps:

- State the hypothesis
- Identify the test statistics
- Specify the level of significance
- State the decision rule
- Collect the data and make the necessary calculations
- Make the statistical decision
- Make the economic decision

1. State the Hypothesis
   Hypotheses are always stated in pares, called null and alternative hypotheses.

   - Null hypothesis ($H_0$/虛無假設): the positive affirmation of the hypothesis that is being tested. The hypothesis is not rejected unless the sample data makes it highly unlikely to be true.
   - Alternative hypothesis ($H_1$/對立假設): is the proposition that must be true if the null hypothesis is not true.

   In all cases, give any probability assumed for the test, one hypothesis will be rejected and the other accepted.

   **Example**: An analyst asserts that the true average stock return for all stock traded in Taiwan Stock Exchange is 10%. Given the previous confidence interval calculations, is this possible?

   **Answer:** The first step is to state the hypothesis in the following way

   Null Hypothesis ($H_0$): $\mu_x = 10\%$
   The true average annual index return is 7%

   Alternative Hypothesis ($H_1$): $\mu_x \neq 10\%$
   The true average annual index return is not 7%

   A hypothesis can be formulated as either one-tailed or two-tailed test

   - Two-tailed test: the hypothesis is stated that the parameter is equal to some value and the alternative hypothesis is formulated that it is not equal to that value

   $$H_0: \quad \mu_x = \mu_0$$

   $$H_1: \quad \mu_x \neq \mu_0$$

   - One-tailed test: the null hypothesis is formulated so that the hypothesized true parameter is greater than or equal to some value.

   $$H_0: \quad \mu_x \geq \mu_0 \qquad \text{or} \qquad H_0: \quad \mu_x \leq \mu_0$$

$H_1$: $\mu_x < \mu_0$ $H_1$: $\mu_x > \mu_0$

2. Identify the Test Statistic
The test statistic is a value calculated from the sample data. This calculated value is used to determine whether to accept or reject the null hypothesis. The general formula is as follows:

$$\text{Test Statistics}_{calc} = \frac{\text{Sample Statistic} - \text{Hypothesized Value } (H_0)}{\text{Standard Error of the Sample Statistics}}$$

The actual test statistic used depends on the type of hypothesis test being completed. Below is the commonly used test statistic for *normally distributed population.*

| Situation | Test Statistic |
|---|---|
| Test a population average (mean) versus some numeric value | t or Z* |
| Test the equality of two population means based on independent samples | t or Z* |
| Test the mean difference between two populations (a paired comparison test) | t or Z* |
| Test the variance (or standard deviation) of a population versus some numeric value) | $X^2$ |
| Test the equality of the variances (or standard deviations) of the two populations | F |

*If the **sample size is very large** (which means the degree of freedom is high) or if **the true standard deviation of the population is known**, a normal distribution Z-test can be used.

**Example**: Calculate the t-statistics for the assertion that the stock return is 10%.

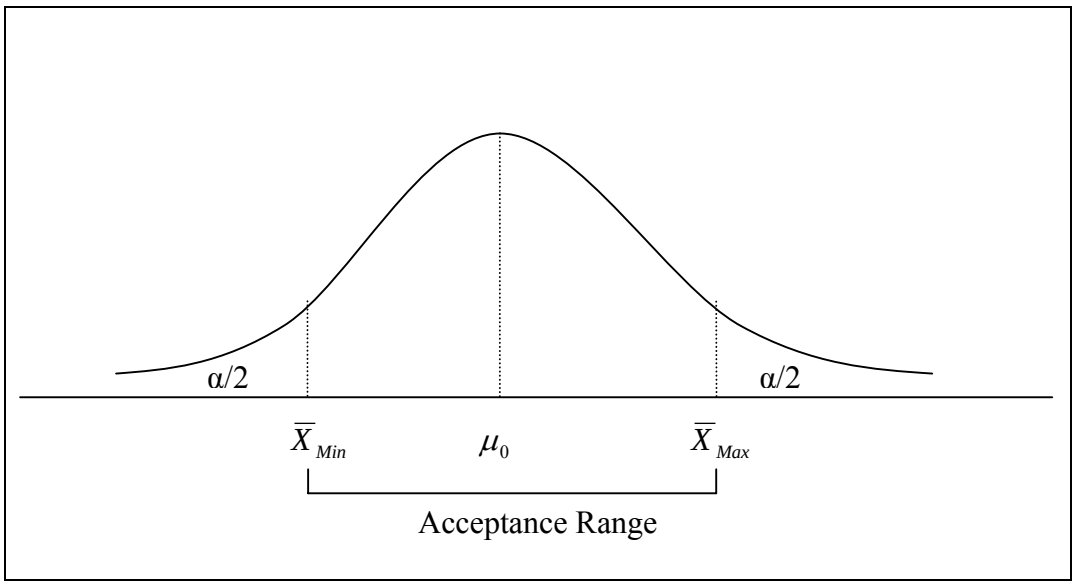
**Answer:** The calculation of the t-statistic for the assertion that the stock return is 10% is as follows:

$$t_{calc} = \frac{\overline{X} - \mu_0}{S_{\overline{X}}} = \frac{\overline{X} - \mu_0}{S_X / \sqrt{n}}$$
$$= \frac{7.6\% - 10\%}{4.8\%\sqrt{5}}$$
$$= -1.12$$

3. Specify the Level of Significance for the Hypothesis Test

Hypothesis test can never prove or disprove a hypothesis. It only shows how likely the hypothesis is true, under the stated level of significance. It is a probability process and thus

- Level of Significance (顯著水準/信賴程度): the probability of rejecting a null hypothesis.
- Type I error is the risk of rejecting the null hypothesis, when it is actually true. The probability of making type I error is equal to the level of significance.
- Type II error is the risk of accepting the null hypothesis while it is actually false. The probability of making the Type II error is unknown.




- There are trade-off between Type I and Type II errors. When the level of significance is set to be high, the risk of making Type I error is low. However, the risk of making Type II error increases due to the increased confidence on not making Type I error.

4. Define and Interpret the Decision Rule
   There are three ways to reach statistic decisions: using acceptance range for test statistics, p-value approach and using an acceptance range around the hypothesized value. All methods will come to the same conclusion.

   **a) Acceptance Range for the Test Statistics**
   - With this approach the acceptance range is stated in terms of maximum and minimum for the test statistic.
   - If the calculated value of the test statistics falls within the acceptance range of test statistic, the null hypothesis cannot be rejected (should be accepted).
   - If the calculated value of the test statistic falls outside the acceptance range of test statistic, the null hypothesis is rejected.

   Calculation of Test of Statistics

$$\text{Test Statistics}_{calc} = \frac{\text{Sample Statistic} - \text{Hypothesized Value } (H_0)}{\text{Standard Error of the Sample Statistics}}$$

   Degree of freedom (d.o.f)=n-1 for t-test

**Example:** In the above example, decide whether to accept the null hypotheses with acceptance rage for test statistics method.

