Introduction

Welcome to Linear Statistical Models

Your Lecturer

Module Convenor:	Dr Ting Wei
School:	Computing and Mathematical Sciences
E-mail:	tw258@leicester.ac.uk
Office Hours:	e-mail me for online appointments

Blackboard Site

- Course material and course announcements will be available on Blackboard.
- Lecture slides posted on Bb before the lectures.
- Lecture recordings uploaded after the lectures, generally the same day.
- Send me e-mail if I forget to upload something!
- Consider using discussion board on Bb to ask questions.
- Some references are provided on Bb if you are interested in further reading, but it is not necessary to succeed in this course.

Blackboard site, cont.

In the tab Learning materials you find:

- Folder Lectures: in each red box you find:
  - Lecture slides posted on Bb before each lecture. All material in the slides is examinable.
  - Lecture recording uploaded after the lecture.
  - Lecture annotations in OneNote.

Blackboard site, cont.

- Workbook (WB) Collection of exercises.
- Folder Feedback Classes (FC) Hints for exercises from WB to try before FC.
- Folder Module Delivery Plan updated along the way.
- Statistical tables are given.

Books

- The following are good books for further reading, but they are not necessary to succeed in this course.
  - Peter Dalgaard, Introductory Statistics with R, Springer.
  - W. Mendenhall, R. L. Scheaffer and D. D. Wackerly, Mathematical Statistics with Applications, Duxbury Press, 2008.
- These books are available in the library, the first one online.

Communication

- Interactive activities during the lectures/FCs.
- Course announcements on Bb. It is your responsability to check this regularly.
- Discussion board: Ask questions on discussion board about the module (not about assessed coursework!). We strongly encourage you to answer questions from fellow students if you think you can help (even if you don't have a complete solution).

Communication, cont.

- E-mail me if you need help. Quick questions sometimes can be discussed over e-mail.
- Office hours: e-mail me for an individual online appointment.
- It is your responsibility to make the most out of the opportunities given above.

Assessment

- 30% marked coursework
- 70% written examination

Coursework

- Coursework:
  - Consisting of 3 problem sheets, to be done individually.
  - Hand in coursework electronically (one file per person).
  - See Blackboard home page for deadlines.
- Exam: 4 questions, 25 points each, exam date TBC.

Workbook

- Good to test and consolidate your understanding of the material in the lectures while the course is in progress, and to revise for the exam.
- Solutions will be released by chapter after coursework on that chapter is handed in.
- However, some questions will be discussed in the feedback class, and suggestions given beforehand.
- Summary table to link exercises and solutions (when released) with lecture material as we go along.
- Please **read the guidelines on the first page.**

Workbook table snapshot

Table of exercise numbers versus main section content needed

Exercise Nr.	Solution page	Main section content needed to perform the exercise
1.1	31	*Section 1.1*
1.2	31	*Section 1.1*
1.3	31	*Section 1.1*
1.4	32	*Section 1.1*
1.5	32	*Section 1.2*
1.6	33	*Section 1.2*
1.7	33	*Sections 1.2, 1.3*
1.8	34	*Sections 1.3, 1.4*
1.9	34	*Section 1.4*
1.10	35	*Section 1.5*
1.11	35	*Section 1.5*
2.1	36	*Sections 2.1, 2.2*

Figure: Screenshot from workbook table

Why taking this module

- Lots of different disciplines use linear regression as a basic form of data analysis: e.g. psychology, sociology, medicine, actuarial science, etc.
- Good for employability skills (basic data analysis, statistical software skills).

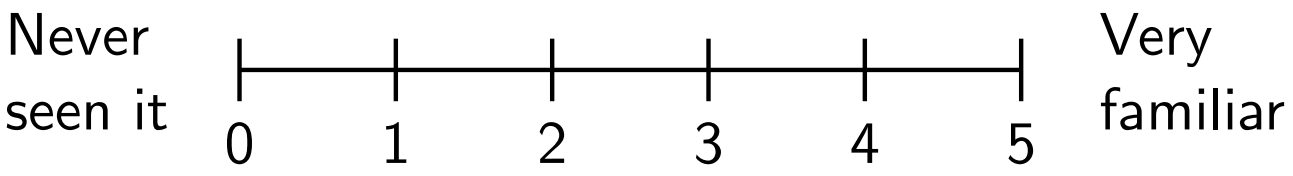
Syllabus

- Topic Outline:
  1. Probability (5 Sections)
  2. Statistical inference (4 Sections)
  3. Simple linear regression (10 Sections)
  4. Multiple linear regression (6 Sections)
  5. One way analysis of variance (3 Sections)
- Detailed syllabus in the course description on Bb.
- Module delivery plan in the folder 'learning materials' on Bb.

How familiar are you with random variables and their distributions ?

Never seen it — 0 — 1 — 2 — 3 — 4 — 5 — Very familiar

How familiar are you with confidence intervals and hypothesis tests?



Chapter 1: Probability

Dr.Ting Wei

Section 1.1: Concepts of probability

Question

- Can you think of an example of a random event?

Sample space

- A sample space is a list of outcomes of a random experiment, normally denoted by Ω.
- Examples:
  a) Tossing a coin: $\Omega = \{H, T\}$.
  b) Rolling a dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  c) Tossing two coins: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$.

Question

- So the sample space is a set.
- For instance, when tossing a coin: $\Omega = \{H, T\}$.
- Lets consider the outcome 'head' that is the set $\{H\}$
- What is the relation between the two sets $\{H\}$ and $\{H, T\}$?

Basic rules of set operations

For any subsets $A, B \subseteq \Omega$,

- Commutative law: $A \cup B = B \cup A$, and $A \cap B = B \cap A$.
- Associative law: $(A \cup B) \cup C = A \cup (B \cup C)$, and $(A \cap B) \cap C = A \cap (B \cap C)$.
- Distributive law: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.
- De Morgan's rules: $(A \cup B)^c = A^c \cap B^c$, and $(A \cap B)^c = A^c \cup B^c$.

Elementary results

- $A = B \Leftrightarrow A \subseteq B$ and $B \subseteq A$,
- $A \cap \emptyset = \emptyset$ and $A \cup \emptyset = A$,
- $A \cap \Omega = A$ and $A \cup \Omega = \Omega$,
- $A \cap A = A \cup A = A$,
- $A \cap B \subseteq A$ and $A \cap B \subseteq B$,
- $A \subseteq B \Rightarrow A \cap B = A$,
- $A \subseteq B^c \Rightarrow A \cap B = \emptyset$,
- $(A^c)^c = A$,
- $A \cup A^c = \Omega$, and $A \cap A^c = \emptyset$.

Events in a sample space

- An event is a subset of a sample space.
- Example: we toss a dice twice.
  $\Omega = \{(1,1),(1,2),(1,3)\ldots,(5,6),(6,6)\}$.

  The event 'the second toss is 6' is
  $A = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6)\}$.

Question

- Suppose you are tossing a dice twice, so $\Omega = \{(1,1),(1,2),(1,3)\ldots,(5,6),(6,6)\}$.
- Can you spot a difference between the pair of events $(1,2)$, $(4,5)$ and $(1,2)$, $(1,3)$?

Disjoint events

- Events $A_1, \cdots A_n$ are disjoint if for all $i \neq j$, $A_i \cap B_j = \emptyset$.
- For instance if we toss a dice twice, the events $(1, 2)$, $(4, 5)$ are disjoint, but the events $(1, 2)$, $(1, 3)$ are not.

Sigma-algebras

- Intuition: given events A, B we want $A \bigcup B$ to be the event 'A or B'. We also want ' A does not happen' to be an event.

- A sigma-algebra on a sample space Ω is a collection $\mathcal{F}$ of subsets of Ω such that

  1) $\Omega \in \mathcal{F}$.

  2) if $A \in \mathcal{F}$, $A^c \in \mathcal{F}$.

  3) If $A_1, A_2, \ldots$ are in $\mathcal{F}$, $\bigcup_{i=1}^{\infty} A_i$ is in $\mathcal{F}$.

Example of a sigma-algebra

- $\{\emptyset, \Omega\}$, is a σ-algebra.
- $\{A, A^c, \emptyset, \Omega\}$ is a σ-algebra.
- The collection of all subsets of Ω is a σ-algebra and called the power set of Ω, which is denoted by $\mathbb{P}(\Omega)$.
- $\mathcal{F} = \mathbb{P}(\Omega)$ is the only σ-algebra we will consider from now on.

Probability spaces

- A probability function is a function $P : \mathcal{F} \to [0, 1]$. This maps each event to its probability of occurrence.

- We also impose the probability axioms:

  1) $P(\emptyset) = 0$.

  2) $P(\Omega) = 1$.

  3) (Countable additivity)

     If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \, .$$

- The triple $(\Omega, \mathcal{F}, P)$ is called a probability space.

Examples of probability spaces

- Fair coin toss $\Omega = \{H, T\}$, $\mathcal{F} = \mathbb{P}(\Omega)$, $P(H) = P(T) = 1/2$, $P(\emptyset) = 0$, $P(\Omega) = 1$.

- Unfair coin toss: choose $0 \leq p \leq 1$ and put $\Omega\{(H, T)\}$, $\mathcal{F} = \mathbb{P}(\Omega)$, and define P by $P(\emptyset) = 0$, $P(\Omega) = 1$ and

$$P(\{H\}) = p, \; P(\{T\}) = 1 - p \; .$$

- 6-sided fair dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathbb{P}(\Omega)$, $P(A) = |A|/6$, where $|A|$ is the size of the subset A.

Basic results on probability

- One can prove the following properties

  Let P be a probability function on a sigma-algebra $\mathcal{F}$ and let $A, B \in \mathcal{F}$. Then:

  i) $P(A^c) = 1 - P(A)$

  ii) If $A \subseteq B$ then $P(A) \leq P(B)$

  iii) $P(B \setminus A) = P(B) - P(B \cap A)$

  iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Independence

- Idea: two events are independent when there is no connection between them.
- For instance, if we toss a coin twice, it is natural to think that the result of the first toss is independent from, or does not affect, the result of the second.
- Two events A and B are called independent if

$$P(A \cap B) = P(A)P(B).$$

Example

- Consider the example of rolling a single 6-sided fair dice. Thus $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathbb{P}(\Omega)$ and $P(\{\omega\}) = 1/6$ for all $\omega \in \Omega$.
- Let $E = \{2, 4, 6\}$ be the event that the roll is even and $T = \{2\}$ be the event that the roll is 2. Is E independent of T?

Mutual independence

- A collection of events are mutually independent if and only if the probability of any intersection is equal to the product of the probabilities of the events included in the intersection.

- Suppose that for each i in some set I we have an event A_i. The family of events $\{A_i : i \in I\}$ is called independent if for any n and any *distinct* elements $i_1, \ldots, i_n$ of I we have

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = \prod_{j=1}^{n} P(A_{i_j}).$$

Conditional probability: motivation

- Motivation: we want to measure the probability of an event A occurring given that another event B occurs.
- Example: In a customer survey we found that 60% of the customers used the brand Favorite (F), while 30% used the brand Super (S). 15% of the customers used both brands.
- How many percent of the Favorite users use Super, and how many percent of the Super users use Favorite?

Conditional probability: definition

- Idea: change the sample space to B and scale all the probabilities accordingly.
- Let A and B be events with $P(B) \neq 0$. The conditional probability of A given B, written $P(A|B)$, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Back to the example

- In a customer survey we found that 60% of the customers used the brand Favorite (F), while 30% used the brand Super (S). 15% of the customers used both brands. How many percent of the Favorite users use Super, and how many percent of the Super users use Favorite?
- We want to find $P(S|F)$ and $P(F|S)$.

Bayes' formula

Figure: Thomas Bayes (c. 1701 – 7 April 1761)

- Let A and B be events with nonzero probability. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of total probability

- Suppose $\mathcal{E} = \{E_1, \ldots, E_n\}$ is a partition of the sample space Ω and A is any event. Then,
- $P(A) = \sum_{i=1}^{n} P(A \cap E_i)$.
- $P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i)$, where the sum is over all i such that $P(E_i) \neq 0$.

Bayes' Theorem

Theorem

Let $\mathcal{E} = \{E_1, \ldots, E_n\}$ be a partition of Ω, and A be an event. Then

$$P(E_j|A) = \frac{P(E_j)P(A|E_j)}{\sum_{i=1}^{n} P(E_i)P(A|E_i)}$$

where $P(E_j) > 0$ for all $1 \leq j \leq n$ and $P(A) > 0$.

Example

- A computer centre has three printers: A, B, and C.
- Documents are routed to printers A, B, and C with probability 0.5, 0.3, and 0.2 respectively.
- Printers A, B, and C, jam with probability 0.04, 0.05, and 0.03, respectively.
- Now you find your program crashes because of printer jamming. What is the probability that printer A is the culprit?

Independence and conditional probability

- If $P(B) \neq 0$, then A and B are independent if and only if $P(A|B) = P(A)$.
- Proof: If A and B are independent,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

If $P(A|B) = P(A)$ then

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B)$$

so by definition A and B are independent.

Summary

- The sample space Ω is the set of all outcomes, and an event is a subset of the sample space.
- A probability space is the triple $(\Omega, \mathcal{F}, P)$ where the probability function $P : \mathcal{F} \to [0, 1]$ satisfies certain axioms.
- Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- Mutual independence of collection of events.
- The conditional probability of A given B is $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Bayes' rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Law of total probability.

Section 1.2: Random variables, Part I

Question

- Can you make up an example of a random variable?

The idea of random variable

- Motivation: we wish to study the numbers which come as a result of random outcomes.
- Example : Let Ω be the sample space of outcomes from two rolls of a dice. Define a function $X(\omega_1, \omega_2) = \omega_1 + \omega_2$ where ω_1 is the result of the first roll and ω_2 is the result of the second roll. Any outcome of the two rolls leads to a real number.
- This example gives a function X on the sample space Ω leading to a real number. We call any such function a random variable.

Definition of random variable

- A random variable X is a function defined on the sample space Ω which returns a well defined real number from any outcome.
- The rigorous definition requires that the above function is 'measurable'. All the functions in this course will be, so we do not elaborate further.

Distribution of a random variable: motivation

- Consider a random variable $X : \Omega \to \mathbb{R}$. Say we want to know the probability that X takes a value less than x. In term of events we have

$$\{\omega : X(\omega) \leq x\} \Leftrightarrow \omega \in X^{-1}((-\infty, x])$$

where X^{-1} is the inverse of the function X.

- The probability of this event is denoted more briefly by $P(X \leq x)$.

Distribution of a random variable

- Let X be a random variable. The cumulative distribution function (CDF) or distribution function of X is the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$F_X(x) = P(X \leq x).$$

PMF of discrete random variables

- Let X be a random variable. We say X is discrete if it is defined on a discrete sample space Ω with values in $\mathbb{R}$.
- Let X be a discrete random variable. The probability mass function or PMF or probability function of X is the function $p_X(x_i) = P(X = x_i)$, $i = 1, 2, \ldots, n$.
- Obviously, $p_X(x_i) \geqslant 0$ and $\sum_{i=1}^{n} p_X(x_i) = 1$.

Infinite but discrete

- A discrete random variable can be infinite but discrete.
- Example: A coin is tossed until a head occurs. Can you draw the sample space and define a random variable on that?

Discussion questions

- Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and

  $P(X = \omega_1) = 0.2$, $P(X = \omega_2) = 0.1$, $P(X = \omega_3) = 0.7$.

  $X(\omega_1) = 1$, $X(\omega_2) = 2$, $X(\omega_3) = 3$.

- Can you draw the CDF ?

- Recall the CDF is $F_X(x) = P(X \leq x)$ and the PMF is $p_X(x_i) = P(X = x_i)$, $i = 1, 2, \ldots, n$.

- What is the relation between CDF and PMF for a discrete random variable X taking values x_i, $i \in \mathbb{N}$?

Expectation of a discrete random variable

- Idea: the expected value of a discrete random variable X is a weighted average of the values that X takes, each value is weighted according to its probability.

- Let X be a discrete random variable with a finite list $\{x_1, ..., x_n\}$ of possible values. The expectation of X is defined by

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i).$$

Discrete uniform distribution

- A discrete random variable X with n possible outcomes $x_1, \ldots, x_n$ is said to have a discrete uniform distribution if the PMF of X is given by

$$P(X = x_i) = \frac{1}{n} \qquad i = 1, \ldots, n\,.$$

- Example: if we roll a fair dice, the outcomes $1, \ldots, 6$ have equal probability of occurring, hence the random variable X 'number of dots observed on the upper face of the dice' has a uniform discrete distribution with PMF $P(X = i) = \dfrac{1}{6}$ $i = 1, \ldots, 6$.
- $E(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Bernoulli random variables

- The Bernoulli distribution is used to model an experiment with only two possible outcomes, often referred to as 'success' and 'failure', or 0 and 1.

- A discrete random variable X has a Bernoulli distribution with parameter $0 \leq p \leq 1$ if its probability mass function is

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

- We write $X \sim \mathrm{Ber}(p)$ if X is a Bernoulli random variable with parameter p, and $E(X) = p$.

Binomial random variables

- A random variable with PMF

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  $k = 0, 1, \ldots, n$ is called a binomial random variable with parameters n and p.

- We write $X \sim \mathrm{Bin}(n, p)$ if X is a binomial random variable with parameters n and p, and $E(X) = np$.

Example

- Suppose we have a coin which has probability p of showing heads. We make n independent tosses of the coin, recording the result, and count the number of heads obtained in the n tosses.
- This is modelled by a binomial random variable with parameters n and p.

Poisson random variable

- Let $\lambda > 0$. A random variable with PMF

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

  for $k = 0, 1, 2, \ldots$ is called a Poisson random variable with parameter λ.

- We write $X \sim \text{Pois}(\lambda)$ if X is a Poisson random variable with parameter λ, and $E(X) = \lambda$.

- Under certain independence assumptions it can be shown that the number of events happening in time t is a Poisson random variable, e.g. for the number of road accidents, industrial accidents in a given unit of time. Also, the number of calls received by a switchboard.

Example

- The number of industrial accidents at a particular manufacturing plant is found to average three per month, and it is likely to follow a Poisson distribution.

Geometric random variables

- A random variable with PMF

$$P(X = k) = (1 - p)^{k-1} p$$

  $k = 1, \ldots, n$ is called a geometric random variable with parameters p.

- We write $X \sim \text{Geom}(p)$ if X is a geometric random variable with parameters p, and $E(X) = \frac{1}{p}$.

- Recall the example: A coin is tossed until a head occurs.

Summary

- Notion of random variable $X : \Omega \to \mathbb{R}$.
- Cumulative distribution function (CDF) $F_X(x) = P(X \leq x)$.
- Probability mass function (PMF) of a discrete random variable.
- Examples of discrete random variables: Uniform, Bernoulli, Binomial, Poisson, Geometric.

Section 1.3: Random variables, Part II

PDF of continuous random variables

- A continuous random variable is a random variable X such that the CDF $F_X(x)$ satisfies

$$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} f_X(y)dy$$

  for some function $f_X : \mathbb{R} \to \mathbb{R}$.

- This function $f_X(x)$ is a non-negative integrable function called the probability density function or density function or PDF of X.

- By fundamental theorem of calculus: $\dfrac{dF_X(x)}{dx} = f_X(x)$.

- For all $x \in \mathbb{R}$, $f_X(x) \geqslant 0$.

PDF of continuous random variables, cont.

- It follows that if X is a continuous random variable with PDF $f_X(x)$ we have

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$

for any a, b with $a \leq b$.

- and

$$\int_{-\infty}^{\infty} f_X(x)dx = 1.$$

True or False?

- From previous slide:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

- Is it true or false that $P(X = a) = f_X(a)$?

Expectation of a continuous random variable

- Let X be a continuous random variable with PDF $f(x)$.
- We define the expected value $E(X)$ as

$$E(X) = \int_{-\infty}^{\infty} x\, f(x) dx \ .$$

Continuous uniform random variable

- If X is a continuous variable with values in the interval $[a, b]$, with density function

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise} \end{cases}$$

  Then X is called a continuous uniform random variable on the interval $[a, b]$.

- We write $X \sim \mathrm{U}(a, b)$ if X is a continuous uniform random variable with parameters a and b, and $E[X] = \frac{a+b}{2}$.

- The cumulative distribution function is

$$F_X(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

Example of continuous uniform random variable

- Suppose a train arrives at a subway station regularly every 10 minutes.
- If a passenger arrives at the station without knowing the timetable, then the waiting time to catch the train is uniformly distributed.

Exponential random variable

- The positive random variable X with density

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

  for some positive parameter λ is called an exponential random variable.

- We write $X \sim \mathrm{Exp}(\lambda)$ if X is an exponential random variable with parameters λ, and $E[X] = \frac{1}{\lambda}$.

- The cumulative distribution function is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Example of exponential random variable

- The lifetime of a certain type of car batteries follows an exponential distribution. It is known that the average lifetime is 5 years. What's the probability that a particular battery of this type lasts for more than 7 years?

Normal (Gaussian) random variables

- A continuous random variable X is said to have a normal (Gaussian) distribution if its density is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  for some μ and positive $\sigma > 0$.

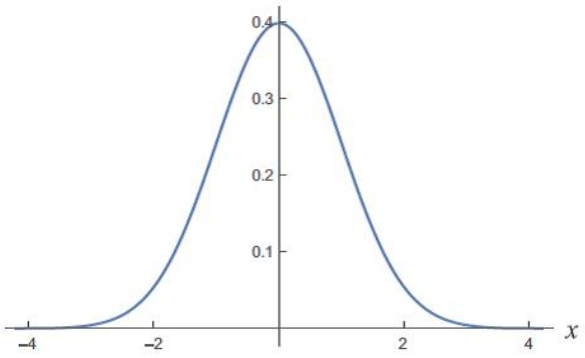
- We write $X \sim N(\mu, \sigma^2)$ if X is a normal random variable with parameters μ and σ, and $E[X] = \mu$.

Normal distribution

- The standard normal random variable has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The graph of PDF is shown here for $\mu = 0$ and $\sigma = 1$ (in this case is $N(0, 1)$).



Normal, cont.

- Thus

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

  The probability that X is between a and b is hence equal to the area under the density function between a and b.

- It can be shown that

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

Normal, cont.

- The CDF of the normal random variable X is

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

There is no explicit formula to solve this integral, it can only be solved by numerical methods.

Find probabilities of standard normal

- We usually use $\Phi(x)$ to denote the CDF of a standard normal random variable,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$$

- It can be shown that $\Phi(1.96) \approx 0.975$. Thus if $X \sim N(0,1)$, $P(X \leq 1.96) \approx 0.975$.
- $P(a \leq X \leq b) = \Phi(b) - \Phi(a)$.
- $\Phi(x) = 1 - \Phi(-x)$.
- Find $\Phi(x)$ using standard normal table or programming.

```
prob <- pnorm(x,mu,sd)
```

Question

- Let $X \sim N(\mu, \sigma^2)$. How to find $P(X \leqslant x)$ from the statistical table?

Answer

- If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$.
- **Example**: Assume $X \sim N(10, 4)$, find $P(8 < X < 11)$.

Variance

- The variance $\mathrm{var}(X)$ of a random variable X is defined by

$$\mathrm{var}(X) = E\left[(X - E[X])^2\right]$$

- The variance is a measure of the deviation from the expected value.

- The standard deviation $\sigma(X)$ of a random variable X is defined as $\sigma(X) = \sqrt{\mathrm{var}(X)}$.

Properties of expectation and variance

Let X, Y be random variables and $a, b, c \in \mathbb{R}$. Then

- $E(aX + bY + c) = aE(X) + bE(Y) + c$.
- $\text{var}(aX + bY + c) = a^2 \,\text{var}(X) + 2ab \,\text{cov}(X, Y) + b^2 \,\text{var}(Y)$ where $\text{cov}(X, Y)$ is the covariance between X and Y .
- We will return to the covariance $\text{cov}(X, Y)$ later. For now just note that the variance is not linear.
- $\text{var}(X) = E(X^2) - E(X)^2$.

Summaries of expectation and variance

- If X follows a discrete uniform distribution, $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2$.
- $X \sim \text{Ber}(p)$, $E(X) = p$ and $\text{var}(X) = p(1-p)$.
- $X \sim \text{Bin}(n, p)$, $E(X) = np$ and $\text{var}(X) = np(1-p)$
- $X \sim \text{Pois}(\lambda)$, $E(X) = \lambda$ and $\text{var}(X) = \lambda$
- $X \sim \text{Geom}(p)$, $E(X) = \frac{1}{p}$ and $\text{var}(X) = \frac{1-p}{p^2}$

Summaries of expectation and variance

- $X \sim \mathrm{U}(a, b)$, $E(X) = \frac{a+b}{2}$ and $\mathrm{var}(X) = \frac{(b-a)^2}{12}$
- $X \sim \mathrm{Exp}(\lambda)$, $E(X) = \frac{1}{\lambda}$ and $\mathrm{var}(X) = \frac{1}{\lambda^2}$
- $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$

Functions of random variables

- Motivating question: given a (reasonably nice) function f and a random variable X whose distribution is known, what is the distribution of $f(X)$?

- Suppose g is differentiable and strictly increasing or strictly decreasing and X is a continuous random variable with density function f_X. Then the density function of $Y = g(X)$ is

$$f_Y(y) = \left| \frac{dg^{-1}}{dy} \right| f_X(g^{-1}(y))$$

Expectation of functions of random variables

- Let X be a discrete random variable and g be a function such that $E[g(X)]$ exists. Then

$$E[g(X)] = \sum_i g(x_i)P(X = x_i) \ .$$

- Let X be a continuous random variable with density function f_X. Suppose g is a function and $E[X]$, $E[g(X)]$ exist. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \ .$$

Example

- A process for refining sugar yields up to 1 ton of pure sugar per day, but the actual amount produced X is a random variable because of machine breakdowns. Suppose that X has density function given by

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- The company is paid at the rate of £ 300 per ton for the refined sugar, but it has also has a fixed overhead cost of £ 100 per day. Thus the daily profit, in hundreds of pounds, is $Y = 3X - 1$.
- Find the probability density function of Y.

Summary

- Probability density function (PDF) of continuous random variables.
- Examples of continuous random variables: Uniform, Exponential, Normal.
- Expectation and variance.
- Functions of random variables.

Section 1.4: Independence and joint distributions

Jointly distributed random variables

- Idea: when we study two random variables X and Y, we first make a list of all possible values X and Y can have. Then we consider all the possible combinations that can occur, and figure out the probability of such combinations.

- Two random variables on the same sample space are called jointly distributed.

Joint cumulative distribution function

- Let X and Y be jointly distributed random variables. The joint CDF $F_{X,Y}$ of X and Y is the function

$$F_{X,Y}(x, y) = P(X \leq x \text{ and } Y \leq y).$$

- This is the probability of the event

$$X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, y]) \ .$$

Independent random variables

- Imposing the condition that these two events are independent we find

$$F_{X,Y}(x,y) = P(X^{-1}(-\infty,x] \cap Y^{-1}(-\infty,y]) =$$
$$= P(X^{-1}(-\infty,x])P(Y^{-1}(-\infty,y]) = F_X(x)F_Y(y)\ .$$

- In this case, the jointly distributed random variables X and Y are called independent

Overview

Given a random variables X and Y,

CDF $P(X \leq x)$	
DISCRETE	**CONTINOUS**
PMF $P(X = x)$	Density $f(x)$

Joint CDF $P(X \leq x \text{ and } Y \leq y)$	
DISCRETE	**CONTINOUS**
Joint PMF	Density $f(x, y)$

Can you guess what goes in place of the dots ?

Jointly distributed discrete random variables

- Discrete random variables on the same sample space are called jointly discrete if they are both discrete.
- Let X and Y be jointly distributed discrete random variables. The joint probability mass function or joint PMF of X and Y is

$$p_{X,Y}(x, y) = P(X = x \text{ and } Y = y).$$

Example

- In a town there are two different firms with about the same number of customers. Let X denote which firm is chosen by a randomly selected customer, and Y the number of days it takes to process an order.
- The joint PMF is

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.1	0.1	0.3
$X = 2$	0.2	0.1	0.2

Marginal distributions: discrete case

- Let X, Y be jointly distributed discrete random variables with joint PMF $p_{X,Y}$.
- The marginal probability mass function of X is

$$p_X(x_i) = P(X = x_i) = \sum_j P(X = x_i \text{ and } Y = y_j).$$

- Fact: X and Y are independent if and only if $p_{X,Y}(x, y) = p_x(x)p_Y(y)$.

Example and Question

- In the previous example the marginals are

	$Y = 1$	$Y = 2$	$Y = 3$	$P(X = x)$
$X = 1$	0.1	0.1	0.3	0.5
$X = 2$	0.2	0.1	0.2	0.5
$P(Y = y)$	0.3	0.2	0.5	

Question: are X and Y independent?

Jointly distributed continuous random variables

- Random variables X and Y on the same sample space are called jointly continuous if there is a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that

$$F_{X,Y}(u,v) = \int_{-\infty}^{u} \int_{-\infty}^{v} f_{X,Y}(x,y)dxdy.$$

  In this case we call $f_{X,Y}$ the joint density function or joint PDF of X and Y.

- We can recover the joint PDF from the joint CDF by differentiation: at points at which $f_{X,Y}$ is continuous we have

$$f_{X,Y} = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}.$$

Marginal distributions: continuous case

- Let X, Y be jointly continuous random variables with joint density function $f_{X,Y}$.
- The marginal density function of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy.$$

- Fact: jointly continuous random variables are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.

Example. The bivariate normal distribution

- Two random variables X and Y are said to have a bivariate normal distribution if their joint probability density function is given by

$$f_{X,Y}(x,y) =$$

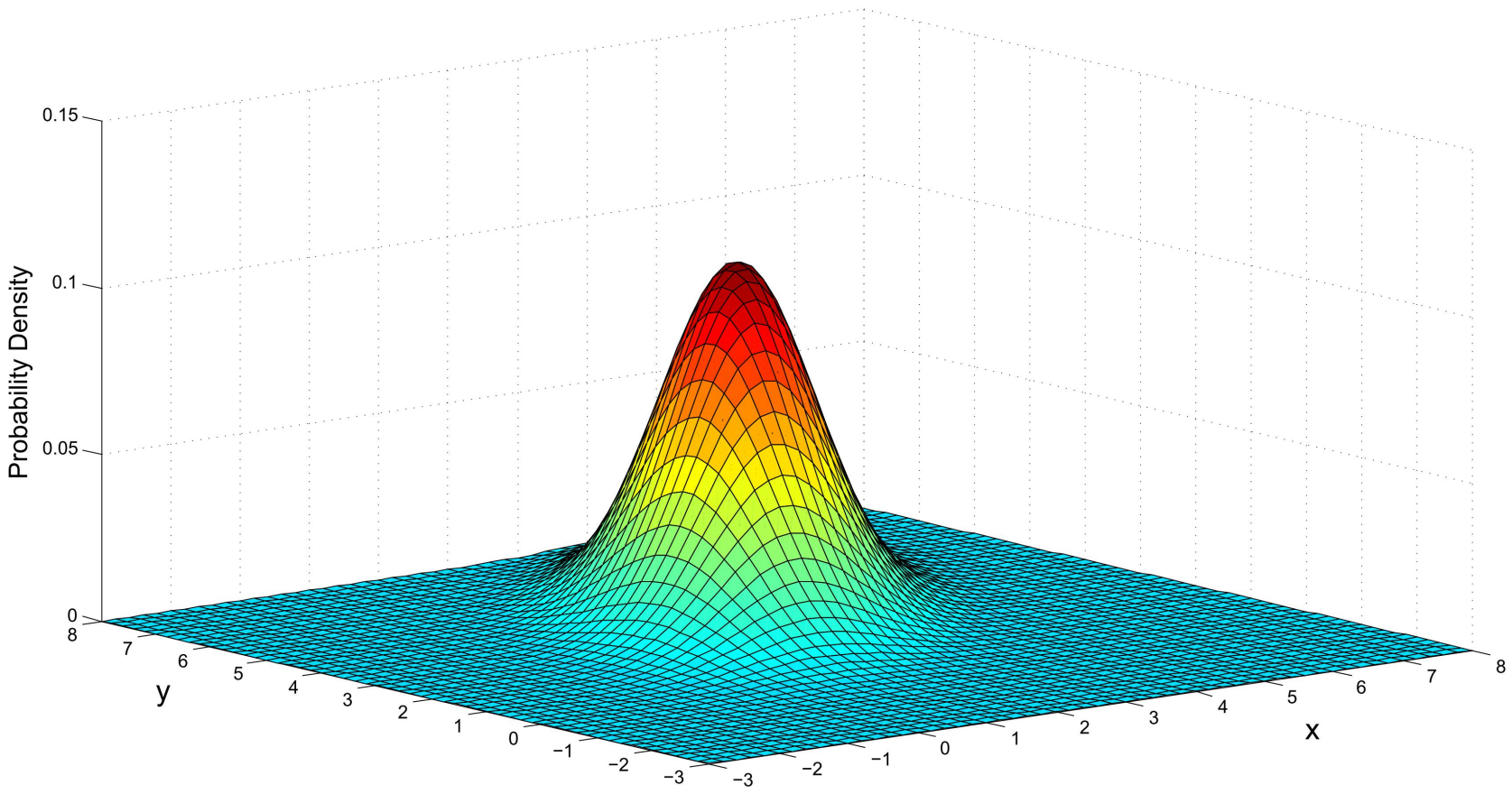
$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \right.\right.$$

$$\left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right]$$

for $-\infty < x, y < \infty$, where ρ is the correlation between X and Y.

