Data and financial modelling

Lianghai Xiao

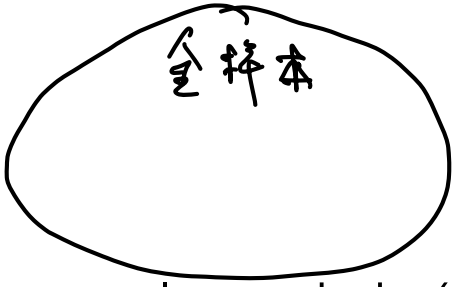
https://github.com/styluck/mat_fin

Outline

- Organizing, Visualizing, and Describing Data
- Probability Concepts
- Common Probability Distributions
- **Sampling and Estimation** 采样、估计
- **Hypothesis Testing** 假设检验
- **Introduction to Linear Regression** 线性回归

SAMPLING AND ESTIMATION

Sampling error 采样误差



- **Sampling error** is the difference between a sample statistic (such as the mean, variance, or standard deviation of the sample) and its corresponding population parameter (the true mean, variance, or standard deviation of the population). For example, the sampling error for the mean is as follows:

- sampling error of the mean =

$$\text{sample mean} - \text{population mean} = \bar{x} - \mu$$

The central limit theorem

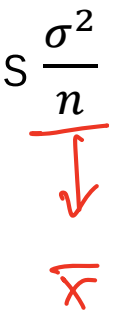
中心极限定理

- The **central limit theorem** states that for simple random samples of size n from a population with a mean μ and a finite variance σ^2, the sampling distribution of the sample mean approaches a normal probability distribution with mean μ and a variance equal to $\frac{\sigma^2}{n}$ as the sample size becomes large.
- The central limit theorem is extremely useful because the normal distribution is relatively easy to apply to hypothesis testing and to the construction of confidence intervals.

置信区间

The central limit theorem

- If the sample size n is sufficiently large ($n \geq 30$), the sampling distribution of the sample means will be **approximately normal**.
- The mean of the population, μ, and the mean of the distribution of all possible sample means are equal.
- The variance of the distribution of sample means is $\frac{\sigma^2}{n}$, the population variance divided by the sample size.



Standard error of sample mean

- The standard error of the sample mean is the standard deviation of the distribution of the sample means.
- When the standard deviation of the population, σ, is known, the standard error of the sample mean is calculated as:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

variance of sample mean = $\frac{\sigma^2}{n}$

where:

$\sigma_{\bar{X}}$ =standard error of the sample mean

σ =standard deviation of the population

- n =size of the sample

Example: Standard error of sample mean

- The mean hourly wage for Iowa farm workers is $13.50 with a population standard deviation of $2.90. Calculate and interpret the standard error of the sample mean for a sample size of 30.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$2.90}{\sqrt{30}} = \$0.53$$

mean hourly wage = $\$13.50 \pm 0.53$

Example: Standard error of sample mean

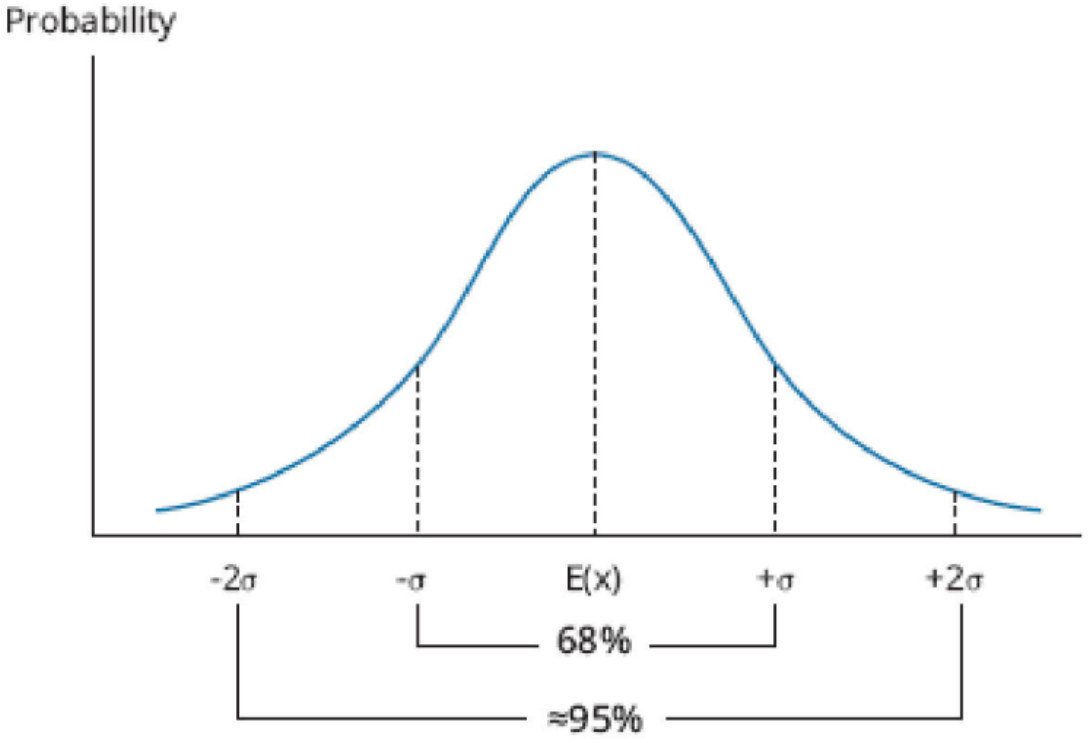
- The mean hourly wage for Iowa farm workers is $13.50 with a population standard deviation of $2.90. Calculate and interpret the standard error of the sample mean for a sample size of 30.
- Answer:
- Because the population standard deviation, σ, is known, the standard error of the sample mean is expressed as:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.90}{\sqrt{30}} = \$0.53$$

Confidence interval

- A confidence interval is a range of values around the expected outcome within which we expect the actual outcome to be some specified percentage of the time.
- A 95% confidence interval is a range that we expect the random variable to be in 95% of the time.
- For a normal distribution, this interval is based on the expected value (sometimes called a point estimate) of the random variable and on its variability, which we measure with standard deviation.

Confidence interval



For any normally distributed random variable, 68% of the outcomes are within one standard deviation of the expected value (mean), and approximately 95% of the outcomes are within two standard deviations of the expected value.

Confidence interval

- In practice, we will not know the actual values for the mean and standard deviation of the distribution, but will have estimated them as X and s. The three confidence intervals of most interest are given by the following:
- The 90% confidence interval for X is
- The 95% confidence interval for X is
- The 99% confidence interval for X is

Confidence interval

- In practice, we will not know the actual values for the mean and standard deviation of the distribution, but will have estimated them as X and s. The three confidence intervals of most interest are given by the following:
- The 90% confidence interval for X is [– 1.65, + 1.65]. $[\mu-1.65\sigma,\ \mu+1.65\sigma]$
- The 95% confidence interval for X is [– 1.96, + 1.96]. $[\mu-1.96\sigma,\ \mu+1.96\sigma]$
- The 99% confidence interval for X is [– 2.58, + 2.58]. $[\mu-2.58\sigma,\ \mu+2.58\sigma]$

黑天鹅事件：发生3-4件事情，出现概率在3σ置信区间外。

Confidence interval for the population mean

- If the population has a normal distribution with a known variance, a confidence interval for the population mean can be calculated as:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

样本均值 → $\bar{X}$　　σ ← 全样本的标准差　　n ← 样本大小

$\bar{X}$ = point estimate of the population mean (sample mean)

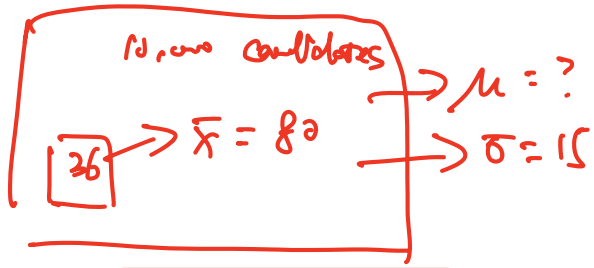
$z_{\alpha/2}$ = reliability factor, a standard normal random variable for which the probability in the right - hand tail of the distribution is $\alpha/2$.

In other words, this is the z-score that leaves of $\alpha/2$ probability in the upper tail.

α → 小概率

$\frac{\sigma}{\sqrt{n}}$ = the standard error of the sample mean where σ is the known standard deviation of the population, and n is the sample size.

Example: Confidence interval



- Consider a practice exam that was administered to 36 candidates. Their mean score on this practice exam was 80. Assuming a population standard deviation equal to 15, construct and interpret a 99% confidence interval for the mean score on the practice exam for all candidates. Note that, in this example, the population standard deviation is known, so we don't have to estimate it.

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$= 80 \pm 2.58 \frac{15}{\sqrt{36}} = 80 \pm 6.45$$

90% : $[-1.65, 1.65]$

95% : $[-1.96, 1.96]$

99% : $[-2.58, 2.58]$

Example: Confidence interval

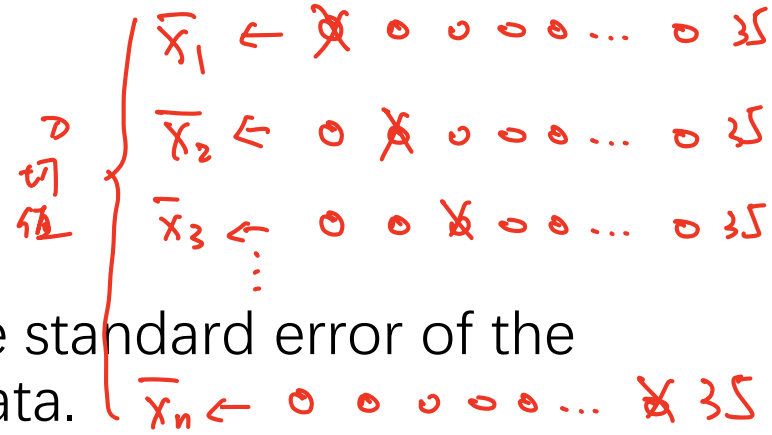
- Consider a practice exam that was administered to 36 candidates. Their mean score on this practice exam was 80. Assuming a population standard deviation equal to 15, construct and interpret a 99% confidence interval for the mean score on the practice exam for all candidates. Note that, in this example, the population standard deviation is known, so we don't have to estimate it.
- At a confidence level of 99%, $z = 2.58$. So, the 99% confidence interval is calculated as follows:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 80 \pm 2.58\frac{15}{\sqrt{36}} = 80 \pm 6.45$$

Resampling: jackknife

刀切法

重采样



- Two alternative methods of estimating the standard error of the sample mean involve resampling of the data.
- The first of these, termed the **jackknife**, calculates multiple sample means, each with one of the observations removed from the sample.
- The standard deviation of these sample means can then be used as an estimate of the standard error of sample means.
- The jackknife is a computationally simple tool and can be used when the number of observations available is relatively small. It can remove bias from statistical estimates.