**Answer:** the tcritical is derived as follows:

d.o.f=5-1=4

$t_{critical}$ (or $Z_{critical}$ Values when samples are large) =
$t_{0.05/2,5-1} = t_{0.025,4} = 2.776$

Thus, if the $t_{calc}$ falls within –2.776 and +2.776, accept the null hypothesis that the average stock return is 10%. Otherwise, if $t_{calc} >$ 2.776 or $t_{calc}$<-2.776, reject the null hypothesis.

In the sample, $t_{calc}$=-1.12, within the acceptance range. Thus we should not reject the null hypothesis that the average stock return for Taiwan Stock Exchange is 10%.

**b) P-value approach**

- The P-value is the lowest level of significance at which the null hypothesis is rejected.
- If the P-value is greater than or equal to the level of significance of the test, the null hypothesis cannot be rejected.
- If the P-value is less than the level of significance of the test, the null hypothesis is rejected.

**Example**: in the example above, use P-value approach to make test decisions:

**Answer**: the P-value for $t_{calc}$=1.12 is 0.16, which is greater than the value of (α/2)=0.025. Thus the null hypothesis cannot be rejected.

**c) Acceptance Range Around the Hypothesized Value**

- With this approach the decision rule is stated as a range of values around the hypothesized value in the null hypothesis. This range is called the *acceptance range*.
- If the value of the statistic from the sample data falls within the acceptance, the null hypothesis cannot be rejected (accept the null hypothesis).
- If the value of the statistics from the sample data falls outside the acceptance range, the null hypothesis should be rejected.

| Acceptance = Hypothesized Value ($H_0$) ± (Reliability Factor x Standard Error) |
|---|

$$\overline{X}_{Min/Max} = \mu_0 \pm (R.F. \times S_{\overline{X}})$$

The reliability factor for a test related to arithmetic means is found using t-value unless the sample size is large of the standard deviation of the population is known.

**Example:** For the above example, use acceptance range around the hypothesized value approach to make the statistic decision.

**Answer:** the range is calculated as follows:

Standard error is calculated as: $S_{\bar{X}} = \frac{4.8\%}{\sqrt{5}} = 2.15\%$

Acceptance Range is

$\bar{X}_{Min/Max} = 10\% \pm (2.776 \times 2.15\%) = 4.03\% \ to \ 15.97\%$

$\bar{X}$ is 7.6% which falls in the range thus the null hypothesis cannot be rejected.

5. Collect the Data and Make the Calculations

6. Make the Statistical and Economic Decisions

Hypothesis testing is based on probability and sampling. Thus the decision based on the can be wrong.

Possible reasons:

i. Probability: with this process we assumed populations are normally distributed.
ii. Sampling: sampling may not collect the most representative items of the populations, for example, outliers. Different samples may yield different results.
iii. History doesn't tell the future: all the statistical process are based on the pattern happened in the past.
iv. Other considerations within the business situation and economic environment should also be considered when making economic decisions.

## TESTING A POPULATION MEAN

**Example:** XYZ fund management company claims that investment methodology produces a average annual return of 15% of its funds under management. A competitor wants to test this claim. The competitor randomly selected a sample of XYZ's funds and fin the following:

Sample Size: 20

Average fund return: 12.5%

Standard Deviation: 5.1%

Using 95% level of confidence, perform a two-tailed hypothesis to determine whether the average return of all XYZ funds equals to 15%.

**Answer:**

1. Draft the hypothesis

   Null Hypothesis : $\mu_0 = 15\%$

   Alternative Hypothesis $\mu_0 \neq 15\%$

2. Decide statistic critical value

   Confidence level = 95%, thus $\alpha$ = (1-95%) = 5%

   t-statistics is used with the degree of freedom of 19 (20-1=19)

   $t_{critical} = t_{\alpha/2,n-1} = t_{0.05/2,20-1} = t_{0.025,19} = 2.093$

   Thus, accept the null hypothesis if $t_{calc}$ falls between –2.093 and +2.093

3. Calculate t-statistic

$$S_{\bar{X}} = \frac{S_X}{\sqrt{n}} = \frac{5.1\%}{\sqrt{20}} = 1.14$$

$$t_{calc} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \frac{12.5\% - 15\%}{1.14} = -2.19$$

4. Make statistical decision

   The t-statistic –2.19 < the lower bound of the acceptance range –2.093 thus the null hypothesis that the XYZ's average fund return is 15% is rejected.

- One-tailed vs. two tailed hypothesis

**Example:** In the example above, if the claim of XYZ company is that the average return of its funds is at least greater than 10%, how can this be tested under 95% of confidence level?

**Answer:**

1. State the hypothesis

   Null Hypothesis : $\mu_0 \geq 10\%$

Alternative Hypothesis $\mu_0 < 10\%$

2. Decide statistic critical value

   Confidence level = 95%, thus $\alpha$ = (1-95%) = 5%

   t-statistics is used with the degree of freedom of 19 (20-1=19)

   $t_{critical} = t_{\alpha,n-1} = t_{0.05,20-1} = t_{0.05,19} = 1.7291$

   Thus, accept the null hypothesis if $t_{calc}$ is not greater than 1.7291

3. Calculate t-statistic

$$S_{\bar{X}} = \frac{S_X}{\sqrt{n}} = \frac{5.1\%}{\sqrt{20}} = 1.14$$

$$t_{calc} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \frac{12.5\% - 10\%}{1.14} = 2.19$$

4. Make statistical decision

   The t-statistic –2.19 < the lower bound of the acceptance range –2.093 thus the null hypothesis that the XYZ's average fund return is 15% is rejected.

## TESTING DIFFERENCE BETWEEN TWO POPULATION MEANS USING INDEPENDENT SAMPLES

**Situation:** to test if the arithmetic means of two independent populations are equal.

**Method:** take independent samples from the two populations and use a hypothesis test to determine if the difference between the Xs of the two samples is significantly different from zero.

**Example:** A professional investors is reviewing the return of two portfolio managers to see if their portfolio performance difference is statistically different. Below is the two portfolio's historical performance

| | Manager A | Manager B |
|---|---|---|
| 10-year average annual return: | 10% | 15% |
| Standard Deviation of Returns: | 13% | 20% |

Under 5% level of significance, test if the return difference of these two managers is statistically different from zero.

**Answer:**

1. State the hypothesis

$H_0$: $\mu_A - \mu_B = 0$

$H_1$: $\mu_A - \mu_B \neq 0$

2. Decide statistic critical value

Confidence level = 95%, thus α = (1-95%) = 5%

Degree of freedom = $n_A + n_B - 2 = 10+10-2 = 18$

$t_{critical} = t_{\alpha/2,n-1} = t_{0.05/2,20-2} = t_{0.025,18} = 2.1$

Thus, reject the null hypothesis if $t_{calc}$ is smaller than –2.1 or greater than +2.1

3. Calculate t-statistic

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} = \frac{(10-1)(13\%)^2 + (10-1)(20\%)^2}{10+10-2} = 2.85\%$$

$$t_{calc} = \frac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{\left(\frac{S_P^2}{n_A}\right) + \left(\frac{S_p^2}{n_B}\right)}} = \frac{(10\% - 15\%) - 0}{\sqrt{\left(\frac{2.85\%}{10}\right) + \left(\frac{2.85\%}{10}\right)}} = -0.66$$

4. Make statistical decision

The t-statistic –0.66 falls in the range of –2.1 and +2.1 thus we cannot reject the null hypothesis that the return difference of the two managers equals to zero.