- There are five parameters: μ_x, μ_y, σ_x, σ_y, ρ.

The bivariate normal distribution



A vertical cross-section through the surface at $x = x_0$ or at $y = y_0$ is a normal density curve.

Example

- Let X, Y have the bivariate normal distribution with joint PDF as on slide of previous example

$$f_{X,Y}(x,y) =$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \right.\right.$$

$$\left.\left. + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right]$$

- The marginal distributions of X and Y are normal with means μ_X, μ_Y and variances σ_X^2 and σ_Y^2.

- If $\rho = 0$ then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, so X and Y are independent.

Conditional distributions: discrete case

Let X, Y be discrete random variables.

- The conditional probability mass function $p_{X|Y}$ is

$$p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)}.$$

  This function only makes sense for y such that $P(Y = y) \neq 0$, since otherwise the conditional probability isn't defined.

- The conditional expectation of X given Y is

$$E(X|Y = y) = \sum_i x_i p_{X|Y}(x_i|y).$$

Example

- In the previous example:

	$Y=1$	$Y=2$	$Y=3$	$P(X=x)$
$X=1$	0.1	0.1	0.3	0.5
$X=2$	0.2	0.1	0.2	0.5
$P(Y=y)$	0.3	0.2	0.5	

- Find $p_{X|Y}(1|1)$, $p_{X|Y}(1|2)$, $p_{X|Y}(1|3)$, $p_{X|Y}(2|1)$, $p_{X|Y}(2|2)$, $p_{X|Y}(2|3)$.

Conditional distributions: continuous case

Let X, Y be continuous random variables.

- The conditional density function is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

- The conditional expectation of X given Y is

$$E(X|Y=y) = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)dx.$$

Summary

- Jointly distributed random variables and their joint CDF.
- Independent random variables.
- Joint PMF of jointly discrete random variables,
- Joint PDF of jointly continuous random variables.
- Marginal distributions.
- Conditional distributions and expectations.

Section 1.5: Random sampling

Question

- We obtain a dataset of ten elements by tossing a fair coin ten times and recording the result of each toss.
- Can you model this experiment by a sequence of random variables?
- If so, what are they and which characteristics do they have?

Random sample: the idea

- Often the elements of a dataset are repeated measurements of the same quantity.
- We interpret the outcome of an experiment as a realization of some random variables.
- If the measurements are obtained under the same experimental conditions, it is justified to assume that the underlying probability distributions are the same.
- If the successive measurements do not influence each others, we can assume that the random variables are mutually independent.

Random sample definition

- A random sample is a collection of random variables $X_1, \ldots, X_n$ that have identical probability distributions and are mutually independent.

Sample statistic: the idea

- Given a dataset $x_1, \ldots, x_n$, many empirical summaries can be written, such as the mean $(x_1 + \cdots + x_n)/n$.
- Such summaries are functions $h(x_1, \ldots, x_n)$.
- Since datasets are modelled as realization of a random sample $X_1, \ldots, X_n$, an object $h(x_1, \ldots, x_n)$ is a realization of the corresponding random variable $h(X_1, \ldots, X_n)$.

Example

- Consider the random sample associated to rolling a fair dice 3 times.
- Suppose in an experiment we obtain the dataset $\{2, 6, 4\}$.
- We are interested in the mean of the first two rolls.
- What is the numerical summary and the corresponding statistics?

Sample statistic: definition

- A statistic is a function of the sample random variables.
- Thus a statistic is itself a random variable. The distribution of a statistic is called its sampling distribution.
- If a statistical model adequately describes the dataset at hand, then the sample statistics corresponding to the empirical summaries should reflect the corresponding features of the model distribution.

Example

1. The sample mean is $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

2. The sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Sample mean from normal population

- Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$.
- The expectation of the sample mean is

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

- Since X_i are independent,

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Sample variance from a normal random sample

- Recall the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- It can be proved that the following important relationship holds:

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

- This can be considered as an example of a χ^2 random variable.

Connecting the t, F and normal distributions

- Let $X_1, \ldots, X_n$ be n independent and identically normally distributed variables with distribution $N(\mu_X, \sigma_X^2)$.
- Let $Y_1, \ldots, Y_m$ be m independent and identically normally distributed variables with distribution $N(\mu_Y, \sigma_Y^2)$. Suppose each X_i is independent of each Y_j.
- Then the following relationships hold
  - $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$.
  - $\dfrac{\bar{X} - \mu_X}{S_X/\sqrt{n}} \sim t_{n-1}$.
  - $\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$.

Convergence in distribution

- Suppose we have a sequence of random variables $X_1, X_2, \ldots$ with CDFs F_{X_1}, F_{X_1}, and X is a random variable with CDF F_X.

- We say the X_i converge in distribution to X as $n \to \infty$ and write

$$X_n \xrightarrow{D} X \quad \text{as} \quad n \to \infty$$

if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$.

Central limit theorem (CLT)

- Central Limit Theorem (CLT): Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ and variance σ^2. Let $\bar{X}_n$ be the sample mean of $X_1, \ldots, X_n$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \to \infty \,.$$

- This is useful because of its generality: it implies that the sample mean of an independent sample from any distribution with finite mean and variance has approximately the $N(\mu, \sigma^2/n)$ distribution for large n.

Central limit theorem (CLT), cont.

- Let $S_n = X_1 + \cdots + X_n$, so $\bar{X}_n = S_n/n$.
- Thus $E(S_n) = n\mu, \quad \text{var}(S_n) = n\sigma^2$.
- We note that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - E(S_n)}{\sqrt{\text{var}(S_n)}}.$$

- Thus by the CLT

$$\frac{S_n - E(S_n)}{\sqrt{\text{var}(S_n)}} \xrightarrow{D} N(0,1).$$

Example

- Assume that 400 customers come to a shop every day and that their purchases are independent. They often buy milk, and we assume that the number of liters bought by each customer is a random variable X with distribution

$$P(X = 0) = 0.3 \quad , P(X = 1) = 0.5, \quad P(X = 2) = 0.2 \ .$$

- Find an approximate value for the probability that the shop sells between 341 and 390 liters.

Example: solution (fill gaps)

- We call the total number of liters sold $S =$ _____

  We need to compute

$$P(341 \leq S \leq 390) = P(_____) - P(_____)\ .$$

- We can find approximate values for these probabilities from the central limit theorem, but we need to know $E(S)$ and $\text{var}(S)$.
- We first compute $E(X)$ and $\text{var}(X)$:

$$E(X) = _____$$
$$E(X^2) = _____$$
$$\text{var}(X) = _____$$

- Therefore

$$E(S) = _____$$

Example: solution, cont.

- Since $X_1, \ldots, X_{400}$ are independent we have

  $\mathrm{var}(S) =$ _____ hence $\sigma =$ _____ .

- As $n = 400$ is large, by the CLT,

  $$\frac{S - E(S)}{\sqrt{\mathrm{var}(S)}} \sim N(0,1).$$

- We calculate

  $$P(S \leq 390) - P(S \leq 340) \approx \Phi\,(_____) - \Phi\,(_____) =$$
  $$= _____$$

  The probability that the customers buy between 341 and 390 liters of milk is hence about 91%.

Normal approximations

- Let $B \sim \text{Bin}(n, p)$. Then

$$\frac{B - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$$

  as $n \to \infty$.

- Note that a binomial distribution can always be viewed as a sum of n independent Bernoulli distributions, and the central limit theorem can be applied directly to the distribution.

Continuity correction

- Continuity correction should be applied for normal approximations of binomial probabilities.
- Let $B \sim \text{Bin}(n, p)$ and $Y \sim N(np, np(1-p))$, continuity correction states
  - $P(B \leq x) \approx P(Y \leq x + 0.5)$
  - $P(B < x) \approx P(Y \leq x - 0.5)$
  - $P(B > x) \approx P(Y \geq x + 0.5)$
  - $P(B \geq x) \approx P(Y \geq x - 0.5)$
- For n sufficiently large, the following result holds

$$P(B \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Example (fill gaps)

- A fair coin is tossed 1000 times. Find the probability P that heads occur exactly 531 times.
- Let B be the random variable counting the number of heads, so $B \sim \text{Bin}(1000, 1/2)$. We then have an exact expression for this probability:

$$P(B = 531) = _____$$

- Unfortunately $\binom{1000}{531}$ is nearly 10^{300}, too large to deal with directly on an ordinary calculator. Instead (from the context of the rest of the question) we should try to use a normal approximation.

Example, cont.

- The normal approximation says that $P(B \leq x) \approx P(Y \leq x + 0.5)$ where $Y \sim N(500, 250)$. Thus the probability we want is

$$\begin{aligned} P(B = 531) &= P(B \leq 531) - P(B \leq 530) \\ &\approx P(\underline{\qquad}) - P(\underline{\qquad}) \\ &= \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \end{aligned}$$

Summary

- Random sample.
- Statistic and sampling distributions.
- Samples from a normal population.
- Connecting the t, F and normal distributions.
- Central limit theorem and normal approximations.

Chapter 2: Statistical inference

Dr.Ting Wei

Section 2.1: Estimation

General setup

- We saw that a dataset can be modelled as a realization of a random sample from a probability distribution.
- Quantities of interest correspond to features of the model distribution, more precisely to one of the parameters of the model distribution or to a function of the parameters.
- Our task is to use the dataset to estimate the quantity of interest.

Model parameters: an example

- Consider arrivals of packages at a network server. We want to know the number of arrivals during one minute.
- Model this by a random variable with Poisson distribution with (unknown) parameter λ.
- Observe the actual process and get a dataset $x_1, \ldots, x_n$ with $x_i =$ number of arrivals in one minute.
- We want to use this dataset to estimate λ.

Model parameters

- The parameters that determine the model distribution are called model parameters.
- We are interested in estimating model parameters or a function of them.
- For instance, in the above example $e^{-\lambda}$ is the probability that no package arrives in one minute. We may want to estimate this as well.

Estimates

- Let $X_1, \ldots, X_n$ be a random sample where X_i follows a distribution with parameter θ. Let $x_1, \ldots, x_n$ be an observation of the random sample.
- An estimate of parameter θ is a value that only depends on the dataset, that is a function $\hat{\theta}(x_1, \ldots, x_n)$.
- Our computation should give us an indication of the true value of the parameter of interest.

Estimators

- Since our dataset $x_1, \ldots, x_n$ is modelled as a realization of a random sample $X_1, \ldots, X_n$ the estimate $\hat{\theta}(x_1, \ldots, x_n)$ is the realization of the random variable $\hat{\theta}(X_1, \ldots, X_n)$.
- The random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called an estimator of parameter θ.
- Do not confuse the estimate with the estimator! The estimate is a number, the estimator is a random variable.

Sampling distributions

- Let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator of parameter θ based on a random sample $X_1, \ldots, X_n$.
- The probability distribution of $\hat{\theta}(X_1, \ldots, X_n)$ is called the sampling distribution.

Example

- In the previous example, we can use the sample mean $\overline{x}_n = \frac{1}{n}(x_1 + \cdots + x_n)$ as a natural estimate for λ.
- Thus an estimate for $e^{-\lambda}$ is $e^{-\overline{x}_n}$.
- The corresponding estimator of λ is the random variable $\hat{\lambda} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- The corresponding estimator of $e^{-\lambda}$ is the random variable $e^{-\hat{\lambda}} = e^{-\frac{1}{n}(X_1 + \cdots + X_n)}$.

Example, cont.

- Fact: the sum of n independent $Pois(\lambda)$ random variables has a $Pois(n\lambda)$ distribution.
- Thus the sampling distribution of $\hat{\lambda}$ in the previous example is

$$P\left(\hat{\lambda} = -\frac{k}{n}\right) = P\left(e^{-\hat{\lambda}} = e^{-\frac{k}{n}}\right)$$

$$= P(X_1 + \cdots + X_n = k) = \frac{(n\lambda)^k e^{-n\lambda}}{k!}$$

Unbiased estimators

- An estimator $\hat{\theta}$ is called an unbiased estimator for the parameter θ if $E(\hat{\theta}) = \theta$.
- The difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.
- If this difference is non-zero, then $\hat{\theta}$ is called a biased estimator.

Estimators of mean and variance

- Let $X_1, \ldots, X_n$ be a random sample with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$.
- We saw in Section 1.5 that $E(\bar{X}) = \mu$ where $\bar{X}$ is the sample mean. Thus the sample mean is an unbiased estimator of μ.
- Recall from Section 1.5 the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.
- It can be shown that the sample variance S^2 is an unbiased estimator of σ^2.

Example

- Let X and Y be two random variables with $E[X] = \mu$, $E[Y] = 2\mu$. Show that for any value of a constant c, the variable

$$Z = (1 - 2c)X + cY$$

is an unbiased estimator of μ.

Example

- Let X be the production at a department randomly selected in a day. We observe
  $X_1 = 210,\ X_2 = 220,\ X_3 = 210,\ X_4 = 225,\ X_5 = 220,\ X_6 = 217.$
  Find unbiased estimates for $\mu = E[X]$ and $\sigma^2 = \text{var}[X]$.

Summary

- Estimate and estimators.
- Sampling distributions.
- Unbiased estimators.
- Estimators of means and variance.

Section 2.2: Maximum Likelihood Estimation

Recap: estimates and estimators

- Let the dataset $x_1, \ldots, x_n$ be modelled as a realization of a random sample $X_1, \ldots, X_n$.
- An estimate is a value that only depends on the dataset, that is a function $\hat{\theta}(x_1, \ldots, x_n)$.
- The estimate $\hat{\theta}(x_1, \ldots, x_n)$ is the realization of the random variable $\hat{\theta}(X_1, \ldots, X_n)$.
- The random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called an estimator.
- Do not confuse the estimate with the estimator! The estimate is a ....., the estimator is .....

Fill in the gaps

The need for a general principle

- Sometimes it is easy construct estimators for parameters of interest because these parameters had a natural analogue, such as expectation versus sample mean.
- However, in many situations such an analogue does not exist.
- We therefore need a general principle to construct estimators.

The maximum likelihood principle: the idea

- The idea of the maximum likelihood principle is that, given a dataset, we choose the parameters of interest in such a way that the data are most likely.
- We need a precise mathematical method to formalize this idea.

The likelihood function

- Let $x_1, \cdots x_n$ be a dataset which is the realization from a random sample $X_1, \cdots X_n$.
- The likelihood function $\mathcal{L}(\theta)$ is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} P(X_i = x_i|\theta)$$

if the X_i are discrete, and

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

if the X_i are continuous.

- Here $P(X_i = x_i|\theta)$ denotes the probability X_i takes the value x_i if the parameter is θ, and $f_{X_i}(x_i|\theta)$ is the density function of X_i if the parameter is θ.

Maximum likelihood estimates (MLE)

- The maximum likelihood estimates (MLE) of the parameter θ is the value $\hat{\theta}(x_1, \ldots, x_n)$ that maximizes the likelihood function $\mathcal{L}(\theta)$.
- The corresponding random variable $\hat{\theta}(X_1, \ldots, X_n)$ is called the maximum likelihood estimator for θ.

Example

- Suppose we have a dataset $x_1, \ldots, x_n$ modelled as a realization of a random sample from an exponential distribution with probability density function

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- The likelihood function is

$$\mathcal{L}(\lambda) = \ldots\ldots\ldots\ldots$$

Fill in the gaps

Example, cont.

- $\mathcal{L}(\lambda) = \lambda^n e^{-\lambda(x_1+\cdots+x_n)}$.
- To find the MLE of λ we compute the derivative

$$\frac{d\mathcal{L}}{d\lambda} = n\lambda^{n-1}e^{-\lambda\sum_{i=1}^n x_i} - \lambda^n\left(\sum_{i=1}^n x_i\right)e^{-\lambda\sum_{i=1}^n x_i} =$$

$$= n\left(\lambda^{n-1}e^{-\lambda\sum_{i=1}^n x_i}\right)\left(1 - \frac{\lambda}{n}\sum_{i=1}^n x_i\right)$$

- Thus $\dfrac{d\mathcal{L}}{d\lambda} = 0$ if and only if .....

Fill in the gaps

Example, cont.

- Thus $\dfrac{d\mathcal{L}}{d\lambda} = 0$ iff $1 - \lambda\overline{x}_n = 0$
- The MLE of λ is ..... while the maximum likelihood estimator is .....
- Checking: As $\dfrac{d^2\mathcal{L}}{d\lambda^2} = -\overline{x}_n < 0$, the MLE of λ is the maximal of $\mathcal{L}(\lambda)$.

Fill in the gaps

Finding the maximum of $\mathcal{L}(\theta)$

- In the previous example it was easy to find the value of the parameter for which the likelihood is maximal.
- However, in general computing the derivative $\dfrac{d\mathcal{L}}{d\theta}$ can be tedious because $\mathcal{L}(\theta)$ is a product of terms, so one needs to use the product rule for differentiation.

The log likelihood

- We thus introduce the log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

  in which products are turned into sums, which are easier to differentiate.

- Since 'log' is a positively increasing monotone function, $\mathcal{L}(\hat{\theta})$ is maximal if and only if $\ell(\hat{\theta})$ is maximal.

MLE for normal distribution

- Suppose that the dataset $x_1, \ldots, x_n$ is a realization from a $N(\mu, \sigma^2)$ distribution.
- In this case θ is the vector (μ, σ) and the likelihood function is a function of two variables

$$\mathcal{L}(\mu, \sigma) = f_{\mu,\sigma}(x_1) f_{\mu,\sigma}(x_2) \cdots f_{\mu,\sigma}(x_n)$$

where

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

MLE for normal distribution, cont.

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log \prod_{i=1}^{n} f_{\mu,\sigma}(x_i) =$$

$$= \sum_{i=1}^{n} \log f_{\mu,\sigma}(x_i) = \sum_{i=1}^{n} \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} \right) =$$

$$= \sum_{i=1}^{n} \left(\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right) =$$

$$= -\, n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2$$

Differentiating the log likelihood

- From previous slide

$$\ell(\mu, \sigma) = -n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \ .$$

- We compute

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{n}{\sigma^2} (\overline{x}_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$= -\frac{n}{\sigma^3} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

Conclusion

- Solving $\dfrac{\partial \ell}{\partial \mu} = 0$ and $\dfrac{\partial \ell}{\partial \sigma} = 0$ yields

$$\hat{\mu} = \overline{x}_n, \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

Invariance property

- The following is an important property of MLE. We do not prove it, but we will use it repeatedly in studying linear statistical models.

- Invariance property of MLE: if $\hat{\theta}$ is the maximum likelihood estimator of a parameter θ and $g(\theta)$ is an invertible function of θ, then $g(\hat{\theta})$ is the maximum likelihood estimator for $g(\theta)$.

- Notation: we will sometimes denote the MLE of θ by $\hat{\theta}$. In this case the above reads $\widehat{g(\theta)} = g(\hat{\theta})$. That is you can 'bring the hat inside the function'.

Summary

- Likelihood function.
- Maximum likelihood estimates and estimators.
- Log likelihood.
- Example: normal distribution.
- Invariance property of MLE.

Section 2.3: Confidence Intervals

The limitations of point estimates

- Suppose we have an estimator $\hat{\theta}(X_1, \ldots, X_n)$ of an unknown parameter θ.
- We use its realization $\hat{\theta}(x_1, \ldots, x_n)$, based on a dataset from an experiment, as our estimate for θ.
- Suppose we repeat the experiment many times: do you expect the estimates to remain the same?

The need for interval estimates

- We cannot say that the estimate $\hat{\theta}(x_1, \ldots, x_n)$ equals the true value of θ, but rather than it is only close to the true θ.
- We want to provide an interval of plausible values for θ and also add a specific statement about how confident we are that the true θ is among them.
- This will be based on the knowledge of the sampling distributions of corresponding estimators.

Confidence intervals

- Suppose a dataset $x_1, \ldots, x_n$ is given, modelled as realization of random variables $X_1, \ldots, X_n$. Let θ be the parameter of interest, and γ a number between 0 and 100.

- If there exist sample statistics $\mathcal{L}_n = g(X_1, \ldots, X_n)$ and $\mathcal{U}_n = h(X_1, \ldots, X_n)$ such that

$$P(\mathcal{L}_n < \theta < \mathcal{U}_n) = \gamma\%$$

  for every value of θ, then (ℓ_n, u_n) where $\ell_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$ is called a $\gamma\%$ confidence interval for θ.

- The number $\gamma\%$ is called the confidence level.

Confidence intervals, cont.

- There is no way of knowing whether an individual confidence interval is correct, in the sense it indeed does cover θ.
- The procedure guarantees that each time we make a confidence interval we have the probability $\gamma\%$ of covering θ.

Example: Normal data (variance known)

- Suppose the data can be seen as the realization of a sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$ distribution and μ is the unknown parameter of interest, while σ^2 is known.
- The mean $\overline{X}_n$ has an $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution.
- Therefore,

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \; .$$

by central limit theorem.

Example, cont.

- If c_ℓ and c_u are chosen such that $P(c_\ell < Z < c_u) = \gamma\%$ for Z a $N(0,1)$ distributed random variable, then

$$\begin{aligned}\gamma\% =& P\left(c_\ell < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < c_u\right) = P\left(c_\ell \frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < c_u \frac{\sigma}{\sqrt{n}}\right) \\ =& P\left(\overline{X}_n - c_u \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n - c_\ell \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

- Thus, if $\ell_n = \overline{x}_n - c_u \frac{\sigma}{\sqrt{n}}$ and $u_n = \overline{x}_n - c_\ell \frac{\sigma}{\sqrt{n}}$ the interval (ℓ_n, u_n) covers μ with probability $\gamma\%$.

Example, cont.

- Thus $\left(\overline{x}_n - c_u \frac{\sigma}{\sqrt{n}}, \overline{x}_n - c_\ell \frac{\sigma}{\sqrt{n}}\right)$ is a $\gamma\%$ confidence interval for parameter μ.

- A common choice is to divide $\alpha = 1 - \gamma\%$ evenly between the two tails of the distribution, that is

$$P(Z \geq c_u) = \alpha/2 \text{ and } P(Z \leq c_\ell) = \alpha/2$$

so that $c_u = Z_{\alpha/2}, \quad c_\ell = -Z_{\alpha/2}$.

- In summary, the $100(1-\alpha)\%$ C.I. for μ is

$$(\overline{x}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- If $\alpha = 0.05$ then $Z_{\alpha/2} = 1.96$.

Example: Normal data (variance unknown)

- Suppose the data can be seen as a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution where both μ and σ^2 are unknown.
- The fact that $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ has the standard normal distribution is not useful, since the corresponding C.I. contains σ, which is unknown.

Example: cont.

- We use the following general fact (without proof):

  For a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution,

  $$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

  where S_n^2 is the sample variance.

- In other words, replacing σ with S_n changes $N(0,1)$ into t_{n-1}.

- Discussion question: what is the corresponding $100(1-\alpha)\%$ confidence interval for μ?

Example, cont.

- The $100(1-\alpha)\%$ C.I. for μ is

$$\left(\overline{x}_n - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}},\ \overline{x}_n + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right)$$

  where $P(T_n \geq t_{\alpha/2,n-1}) = \alpha/2$ and $P(T_n \leq -t_{\alpha/2,n-1}) = \alpha/2$.

- Find the value of $t_{\alpha/2,n-1}$ from the t-distribution table.

- Or using R programming, $t_{\alpha/2,n-1} =$

```
-qt(alpha/2,df)
```

Approximate C.I.

- If we have a random sample whose distribution we approximately know, we can use our confidence interval procedure to derive approximate C.I. for the parameters.
- For large datasets the central limit theorem ensures that this method provides confidence intervals with approximately correct confidence levels.

Example

- Suppose we know that $B \sim \text{Bin}(n, p)$, but we don't know p and would like to produce a 95% confidence interval for it.
- In this case we can use the normal approximation to the binomial for large n, say

$$\frac{B - np}{\sqrt{np(1-p)}} \sim N(0, 1) \ .$$

Example

- Recall that if $Z \sim N(0,1)$ then $P(-1.96 < Z < 1.96) = 0.95$. Therefore

$$P\left(-1.96 < \frac{B - np}{\sqrt{np(1-p)}} < 1.96\right) \approx 0.95$$

- Suppose we observe a value b of B, so that the approximate 95% confidence interval for p is based on this observation, that is:

$$-1.96 < \frac{b - np}{\sqrt{np(1-p)}} < 1.96.$$

Example, cont.

- Squaring the last inequalities we obtain

$$\frac{(b - np)^2}{np(1-p)} < 1.96^2$$

- Rearranging as a function of p we have

$$p^2 - \frac{n(2b + 1.96^2)}{1 + 1.96^2 n} p + b^2 < 0.$$

- The left hand side is a quadratic function of p, so the interval we want is the interval between the two roots of the quadratic.

Summary

- The need for interval estimates.
- Confidence intervals.
- Normal data (variance known).
- Normal data (variance unknown).
- Approximate confidence intervals.

Section 2.4: Testing Hypotheses

Null and alternative hypotheses

- The first of the two competing propositions is called the null hypothesis, denoted H_0 and the second one is called the alternative hypothesis, denoted H_1.
- The null hypothesis is presumed to be true until the data provide convincing evidence against it.
- If we reject the null hypothesis we will accept H_1.
- The next step is a criterion that provides an indication about whether H_0 is false. This involves a test statistic.

Test statistic

- Suppose that the dataset is modelled as the realization of random variables $X_1, \ldots, X_n$.
- A test statistic is any sample statistic $\mathrm{T} = h(X_1, \ldots, X_n)$ whose numerical value is used to decide whether we reject H_0.
- The values of the test statistic can be viewed on a credibility scale for H_0, and we must determine which of these values provide evidence in favor of H_0, and which provide evidence in favor of H_1.

Test for a single mean (variance known)

- For a random sample $X_1, \ldots, X_n$, let $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$. Let the sample statistic T be the sample mean $\bar{X}$.
- We want to test if μ is equal to a constant μ_0 or not. When σ^2 is known, $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$,
  - $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$,

    If $P\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right) \leqslant \alpha$, reject H_0.

Test for a single mean (variance unknown)

- We want to test if μ is equal to a constant μ_0 or not. When σ^2 is unknown, and let S^2 be the sample variance, $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$,
  - $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > \frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$,

    If $P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < \frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right) \leqslant \alpha$, reject H_0.
  - $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$,

    If $P\left(\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > \left|\frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right|\right) \leqslant \alpha$, reject H_0.

Example

- Suppose that we know that blood pressures in humans from Honolulu follow a normal distribution, but we don't know the mean μ.
- For the rest of the U.S. The mean is known to be 120 mm Hg, and the standard deviation is 10 mm Hg.
- Some people think that Honolulans have different blood pressure on average with other Americans, so we want to test the hypothesis H_0 that $\mu = 120$ against the alternative H_1 that $\mu \neq 120$.

Example, cont.

- We measure the blood pressure of 100 Honolulans selected independently at random and compute the mean $\bar{x}$, which turns out to be 130.1 mm Hg.
- We make the assumption that the standard deviation for blood pressure of Honolulans is also 10mm Hg.
- Since we have 100 independent observations of a $N(\mu, 100)$ random variable, we have

$$\frac{\bar{X} - \mu}{10/\sqrt{100}} \sim N(0, 1)$$

where $\bar{X}$ is the sample mean.

Example, cont.

- If the mean is really 120 (that is, H_0 is true) then

$$Z = \frac{\bar{X} - 120}{10/\sqrt{100}} \sim N(0,1)$$

- Hence the probability of observing a sample mean of not 130.1 mm Hg

$$P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) = P\left(\left|\frac{\bar{X} - 120}{10/\sqrt{100}}\right| > \left|\frac{130.1 - 120}{10/\sqrt{100}}\right|\right)$$
$$= P\left(Z > \frac{130.1 - 120}{10/\sqrt{100}}\right) + P\left(Z < -\frac{130.1 - 120}{10/\sqrt{100}}\right)$$
$$= P(Z > 10.1) + P(Z < -10.1) < 0.001$$

- This probability is called the *p*-value of our test. Since it is very small, we can regard it as evidence to reject H_0 in favour of our alternative that $\mu \neq 120$.

Error types

- There are two situations in which the decision made on the basis of data is wrong:
  - The null hypothesis H_0 may be true, whereas data lead to rejection of H_0.
  - The alternative hypothesis H_1 may be true, whereas we do not reject H_0 on the basis of the data,
- A type I error occurs if we falsely reject H_0. A type II error occurs if we falsely do not reject H_0.
- The question is: what should be the probability of committing a type I error, i.e. for which values of T should we reject H_0?

Significance level

- The significance level is the largest acceptable probability of committing a type I error and is denoted by α, where $0 < \alpha < 1$.
- We speak of 'performing the test at level α' as well as 'rejecting H_0 in favor of H_1 at level α'.
- We usually take $\alpha = 0.05$.

Critical region and critical values

- Suppose we test H_0 against H_1 at significance level α by means of a test statistic T.

- The set $K \subset \mathbb{R}$ that corresponds to all values of T for which we reject H_0 in favour of H_1 is called the critical region. Values on the boundary of the critical region are called critical values.

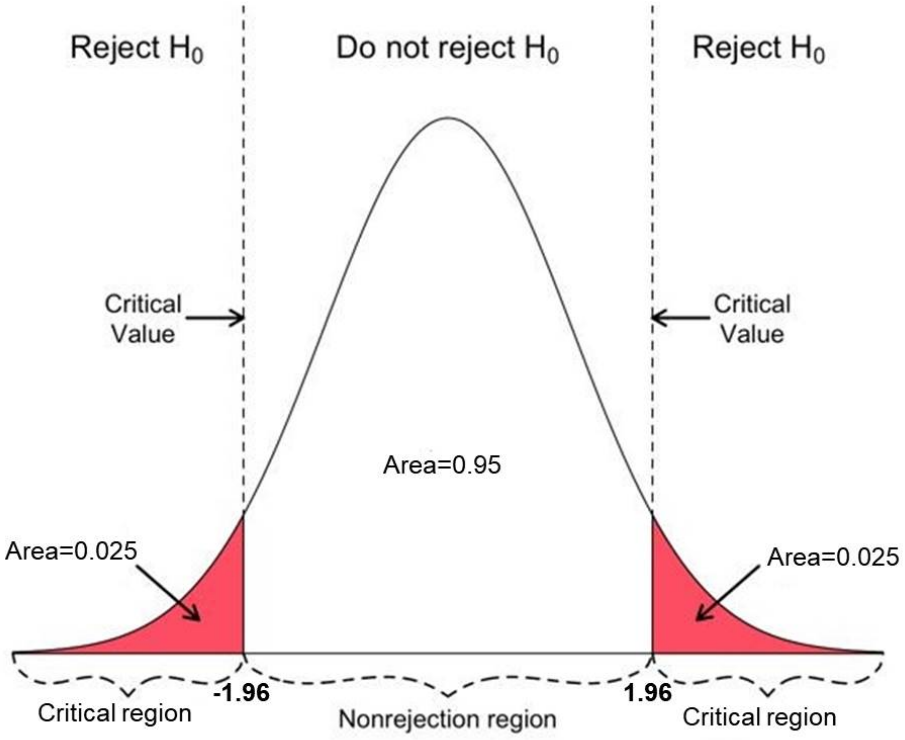
- The precise shape of the critical region depends on both the chosen significance level α and the test statistic T that is used. But it will always be such that the probability that $\mathrm{T} \in K$ satisfies

$$P(\mathrm{T} \in K) \leq \alpha \qquad \text{in the case that } H_0 \text{ is false.}$$

Example

- Suppose the test statistic is $N(0,1)$ and the significance level is $\alpha = 0.05$. Then the critical values are ± 1.96, while the critical region is $(-\infty, -1.96) \cup (1.96, +\infty)$.

- In picture: the sum of the areas of the two tails (red) is 0.05.



p-values

- If the observed value of the statistic falls in the critical region, we reject the null hypothesis H_0.
- The 2-sided p-value is the sum of the areas of the two tail probabilities $P(\mathrm{T} \leq -\mathrm{t}) + P(\mathrm{T} \geq \mathrm{t})$. The left-sided p-value is $P(\mathrm{T} \leq -\mathrm{t})$ and the right-sided p-value is $P(\mathrm{T} \geq \mathrm{t})$.
- The p-value expresses how likely is to obtain a value of the test statistic T at least as extreme as the value t obtained for the data.
- The smaller the p-value, the stronger evidence the observed value t bears against H_0.

Test for a single mean, unknown variance

- Suppose we can take independent samples from a normal distribution $N(\mu, \sigma^2)$ in which both μ and σ^2 are unknown. Then

$$\mathrm{T} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- Suppose that we observe $\bar{x} = 102$, $s = 4.7$ with a sample size $n = 25$. We want to test the null hypothesis H_0 that $\mu = 100$ against the alternative H_1 that $\mu > 100$ at significance level $\alpha = 0.05$.

Test for a single mean, unknown variance, cont.

- We have $\frac{102-100}{4.7/5} = 2.13 \sim t_{24}$.
- The p-value is $P\left(\mathrm{T} > \frac{102-100}{4.7/5}\right) = 0.0218 < 0.05$.
- Alternatively, read from the t distribution table, the critical value is 1.711, the critical region is $(1.711, +\infty)$.
- Therefore we reject the null hypothesis.