Resampling: bootstrap

- A **bootstrap** method is more computationally demanding but has some advantages.
- To estimate the standard error of the sample mean, we draw repeated samples of size n from the full data set (replacing the sampled observations each time).
- We can then directly calculate the standard deviation of these sample means as our estimate of the standard error of the sample mean.

Sample Bias

- We have seen so far that a larger sample reduces the sampling error and the standard deviation of the sample statistic around its true (population) value.
- Confidence intervals are narrower when samples are larger and the standard errors of the point estimates of population parameters are less.

$n \uparrow \Rightarrow$ standard error $= \frac{\sigma}{\sqrt{n}} \downarrow$

$n \uparrow \Rightarrow$ confidence interval $\mu \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \downarrow$

数据窥探

Data snooping → 过度的拟合有限的数据

使得模型在样本内出现但样本数据上表现不佳

- **Data snooping** occurs when analysts repeatedly use the same database to search for patterns or trading rules until one that "works" is discovered.
- For example, empirical research has provided evidence that value stocks appear to outperform growth stocks. Some researchers argue that this anomaly is actually the product of data snooping. Because the data set of historical stock returns is quite limited, it is difficult to know for sure whether the difference between value and growth stock returns is a true economic phenomenon or simply a chance pattern that was stumbled upon after repeatedly looking for any identifiable pattern in the data.

样本选择偏误

Sample selection bias

- **Sample selection bias** occurs when some data is systematically excluded from the analysis, usually because of the lack of availability. This practice renders the observed sample to be non-random, and any conclusions drawn from this sample can't be applied to the population because the observed sample and the portion of the population that was not observed are different.

规模以上企业　营收 > 1000万

3.3万亿　3万亿

2024年　2025年

-> 同比增长 13%

幸存者偏差

Survivorship bias

- **Survivorship bias** is the most common form of sample selection bias. A good example of the existence of survivorship bias in investments is the study of mutual fund performance. Most mutual fund databases, only include funds currently in existence—the “survivors.” They do not include funds that have ceased to exist due to closure or merger.

预见偏差（未来数据）

Look-ahead bias

金融数据：时间序列数据

- **Look-ahead bias** occurs when a study tests a relationship using sample data that was not available on the test date. For example, consider the test of a trading rule that is based on the price-to-book ratio at the end of the fiscal year. Stock prices are available for all companies at the same point in time, while end-of-year book values may not be available until 30 to 60 days after the fiscal year ends. In order to account for this bias, a study that uses price-to-book value ratios to test trading strategies might estimate the book value as reported at fiscal year end and the market value two months later

避免 Look-ahead bias：严格遵守时间顺序.

Time-period bias 时间期限偏差

康波周期 ~60年

- **Time-period bias** can result if the time period over which the data is gathered is either too short or too long. If the time period is too short, research results may reflect phenomena specific to that time period, or perhaps even data mining. If the time period is too long, the fundamental economic relationships that underlie the results may have changed.
- For example, research findings may indicate that small stocks outperformed large stocks during 1980–1985. This may well be the result of time-period bias—in this case, using too short a time period. It's not clear whether this relationship will continue in the future or if it is just an isolated occurrence.

假设 检验

HYPOTHESIS TESTING

Hypothesis testing

- A hypothesis is a statement about the value of a population parameter developed for the purpose of testing a theory or belief. Hypotheses are stated in terms of the population parameter to be tested, like the population mean, μ.
- For example, a researcher may be interested in the mean daily return on stock *option*s. Hence, the hypothesis may be that the mean daily return on a portfolio of stock options is positive.

Hypothesis testing

- Hypothesis testing procedures, based on sample statistics and probability theory, are used to determine whether a hypothesis is
  - a reasonable statement and should not be rejected, or
  - if it is an unreasonable statement and should be rejected.

Hypothesis testing

- State the hypothesis
- Select the appropriate test statistic
- Specify the level of significance
- State the decision rule regarding the hypothesis
- Collect the sample arid calculate the sample statistics
- Make a decision regarding the hypothesis
- Make a decision based on the results of the test

The null hypothesis and The alternative hypothesis

零假设　　备选假设

- The null hypothesis, designated H_0, is the hypothesis that is actually tested and is the basis for the selection of the test statistics. The null is generally stated as a simple statement about a population parameter. Typical statements of the null hypothesis for the population mean include

$$H_0\colon \mu = \mu_0,\ H_0\colon \mu \leq \mu_0, \text{ and } H_0\colon \mu \geq \mu_0,$$

- where μ is the population mean and μ_0 is the hypothesized value of the population mean.

H_0：等号倒闭

H_1：等号开口

$H_0: \mu = \mu_0$　　$H_1: \mu \neq \mu_0$

$H_0: \mu \leq \mu_0$　　$H_1: \mu > \mu_0$

$H_0: \mu \geq \mu_0$　　$H_1: \mu < \mu_0$

The null hypothesis and The alternative hypothesis

- The null hypothesis always includes the “equal to” condition. The alternative hypothesis, designated H_a, is what is concluded if there is sufficient evidence to reject the null hypothesis. It is usually the alternative hypothesis that you are really trying to assess.
- Why? Because you can never really prove anything with statistics, when the null hypothesis is discredited, the implication is that the alternative hypothesis is valid.

one-tailed test and two-tailed test

- A two-tailed test for the population mean may be structured as:

$$H_0: \mu = \mu_0 \text{ versus } H_a: \mu \neq \mu_0$$

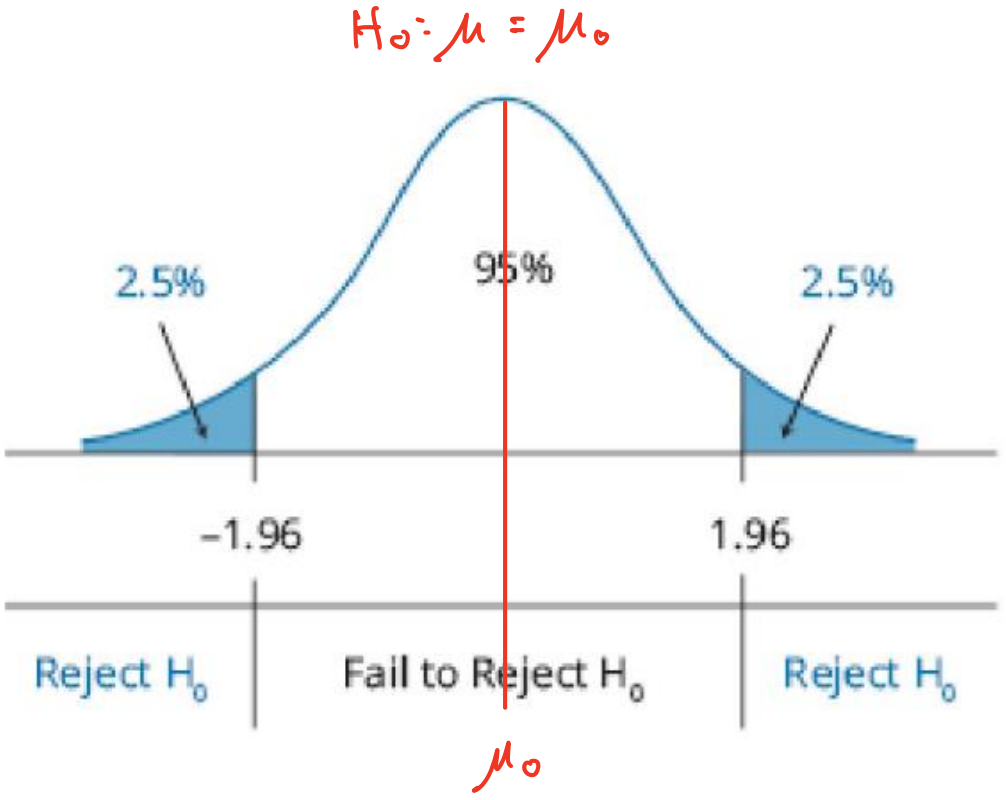
- Since the alternative hypothesis allows for values above and below the hypothesized parameter, a two-tailed test uses two critical values (or rejection points).