## TESTING THE MEAN DIFFERENCE BETWEEN POPULATIONS WHEN THE SAMPLES ARE NOT INDEPENDENT (PAIRED COMPARISON TESTS)

**Situation:** hypothesis test when the samples are not independent. This type of test is known as paired comparison test. This a test of the difference in averages between two populations where each of the observations is mad as a pair, resulting in only one sample of data.

**Example:** An investment advisory firm claims that their stock selection techniques enable it to find stock that, on average, outperform the Taiwan Weighted Stock Index by 5% a year. How can the claim be tested based on the sampled data below?

Number of Samples: 50 stocks selected by the company

Average excess return (compared with the Index): 3.4%

Standard deviation of the excess return: 5.2%

Should the claim be challenged at the 5% level of confidence?

**Answer:**

1. State the hypothesis

$H_0$: $\mu_d = 5\%$

$H_1$: $\mu_d \neq 5\%$

2. Decide statistic critical value

Confidence level = 95%, thus $\alpha$ = (1-95%) = 5%

Degree of freedom = 50-1=49

$$t_{critical} = t_{\alpha/2,n-1} = t_{0.05/2,50-1} = t_{0.025,49} = 2.01$$

Thus, reject the null hypothesis if $t_{calc}$ is smaller than –2.01 or greater than +2.01

3. Calculate t-statistic

$$S_{\bar{d}} = \frac{S_d}{\sqrt{n}} = \frac{5.2\%}{\sqrt{50}} = 0.735\%$$

$$t_{calc} = \frac{\bar{d} - \mu_0}{S_{\bar{d}}} = \frac{3.4\% - 5\%}{0.735\%} = -2.17$$

4. Make statistical decision

The t-statistic –2.17 falls outside the range of –2.01 and +2.01 thus we reject the null hypothesis that the excess return equals to 5%.

**TESTING THE VALUE OF THE VARIANCE OF A POPULATION**

**Situation:** to test if the variance of a population equals, is greater than, or is less than a specific value.

**Method:** the test statistics that is used to decide whether to reject the null hypothesis is the chi-square statistics, which is defined as:

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

d.o.f = n-1

- Chi-square distributions is not symmetrical
- $\chi^2$ value cannot be negative—it ranges from 0 to ∞ (Variances are always positive)
- $\chi^2$ statistics is an adequate test statistic for testing hypotheses concerning population variance for normally distributed populations only.

**Example:** a pension fund sponsor would like to evaluate the risk level of the pension under management. The original statement of contract states that the standard deviation of fund portfolio return should be controlled to 20%. The pension sponsor would like to perform a hypothesis test to test if the variance of the portfolio is the past ten years shows that the risk policy is confined, with the data below:

Sampled years: 10 years

Average annual portfolio variances: 24.5%

Level of confidence: 5%

**Answer:**

1. State the hypothesis

   *Variance is standard deviation squared.

   $H_0$: $\sigma_X^2 \leq 400$

   $H_1$: $\sigma_X^2 > 400$

2. Decide statistic critical value

   Confidence level = 95%, thus α = (1-95%) = 5%

   Degree of freedom = 10-1=9

   $\chi^2_{critical} = \chi^2_{0.05,9} = 16.919$

   Thus, reject the null hypothesis if $\chi^2_{calc}$ is greater than 16.919.

3. Calculate $\chi^2$-statistic

$$\chi^2_{calc} = \frac{(n-1)S_X^2}{\sigma_0^2} = \frac{(10-1)(24.5)^2}{20^2} = 13.51$$

4. Make statistical decision

   The $\chi^2$-statistic 13.51 which is smaller than 16.919, thus the null hypothesis cannot be rejected that the standard deviation of the pension portfolio is less than or equal to 20%, at level of 5% significance.

## TESTING THE EQUALITY OF TH VARIANCES OF TWO POPULATIONS

**Situation:** This is used to test if the variances of two populations are equal, or if the variance of one population is greater than (or is less than) the variance of another population.

| | | | | | |
|---|---|---|---|---|---|
| $H_0:$ | $\sigma_1^2 = \sigma_2^2$ | $H_0:$ | $\sigma_1^2 \geq \sigma_2^2$ | $H_0:$ | $\sigma_1^2 \leq \sigma_2^2$ |
| $H_1:$ | $\sigma_1^2 \neq \sigma_2^2$ | $H_1:$ | $\sigma_1^2 < \sigma_2^2$ | $H_1:$ | $\sigma_1^2 > \sigma_2^2$ |

**Method:** The test statistic that is used to decide whether to reject these types of null or alternative hypotheses is the F-statistic, which is defined as:

$$F = \frac{S_1^2}{S_2^2}$$

The F-distribution is a family of distributions that are defined by two separate degrees of freedom: the number of degrees of freedom in the numerator and the number of degree of freedom in the denominator.

$$\text{d of } f_n = n_1 - 1$$

$$\text{d of } f_d = n_2 - 1$$

- Like Chi-square distributions, F-distribution is not symmetrical.
- F value cannot be negative because variance cannot be negative, and ranges between 0 and infinity. Furthermore, the expected value of the F-distribution is 1.0, if $\sigma_1^2 = \sigma_2^2$.
- F statistics is an adequate test statistic for testing hypotheses concerning population variance for normally distributed populations only.

**Example:** Fund management firm is to compare the 10-year performance records of two managers. The company wonders if the two manager's levels of risk are materially different. Perform a hypothesis test at 5% of level of confidence to determine whether the manager's standard deviations of returns are equal.

| | Manager A | Manager B |
|---|---|---|
| 10-year annual return | 10% | 20% |
| Standard Deviation of Returns | 16% | 33% |
| Variance | 256 | 1089 |

**Answer**:

1. State the hypothesis

   $H_0$: $\sigma_A^2 = \sigma_B^2$

   $H_1$: $\sigma_A^2 \neq \sigma_B^2$

2. Decide statistic critical value

   $\alpha = 5\%$

   Degree of freedom for $f_n$=10-1=9
   Degree of freedom for $f_d$=10-1=9

   $F_{0.05/2,9/9}=F_{0.025,9/9} = 4.03$

   Thus, accept the null hypothesis if $F_{calc}$ ranges between 0 and 4.03.

3. Calculate F-statistic

$$F_{calc} = \frac{S_B^2}{S_A^2} = \frac{(33\%)^2}{(16\%)^2} = \frac{1089}{256} = 4.25$$

4. Make statistical decision

   The F-statistic 4.25 which is greater than 4.03, thus the null hypothesis should be rejected that the two managers risk levels are equal.

# Part II: Correlation and Regression (相關與迴歸)

Correlation and regression is a statistical process discovering statistical relationship between variables.