Hypothesis tests and confidence intervals

- Hypothesis tests and confidence intervals are equivalent ways to do interval estimation.
- Suppose that for some parameter θ we test $H_0 : \theta = \theta_0$. Hence we reject $H_0 : \theta = \theta_0$ in favour of H_1 at level α if and only if θ_0 is not in the $100(1-\alpha)\%$ C.I. for θ.
- Note: If the hypothesis test and the C.I. give contradictory results, it means you have made a calculation mistake!

Summary

- Null and alternative hypotheses.
- Test statistic for hypothesis testing.
- Type I and type II errors.
- Significance level.
- Critical values and critical regions.
- p-values.
- Example: test for single mean, unknown variance.
- Relation between hypothesis tests and confidence intervals.

Chapter 3: Simple Linear Regression, Part 1

Dr.Ting Wei

Section 3.1:
Relationship between variables

Relationship between random variables.

- Two random variables can be independent or dependent.
- If they are dependent, the relationship between them can be strong or weak.
- Question: how can we find a numerical measure of the strength of a relationship between two random variables?

Question

Consider the following two experiments:

- Experiment 1: measure X and Y where
  $X =$ weight of a sample of water,
  $Y =$ volume of the same sample of water.

- Experiment 2: measure X and Y where
  $X =$ body weight of a human,
  $Y =$ same human height.

- Question: in which experiment do you think the relationship between X and Y be strongest?

Measuring strength of relationship.

- Main idea: the covariance and correlation are two measures that quantify the difference in strength of a relationship between two random variables.

Covariance

- Let X, Y be random variables with means $E(X) = \mu_X$ and $E(Y) = \mu_Y$. The covariance between them, often denoted σ_{XY}, is defined as

$$\sigma_{XY} = \text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- Note: $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.

The correlation coefficient

- Problem with the covariance: it is measured in units of X times units of Y. We want to make this scale-free. We do this by introducing the coefficient of correlation.

- The coefficient of correlation is defined by

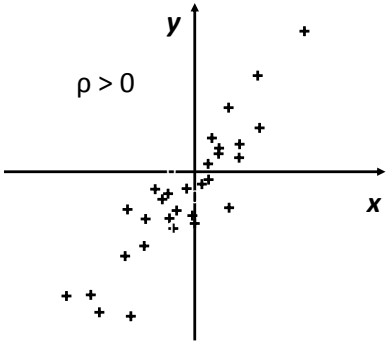
$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and

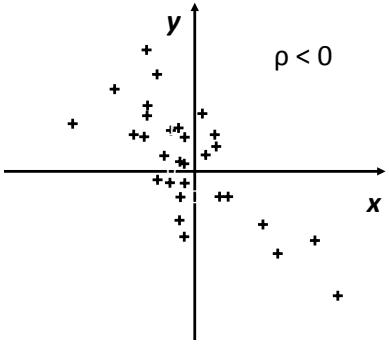
$$-1 \leqslant \rho \leqslant 1$$

The sign of the correlation coefficient

- The sign of ρ is the sign of $\sigma_{XY} = \text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.



$\rho > 0$: we tend to have more positive values of $(X - \mu_X)(Y - \mu_Y)$: X and Y tends to increase together.



$\rho < 0$: we tend to have more negative values of $(X - \mu_X)(Y - \mu_Y)$: as X increases, Y tends to decrease and vice-versa.

Estimating the correlation coefficient

- We make n pairs of observations

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

of random variables X and Y.

- We want an estimate $\hat{\rho}$ of ρ from these observations.

- By definition, $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$, therefore

$$\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Estimating the correlation coefficient, cont.

- The estimates of σ_X, σ_Y, σ_{XY} are:

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 \ ,$$
$$\hat{\sigma}_Y^2 = s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 \ ,$$
$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \ .$$

- Therefore

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- This is called the sample correlation coefficient between X and Y.

Some useful notation and formulas

- We denote

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \ ,$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \ = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n},$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \ .$$

- Therefore, in this notation

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly).

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

Example: Pulse Data, cont.

- Let X = first pulse rate, Y = second pulse rate

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$
$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$
$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$\sum_{i=1}^{n} x_i = 890$; $\sum_{i=1}^{n} x_i^2 = 73780$; $S_{xx} = 1770.909$.
$\sum_{i=1}^{n} y_i = 1241$; $\sum_{i=1}^{n} y_i^2 = 141653$; $S_{yy} = 1645.636$.
$\sum_{i=1}^{n} x_i y_i = 101404$; $S_{xy} = 995.818$.

Therefore

$$\hat{\rho} = \frac{995.818}{\sqrt{1770.909 \times 1645.636}} = 0.583$$

Correlation as a measure of linear relationship

- Theorem: *The sample correlation coefficient satisfies the inequality* $|\hat{\rho}| \leq 1$, *with equality if and only if the points* $\{(x_i, y_i)\}$ *lie on a straight line.*

- Therefore correlation is a measure of linear association

$\hat{\rho} = 0$	no linear association.
$\hat{\rho} = 1$	perfect positive linear association; x and y increase together.
$\hat{\rho} = -1$	perfect negative linear association; y decreases as x increases.

The meaning of zero correlation

- Let $X \sim N(\mu, \sigma^2)$ and let $Y = (X - \mu)^2$. Then $E(Y) = \text{var}(X) = \sigma^2$.
- Since $E(X - \mu) = E[(X - \mu)^3] = 0$ we have

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] =$$
$$= E\left[(X - \mu)\left((X - \mu)^2 - \sigma^2\right)\right] = 0$$

And thus, $\rho = 0$.

Inferences about the population correlation

- The sample correlation coefficient $\hat{\rho}$ is an estimate of the population correlation coefficient ρ.
- Hence, if $\rho = 0$ is true, we expect $\hat{\rho} \approx 0$.
- We want to test the hypotheses $H_0 : \rho = 0$, and compute a confidence interval for ρ.

Testing the hypothesis $\rho = 0$

- We want to test the hypothesis: $H_0 : \rho = 0$, $H_1 : \rho \neq 0$.
- We can use the t-statistics

$$\mathrm{T} = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$$

- Under the null hypothesis this has a t distribution with $n-2$ degrees of freedom.

Example: Pulse data

- We computed the sample correlation coefficient for the Pulse data and found it to be $\hat{\rho} = 0.583$.
- Using the formula $\mathrm{T} = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$ with $\hat{\rho} = 0.583$ and $n = 11$, we obtain

$$\mathrm{T} = 0.5833\sqrt{\frac{9}{1 - 0.5833^2}} = 2.1545 \ .$$

- From the table of the t-distribution with 9 df, we see that critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$. Hence the hypothesis $\rho = 0$ is acceptable.

Confidence intervals for ρ

- It can be shown that the 95% confidence interval for ρ is given by (l_1, l_2) where

$$l_1 = \frac{e^{-\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{-\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1}, \qquad l_2 = \frac{e^{\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} - 1}{e^{\frac{2\times 1.96}{\sqrt{n-3}}} \cdot \frac{1+\hat{\rho}}{1-\hat{\rho}} + 1}$$

- Thus if we repeat the experiment many times, the CI covers the true value of ρ with 95% probability.

Pulse data

- For the pulse data we calculated $\hat{\rho} = 0.5833$; also $n = 11$.
- Hence 95% confidence interval is

$$(l_1,\ l_2) = \left(\frac{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{-2\times 1.96}{\sqrt{8}}\right) - 1}{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{-2\times 1.96}{\sqrt{8}}\right) + 1},\ \frac{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{2\times 1.96}{\sqrt{8}}\right) - 1}{\left(\frac{1+0.5833}{1-0.5833}\right) \exp\left(\frac{2\times 1.96}{\sqrt{8}}\right) + 1} \right)$$
$$= (-0.026,\ 0.877)\ .$$

- Note: the interval contains 0, as expected.

Summary

- Coefficient of correlation $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.
- Estimation: ρ is estimated by the sample correlation coefficient $\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$.
- Theorem: $|\hat{\rho}| \leq 1$ and $|\hat{\rho}| = 1$ iff $\{(x_i, y_i)\}$ lie on a straight line. Thus $\hat{\rho}$ (and ρ) is a measure of linear association.
- We showed in an example that zero covariance between two random variables does not imply their independence.
- We discussed the hypothesis test $\rho = 0$ using the t-statistics, and we discussed how to compute the corresponding confidence interval.
- Example: Pulse data.

Section 3.2: Least Squares Estimation

Introduction to regression: response versus explanatory variables

- Regression analysis is used for explaining or modelling the relationship between a single variable Y, called the response variable, and one or more variables $X_1, \ldots, X_k$, called predictors or explanatory variables.
- In regression analysis we treat the X variables as fixed constants. We are interested in the effect of the predictors on the response.
- If $k = 1$ we have a simple regression model.

Correlation versus simple regression

- Regression is different from correlation.
- Correlation describes the joint distribution between random variables X and Y. The correlation between X and Y is the same as the correlation between Y and X.
- In simple regression, one of the variables (Y) is the response, which we predict from the X. So regression describes the conditional distribution of Y given X.

Introduction to regression: objectives

- Possible objectives of regression analysis include:
  - Prediction of future observations.
  - Assessment of the effect of, or relationship between, explanatory variables and the response.
  - General description of data structure.

Linear models

- Suppose we want to model the response Y in terms of three predictors X_1, X_2, X_3. One general form of the model could be

$$Y = f(X_1, X_2, X_3) + \varepsilon$$

where f is an unknown function and ε is an error term.

- In a linear model , the function f has the following form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where β_i, $i = 0, 1, 2, 3$, are unknown parameters.

The simple linear regression model

- The simple linear regression model has the form

$$Y = a + bX + \varepsilon$$

  where ε is a random variable with a specified probability distribution with mean 0.

- It follows that $E(Y) = a + bX$.

- We call $a + bX$ the systematic (or deterministic) part of the model, while ε is the random component.

Assumptions in the simple linear regression model

- In the simple linear regression model $Y = a + bX + \varepsilon$ we assume

  a) ε has a normal distribution with $E(\varepsilon) = 0$

  b) $\text{var}(\varepsilon) = \sigma^2$ is independent of X.

- If n independent observations are made on this model, we have n independent random variables

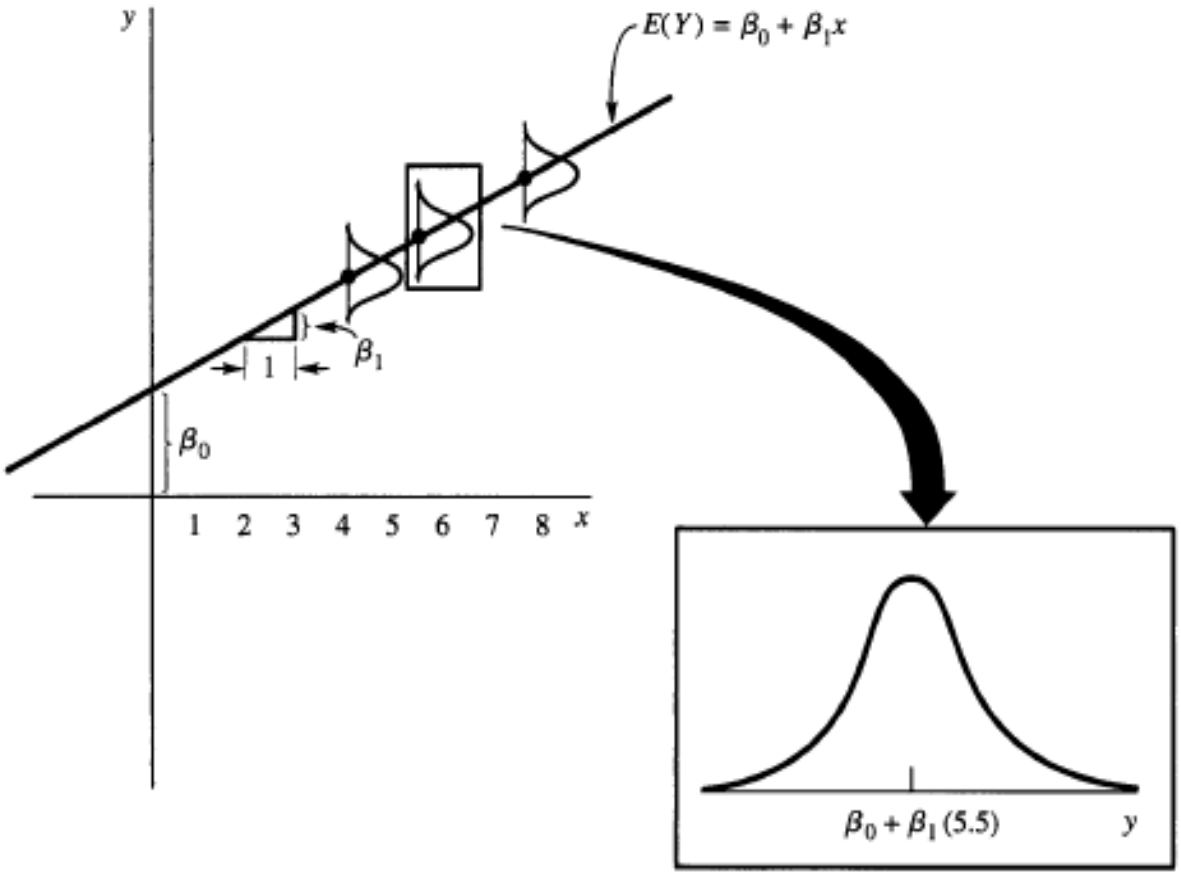
$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \ldots, n$$

  where ε_i are independent random variables with normal distribution $N(0, \sigma^2)$

- Equivalently: $Y_i \sim N(a + bX_i, \sigma^2)$.

Graphical representation

For each X there is a population of possible values of Y, which is normally distributed with mean $a + bX$ and variance σ^2.



Estimation of parameters: main aim

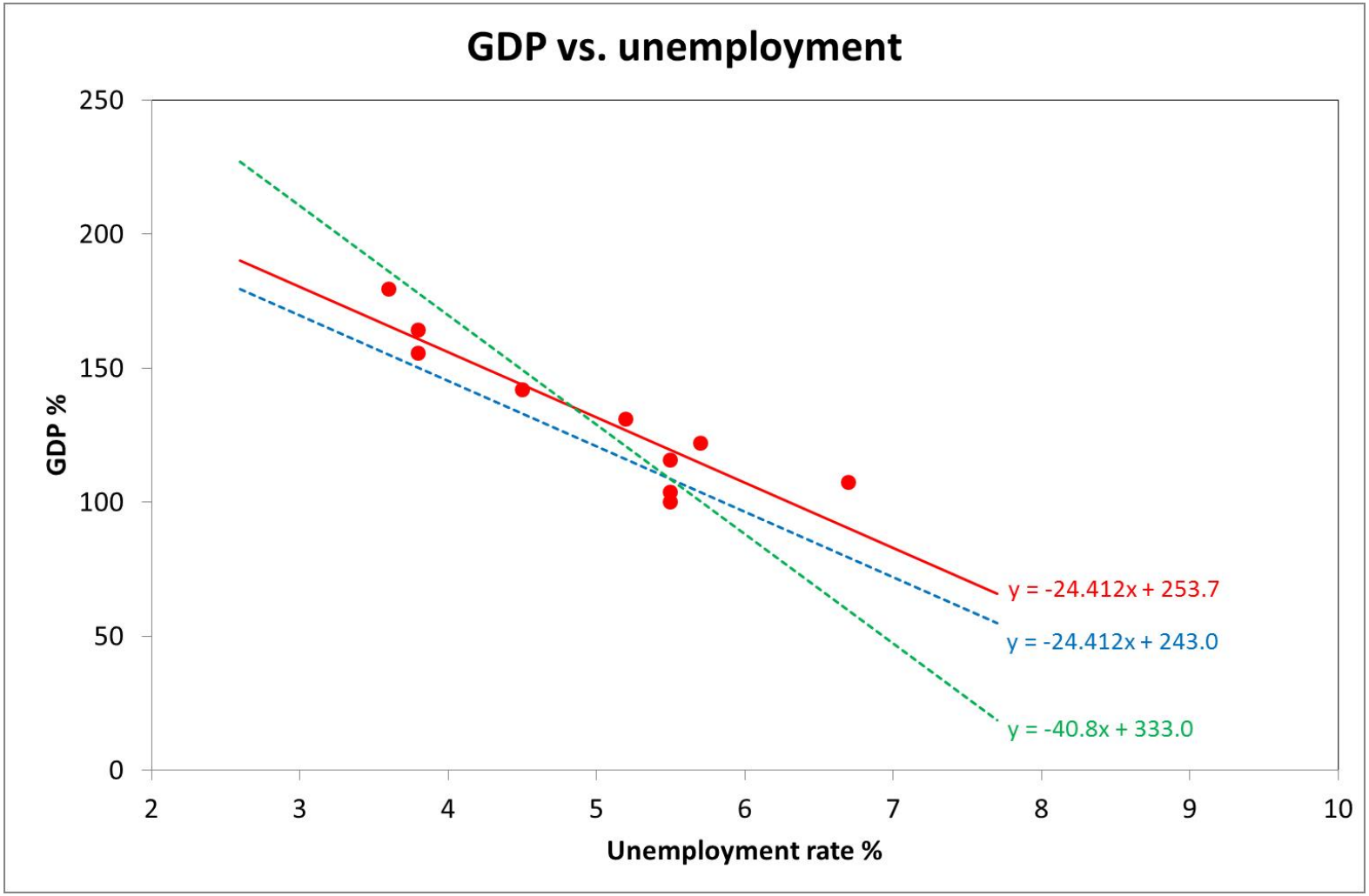
- Our main aim is to estimate the model parameters a and b. That is, to produce the line $E(Y) = \hat{a} + \hat{b}X$ which best fits the data points.
- Then $\hat{y}_i = \hat{a} + \hat{b}x_i$ is called the ith fitted value. It is the predictor of y_i given the value x_i.

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100.0
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

Question



Question: Which of the three lines best fit the data, and why?

MLE for simple linear regression

- Recall that in the simple linear regression model $Y_i \sim N(a + bX_i, \sigma^2)$.

- Likelihood $= \mathcal{L}(a, b, \sigma^2 | \text{data}) =$

$$= \prod_{i=1}^{n} f(y_i | a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(y_i - a - bx_i)^2}{2\sigma^2} \right\}$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \right\} .$$

- To maximize the likelihood we need to minimize

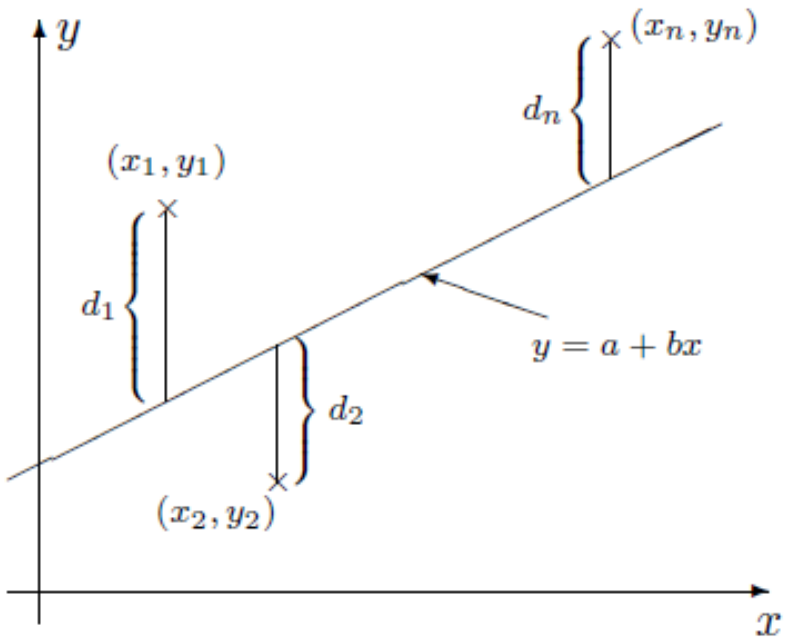
$$SS = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

- Thus the maximum likelihood estimates are those values which minimise the sum of squared distances between the observed and fitted values.

Least square estimation

- The estimates obtained by minimizing SS are also called the least square estimates.

- The line $\hat{a} + \hat{b}x$ is the one that minimizes the sum of squares of the vertical distances between the observed points and the line



Minimizing SS

- $SS = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ is a function of the two variables a and b.
- To minimize it, we need to equate to zero the partial derivatives

$$\frac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) \quad \text{and}$$
$$\frac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) \ .$$

Some calculations

- $\dfrac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) = 0$ implies

$$\sum_{i=1}^{n} y_i = n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i \quad \text{or} \quad \bar{y} = \hat{a} + \hat{b}\bar{x}\ .$$

- $\dfrac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) = 0$ implies

$$\sum_{i=1}^{n} x_i y_i = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2\ .$$

- These equations are called the likelihood equations.

Some calculations, cont.

- From previous slide

$$\bar{y} = \hat{a} + \hat{b}\bar{x} \ , \qquad \sum_{i=1}^{n} x_i y_i = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 \ .$$

- Therefore

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \hat{b}\bar{x})\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 = n\bar{y}\bar{x} - \hat{b}\bar{x}^2 n + \hat{b}\sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \hat{b}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Equivalently

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Some calculations, cont.

- From previous slide:

$$\bar{y} = \hat{a} + \hat{b}\bar{x} \ , \qquad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

- In conclusion, the MLE of a and b are

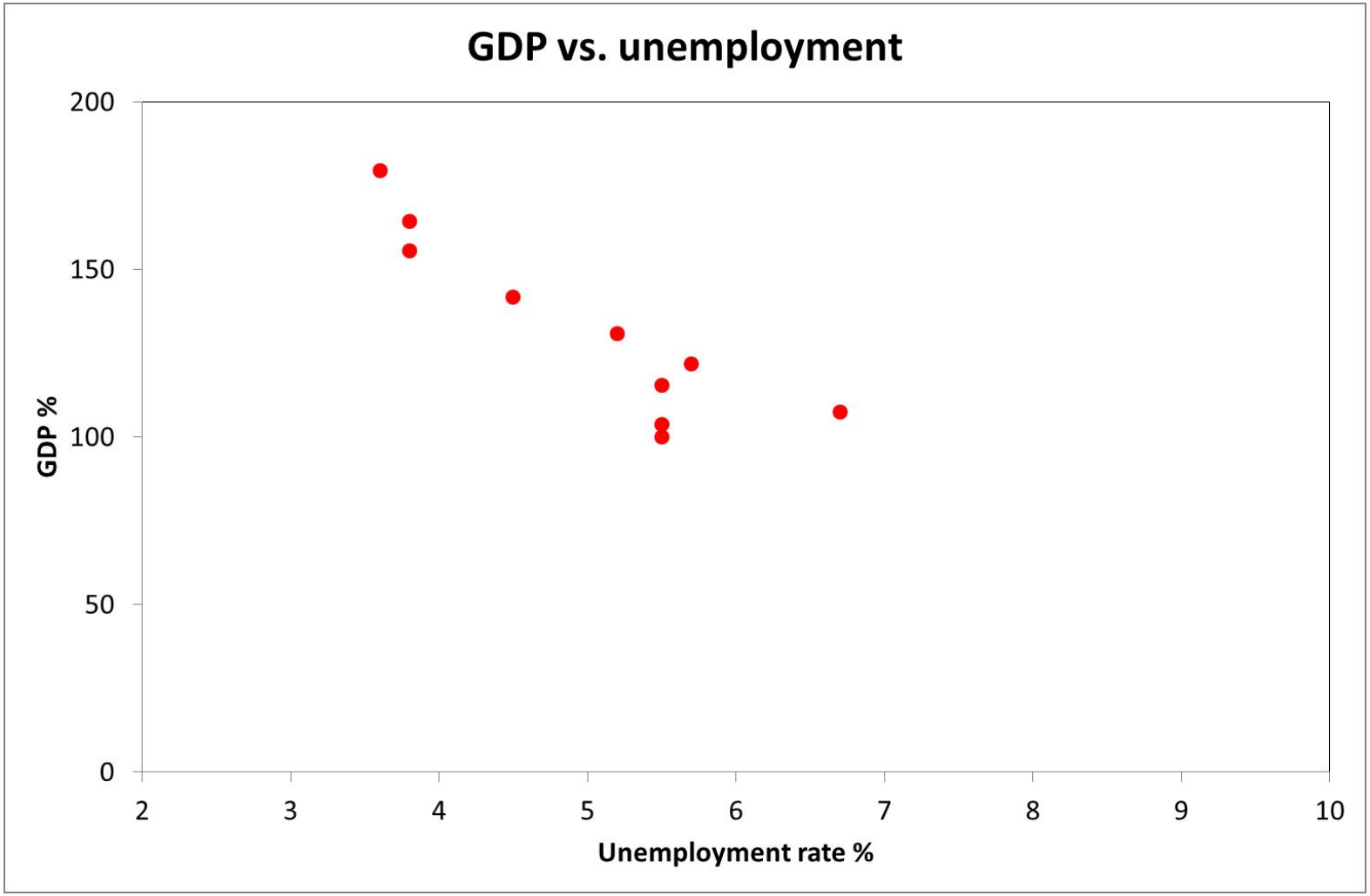
$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \ .$$

Residuals

- The ith residual is the difference between the ith observed and fitted values

  $r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$.

Example: Unemployment and GDP, cont.



Plot of the GDP against the unemployment rate rate

Example: Unemployment and GDP, cont.

- We can calculate as done in lecture 10:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = -225.954$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 9.256$$
$$\sum_i x_i = 49.8\,, \quad \bar{x} = 4.98\,, \quad \sum_i y_i = 1321.3\,, \quad \bar{y} = 132.13$$

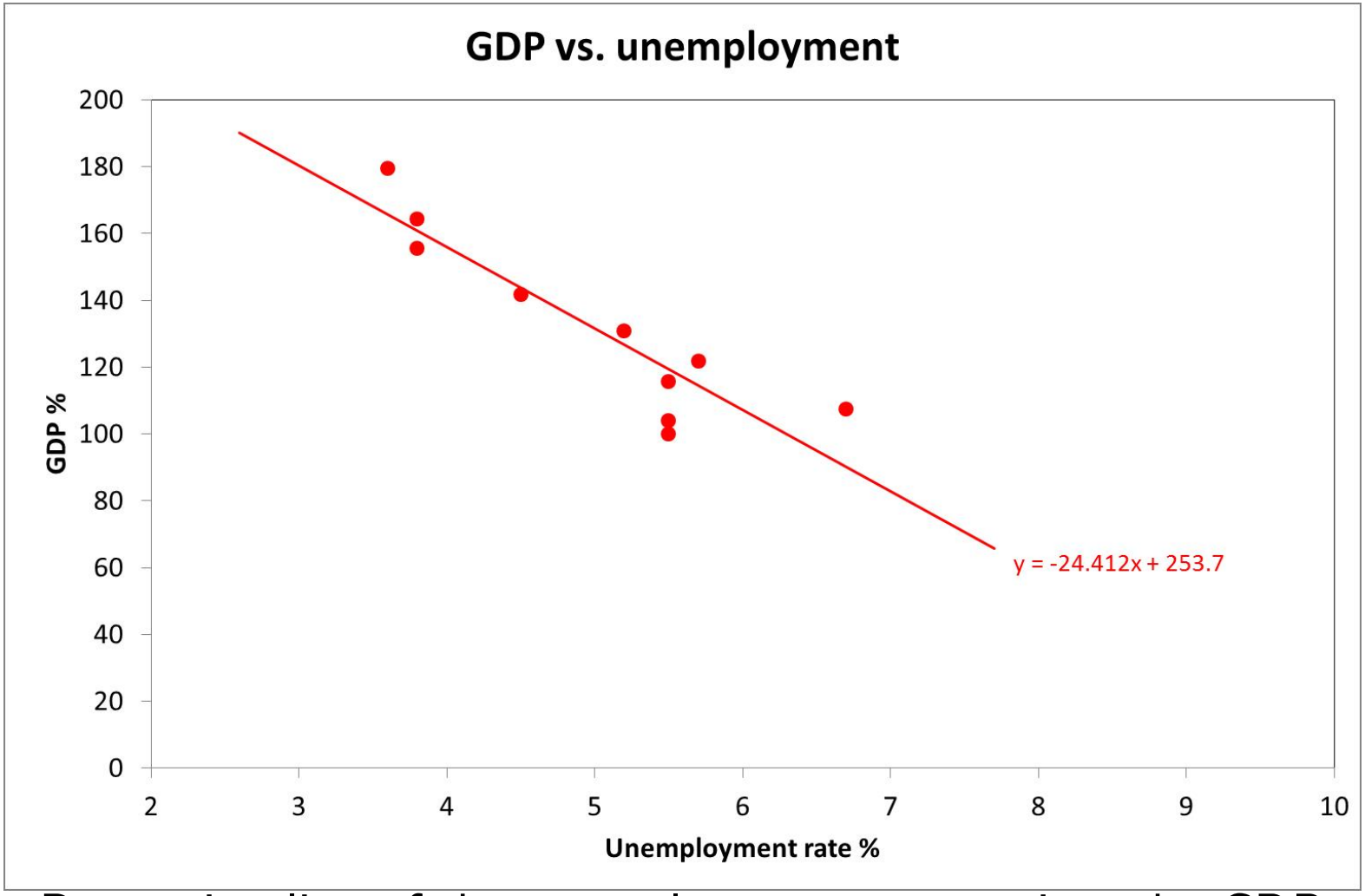
- Therefore

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = -24.4116 \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 253.7$$

- The regression line is

$$y = \hat{a} + \hat{b}x = 253.7 - 24.4116x\ .$$

Example: Unemployment and GDP, cont.



Regression line of the unemployment rate against the GDP percentage.

Example: Unemployment and GDP, cont.

- Calculating and comparing RSS for the three lines.

$$y = -24.412x + 253.7 \quad \text{Equation A (regression)}$$
$$y = -24.412x + 243.0 \quad \text{Equation B}$$
$$y = -40.8x + 333.0 \quad \text{Equation C}$$

Data		Fitted values			Square of residuals		
UN.%	GDP %	Eq. A	Eq. B	Eq. C	Eq. A	Eq. B	Eq. C
5.5	100.0	119.436	108.736	108.6	377.766	76.321	73.960
5.5	103.9	119.436	108.736	108.6	241.374	23.389	22.090
6.7	107.5	90.142	79.442	59.64	301.290	787.236	2290.58
5.5	115.6	119.436	108.736	108.6	14.716	47.112	49.000
5.7	121.9	114.554	103.854	100.44	53.965	325.662	460.532
5.2	130.9	126.760	116.060	120.84	17.142	220.235	101.204
4.5	141.9	143.848	133.148	149.4	3.794	76.601	56.250
3.8	155.6	160.936	150.236	177.96	28.472	28.773	499.970
3.8	164.4	160.936	150.236	177.96	12.000	200.621	183.874
3.6	179.6	165.818	155.118	186.12	189.937	599.357	42.510
				RSS	1240.46	2385.31	3779.97

Summary

- The simple linear regression model is used for modeling the relationship between a response variable Y and an explanatory variable X.
- The model has the form $Y = aX + b + \varepsilon$ where ε is a random variable with mean 0 and constant variance σ^2.
- If n independent observations are made on this model, then $Y_i \sim N(a + bX_i, \sigma^2)$.
- The MLE of the parameters a and b correspond to the least square estimators and are given by:

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \ .$$

- The regression line $y = \hat{a}x + \hat{b}$ is the best line fitting the data points. The value $\hat{y}_i = \hat{a}x_i + \hat{b}$ is the ith fitted value while $y_i - \hat{y}_i$ is the ith residual.

Section 3.3: Point estimation of sigma

Why estimating σ^2?

Reasons to find an estimate of σ^2 include:

a) We want to obtain an indication of the variability of the probability distribution of Y.

b) As we will see later on, a variety of inferences concerning the regression function and the prediction of Y require an estimate of σ^2.

Residuals

- The ith residual is the difference between the ith observed and fitted values

  $r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$

An estimate of σ^2 for the simple regression model

- The residual sum of squares:

$$RSS = \sum_{i=1}^{n} r_i^2$$

- The estimate of the variance is:

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}$$

Properties of $\hat{\sigma}^2$

- Recall that a point estimator $\hat{\theta}$ of a parameter θ is called unbiased if $E(\hat{\theta}) = \theta$. Otherwise, it is called biased.
- It can be shown that $\hat{\sigma}^2 = \dfrac{RSS}{(n-2)}$ is unbiased.
- One can prove that the maximum likelihood estimator of σ^2 is $\dfrac{RSS}{n}$ and this is a biased estimator.
- Hence $\hat{\sigma}^2$ is a bias-corrected version of the MLE.

Properties of RSS

- Recall that in the simple linear regression model, $Y_i = a + bX_i + \varepsilon_i$ where ε_i are normal independent random variables with mean 0 and variance σ^2.
- Hence $\frac{\varepsilon_i}{\sigma}$ has the standard normal distribution.
- It follows that $\sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi_n^2$ since it is sum of n independent standard normal distributions.
- What about $\frac{RSS}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{r_i}{\sigma}\right)^2$?

Properties of RSS, cont.