one-tailed test and two-tailed test

Reject H_0 if:

test statistic < –1.96 or

test statistic > 1.96



one-tailed test and two-tailed test

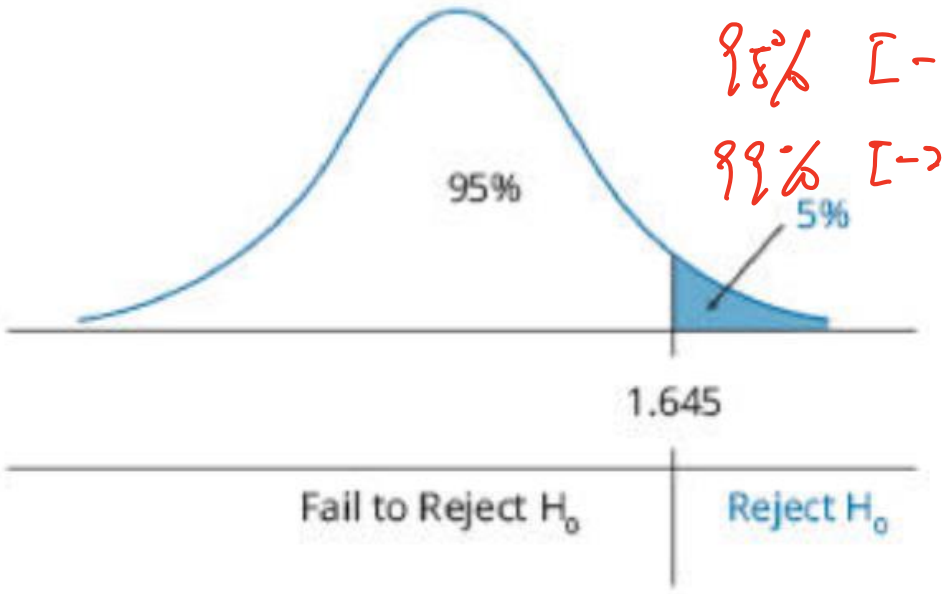
- For a one-tailed hypothesis test of the population mean, the null and alternative hypotheses are either:
  - Upper tail: H_0: $\mu \leq \mu_0$ versus H_a: $\mu > \mu_0$, or
  - Lower tail: H_0: $\mu \geq \mu_0$ versus H_a: $\mu < \mu_0$

one-tailed test and two-tailed test

90% [-1.65, 1.65]

95% [-1.96, 1.96]

99% [-2.58, 2.58]



P-值

P-VALUES AND TESTS OF MEANS

- The p-value is the probability of obtaining a test statistic that would lead to a rejection of the null hypothesis, assuming the null hypothesis is true. It is the smallest level of significance for which the null hypothesis can be rejected.
- For one-tailed tests, the p-value is the probability that lies above the computed test statistic for upper tail tests or below the computed test statistic for lower tail tests.
- 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.

P-VALUES AND TESTS OF MEANS

statistics = 2.3 > 1.96

⇒ Probability = 1.07%

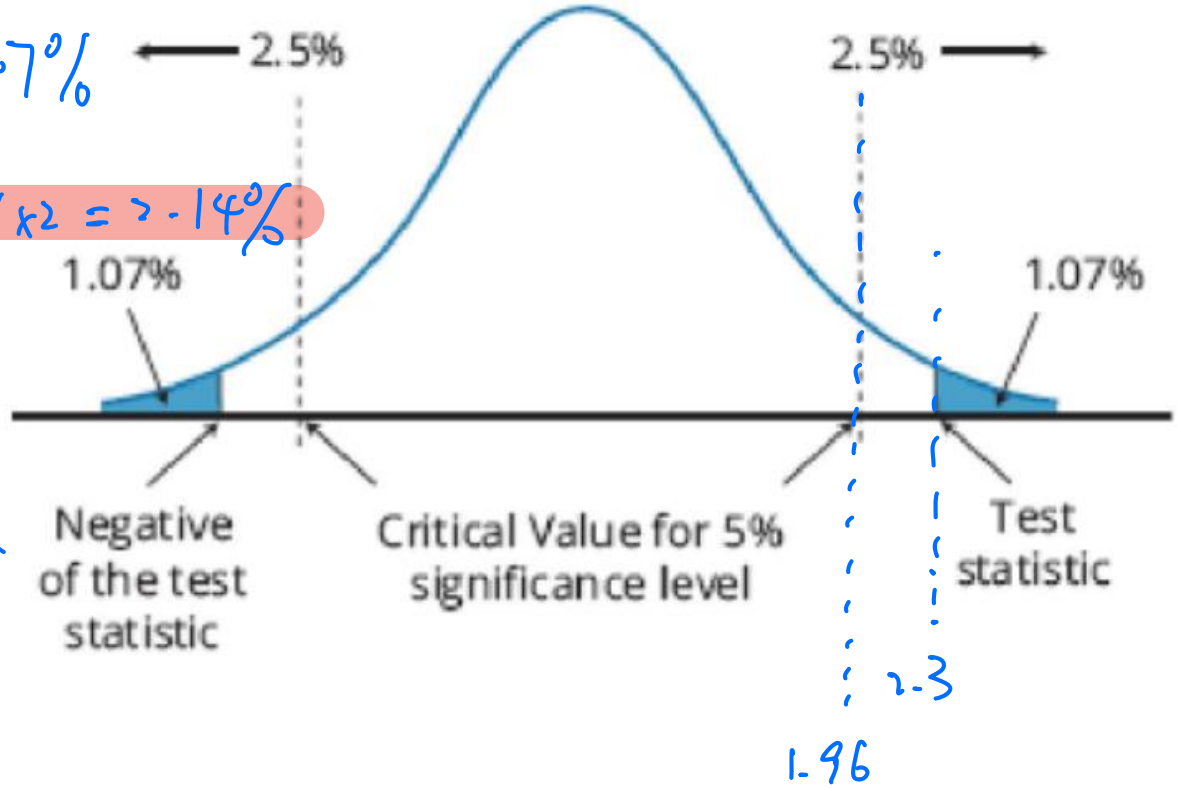
⇒ P-value = 1.07% × 2 = 2.14%

3% 4% 5%

显著性水平下，
我们拒绝零假设

1%，2%

显著性水平下，
我们接受零假设



test statistic

- A test statistic is calculated by comparing the point estimate of the population parameter with the hypothesized value of the parameter (i.e., the value specified in the null hypothesis).
- With reference to our *option* return example, this means we are concerned with the difference between the mean return of the sample (i.e., = 0.001) and the hypothesized mean return (i.e., $\mu_0 = 0$).
- As indicated in the following expression, the test statistic is the difference between the sample statistic and the hypothesized value, scaled by the standard error of the sample statistic.

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

The t-Test

t-检验

- The t-test is a widely used hypothesis test that employs a test statistic that is distributed according to a t-distribution. Following are the rules for when it is appropriate to use the t-test for hypothesis tests of the population mean.
- Use the t-test if the population variance is unknown and either of the following conditions exist:
- The sample is large ($n \geq 30$).
- The sample is small (less than 30), but the distribution of the population is normal or approximately normal.

The t-Test

- If the sample is small and the distribution is nonnormal, we have no reliable statistical test. The computed value for the test statistic based on the t-distribution is referred to as the t-statistic.
- For hypothesis tests of a population mean, a t-statistic with n – 1 degrees of freedom is computed as:

$$t - statstic = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

$\bar{X}$ = sample mean

μ_0 = hypothesized population mean (i.e., the null)

s = standard deviation of the sample

n = sample size

The z-Test

- The z-test is the appropriate hypothesis test of the population mean when the population is normally distributed with known variance. The computed test statistic used with the z-test is referred to as the z-statistic. The z-statistic for a hypothesis test for a population mean is computed as follows:

$$Z - statistic = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$t\text{-statistic} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

$\bar{X}$ = sample mean

μ_0 = hypothesized population mean (i.e., the null)

σ = standard deviation of the *population*

n = sample size

The z-Test

- When the sample size is large and the population variance is unknown, the z-statistic is:

$$Z - statistic = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

$\bar{X}$ = sample mean

μ_0 = hypothesized population mean (i.e., the null)

s = standard deviation of the *sample*

n = sample size

z-test or t-test?

- Z-test: The calculation is based on the standard normal distribution (with a mean of 0 and a standard deviation of 1).
- T-test: The calculation is based on the t-distribution. The t-distribution is similar to the standard normal distribution but has thicker tails and is suitable for small sample sizes. The T-test requires consideration of degrees of freedom (df), which is typically n−1. Degrees of freedom affect the shape of the t-distribution, thereby influencing the outcome of the test.
- The Z-test does not involve the concept of degrees of freedom because it is based on the standard normal distribution. In the Z-test, the obtained z-value is directly related to the cumulative probability of the standard normal distribution, and the corresponding p-value can be directly looked up. In the T-test, the obtained t-value requires looking up the corresponding p-value in the t-distribution table based on the degrees of freedom.

t-table

自由度

Z-table

DIFFERENCE IN MEANS

- Up to this point, we have been concerned with tests of a single population mean.
- In practice, we frequently want to know if there is a difference between the means of two populations.
- The t-test for differences between means requires that we are reasonably certain that our samples are independent and that they are taken from two populations that are normally distributed.

DIFFERENCE IN MEANS

- A pooled variance is used with the t-test for testing the hypothesis that the means of two normally distributed populations are equal, when the variances of the populations are unknown but assumed to be equal.
- Assuming independent samples, the t-statistic is computed as:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

s_1^2 = variance of the first sample
s_2^2 = variance of the second sample
n_1 = number of observations in the first sample
n_2 = number of observations in the second sample

Example: Difference between means – equal variances

- You are investigating whether the abnormal returns for acquiring firms during merger announcement periods differ for horizontal and vertical mergers. You estimate the abnormal returns for a sample of acquiring firms associated with horizontal mergers and a sample of acquiring firms involved in vertical mergers. You find that abnormal returns from horizontal mergers have a mean of 1.0% and a standard deviation of 1.0%, while abnormal returns from vertical mergers have a mean of 2.5% and a standard deviation of 2.0%.
- You assume that the samples are independent, the population means are normally distributed, and the population variances are equal. You calculate the t-statistic as –5.474 and the degrees of freedom as 120. Using a 5% significance level, should you reject or fail to reject the null hypothesis that the abnormal returns to acquiring firms during the announcement period are the same for horizontal and vertical mergers?