## DEPENDENT AND INDEPENDENT VARIABLES

In a simple linear relationship, the dependent and independent variables can be expressed in the following mathematical form:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

where

$Y_i$ is the dependent
$X_i$ is the independent
$b_0$ is the regression intercept (Regression parameters)
$b_1$ is the regression coefficient (Regression parameters)
$\varepsilon_i$ is the error term or residual

Correlation and Regression Analysis can be used to

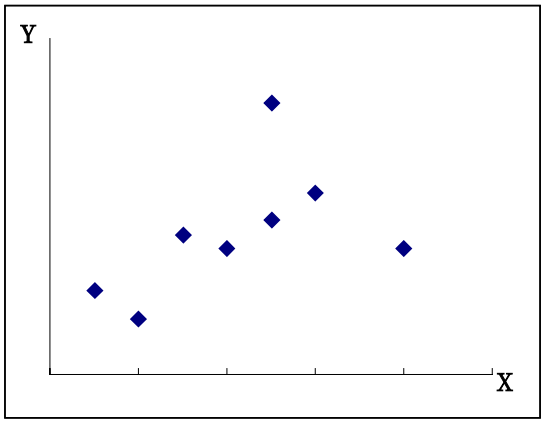
1. Verify that there is a relationship between X and Y by applying **correlation analysis**.
2. Determine the mathematical form of the relationship-linear, log-linear, parabolic or other general mathematical forms. (Only linear relationships are considered for the exam.)
3. Determine the précised form of relationship—this is to determine the actual value of regression parameters—intercept ($b_0$) and regression coefficients ($b_1$).

**Example:** in modern portfolio theory, CAPM captures the relationship of individual stock price change in correspondence to overall market change. Statistically speaking, there is a correlation relationship between the index change and individual stock price change.

## CORRELATION ANALYSIS

The first step of the analysis is to determine if there is a relationship between two variables. The analysis is called correlation analysis.

1. Scatter Plot (XY chart)




2. Calculate Covariance(共變數) and Correlation Coefficient(相關係數)

   The statistical measure of the direction and degree of linear association between two random variables is the correlation coefficient. The formula of correlation coefficient (denoted as $r_{xy}$) is as follows:

$$r_{xy} = \frac{COV_{xy}}{S_x S_y}$$

   where

$$COV_{xy} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{n-1}$$

   **Example:** Calculate the $r_{xy}$ according to the information below:

$$Cov_{XY} = 102.5$$
$$S_Y^2 = 145$$
$$S_X^2 = 80$$
$$n = 5$$

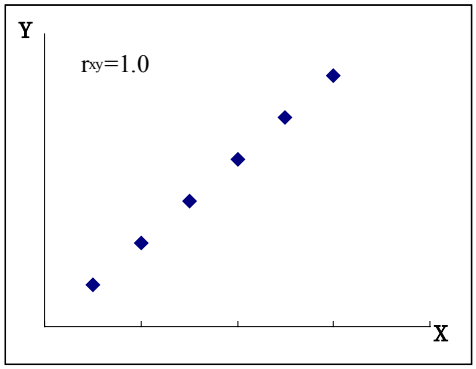
   **Answer:**

$$r_{XY} = \frac{Cov_{XY}}{S_Y S_X} = \frac{102.5}{\sqrt{145}\sqrt{80}} = 0.952$$

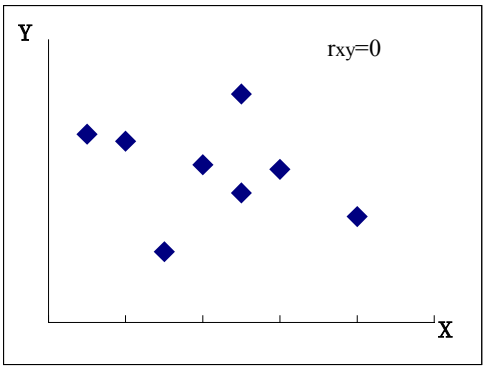
3. Interpret a Correlation Coefficient

   Correlation coefficient measures the direction and degree of linear relationship between two random variables, which is a number ranges between –1 and +1.

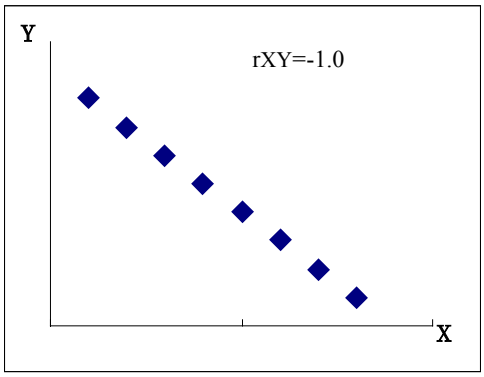
   a. A correlation coefficient of +1 ($r_{xy}$=+1) means that two random variables X and Y are *perfectly positively* related in a linear manner.




b. A correlation coefficient of zero ($r_{xy}$=0) indicates that the two random variables are not related at all.




c. A correlation coefficient of –1 ($r_{xy}$=-1) indicates that the two random variables are perfectly negatively related.




4. Coefficient of Determination

The square of the correlation coefficient is called the coefficient of determination, which measures the percentage of the total variation in the dependent variable (Y) that is explained by the variation in independent variable (X).

$$R_{xy}^2 = (r_{xy})^2$$

**Example:** In the example above, calculate the coefficient of determination

**Answer:**

$$R_{XY}^2 = (r_{XY})^2 = (0.952)^2 = 0.91$$

5. Testing the Significance of a Correlation Coefficient

It is important that the correlation coefficient be significant, meaning that it is not likely to be zero. If it is not significant, there may be no relationship between the dependent and the independent variables.

The null and the alternative hypothesis is formulated as follows:

$$H_0: \ r_{xy} = 0$$

$$H_1: \ r_{xy} \neq 0$$

Test statistic used for test hypothesis about correlation coefficient is as follows:

$$t_{calc} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Degree of freedom=n-2 (for sample linear regressions)

- If $|t_{calc}| \leq t_{critical}$, do not reject (accept) the null hypothesis that the $r_{xy}=0$; relationship between dependent and independent variables is not significant.
- If $|t_{calc}| > t_{critical}$, reject the null hypothesis that the $r_{xy}=0$; relationship between dependent and independent variables is significant.

**Example:** Determine whether or not the correlation coefficient calculated above is significantly different from zero.

**Answer:**

1. State the hypothesis

$H_0$: $r_{XY} = 0$

$H_1$: $r_{XY} \neq 0$

2. Decide statistic critical value

$$t_{criterial} = t_{\alpha/2,n-2} = t_{0.05/2,5-2} = t_{0.025,3} = 3.182$$

3. Calculate test statistics

$$t_{calc} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} = \frac{0.952\sqrt{5-2}}{\sqrt{1-(0.952)^2}} = 5.39$$

4. Make statistical decision

   The calculated statistics is greater than the critical value of the test statistic. Therefore, the null hypothesis that $r_{XY}=0$ is rejected.