- In Section 3.2 we showed that, equating to zero the partial derivatives of SS we found

$$\frac{\partial SS}{\partial a} = \sum_{i=1}^{n} 2(-1)(y_i - a - bx_i) = 0$$
$$\frac{\partial SS}{\partial b} = \sum_{i=1}^{n} 2(-x_i)(y_i - a - bx_i) = 0$$

- Since $r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$ this is same as

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} x_i r_i = 0$$

Properties of RSS, cont.

- Hence the r_i are not independent random variables as they satisfy two linear restrictions.
- It can be shown that

$$\frac{RSS}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{r_i}{\sigma}\right)^2 \sim \chi^2_{n-2}$$

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

- In lecture 11 we calculated:
  $\bar{x} = 4.98 \quad S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954.$
  $\hat{a} = 253.7, \qquad \hat{b} = -24.4116\ .$

Example: Unemployment and GDP, cont.

- From previous slide the regression line equation is: $y = 253.7 - 24.4116x$.

- Hence we can calculate the residuals:

Year	Unempl. rate % (x)	GDP % (y)	Residuals (r_i)
1	5.5	100	-19.436
2	5.5	103.9	-15.536
3	6.7	107.5	17.358
4	5.5	115.6	-3.836
5	5.7	121.9	7.346
6	5.2	130.9	4.140
7	4.5	141.9	-1.948
8	3.8	155.6	-5.336
9	3.8	164.4	3.464
10	3.6	179.6	13.782

- Hence $RSS = \sum_i r_i^2 = 1240.457$, $\hat{\sigma}^2 = RSS/(n-2) = 1240.457/8 = 155.057$

Computing $\hat{\sigma}^2$: alternative formulas

- Since $\hat{\sigma}^2 = \dfrac{RSS}{(n-2)}$, we need to compute RSS in order to compute $\hat{\sigma}^2$.
- Computing RSS directly is tedious.
- There is an alternative formula for RSS, as follows:

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

RSS and sample correlation

- From previous slide: $RSS = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$

- From Section 3.1: $\hat{\rho} = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, that is

$$\hat{\rho}^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

- Therefore

$$RSS = S_{yy} - \frac{\hat{\rho}^2 S_{xx}S_{yy}}{S_{xx}} = S_{yy}(1 - \hat{\rho}^2)$$

- Equivalently

$$\hat{\rho}^2 = \frac{S_{yy} - RSS}{S_{yy}}$$

An interpretation

- From previous slide: $\hat{\rho}^2 = \dfrac{(S_{yy} - RSS)}{S_{yy}}$ (*)

- $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total variation in y about its mean. RSS is the residual variation in y after fitting the regression line (that is, not explained by x).
  $S_{yy} - RSS$ is the amount of variation in y explained by x.

- Therefore the formula (*) says that $\hat{\rho}^2$ is the proportion of variation in the response y explained by the explanatory variable x.

An interpretation, cont.

- From previous slide: $\hat{\rho}^2 = \dfrac{(S_{yy} - RSS)}{S_{yy}}$
- Some special cases:

$\hat{\rho} = \pm 1$	$RSS = 0$	all variation in y explained as a function of x
$\hat{\rho} = 0$	$RSS = S_{yy}$	no variation in y explained as a function of x
$\hat{\rho} = 0.8$	64%	of variation in y explained as a function of x

where 'as a function of x' means the simple linear regression line.

Example: Unemployment and GDP, cont.

- In slide 11 we calculated:

$$S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954.$$

- Therefore we compute

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 6756.361 - \frac{(-225.954)^2}{9.256} = 1240.457$$

$$\hat{\rho}^2 = \frac{S_{yy} - RSS}{S_{yy}} = \frac{6756.361 - 1240.457}{6756.361} = 0.8164$$

- Hence 81.64% of the variation in y is explained by x.

Summary

- We discussed the unbiased estimator $\hat{\sigma}^2 = \frac{RSS}{(n-2)}$ of the parameter σ, where RSS is the residual sum of squares.
- We discussed the distribution of $\frac{RSS}{\sigma^2}$.
- We showed that the sample correlation coefficient and RSS are related by $\hat{\rho}^2 = \frac{S_{yy}-RSS}{S_{yy}}$.
- We deduced that $\hat{\rho}^2$ is the proportion of variation in the response y explained by the explanatory variable x.
- We illustrated this in the GDP example.

Section 3.4: Sampling distributions

The idea of sampling distributions

- Think of repeating the experiment many times. Each time you have new observations, hence new estimates $\hat{a}, \hat{b}$.
- We want to analyze this variability in the estimates. That is we want to find $E(\hat{b})$, $var(\hat{b})$, $E(\hat{a})$, $var(\hat{a})$ and then the distributions of $\hat{b}, \hat{a}$, which are called sampling distributions.

The estimate $\hat{b}$

- From Section 3.2:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- On the other hand, $\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \bar{y}(\sum_{i=1}^{n} x_i - n\bar{x}) = 0$.

- Therefore

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Discussion

- Consider the expression of the previous slide

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Which part of this expression is random, which part is not?
- Is $\hat{b}$ a linear combination of random variables?

Answer to discussion questions

- From previous slide:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Only y_i's are random.

- Since x_i and $\bar{x}$ are constant, this has the form

$$\hat{b} = \sum_{i=1}^{n} c_i y_i \qquad c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

a linear combination of independent random variables $y_1, \ldots, y_n$.

Reminder

- The expectation of a linear combination of random variables is given by $E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i)$.
- The variance of a linear combination of *independent* random variables is

$$\text{var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \,\text{var}(y_i)$$

Expectation of $\hat{b}$

- In our case $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad E(y_i) = a + bx_i$ therefore

$$
\begin{aligned}
E(\hat{b}) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a + bx_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})bx_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{since } \sum_{i=1}^{n}(x_i - \bar{x})a{=}0 \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} b \qquad \text{since } \sum_{i=1}^{n}(x_i - \bar{x})\bar{x}{=}0 \\
&= b
\end{aligned}
$$

- Hence $\hat{b}$ is an unbiased estimator of b.

Variance of $\hat{b}$

- In our case $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad \text{var}(y_i) = \sigma^2$.
- Therefore

$$\begin{aligned}
\text{var}(\hat{b}) &= \text{var}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{1}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \,\text{var}(y_i) \\
&= \frac{\sigma^2}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Distribution of $\hat{b}$

- We calculated $E(\hat{b}) = b$ and $\text{var}(\hat{b}) = \dfrac{\sigma^2}{S_{xx}}$.
- Since y_i are normally distributed, we conclude that

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

Expectation of $\hat{a}$

- Recall from Section 3.2:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

- Therefore

$$\begin{aligned} E(\hat{a}) &= E(\bar{y} - \hat{b}\bar{x}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) - E(\hat{b})\bar{x} \\ &= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) - b\bar{x} \\ &= a + b\bar{x} - b\bar{x} \\ &= a \end{aligned}$$

- Hence $\hat{a}$ is an unbiased estimator of a.

Distribution of $\hat{a}$

- One can calculate that

$$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

- If y_i are normally distributed, we conclude

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

Example: Unemployment rate and Gross Domestic Product (GDP)

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

- In Section 3.2 we calculated:
  $\bar{x} = 4.98 \quad S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954$
  $\hat{a} = 253.7, \quad \hat{b} = -24.4116, \quad RSS = 1240.457$

Example: Unemployment and GDP

- We want to calculate the standard errors for $\hat{a}$ and $\hat{b}$ using the formulas:

$$\text{var}(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$$

σ^2 is unknown, so we use the estimate from Section 3.3

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}$$

- For unemployment and GDP example, we previously calculated

$$\hat{\sigma} = \sqrt{\frac{RSS}{(n-2)}} = \sqrt{\frac{1240.457}{8}} = 12.452$$

Example: Unemployment and GDP, cont.

- Therefore

$$\text{s.e.}(\hat{a}) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = 12.452\sqrt{\left(\frac{1}{10} + \frac{4.98^2}{9.256}\right)} = 20.759$$

$$\text{s.e.}(\hat{b}) = \hat{\sigma}\sqrt{\left(\frac{1}{S_{xx}}\right)} = 12.452\sqrt{\left(\frac{1}{9.256}\right)} = 4.093\ .$$

Distribution of residuals

- The ith residual is $r_i = y_i - \hat{a} - \hat{b}x_i$, therefore

$$E(r_i) = E(y_i) - E(\hat{a}) - E(\hat{b})x_i = a + bx_i - a - bx_i = 0$$

- It can be shown that

$$\text{var}(r_i) = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\sigma^2$$

Standardized residuals

- The standardized residual is defined as the residual divided by its standard error

$$\frac{r_i}{\hat{\sigma}\sqrt{\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$

- Standardized residuals immediately tell us how many estimated standard deviations any point is away from the fitted regression model.

- For large samples these *standardized* residuals will approximately have a standardized Normal distribution and hence values larger than 2 or smaller than -2 are considered to be *unusual*.

Summary

- The estimators $\hat{a}$ and $\hat{b}$ are unbiased with variances

$$\mathrm{var}(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad \mathrm{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$$

  If we assume normal errors, $\hat{a}$ and $\hat{b}$ are normally distributed.

- The residual r_i is normally distributed with mean 0 and variance

$$\mathrm{var}(r_i) = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \sigma^2$$

- The standardized residuals are

$$\frac{r_i}{\hat{\sigma}\sqrt{\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$

  and have a standard normal distribution.

Section 3.5: Confidence intervals and hypothesis tests

Confidence intervals: recap

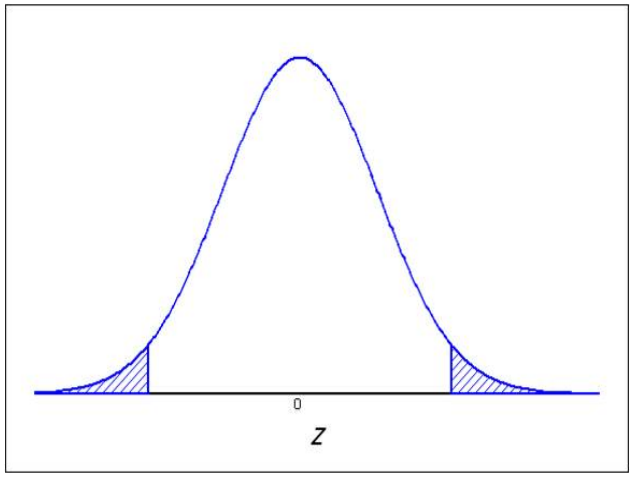
- Fact: Let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. Then

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1)$$

by central limit theorem.

Confidence intervals: recap, cont.

- Let $Z_{0.025}$ and $-Z_{0.025}$ be as follows



where $P(-Z_{0.025} \leq Z \leq Z_{0.025}) = 0.95$.

- The 95% confidence interval for θ is

$$(\hat{\theta} - 1.96\sigma_{\hat{\theta}}\,,\ \hat{\theta} + 1.96\sigma_{\hat{\theta}})$$

Discussion

- We showed in Section 3.4 that, under the normal error assumption

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

- What is the corresponding standard normal distribution?
- What is the corresponding 95% C.I?

Confidence intervals for a

- We showed in Section 3.4 that, under the normal error assumption

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right) .$$

- Hence a 95% confidence interval for a is

$$\left(\hat{a} - 1.96\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\ ,\ \hat{a} + 1.96\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- However, σ is usually unknown.

Confidence intervals for a, using t-distribution

- We are going to replace σ with its estimator $\hat{\sigma}$.
- One obtains

$$\frac{\hat{a} - a}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

where t_{n-2} is the t-distribution with $n-2$ degrees of freedom.

Confidence intervals for a, using t-distribution, cont.

- Hence a 95% confidence interval for a is

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\ ,\ \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- That is, replacing σ^2 by the unbiased estimate $\hat{\sigma}^2$ merely changes the normal percentage point to the corresponding t percentage point.

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly). Let X = first pulse rate, Y = second pulse rate,

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

We have $\bar{x} = 80.909$, $\bar{y} = 112.828$, $S_{xx} = 1770.909$, $S_{yy} = 1645.636$, $S_{xy} = 995.818$.

Example: Pulse data

- In this example, $n = 11$ and we calculate $\hat{a} = 67.323$, $\hat{\sigma} = 10.983$, $\bar{x} = 80.909$, $S_{xx} = 1770.909$.
- From the t-distribution table, $t_{0.025,9} = 2.262$.
- Hence a 95% C.I. for a is

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\ ,\ \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

$$= 67.32 \pm 2.262 \times 10.98\sqrt{\frac{1}{11} + \frac{80.91^2}{1770.91}} = (18.97, 115.67)\ .$$

Hypothesis test for a

- We want to test the hypothesis $H_0 : a = a_0$ against $H_1 : a \neq a_0$.
- If H_0 is true, then

$$t = \frac{\hat{a} - a_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

  is an observation from a t_{n-2} distribution.
- We choose significance level 5%. The corresponding critical region for this two-sided test is

$$(-\infty, -t_{0.025,n-2}) \cup (t_{0.025,n-2}, +\infty)$$

Example, cont.

- We want to test the hypothesis $a = 0$, that is, the regression line goes through the origin.
- As before, $\hat{a} = 67.323$. $\hat{\sigma} = 10.983$, $\bar{x} = 80.909$, $n = 11$, $S_{xx} = 1770.909$.
- The t-statistic is

$$t = \frac{\hat{a} - a_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} = \frac{67.323 - 0}{10.983\sqrt{\frac{1}{11} + \frac{80.909^2}{1770.909}}} = 3.15$$

- With a significance level of 5%, the critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$ which contains 3.15. Hence we reject the null hypothesis.

Inference for b

- Inference for the slope of regression line is very often very useful in practice.
- When testing the hypothesis $H_0 : b = 0$ we want to assess if the explanatory variable X has really an influence on the response Y, or else is not needed.
- If $H_0 : b = 0$ is rejected, we say the explanatory variable is statistically significant to the response Y.

Confidence intervals for b

- We showed in Section 3.4 that, under the normal errors assumption,

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

- Hence a 95% C.I. for b has limits

$$\left(\hat{b} - 1.96\frac{\sigma}{\sqrt{S_{xx}}}\ ,\ \hat{b} + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

- As in the case of $\hat{a}$, there is the problem that σ is in general unknown.

Confidence intervals for b, cont.

- We substitute σ with its estimate $\hat{\sigma}$.
- Similarly to the case for a, this changes the limits of the C.I. from the normal percentage point $Z_{0.025} = 1.96$ to the t-percentage point $t_{0.025,n-2}$.
- Hence a 95% C.I. for b is

$$\left(\hat{b} - t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \ , \ \hat{b} + t_{0.025,n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)$$

Hypothesis test for b

- We want to test the hypothesis $H_0 : b = b_0$ against $H_1 : b \neq b_0$.
- If H_0 is true, then

$$t = \frac{\hat{b} - b_0}{\hat{\sigma}/\sqrt{S_{xx}}}$$

  is an observation from a t_{n-2} distribution.
- We choose significance level 5%. The corresponding critical region for this two-sided test is

$$(-\infty, -t_{0.025,n-2}) \cup (t_{0.025,n-2}, +\infty)$$

Example: Pulse Data

- For Pulse Data we have:

  $n = 11,\ \hat{b} = 0.5623,\ \hat{\sigma} = 10.983,\ S_{xx} = 1770.909.$

- A 95% C.I. for b has limits

$$\left(\hat{b} - t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\ ,\ \hat{b} + t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$
$$= 0.5623 \pm 2.262 \times 10.983\sqrt{\frac{1}{1770.909}}$$
$$= (-0.03, 1.15)$$

- Note: this interval includes 0.

Example: Pulse Data, cont.

- We want to test the hypothesis $b = 0$.
- As before, $\hat{b} = 0.5623$, $\hat{\sigma} = 10.983$, $S_{xx} = 1770.909$.

  We calculate the t-statistic

$$t = \frac{\hat{b} - b_0}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \frac{0.5623\sqrt{1770.909}}{10.983} = 2.15$$

- With a significance level of 5%, the critical region is $(-\infty, -2.262) \cup (2.262, +\infty)$ which doesn't contain 2.15. Hence we accept the null hypothesis $b = 0$.
- We conclude that the explanatory variable X has no influence on the response Y.

Confidence intervals for σ^2

- From Section 3.3,

$$\hat{\sigma}^2 = \frac{RSS}{(n-2)}, \qquad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

- We have

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

- Hence,

$$P\left(\chi^2_{0.975,n-2} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi^2_{0.025,n-2}\right) = 0.95$$

Confidence intervals for σ^2, cont.

- The 95% C.I. for σ^2 is

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}}\right)$$

- The corresponding C.I. for σ is

$$\left(\hat{\sigma}\sqrt{\frac{n-2}{\chi^2_{0.025,n-2}}}, \hat{\sigma}\sqrt{\frac{n-2}{\chi^2_{0.975,n-2}}}\right)$$

Example: Pulse data

- In this case, $n = 11$, $RSS = (n-2)\hat{\sigma}^2 = 1085.67$, $\chi^2_{0.025,9} = 19.02$, $\chi^2_{0.975,9} = 2.7$.
- Hence a 95% C.I. for σ^2 is given by

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}}\right) = \left(\frac{1085.67}{19.02}, \frac{1085.67}{2.7}\right) = (57.04, 401.87)$$

- The corresponding C.I. for σ is

$$(\sqrt{57.04}, \sqrt{401.87}) = (7.55, 20.05)$$

Summary

- We established the limits of 95% C.I. for a, b and σ^2 to be

$$\left(\hat{a} - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\ ,\ \hat{a} + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

$$\left(\hat{b} - t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\ ,\ \hat{b} + t_{0.025,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.025,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{0.975,n-2}}\right)$$

- The corresponding hypothesis tests require the critical value from the t-distribution.
- Of particular interest is the hypothesis test for $H_0 : b = 0$ which assesses the need for the explanatory variable.

Section 3.6:
Confidence and prediction intervals

The need of confidence intervals for predictions

- Given a value x_0 of the predictor, the predicted response is $\hat{y}_0 = \hat{a} + \hat{b}x_0$.
- We want to assess the uncertainty in this prediction, as often needed in decision making.
- If the prediction has a wide C.I., we need to allow for outcomes far from the point estimate.

Different types of prediction

- We consider two different types of prediction:
  1. Predictions of future mean response.
  2. Prediction of future observation.

Discussion

Suppose we build a simple linear regression model where Y is the selling price of houses in a given area and X is the number of bedrooms. We can make two kinds of predictions for houses with 4 bedrooms:

1. We want to know what would a house with 4 bedrooms sell for on average.

2. Suppose that a specific house with 4 bedrooms comes on the market, we want to predict its price.

Question: for which type of prediction of the selling price do you think the error is greater?

Answer to discussion question

Suppose we build a simple linear regression model where Y is the selling price of homes in a given area and X is the number of bedrooms. We can make two kinds of predictions for a given x_0:

1. We want to know the average price of houses with characteristic x_0. This average selling price is $a + bx_0$ since $E(\varepsilon) = 0$, and this is predicted by $\hat{a} + \hat{b}x_0$, so only the variances of $\hat{a}$ and $\hat{b}$ need to be taken into account.

2. Suppose that a specific house comes on the market with characteristic x_0. Its selling price is $a + bx_0 + \varepsilon$.
   The predicted price of this house is still $\hat{a} + \hat{b}x_0$, but in assessing the variance of this prediction, we must include the variance of ε.

Confidence intervals for expectations

- When $x = x_0$, then $E(y_0) = a + bx_0$, which has unbiased estimator

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

- Recalling that $\hat{a} = \bar{y} - \hat{b}\bar{x}$, we obtain

$$\hat{y}_0 = \bar{y} - \hat{b}\bar{x} + \hat{b}x_0 = \bar{y} + \hat{b}(x_0 - \bar{x})$$

- We now use this expression to calculate the variance of $\hat{y}_0$.

Confidence intervals for expectation, cont.

- From previous slide:

$$\hat{y}_0 = \bar{y} + \hat{b}(x_0 - \bar{x})$$

- Therefore

$$\mathrm{var}(\hat{y}_0) = \mathrm{var}(\bar{y}) + 2(x_0 - \bar{x})\,\mathrm{cov}(\bar{y}, \hat{b}) + (x_0 - \bar{x})^2\,\mathrm{var}(\hat{b})$$

Confidence intervals for expectation, cont.

- Recall $\text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}$ and $\text{var}(\bar{y}) = \frac{\sigma^2}{n}$, and

$$\text{cov}(\bar{y}, \hat{b}) = \text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} c_i \,\text{var}(y_i)$$
$$= \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

- Therefore,

$$\text{var}(\hat{y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

- Question: When does $\text{var}(\hat{y}_0)$ obtain the minimum?

Remark

- We calculated

$$\mathrm{var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

- We see that $\mathrm{var}(\hat{y}_0)$ is minimal at $x_0 = \bar{x}$ and increases as x_0 moves away from $\bar{x}$.
- Thus the regression line is best estimated at the average value of the predictor and poorly estimated for extreme values of the predictor.

Confidence intervals for expectation, cont.

- We established that $\hat{y}_0 = \hat{a} + \hat{b}x_0$ is normally distributed with mean $a + bx_0$ and variance $\mathrm{var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$.

- It follows that

$$\frac{\hat{y}_0 - a - bx_0}{\sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

has a ..... distribution.

Fill in the gaps

Confidence intervals for expectation, cont.

- We substitute σ^2 with its estimate $\hat{\sigma}^2$. It can be shown that

$$\frac{\hat{y}_0 - a - bx_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}}$$

has a ..... distribution.

- A 95% C.I. for $a + bx_0$ has therefore limits

$$\left(\hat{a} + \hat{b}x_0 - t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\right.$$
$$\left.\hat{a} + \hat{b}x_0 + t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

Fill in the gaps

Example: Acid Rain

- Pure water has a pH of 7.0. Normal rain is slightly acidic because carbon dioxide (CO2) dissolves into it forming weak carbonic acid, with pH of approximately 5.5.
- A laboratory is equipped to make pH measurements.
- The following slide shows the results of measurements of pH made by a laboratory on 18 samples of water. Each sample is prepared by a chemical engineer and have a *known* level of pH.
- We want to compare the "True" values (x) with the laboratory measurements (y).

Example: Acid Rain, cont.

- The following table contains the data of the experiment

Sample	True pH	Measur. pH	Sample	True pH	Measur. pH
1	7.50	7.37	10	5.74	5.52
2	6.20	6.36	11	4.94	4.71
3	6.20	6.12	12	7.10	7.12
4	7.10	6.96	13	6.55	6.43
5	6.20	6.25	14	7.20	6.76
6	5.95	5.93	15	5.34	5.11
7	6.55	6.38	16	9.50	8.50
8	6.55	6.34	17	8.00	8.50
9	5.34	4.93	18	10.0	9.50

- For these data we calculate:
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $\bar{y} = 6.599, \quad S_{xx} = 30.626, \quad S_{yy} = 27.400, \quad S_{xy} = 28.205,$
  $RSS = 1.4254, \quad n = 18.$

Example: Acid Rain, cont.

- From previous slide
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $S_{xx} = 30.626,\ n = 18.$

- Suppose we want the 95% C.I. at $x_0 = 5$. Since $t_{0.025,16} = 2.12$, we have

$$\begin{aligned}
&\hat{a} + \hat{b}x_0 \pm t_{0.025,16}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \\
&\quad = 0.3595 + 0.921 \times 5 \pm 2.12 \times 0.29848\sqrt{\frac{1}{18} + \frac{(5 - 6.776)^2}{30.626}} \\
&\quad = (4.713, 5.216)
\end{aligned}$$

Prediction intervals for new observation

- If y_0 is the actual value, then

$$y_0 = a + bx_0 + \varepsilon_0 \ ,$$

where ε_0 is some unknown error and $\varepsilon_0 \sim N(0, \sigma^2)$.

- A point estimate of y_0 is $\hat{y}_0 = \hat{a} + \hat{b}x_0$ and

$$\hat{y}_0 - y_0 = \hat{y}_0 - a - bx_0 - \varepsilon_0$$

- It follows that $E(\hat{y}_0 - y_0) = 0$.

Prediction intervals for new observation, cont.

- It can be shown that

$$\mathrm{var}(\hat{y}_0 - y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

- In conclusion, $\hat{y}_0 - y_0$ is normally distributed with mean 0 and variance $\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$.
- It follows that

$$\frac{\hat{y}_0 - y_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0, 1)$$

- However, σ is in general unknown.

Prediction intervals for new observation, cont.

- We substitute the unknown σ with its estimate $\hat{\sigma}$ and find

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

- The 95% prediction interval for y_0 has limits

$$\left(\hat{a} + \hat{b}x_0 - t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\right.$$
$$\left.\hat{a} + \hat{b}x_0 + t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

Example: Acid Rain

- We want to calculate the 95% prediction interval for Y at the value $x = 5$ for the Acid Rain.

- We previously calculated $\hat{a} = 0.3595, \quad \hat{b} = 0.921,$ $\hat{\sigma} = 0.29848, \quad \bar{x} = 6.776, \quad S_{xx} = 30.626,\ n = 18,$ $t_{0.025,16} = 2.12.$

- Therefore

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,16}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 0.3595 + 0.921 \times 5 \pm 2.12 \times 0.29848\sqrt{1 + \frac{1}{18} + \frac{(5 - 6.776)^2}{30.626}}$$

$$= (4.283, 5.646)$$

Confidence interval versus prediction interval

- We showed that the C.I. for the mean response when $x = x_0$ is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \ ,$$

  while the prediction interval for a new observation at $x = x_0$ is

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \ .$$

- Which interval is wider?

Summary

- We gave the notions of CI for mean response and prediction interval for new observations, and illustrated their differences.
- We gave the corresponding formulas

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \ ,$$

$$\hat{a} + \hat{b}x_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \ .$$

- We illustrated this on the Acid Rain example.

Chapter 3: Simple Linear Regression, Part 2

Dr.Ting Wei

Section 3.7:
Checking the model assumptions

Using R software

- The document 'R Instruction' for guidance.
- You can get R from http://www.r-project.org/. Versions are provided for all platforms.
- RStudio is a very useful interface that allows the user to run R in a more user-friendly environment. It is open-source (i.e. free) and available at http://www.rstudio.com/.
- RStudio Cloud is a cloud-based solution that allows anyone to access R online. Individual users, instructors and students only need a browser to do, share, teach and learn data science, which is available at https://rstudio.cloud/.

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly). Let X = first pulse rate, Y = second pulse rate,

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

We have $\bar{x} = 80.909$, $\bar{y} = 112.828$, $S_{xx} = 1770.909$, $S_{yy} = 1645.636$, $S_{xy} = 995.818$.
$\hat{b} = 0.5623$, $\hat{a} = 67.32$, $\hat{\sigma} = 10.983$.

The function `lm`

- The function `lm` is the main R function for linear regression.
- For the Pulse data, the regression of 'second pulse rate' (response variable) against 'first pulse rate' (explanatory variable) can be done with the following code:

```
> x=c(96,62,78,82,100,68,96,78,88,62,80)
> y=c(140,100,104,100,115,112,116,118,110,98,128)
> modPulse=lm(y ~ x)
```

The function lm, cont.

- Main results

```
> summary(modPulse)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-13.432  -6.994  -4.185   6.629  18.696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.3214    21.3748   3.150   0.0117 *
x             0.5623     0.2610   2.155   0.0596 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.98 on 9 degrees of freedom
Multiple R-squared: 0.3403,     Adjusted R-squared: 0.267
F-statistic: 4.642 on 1 and 9 DF,  p-value: 0.05959
```

The function `lm`: commented results.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.3214    21.3748   3.150   0.0117 *
x             0.5623     0.2610   2.155   0.0596 .

Residual standard error: 10.98 on 9 degrees of freedom
```

- Column 1 of Coefficients table gives $\hat{a}$ and $\hat{b}$ estimates respectively.
- Column 2 of Coefficients table gives $\sqrt{\text{var}(\hat{a})}$ and $\sqrt{\text{var}(\hat{b})}$ estimates respectively.

The function `lm`: commented results.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.3214    21.3748   3.150   0.0117 *
x             0.5623     0.2610   2.155   0.0596 .

Residual standard error: 10.98 on 9 degrees of freedom
```

- Column 3 gives the estimates $\frac{\hat{a}}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$ and $\frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} \sim t_{n-2}$ respectively.
- Column 4 gives the p-values corresponding to the t_{n-2} values estimated in column 3.
- `Residual standard error: 10.98 on 9 degrees of freedom` means $\hat{\sigma} = 10.98$, i.e. $\hat{\sigma}^2 = 120.6$.

Fitted values

- Recall that for the example: $\hat{y} = \hat{a} + \hat{b}x = 67.3214 + 0.5623x$. To extract the fitted values from the model,use the command

```
> modPulse$fitted.values
```

Note the symbol $ after the name of the model.

```
       1        2        3        4        5        6
121.3041 102.1852 111.1823 113.4316 123.5534 105.5591
       7        8        9       10       11
121.3041 111.1823 116.8055 102.1852 112.3070
```

Residuals

- To compute residuals two different R commands can be used:

```
> modPulse$residuals

> residuals(modPulse)
```

```
       1          2          3          4          5
18.695893  -2.185216  -7.182341 -13.431622  -8.553388
       6          7          8          9         10
 6.440862  -5.304107   6.817659  -6.805544  -4.185216
      11
15.693018
```

Standardized residuals

- Recall that the standardized residuals are

$$\frac{r_i}{\hat{\sigma}\sqrt{\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$

- To compute residuals, the R commands is:

```
> rstandard(modPulse)
```

```
        1          2          3          4          5
1.9267909 -0.2365920 -0.6876693 -1.2830932 -0.9286353
        6          7          8          9         10
0.6495922 -0.5466390  0.6527531 -0.6602709 -0.4531308
       11
1.4989500
```

Example: Acid Rain

- The following table contains the data of the experiment

Sample	True pH (x)	Measur. pH (y)	Sample	True pH (x)	Measur. pH (y)
1	7.50	7.37	10	5.74	5.52
2	6.20	6.36	11	4.94	4.71
3	6.20	6.12	12	7.10	7.12
4	7.10	6.96	13	6.55	6.43
5	6.20	6.25	14	7.20	6.76
6	5.95	5.93	15	5.34	5.11
7	6.55	6.38	16	9.50	8.50
8	6.55	6.34	17	8.00	8.50
9	5.34	4.93	18	10.0	9.50

- For these data we calculate:
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $\bar{y} = 6.599, \quad S_{xx} = 30.626, \quad S_{yy} = 27.400, \quad S_{xy} = 28.205.$

Example: Acid Rain, cont.

- You can store the dataset in a '.txt' file and let R read it:

```
> T=read.table("AcidRain.txt")
> x=T[,1]
> y=T[,2]
```

- We then fit the simple linear regression model:

```
> modAcid=lm(y ~ x)
> summary(modAcid)
```

Example: Acid Rain, cont.

- Main results

```
> summary(modAcid)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60851 -0.15697  0.01328  0.10025  0.77291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35953    0.37215   0.966    0.348
x            0.92095    0.05393  17.075 1.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2985 on 16 degrees of freedom
Multiple R-squared:  0.948,Adjusted R-squared:  0.9447
F-statistic: 291.6 on 1 and 16 DF,  p-value: 1.079e-11
```

Recap: model assumptions

- When we use a simple linear regression model, we implicitly make a series of assumptions.

  More precisely:

  (i) model equation is a linear function of x plus random error;

  (ii) errors have constant variance;

  (iii) errors are independent;

  (iv) errors are normally distributed.

The need for model diagnostics

- Question: given a set of real data, are these assumptions valid?
- It makes sense to base inferences or conclusions only on valid models.
- Looking only at numerical regression output may lead to very misleading conclusions about the data, and lead to adopting the wrong model.
- Model diagnostics is a series of tools to check each assumption.

A motivating example

- Consider four data sets constructed by Anscombe (1973).

case	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

A motivating example, cont.

- Data Set 1:

```
> m1 <- lm(y1 ~ x1)
> summary(m1)

Call:
lm(formula = y1 ~ x1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665,     Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

A motivating example, cont.

- Data Set 2:

```
> m2 <- lm(y2 ~ x2)
> summary(m2)

Call:
lm(formula = y2 ~ x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662,     Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

A motivating example, cont.

- Data Set 3:

```
> m3 <- lm(y3 ~ x3)
> summary(m3)

Call:
lm(formula = y3 ~ x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663,     Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

A motivating example, cont.

- Data Set 4:

```
> m4 <- lm(y4 ~ x4)
> summary(m4)

Call:
lm(formula = y4 ~ x4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667,     Adjusted R-squared: 0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

A motivating example, cont.