Example: Difference between means – equal variances

Partial *t*-Table

df	One-Tailed Probabilities (p)		
	p = 0.10	p = 0.05	p = 0.025
110	1.289	1.659	1.982
120	1.289	1.658	1.980
200	1.286	1.653	1.972

$t = -5.679$

Example: Difference between means – equal variances

μ_1 → horizontal

μ_2 → vertical

- Answer:
- Since this is a two-tailed test, the structure of the hypotheses takes the following form:

$$H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_a: \mu_1 - \mu_2 \neq 0$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} = -5.474$$

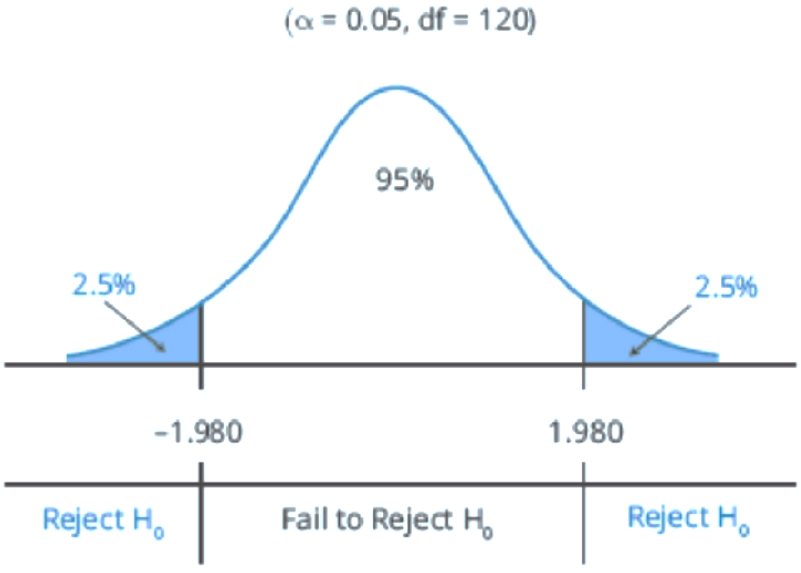
μ_1 = the mean of the abnormal returns for the horizontal mergers

μ_2 = the mean of the abnormal returns for the vertical mergers

- From the following t-table segment, the critical t-value for a 5% level of significance at $\alpha / 2 = p = 0.025$ with $df = 120$, is 1.980.

Example: Difference between means – equal variances

- Thus, the decision rule can be stated as:
    - Reject H_0 if t-statistic < –1.980 or t-statistic > 1.980



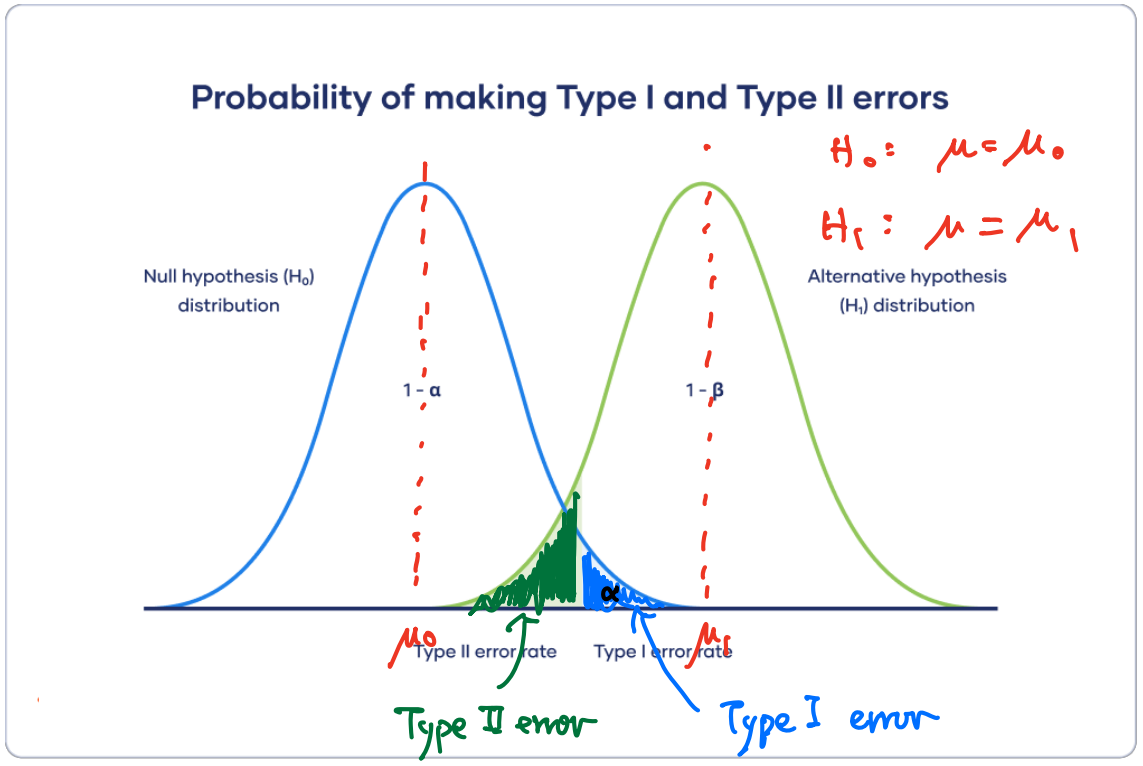
Type I and Type II errors

- Keep in mind that hypothesis testing is used to make inferences about the parameters of a given population on the basis of statistics computed for a sample that is drawn from that population.
- We must be aware that there is some probability that the sample, in some way, does not represent the population, and any conclusion based on the sample about the population may be made in error.
- When drawing inferences from a hypothesis test, there are two types of errors:
- **Type I error**: the rejection of the null hypothesis when it is actually true.
- **Type II error**: the failure to reject the null hypothesis when it is actually false.

Type I and Type II errors

Type I and Type II Error		
Null hypothesis is ...	**True**	**False**
Rejected	**Type I error** False positive Probability = **α**	**Correct decision** True positive Probability = 1 - **β**
Not rejected	**Correct decision** True negative Probability = 1 - **α**	**Type II error** False negative Probability = **β**

Type I and Type II errors

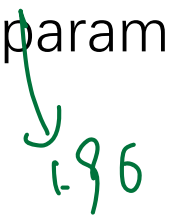


The Relation Between Confidence Intervals and Hypothesis Tests

- A confidence interval is determined as:

$$[\text{sample statistic} - (\text{critical value})(\text{standard error})]$$
$$\leq \text{population parameter}$$
$$\leq [\text{sample statistic} + (\text{critical value})(\text{standard error})]$$

- The interpretation of a confidence interval is that for a level of confidence of 95%; for example, there is a 95% probability that the true population parameter is contained in the interval.



The Relation Between Confidence Intervals and Hypothesis Tests

- From the previous expression, we see that a confidence interval and a hypothesis test are linked by the critical value. To see this relationship more clearly, the expression for the confidence interval can be manipulated and restated as:

$$-\text{critical value} \leqslant \text{test statistic} \leqslant +\text{critical value}$$

- This is the range within which we fail to reject the null for a two-tailed hypothesis test at a given level of significance.

Example: Confidence intervals and two-tailed hypothesis tests

- A researcher has gathered data on the daily returns on a portfolio of call options over a recent 250-day period. The mean daily return has been 0.1%, and the sample standard deviation of daily portfolio returns is 0.25%. The researcher believes that the mean daily portfolio return is not equal to zero.
- 1. Construct a 95% confidence interval for the population mean daily return over the 250-day sample period.
- 2. Construct a hypothesis test of the researcher's belief.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.25\%}{\sqrt{250}} = 0.0158\%$$

$$\bar{x} \pm 1.96 s_{\bar{x}} = 0.1\% \pm 1.96 \times 0.0158\% = 0.1\% \pm 0.0310\%$$

Confidence interval : $[0.1\% - 0.031\%,\ 0.1\% + 0.031\%]$

Example: Confidence intervals and two-tailed hypothesis tests

- Answer:
- 1. Given a sample size of 250 with a standard deviation of 0.25%, the standard error can be computed as

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.25\%}{\sqrt{250}} = 0.0158\%$$

2. Null hypothesis : $H_0: \mu = 0$, Alternative Hypothesis : $H_a: \mu \neq 0$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{0.1\%}{0.0158\%} = 6.38 > 2.58$$

∴ at 1% confidence level, null hypothesis is rejected.

Whether a Statistically Significant Result is Also Economically Meaningful

- Statistical significance does not necessarily imply economic significance. For example, we may have tested a null hypothesis that a strategy of going long all the stocks that satisfy some criteria and shorting all the stocks that do not satisfy the criteria resulted in returns that were less than or equal to zero over a 20-year period.
- Assume we have rejected the null in favor of the alternative hypothesis that the returns to the strategy are greater than zero (positive). This does not necessarily mean that investing in that strategy will result in economically meaningful positive returns. Several factors must be considered.

Revisiting example:

- Joe Andrews is examining changes in estimated betas for the common stock of companies in the telecommunications industry before and after deregulation. Andrews believes that the betas may decline because of deregulation since companies are no longer subject to the uncertainties of rate regulation or that they may increase because there is more uncertainty regarding competition in the industry. Andrews calculates a t-statistic of 10.26 for this hypothesis test, based on a sample size of 39. Using a 5% significance level, determine whether there is a change in betas.