6. Limitations of Correlation Analysis

   There are several limitations of correlation analysis:

   a. The correlation coefficient assumes the relationship between the two variables is linear. In case there is a nonlinear relationship, the correlation coefficient might suggest a weak relationship.

   b. The presence of outliers can distort the correlation coefficient. Outliers are the small number of observations that ate atypical relative to the rest of the observations. One way to measure the distorting effect of outliers is to calculate the correlation coefficient for data sets with and without the outlier data.

   c. The existence of spurious correlation, which occurs when sample data produce a correlation coefficient that is reasonably high (and even significant), but there really is no relationship between the random variables. The high correlation coefficient is derived by coincidence. Data snooping and mining may produce such a spurious correlation when large amount of data are performed with such a analysis. Two rules can protect the analyst from falling into the trap:

      - Correlation doe not imply causation but only a statistical result.
      - Correlation without theoretical basis should be suspect.

**SIMPLE LINEAR REGRESSIN ANALYSIS**

Simple linear regression assumes a relationship between a dependent (Y) and independent (X) in the following form:

$$\hat{Y}_i = b_0 + b_1 X_i + \varepsilon_i$$

where

$\hat{Y}_i$ is the calculated value of Y for a specified value of Xi

$b_0$ is the Y-intercept. It is the value of dependent variable Y when the independent variable X equals to zero.

$b_1$ is the regression coefficient—the slop of the regression line. This regression coefficient measures the degree of change of Y at the percentage change of X.

$$b_1 = \frac{\Delta Y}{\Delta X}$$

$\varepsilon_i$ is the error term—the difference between actual value of Y and the theoretical value of Y according to the regress equation for any given value of $X_i$. The error term is also called the residual. The error term can be denoted in the following form:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{b}_0 + \hat{b}_1 X_i)$$

1. Determining the Regression Coefficient and Y-intercept with the Method of **Least Squares Estimation** (最小平方法)

   Simple linear regression analysis attempts to find the best straight line that fits the scatter plot—the method used is the least square method.

   With the method, the regression parameters are calculated as follows:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

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

   When the regression equation is derived from the data observed. We are able to predict the value of Y (dependent) at a give value of X (independent).

   **Example:** with a linear regression equation we derived there is a relation between stock ABC and S&P 500 Index. The linear regression is as follows:

$$\hat{Y}_i = 3.35 + 1.15 X_i$$

   where the intercept of the regression is 3.35 and the regression coefficient (slop) is 1.15.

   Use the simple linear regression derived, predict the return of stock ABC if S&P 500 Index return is 12%.

**Answer:**

$$\hat{Y}_i = 3.35 + 1.15X_i + \varepsilon_i = 3.35 + 1.15(12\%) + 0 = 13.83\%$$

If S&P 500 Index is to change by 12%, the price change of stock ABC is predicted to be 13.83%.

2. The Standard Error of Estimate

As we have seen, the regression equations do not relate the dependent and independent variables perfectly. Thus, the standard error of estimates (SEE) is a measure of how imperfect the regression model is in predicting the dependent variables.

SEE formula is as follows:

$$SEE = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}}$$

Note: Interpretation of the standard error of estimate is somewhat similar to the standard deviation. The standard error of estimate measures deviations of data around a regression line, whereas a standard deviation measures deviations of data around the arithmetic mean.

3. Prediction Intervals on the Dependent Variable

There are two source of error, in addition to the standard error of estimate, in predictions made from a regression equation.

- Regression equations are estimated from sample data, thus the regression parameters b0 and b1 are only estimates of their true values. Thus there is a variance associated with the position of the regression line associated with any value of the independent variable.

- There is also a variance associated with the actual value of the dependent variable (Yi) around the position of the regression line.

Taking these two source into account, as well as the error associated with the standard error of estimate, the variance associated with a predicted (or future)value of the dependent variable (Sf) for a specified value of the independent variable (Xi) is

$$S_f^2 = (SEE)^2 \left[1 + \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}\right]$$

Therefore, the following formula can be used to place a 1-α confidence interval around a predicted value of a dependent variable determined from regression model:

$$\hat{Y}_{i_{\max/\min}} = \hat{b}_0 + \hat{b}_1 X_i \pm t_{\alpha/2,n-2} S_f$$

**Example:** in the example above, use the following data to estimated the return range of stock ABC when return on S&P 500 Index is 12%, at 95% confidence interval.

SSE=4.27

$S_f^2 = 28.7738$

**Answer:**

$S_f = \sqrt{28.7738} = 5.36$

$Y_{i_{\max/\min}} = \hat{b}_0 + \hat{b}_1 X_i \pm t_{\alpha/2,n-2} S_f$

$t_{\alpha/2,n-2} = t_{.025,3} = 3.182$

$Y_{i_{\max/\min}} = 3.35 + 1.15(12\%) \pm 3.182(5.36) = 0.0945\% \ to \ 34.31\%$

4. Confidence Interval on Regression Parameters

The confidence interval of regression parameters are calculated the same way as the confidence interval for mean of the population. 1-α confidence interval can be placed on the regression formula in the following form:

$$b_{0_{\min/\max}} = \hat{b}_0 \pm t_{\alpha/2,n-2} S_{\hat{b}_0}$$
$$b_{1_{\min/\max}} = \hat{b}_1 \pm t_{\alpha/2,n-2} S_{\hat{b}_1}$$

**Example:** In the example, compute the 95% confidence interval of the intercept and the regression coefficient according to the sample standard error of the regression parameters.

Standard error of Y-intercept: $S_{\hat{b}_0} = 1.92$

Standard error of regression coefficient: $S_{\hat{b}_1} = 0.239$

**Answer:**

Regression equation is $Y_i = 3.35 + 1.15 X_i + \hat{\varepsilon}_i$

$t_{\alpha/2,n-2} = t_{0.025,3} = 3.182$

$$b_0 = \hat{b}_0 \pm t_{\alpha/2,n-2} S_{\hat{b}_0} = 3.35 \pm 3.182(1.92) = -2.76 \ to \ 9.46$$
$$b_1 = \hat{b}_1 \pm t_{\alpha/2,n-2} S_{\hat{b}_1} = 1.15 \pm 3.182(0.239) = 0.39 \ to \ 1.91$$

5. Testing the Significance of Regression Parameters

Regression analysis is based on sample data, it is always possible that the parameters of the regression equation may not be significant—they may be equal to zero. It is particularly important that the regression coefficient ($b_1$) to be significant thus, if the regression coefficient is not significant ($b_1$=0), there is no relationship between dependent and independent variables.

The null and alternative hypothesis can be formulated as follows:

| | | |
|---|---|---|
| $H_0$: $b_0 = \beta_0$ | Or | $H_0$: $b_1 = \beta_1$ |
| $H_1$: $b_0 \neq \beta_0$ | | $H_1$: $b_1 \neq \beta_1$ |

- t-Test applied.

$$t_{calc} = \left| \frac{\hat{b}_0 - \beta_0}{S_{\hat{b}_0}} \right| \quad or \quad t_{calc} = \left| \frac{\hat{b}_1 - \beta_1}{S_{\hat{b}_1}} \right|$$

- Get $t_{critical}$ value
  For simple linear regression (two estimated regression parameters), there are n-2 degree of freedom.

$$t_{critical} = t_{\alpha/2,n-2}$$

- Decision rule
  If $|t_{calc}| \leq t_{critical}$, do not reject null hypothesis ($b=\beta$)
  If $|t_{calc}| > t_{critical}$, reject null hypothesis ($b \neq \beta$)

**Example:** In the example, perform a hypothesis test to see if the beta of stock ABC equals to market beta (equals to 1) at 5% level of confidence.