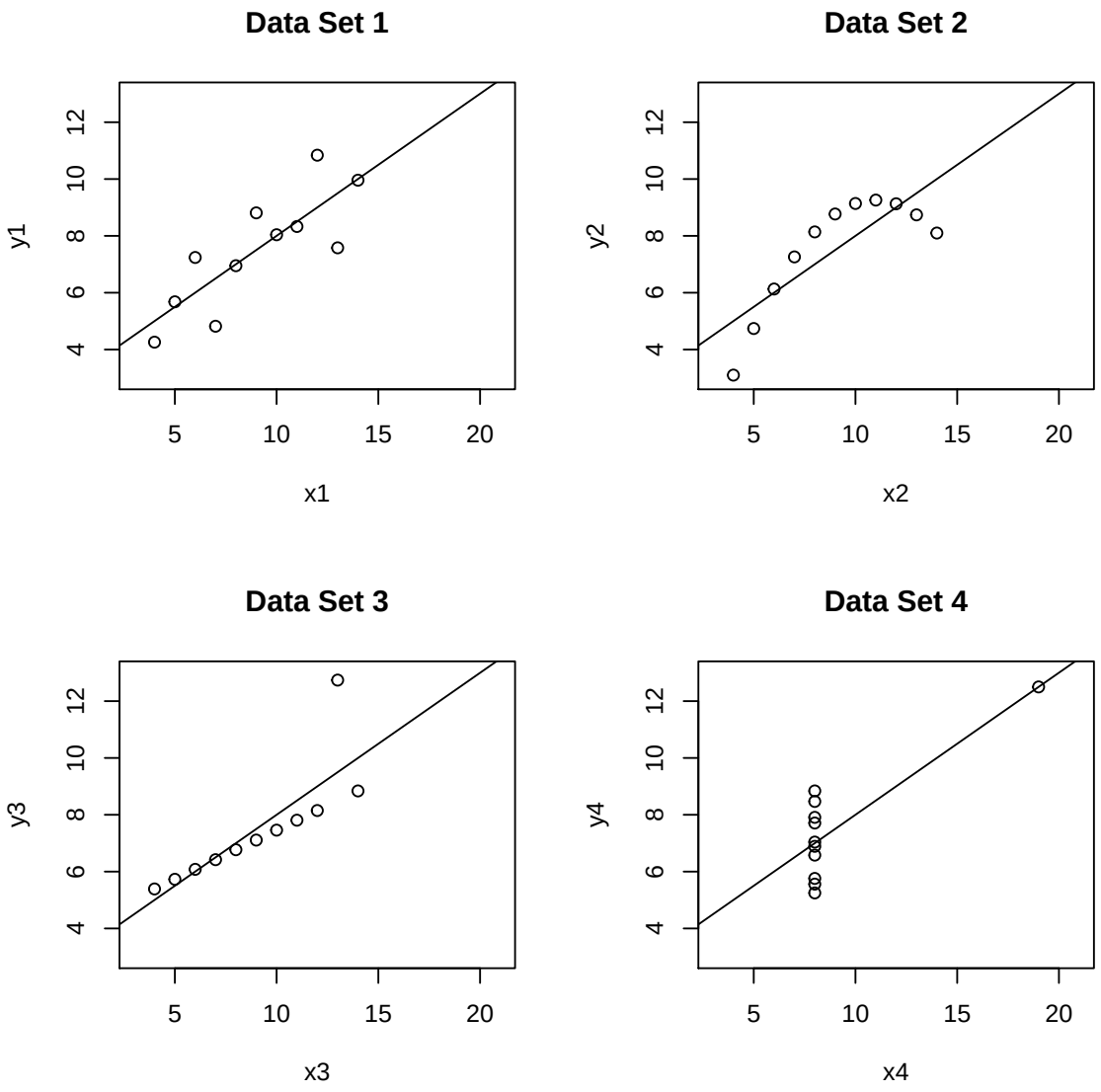
- When a regression model is fitted to data sets 1, 2, 3, 4 in each case the fitted regression model is:

$$\hat{y} = 3 + 0.5x$$

- The regression output for the four data sets is identical (to two decimal places).

A motivating example, cont.

- The following are plots of Anscombe's four data Sets



Checking the structure of the model

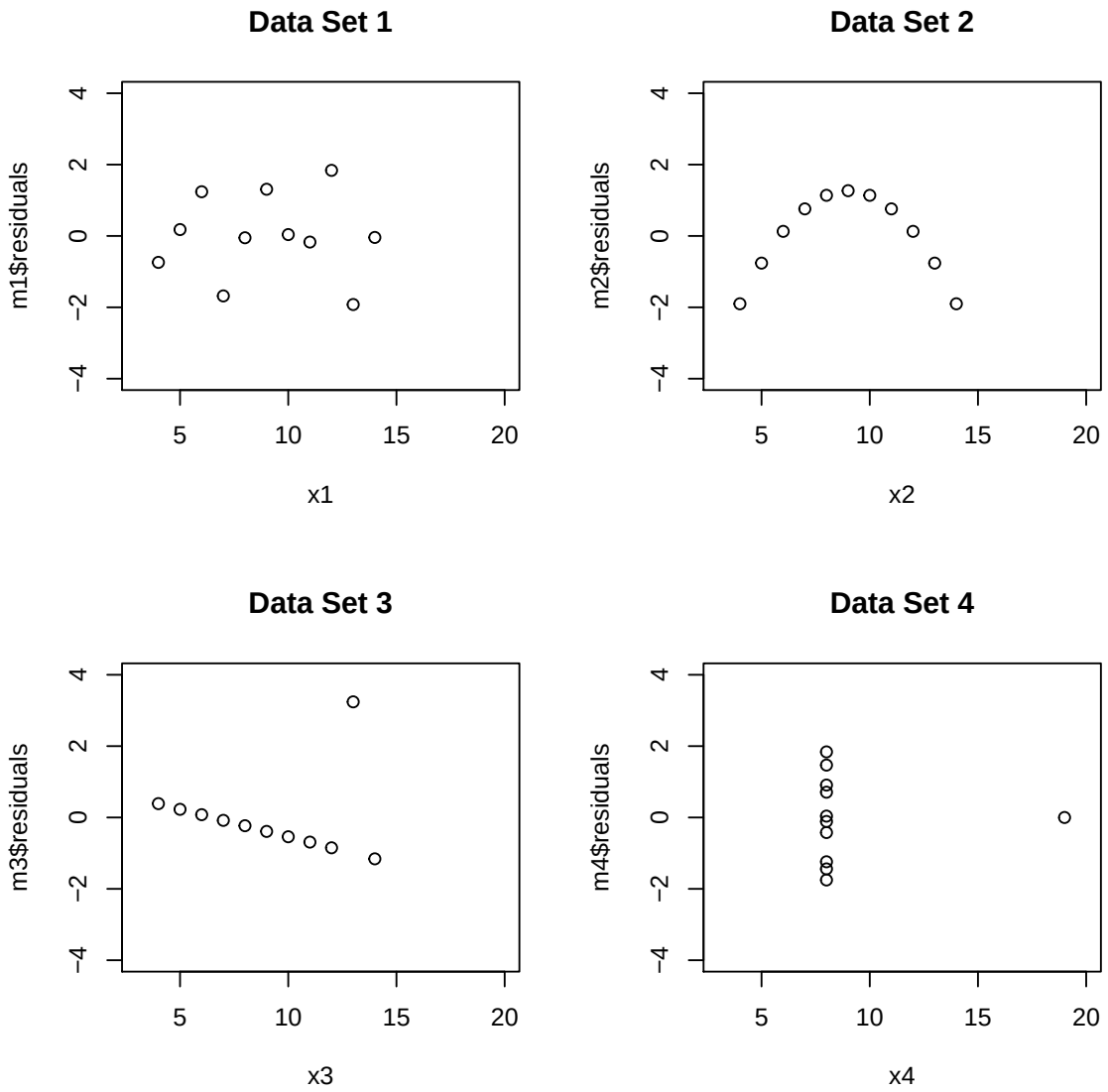
- We can also plot the residuals against x and look for patterns.
- If no pattern is found, then the structure of the model is adequate.
- In fact, if $y_i = a + bx_i + \varepsilon_i$ is correct, and the least square estimates $\hat{a}$, $\hat{b}$ are close to a, b, we have

$$\begin{aligned} r_i = y_i - \hat{y}_i &= a + bx_i + \varepsilon_i - \hat{a} - \hat{b}x_i \\ &= (a - \hat{a}) + (b - \hat{b})x_i + \varepsilon_i \approx \varepsilon_i \end{aligned}$$

that is, the residuals resemble random errors.

Checking the structure of the model, cont.

- The plots of residuals r_i against x.

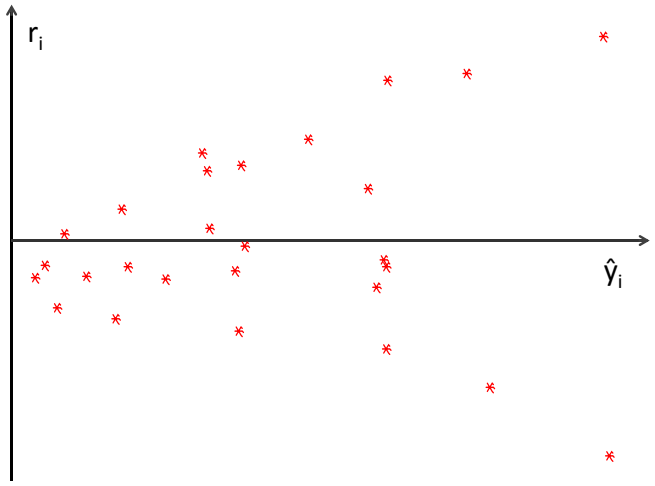


Checking the error assumptions

- We want to check constant variance, independence and normality of the errors.
- The errors are not observable, but we can examine the residuals.
- As explained in a previous slide, we can reasonably assume that $r_i \approx \varepsilon_i$.

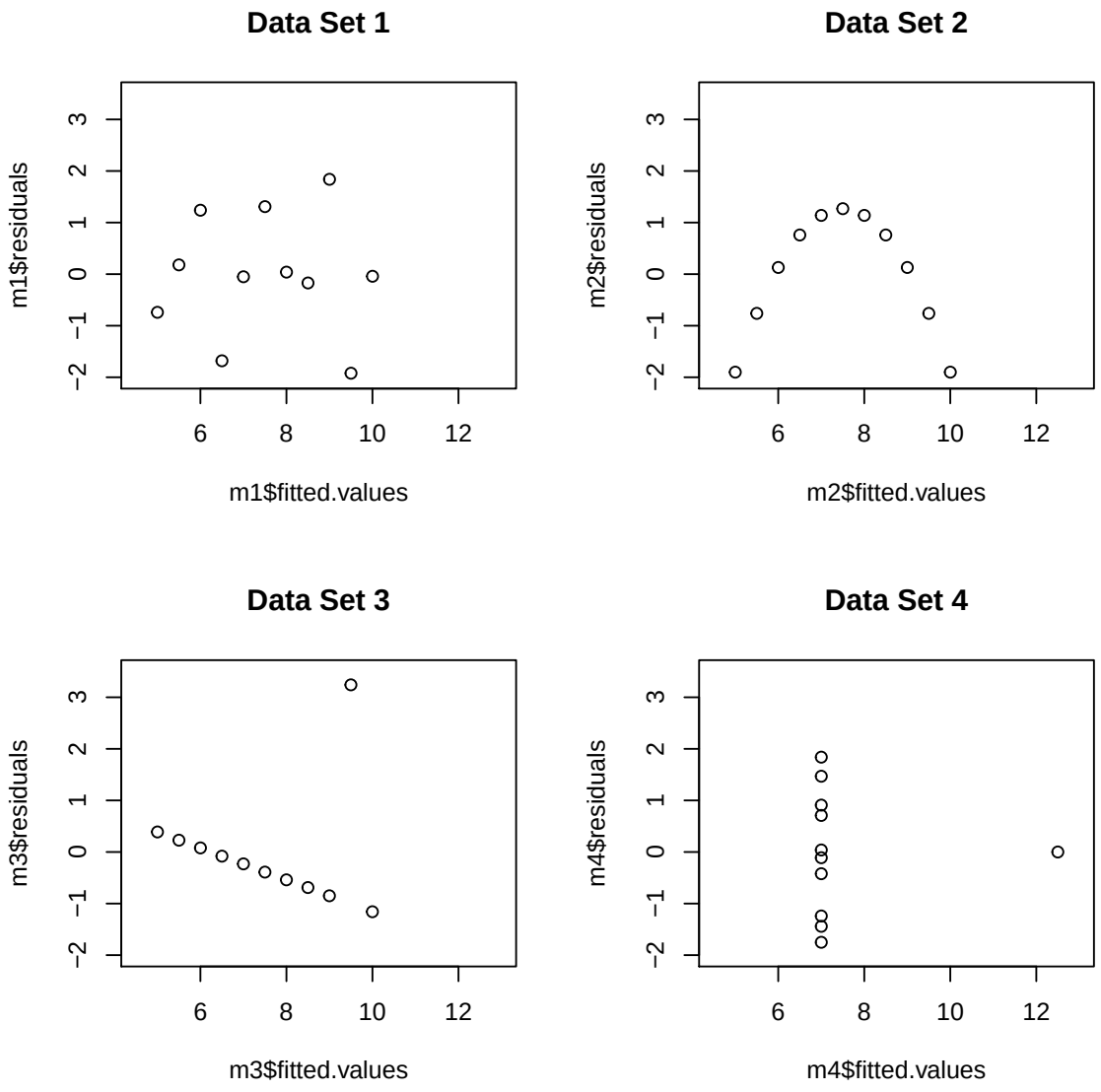
Checking the error assumptions, cont.

- Plot the residuals r_i against the fitted values $\hat{y}_i$.
- If the model is correct then these plots should show no relationship between r_i and $\hat{y}_i$.
- If all is well, we should see constant variance in the vertical direction and the scatter should be approximately symmetric about zero (x-axis).
- Especially, a 'fan'-shaped plot of residuals such as the following suggests non-constant variance.



A motivating example, cont.

- The plots of residuals r_i against the fitted value $\hat{y}_i$.

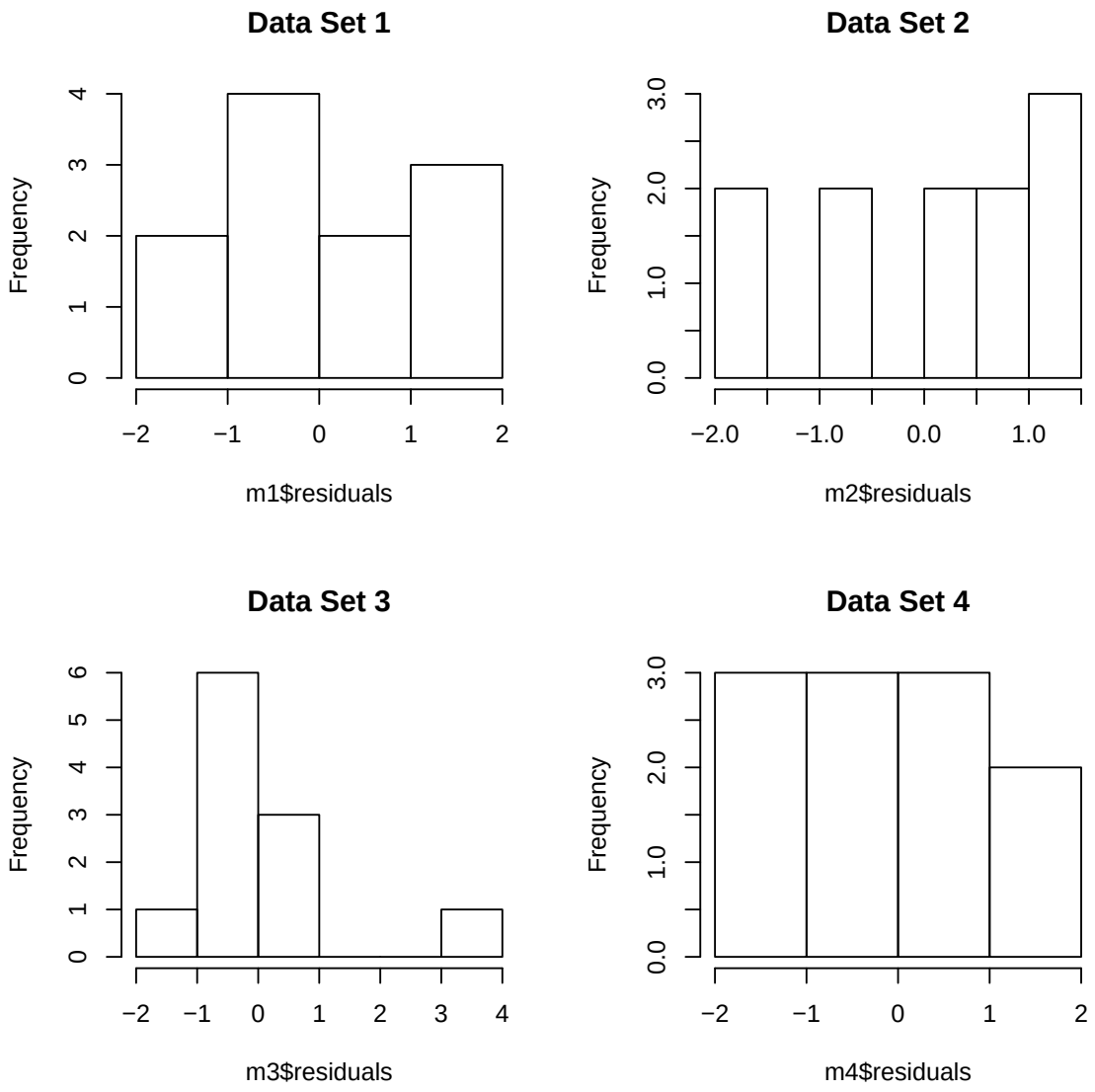


Checking for normality

- The hypothesis tests and confidence intervals we used are based on the assumption of normal errors.
- To check for normality, we can histogram the residuals.
- A more sophisticated approach is provided by the probability plot, which is usually called quantile-quantile plot (Q-Q plot).
- A plot of residuals r_i against $\Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$ should be approximately a straight line with slope σ.

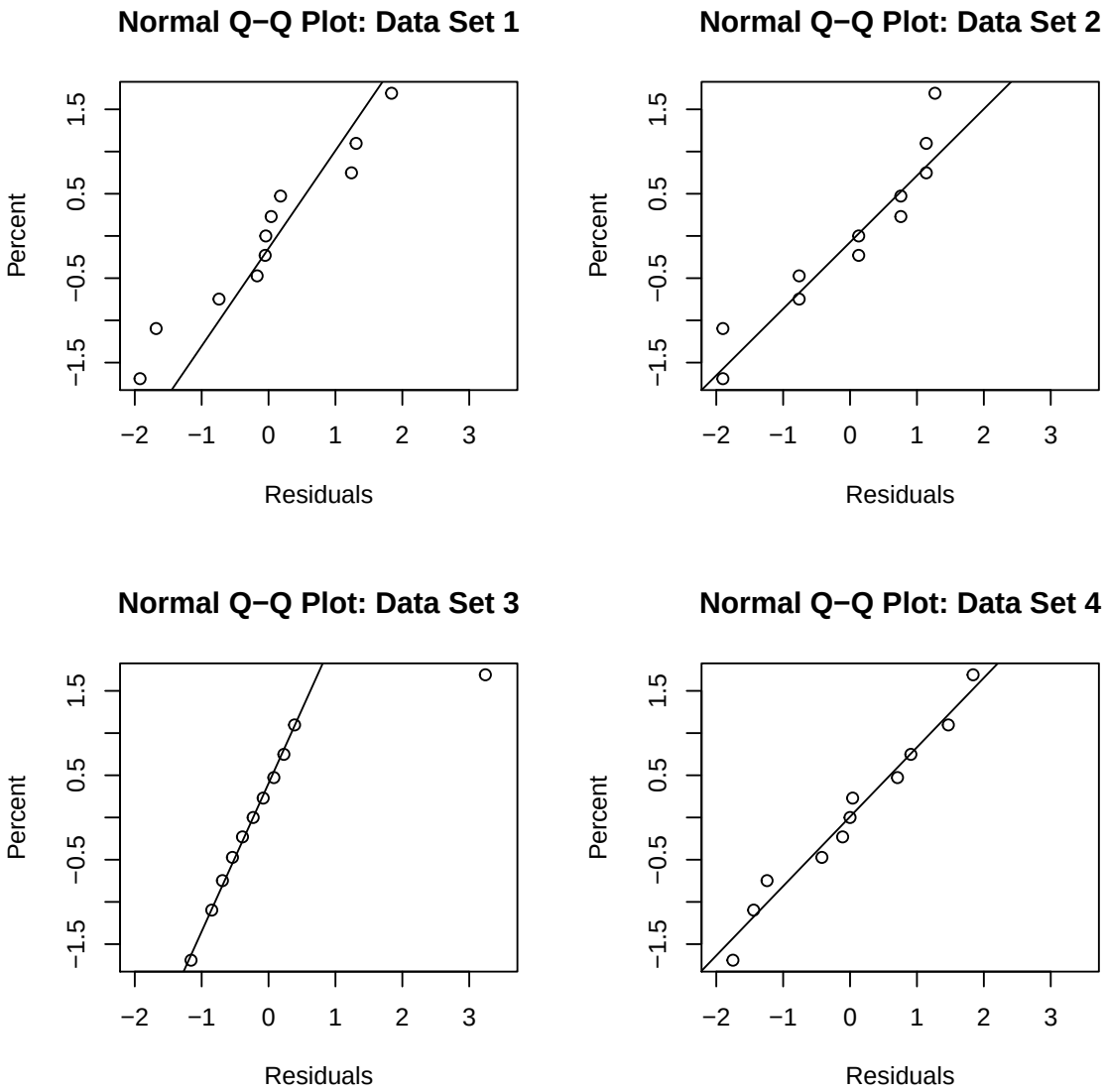
Checking for normality, cont.

- The histogram plots of residuals.



Checking for normality, cont.

- The Q-Q plots of residual probabilities.



Summary

- We discussed the need for model diagnostic by illustrating the Anscombe data set.
- We discussed how to check the different model assumptions:
  - Checking the structure of the model (plot y against x and residuals against x).
  - Checking assumption of constant variance (plot residuals against fitted values).
  - Checking normality of the residuals (normal probability plots of residuals).

Section 3.8: Analysis of variance

Discussion

What is the meaning of the following sums of squares?

- $SST = S_{yy} = \sum_i (y_i - \bar{y})^2$,
- $RSS = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $SST - RSS$

Answer to discussion question

- The total variation

$$SST = S_{yy} = \sum_i (y_i - \bar{y})^2$$

represents the total variation in y about its mean.

- The residual sum of squares, also called error variation

$$RSS = SSE = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

represents the residual variation in y after fitting the regression line.

- The model variation

$$SSM = SST - SSE$$

represents the amount of variation in y explained by x.

Distributions of sums of squares in the simple linear regression model

- As discussed in Section 3.3,

$$\frac{SSE}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

- Recall that

$$RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \qquad \hat{b} = \frac{S_{xy}}{S_{xx}}$$

- Therefore,

$$SSM = SST - RSS = S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \hat{b}^2 S_{xx}$$

Distributions of sums of squares in the simple linear regression model, cont.

- From previous slide: $SSM = (\hat{b})^2 S_{xx}$.
- From Section 3.4, $\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$, we deduce

$$E(\hat{b}^2) = \text{var}(\hat{b}) + \left(E(\hat{b})\right)^2 = \frac{\sigma^2}{S_{xx}} + b^2$$

- Therefore,

$$\begin{aligned} E(SSM) &= E(\hat{b}^2 S_{xx}) = E(\hat{b}^2) S_{xx} \\ &= \left(\frac{\sigma^2}{S_{xx}} + b^2\right) S_{xx} = \sigma^2 + b^2 S_{xx} \end{aligned}$$

Distributions of sums of squares in the simple linear regression model, cont.

- We calculated: $E(SSM) = \sigma^2 + b^2 S_{xx}$.
- Hence,

$$E\left(\frac{SSM}{\sigma^2}\right) = \begin{cases} = 1, & \text{if } b = 0 \\ > 1, & \text{if } b \neq 0 \end{cases}$$

- If the hypothesis $H_0 : b = 0$ is true, then $\hat{b} \sim N\left(0, \frac{\sigma^2}{S_{xx}}\right)$ so that

$$\frac{\hat{b}\sqrt{S_{xx}}}{\sigma} \sim N(0, 1)$$

which implies

$$\frac{SSM}{\sigma^2} = \frac{\hat{b}^2 S_{xx}}{\sigma^2} \sim \chi^2_1$$

Distributions of sums of squares in the simple linear regression model, cont.

- In conclusion we showed that, if the hypothesis $b = 0$ is true

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}, \qquad \frac{SSM}{\sigma^2} \sim \chi^2_1$$

- Further, it can be shown that $\frac{SSE}{\sigma^2}$ and $\frac{SSM}{\sigma^2}$ are independent.

- It follows that

$$\frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSM}{\sigma^2} = \chi^2_{n-2} + \chi^2_1 = \chi^2_{n-1}$$

Mean squares

- Dividing the sums of squares by their degrees of freedom we obtain the corresponding 'mean squares'.
- The mean squared error is

$$MSE = \frac{SSE}{n-2} \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2}$$

- The mean squared model variation under $H_0 : b = 0$ is

$$MSM = \frac{SSM}{1} \sim \sigma^2 \chi^2_1$$

The F-distribution

- If $X_1 \sim \chi^2_m$, $X_2 \sim \chi^2_n$ are two independent random variables then

$$Y = \frac{X_1/m}{X_2/n}$$

  is said to have an F-distribution with (m, n) degrees of freedom and write $Y \sim F_{m,n}$.

- Please note: $F_{m,n} \neq F_{n,m}$.

- In the special case $m = 1$ we have

$$F_{1,n} = \frac{\chi^2_1}{\chi^2_n/n} = \frac{[N(0,1)]^2}{[\sqrt{\chi^2_n/n}]^2} = t^2_n$$

The F statistic

- From previous slide, under the hypothesis $H_0 : b = 0$,

$$MSE = \frac{\sigma^2 \chi^2_{n-2}}{n-2}, \qquad MSM = \sigma^2 \chi^2_1$$

- Hence, if $H_0 : b = 0$ is true, the ratio

$$F = \frac{MSM}{MSE} \sim \frac{\sigma^2 \chi^2_1}{\sigma^2 \chi^2_{n-2}/(n-2)} = F_{1,n-2}$$

has a $F_{1,n-2}$ distribution.

The F statistic

- This can be used to test the hypothesis $H_0 : b = 0$. For a 5% significance level, we reject H_0 if F exceeds the critical value of $F_{1,n-2}$, which can be found from the F-distribution table corresponding to 5% points.
- Alternatively, you can find the p-value $P(> F)$ (upper tail probability). If p-value< 0.05, we reject $H_0 : b = 0$.

Relation between F-statistic and t-statistic

- In Section 3.5 we also tested the null hypothesis $H_0 : b = 0$ using the t-distribution:

$$t = \frac{\hat{b}}{SE(\hat{b})} = \frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} \sim t_{n-2}$$

- This is equivalent to the test described above because, recalling that $SSM = (\hat{b})^2 S_{xx}$ and $\hat{\sigma}^2 = \frac{RSS}{(n-2)} = MSE$,

$$t^2 = \frac{\hat{b}^2 S_{xx}}{\hat{\sigma}^2} = \frac{MSM}{MSE} = F$$

Example: Pulse Data

- The following data gives results from 11 people in three columns: first pulse rate (at rest), second pulse rate (after exercise) and smoking indicator (1=smokes regularly, 2=does not smoke regularly). Let X = first pulse rate, Y = second pulse rate,

First pulse rate	Second pulse rate	Smoking indicator
96	140	2
62	100	2
78	104	1
82	100	2
100	115	1
68	112	2
96	116	2
78	118	2
88	110	1
62	98	1
80	128	2

We have $\bar{x} = 80.909$, $\bar{y} = 112.828$, $S_{xx} = 1770.909$, $S_{yy} = 1645.636$, $S_{xy} = 995.818$.

Example: F statistic for Pulse Data

- We have

$$SSE = RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 1645.636 - \frac{995.818^2}{1770.909} = 1085.67$$

$$MSE = \frac{SSE}{n-2} = 1085.67/9 = 120.63$$

$$MSM = \frac{SSM}{1} = \frac{S_{xy}^2}{S_{xx}} = \frac{995.818^2}{1770.909} = 559.97$$

$$F = \frac{MSM}{MSE} = \frac{559.97}{120.63} = 4.64 \sim F_{1,9}$$

$$\text{Critical region} = (5.117, +\infty)$$

  Hence we accept the null hypothesis.

- Note: the same conclusion was reached in Section 3.5 using the t-test for 2.15 (to be compared with t_9 distribution). The two tests are equivalent since $F_{1,9} = t_9^2$.

Example: Acid Rain, cont.

- The following table contains the data of the experiment

Sample	True pH	Measur. pH	Sample	True pH	Measur. pH
1	7.50	7.37	10	5.74	5.52
2	6.20	6.36	11	4.94	4.71
3	6.20	6.12	12	7.10	7.12
4	7.10	6.96	13	6.55	6.43
5	6.20	6.25	14	7.20	6.76
6	5.95	5.93	15	5.34	5.11
7	6.55	6.38	16	9.50	8.50
8	6.55	6.34	17	8.00	8.50
9	5.34	4.93	18	10.0	9.50

- For these data we calculate:
  $\hat{a} = 0.3595, \quad \hat{b} = 0.921, \quad \hat{\sigma} = 0.29848, \quad \bar{x} = 6.776,$
  $\bar{y} = 6.599, \quad S_{xx} = 30.626, \quad S_{yy} = 27.400, \quad S_{xy} = 28.205,$
  $RSS = 1.4254, \quad n = 18.$

Example: F statistic for Acid Rain

- We have

$$SSE = RSS = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 27.4 - \frac{28.205^2}{30.626} = 1.425$$

$$MSE = \frac{SSE}{n-2} = 1.425/16 = 0.089$$

$$MSM = \frac{SSM}{1} = \frac{S_{xy}^2}{S_{xx}} = \frac{28.205^2}{30.626} = 25.975$$

$$F = \frac{MSM}{MSE} = \frac{25.975}{0.089} = 291.858 \sim F_{1,16}$$

$$\text{Critical region} = (4.494, +\infty)$$

  Hence we reject the null hypothesis $H_0 : b = 0$.

- The same conclusion can be reached using the t-test for $t = \frac{\hat{b}\sqrt{S_{xx}}}{\hat{\sigma}} = \frac{0.921\sqrt{30.626}}{0.29848} = 17.08 \sim t_{n-2}$, that, compared with t_{16} distribution, gives critical region $(-\infty, -2.12) \cup (2.12, +\infty)$. The two tests are equivalent since $F_{1,16} = t_{16}^2$.

The analysis of variance (ANOVA) table

- The analysis of variance table summarizes the main results of simple linear regression. It is an output of every regression software.
- The table has four columns that report the sum of squares (SS), the degrees of freedom of their distributions, the mean squares and the F-statistics.

The analysis of variance (ANOVA) table

The analysis of variance table summarizes the main results of simple linear regression. It is an output of every regression software.

	Source	df	SS	MS	F
SSM	Regression (Model)	1	$\frac{(S_{xy})^2}{S_{xx}} = (\hat{b})^2 S_{xx}$	$\frac{SSM}{1}$	$F_{1,n-2}$
SSE	Error (Residual)	$n-2$	$S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$	$\frac{SSE}{(n-2)}$	

R output for Pulse data

```
> summary(modPulse)

Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.3214    21.3748   3.150   0.0117 *
x             0.5623     0.2610   2.155   0.0596 .

Residual standard error: 10.98 on 9 degrees of freedom
Multiple R-squared: 0.3403,     Adjusted R-squared: 0.267
F-statistic: 4.642 on 1 and 9 DF,  p-value: 0.05959
```

```
> anova(modPulse)

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          1  559.97  559.97   4.642 0.05959 .
Residuals  9 1085.67  120.63
```

R output for Acid Rain

```
> summary(modAcid)
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35953    0.37215   0.966    0.348
x            0.92095    0.05393  17.075 1.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2985 on 16 degrees of freedom
Multiple R-squared:  0.948,Adjusted R-squared:  0.9447
F-statistic: 291.6 on 1 and 16 DF,  p-value: 1.079e-11
```

```
> anova(modAcid)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 25.9749 25.9749  291.56 1.079e-11 ***
Residuals 16  1.4254  0.0891
```

Summary

- The model variation is $SSM = SST - SSE$.
- Dividing the sums of squares by their degrees of freedom gives the mean squares.
- The F statistic is

$$F = \frac{MSM}{MSE} \sim \frac{\sigma^2 \chi_1^2}{\sigma^2 \chi_{n-2}^2/(n-2)} = F_{1,n-2}$$

  has a $F_{1,n-2}$ distribution and is used to test $b = 0$.
- Summary in the ANOVA table.

Section 3.9:
Repeated observations, Part 1

Example: Unemployment rate and GDP

- The following table shows the behaviour of the Gross Domestic Product (GDP) in USA, for the decade 1960-1969, as a function of the unemployment rate percentage. GDP is expressed in percent of the GDP value of the initial year 1960.