Null hypothesis H_0: β is unchanged. $\mu = 0$

H_a: β is changed $\mu \neq 0$

Revisiting example:

$df = n - 1 = 39 - 1 = 38$

Partial *t*-Table

df	One-Tailed Probabilities (p)		
	p = 0.10	p = 0.05	p = 0.025
38	1.304	1.686	2.024
39	1.304	1.685	2.023
40	1.303	1.684	2.021

$t = 10.26 \geq 2.024$

Revisiting example:

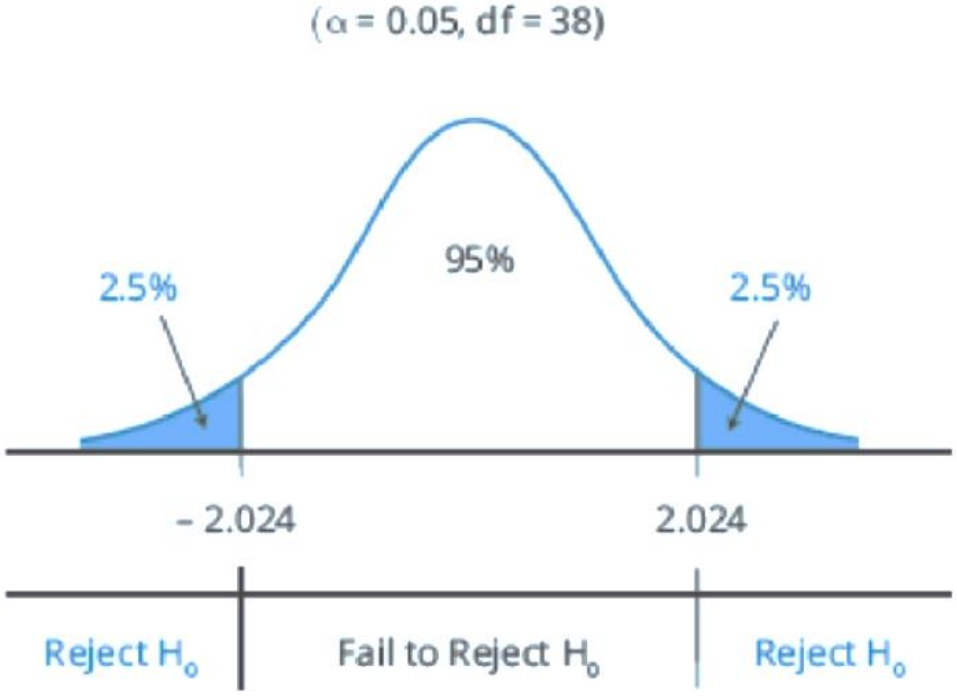
- Because the mean difference may be positive or negative, a two-tailed test is in order here. Thus, the hypotheses are structured as:

$$H_0: \mu_d = 0 \; versus \; H_a: \mu_d \neq 0$$

- There are 39 – 1 = 38 degrees of freedom. Using the t-distribution, the two-tailed critical t-values for a 5% level of significance with df = 38 is ±2.024. As indicated in the following table, the critical t-value of 2.024 is located at the intersection of the p = 0.025 column and the df = 38 row. The one-tailed probability of 0.025 is used because we need 2.5% in each tail for 5% significance with a two-tailed test.

Revisiting example:

- Thus, the decision rule becomes:
    - Reject H_0 if t-statistic < –2.024, or t-statistic > 2.024



线性回归

INTRODUCTION TO LINEAR REGRESSION

金融数据 传感收集.

simple linear regression

- The purpose of simple linear regression is to explain the variation in a dependent variable in terms of the variation in a single independent variable. Here, the term "variation" is interpreted as the degree to which a variable differs from its mean value. Don't confuse variation with variance—they are related but are not the same.

$$\text{variation in } Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

simple linear regression

因变量

- The **dependent variable** is the variable whose variation is explained by the independent variable. We are interested in answering the question, “What explains fluctuations in the dependent variable?” The dependent variable is also referred to as the explained variable, the endogenous variable, or the predicted variable.

被解释变量，内生变量，被预测变量

自变量

- The **independent variable** is the variable used to explain the variation of the dependent variable. The independent variable is also referred to as the explanatory variable, the exogenous variable, or the predicting variable.

解释变量，外生变量，预测变量

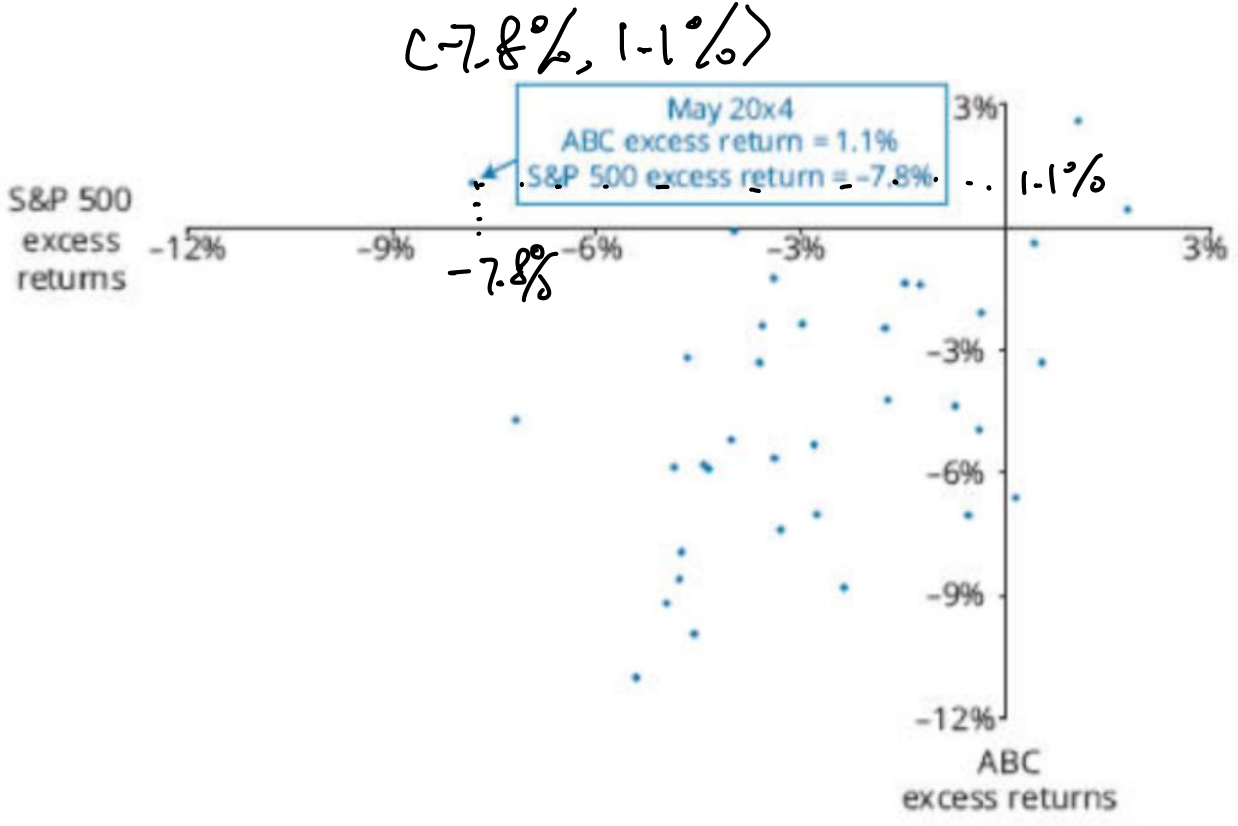
simple linear regression

- Suppose we want to use excess returns on the S&P 500 (the independent variable) to explain the variation in excess returns on ABC common stock (the dependent variable). For this model, we define excess return as the difference between the actual return and the return on 1-month Treasury bills.

$(R_{S\&P500} - R_{risk\text{-}free})$ → independent variable

$(R_{ABC} - R_{risk\text{-}free})$ → dependent variable

simple linear regression



simple linear regression

- The following linear regression model is used to describe the relationship between two variables, X and Y:

$$Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$

where:
Y_i = ith observation of the dependent variable, Y
X_i = ith observation of the independent variable, X
b_0 = regression intercept term 回归截距项
b_1 = regression slope coefficient 回归斜率系数
ϵ_i = residual for the ith observation (also referred to as the disturbance term or error term) ↳误差项

残差

simple linear regression

- Based on this regression model, the regression process estimates an equation for a line through a scatter plot of the data that "best" explains the observed values for Y in terms of the observed values for X. The linear equation, often called the line of best fit or regression line, takes the following form:

$$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1} X_i, \quad i = 1,2,3, \ldots, n$$

where:

$\widehat{Y_i}$ = estimated value of Y_i given X_i

$\widehat{b_0}$ = estimated intercept term

$\widehat{b_1}$ = estimated slope coefficient

ordinary least square (OLS)

simple linear regression

- The sum of the squared vertical distances between the estimated and actual Y-values is referred to as **the sum of squared errors (SSE).**

$$\text{error}: \sum_{i=1}^{n} |y_i - \hat{y}_i|^2$$

- Thus, the regression line is the line that minimizes the SSE. This explains why simple linear regression is frequently referred to as **ordinary least squares (OLS) regression**, and the values determined by the estimated regression equation, , are called least squares estimates.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\min_{\beta_0, \beta_1} SSE \iff \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \iff \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$y = \beta_0 + \beta_1 x$ $\quad$ y : ABC excess return

x : S&P 500 excess return

가장 간단: Ordinary least square (OLS)

$\min_{\beta_0, \beta_1} \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2$ $\quad$ $\hat{y}_i - y_i$ = estimation error

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\Rightarrow \min_{\beta_0, \beta_1} \frac{1}{N}\sum_{i=1}^{N} (\beta_0 + \beta_1 x_i - y_i)^2$

$= \frac{1}{N}\left(\sum_{i=1}^{N} (y_i - \beta_1 x_i)^2 - 2\beta_0 \sum_{i=1}^{N} (y_i - \beta_1 x_i) + N\beta_0^2 \right)$

$\Rightarrow -\frac{2}{N}\sum_{i=1}^{N} (y_i - \beta_1 x_i) + 2\beta_0 = 0$ $\quad \Rightarrow \beta_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i \right)$

$\Rightarrow \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$

$\Rightarrow \min_{\beta_0, \beta_1} \frac{1}{N}\sum_{i=1}^{N} (\beta_0 + \beta_1 x_i - y_i)^2$

$= \frac{1}{N}\left(\sum_{i=1}^{N} (y_i - \beta_0)^2 - 2\sum_{i=1}^{N} (y_i - \beta_0) x_i \beta_1 + \sum_{i=1}^{N} (x_i \beta_1)^2 \right)$