Answer:

1. State the hypothesis

   $H_0$: $b_1 = 1$

   $H_1$: $b_1 \neq 1$

2. Decide statistic critical value

$$t_{critical} = t_{\alpha/2,n-2} = t_{0.025,3} = 3.182$$

3. Calculate t-statistics

$$t_{calc} = \left( \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}} \right) = \left( \frac{1.15 - 1.0}{0.239} \right) = 0.628$$

4. Make statistic decision

   Since the calculated t-statistic is less than 3.182, thus the null hypothesis should be accepted. (The beta is not significantly different from 1.)

## THE COEFFICIENT OF DETERMINATION (判定係數)

The coefficient of determination measures the percentage of the variation in the dependent variable (Y) that s explained by the regression equation. The value of coefficient of determination ranges from 0 to 1.

$$0 \le R_{XY}^2 \le 1$$

The correlation coefficient is the square root of the coefficient of determination

$$r_{XY} = \sqrt{R_{XY}^2}$$

**Example:** Calculate the coefficient of determination for stock ABC and interpret the meaning of the coefficient of determination.

**Answer:**

$$R_{XY}^2 = (r_{XY})^2 = (0.952)^2 = 0.91$$

The coefficient determination of 0.91 means that the variation of S&P 500 Index returns explains 91% of the variation of stock ABC returns.

## THE ANALYSIS OF VARIANCE (ANOVA ANALYSIS)

The ANOVA summarizes the total variation and attributes this variation to its different sources.

| ANOVA | d.o.f | Sum of Squares | Mean sum of Squares |
|---|---|---|---|
| Regression | 1 | $RSS = \sum_{i=1}^{n} \left( \hat{Y}_i - \bar{Y} \right)^2$ | $MSS_R = RSS/1$ |
| Residuals (errors) | n-2 | | $MSS_E = SSE/(n-2) = SEE^2$ |

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Total n-1 $TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ $MSS_T=TSS/(n-1)$

- Regression
  The degree of freedom for simple regression=1 (only one independent variable). The regression sum of squares (RSS) measures the amount of the variation in the dependent variable that is explained by the variation in the independent variable. The mean sum of the squares for the regression ($MSS_R$) equals the regression sums of squares divided by the degrees of freedom.

- Residuals
  The degree of freedom for the residuals is equal to the number of observations (data points) minus 2. The sum of the squares of the error (SSE) measures the amount of the variation in the dependent variables that is not explained by the variation in the independent variable by the regression equation.

- Total
  The total sum of the squares (TSS) is the total amount of variation in the dependent variable, which is equal to RSS + SSE. The mean sum of the squares in total ($MSS_T$) equals the total sums of squares divided by the degrees of freedom (n-1).

- Calculate Coefficient of Determination from ANOVA
  With ANOVA analysis, we can derive the coefficient of determination, which measures the percentage of total variation in the dependent variable that is explained by the regression equation, through the following way:

$$R_{XY}^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

**Example:** in the example, the ANOVA table is as blow:

| ANOVA | d of f | Sum of Squares | Mean Sum of Squares |
|---|---|---|---|
| Regression | 1 | 525.312 | 525.312 |
| Residual (errors) | 3 | 54.688 | 18.229 |
| Total | 4 | 580.000 | 145.000 |

Using the ANOVA table above, determine the coefficient of determination, correlation coefficient of the regression, and standard error of estimate.

**Answer:**

Coefficient of determination ($R^2$):

$$R^2_{XY} = \frac{RSS}{TSS} = \frac{525.312}{580} = 0.91 \text{ or } R^2_{XY} = 1 - \frac{SSE}{TSS} = 1 - \frac{54.688}{580} = 0.91$$

Correlation coefficient (r)

$$r_{XY} = \sqrt{R^2} = \sqrt{0.91} = 0.95$$

Standard error of estimate (SEE):

$$(SEE)^2 = MSS_E = 18.229$$
$$SSE = \sqrt{18.229} = 4.27$$

## LIMITATIONS OF REGRESS ANALYSIS

- The relationship of two variables changes over time. Regression explains historical data but not predict future behavior.
- Assumptions underlying regression analysis may not be true, and thus the inference of the analysis. For example, assumption of normal distribution of parameters is usually used in the hypothesis testing to test the validation of regression parameters.
- One the relationship is discovered and applied, the effectiveness of the analysis turns into useless soon.

## Part III: Multiple Linear Regression (多元迴歸)

Multiple linear regression describes variables in the following form:

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \Lambda + b_n X_{nt} + \varepsilon_t$$

where

$Y_t$ is the dependent variable

$X_{it}$ are the independent variables

$b_0$ is the intercept; i.e., the value of the dependent variable (Y) when all of the values of the independent variables equals zero.

$b_i$ are the regression coefficients of each of the independent variables. Holding all other independent variables constant, the regression coefficient equals to

$$b_i = \frac{\partial Y_t}{\partial b_i}$$

**ASSUMPTIONS OF A MULTIPLE LINEAR REGRESSION MODEL**

- The relationship between the dependent variable and each independent variable in the regression model is linear.
- The independent variables are not random and no exact linear relation exists between two or more of the independent variables. If it happens, then there is "multicollinearity" in the independent variables. (will be discussed latter.)
- The expected value of error term is zero. $\sum_{t=1}^{n} \varepsilon_t = 0$
- The variance (or standard deviation) of the error term is the same for all observations, i.e., the error term is homoskedastic.
- The error term is uncorrelated across observations; i.e., there is no serial correlation (or autocorrelation) in the error terms. $r_{\varepsilon_t \varepsilon_{t-n}} = 0$
- The error term is normally distributed. All These assumptions permit confidence limits to be placed on, and hypothesis tests to be performed on, the regression parameters using standard t-tests (or Z-test).

In a multiple linear regression model, lease squares method is also applied to derive regression coefficients. However, the calculation is much more complex and thus is beyond the scope of exam.