Year	Unempl. rate % (x)	GDP % (y)
1	5.5	100
2	5.5	103.9
3	6.7	107.5
4	5.5	115.6
5	5.7	121.9
6	5.2	130.9
7	4.5	141.9
8	3.8	155.6
9	3.8	164.4
10	3.6	179.6

- We calculated:
  $\bar{x} = 4.98, \quad S_{xx} = 9.256, \quad S_{yy} = 6756.361, \quad S_{xy} = -225.954.$
  $\hat{a} = 253.7, \quad \hat{b} = -24.4116, \quad RSS = 1240.457, \quad \hat{\sigma}^2 = 155.057.$

The formal set-up

- Let y_{ij} be the jth observation at x_i, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$:

$$\begin{array}{lll} \text{at } x = x_1 & \text{observations} & y_{11} \ldots y_{1n_1} \\ \text{at } x = x_2 & \text{observations} & y_{21} \ldots y_{2n_2} \\ \vdots & & \\ \text{at } x = x_k & \text{observations} & y_{k1} \ldots y_{kn_k} \end{array}$$

- The total number of observations is $N = \sum_{i=1}^{k} n_i$.

The formal set-up, cont.

- The model is

$$y_{ij} = a + bx_i + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, n_i,$$

where the ε_{ij} are independent random variables with the normal distribution $N(0, \sigma^2)$, and a, b, and σ^2 are unknown parameters, σ^2 does not depend on x.

Total variation

- Let $\bar{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k} n_i\bar{y}_i$ be the overall mean.
- Total sum of squares for the model is

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

Discussion

- What is the mean value of observations at x_i? Recall there are observations $y_{i1} \dots y_{in_i}$ for $x = x_i$.
- What is the variation in the response within the ith observation?

Answer to discussion question

- $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ is the mean value of observations at x_i.
- The variation in the response within the ith observation is given by the sum of squares

$$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Within groups variation

- Define

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- This is called pure error sum of squares and is based on the variation within groups of repeated observations.

- The degrees of freedom associated with SSW is $N - k$ because there are $n_1 - 1$ degrees of freedom associated for the sums of squares for x_1, $n_2 - 1$ degrees of freedom associated for the sums of squares for x_2 and so on.

Between groups variation

- Define

$$SSB = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

- This is based on the variation between groups of repeated observations.

- The degrees of freedom associated with SSB is $k - 1$.

Splitting the total variation

- It can be shown that

$$SST = SSW + SSB$$

- That is, the total variation is split between the variation within each group SSW and the variation between the groups SSB.

Distributions of sums of squares

- It follows from the previous slide that

$$\frac{SSW}{\sigma^2} + \frac{SSB}{\sigma^2} = \frac{SST}{\sigma^2}$$

- From Section 3.8,

$$\frac{SST}{\sigma^2} \sim \chi^2_{N-1}$$

Distributions of sums of squares, cont.

- It can be shown that, if the hypothesis $H_0 : b = 0$ is true, then SSW and SSB are independently distributed and

$$\frac{SSW}{\sigma^2} \sim \chi^2_{N-k} \qquad \frac{SSB}{\sigma^2} \sim \chi^2_{k-1}$$

Mean sums of squares

- We define the mean sums of squares by dividing the sums of squares by their degrees of freedom:

$$MSB = \frac{SSB}{k-1} \qquad MSW = \frac{SSW}{N-k}$$

Mean sums of squares and F-test

- If the hypothesis $H_0 : b = 0$ is true, it follows that

$$\frac{MSB}{MSW} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(N-k)} \sim \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{N-k}/(N-k)} = F_{k-1,N-k}$$

- Hence this can be used to test the hypothesis $b = 0$.

Example: Age and Fatness

- A random group of 18 people, aged between 25 and 60 years, is selected. Their weight (in kg) and their height (in meters) is measured. This determines the Body Mass Index (BMI), which is expressed by the relation

  $$BMI = \frac{body\ weight\ (kg)}{(height\ (m))^2}$$

  We obtain the following table:

Case	Age	BMI		Case	Age	BMI
1	25	21.8		10	50	24.9
2	28	20.1		11	53	26.6
3	35	21.8		12	53	25.5
4	35	23.1		13	57	29.4
5	40	24.2		14	58	27.5
6	40	23.3		15	58	26.4
7	43	25.4		16	58	27.1
8	45	24.5		17	60	27.7
9	50	27.6		18	60	30.1

Example: Age and Fatness, cont.

- In the previous table some ages are repeated. The table can be re-arranged as follows:

Case	Age	BMI		
1	25	21.8		
2	28	20.1		
3	35	21.8	23.1	
4	40	24.2	23.3	
5	43	25.4		
6	45	24.5		
7	50	27.6	24.9	
8	53	26.6	25.5	
9	57	29.4		
10	58	27.5	26.4	27.1
11	60	27.7	30.1	

Example: Age and Fatness, cont.

- We calculate: $\bar{y} = 25.389$
- $SSW = \sum_{i=1}^{8} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = 9.0$
- $SSB = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 = 115.1778$

Example: Age and Fatness, cont.

- $MSB = \dfrac{SSB}{10} = 11.518$

  $MSW = \dfrac{SSW}{7} = 1.2857$

- $\dfrac{MSB}{MSW} = \dfrac{11.518}{1.2857} = F_{10,7} = 8.958,$

  The critical region is $(3.637, +\infty)$. Hence the hypothesis $b = 0$ is rejected.

Summary

- We introduced the pure error sums of squares SSW (representing the within groups variation) and SSB (representing the between groups variation).
- We proved that $SST = SSW + SSB$.
- We introduced the mean sums of squares $MSB = \frac{SSB}{k-1}$ and $MSW = \frac{SSW}{N-k}$.
- We discussed the F statistics $\frac{MSB}{MSW} = F_{k-1,N-k}$ to test the hypothesis $b = 0$.

Section 3.10:
Repeated observations, Part 2

Testing for lack of fit

- We consider a formal test for determining whether or not a specific regression function adequately fits the data.
- This test assumes that the assumptions of the simple linear regression model hold.

The general idea

- If the model is correct, then $\hat{\sigma}^2$ should be an unbiased estimate of σ^2. The problem, however, is that σ^2 is in general unknown.
- The $\hat{\sigma}^2$ based on the chosen regression model needs to be compared to some model-free estimate of σ^2.
- We can do this if we have replications, that is more than one observation y for some fixed values of x. Such repeated measures would reveal the measurement error, while the variability between groups reflects the σ^2 described in the model.

Discussion

- What are the residuals within the ith group of observations? Recall there are observations $y_{i1} \dots y_{in_i}$ for $x = x_i$.

Answer to discussion question

- The fitted value at x_i is $\hat{y}_i$ and is the same for all observations at x_i.
- Hence the residuals within the ith group of observations are

$$r_{ij} = y_{ij} - \hat{y}_i$$

Splitting the residuals

- The average value of observations at x_i is

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

- The residuals $r_{ij} = y_{ij} - \hat{y}_i$ can be written as

$$r_{ij} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

- The difference $(y_{ij} - \bar{y}_i)$ measures the random variation at x_i and is called pure error.
- The difference $(\bar{y}_i - \hat{y}_i)$ arises because the model may not be a good one, so it measures the lack of fit at x_i.

Lack of fit sums of squares

- The error variation for the model is

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$$

- Recall from the Section 3.9 that the pure error sum of squares is

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- We define the lack of fit sum of squares be

$$SSL = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \hat{y}_i)^2$$

Splitting the error variation

- Fact:

$$SSE = SSW + SSL$$

Distribution of residual sum of squares

- From Section 3.8, if the simple regression model is true, then

$$\frac{SSE}{\sigma^2} \sim \chi^2_{N-2}$$

- Also, from Section 3.9,

$$\frac{SSW}{\sigma^2} \sim \chi^2_{N-k}$$

- From the previous slide we have

$$\frac{SSE}{\sigma^2} = \frac{SSW}{\sigma^2} + \frac{SSL}{\sigma^2}$$

- When the model is true, then SSL and SSW are independent and

$$\frac{SSL}{\sigma^2} \sim \chi^2_{k-2}$$

F-statistic for lack of fit

- We showed that, if the model $E(Y_{ij}) = a + bx_i$ is true, then

$$\frac{SSW}{\sigma^2} \sim \chi^2_{N-k} \qquad\qquad \frac{SSL}{\sigma^2} \sim \chi^2_{k-2}$$

  and the two random variables are independent.

- It follows that, under the hypothesis H_0 : 'the model is true',

$$\frac{SSL/\sigma^2(k-2)}{SSW/\sigma^2(N-k)} \sim \frac{\chi^2_{k-2}/(k-2)}{\chi^2_{N-k}/(N-k)} = F_{k-2,N-k}$$

- This statistic can therefore be used to test the lack of fit of the model. Large values of F lead to the rejection of the null hypothesis that the model is true.

ANOVA table

- These results can be summarized using an analysis of variance table, where we recall that

$$SST = SSM + SSE = SSM + SSW + SSL\,.$$

	Source	df	SS	MS	F
SSM	Regression (Model)	1	$\frac{((\sum_{i=1}^{k} n_i x_i(\bar{y}_i - \bar{y}))^2}{\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}$	$\frac{SSM}{1}$	
SSL	Lack of fit	$k-2$	$\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{a} - \hat{b}x_i)^2$	$\frac{SSL}{(k-2)}$	$F_{k-2,N-k}$
SSW	Pure error	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$\frac{SSW}{(N-k)}$	
SST	Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$		

Example: Age and Fatness

- For the Age and Fatness data, we calculated:
  $SSW = 9.0, \quad SSB = 115.1778, \quad SST = 124.1778;$
  $SSE = 20.0175, \quad SSM = 104.1603.$

- Hence

$$SSL = SSE - SSW = 20.0175 - 9.0 = 11.0175$$

- Thus $F_{9,7} = \dfrac{SSL/9}{SSW/7} = \dfrac{MSL}{MSW} = 0.9521.$

- The critical region is $(3.677, +\infty)$. We conclude that the null hypothesis $E(y_{ij}) = a + bx_i$ is acceptable.

Example: Age and Fatness, cont.

Analysis of Variance table

	Source	df	*SS*	*MS*	*F*
SSM	Regression (Model)	1	104.16	104.16	
SSL	Lack of fit	9	11.0175	1.2242	0.9521
SSW	Pure error	7	9.0	1.2857	
SST	Total	17	124.18		

Example: Age and Fatness, results obtained with R

- We firstly fit the simple linear regression model,

```
> T=read.table("fatness.txt")
> Age=T[,1]
> BMI=T[,2]
> modFat=lm(BMI ~ Age)
> summary(modFat)
```

```
Call:
lm(formula = BMI ~ Age)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.99956    1.16875  12.834 7.73e-10 ***
Age          0.22053    0.02417   9.124 9.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.119 on 16 degrees of freedom
Multiple R-squared:  0.8388,Adjusted R-squared:  0.8287
F-statistic: 83.26 on 1 and 16 DF,  p-value: 9.675e-08
```

Example: Age and Fatness, results obtained with R, cont.

- Two types of ANOVA tables:

```
> anova(modFat)
Analysis of Variance Table

Response: BMI
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 104.160 104.160  83.256 9.675e-08 ***
Residuals 16  20.017   1.251


> library(EnvStats)
> anovaPE(modFat)
               Df  Sum Sq Mean Sq F value    Pr(>F)
Age             1 104.160 104.160 81.0136 4.263e-05 ***
Lack of Fit     9  11.017   1.224  0.9521    0.5386
Pure Error      7   9.000   1.286
```

- To draw the second type of ANOVA table, the R package 'EnvStats' should be installed.

Summary

- We proved that $SSE = SSW + SSL$ where SSL is the lack of fit sum of squares.
- We discussed the F-statistics $F_{k-2,N-k}$ to assess the lack of fit of the model.
- We illustrated this on the Age and Fatness data example.

Chapter 4: Multiple Linear Regression

Dr.Ting Wei

Section 4.1: General Linear Model

The need for several explanatory variables

- In the simple linear regression model we had only one explanatory variable x.
- In many practical situations, the response is influenced by more than one explanatory variable; then the simple linear regression model does not give an adequate fit to the data.
- For example, the responsiveness to a drug could be influenced both by the dosage and by the age of the patients.
- Multiple regression extends simple linear regression to account for more than one explanatory variable.

Example: first-order model with two independent variables

- The first-order model with two independent variables is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n$$

  where $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta_0, \beta_1, \beta_2, \sigma^2$ are unknown parameters.

- As usual, y_i is the response in the ith trial, while x_{i1} and x_{i2} are the values of the two explanatory variables in the ith trial.

Discussion

- From previous slide:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n$$

  where $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta_0, \beta_1, \beta_2, \sigma^2$ are unknown parameters.

- For the above model, what is $E(y)$?

- For the above model, what is the regression function?

Answer to discussion question

- For the above model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

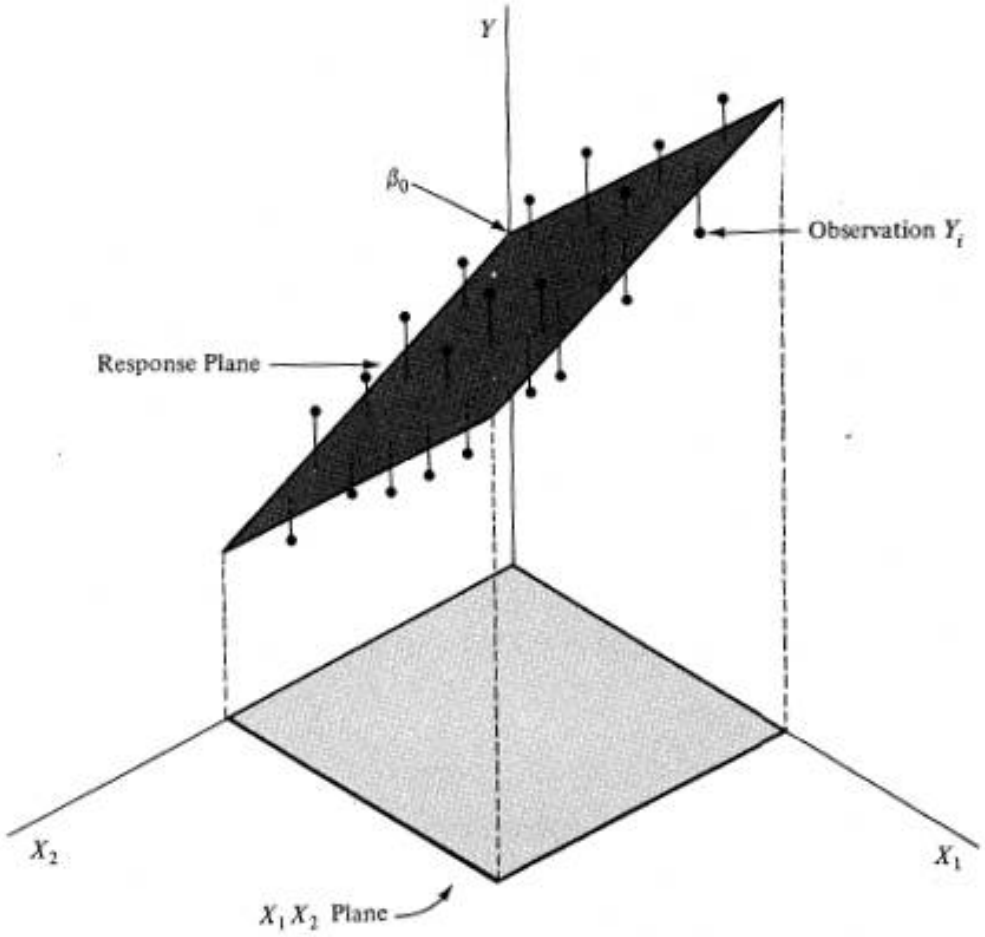
- The regression function has the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

for certain numerical estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ of the parameter.

Answer to discussion question, cont.

- The graph of the regression function is a plane ('response plane').



Simple linear regression model in matrix terms

- The simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

implies

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

- We want to re-write this in matrix notation.

Simple linear regression model in matrix terms, cont.

- Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Then

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}$$

- Hence the model in matrix term is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Model assumptions in matrix terms

- Recall the model assumptions for simple linear regression:
  i) $E(\varepsilon_i) = 0$;
  ii) $\text{var}(\varepsilon_i) = \sigma^2$;
  iii) ε_i are independent normal random variables.
- In matrix terms, condition (i) is

$$E\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix} = \begin{bmatrix}E(\varepsilon_1)\\ \vdots\\ E(\varepsilon_n)\end{bmatrix} = \begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix} = \mathbf{0}.$$

Model assumptions in matrix terms, cont.

- Condition (iii) implies $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ and with condition (ii) this gives

$$\text{VarCov}(\boldsymbol{\varepsilon}) \underset{def}{=} \begin{pmatrix} \text{var}(\varepsilon_1) & \text{cov}(\varepsilon_1\varepsilon_2) & \cdots & \text{cov}(\varepsilon_1\varepsilon_n) \\ \text{cov}(\varepsilon_2\varepsilon_1) & \text{var}(\varepsilon_2) & \cdots & \text{cov}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ \text{cov}(\varepsilon_n\varepsilon_1) & \text{cov}(\varepsilon_n\varepsilon_2) & \cdots & \text{var}(\varepsilon_n) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_n$$

Model assumptions in matrix terms, cont.

- In conclusion, the model in matrix terms is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a vector of independent random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{VarCov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

General linear model formulation

- The general linear regression model for a response variable Y and explanatory variables $x_1, \ldots, x_{p-1}$ is

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i, \quad i = 1, \ldots, n$$

- In matrix notation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{p-1,n} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

General linear model formulation, cont.

- The assumptions of the model are: ε is a vector of independent normal random variables with expectation $E(\varepsilon) = \mathbf{0}$ and variance-covariance matrix $\text{VarCov}(\varepsilon) = \sigma^2 \mathbf{I}_n$.

- It follows that

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \varepsilon) = E(\mathbf{X}\boldsymbol{\beta}) + E(\varepsilon) = \mathbf{X}\boldsymbol{\beta}$$
$$\text{VarCov}(\mathbf{Y}) = \text{VarCov}(\varepsilon) = \sigma^2 \mathbf{I}_n$$

- The regression function, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1}$ is a hyperplane.

- The case $p = 2$ recovers the simple linear regression model.

Least squares estimates

- The sum of squares of errors for the general linear model is given by

$$
\begin{aligned}
SS &= \sum_{i=1}^{n}(y_i - E(y_i))^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_{p-1}x_{p-1,i})^2 \\
&= \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{Y}^T - \boldsymbol{\beta}^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

where in the last equality we used

$$\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} = (\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta})^T = \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}$$

as it is a scalar.

Least squares estimates, cont.

- To obtain the least squares estimators for the parameter vector $\boldsymbol{\beta}$ we minimize SS.

- From previous slide

$$SS = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

- Differentiating with respect to β_k gives

$$\frac{\partial SS}{\partial \beta_k} = -2(kth \text{ entry of } \mathbf{X}^T\mathbf{Y}) + 2(kth \text{ entry of } \mathbf{X}^T\mathbf{X}\boldsymbol{\beta})$$

which is equivalent to

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Least squares estimates, cont.

- Therefore

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

  these are called normal equations.

- If $\mathbf{X}^T\mathbf{X}$ is non-singular, then the normal equations have a solution, which is the LSE, given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Example: Portrait sales

- Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children.
- The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2).
- Data on these variables for the most recent year for the 21 cities in which Dwaine Studios is now operating are shown in next slide (numbers expressed in thousands).

Table 1: Portrait sales data

X_1	X_2	Y
68.5	16.7	174.4
45.2	16.8	164.4
91.3	18.2	244.2
47.8	16.3	154.6
46.9	17.3	181.6
66.1	18.2	207.5
49.5	15.9	152.8
52.0	17.2	163.2
48.9	16.6	145.4
38.4	16.0	137.2
87.9	18.3	241.9
72.8	17.1	191.1
88.4	17.4	232.0
42.9	15.8	145.3
52.5	17.8	161.1
85.7	18.4	209.7
41.3	16.5	146.4
51.7	16.3	144.0
89.6	18.1	232.6
82.7	19.1	224.1
52.3	16.0	166.5

Example: Portrait sales, cont.

- The $\mathbf{X}$ and $\mathbf{Y}$ matrices are:

$$\mathbf{X} = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}$$

- We calculate

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 21.0 & 1302.4 & 360.0 \\ 1302.4 & 87707.9 & 22609.2 \\ 360.0 & 22609.2 & 6190.3 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 3820 \\ 249643 \\ 66073 \end{bmatrix}$$

Example: Portrait sales, cont.

- The least square estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
$$= \begin{bmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{bmatrix} \begin{bmatrix} 3820 \\ 249643 \\ 66073 \end{bmatrix}$$

- Hence

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix}$$

- The estimated regression function is

$$\hat{y} = -68.857 + 1.455x_1 + 9.366x_2$$

Properties of the estimator $\hat{\beta}$

- Given that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Recalling that $E(\mathbf{Y}) = \mathbf{X}\beta$ we deduce

$$\begin{aligned} E(\hat{\beta}) &= E\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta \end{aligned}$$

- Hence the estimator $\hat{\beta}$ is unbiased.

Properties of the estimator $\hat{\beta}$, cont.

- General fact: if A is a matrix of constants, then

$$\text{VarCov}(\boldsymbol{AY}) = \mathbf{A}(\text{VarCov}(\mathbf{Y}))\mathbf{A}^T$$

- Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, it follows that

$$\begin{aligned}
\text{VarCov}(\hat{\boldsymbol{\beta}}) = & \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\text{VarCov}(\mathbf{Y})\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) \\
= & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
= & \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

Distribution of the estimator $\hat{\beta}$

- In conclusion, if $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and variance as above:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

Summary

- The general linear model in matrix terms is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a vector of independent random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{VarCov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.
- The estimated parameter vector is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Distribution of the estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Section 4.2: Polynomial regression

Discussion

- The following is a polynomial regression model with one predictor variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \varepsilon_i$$

- Is this a special case of the general linear regression model?
- What is the fitted regression line?

Answer to discussion question

- The following is a polynomial regression model with one predictor variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \varepsilon_i$$

- This is a special case of the general linear regression model. If we let $x_{1i} = x_i$ and $x_{2i} = {x_i}^2$ We can write the model as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- The fitted regression line is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 {x_i}^2$$

which is a parabola.

Type of polynomial regression

- Polynomial regression models may contain one, two, or more than two predictor variables. Further, each predictor variable may be present in various powers.
- Polynomial regression models are special cases of the general linear regression model.
- They contain squared and higher-order terms of the predictor variable(s), making the response function curvilinear.

Interaction Effects

- When the effects of the predictor variables on the response variable are not additive, the effect of one predictor variable depends on the levels of the other predictor variables.

- An example of a nonadditive regression model with two predictor variables x_1 and x_2 is the following

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

- This is a special case of the general linear regression model. Let $x_{3i} = x_{1i} x_{2i}$ and then write the model as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Combination of Cases

- A regression model may combine several of the elements we have just noted and still be treated as a general linear regression model.
- Consider the following regression model containing linear and quadratic terms for each of two predictor variables and an interaction term represented by the cross-product term

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + \varepsilon_i$$

- Question: How can we view this as a case of the general linear model?

Combination of Cases, etc.

- Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + \varepsilon_i$$

- Let us define

$$z_{1i} = x_{1i} \quad z_{2i} = x_{1i}^2 \quad z_{3i} = x_{2i} \quad z_{4i} = x_{2i}^2 \quad z_{5i} = x_{1i} x_{2i}$$

- Then we re-write the model as

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i} + \beta_5 z_{5i} + \varepsilon_i$$

Meaning of Linear in General Linear Regression Model.

- As illustrated in the various examples the general linear regression model is not restricted to linear response surfaces.
- The term linear model refers to the fact that model is linear in the parameters; it does-not refer to the shape of the response surface.

Example: cafeteria

- A staff analyst for a cafeteria chain wishes to investigate the relation between the number of self-service coffee dispensers in a cafeteria lines and sales of coffee.
- Fourteen similar cafeterias are chosen in the experiment. The number of self-service dispensers that are placed in the test cafeterias varies from zero (coffee is dispensed here by a line attendant) to six.

Data for cafeteria coffee sales

Cafeteria i	Number of Dispensers X_i	Coffee Sales (hundred gallons) Y_i
1	0	508.1
2	0	498.4
3	1	568.2
4	1	577.3
5	2	651.7
6	2	657.0
7	3	713.4
8	3	697.5
9	4	755.3
10	4	758.9
11	5	787.6
12	5	792.1
13	6	841.4
14	6	831.8

Example: cafeteria, cont.

- The analyst fits a quadratic polynomial regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Example: cafeteria, cont.

- In matrix form, the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\mathbf{Y} = \begin{bmatrix} 508.1 \\ 498.4 \\ 568.2 \\ 577.3 \\ 651.7 \\ 657.0 \\ 713.4 \\ 697.5 \\ 755.3 \\ 758.9 \\ 787.6 \\ 792.1 \\ 841.4 \\ 831.8 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 6 & 36 \end{bmatrix}$$

Example: cafeteria, cont.

- We calculate

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 6 & 36 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & 0 & 1 & \cdots & 5 & 6 & 6 \\ 0 & 0 & 1 & \cdots & 25 & 36 & 36 \end{bmatrix} = \begin{bmatrix} 14 & 42 & 182 \\ 42 & 182 & 882 \\ 182 & 882 & 4550 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.38095 & -0.23214 & 0.029762 \\ -0.23214 & 0.23214 & -0.035714 \\ 0.029762 & -0.035714 & 0.0059524 \end{bmatrix}$$

Example: cafeteria, cont.

- We calculate

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} =$$

$$= \begin{bmatrix} 0.38095 & -0.23214 & 0.029762 \\ -0.23214 & 0.23214 & -0.035714 \\ 0.029762 & -0.035714 & 0.0059524 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & 0 & 1 & \cdots & 5 & 6 & 6 \\ 0 & 0 & 1 & \cdots & 25 & 36 & 36 \end{bmatrix} \begin{bmatrix} 508.1 \\ 498.4 \\ 568.2 \\ 577.3 \\ 651.7 \\ 657.0 \\ 713.4 \\ 697.5 \\ 755.3 \\ 758.9 \\ 787.6 \\ 792.1 \\ 841.4 \\ 831.8 \end{bmatrix}$$

Example: cafeteria, cont.

- We obtain

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 502.556 \\ 80.386 \\ -4.249 \end{bmatrix}$$

- The fitted regression function is:

$$\hat{y} = 502.556 + 80.386x - 4.249x^2$$

Example: cafeteria, R treatment

- Input the problem data into R as follows:

```
> T=read.table("cafeteria.txt")
> x=T[,1]
> x2=x^2
> y=T[,2]
> cafeteria_fit <- lm(y ~ x + x2)
> summary(cafeteria_fit)
```

Example: cafeteria, R treatment, results

- R outputs:

```
Call:
lm(formula = y ~ x + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 502.5560     4.8500 103.619  < 2e-16 ***
x            80.3857     3.7861  21.232 2.81e-10 ***
x2           -4.2488     0.6063  -7.008 2.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.858 on 11 degrees of freedom
Multiple R-squared:  0.9961,Adjusted R-squared:  0.9953
F-statistic:  1391 on 2 and 11 DF,  p-value: 5.948e-14
```

Example: cafeteria, R treatment, cont.

- What is the matrix **X**:

```
> X=model.matrix(cafeteria_fit)
   (Intercept) x x2
1            1 0  0
2            1 0  0
3            1 1  1
4            1 1  1
5            1 2  4
6            1 2  4
7            1 3  9
8            1 3  9
9            1 4 16
10           1 4 16
11           1 5 25
12           1 5 25
13           1 6 36
14           1 6 36
```

Example: cafeteria, R treatment, cont.

- Then the $\mathbf{X}^T\mathbf{X}$:

```
> t(X) %*% X
            (Intercept)   x   x2
(Intercept)          14  42  182
x                    42 182  882
x2                  182 882 4550
```

- And $(\mathbf{X}^T\mathbf{X})^{-1}$:

```
> solve(t(X) %*% X)
            (Intercept)           x           x2
(Intercept)   0.3809524 -0.23214286  0.029761905
x            -0.2321429  0.23214286 -0.035714286
x2            0.0297619 -0.03571429  0.005952381
```

Example: cafeteria, R treatment, cont.

- Vector **Y**:

```
> Y=as.matrix(y)
       [,1]
 [1,] 508.1
 [2,] 498.4
 [3,] 568.2
 [4,] 577.3
 [5,] 651.7
 [6,] 657.0
 [7,] 713.4
 [8,] 697.5
 [9,] 755.3
[10,] 758.9
[11,] 787.6
[12,] 792.1
[13,] 841.4
[14,] 831.8
```

Example: cafeteria, R treatment, cont.

- We calculate $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$:

```
> solve(t(X) %*% X) %*% t(X) %*% Y
                 [,1]
(Intercept) 502.55595
x            80.38571
x2           -4.24881
```

Example: Portrait sales data

X_1	X_2	Y
68.5	16.7	174.4
45.2	16.8	164.4
91.3	18.2	244.2
47.8	16.3	154.6
46.9	17.3	181.6
66.1	18.2	207.5
49.5	15.9	152.8
52.0	17.2	163.2
48.9	16.6	145.4
38.4	16.0	137.2
87.9	18.3	241.9
72.8	17.1	191.1
88.4	17.4	232.0
42.9	15.8	145.3
52.5	17.8	161.1
85.7	18.4	209.7
41.3	16.5	146.4
51.7	16.3	144.0
89.6	18.1	232.6
82.7	19.1	224.1
52.3	16.0	166.5

Example: Portrait sales data, R treatment

- Input the problem data into R as follows:

```
> T=read.table("Portrait.txt")
> number=T[,1]
> income=T[,2]
> sales=T[,3]
> Portrait_fit <- lm(sales ~ number + income)
> summary(Portrait_fit)
```

Example: Portrait sales data, R treatment, results

- R outputs:

```
Call:
lm(formula = sales ~ number + income)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
number        1.4546     0.2118   6.868    2e-06 ***
income        9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

Example: Portrait sales data, R treatment, cont.

- What is the matrix **X**:

```
> X=model.matrix(Portrait_fit)
   (Intercept) number income
1            1   68.5   16.7
2            1   45.2   16.8
3            1   91.3   18.2
4            1   47.8   16.3
5            1   46.9   17.3
6            1   66.1   18.2
7            1   49.5   15.9
8            1   52.0   17.2
9            1   48.9   16.6
10           1   38.4   16.0
11           1   87.9   18.3
12           1   72.8   17.1
13           1   88.4   17.4
14           1   42.9   15.8
15           1   52.5   17.8
16           1   85.7   18.4
17           1   41.3   16.5
18           1   51.7   16.3
19           1   89.6   18.1
20           1   82.7   19.1
21           1   52.3   16.0
```

Example: Portrait sales data, R treatment, cont.

- Then the $\mathbf{X}^T\mathbf{X}$:

```
> t(X) %*% X
            (Intercept)   number   income
(Intercept)        21.0  1302.40   360.00
number           1302.4 87707.94 22609.19
income            360.0 22609.19  6190.26
```

- And $(\mathbf{X}^T\mathbf{X})^{-1}$:

```
> solve(t(X) %*% X)
            (Intercept)        number       income
(Intercept) 29.72892348  0.0721834719 -1.992553186
number       0.07218347  0.0003701761 -0.005549917
income      -1.99255319 -0.0055499173  0.136310637
```

Example: Portrait sales data, R treatment, cont.

- Vector **Y**:

```
> Y=as.matrix(sales)
       [,1]
 [1,] 174.4
 [2,] 164.4
 [3,] 244.2
 [4,] 154.6
 [5,] 181.6
 [6,] 207.5
 [7,] 152.8
 [8,] 163.2
 [9,] 145.4
[10,] 137.2
[11,] 241.9
[12,] 191.1
[13,] 232.0
[14,] 145.3
[15,] 161.1
[16,] 209.7
[17,] 146.4
[18,] 144.0
[19,] 232.6
[20,] 224.1
[21,] 166.5
```

Example: Portrait sales data, R treatment, cont.

- We calculate $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$:

```
> solve(t(X) %*% X) %*% t(X) %*% Y
                 [,1]
(Intercept) -68.85707
number        1.45456
income        9.36550
```

Summary

- Polynomial regression model with one predictor variable is a special case of the general linear regression model.
- Polynomial regression model with higher-order terms and interaction effects.
- Examples: 'Cafeteria' and 'Portrait sales'.

Section 4.3: Confidence intervals and hypothesis tests

Recap

- The GLM model in matrix forms is ....., where ε is a vector of independent random variables with $E(\varepsilon) =$ ..... and $\text{VarCov}(\varepsilon) =$ .....

Inferences about individual elements of β

- Recall from Section 4.1 that the LSE $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- Hence the kth-entry of $\boldsymbol{\beta}$, β_k is distributed as

$$\hat{\beta}_k \sim N\left(\beta_k, K^2\sigma^2\right)$$

where $K^2 = [(\mathbf{X}^T\mathbf{X})^{-1}]_{k+1,k+1}$ and $K\sigma$ is the standard error of $\hat{\beta}_k$.

Inferences about individual elements of β

- It follows that

$$\frac{\hat{\beta}_k - \beta_k}{K\sigma} \sim N(0,1)$$

from which we obtain the 95% C.I.

$$\left(\hat{\beta}_k - 1.96K\sigma,\ \hat{\beta}_k + 1.96K\sigma\right)$$

- However, σ^2 is in general unknown.

Point estimate of σ^2

- Recall from the simple linear regression that the estimate of the variance is:

$$\hat{\sigma}^2 = \frac{SSE}{(n-2)}$$

- For multiple linear regression, the estimate of the variance is:

$$\hat{\sigma}^2 = \frac{SSE}{(n-p)}$$

where $SSE = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Confidence interval for β_k

- From previous slide, we can show that

$$\frac{\hat{\beta}_k - \beta_k}{K\hat{\sigma}} \sim t_{n-p}$$

- Therefore a 95% confidence interval for β_k is

$$\left(\hat{\beta}_k - t_{0.025,n-p}K\hat{\sigma},\ \hat{\beta}_k + t_{0.025,n-p}K\hat{\sigma}\right)$$

Hypothesis tests for β_k

- We want to test the hypothesis $H_0 : \beta_k = \beta$ against $H_1 : \beta_k \neq \beta$.
- If the hypothesis $H_0 : \beta_k = \beta$ is true, then

$$t = \frac{\hat{\beta}_k - \beta}{K\hat{\sigma}} = \frac{\hat{\beta}_k - \beta}{\text{s.e.}(\hat{\beta}_k)}$$

  should be a typical observation from a t_{n-p} distribution.

- Hence we can find the critical region or calculate the p-value, and decide to accept/reject H_0 with the given significance level.

Discussion

- What is the meaning of the null hypothesis $H_0 : \beta_k = 0$?
- Hint: think in terms of the effect of the kth predictor as well as in terms of models comparison.

Answer to discussion question

- The null hypothesis $H_0 : \beta_k = 0$ says that the kth predictor is unnecessary and can be omitted from the model.
- This means that, conditional on $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{p-1}$ being in the model, then x_k has no additional explanatory power.

Hypothesis tests for β_k, cont.

- We can think about this as a test for comparing models

$$H_0 : E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{k-1} x_{k-1,i} + \beta_{k+1} x_{k+1,i} + \cdots + \beta_{p-1} x_{p-1,i}$$

against

$$H_1 : E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i}$$

- Note: this does not say anything about comparing models $E(y_i) = \beta_0$ and $E(y_i) = \beta_0 + \beta_k x_{ki}$.

Example: Salary data

- The following data gives results from 10 employees in three columns: salary in thousands of pounds, age in years and gender(0 =male,1 =female).

Row	Salary	Age	M/F
1	38.985	52	0
2	28.938	39	1
3	32.920	45	1
4	29.548	30	0
5	31.138	46	1
6	24.749	26	1
7	41.889	63	0
8	31.528	35	0
9	38.791	48	0
10	39.828	47	1

Example: Salary data, cont.

- Let Y = salary, X_1 = age, X_2 = gender. We fit a multiple regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- We have

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 431 & 5 \\ 431 & 19649 & 203 \\ 5 & 203 & 5 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{25260} \begin{bmatrix} 57036 & -1140 & -10752 \\ -1140 & 25 & 125 \\ -10752 & 125 & 10729 \end{bmatrix}$$

$$SSE = 47.898, \quad \hat{\sigma} = \sqrt{\frac{47.898}{7}} = 2.616$$

Example: Salary data, cont.

- From the estimate $\hat{\sigma}$ and the diagonal entries of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ we obtain

$$\text{s.e.}(\hat{\beta}_0) = 2.616\sqrt{\frac{57036}{25260}} = 3.931$$

$$\text{s.e.}(\hat{\beta}_1) = 2.616\sqrt{\frac{25}{25260}} = 0.082$$

$$\text{s.e.}(\hat{\beta}_2) = 2.616\sqrt{\frac{10729}{25260}} = 1.705$$

Example: Salary data, cont.

- We calculate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 16.445 \\ 0.43209 \\ -2.4732 \end{bmatrix}$$

- The 95% C.I. for β_2 is

$$\begin{aligned} &\left(\hat{\beta}_2 - t_{0.025,7}\,\mathrm{s.e.}(\hat{\beta}_2),\ \hat{\beta}_2 + t_{0.025,7}\,\mathrm{s.e.}(\hat{\beta}_2)\right) \\ &\quad = (-2.4732 \pm 2.365 \times 1.705) \\ &\quad = (-6.504,\ 1.558) \end{aligned}$$

Example: Salary data, cont.

- To test the hypothesis $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, we calculate

$$t = \frac{\hat{\beta}_2}{\mathrm{s.e.}(\hat{\beta}_2)} = -\frac{2.4732}{1.705} = -1.451 \sim t_7$$

- The critical region is $(-\infty, -2.365) \cup (2.365, +\infty)$. Hence the hypothesis is accepted, and the variable X_2 is unnecessary. Thus there is no significant difference between males and females.

Example: Salary data, R results

- R outputs:

```
> salary_fit <- lm(y ~ x1 + x2)
> summary(salary_fit)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.44499    3.93067   4.184  0.00412 **
x1           0.43209    0.08229   5.251  0.00119 **
x2          -2.47316    1.70479  -1.451  0.19015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.616 on 7 degrees of freedom
Multiple R-squared:  0.835,Adjusted R-squared:  0.7878
F-statistic: 17.71 on 2 and 7 DF,  p-value: 0.001826

> confint(salary_fit, "x2", level = 0.95)

       2.5 %  97.5 %
x2 -6.504351 1.55803
```

Summary

- The 95% confidence interval for β_k is

$$\hat{\beta}_k \pm t_{0.025,n-p} K \hat{\sigma} \ , \qquad K^2 = [(\mathbf{X}^T\mathbf{X})^{-1}]_{k+1k+1}$$

- If the hypothesis $H_0 : \beta_k = \beta$ is true, then

$$t = \frac{\hat{\beta}_k - \beta}{K\hat{\sigma}} = \frac{\hat{\beta}_k - \beta}{\mathrm{s.e.}(\hat{\beta}_k)}$$

should be a typical observation from a t_{n-p} distribution.

Section 4.4: Confidence intervals and prediction intervals

Different type of prediction

- We want to use the fitted linear regression model to make predictions.
- As in the case of simple linear regression (see Section 3.6) we have two types of predictions:

  1) Prediction of a future mean response.

  2) Prediction of a future observation.

Inferences about linear functions of the parameters

- Sometimes we are interested in making inferences about a linear combination of the model parameters $\beta_0, \ldots, \beta_{p-1}$.
- For example, we may want to estimate

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

where $E(y)$ is the mean yield of a chemical process for setting of controlled process variables $x_1, \ldots, x_{p-1}$.

Inferences about linear functions of the parameters, cont.

- We want to make an inference about the linear function

$$a_0\beta_0 + a_1\beta_1 + \cdots + a_{p-1}\beta_{p-1}$$

  where a_i are constants.

- In matrix form this is $\mathbf{a}^T\boldsymbol{\beta}$ where $\mathbf{a}^T = (a_0, \ldots, a_{p-1})$.

Inferences about linear functions of the parameters, cont.

- By the invariance property of maximum likelihood estimators,

$$\widehat{\mathbf{a}^T\boldsymbol{\beta}} = a_0\hat{\beta}_0 + \cdots + a_{p-1}\hat{\beta}_{p-1} = \mathbf{a}^T\hat{\boldsymbol{\beta}}$$

- Hence

$$\begin{aligned} E(\widehat{\mathbf{a}^T\boldsymbol{\beta}}) &= E(a_0\hat{\beta}_0 + \cdots + a_{p-1}\hat{\beta}_{p-1}) \\ &= a_0E(\hat{\beta}_0) + \cdots + a_{p-1}E(\hat{\beta}_{p-1}) \\ &= a_0\beta_0 + \cdots + a_{p-1}\beta_{p-1} = \mathbf{a}^T\boldsymbol{\beta} \end{aligned}$$

Inferences about linear functions of the parameters, cont.

- From previous slide: $\widehat{\mathbf{a}^T\boldsymbol{\beta}} = a_0\hat{\beta}_0 + \cdots + a_{p-1}\hat{\beta}_{p-1}$.
- Therefore

$$\mathrm{var}(\widehat{\mathbf{a}^T\boldsymbol{\beta}}) = \sum_{i=0}^{p-1} a_i^2\,\mathrm{var}(\beta_i) + 2\sum_{0\le i<j\le p-1} a_i a_j\,\mathrm{cov}(\beta_i,\beta_j) =$$
$$= \mathbf{a}^T\,\mathrm{VarCov}(\boldsymbol{\beta})\,\mathbf{a} = \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{a}$$

- Also $\widehat{\mathbf{a}^T\boldsymbol{\beta}}$ is normally distributed since each $\hat{\beta}_k$ is.

$$\boldsymbol{a}^T\hat{\beta} \sim N\left(\boldsymbol{a}^T\beta, \sigma^2\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}\right)$$

Inferences about linear functions of the parameters, cont.

- We showed that $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{a}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a})$.
- Hence

$$\frac{\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\beta}{\sigma\sqrt{\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a}}} \sim N(0,1)$$

Inferences about linear functions of the parameters, cont.

- Replacing σ with $\hat{\sigma}$ we have

$$\frac{\mathbf{a}^T\hat{\beta} - \mathbf{a}^T\beta}{\hat{\sigma}\sqrt{\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a}}} \sim t_{n-p}$$

- So a 95% C.I. for $\mathbf{a}^T\beta$ is

$$\left(\mathbf{a}^T\hat{\beta} - t_{0.025,n-p}\hat{\sigma}\sqrt{\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a}},\right.$$
$$\left.\mathbf{a}^T\hat{\beta} + t_{0.025,n-p}\hat{\sigma}\sqrt{\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a}}\right)$$

Methodology

- We showed that

$$\boldsymbol{a}^T \hat{\beta} \sim N\left(\boldsymbol{a}^T \beta, \sigma^2 \boldsymbol{a}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{a}\right)$$

- To predict $E(y_0)$ corresponding to $\mathbf{x}_0$ we can apply this to the linear combination $\mathbf{x}_0^T \beta = E(y_0)$ for a preassigned value $\mathbf{x} = \mathbf{x}_0$.

Confidence interval for expectation

- We deduce that $\mathbf{x}_0^T \boldsymbol{\beta}$ has a normal distribution with mean and variance given by

$$\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}, \qquad \text{var}(\hat{y}_0) = \sigma^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0$$

- Therefore

$$\frac{\hat{y}_0 - \boldsymbol{x}_0^T \boldsymbol{\beta}}{\hat{\sigma}\sqrt{\boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \sim t_{n-p}$$

Confidence interval for expectation, cont.

- The 95% C.I. for $E(y_0) = \mathbf{x}_0^T \boldsymbol{\beta}$ is therefore

$$\left(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - t_{0.025,n-p} \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{x}_0}, \right.$$
$$\left. \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} + t_{0.025,n-p} \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{x}_0} \right)$$

- This is the confidence interval for the future expected value at $\mathbf{x} = \mathbf{x}_0$.

Prediction interval for future observation

- If y_0 is the actual value observed at x_0, then

$$y_0 = \boldsymbol{x}_0^T \boldsymbol{\beta} + \varepsilon_0$$

where $\varepsilon_0 \sim N(0, \sigma^2)$.

- The point estimate of y_0 is $\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$ and, since $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,

$$E(\hat{y}_0 - y_0) = E(\hat{y}_0) - E(y_0) = \mathbf{x}_0^T E(\hat{\boldsymbol{\beta}}) - \mathbf{x}_0^T \boldsymbol{\beta} = 0$$

Prediction interval for future observation, cont.

- Since $\hat{y}_0$ and ε_0 are independent,

$$\text{var}(\hat{y}_0 - y_0) = \text{var}(\hat{y}_0) + \text{var}(\varepsilon_0) = \sigma^2 \left(1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0\right)$$

- In conclusion

$$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2 (1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0)\right)$$

and independent on $\hat{\sigma}^2$.

Prediction interval for future observation, cont.

- Hence

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p}$$

- The 95% prediction interval for y_0 is therefore

$$\left(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - t_{0.025,n-p}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0},\right.$$
$$\left.\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + t_{0.025,n-p}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{x}_0}\right)$$

Comparison

- In conclusion,

Confidence interval for future expected value $E(y) = \mathbf{x}_0^T \boldsymbol{\beta}$ at $\mathbf{x}_0$	$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,n-p} \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}$	(1)
Prediction interval for future observed value $y_0 = \mathbf{x}_0^T \boldsymbol{\beta} + \varepsilon_0$ at $\mathbf{x}_0$	$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,n-p} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}$	(2)

- Note that the intervals (1) and (2) have the same mid-point, but the interval (2) is wider.

Salary data

- Recall the Salary data from Section 4.3: the table gives results from 10 employees in three columns: salary in thousands of pounds, age in years and gender (0=male, 1=female).

Row	Salary	Age	M/F
1	38.985	52	0
2	28.938	39	1
3	32.920	45	1
4	29.548	30	0
5	31.138	46	1
6	24.749	26	1
7	41.889	63	0
8	31.528	35	0
9	38.791	48	0
10	39.828	47	1

- Question: predict the expected salary for a male aged 30 years.

Example: Salary Data, cont.

- Since $X_1 =$ age, $X_2 =$ gender with $X_2 = 0$ for males, we have $\mathbf{x}_0^T = (1, 30, 0)$.
- We previously calculated

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 16.445 \\ 0.43209 \\ -2.4732 \end{bmatrix}$$

Therefore

$$\hat{y}_0 = \mathbf{x}_0^T\boldsymbol{\beta} = 16.445 + 0.43209 \times 30 = 29.408$$

Example: Salary Data, cont.

- We previously calculated:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{25260}\begin{bmatrix} 57036 & -1140 & -10752 \\ -1140 & 25 & 125 \\ -10752 & 125 & 10729 \end{bmatrix}$$

- $SSE = 47.898, \quad \hat{\sigma} = \sqrt{\frac{47.898}{7}} = 2.616$

- Also, $t_{0.025,7} = 2.365$.

Example: Salary Data, cont.

- The confidence interval for the expected salary for a male aged 30 is:

$$\begin{aligned} \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,7} \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0} \\ = (25.301,\ 33.515) \end{aligned}$$

- The prediction interval for the actual salary for a male aged 30 is:

$$\begin{aligned} \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,7} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0} \\ = (21.983,\ 36.832) \end{aligned}$$

Example: Salary data, using R

- R outputs:

```
> salary_fit <- lm(y ~ x1 + x2)
> newdata = data.frame(x1=30,x2=0)

> predict(salary_fit, newdata, interval="confidence")
       fit      lwr      upr
1 29.40763 25.30069 33.51457
> predict(salary_fit, newdata, interval="prediction")
       fit     lwr      upr
1 29.40763 21.9829 36.83236
```

Summary

- We derived the formulas for the CI for mean response and prediction interval for new observation $\mathbf{x}_0$.

Confidence interval for future expected value $E(y) = \mathbf{x}_0^T \boldsymbol{\beta}$ at $\mathbf{x}_0$	$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,n-p} \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}$	(1)
Prediction interval for future observed value $y_0 = \mathbf{x}_0^T \boldsymbol{\beta} + \varepsilon_0$ at $\mathbf{x}_0$	$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{0.025,n-p} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}$	(2)

- As in the case of the simple linear regression model, the prediction interval is wider, as it takes into account the variance of the error term.

Section 4.5: Anova for the general linear model

Motivation

- Suppose we have a general linear models with predictors $x_1, \cdots, x_{p-1}$.
- Question: does at least one of the predictors affects the response?
- The ANOVA for the model will answer this question.

The hat matrix and the fitted values

- The matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

  is called the hat matrix.

- Recalling that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ we have the following expression for the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{Y}$$

- It can be easily shown that $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^T = \mathbf{H}$.

Residual vector

- From the previous slide

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

where $\mathbf{H}$ is the hat matrix, satisfying $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^T = \mathbf{H}$.

- Therefore the residual vector is

$$\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Residual sum of squares

- Therefore the residual sum of squares (error variation) is

$$
\begin{aligned}
SSE &= \sum_{i=1}^{n} r_i^2 = \mathbf{r}^T\mathbf{r} \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}
\end{aligned}
$$

- It follows that $(\mathbf{I}-\mathbf{H})^2 = \mathbf{I}-\mathbf{H}$.

ANOVA for simple linear regression

- Recall from Section 3.8 that we divided the total sum of squares in the simple linear regression model into two parts:

$$SST = SSM + SSE$$

where

$$SST = \text{Total variation} = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSM = \text{Model variation} = (\hat{b})^2 S_{xx} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SSE = \text{Error variation} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ANOVA for simple linear regression, cont

- Therefore the equation

$$SST = SSM + SSE$$

can be rewritten as:

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

Geometric interpretation

- From last slide

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

- Let $\mathbf{Y} = (y_1, \ldots, y_n)^T, \quad \mathbf{1} = (1, \ldots, 1)^T, \quad \hat{\mathbf{Y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$

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

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$$

- Hence (1) is the same as:

$$\|\mathbf{Y} - \bar{y}\mathbf{1}\|^2 = \|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \tag{2}$$

ANOVA for general linear model with intercept

- For the general linear model, we have a similar decomposition of the total sum of squares provided that we have an intercept.
- We are going to write the various sums of squares in matrix notation.
- Recall that $\mathbf{Y} = (y_1, \ldots, y_n)^T$ is the response matrix, $\hat{\mathbf{Y}} = \mathbf{HY}$ where $\mathbf{H}$ is the hat matrix and $\boldsymbol{J}$ is the $n \times n$ matrix with each entry equal to 1.
- Thus

$$\mathbf{Y}^T \frac{1}{n} \boldsymbol{J}\, \mathbf{Y} = (y_1, \ldots, y_n) \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{y} \sum_{i=1}^{n} y_i = n\bar{y}^2$$

ANOVA for general linear model with intercept, cont.

- The total sum of squares in matrix notation is

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 + n\bar{y}^2 - 2\bar{y}\sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} y_i^2 + n\bar{y}^2 - 2n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\frac{1}{n}\boldsymbol{J}\mathbf{Y} = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\boldsymbol{J}\right)\mathbf{Y}
\end{aligned}
$$

- The residual sum of squares is

$$SSE = \mathbf{Y}^T(\mathbf{I} - \boldsymbol{H})\mathbf{Y}$$

- Hence the model variation is

$$SSM = SST - SSE = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\boldsymbol{J}\right)\mathbf{Y} - \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^T\left(\mathbf{H} - \frac{1}{n}\boldsymbol{J}\right)\mathbf{Y}$$

Distributions of sums of squares

- It can be shown that

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$$

and

$$\frac{SST}{\sigma^2} \sim \chi^2_{n-1}$$

- Further, if the first column of $\mathbf{X}$ is $\mathbf{1}$ and $\boldsymbol{\beta}^T = (\beta_0, \mathbf{0}^T)$, then

$$\frac{SSM}{\sigma^2} \sim \chi^2_{p-1}$$

and $\frac{SSE}{\sigma^2}$, $\frac{SSM}{\sigma^2}$ are independent.

Mean squares

- Dividing the sums of squares by their degrees of freedom we obtain the corresponding 'mean squares'.

- The mean squared error is

$$MSE = \frac{SSE}{n-p} \sim \frac{\sigma^2 \chi^2_{n-p}}{n-p}$$

- The mean squared model variation under $H_0 : \boldsymbol{\beta}^T = (\beta_0, \mathbf{0}^T)$ is

$$MSM = \frac{SSM}{p-1} \sim \frac{\sigma^2 \chi^2_{p-1}}{p-1}$$

The F statistic

- From the previous slide: under the hypothesis $H_0 : \boldsymbol{\beta}^T = (\beta_0, \mathbf{0}^T)$

$$MSE \sim \frac{\sigma^2 \chi^2_{n-p}}{n-p}, \qquad MSM \sim \frac{\sigma^2 \chi^2_{p-1}}{p-1}$$

- Hence, if H_0 is true, the ratio

$$F = \frac{MSM}{MSE} \sim \frac{\sigma^2 \chi^2_{p-1}/p-1}{\sigma^2 \chi^2_{n-p}/n-p} = F_{p-1,n-p}$$

  has a $F_{p-1,n-p}$ distribution.

- This can be used to test the hypothesis $H_0 : \boldsymbol{\beta}^T = (\beta_0, \mathbf{0}^T)$. If H_0 is rejected, we conclude that at least one of the predictors in the model affects the response.

The coefficient of multiple determination

- The **coefficient of multiple determination** is defined as

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST}$$

- This measures the proportion of total variation in the response y that is explained by the simultaneous predicting power of all the explanatory variables, through the multiple regression model.

- From Section 3.3, in the case of the simple linear regression model, $R^2 = \hat{\rho}^2$, where $\hat{\rho}$ is the sample correlation coefficient between X and Y.

The coefficient of multiple determination, cont.

- Since $SSM = SST - SSE$ we have

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- We have:
  - $0 \leq R^2 \leq 1$;
  - $R^2 = 0$ when all $\beta_k = 0$, $k = 1 \ldots, p-1$;
  - $R^2 = 1$ when $y_i = \hat{y}_i$ for all i.

The ANOVA table for the intercept model

	Source	df	SS	MS	F
SSM	Regression (Model)	$p-1$	$\mathbf{Y}^T\left(\mathbf{H}-\frac{1}{n}\boldsymbol{J}\right)\mathbf{Y}$	$\dfrac{SSM}{p-1}$	$F_{p-1,n-p}$
SSE	Error (Residual)	$n-p$	$\mathbf{Y}^T(\mathbf{I}-\boldsymbol{H})\mathbf{Y}$	$\dfrac{SSE}{n-p}$	
SST	Total	$n-1$	$\mathbf{Y}^T\left(\mathbf{I}-\frac{1}{n}\boldsymbol{J}\right)\mathbf{Y}$		

Example: Salary data

Education (x_1)	Experience (x_2)	Age (x_3)	Salary (y)
8	12	42	95
13	19	48	145
16	20	51	164
14	24	53	186
15	30	61	197
11	16	46	139
13	19	50	163
15	15	45	125
16	21	56	133
13	26	58	173
15	18	55	143
9	5	35	75
13	19	48	165
13	22	52	167
14	28	57	187

- We want to fit a multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad i = 1, \ldots, 15$$

Example: Salary data, R calculation

- The multiple regression with R gives the following result

```
> T=read.table("salary2.txt")
> x1=T[,1]
> x2=T[,2]
> x3=T[,3]
> y=T[,4]
> salary_fit <- lm(y ~ x1 + x2 + x3)
> summary(salary_fit)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.028     43.545   2.688 0.021123 *
x1             2.469      1.974   1.251 0.237025
x2             7.305      1.503   4.861 0.000502 ***
x3            -2.820      1.558  -1.810 0.097707 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 11 degrees of freedom
Multiple R-squared:  0.9052,Adjusted R-squared:  0.8794
F-statistic: 35.02 on 3 and 11 DF,  p-value: 6.369e-06
```

Example: Salary data, R calculation, cont.

- The ANOVA result calculated by R,

```
> null_model=lm(y ~ 1)
> anova(null_model, salary_fit)

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     14 16037.7
2     11  1519.9  3     14518 35.024 6.369e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary

- ANOVA table for the general linear model.
- The F statistics to test the hypothesis $H_0 : \boldsymbol{\beta}^T = (\beta_0, \mathbf{0}^T)$.
- The coefficient of multiple determination $R^2 = \frac{SSM}{SST}$ and its meaning.

Section 4.6: Extra sum of squares

Main question

- In Section 4.5 we discussed the F-statistic to test the hypothesis $H_0 : \boldsymbol{\beta}^T = (\boldsymbol{\beta}_0, \mathbf{0}^T)$.
- If the hypothesis is rejected, we know that at least one of the explanatory variables affects the response.
- However, we may still be able to delete several variables.

Discussion

- We ask if we can reduce the set of explanatory variables $x_1, \ldots, x_{p-1}$ to $x_1, \ldots, x_{q-1}$ where $q < p$, by omitting $x_q, \ldots, x_{p-1}$.
- Questions:
  - How would you state the null hypothesis in this case?
  - How do you express this in terms of model comparison?

Answer to discussion question

- The null hypothesis is $H_0 : \beta_q = \cdots = \beta_{p-1} = 0$.
- The model comparison is between the full model with all predictors and the reduced model with predictors $x_1, \ldots, x_{q-1}$ only.
  - Reduced model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1} + \varepsilon$$

  - Full model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1} + \beta_q x_q + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

Methodology

- Our methodology consists in determining if the predictors $x_q, \ldots, x_{p-1}$ provide a significant increase in the overall regression sum of squares, or equivalently a significant decrease in the residual sum of squares.

Extra sum of squares

- The extra sum of squares due to $x_q, \ldots, x_{p-1}$ given $x_1, \ldots, x_{q-1}$ are already in the model is defined by

$$\begin{aligned} & SSM(x_q, \ldots, x_{p-1} | x_1, \ldots, x_{q-1}) \\ & = SSM(x_1, \ldots, x_{p-1}) - SSM(x_1, \ldots, x_{q-1}) \\ & = \begin{array}{c}\text{model variation}\\ \text{of full model}\end{array} - \begin{array}{c}\text{model variation}\\ \text{of reduced model}\end{array} \end{aligned}$$

- In short, $SS_{extra} = SSM_F - SSM_R$.

Extra sum of squares, cont.

- Recall $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $SST = SSM + SSE$.
- In particular, for the two models, we have

$$SSM(x_1, \ldots, x_{p-1}) = SST - SSE(x_1, \ldots, x_{p-1})$$
$$SSM(x_1, \ldots, x_{q-1}) = SST - SSE(x_1, \ldots, x_{q-1})$$

- Therefore

$$\begin{aligned} & SSM(x_q, \ldots, x_{p-1} | x_1, \ldots, x_{q-1}) \\ & = SSE(x_1, \ldots, x_{q-1}) - SSE(x_1, \ldots, x_{p-1}) \\ & = \begin{array}{c}\text{error variation}\\ \text{of reduced model}\end{array} - \begin{array}{c}\text{error variation}\\ \text{of full model}\end{array} \end{aligned}$$

- In short, $SS_{extra} = SSE_R - SSE_F$.

Distributions

- To determine if the extra sum of squares is significant, we test the hypothesis

$$H_0 : \beta_q = \cdots = \beta_{p-1} = 0$$

- It can be shown that, if H_0 is true, then

$$SSE_R/\sigma^2 \sim \chi^2_{n-q} \quad SSE_F/\sigma^2 \sim \chi^2_{n-p} \quad SS_{extra}/\sigma^2 \sim \chi^2_{p-q}$$

- Further, $\dfrac{SSE_F}{\sigma^2}$ and $\dfrac{SS_{extra}}{\sigma^2}$ are independent.

The F statistic

- It follows from the results of the previous slide that, if H_0 is true, then

$$\frac{\frac{SS_{extra}}{\sigma^2}/(p-q)}{\frac{SSE_F}{\sigma^2}/(n-p)} = \frac{SS_{extra}/(p-q)}{SSE_F/(n-p)} \sim F_{p-q,n-p}$$

- This statistic is used to test the hypothesis $H_0 : \beta_q = \cdots = \beta_{p-1} = 0$

Remarks

- In the simple linear regression model it is $p = 2$, $q = 1$. We therefore recover the $F_{1,n-2}$ test introduced earlier.

- This test therefore compare the two models:

  reduced model (null model): $\quad y = \beta_0 + \varepsilon$,

  full model: $\quad y = \beta_0 + \beta_1 x + \varepsilon$.

Remarks, cont.

- In the multiple linear regression model, if $q = 1$, then we recover the $F_{p-1,n-p}$ test introduced in Section 4.5.
- This test therefore compares the two models:

  reduced model (null model): $\quad y = \beta_0 + \varepsilon$,

  full model: $\quad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$.

Remarks, cont.

- For the null model $y = \beta_0 + \varepsilon$, we always have $SSE_R = SST$.
- To find the SST, we can always use the following R commands:

```
> null_model=lm(y ~ 1)
> anova(null_model)
```

Then the SSE in this anova output is equal to the SST.

Example: Salary data

Education (x_1)	Experience (x_2)	Age (x_3)	Salary (y)
8	12	42	95
13	19	48	145
16	20	51	164
14	24	53	186
15	30	61	197
11	16	46	139
13	19	50	163
15	15	45	125
16	21	56	133
13	26	58	173
15	18	55	143
9	5	35	75
13	19	48	165
13	22	52	167
14	28	57	187

- We firstly fit a multiple regression model (Mod123)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Example: Salary data, R calculation

- Multiple linear regression model with variables x_1, x_2, x_3.

```
> T=read.table("salary2.txt")
> x1=T[,1]
> x2=T[,2]
> x3=T[,3]
> y=T[,4]
> Mod123 <- lm(y ~ x1 + x2 + x3)
> summary(Mod123)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.028     43.545   2.688 0.021123 *
x1             2.469      1.974   1.251 0.237025
x2             7.305      1.503   4.861 0.000502 ***
x3            -2.820      1.558  -1.810 0.097707 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 11 degrees of freedom
Multiple R-squared:  0.9052,    Adjusted R-squared:  0.8794
F-statistic: 35.02 on 3 and 11 DF,  p-value: 6.369e-06
```

Example: Salary data, SST

- We fit a null model,

```
> null_model=lm(y ~ 1)
> anova(null_model)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 14  16038  1145.5
```

The SSE for null model is 16038. Therefore, $SST = 16038$.

Example: Salary data

- The ANOVA analysis for comparing Mod123 with null model,

```
> anova(null_model, Mod123)

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     14 16037.7
2     11  1519.9  3     14518 35.024 6.369e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Since F-test for the model with x_1, x_2, x_3 has a very small p-value, the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is rejected.
- Thus the three predictors together are significant. We ask if we can delete any of them.

Example: Salary data

- We compare the model (Mod123)
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the following models:
  - Deleting x_1 (Mod23): $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
  - Deleting x_2 (Mod13): $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$
  - Deleting x_3 (Mod12): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - Deleting x_2, x_3 (Mod1): $y = \beta_0 + \beta_1 x_1 + \varepsilon$
  - Deleting x_1, x_3 (Mod2): $y = \beta_0 + \beta_2 x_2 + \varepsilon$
  - Deleting x_1, x_2 (Mod3): $y = \beta_0 + \beta_3 x_3 + \varepsilon$

Example: Salary data

- Mod123: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$,

```
> anova(Mod123)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 6224.5  6224.5 45.0487 3.323e-05 ***
x2         1 7840.8  7840.8 56.7471 1.152e-05 ***
x3         1  452.6   452.6  3.2753   0.09771 .
Residuals 11 1519.9   138.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SSE_F = 1519.9$ with d.f.$= n - p = 15 - 4 = 11$.

Example: Salary data

- Mod23: $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$,

```
> Mod23 <- lm(y ~ x2 + x3)
> anova(Mod23)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x2         1 14038.4 14038.4 97.0401 4.211e-07 ***
x3         1   263.3   263.3  1.8204    0.2022
Residuals 12  1736.0   144.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 1736 - 1519.9 = 216.1$
  with d.f.$= p - q = 4 - 3 = 1$.
- $F = \frac{SS_{extra}/(4-3)}{SSE_F/(15-4)} = 1.564 \sim F_{1,11}$

Example: Salary data

- Mod13: $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$,

```
> Mod13 <- lm(y ~ x1 + x3)
> anova(Mod13)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 6224.5  6224.5   15.61 0.001925 **
x3         1 5028.2  5028.2   12.61 0.003989 **
Residuals 12 4785.1   398.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 4785.1 - 1519.9 = 3265.2$ with d.f.$= p - q = 4 - 3 = 1$.
- $F = \frac{SS_{extra}/(4-3)}{SSE_F/(15-4)} = 23.631 \sim F_{1,11}$

Example: Salary data

- Mod12: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$,

```
> Mod12 <- lm(y ~ x1 + x2)
> anova(Mod12)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 6224.5  6224.5  37.868 4.918e-05 ***
x2         1 7840.8  7840.8  47.702 1.636e-05 ***
Residuals 12 1972.4   164.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 1972.4 - 1519.9 = 452.5$ with d.f.$= p - q = 4 - 3 = 1$.
- $F = \frac{SS_{extra}/(4-3)}{SSE_F/(15-4)} = 3.2753 \sim F_{1,11}$

Example: Salary data

- Mod1: $y = \beta_0 + \beta_1 x_1 + \varepsilon$,

```
> Mod1 <- lm(y ~ x1)
> anova(Mod1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x1         1 6224.5  6224.5  8.2458 0.0131 *
Residuals 13 9813.3   754.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 9813.3 - 1519.9 = 8293.4$ with d.f.$= p - q = 4 - 2 = 2$.
- $F = \frac{SS_{extra}/(4-2)}{SSE_F/(15-4)} = 30.011 \sim F_{2,11}$

Example: Salary data

- Mod2 $y = \beta_0 + \beta_2 x_2 + \varepsilon$,

```
> Mod2 <- lm(y ~ x2)
> anova(Mod2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x2         1 14038.4 14038.4   91.28 3.047e-07 ***
Residuals 13  1999.3   153.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 1999.3 - 1519.9 = 479.4$ with d.f.$= p - q = 4 - 2 = 2$.
- $F = \frac{SS_{extra}/(4-2)}{SSE_F/(15-4)} = 1.735 \sim F_{2,11}$

Example: Salary data

- Mod3 $y = \beta_0 + \beta_3 x_3 + \varepsilon$,

```
> Mod3 <- lm(y ~ x3)
> anova(Mod3)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x3         1 11238.6 11238.6  30.443 9.915e-05 ***
Residuals 13  4799.1   369.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $SS_{extra} = SSE_R - SSE_F = 4799.1 - 1519.9 = 3279.2$ with d.f.$= p - q = 4 - 2 = 2$.
- $F = \frac{SS_{extra}/(4-2)}{SSE_F/(15-4)} = 11.867 \sim F_{2,11}$

Example: Salary data

Model	SSE	SS_{extra}	F	Critical Region
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	1519.9			
$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	1736.0	216.1	$F_{1,11} = 1.564$	$(4.844, +\infty)$
$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$	4785.1	3265.2	$F_{1,11} = 23.631$	$(4.844, +\infty)$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	1972.4	452.5	$F_{1,11} = 3.2753$	$(4.844, +\infty)$
$y = \beta_0 + \beta_1 x_1 + \varepsilon$	9813.3	8293.4	$F_{2,11} = 30.011$	$(3.982, +\infty)$
$y = \beta_0 + \beta_2 x_2 + \varepsilon$	1999.3	479.4	$F_{2,11} = 1.735$	$(3.982, +\infty)$
$y = \beta_0 + \beta_3 x_3 + \varepsilon$	4799.1	3279.2	$F_{2,11} = 11.867$	$(3.982, +\infty)$

Example: Salary data, model comparison in R

- We can achieve the same result in R by comparing the pairs of model as follows:

```
> anova(Mod23, Mod123)
Analysis of Variance Table

Model 1: y ~ x2 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     12 1736.0
2     11 1519.9  1     216.1 1.564  0.237

> anova(Mod13, Mod123)
Analysis of Variance Table

Model 1: y ~ x1 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     12 4785.1
2     11 1519.9  1    3265.2 23.631 0.0005017 ***
```

Example: Salary data, model comparison in R

- We can achieve the same result in R by comparing the pairs of model as follows:

```
> anova(Mod12, Mod123)
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     12 1972.4
2     11 1519.9  1    452.56 3.2753 0.09771 .

> anova(Mod1, Mod123)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 9813.3
2     11 1519.9  2    8293.4 30.011 3.507e-05 ***
```

Example: Salary data, model comparison in R

- We can achieve the same result in R by comparing the pairs of model as follows:

```
> anova(Mod2, Mod123)
Analysis of Variance Table

Model 1: y ~ x2
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     13 1999.3
2     11 1519.9  2    479.45 1.735 0.2214

> anova(Mod3, Mod123)
Analysis of Variance Table

Model 1: y ~ x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     13 4799.1
2     11 1519.9  2    3279.2 11.867 0.001793 **
```

Example: Salary data

- From the above comparisons, we conclude only variable 'Experience (x_2)' is significant.
- We conclude that we can delete variables 'Education (x_1)' and 'Age (x_3)' from the model.