$= \frac{1}{N}\left(\sum_{i=1}^{N} (y_i - \beta_0)^2 - 2\beta_1 \sum_{i=1}^{N} (y_i - \beta_0) x_i + \beta_1^2 \sum_{i=1}^{N} x_i^2 \right)$

$\Rightarrow -\frac{2}{N}\sum_{i=1}^{N} (y_i - \beta_0) x_i + 2\beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$

$\Rightarrow -\frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) x_i + \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 0$

$\Rightarrow \beta_1 \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} \bar{x} x_i \right) = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y}) x_i$

$\Rightarrow \beta_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y}) x_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} \bar{x} x_i} = \frac{n\,\mathrm{cov}(x,y)}{n\,\mathrm{var}(x)} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$

$\sum x_i = n\bar{x}$

$\bar{x} = \frac{1}{n}\sum x_i$

simple linear regression

- The estimated slope coefficient 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:

$$\hat{b}_1 = \frac{\text{Cov}_{XY}}{\sigma_X^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y}) / n-1}{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

simple linear regression

- The intercept term $\hat{b}_0$ 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:

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

where:

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

Example: Computing the slope coefficient and intercept term

- Compute the slope coefficient and intercept term for the ABC regression example using the following information:

Cov(S&P 500, ABC)	= 0.000336	Mean return, S&P 500	= –2.70%
Var(S&P 500)	= 0.000522	Mean return, ABC	= –4.05%

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = -4.05\% - 0.64 \cdot (-2.70\%) = -2.3\%$$

$$\hat{b}_1 = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{0.000336}{0.000522} \approx 0.64$$

Example: Computing the slope coefficient and intercept term

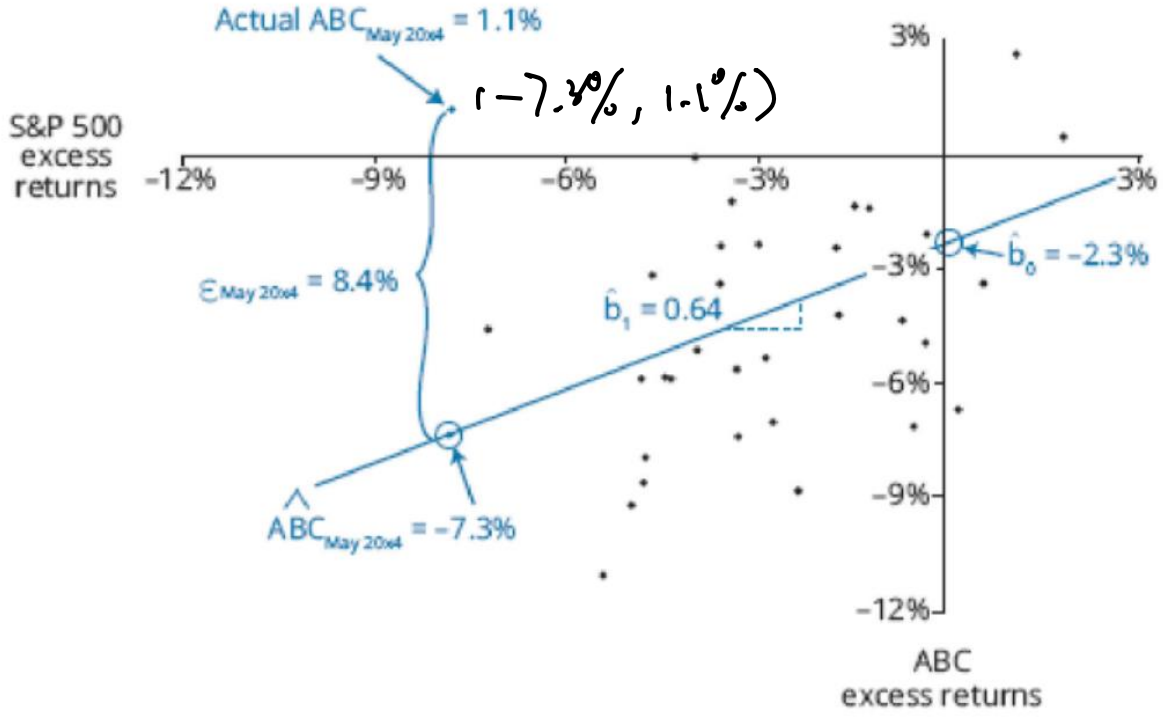
- Compute the slope coefficient and intercept term for the ABC regression example using the following information:
- Answer:
- The slope coefficient is calculated as = 0.000336 / 0.000522 = 0.64. The intercept term is

$$\hat{b}_0 = \overline{\text{ABC}} - \hat{b}_1 \overline{\text{S\&P 500}} = -4.05\% - 0.64(-2.70\%) = -2.3\%$$

simple linear regression

$Y = \beta_0 + \beta_1 X$

$ABC = -2.3\% + 0.64 \cdot S\&P500$



Example: Interpreting regression coefficients

- In the ABC regression example, the estimated slope coefficient was 0.64 and the estimated intercept term was –2.3%. Interpret each coefficient estimate.

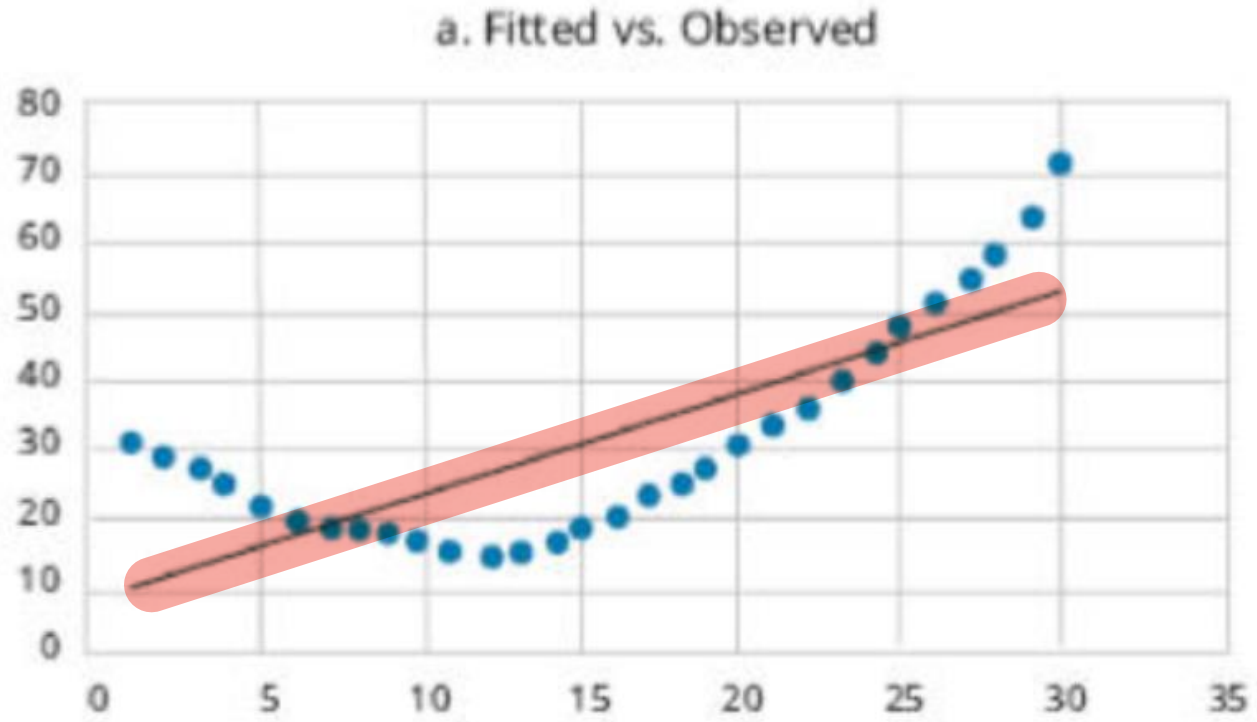
Example: Interpreting regression coefficients

- In the ABC regression example, the estimated slope coefficient was 0.64 and the estimated intercept term was –2.3%. Interpret each coefficient estimate.
- Answer:
- The slope coefficient of 0.64 can be interpreted to mean that when excess S&P 500 returns increase (decrease) by 1%, ABC excess return is expected to increase (decrease) by 0.64%.
- The intercept term of –2.3% can be interpreted to mean that when the excess return on the S&P 500 is zero, the expected return on ABC stock is –2.3%.

