GE2262 Business Statistics - Review

Understanding the Formulae Sheet

(1)

- $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
- $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Reminder: N, μ, σ^2, π are population parameters; $n, \bar{X}, s^2, p$ are sample statistics. β_0, β_1 is for true regression line, b_0, b_1 is for estimated regression line.

(2)

- $\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i)$
- $\sigma^2 = Var(X) = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)$

Corollary: $\sigma^2 = E(X^2) - E(X)^2 = \sum_{i=1}^{N} x_i^2 P(x_i) - \mu^2$

(3)

$X \sim \mathcal{B}(n, \pi)$

- $P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}$
- $\mu = E(X) = n\pi$
- $\sigma^2 = Var(X) = n\pi(1-\pi)$

(4)

- $P(A) = P(A \wedge B_1) + P(A \wedge B_2) + \cdots + P(A \wedge B_n)$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- $P(A|B) = \frac{P(A \wedge B)}{P(B)}$
- $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$

Reminder: to prove two events are independent, show $P(A|B) = P(A)$.

(5)

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1^2)$

(6)

- $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1^2)$

Reminder: to prove the sample mean is normally distributed, one of the following conditions must be satisfied:

- the sample is drawn from a normal population
- sample size $n \geq 30$ and apply Central Limit Theorem
- the population is unknown and $n < 30$, assume the population is normal

(7)

- $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- $\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

Reminder: $\frac{\sigma}{\sqrt{n}}$ is the standard error. $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the margin of error / sampling error.

(8)

- $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
- $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Reminder: 3 types of hypothesis testing:

Type	H_0	H_1	CV	RR	P-value
Two-tail	$\mu = \mu_0$	$\mu \neq \mu_0$	$\pm Z_{\alpha/2}$	$Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$	$P(z < -\|Z\|) + P(z > \|Z\|)$
Lower-tail	$\mu \geq \mu_0$	$\mu < \mu_0$	$-Z_\alpha$	$Z < -Z_\alpha$	$P(z < Z)$
Upper-tail	$\mu \leq \mu_0$	$\mu > \mu_0$	Z_α	$Z > Z_\alpha$	$P(z > Z)$

(9)

- $p \sim \mathcal{N}(\pi, \frac{\pi(1-\pi)}{n})$
- $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathcal{N}(0, 1^2)$
- $p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- $E = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$
- $Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

(10)

- $S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$ (Sample Covariance)
- $r_{XY} = \frac{S_{XY}}{S_X S_Y}$ (Coefficient of Correlation)
- SST = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (Sum of Squares Total)
- SSR = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ (Sum of Squares Regression)
- SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ (Sum of Squares Error)
- SST = SSR + SSE
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ (Coefficient of Determination)
- $Y = \beta_0 + \beta_1 X + \epsilon$ (True Regression Line)
- $b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = r_{XY}\frac{S_Y}{S_X}$ (Least Squares Estimator)
- $b_0 = \bar{Y} - b_1\bar{X}$
- $b_1 \pm t_{\alpha/2,n-2}S_{b_1}$ (Confidence Interval for Slope)
- $t = \frac{b_1-\beta_1}{S_{b_1}}$ (Hypothesis Testing for Slope)
- $S_{b_1}^2 = \frac{SSE}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

1. Introduction to Statistics

- Statistics: the branch of mathematics that transforms data into useful information for decision making.
- Descriptive Statistics: collecting, summarizing (numbers, tables, graphs), and describing data (distribution, central tendency, variability).
- Inferential Statistics: making inferences about **population** based on **sample** data.

Process of a Statistical Study:

Population → Sample → Sample Statistics → Population Parameters

- Variable: a characteristic, number or quantity that can be measured or counted.
- Data: the values measured or collected for each variable.