Below is a multiple regression output

**Multiple Regression Output**

No. of Variables……..3
No. of Observations……100

| Regression Coefficient | Value | Standard Error | t-value | P-Value |
|---|---|---|---|---|
| 1 | -0.38 | 0.08 | -4.75 | 0.000 |
| 2 | 1.25 | 1.15 | 1.09 | 0.26 |
| 3 | 0.05 | 0.01 | 5.00 | 0.000 |

Intercept ……...0.157
Standard Error of Intercept……...…0.083
Coefficient of Multiple Correlation (r)……..0.705

Durbin-Watson Statistic……..1.95

**Analysis of Variance**

| Source | DF | SS | MSS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 3 | 35,687.67 | 11,895.89 | 31.62 | 0.0000 |
| Error | 96 | 36,118.51 | 376.23 | | |
| Total | 99 | 71,806.18 | | | |

**Correlation Matrix**

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $X_1$ | 1.00 | | |
| $X_2$ | -.30 | 1.00 | |
| $X_3$ | -.15 | .20 | 1.00 |

1. Writing Multiple Regression Equation

   **Example:** Use the multiple regression output table above, formulate the regression equation.

   **Answer:**

$$Y_t = 0.157 - 0.38X_{1t} + 1.25X_{2t} + 0.05X_{3t} + \varepsilon_t$$

2. Use Multiple Regression Equation to Predict the Value of Dependent Variable

**Example:** predict the Y value under a set of specific independent variables as below:

$X_{1t}$……..10
$X_{2t}$……..15
$X_{3t}$……185

**Answer:**

The regression equation is: $Y_t = 0.157 - 0.38X_{1t} + 1.25X_{2t} + 0.05X_{3t} + \varepsilon_t$

Predicted value is : $\hat{Y}_t = 0.157 - 0.38(10) + 1.25(15) + 0.05(185) + 0 = 24.357$

3. Testing the Overall Validity of the Regression Model

The validity of multiple regression equations can be tested in many ways. A global test is used to determine whether or not all of the regression coefficients as a group are insignificant.

The null and alternative hypotheses are:

$H_0$: $b_1=b_2=b_3=\ldots=b_k=0$

$H_1$: At least one slope coefficient does not equal zero.

F-statistic is used to perform the test

a. Calculate F-Score (with data in ANOVA table)

$$F_{calc} = \frac{RSS/k}{SSE/(n-k-1)} = \frac{MSS_R}{MSS_E}$$

where n is total number of observations, k is the total number of regression coefficients.

b. Find the critical value of the F-statistic

Degree of freedom for numerator ($MSS_R$)=k
Degree of freedom for denominator ($MSS_E$)=n-k-1

$$F_{critical}=F_{\alpha,\, k/(n-k-1)}$$

c. The Decision Rules

If $F_{calc} \leq F_{critical}$, do not reject null hypothesis that all regression coefficients are not significantly different from zero.

If $F_{calc} > F_{critical}$, reject null hypothesis; one or one more regression coefficients are considered to be significant.

**Example:** use the information in the regression output table, calculate if all regression coefficients are not significantly different from zero at 5% level of confidence.

**Answer:**

1. State the hypothesis

   $H_0$: $b_1=b_2=b_3=0$

   $H_1$: At least one slope coefficient does not equal zero.

2. Decide F-statistic critical value

$$d\ of\ \ f_{MSS_R} = k = 3$$
$$d\ of\ \ f_{MSS_E} = n-k-1 = 100-3-1 = 96$$
$$F_{critical} = F_{\alpha,k/(n-k-1)} = F_{0.05,3/96} = 2.7$$

3. Calculate F-statistics

$$F_{calc} = \frac{RSS/K}{SSE/(n-k-1)} = \frac{MSS_R}{MSS_E} = \frac{11{,}895.89}{376.23} = 31.62$$

4. Make statistic decision

   The calculated F-statistic is greater its critical value, thus the null hypothesis is rejected, meaning that at least one regression coefficient is significant.

4. Testing Significance of the Intercept and the Regression Coefficients

We will also test the validity of the regression intercept and each individual regression coefficient.

The null and alternative hypotheses are as follows:

$H_0$: $b_i = 0$

$H_1$: $b_i \neq 0$

**t-statistic** is used to perform the testing under assumptions of multiple linear regression

a. Calculate t-statistic

$$t_{calc} = \frac{\hat{b}_i - b_i}{S_{\hat{b}_i}}$$

b. Determine the critical value of the test statistic

Degree of freedom=n-k-1

$t_{critical} = t_{\alpha/2,n-k-1}$

Two-tailed hypothesis test is applied that we are trying to test whether the regression coefficient is equal to zero.

c. The decision rules

If $|t_{calc}| \leq t_{critical}$, do not reject null hypothesis that $b_i=0$ (insignificant).

If $|t_{calc}| > t_{critical}$, reject null hypothesis that $b_i=0$ (significant).

**Example:** test each of the regression parameters at 5% of confidence if they are significantly different from zero

**Answer:**

1. Decide t-statistic critical value for tests

$$d\ of\ \ f = n-k-1 = 96$$
$$t_{critical} = t_{\alpha/2,(n-k-1)} = t_{0.025,96} \approx 2.0$$

2. Calculate t-statistics for each of regression parameters

$b_o$ : $t_{calc} = \frac{\hat{b}_0 - b_o}{S_{\hat{b}_0}} = \frac{0.157-0}{0.083} = 1.89$ $\hat{b}_0$ is not significant.

$b_1$ : $t_{calc} = \frac{\hat{b}_1 - b_1}{S_{\hat{b}_1}} = \frac{-0.38-0}{0.08} = -4.75$ $\hat{b}_1$ is significant

$b_2$ : $t_{calc} = \frac{\hat{b}_2 - b_2}{S_{\hat{b}_2}} = \frac{1.25-0}{1.15} = 1.09$ $\hat{b}_2$ is not significant

$b_3$ : $t_{calc} = \frac{\hat{b}_3 - b_3}{S_{\hat{b}_3}} = \frac{0.05-0}{0.01} = 5.00$ $\hat{b}_3$ is significant

5. Test whether a regression coefficient is significantly different from a specified value

One-tailed and two-tailed hypothesis can be used to test the significance of a regression parameter is less than, equal to, or greater than a specified value.

6. Determine Confidence Intervals for Regression Parameters

It is possible to place 1-α confidence intervals on the regression parameters.

$$\text{Confidence Interval for } (1-\alpha)_{b_i} = \hat{b}_i \pm t_{\alpha/2,(n-k-1)} S_{\hat{b}_i}$$

7. Determine the Standard Error of Estimate of a Multiple Regression Model

The standard error of estimate of a multiple regression model is the standard error (or standard deviation) of the residuals of the regression model.

$$\varepsilon_t = Y_t - \hat{Y}_t = Y_t - (\hat{b}_0 + \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \Lambda + \hat{b}_k X_{kt})$$

The formula for the variance of the residuals is:

$$(SEE)^2 = S_\varepsilon^2 = \frac{\sum_{t=1}^{n} (\hat{\varepsilon}_t - \bar{\varepsilon})^2}{n-k-1}$$

where the degree of freedom for $MSS_E$=n-k-1

The standard error (standard deviation) for residuals is as follows:

$$SEE = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSS_E}$$

**Example:** Calculate the standard error of estimate with the data in the ANOVA table.

**Answer:**

$$SEE = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSS_E} = \sqrt{\frac{36{,}118.51}{96}} = \sqrt{376.23} = 19.4$$

8. Determine the Coefficient of Determination and the Correlation Coefficient of a Multiple Regression

The coefficient of determination of a multiple regression model ($R^2$) is the percentage of the total variation in the dependent variable (Y) that is explained by the regression equation.