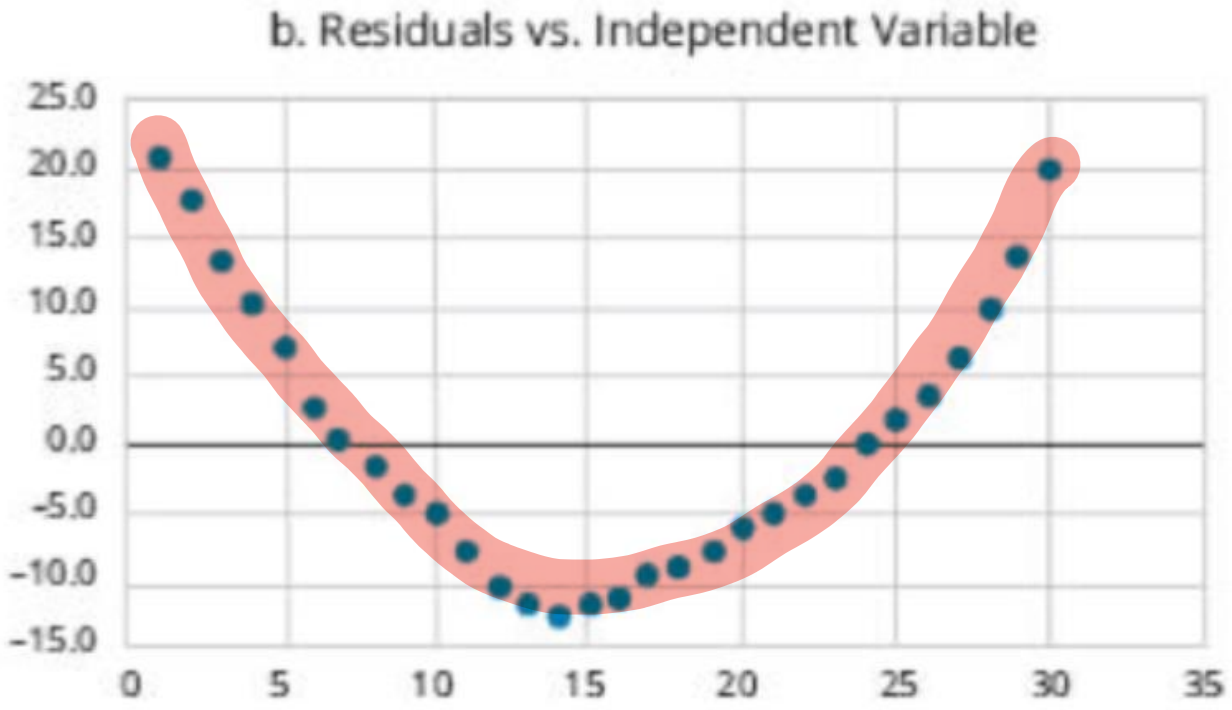
Assumptions

- Linear regression assumes the following:
- 1. A linear relationship exists between the dependent and the independent variables.
- 2. The variance of the residual term is constant for all observations (homoskedasticity). 同方差假设
- 3. The residual term is independently distributed; that is, the residual for one observation is not correlated with that of another observation.
- 4. The residual term is normally distributed.

Linear Relationship



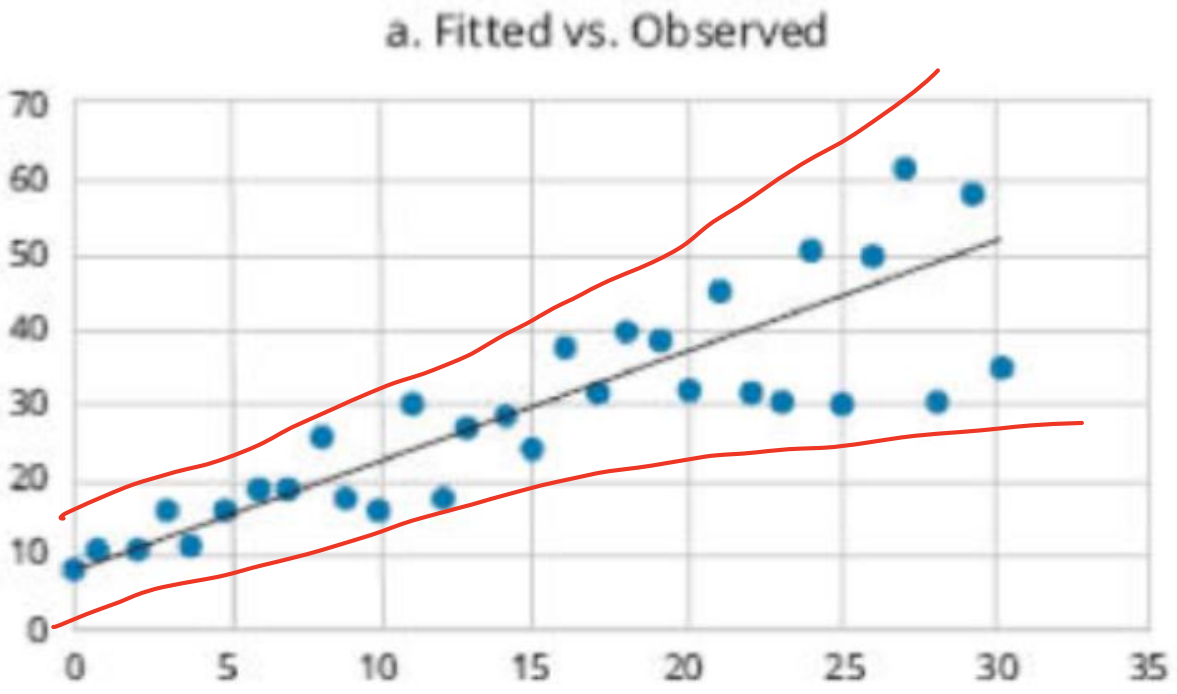
Linear Relationship



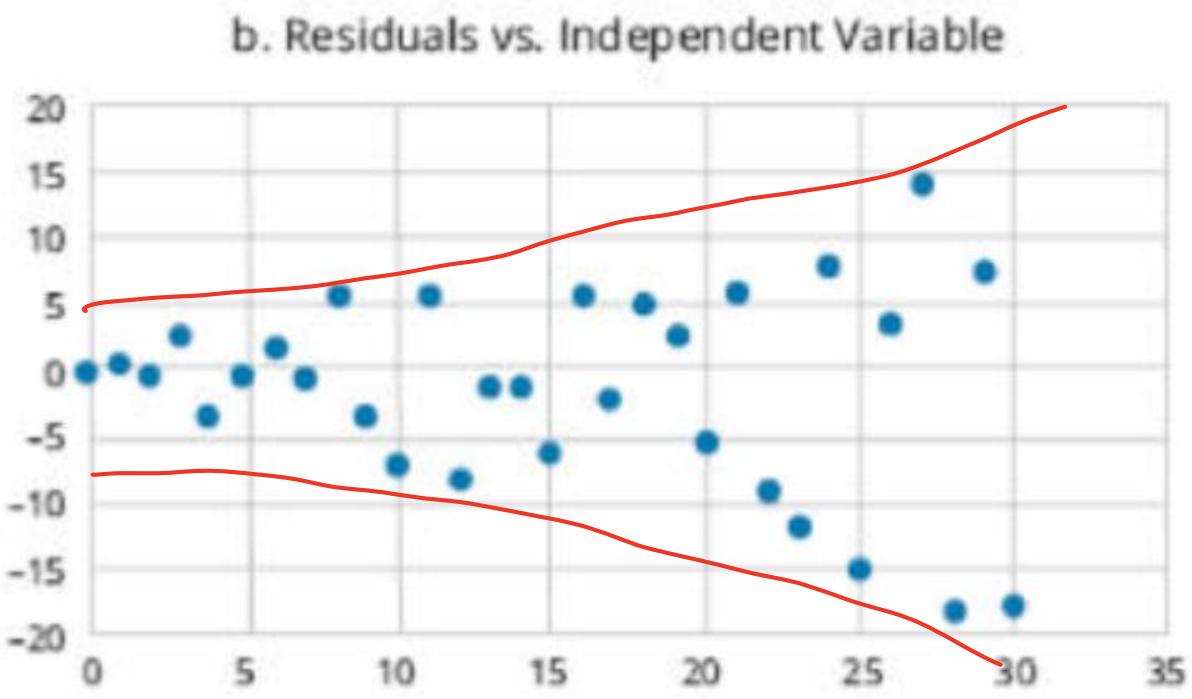
Homoskedasticity

- **Homoskedasticity** refers to the case where prediction errors all have the same variance.
- Heteroskedasticity refers to the situation when the assumption of homoskedasticity is violated.