Types of Variables:

- Numerical variables (counted or measured)
  - Discrete: finite number of values
  - Continuous: infinite number of values
- Categorical variables (defined by categories)
  - Nominal: no order
  - Ordinal: order

1.1. Data Tables and Graphs

- **Summary table**: a table that lists the number of observations for each category (e.g. red, blue, green).
  - Applicable: categorical variables
- **Bar chart**: a chart that uses bars to represent the frequency of each category.
  - Applicable: categorical variables
  - Can use **frequency** or **relative frequency**. Values should not be marked on the bars.
  - Bars are not connected.
- **Pie chart**: a chart that uses slices to represent the proportion of each category.
  - Applicable: categorical variables
  - Should use **relative frequency**. Values should be marked on the slices.
- **Frequency distribution**: a table that lists the number of observations within each interval. (e.g. $[0, 100), [100, 200), \cdots$)
  - Applicable: numerical variables
  - Can use **frequency** or **relative frequency**.
- **Histogram**: a chart that uses bars to represent the frequency of each interval.
  - Applicable: numerical variables
  - Bars are connected (i.e. no gaps between bars). Values should not be marked on the bars.

Principles of a Good Graph:

- uniform and appropriate (i.e. not too compressed or too spread out) scale
- properly labeled axes
- y-axis starts at 0 (and goes down for negative values)
- title

1.2. Central Tendency, Variability, and Shape

Central Tendency:

- **Mean**
  - **Sample mean**: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
  - **Population mean**: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- **Median**:
  - $a_{\frac{n+1}{2}}$ if n is odd
  - $\frac{1}{2}(a_{\frac{n}{2}} + a_{\frac{n}{2}+1})$ if n is even
- **Mode**: the value that occurs most frequently
  - In a continuous distribution, the mode is the peak of the distribution.
  - No mode if all values are unique.
  - Several modes if multiple values occur with the same frequency.

Variability:

- **Range**: $R = X_{max} - X_{min}$
- **IQR**: $Q_3 - Q_1$
  - $Q_1 = a_{\frac{n+1}{4}}$ is the lower quartile
  - $Q_3 = a_{\frac{3(n+1)}{4}}$ is the upper quartile
  - e.g. when $n = 10$, $Q_1 = a_{\frac{11}{4}} = \frac{3}{4}a_3 + \frac{1}{4}a_4, Q_3 = a_{\frac{33}{4}} = \frac{3}{4}a_8 + \frac{1}{4}a_9$
  - Outliers: $(-\infty, Q_1 - 1.5IQR) \cup (Q_3 + 1.5IQR, +\infty)$
- **Variance**:
  - **Sample variance**: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
  - **Population variance**: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- **Standard deviation**: square root of variance

Shape:

	Left Skew	Symmetric	Right Skew
Longer tail	Left	None	Right
	Mean < Median	Mean = Median	Mean > Median
Skewness	< 0	0	> 0

Boxplot (Five-number summary):

- X_{min}, Q_1, Median, Q_3, X_{max}
- Each interval contains 25% of the data.

X_{min} $\Rightarrow$ Whisker $\Rightarrow$ Q_1 $\Rightarrow$ Box $\Rightarrow$ Median $\Rightarrow$ Box $\Rightarrow$ Q_3 $\Rightarrow$ Whisker $\Rightarrow$ X_{max}

2. Probability

- **Outcome**: a possible result of an experiment.
- **Event**: a collection of outcomes.
- **Complement event**: $A' = \{x \in S : x \notin A\}$

- **Sample space** S: the set of all possible outcomes.
- **Probability**: the likelihood of an event occurring.
  - A priori probability: computed theoretically
  - Empirical probability: computed with experimental data
  - Subjective probability: based on personal judgment
- **Joint probability**: probability of two or more events occurring together.
- **Marginal probability**: probability of a single event occurring. (simple probability)
  - $P(A) = P(A \wedge B_1) + P(A \wedge B_2) + \cdots + P(A \wedge B_n)$
  - where $B_1, B_2, \cdots, B_n$ are mutually exclusive and collectively exhaustive.
- **Mutually exclusive events**: $P(A \wedge B) = 0$
- **Collectively exhaustive events**: $P(A \vee B) = 1$

2.1. Rules of Probability

- **Axioms**:
  - $0 \leq P(A) \leq 1$
  - $P(S) = 1$
- **General addition rule**:
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - When A and B are mutually exclusive, $P(A \vee B) = P(A) + P(B)$
- **Conditional probability**:
  - $P(A|B) = \frac{P(A \wedge B)}{P(B)}$ is the probability of A given B. ($P(B) > 0$)
- **General multiplication rule**:
  - $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$
- **Independent events**:
  - To prove A and B are independent, show $P(A|B) = P(A)$ or $P(B|A) = P(B)$.
  - Corollary: $P(A \wedge B) = P(A)P(B)$