The total variation in the dependent variable is the total sum of the squared difference between the actual values of the dependent variable and its mean

value, which can be broken down into two components: the sum of the squared differences that is explained by the regression and the sum of the squared differences that is not explained by the regression.

| Total Variation | Explained Variation | Unexplained Variation |
|---|---|---|
| $\sum_{t=1}^{n}\left(Y_t-\bar{Y}\right)^2$ | $\sum_{t=1}^{n}\left(\hat{Y}_t-\bar{Y}\right)^2$ | $\sum_{t=1}^{n}\left(Y_t-\hat{Y}_t\right)^2$ |

Total Variation=Explained Variation + Unexplained Variation
(TSS) (RSS) (SSE)

The coefficient of determination is derived from the following formula:

$$R^2=\frac{Total\ \ Variation-Unexplained\ \ Variation}{Total\ \ Variation}=\frac{TSS-SSE}{TSS}=1-\frac{SSE}{TSS}$$

$$R^2=\frac{Explained\ \ Variation}{Total\ \ Variation}=\frac{RSS}{TSS}$$

**Example:** Calculate the coefficient of determination and correlation coefficient with data in the ANOVA table.

**Answer:**

$$R^2=1-\frac{SSE}{TSS}=\frac{RSS}{TSS}=1-\frac{36{,}118.51}{71{,}806.18}=\frac{35{,}687.67}{71{,}806.18}=0.497$$

$$R=\sqrt{R^2}=\sqrt{0.497}=0.705$$

Coefficient of determination is 0.497 and correlation coefficient of the regression is 0.705

a. $R^2$ and Adjusted $R^2$

Problems associated with $R^2$: every time a new independent variable is added to a multiple regression model, the amount of unexplained variation will decrease if the new independent variable explains any of the unexplained variation in the regression. Thus, the $R^2$ of a multiple regression model will increase every time a new independent variable is added, even when the new model provides no additional information that is useful and does not predict any better than before.

Alternative measure of how well the independent variables fit the dependent variable—adjusted $R^2$: adjusted $R^2$ will be smaller than $R^2$ and does not automatically increase when more independent variables are added to the regression.

Formula for adjusted $R^2$ is:

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

**Example:** Calculate the adjusted $R^2$.

**Answer:**

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1-R^2\right) = 1 - \left(\frac{100-1}{100-3-1}\right)\left(1-0.497\right) = 0.481$$

## ISSUES ASSOCIATED WITH MULTIPLE REGRESSION MODEL

1. Heteroskedasticity (異質變異數) In Regression Model

One of the assumptions of multiple linear regression model is that the variance of residuals are constant; however, if the variances of residuals are not constant, heteroskedasticity arises.

- There are two type of heteroskedasticity:

  a. Unconditional Heteroskedasticity

     The variance of the error terms changes over the observation range, but in an unsystematic manner that is not correlated with the value of independent variables.

  b. Conditional Heteroskedasticity

     The variance of the error terms changes in a systematic manner that is correlated with the values of the independent variables.

Only conditional heteroskedasticity needs to be addressed for that their standard error will be underestimated, producing t-scores that are deceptively high.

- Test of Conditional Heteroskedasticity (beyond the CFA exam scope)

  The Breusch-Pagan test can be used to test conditional heteroskedastic.

- Correct for Conditional Heteroskedasticity (beyond the CFA exam scope)

  Compute the robust standard errors which are larger than original errors.

2. Serial Correlation (序列相關) In Regression Residuals (Autocorrelation;自我相關)

Another assumption of regression analysis is that the error terms are not serially correlated; i.e. $r_{\varepsilon_t \varepsilon_{t-1}} = 0$. However, if it is not true, serial correlation (autocorrelation) arises.

The effect of serial correlation is that the standard errors of the regression coefficients may not be correct.

- Testing for Serial Correlation in Regression Residuals

  Most commonly used method for testing serial correlation is the Durbin-Watson statistic (DW):

$$DW = \frac{\sum_{t=1}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

  For large sample sizes, this calculation can be approximated as:

$$DW \approx 2(1 - r_{\varepsilon_t, \varepsilon_{t-1}})$$

  Thus, if there is no autocorrelation, the DW statistic should approximately equal to 2.0.

  If residuals are positively correlated, the DW statistic should be less than 2.0

  If the residuals are negatively correlated, the DW statistic should be greater than 2.0.

- Correcting for Serial Correlation

  a. Adjust standard errors upward to account for the impact of serial correlation (derived by Hansen 1982).

  b. Change the regression equation itself in some ways to eliminate the problem.

3. Multicollinearity (多元共線) Among the Independent Variables of a Regression

   Multicollinearity happens when two or more of the independent variables (or combinations of independent variables) in a regression model are highly (but not perfectly) correlated with each other. When multicollinearity exists, the regression equation can be estimated, but its economic meaning becomes problematic because the same variables has been counted twice in the regression model under two different “names”.

- Testing for Multicollinearity

  Correlation Matrix is usually used for examining correlation coefficients among independent variables. As a common rule, the correlation coefficients between any two independent variables should be 0.7 or higher. Another sign for multicollinearity is when the $R^2$ of the regression model is high, but the regression coefficients are all statistically insignificant.

- Correcting for Multicollinearity

  The most common method for correctingmulticollinearity is to reformulate the regression model, leaving out variables that appear to be redundant, based on their correlations with the other independent variables.

4. Dummy Variables in Regression

   Dummy variables are used to formulate "qualitative" variables rather than "quantitative" variables.

   There are two ways to address qualitative variables:

   a. Find a quantitative measurable proxy for the qualitative variables.

   b. Use a dummy variable to represent the quality being captured by the qualitative variable. The dummy variables are usually assigned the value "1" and "0".

5. Models with Qualitative Dependent Variables

   Qualitative dependent variables, such as default of bonds (value to be 0 if go bankruptcy), are not suited to perform with linear regression analysis. Other models apply.

6. Statistical Modeling

   In order for a multiple regression model to be economically meaningful, two conditions must be met:

   a. The model should have a good theoretical basis.

   b. The model should be able to pass the most stringent statistical tests.

## Appendix

### Use of F test for global test (testing the overall validity of regression)

F test is used for testing validity in the following way

$H_0$: $b_1=b_2=b_3=\ldots=b_k=0$
$H_1$: At least one slope coefficient does not equal zero.

$$F_{calc} = \frac{RSS / k}{SSE /(n-k-1)}$$

**One-tailed test is used to perform the test when finding the $F_{critical}$ value, because of the following reason.**

**The test hypothesis is written in the form of testing whether all regression coefficients are significantly different from zero. However, with the F-test method we actually are testing the explaining power of the regression model – RSS relative to ESS (see the formula above). Which means, if RSS (the variation explained by the regression model) is significant enough (larger than zero), it is impossible that none of the regression coefficient is meaningful (at least one or more coefficients must explains the variation).**

**Thus, we try to prove that the F value deviate from zero (where RSS explains nothing) at a certain confidence level. F value ranges from zero to infinity, therefore one-tailed test is used to perform the test.**