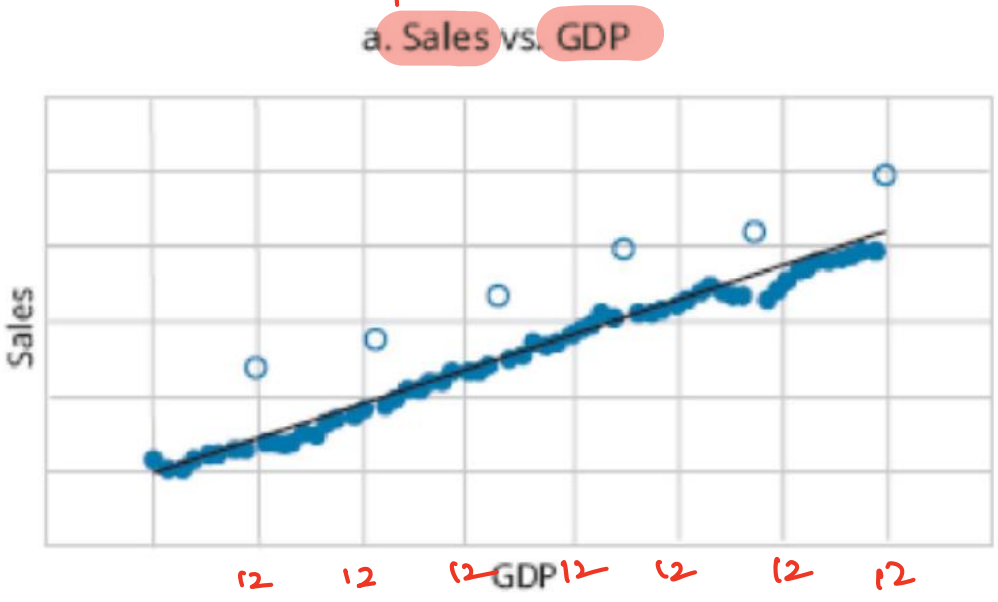
Homoskedasticity

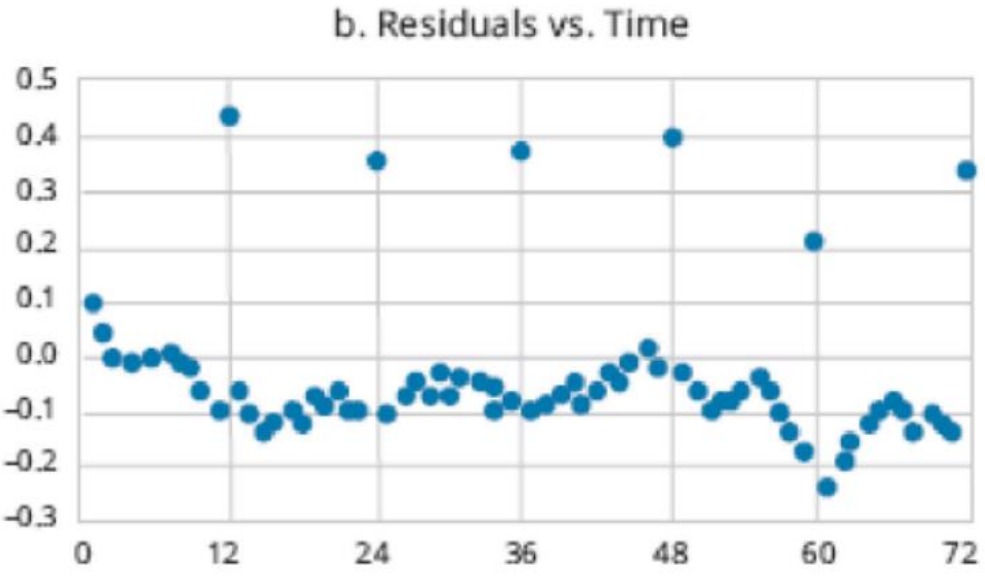


Homoskedasticity

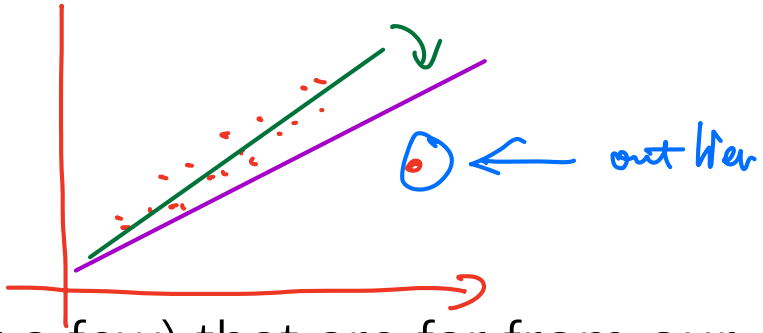


Independence





Outliers 异常值



- Outliers are observations (one or a few) that are far from our regression line (have large prediction errors or X values that are far from the others). Outliers will influence our parameter estimates so that the OLS model will not fit the other observations well.

analysis of variance

- **Analysis of variance (ANOVA)** is a statistical procedure for analyzing the total variability of the dependent variable. Let's define some terms before we move on to ANOVA tables:
- **Total sum of squares (SST)** measures the total variation in the dependent variable. SST is equal to the sum of the squared differences between the actual Y-values and the mean of Y.

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

analysis of variance

回归法

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **Sum of squares regression (SSR)** measures the variation in the dependent variable that is explained by the independent variable. SSR is the sum of the squared distances between the predicted Y-values and the mean of Y.

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

误差方差

- **Sum of squared errors (SSE)** measures the unexplained variation in the dependent variable. It's also known as the sum of squared residuals or the residual sum of squares. SSE is the sum of the squared vertical distances between the actual Y-values and the predicted Y-values on the regression line.

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

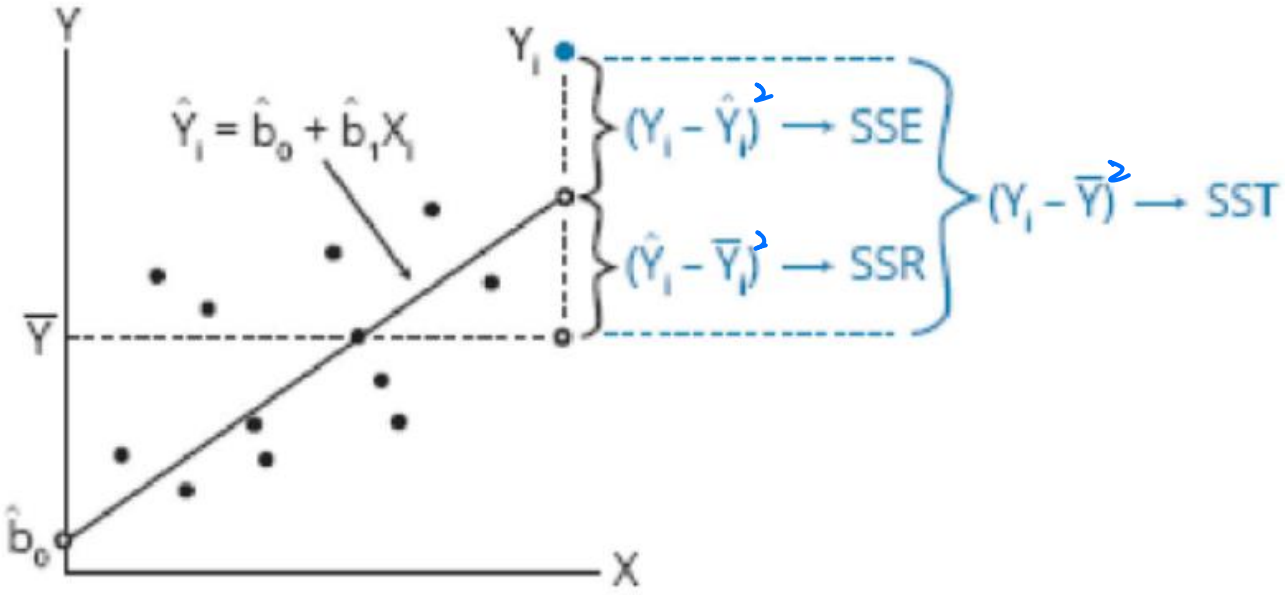
analysis of variance

- You probably will not be surprised to learn that:
- total variation = explained variation + unexplained variation
- or:

$$SST = SSR + SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

analysis of variance



analysis of variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	SSR	$MSR = \frac{SSR}{k} = \frac{SSR}{1} = SSR$
Error (unexplained)	$n - 2$	SSE	$MSE = \frac{SSE}{n-2}$
Total	$n - 1$	SST	

自由度

number of independent variable
number of x

n: sample size

$n-k-1$, k: number of x

Standard Error of Estimate (SE E)

- SEE for a regression is the standard deviation of its residuals. The lower the SEE, the better the model fit.

$$SEE = \sqrt{MSE}$$

coefficient of determination (R^2)

R-squared

- The coefficient of determination (R^2) is defined as the percentage of the total variation in the dependent variable explained by the independent variable. For example, an R^2 of 0.63 indicates that the variation of the independent variable explains 63% of the variation in the dependent variable.

$$R^2 = SSR / SST$$

SSR: sum of squared regression $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

SST: total sum of square $\sum_{i=1}^{n} (y_i - \bar{y})^2$

Example: Using the ANOVA table

n = 36

- Complete the ANOVA table for the ABC regression example and calculate the R^2 and the standard error of estimate (SEE).

Partial ANOVA Table for ABC Regression Example

	Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
SSR	Regression (explained)	? 1	0.00756	?
SSE	Error (unexplained)	? n-2=34	0.04064	?
SST	Total	? 35	?0.04820	

SST = SSR + SSE

$R^2 = \frac{SSR}{SST} = \frac{0.00756}{0.04820} = 15.8\%$

Example: Using the ANOVA table

- Recall that the data included three years of monthly return observations, so the total number of observations (n) is 36.

Completed ANOVA Table for ABC Regression Example

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	**1**	0.0076	**0.0076**
Error (unexplained)	**34**	0.0406	**0.0012**
Total	**35**	**0.0482**	

$$R^2 = \frac{\text{explained variation (SSR)}}{\text{total variation (SST)}} = \frac{0.0076}{0.0482} = 0.158 \text{ or } 15.8\%$$

$$\text{SEE} = \sqrt{\text{MSE}} = \sqrt{0.0012} = 0.035$$

The F-Statistic

- An F-test assesses how well a set of independent variables, as a group, explains the variation in the dependent variable.
- The F-statistic is calculated as:

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/n-k-1}$$

where:

MSR = mean regression sum of squares

MSE = mean squared error

The F-Statistic

- or simple linear regression, there is only one independent variable, so the F-test is equivalent to a t-test for statistical significance of the slope coefficient:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0$$

- calculated F-statistic is compared with the critical F-value, Fc, at the appropriate level of significance.
- The degrees of freedom for the numerator and denominator with one independent variable are:

$$df_{numerator} = k = 1$$
$$df_{denominator} = n - k - 1 = n - 2$$

where:
n = number of observations

Example: Calculating and interpreting the F-statistic

- Use the completed ANOVA table from the previous example to calculate and interpret the F-statistic. Test the null hypothesis at the 5% significance level that the slope coefficient is equal to 0.

Example: Calculating and interpreting the F-statistic

$$F = \frac{MSR}{MSE} = \frac{0.0076}{0.0012} = 6.33$$

$$df_{numerator} = k = 1$$

$$df_{denominator} = n - k - 1 = 36 - 1 - 1 = 34$$

- The null and alternative hypotheses are:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0.$$

- The critical F-value for 1 and 34 degrees of freedom at a 5% significance level is approximately 4.1. (Remember, it's a one-tail test, so we use the 5% F-table!) Therefore, we can reject the null hypothesis and conclude that the slope coefficient is significantly

t-statistic

- A t-test may also be used to test the hypothesis that the true slope coefficient, b_1, is equal to a hypothesized value. Letting $\hat{b}_1$ be the point estimate for b_1, the appropriate test statistic with n −2 degrees of freedom is:

$$t_{b_1} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

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

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

t-statistic

- Rejection of the null supports the alternative hypothesis 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 null hypothesis is that the true slope is zero ($b_1 = 0$). The appropriate test structure for the null and alternative hypotheses is:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0$$

Example: Hypothesis test for significance of regression coefficients

- The estimated slope coefficient from the ABC example 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.

Example: Hypothesis test for significance of regression coefficients

- The calculated test statistic is:
- $$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.64 - 0}{0.26} = 2.46$$
- 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 for a simple linear regression is equivalent to a t-test for the correlation coefficient between x and y:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$