2.2. Counting Techniques

- **Permutation**: the number of ways to arrange r objects from n objects.
  - ${}_nP_r = \frac{n!}{(n-r)!}$
- **Combination**: the number of ways to choose r objects from n objects.
  - ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

3. Probability Distributions

- **Random variable**: a variable whose value is determined by the outcome of a random experiment.
  - **Discrete random variable**: counted values
  - **Continuous random variable**: measured values
- **Probability distribution**: a mutually exclusive listing of all possible numerical outcomes and their corresponding probabilities.

3.1. Discrete Probability Distributions

Axioms:

- $0 \leq P(X = x_i) \leq 1$
- $\sum_{i=1}^{N} P(X = x_i) = 1$

Statistics:

- Expected value: $\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i)$
- Variance: $\sigma^2 = Var(X) = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)$

Binomial Distribution:

- $X \sim \mathcal{B}(n, \pi)$
  - where n is the number of trials and π is the probability of success.
  - applies when there are n **independent** trials and each trial has a **constant** probability of success π. There are **two mutually exclusive outcomes** for each trial.
- $P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x}$
- $\mu = E(X) = n\pi$
- $\sigma^2 = Var(X) = n\pi(1 - \pi)$
- When $\pi < 0.5$, the distribution is left-skewed. When $\pi > 0.5$, the distribution is right-skewed.

3.2. Continuous Probability Distributions

- $P(X = x) = 0$
- **Probability density function** $f(x)$: the function that describes the relative likelihood of a continuous random variable taking on a particular value.
  - $f(x) \geq 0$
  - $\int_{-\infty}^{+\infty} f(x) dx = 1$
- **Cumulative distribution function** $F(x)$: the probability that a continuous random variable is less than or equal to a certain value.
  - $F(x) = P(X \leq x)$
  - $F(x) = \int_{-\infty}^{x} f(t) dt$

Normal Distribution:

- $X \sim \mathcal{N}(\mu, \sigma^2)$
  - where μ is the mean and σ^2 is the variance.
- Normal density function: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Mean = Median = Mode = μ
  - Empirical rule: 68-95-99.7 rule
- $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$
- **Standardized normal distribution**: $Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1^2)$
  - $P(x \leq x_0) = P(z \leq \frac{x_0 - \mu}{\sigma})$

4. Sampling Distributions

Sampling distribution: the probability distribution of a sample statistic from all possible samples of a given size.

Sampling distribution of the sample mean:

Given that $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

- Mean $\mu_{\bar{X}} = \mu$
- Standard deviation or **standard error** $SE = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem:

- If the sample size n is large enough, the sampling distribution of the sample mean will be approximately normally distributed.
- Applicable when $n \geq 30$.

5. Confidence Intervals of Population Mean

Confidence interval is a range of values, based on one sample taken from a population, that is likely to contain the true population parameter.

Interpretation: for a $100(1-\alpha)\%$ CI,

- We are $100(1-\alpha)\%$ confident that the true population parameter lies within the interval.
- If all samples of size n are taken, $100(1-\alpha)\%$ of the intervals will contain the true population parameter.

5.1. Z-Distribution

Applicable when the population standard deviation σ is known.

$100(1-\alpha)\%$ CI for μ:

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

where

- $Z_{\alpha/2}$ is the **critical value**.
- $E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is the **margin of error** or **sampling error**.

Factors that affect the width of the CI:

- When σ increases, margin of error increases. $\sigma \uparrow \Rightarrow \frac{\sigma}{\sqrt{n}} \uparrow$
- When n increases, margin of error decreases. $n \uparrow \Rightarrow \frac{\sigma}{\sqrt{n}} \downarrow$
- When level of confidence $1-\alpha$ increases, margin of error increases. $1-\alpha \uparrow \Rightarrow \alpha \downarrow \Rightarrow Z_{\alpha/2} \uparrow$

For a population with known σ, the width of the CI is the same regardless of the actual sample, holding all other factors constant.