R: sequential sum of squares

- The SeqSS for a predictor measures the increase in the regression sum of squares when the predictor is added to a model involving only the predictors listed before it.

- Thus if we have a model with variables $(x_1, x_2, x_3, \ldots, x_{p-1})$ the SeqSS lists the extra sum of squares that compare:

  (null model) with (model with x_1)

  (model with x_1) with (model with x_1, x_2)

  (model with x_1, x_2) with (model with x_1, x_2, x_3)

  ...

  (model with $x_1, \ldots, x_{p-2}$) with (model with $x_1, \ldots, x_{p-1}$)

- Note: if the predictors are entered in a different order (e.g. $x_3, x_1, x_2, \ldots$) then SeqSS changes accordingly.

R: sequential sum of squares, cont.

- In the example 'Salary data',
  Mod123: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

```
> Mod123 <- lm(y ~ x1 + x2 + x3)
> anova(Mod123)

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 6224.5  6224.5 45.0487 3.323e-05 ***
x2         1 7840.8  7840.8 56.7471 1.152e-05 ***
x3         1  452.6   452.6  3.2753   0.09771 .
Residuals 11 1519.9   138.2
```

$$6224.5 = SSE\begin{pmatrix}\text{null model}\\ y = \beta_0 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_1 x_1 + \varepsilon\end{pmatrix}$$

$$7840.8 = SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_1 x_1 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\end{pmatrix}$$

$$452.6 = SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon\end{pmatrix}$$

R: sequential sum of squares, cont.

- Recall that $SSE\begin{pmatrix}\text{null model}\\ y=\beta_0+\varepsilon\end{pmatrix}=SST$. We have

$$\begin{aligned}
14517.9 &= 6224.5 + 7840.8 + 452.6\\
&= SSE\begin{pmatrix}\text{null model}\\ y=\beta_0+\varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon\end{pmatrix}\\
&= SST - SSE\begin{pmatrix}\text{model}\\ y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon\end{pmatrix}\\
&= SSM\begin{pmatrix}\text{model}\\ y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon\end{pmatrix}
\end{aligned}$$

- Therefore, sum of all SeqSS is equal to SSM.

R: sequential sum of squares, cont.

- In the same dataset, we can enter the predictors in a different order

  Mod312: $y = \beta_0 + \beta_3 x_3 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, then we obtain:

```
> Mod312 <- lm(y ~ x3 + x1 + x2)
> summary(Mod312)

Call:
lm(formula = y ~ x3 + x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.028     43.545   2.688 0.021123 *
x3            -2.820      1.558  -1.810 0.097707 .
x1             2.469      1.974   1.251 0.237025
x2             7.305      1.503   4.861 0.000502 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 11 degrees of freedom
Multiple R-squared:  0.9052,    Adjusted R-squared:  0.8794
F-statistic: 35.02 on 3 and 11 DF,  p-value: 6.369e-06
```

R: sequential sum of squares, cont.

- Then the ANOVA and SeqSS

```
> anova(Mod312)
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x3         1 11238.6 11238.6 81.3380 2.054e-06 ***
x1         1    14.0    14.0  0.1016 0.7558953
x2         1  3265.2  3265.2 23.6315 0.0005017 ***
Residuals 11  1519.9   138.2
```

$$11238.6 = SSE\begin{pmatrix}\text{null model}\\ y = \beta_0 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \varepsilon\end{pmatrix}$$

$$14.0 = SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \beta_1 x_1 + \varepsilon\end{pmatrix}$$

$$3265.2 = SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \beta_1 x_1 + \varepsilon\end{pmatrix} - SSE\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\end{pmatrix}$$

$$SSM\begin{pmatrix}\text{model}\\ y = \beta_0 + \beta_3 x_3 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\end{pmatrix} = 11238.6 + 14 + 3265.2 = 14517.8$$

Summary

- We want to compare the full model with predictors $x_1, \ldots, x_{p-1}$ to the reduced model with predictors $x_1, \ldots, x_{q-1}$ where $q < p$.
- The extra sum of squares is $SS_{extra} = SSM_F - SSM_R = SSE_R - SSE_F$.
- $F = \frac{SS_{extra}/(p-q)}{SSE_F/(n-p)} \sim F_{p-q,n-p}$ is used to test the hypothesis $H_0 : \beta_q = \cdots = \beta_{p-1} = 0$.
- SeqSS in R.

Chapter 5: One Way ANOVA

Dr.Ting Wei

Section 5.1: One way ANOVA, part 1

Discussion

- Consider the following experiments:

  a) A clinical trials tests the effectiveness of a drug as explained by the dose and the age of the patients.

  b) A survey of consumer preferences assess a new product according to weather customer ever bought it and how they rate it on a scale of 1 (bad) to 5 (excellent).

  c) A financial institution investigates income in a population according to 4 different areas of a city and political preference (conservative/liberal).

- Do you notice a difference in the explanatory variables in these examples, and if so, what is it?

Categorical variables

- Sometimes we need to use qualitative or categorical variables as explanatory variables in regression.
- These have no natural scale or measurement, and we must assign a set of levels to account for the effect on the response. Eg. employment status (employed/unemployed), sex (male/female), etc.
- Notice that categorical explanatory variables can be either numerical (e.g. 1,2) of nominal (e.g. conservative/liberal).

Assumptions for one-way ANOVA

- The data for one-way ANOVA consist of k samples coming from normal populations

  Sample 1 : $y_{11}, y_{21}, \ldots, y_{n_1 1} \sim iid\ N(\mu_1, \sigma^2)$

  $\vdots$

  Sample k : $y_{1k}, y_{2k}, \ldots, y_{n_k k} \sim iid\ N(\mu_k, \sigma^2)$

- We assume that the samples are normal, the population variances σ^2 are equal and the samples are independent.

- The total number of observations is $N = \sum_{j=1}^{k} n_j$.

Factors in one-way ANOVA

- In one-way ANOVA, experimental units are assigned to various levels of single experimental factor.
- Example: patients are divided into k groups and each group is given a different drug.
  There is a single factor "Drug" which has k different levels (Drug 1, Drug 2, ..., Drug k).

One-way ANOVA in regression form

- We wrote the one-way ANOVA model as:

  Sample 1 : $y_{11}, y_{21}, \ldots, y_{n_1 1} \sim iid\ N(\mu_1, \sigma^2)$

  $\vdots$

  Sample k : $y_{1k}, y_{2k}, \ldots, y_{n_k k} \sim iid\ N(\mu_k, \sigma^2)$

- An equivalent way to write this model is

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

  where $\mu_1, \ldots, \mu_k$ are fixed unknown parameters and ε_{ij} are independent normal random variable $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- μ_j are called cell means and represent the expected response in group j.

Another way to express the model

- Another common way to write the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

  where $\mu, \alpha_1, \ldots, \alpha_k$ are unknown constants and ε_{ij} are *iid* $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- This is the default ANOVA model in R software and is often used in statistical analysis.

- The model as written is not identified, as there are k groups and $k + 1$ parameters. Hence we need to introduce constraints among the α's.

Parameter constraints

- The usual parameter constraint for the one-way ANOVA is

  $$\alpha_1 = 0$$

  which is the default constraint in R software.

- In this case, 1 becomes the reference level, μ the expected response in that level, and α_j is the effect of level j of the factor compared to the reference level.

One-way ANOVA in matrix form

- We firstly consider the model
  $y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k.$
- Stack the response from the k samples together to form a single vector $\mathbf{Y}$ of length $N = \sum_{j=1}^{k} n_j$:

$$\mathbf{Y} = (y_{11}, \cdots, y_{n_1 1}, \cdots \cdots, y_{1k}, \cdots, y_{n_k k})^T$$

- The error vector is

$$\boldsymbol{\varepsilon} = (\varepsilon_{11}, \cdots, \varepsilon_{n_1 1}, \cdots \cdots, \varepsilon_{1k}, \cdots, \varepsilon_{n_k k})^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I_N})$$

- The parameter vector is

$$\boldsymbol{\beta} = (\mu_1, \ldots, \mu_k)^T$$

One-way ANOVA in matrix form, cont.

- Let $\mathbf{X}$ be the $N \times k$ matrix whose column $\mathbf{x}_j$ has 1's in all positions corresponding to the observations in sample j and 0's elsewhere:

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

  where the $\mathbf{x}_j$ are called indicators vectors.

- The one-way ANOVA model in matrix form is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least squares estimates

- From Section 4.1 the least squares estimates of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- From the form of $\mathbf{X}$ one easily calculates

$$(\mathbf{X}^T\mathbf{X}) = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k \end{bmatrix}$$

Least squares estimates, cont.

- We deduce

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix}$$

- Let $y_{+j} = \sum_{i=1}^{n_j} y_{ij}$ and $\bar{y}_j = \frac{y_{+j}}{n_j}$.
- Then

$$\mathbf{X}^T\mathbf{Y} = (y_{+1}, y_{+2}, \ldots, y_{+k})^T$$

Least squares estimates, cont.

- In conclusion the LSE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix} \begin{bmatrix} y_{+1} \\ y_{+2} \\ \vdots \\ y_{+k} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix}$$

Least squares estimates, cont.

- From Section 4.1, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{VarCov}\,\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k} \end{bmatrix}$$

Another way to express the model

- We want to write the one-way ANOVA model in the form

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

- Since $\alpha_1 = 0$, we have

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_k \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ 0 \\ \hat{\mu}_2 - \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k - \hat{\mu}_1 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ 0 \\ \bar{y}_2 - \bar{y}_1 \\ \vdots \\ \bar{y}_k - \bar{y}_1 \end{bmatrix}$$

Example: Kenton Food Company

The Kenton Food Company wished to test four different package designs for a new breakfast cereal. Ten stores, with approximately equal sales volumes, were selected as the experimental units. Each store was randomly assigned one of the package designs.

The stores were chosen to be comparable in location and sales volume. Other relevant conditions that could affect sales were kept the same for all of the stores in the experiment.

Sales, in number of cases, were observed for the study period, and the results are recorded

Example: Kenton Food Company data

(a) Sample Data

Package Design	Store 1	Store 2	Store 3	Total	Mean	Number of Stores
1	12	18		30	15	2
2	14	12	13	39	13	3
3	19	17	21	57	19	3
4	24	30		54	27	2
All designs				180	18	10

(b) Symbolic notation

Factor Level (i)	Sample unit (j) 1	2	3	Total	Mean	Nr. of Sample units
1	y_{11}	y_{21}		y_{+1}	$\bar{y}_1$	n_1
2	y_{12}	y_{22}	y_{32}	y_{+2}	$\bar{y}_2$	n_2
3	y_{13}	y_{23}	y_{33}	y_{+3}	$\bar{y}_3$	n_3
4	y_{14}	y_{24}		y_{+4}	$\bar{y}_4$	n_4
All factor levels				y_{++}	$\bar{y}$	N

Example: Kenton Food Company, cont.

- For this example, the response vector is

$$\mathbf{Y} = (12, 18, 14, 12, 13, 19, 17, 21, 24, 30)^T$$

- The error vector is

$$\boldsymbol{\varepsilon} = (\varepsilon_{11}, \varepsilon_{21}, \varepsilon_{12}, \varepsilon_{22}, \varepsilon_{32}, \varepsilon_{13}, \varepsilon_{23}, \varepsilon_{33}, \varepsilon_{14}, \varepsilon_{24})$$

- The parameter vector is

$$\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3, \mu_4)^T$$

Example: Kenton Food Company, cont.

- The $\mathbf{X}$ matrix for this example is the 10×4 matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- The model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Example: Kenton Food Company, cont.

- Since $\mathbf{X}\boldsymbol{\beta} = (\mu_1, \mu_1, \mu_2, \mu_2, \mu_2, \mu_3, \mu_3, \mu_3, \mu_4, \mu_4)^T$ the model explicitly is

$$
\begin{aligned}
y_{11} &= \mu_1 + \varepsilon_{11} \\
y_{21} &= \mu_1 + \varepsilon_{21} \\
y_{12} &= \mu_2 + \varepsilon_{12} \\
y_{22} &= \mu_2 + \varepsilon_{22} \\
y_{32} &= \mu_2 + \varepsilon_{32} \\
y_{13} &= \mu_3 + \varepsilon_{13} \\
y_{23} &= \mu_3 + \varepsilon_{23} \\
y_{33} &= \mu_3 + \varepsilon_{33} \\
y_{14} &= \mu_4 + \varepsilon_{14} \\
y_{24} &= \mu_4 + \varepsilon_{24}
\end{aligned}
$$

Example: Kenton Food Company, cont.

- The least squared estimate for the example is

$$\hat{\beta} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{bmatrix} = \begin{bmatrix} 15 \\ 13 \\ 19 \\ 27 \end{bmatrix}$$

- Thus, the mean sales per store with package 1 are estimated to be 15, and the fitted values for the observations for package 1 are $\hat{y}_{11} = \hat{y}_{21} = 15$.

- Similarly, the fitted values for the observations for packages 2, 3, 4 are

$$\hat{y}_{12} = \hat{y}_{22} = \hat{y}_{32} = 13$$
$$\hat{y}_{13} = \hat{y}_{23} = \hat{y}_{33} = 19$$
$$\hat{y}_{14} = \hat{y}_{24} = 27$$

Example: Kenton Food Company, cont.

- Then the model can be written as

$$
\begin{aligned}
y_{11} &= \mu + \alpha_1 + \varepsilon_{11} \\
y_{21} &= \mu + \alpha_1 + \varepsilon_{21} \\
y_{12} &= \mu + \alpha_2 + \varepsilon_{12} \\
y_{22} &= \mu + \alpha_2 + \varepsilon_{22} \\
y_{32} &= \mu + \alpha_2 + \varepsilon_{32} \\
y_{13} &= \mu + \alpha_3 + \varepsilon_{13} \\
y_{23} &= \mu + \alpha_3 + \varepsilon_{23} \\
y_{33} &= \mu + \alpha_3 + \varepsilon_{33} \\
y_{14} &= \mu + \alpha_4 + \varepsilon_{14} \\
y_{24} &= \mu + \alpha_4 + \varepsilon_{24}
\end{aligned}
$$

with point estimates of parameters

$$
\begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\alpha}_4 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ 0 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_3 - \bar{y}_1 \\ \bar{y}_4 - \bar{y}_1 \end{bmatrix} = \begin{bmatrix} 15 \\ 0 \\ -2 \\ 4 \\ 12 \end{bmatrix}
$$

Example: Kenton Food Company, R results.

- We fit the model in R,

```
> Sales=c(12,18,14,12,13,19,17,21,24,30)
> Design=c(1,1,2,2,2,3,3,3,4,4)
> Model=lm(Sales ~ factor(Design))
> summary(Model)

Call:
lm(formula = Sales ~ factor(Design))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       15.000      1.958   7.661 0.000258 ***
factor(Design)2   -2.000      2.528  -0.791 0.458922
factor(Design)3    4.000      2.528   1.583 0.164620
factor(Design)4   12.000      2.769   4.334 0.004908 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.769 on 6 degrees of freedom
Multiple R-squared:  0.8487,Adjusted R-squared:  0.773
F-statistic: 11.22 on 3 and 6 DF,  p-value: 0.007135
```

Summary

- Categorical explanatory variables.
- Different equivalent formulations of one-way ANOVA.
- Parameter constraints.
- One-way ANOVA in matrix form.
- Least squares estimates of the parameters.
- Example: Kenton food company.

Section 5.2: One way ANOVA, part 2

Assumptions for one-way ANOVA

Recall from last time:

- The data for one-way ANOVA consist of k samples coming from normal populations

  Sample 1 : $y_{11}, y_{21}, \ldots, y_{n_1 1} \sim iid\ N(\mu_1, \sigma^2)$

  $\vdots$

  Sample k : $y_{1k}, y_{2k}, \ldots, y_{n_k k} \sim iid\ N(\mu_k, \sigma^2)$

- We assume that the samples are normal, the population variances σ^2 are equal and the samples are independent.

- The total number of observations is $N = \sum_{j=1}^{k} n_j$.

Model equation for one-way ANOVA

Recall from Section 5.1:

- An equivalent way to write this model is

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

  where $\mu_1, \ldots, \mu_k$ are fixed unknown parameters and ε_{ij} are independent normal random variable $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Another common way to write the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

  where $\alpha_1 = 0$, and ε_{ij} are *iid* $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- We also discussed the matrix form of the model equation.

Parameter estimates for one-way ANOVA

Recall from Section 5.1:

- Using the matrix form of the model and the general methodology for the general linear model, we proved that the parameter estimates are given by

$$\hat{\mu}_j = \bar{y}_j = \frac{y_{+j}}{n_j} \qquad j = 1, \ldots, k$$

Example: Kenton Food Company data

(a) Sample Data

Package Design	Store 1	Store 2	Store 3	Total	Mean	Number of Stores
1	12	18		30	15	2
2	14	12	13	39	13	3
3	19	17	21	57	19	3
4	24	30		54	27	2
All designs				180	18	10

(b) Symbolic notation

Factor Level (i)	Sample unit (j) 1	2	3	Total	Mean	Number of Sample units
1	y_{11}	y_{21}		y_{+1}	$\bar{y}_1$	n_1
2	y_{12}	y_{22}	y_{32}	y_{+2}	$\bar{y}_2$	n_2
3	y_{13}	y_{23}	y_{33}	y_{+3}	$\bar{y}_3$	n_3
4	y_{14}	y_{24}		y_{+4}	$\bar{y}_4$	n_4
All factor levels				y_{++}	$\bar{y}$	N

Main question

- Are the differences in mean sales between the package designs due to chance or are they statistically significant?
- We want to test the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Methodology

- To test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = \mu$, we can compare models

$$\begin{aligned} \text{Null model: } & y_{ij} = \mu + \varepsilon_{ij} \\ \text{One-way model: } & y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \end{aligned}$$

  for $i = 1, \ldots, n_j$ and $j = 1, \ldots, k$.

- Both models differ by α_j. If the null model is a better fit, we can conclude $\alpha_j = 0$, which indicates H_0 is true.

- To write the corresponding F statistics, we will split the total variation of the bigger model as

$$SST = SSB + SSE$$

Total variation

- The total variability of the y_{ij} observations, not using any information about factor levels, is measured in terms of the deviation of each observation y_{ij} around the overall mean $\bar{y}$, where

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}$$

- Hence

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

Partitioning the total sum of squares

- SSB (between groups variation) is a measure of the extent of differences between factor level sample means, based on the deviation of the factor level sample means $\bar{y}_j$ around the overall mean $\bar{y}$.

$$SSB = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

- SSW (within group variation) is a measure of the random variation of the observations around the respective factor level sample means.

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

- As shown in Section 3.9, we can partition the total variation as

$$SST = SSB + SSW$$

Partitioning the total sum of squares, cont.

- SSE (error variation) is a measure of the random variation caused by noise.

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2$$

- From Section 5.1, we derived $\hat{y}_{ij} = \bar{y}_j$, $i = 1, \cdots, n_j$, and then we have

$$SSE = SSW$$

- Hence, the partition of the total variation is

$$SST = SSB + SSE$$

Some computationally useful formulas

- We introduce the correction factor for the mean

$$CF = \frac{(\text{Sum of all observations})^2}{N} = \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij})^2}{N}$$

- Since $\bar{y} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$, we have $\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = N\bar{y}$.

- Hence

$$CF = \frac{\bar{y}^2 N^2}{N} = N\bar{y}^2$$

Some computationally useful formulas, cont.

- One can show

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}^2 - CF$$

and

$$SSB = \sum_{j=1}^{k} n_j \bar{y}_j^2 - CF$$

The F statistic for model comparison

Recall from Section 4.6:

- The F statistics for comparing two nested general linear models: a reduced model with q parameters and a full model with p parameters is

$$\frac{SS_{extra}/(p-q)}{SSE_F/(n-p)} \sim F_{p-q,n-p}$$

where $SS_{extra} = SSE_R - SSE_F$.

The F statistic

- In our case the two models are

$$y_{ij} = \mu + \varepsilon_{ij} \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad \alpha_1 = 0, \quad i = 1, \ldots, n_j \quad j = 1, \ldots, k$$

- Therefore $q = 1, p = k, n = N$ and $SSE_R = SST$, therefore

$$SS_{extra} = SSE_R - SSE_F = SST - SSE_F = SSB$$

The F statistic, cont.

- Define the mean squares:

$$MSB = \frac{SSB}{k-1}, \qquad MSE = \frac{SSE}{N-k}$$

- Then, if $\mu_1 = \cdots = \mu_k = \mu$ the quantity

$$F = \frac{MSB}{MSE} \sim \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{N-k}/(N-k)} \sim F_{k-1,N-k}$$

  has a $F_{k-1,N-k}$ distribution.

- This can be used to test the hypothesis
  $H_0 : \mu_1 = \cdots = \mu_k = \mu$ against H_1 : at least one μ_j different.

ANOVA table

- We can summarize all the relevant information in the following ANOVA table:

	Source	df	SS	MS	F
SSB	Between Groups	$k-1$	$\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$	$\frac{SSB}{(k-1)}$	$F_{k-1,N-k}$
SSE	Error (Within Groups)	$N-k$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$	$\frac{SSE}{(N-k)}$	
SST	Total	$N-1$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$		

Example: Kenton Food Company data

(a) Sample Data

Package Design	Store 1	Store 2	Store 3	Total	Mean	Number of Stores
1	12	18		30	15	2
2	14	12	13	39	13	3
3	19	17	21	57	19	3
4	24	30		54	27	2
All designs				180	18	10

Example: Kenton Food Company

- For the Kenton Food Company example of Section 5.1, we have $\bar{y}_1 = 15$, $\bar{y}_2 = 13$, $\bar{y}_3 = 19$, $\bar{y}_4 = 27$, $\bar{y} = 18$, $N = 10$.
- Therefore $CF = N\bar{y}^2 = 10 \times 18^2 = 3240$

Example: Kenton Food Company, cont.

- We calculate

$$
\begin{aligned}
SST &= \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}^2 - CF \\
&= 12^2 + 18^2 + 14^2 + 12^2 + 13^2 + \\
&\quad + 19^2 + 17^2 + 21^2 + 243^2 + 30^2 - 3240 = 304
\end{aligned}
$$

$$
\begin{aligned}
SSB &= \sum_{j=1}^{k} n_j \bar{y}_j^2 - CF \\
&= 2 \times 15^2 + 3 \times 13^2 + 3 \times 19^2 + 2 \times 27^2 - 3240 = 258
\end{aligned}
$$

Example: Kenton Food Company, cont.

- We have

$$SSE = SST - SSB = 304 - 258 = 46$$

- We see that much of the total variation is associated with variation between the factor level sample means.

Example: Kenton Food Company, cont.

- The mean squares are:

$$MSB = \frac{SSB}{k-1} = \frac{258}{3} = 86, \qquad MSE = \frac{SSE}{N-k} = \frac{46}{6} = 7.67$$

- Therefore

$$F = \frac{MSB}{MSE} = \frac{86}{7.67} = 11.2 \sim F_{3,6}$$

- The critical region is $(4.757, +\infty)$. Hence we reject the null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4$. That is, at least one package designs has a statistically significant effect on the mean sales volume.

Example: Kenton Food Company, cont.

- R results:

```
> Sales=c(12,18,14,12,13,19,17,21,24,30)
> Design=c(1,1,2,2,2,3,3,3,4,4)
> Model=lm(Sales ~ factor(Design))
> anova(Model)

Analysis of Variance Table

Response: Sales
               Df Sum Sq Mean Sq F value   Pr(>F)
factor(Design)  3    258  86.000  11.217 0.007135 **
Residuals       6     46   7.667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary

- We split the total variation SST for the one-way ANOVA model as $SST = SSB + SSE$ where SSB is the between group variation and SSE is the error variation.
- We defined the mean squares MSB and MSE and the F-statistic to test the hypothesis $H_0 : \mu_1 = \cdots = \mu_k = \mu$.

Section 5.3: Contrasts

Motivation

- We discussed the F test for determining whether or not the factor level means μ_j differ.
- When this test leads to the acceptance of the hypothesis that the factor level means μ_j are equal, the ANOVA model is inappropriate.
- Hence no relationship exists between the factors and the response variable, and usually no further analysis of factor means is therefore indicated.

Motivation, cont.

- On the other hand, when the F test leads to the conclusion that the factor level means differ, a relationship between the factor and the response variable is present.
- In this latter case, a thorough analysis of the nature of the factor level effects is usually performed. The technique of contrasts precisely achieves this.

Example: Kenton Food Company data

(a) Sample Data

Package Design	Store 1	Store 2	Store 3	Total	Mean	Number of Stores
1	12	18		30	15	2
2	14	12	13	39	13	3
3	19	17	21	57	19	3
4	24	30		54	27	2
All designs				180	18	10

Discussion

- Last time we rejected the null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4$. That is, the four different package designs do not lead to the same mean sales volume.
- Suppose we are interested in understanding the difference (if any) between package 1 and 2. How would you phrase this in terms of a hypothesis test?

Answer to discussion question

- We want to test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$.
- The means difference $\mu_1 - \mu_2$ is an example of a contrast.

Definition of contrast

- A contrast is a comparison involving two or more factor level means.
- A contrast $\mathcal{C}$ is defined as a linear combination of the factor level means μ_j where the coefficients c_j sum to zero:

$$\mathcal{C} = \sum_{j=1}^{k} c_j \mu_j \quad \text{with} \quad \sum_{j=1}^{k} c_j = 0$$

Illustrations of contrasts

- We illustrate some contrasts by reference to the Kenton Food Company example.
- In that example, package designs 1 and 2 used 3-color printing and designs 3 and 4 used 5-color printing.
- In addition, package designs 1 and 3 utilized cartoons while no cartoons were utilized in designs 2 and 4.

Illustrations of contrasts, cont.

- Comparison of the mean sales for the two 3-color designs:

  $$\mathcal{C} = \mu_1 - \mu_2$$

  Here $c_1 = 1$, $c_2 = -1$, $c_3 = c_4 = 0$.

- This contrast is simply a pairwise comparison.

Illustrations of contrasts, cont.

- Comparison of the mean sales for the 3-color and 5-color designs:

$$\mathcal{C} = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

Here $c_1 = c_2 = \frac{1}{2}, \quad c_3 = c_4 = -\frac{1}{2}.$

- We compare averages of several factor level means.

Illustrations of contrasts, cont.

- Comparison of the mean sales for designs with and without cartoons:

$$\mathcal{C} = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

Here $c_1 = c_3 = \frac{1}{2}, \quad c_2 = c_4 = -\frac{1}{2}.$

- Again we compare averages of several factor level means.

Confidence interval for $\mathcal{C}$

- An unbiased estimator of $\mathcal{C}$ is:

$$\hat{\mathcal{C}} = \sum_{j=1}^{k} c_j \bar{y}_j$$

- Since the $\bar{y}_j$ are independent, the variance of $\hat{\mathcal{C}}$ is:

$$\text{var}(\hat{\mathcal{C}}) = \sum_{j=1}^{k} c_j^2 \,\text{var}(\bar{y}_j) = \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

- An unbiased estimation of this variance is:

$$s^2(\hat{\mathcal{C}}) = \hat{\sigma}^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j} = MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Confidence interval for $\mathcal{C}$, cont.

- $\hat{\mathcal{C}}$ is normally distributed because it is a linear combination of independent normal random variables.
- It can be shown that for the one-way ANOVA model, $y_{ij} = \mu_j + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim N(0, \sigma^2)$:

$$\frac{\hat{\mathcal{C}} - \mathcal{C}}{s(\hat{\mathcal{C}})} \sim t_{N-k}$$

- The 95% C.I. for $\mathcal{C}$ is

$$\left(\hat{\mathcal{C}} - t_{0.025,N-k}\, s(\hat{\mathcal{C}}),\, \hat{\mathcal{C}} + t_{0.025,N-k}\, s(\hat{\mathcal{C}})\right)$$

Example: Kenton Food Company

- We wish to estimate $\mathcal{C} = \mu_1 - \mu_2$ and find the 95% C.I. .
- $\hat{\mathcal{C}} = \bar{y}_1 - \bar{y}_2 = 15 - 13 = 2$

  $$s^2(\hat{\mathcal{C}}) = MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j} = 7.67 \left(\frac{1}{2} + \frac{1}{3} \right) = 6.39$$

  $$s(\hat{\mathcal{C}}) = 2.53$$
- Since $t_{0.025,6} = 2.447$, the 95% C.I. for $\mathcal{C}$ is:

$$\begin{aligned} &\left(\hat{\mathcal{C}} - t_{0.025,6} s(\hat{\mathcal{C}}),\ \hat{\mathcal{C}} + t_{0.025,6} s(\hat{\mathcal{C}}) \right) \\ &\qquad = (2 - 2.447 \times 2.53,\ 2 + 2.447 \times 2.53) \\ &\qquad = (-4.19,\ 8.19) \end{aligned}$$

Example: Kenton Food Company, cont.

- We wish to estimate $\mathcal{C} = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$ and find the 95% C.I. .
- The point estimate of $\mathcal{C}$ is:

$$\hat{\mathcal{C}} = \frac{\bar{y}_1 + \bar{y}_2}{2} - \frac{\bar{y}_3 + \bar{y}_4}{2} = \frac{15 + 13}{2} - \frac{19 + 27}{2} = -9$$

Example: Kenton Food Company, cont.

- Since $c_1 = c_2 = \frac{1}{2}$ and $c_3 = c_4 = -\frac{1}{2}$ we obtain:

$$\sum_{j=1}^{k} \frac{c_j^2}{n_j} = \frac{(0.5)^2}{2} + \frac{(0.5)^2}{3} + \frac{(-0.5)^2}{3} + \frac{(-0.5)^2}{2} = \frac{5}{12}$$

- $s^2(\hat{\mathcal{C}}) = MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j} = 7.67 \times \frac{5}{12} = 3.196$

$s(\hat{\mathcal{C}}) = 1.79$

Example: Kenton Food Company, cont.

- Since $t_{0.025,6} = 2.447$ we conclude that the 95% C.I. for $\mathcal{C}$ is

$$\begin{aligned}\left(\hat{\mathcal{C}} - t_{0.025,6}s(\hat{\mathcal{C}}),\, \hat{\mathcal{C}} + t_{0.025,6}s(\hat{\mathcal{C}})\right) \\ &= (-9 - 2.447 \times 1.79,\, -9 + 2.447 \times 1.79) \\ &= (-13.38,\, -4.62)\end{aligned}$$

- Therefore, we can conclude that the mean sales for the 3-color design fall below those for the 5-color design by somewhere between 4.62 and 13.38 cases per store.

Summary

- Contrasts are used to study the relationship between the factors and the response variable when the F-test establishes that such a relationship exists.
- Definition of contrast $\mathcal{C}$.
- Point estimate, standard error and 95% C.I. for $\mathcal{C}$.
- Example: Kenton Food Company.