1-α	$Z_{\alpha/2}$
0.90	1.645
0.95	1.96
0.99	2.576

$Z_{\alpha/2}$ is defined as the value of Z such that $P(Z \leq z) = \alpha/2$. Do note that since $\alpha/2 < 0.5$, z is negative, but we take the absolute value. Therefore when α decreases, z decreases and $Z_{\alpha/2}$ increases.

As level of confidence $1 - \alpha$ increases, $Z_{\alpha/2}$ increases.

5.2. t-Distribution

For a normal population with unknown σ, the sample standard deviation s is used.

The t-statistic is defined as $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$.

t follows a t-distribution with $n - 1$ degrees of freedom. $t \sim t(n-1)$.

Degrees of freedom:

- Total number of observations minus the number of parameters estimated from the data.
- For calculation of sample variance, the degrees of freedom are $n - 1$.
- For combining multiple t-distributions, the degrees of freedom are the sum of the individual degrees of freedom. (e.g. $n - 1 + m - 1 = n + m - 2$)
- For least squares estimation, the degrees of freedom are $n - 2$.

t-distribution is a bell-shaped curve that is symmetric about the mean $\mu = 0$. It is flatter and has heavier tails than the standard normal distribution. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution.

$t_{\alpha/2, n-1}$ is the critical value of the t-distribution. It is the value of t such that $P(T \leq t) = \alpha/2$.

- $t_{\alpha/2, n-1}$ is larger than $Z_{\alpha/2}$.
- As level of confidence $1 - \alpha$ increases, $t_{\alpha/2, n-1}$ increases.
- As degree of freedom $n - 1$ increases, $t_{\alpha/2, n-1}$ decreases. (Starting from $+\infty$ to the normal distribution)

$100(1 - \alpha)\%$ CI for μ:

$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$

where

- $t_{\alpha/2, n-1}$ is the critical value.
- $E = t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$ is the margin of error.

Factors that affect the width of the CI:

- s is a random variable; it does not affect the width of the CI. Instead, it decides the width.
- When n increases, margin of error decreases.
  - $n \uparrow \Rightarrow \frac{s}{\sqrt{n}} \downarrow$
  - $n \uparrow \Rightarrow t_{\alpha/2, n-1} \downarrow$
- When level of confidence $1 - \alpha$ increases, margin of error increases.

5.3. Sample Size Determination

When determining the sample size n for a given level of confidence $1 - \alpha$ and margin of error $\pm E$:

$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{Z_{\alpha/2}\sigma}{E} \right)^2$

6. Hypothesis Testing of Population Mean

Hypothesis: a statement about a **population parameter** (μ, σ^2, π).

Null hypothesis H_0: always contains the equality sign.

Alternative hypothesis H_1: the complement of the null hypothesis.

Rejection region: the range of values that lead to the rejection of H_0.

- Two-tail test: $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
- Lower-tail test: $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$
- Upper-tail test: $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$

Decision \ Truth	H_0 is True	H_0 is False
DNR H_0	Level of significance $1 - \alpha$	Type II error β
Reject H_0	Type I error α	Power of the test $1 - \beta$

How to reduce Type II error β:

- By increasing α, β decreases.
- By increasing sample size n, β decreases.

6.1. Z-Test

Applicable when the population standard deviation σ is known.

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

- Two-tail test: CV = $\pm Z_{\alpha/2}$
  - If $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$, reject H_0.
  - If $P(z < -|Z|) + P(z > |Z|) < \alpha$, reject H_0.
- Lower-tail test: CV = $-Z_\alpha$
  - If $Z < -Z_\alpha$, reject H_0.
  - If $P(z < Z) < \alpha$, reject H_0.
- Upper-tail test: CV = Z_α
  - If $Z > Z_\alpha$, reject H_0.
  - If $P(z > Z) < \alpha$, reject H_0.

6.2. t-Test

Applicable when the population standard deviation σ is unknown.

t-statistic: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

(p-value approach can only yield a range of values from the t-distribution.)

7. Proportion

Variable of interest: the variable is a two-level categorical variable.

Sample proportion $p = \frac{Y}{n}$, where Y is the number of successes and n is the sample size.

Population proportion π is the probability of success.

By binomial distribution, $p \sim \mathcal{N}(\pi, \frac{\pi(1-\pi)}{n})$

Sampling distribution of the sample proportion:

- $\mu_p = \pi$
- $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

According to Central Limit Theorem, we approximate the sampling distribution of the sample proportion to a normal distribution, if all of the following conditions are satisfied:

- $n \geq 30$
- $n\pi \geq 5$ (or $np \geq 5$)
- $n(1-\pi) \geq 5$ (or $n(1-p) \geq 5$)

Z-statistic: $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

7.1. Confidence Intervals of Population Proportion

Unlike the case of population mean, the standard deviation of p is dependent on π. Therefore, we always estimate the standard deviation of p using the sample proportion p:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

$100(1-\alpha)\%$ CI for π:

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where

- $Z_{\alpha/2}$ is the critical value.
- $E = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ is the margin of error.

Remarks:

- If $p - E < 0$, set $p - E = 0$.
- If $p + E > 1$, set $p + E = 1$.

Factors that affect the width of the CI:

- When n increases, margin of error decreases.
- When level of confidence $1-\alpha$ increases, margin of error increases.
- When $0 < p < 0.5$ and p increases, margin of error increases. When $0.5 < p < 1$ and p increases, margin of error decreases.

Sample size determination:

$$E = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \Rightarrow n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 \pi(1-\pi)$$

If neither π nor p is known (which is usually the case), use $\pi = 0.5$ which yields the largest sample size.

7.2. Hypothesis Testing of Population Proportion

Z-statistic: $Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

8. Simple Linear Regression

8.1. Coefficient of Correlation

Covariance:

- Population covariance: $\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{N}$
- Sample covariance: $S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

Dividing the XY plane to four quadrants:

- If covariance is positive, the points are in the Q1 and Q3 quadrants.
- If covariance is negative, the points are in the Q2 and Q4 quadrants.
- Covariance can only measure linear association.

Coefficient of Correlation:

- Population coefficient of correlation: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$
- Sample coefficient of correlation: $r_{XY} = \frac{S_{XY}}{S_X S_Y}$
- ρ_{XY} (or r_{XY}) ranges from -1 to 1, and has the same sign as the covariance.

8.2. Simple Linear Regression Model

Linear regression model: $E(Y|X=x) = \beta_0 + \beta_1 x$

Discrepancy: $\epsilon_i = Y_i - E(Y_i|X_i) = Y_i - (\beta_0 + \beta_1 X_i)$

True regression line: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Error or residual $e = Y_i - \hat{Y}_i$

Least squares estimation: Minimize SSE

- SSE = Sum of Squares Error: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$
- Sample regression line: $\hat{Y}_i = b_0 + b_1 X_i$
- $b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = r_{XY}\frac{S_Y}{S_X}$ has the same sign as r_{XY}.
- $b_0 = \bar{Y} - b_1\bar{X}$

The linear regression model is only valid within the range of the data.

Coefficient of Determination:

- SST = Sum of Squares Total: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- SSR = Sum of Squares Regression: $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Coefficient of Determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- In simple linear regression, $R^2 = (r_{XY})^2$

8.3. Confidence Intervals and Hypothesis Testing of Slope

Given the population regression line $Y = \beta_0 + \beta_1 X + \epsilon$:

Assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Sampling distribution of the slope b_1:

- $b_1 \sim \mathcal{N}(\beta_1, \sigma^2_{b_1})$
- $E(b_1) = \beta_1$
- $s^2_{b_1} = \frac{s^2_\epsilon}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{SSE/(n-2)}{\sum_{i=1}^n (X_i - \bar{X})^2}$

$100(1-\alpha)\%$ CI for β_1:

$b_1 \pm t_{\alpha/2, n-2} S_{b_1}$

Hypothesis testing for the slope:

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ (to test if linear relationship exists)

Or ≥ 0 or ≤ 0 for one-tail test.

T-statistic: $t = \frac{b_1 - \beta_1}{S_{b_1}}$

- Critical value: reject H_0 if $t < -t_{\alpha/2, n-2}$ or $t > t_{\alpha/2, n-2}$
- P-value: reject H_0 if $P(t < -|t|) + P(t > |t|) < \alpha$